SPEAKER ADAPTATION USING COMBINED TRANSFORMATION AND BAYESIAN METHODS

Vassilios Digalakis and Leonardo Neumeyer

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

ABSTRACT

The performance and robustness of a speech recognition system can be improved by adapting the speech models to the speaker, the channel and the task. In continuous mixture-density hidden Markov models the number of component densities is typically very large, and it may not be feasible to acquire a large amount of adaptation data for robust maximum-likelihood estimates. To solve this problem, we propose a constrained estimation technique for Gaussian mixture densities, and combine it with Bayesian techniques to improve its asymptotic properties. We evaluate our algorithms on the large-vocabulary Wall Street Journal corpus for nonnative speakers of American English. The recognition error rate is comparable to the speaker-independent accuracy achieved for native speakers.

1. INTRODUCTION

Two families of adaptation schemes have been proposed in the past. One transforms the speaker's feature space to "match" the space of the training population [1],[2],[3]. The transformation can be applied either directly to the features, or to the speech models [4]. This approach has the advantage of simplicity and, if the number of free parameters is small, then transformation techniques adapt to the user with only a small amount of adaptation speech (quick adaptation). Disadvantages of transformation methods are that they are usually text-dependent and that they may not take full advantage of large amounts of adaptation data. The second main family of adaptation algorithms follows a Bayesian approach, where the speaker-independent information is encapsulated in the prior distributions [5][6]. The Bayesian approach is text-independent, and has nice asymptotic properties: speaker-adaptive performance will converge to speaker-dependent performance as the amount of adaptation speech increases. However, the adaptation rate is usually slow.

In this paper we present adaptation schemes that combine the quick adaptation characteristics of transformation-based methods with the nice asymptotic properties of Bayesian methods. We first introduce a transformation-based method for continuous mixture-density hidden Markov models (HMMs). Adaptation is achieved via a transformation of the speaker-independent observation densities, and the transformation parameters are obtained using the maximum-likelihood (ML) criterion. The number of transformation parameters can be adjusted to achieve quick adaptation. We will then show how this algorithm can be combined with Bayesian techniques. The combined method adapts to a new speaker with small amounts of adaptation data, but also has nice asymptotic properties and takes full advantage of large amounts of adaptation data.

2. TRANFORMATION-BASED ADAPTATION

Transformation-based approaches to speaker adaptation are typically text-dependent and require the new speaker to record some predetermined sentences. These utterances are aligned to ones recorded by reference speakers, and mappings between the new-speaker and the reference-speaker acoustic spaces are obtained using regression techniques [2][3].

We have developed a novel transformation-based approach to speaker adaptation for continuous mixture-density HMMs [7]. We apply the transformation at the distribution level, instead of transforming the feature vectors directly, since we can then use the expectation-maximization (EM) algorithm [8] to estimate the transformation parameters by maximizing the likelihood of the adaptation data. Using this approach, we are not required to time-align the new- and reference-speaker data, and the transformation parameters can be estimated using new-speaker data alone. Our scheme can also be viewed as a constrained estimation of Gaussian mixtures, since we apply the same transformation to all the components of a particular mixture (or a group of mixtures, if there is tying of transformations) instead of independently reestimating them. It achieves quick adaptation by adapting Gaussians for which there were no observations in the training data, based on data that were most likely generated by other Gaussians of the same or neighboring mixtures.

Specifically, we assume that the speaker-independent (SI) HMM model for the SI vector process $[y_t]$ has observation densities of the form

$$p_{SI}(y_t \mid s_t) = \sum_i p(\omega_i \mid s_t) N(y_t; \mu_{ig}, \Sigma_{ig}) \quad , \qquad (1)$$

where g is the index of the Gaussian codebook used by state s_t. Adaptation of this system can be achieved by jointly transforming all the Gaussians of each mixture. We assume that, given the HMM state index s_t, the speaker-dependent vector process $[x_t]$ can be obtained by an underlying process $[y_t]$ through the transformation

$$x_t = A_g y_t + b_g \quad . \tag{2}$$

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

$$p_{SA}(x_t|s_t) = \sum_i p(\omega_i|s_t) N(x_t; A_g\mu_{ig} + b_g, A_g\Sigma_{ig}A_g^T) \tag{3}$$

and only the parameters $A_g, b_g, g = 1, \ldots, N_g$ need to be estimated during adaptation, where N_g is the number of distinct transformations. The same transformations can be applied to different HMM states, and this tying of transformations can be used to optimize performance based on the amount of available adaptation data. The transformation parameters can be estimated using the EM algorithm. The reestimation formulae are derived in [7] and are summarized below:

1. Initialize all transformations with $A_g(0) = I, b_g(0) = 0, g = 1, \ldots, N$. Set *k=0.*
2. **E-step:** Perform one iteration of the forward-backward algorithm on the speech data, using Gaussians transformed with the current value of the transformations $A_g(k), b_g(k)$. For all component Gaussians and all mixtures *g*, collect the sufficient statistics

$$\bar{\mu}_{ig} = \frac{1}{n_{ig}} \sum_{t, s_t} \gamma_t(s_t)\phi_{it}(s_t)x_t$$
$$\bar{\Sigma}_{ig} = \frac{1}{n_{ig}} \sum_{t, s_t} \gamma_t(s_t)\phi_{it}(s_t)(x_t - \bar{\mu}_{ig})(x_t - \bar{\mu}_{ig})^T \tag{4}$$
$$n_{ig} = \sum_{t, s_t} \gamma_t(s_t)\phi_{it}(s_t)$$

   where $\gamma_t(s_t)$ is the probability of being at state s_t at time *t* given the current HMM parameters, the summation is over all times and HMM states that share the same mixture components, and $\phi_{it}(s_t)$ is the posterior probability

$$\phi_{it}(s_t) = p(\omega_{ig} \mid A_g(k), b_g(k), x_t, s_t) \quad . \tag{5}$$

3. **M-step:** Compute the new transformation parameters. Under the assumption of diagonal covariance and transformation matrices, the elements *a* and *b* of $A_g(k+1), b_g(k+1)$ can be obtained by solving the following equations for each *g*

$$b = \left(\sum_i \frac{n_i\bar{\mu}_i}{\sigma_i^2} - a\sum_i \frac{n_i\mu_i}{\sigma_i^2}\right) \Big/ \left(\sum_i \frac{n_i}{\sigma_i^2}\right)$$
$$\left(\sum_i n_i\right)a^2 - \left(\sum_i \frac{n_i}{\sigma_i^2}\right)b^2 - \left(\sum_i \frac{n_i\mu_i}{\sigma_i^2}\right)ab \tag{6}$$
$$+ \left(\sum_i \frac{n_i\bar{\mu}_i\mu_i}{\sigma_i^2}\right)a + \left(2\sum_i \frac{n_i\bar{\mu}_i}{\sigma_i^2}\right)b - \left(\sum_i n_i \frac{\bar{\mu}_i^2 + \bar{\sigma}_i^2}{\sigma_i^2}\right) = 0$$

   where for simplicity we have dropped the dependence on *g.* The variables $\mu_i, \sigma_i^2, \bar{\mu}_i, \bar{\sigma}_i^2$ are elements of the vectors and diagonal matrices $\mu_{ig}, \Sigma_{ig}, \bar{\mu}_{ig}, \bar{\Sigma}_{ig}$, respectively.

4. If the convergence criterion is not met, go to step 2.

Once the transformation parameters are determined, the constrained ML estimates for the means and covariances can be obtained using

$$\mu_{ig}^{CML} = A_g\,\mu_{ig} + b_g$$
$$\Sigma_{ig}^{CML} = A_g\Sigma_{ig}A_g^T \quad . \tag{7}$$

3. COMBINED TRANSFORMATION AND BAYESIAN-BASED ADAPTATION

In Bayesian adaptation techniques the limited amount of adaptation data is optimally combined with the prior knowledge. With the appropriate choice of the prior distributions, the maximum *a posteriori* (MAP) estimates for the means and covariances of HMMs with Gaussian observation densities can be obtained using linear combinations of the speaker-dependent sufficient statistics (counts) and some quantities that depend on the parameters of the prior distributions [5][6]. Based on the reestimation formulae for the MAP estimates of the means and covariances of HMM with continuous mixture densities that are derived in [6], a simplified version of Bayesian estimation can be implemented by linearly combining the speaker-independent and the speaker-dependent counts for each component density

$$\langle x\rangle_{ig}^{SA} = \lambda\langle x\rangle_{ig}^{SI} + (1-\lambda)\langle x\rangle_{ig}^{SD}$$
$$\langle xx^T\rangle_{ig}^{SA} = \lambda\langle xx^T\rangle_{ig}^{SI} + (1-\lambda)\langle xx^T\rangle_{ig}^{SD} \quad , \tag{8}$$
$$n_{ig}^{SA} = \lambda n_{ig}^{SI} + (1-\lambda)n_{ig}^{SD}$$

where the superscripts denote the data over which the following statistics are collected during one iteration of the forward-backward algorithm

$$\langle x\rangle_{ig} = \sum_{t, s} \gamma_t(s)\phi_{it}(s)x_t$$
$$\langle xx^T\rangle_{ig} = \sum_{t, s} \gamma_t(s)\phi_{it}(s)x_t x_t^T \quad . \tag{9}$$
$$n_{ig} = \sum_{t, s} \gamma_t(s)\phi_{it}(s)$$

We will refer to this method as approximate Bayesian adaptation. The weight λ controls the adaptation rate. Using the combined counts, we can compute the approximate MAP (AMAP) estimates of the means and covariances of each Gaussian component density from

$$\mu_{ig}^{AMAP} = \frac{\langle x\rangle_{ig}^{SA}}{n_{ig}^{SA}}$$
$$\Sigma_{ig}^{AMAP} = \frac{\langle xx^T\rangle_{ig}^{SA}}{n_{ig}^{SA}} - \mu_{ig}^{AMAP}(\mu_{ig}^{AMAP})^T \quad . \tag{10}$$

Similar adaptation schemes have also appeared for discrete HMMs [9], and can be used to adapt the mixture weights in the approximate Bayesian scheme described here.

In Bayesian adaptation schemes, only the Gaussians of the speaker-independent models that are most likely to have generated some of the adaptation data will be adapted to the speaker. These Gaussians may represent only a small fraction of the total number in continuous HMMs with a large number of Gaussians. On the other hand, as the amount of adaptation data increases, the speaker-dependent statistics will dominate the speaker-independent priors and Bayesian techniques will approach speaker-dependent performance. We should, therefore, aim for an adaptation scheme that retains the nice properties of Bayesian schemes for large amounts of adaptation data, and has improved performance for small amounts of adaptation data. We can achieve this by using our transformation-based adaptation as a preprocessing step to transform the speaker-independent models so that they better match the new speaker characteristics and improve the prior information in MAP estimation schemes. In the approximate Bayesian adaptation, this can be accomplished by first transforming the speaker-independent counts using the method described in Section 2 and then combining them with the speaker-dependent counts collected using the adaptation data.

4. EXPERIMENTAL RESULTS

We evaluated our adaptation algorithms on the “spoke 3” task of the phase-1, large-vocabulary Wall Street Journal (WSJ) corpus [10], trying to improve recognition performance for non-native speakers of American English. Experiments were carried out using SRI’s DECIPHER^{TM} speech recognition system configured with a six-feature front end that outputs 12 cepstral coefficients, cepstral energy, and their first- and second-order differences. 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 hidden Markov models with an arbitrary degree of Gaussian sharing across different HMM states as described in [11]. The speaker-independent 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 with 32 Gaussian components per codebook. For fast experimentation, we used the progressive search framework [12]: an initial, speaker-independent recognizer with a bigram language model outputs word lattices for all the utterances in the test set. These word lattices are then rescored using speaker-adapted models. We used the baseline 5,000-word, closed-vocabulary bigram and trigram language models provided by the MIT Lincoln Laboratory. The trigram language model was implemented using the N-best rescoring paradigm, by rescoring the list of the N-best sentence hypotheses generated using the bigram language model.

In the first series of experiments we used the bigram language model. We first evaluated the performance of the transformation-based adaptation for various numbers of transformations and amounts of adaptation data. As we can see in Figure 1, where we have plotted the word error rate as a function of the number of adaptation sentences, multiple transformations outperform very constrained schemes that use 1 or 2 transformations. The performance with 20 and 40 transformations is similar, and is better than the less constrained case of 160 transformations. However, as the amount of adaptation data increases, the 160 transformations take advantage of the additional data and outperform the more constrained schemes. A significant decrease in error rate is obtained with as few as 5 adaptation sentences. When adapting using a single sentence, the performance is similar for different numbers of transformations, except for the case of 2 transformations. The reason is that in our implementation a transformation is reestimated only if the number of observations is larger than a threshold; otherwise, we use a global transformation estimated from all data. Since most of the transformations are backed off to the global transformation for the case of a single adaptation sentence, the cases with different numbers of transformations exhibit similar performance.

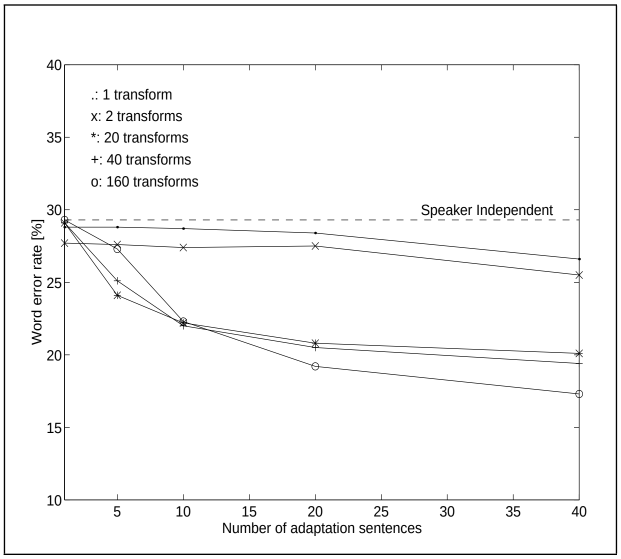

Figure 1: Word error rates for various numbers of transformations for the transformation-based adaptation

In Figure 2 we compare the word error rates of the transformation-only method with 20 and 160 transformations, the approximate Bayesian method with conventional priors, and the combined method for various amounts of adaptation data. In the latter, the number of transformations was optimized according to the available amount of adaptation data. The transformation-only method with 20 transformations outperforms the Bayesian scheme with conventional priors when fewer than 10 sentences are used for adaptation, whereas the situation reverses as more adaptation sentences are used. This is consistent with our claim that transformation-based methods adapt faster, whereas Bayesian schemes have better asymptotic properties. The performance of the transformation approach for large amounts of adaptation data can be improved by increasing the number of transformations. We can also see in the same figure the success of the combined method, which significantly outperforms the first two methods over the whole range of adaptation sentences that we

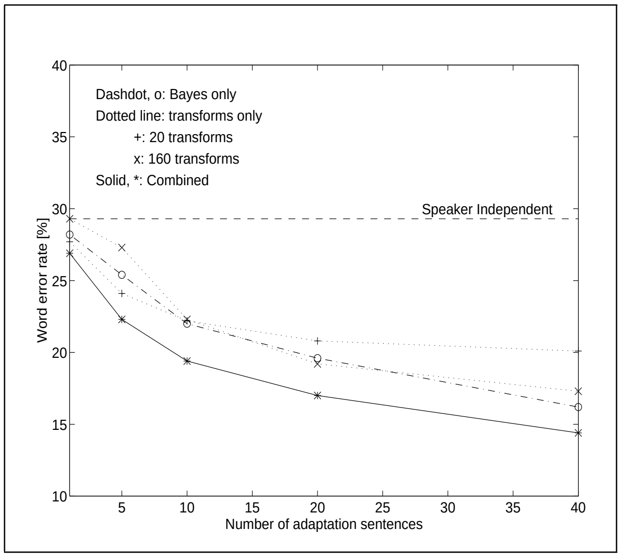

Figure 2: Word error rates for transformation-only, Bayesian-only, and combined schemes.

examined. The transformation step provides quick adaptation when few adaptation sentences are used, and the Bayesian reestimation step improves the asymptotic performance.

Finally, we evaluated the word error rate of our best-performing configuration on the 1993 Spoke-3 development and evaluation sets, and the 1994 evaluation set of the WSJ corpus using a trigram language model. Our results for the 1993 test sets, presented in Table 1, represent the best reported results to date on this task [13][1]. The speaker-independent word error rate for nonnative speakers is reduced by a factor of 2 using only 40 adaptation sentences. Using 200 adaptation sentences, the speaker-adapted error rate of nonnative speakers is comparable to the native speaker-independent word error rate of the same recognition system which is 7.2% and 8.1% on the 1993 development and 1994 evaluation sets, respectively.

Test Set	Adaptation Sentences	SI rate (%)	SA rate (%)
Dev. 93	40	23.5	10.3
Eval. 93	40	16.5	10.0
Eval. 94	40	23.2	11.3
	100		9.4
	200		8.2

Table 1. Speaker Independent (SI) and Speaker Adapted (SA) word error rates on various test sets of nonnative speakers using different amounts of adaptation data.

Acknowledgments

We gratefully acknowledge support for this work from ARPA through Office of Naval Research Contracts N00014-93-C-0142 and N00014-92-C-0154. The 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. J. Bellegarda *et al.*, “Robust Speaker Adaptation Using a Piecewise Linear Acoustic Mapping,” 1992 IEEE ICASSP, pp. I-445—I-448.
2. K. Choukri, G. Chollet and Y. Grenier, “Spectral Transformations through Canonical Correlation Analysis for Speaker Adaptation in ASR,” 1986 IEEE ICASSP, pp. 2659—2662.
3. S. Nakamura and K. Shikano, “A Comparative Study of Spectral Mapping for Speaker Adaptation,” 1990 IEEE ICASSP, pp. 157—160.
4. A. Sankar and C.-H. Lee, “Stochastic Matching for Robust Speech Recognition,” *IEEE Signal Processing Letters,* Vol.1, No.8, August 1994.
5. 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.
6. C.-H. Lee and J.-L. Gauvain, “Speaker Adaptation Based on MAP Estimation of HMM Parameters,” 1993 IEEE ICASSP, pp. II-558 — II-561.
7. V. Digalakis, D. Rtischev and L. Neumeyer, “Fast Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures,” submitted to *IEEE Trans. on Speech and Audio Processing,* April 1994.
8. 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.
9. X. Huang and K.-F. Lee, “On Speaker-Independent, Speaker-Dependent and Speaker-Adaptive Speech Recognition,” *IEEE Trans. on Speech and Audio Processing,* Vol. 1, No. 2, pp. 150—157, April 1993.
10. D. Paul and J. Baker, “The Design for the Wall Street Journal-based CSR corpus,” 1992 DARPA Speech and Natural Language Workshop, pp. 357—362.
11. V. Digalakis and H. Murveit, “Genones: Optimizing the Degree of Tying in a Large Vocabulary HMM-based Speech Recognizer,” 1994 IEEE ICASSP, pp. I-537—I-540.
12. 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.
13. D. Pallet *et al.*, “1993 Benchmark Tests for the ARPA Spoken Language Program,” 1994 ARPA HLT Workshop.

[1] The 1994 official ARPA benchmark results were not available when this paper was written.