Speaker Adaptation Using Combined Transformation and Bayesian Methods

V. Digalakis
TEL +30-821-46566 x226
FAX +30-821-58708
vas@ced.tuc.gr
Electronic and Computer Engineering
Dept., Technical University of Crete
Kounoupidiana, Chania, 73100
GREECE

L. Neumeyer
TEL +1-415-859-4522
FAX +1-415-859-5984
leo@speech.sri.com
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025, USA

November 23, 1994

EDICS SA 1.6.7

ABSTRACT

Adapting the parameters of a statistical speaker-independent continuous-speech recognizer to the speaker and the channel can significantly improve the recognition performance and robustness of the system. 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 sufficient amount of adaptation data for robust maximum-likelihood estimates. To solve this problem, we have recently proposed a constrained estimation technique for Gaussian mixture densities. To improve the behavior of our adaptation scheme for large amounts of adaptation data, we combine it here with Bayesian techniques. We evaluate our algorithms on the large-vocabulary Wall Street Journal corpus for nonnative speakers of American English. The recognition error rate is approximately halved with only a small amount of adaptation data, and it approaches the speaker-independent accuracy achieved for native speakers.

1 INTRODUCTION

Automatic speech recognition performance degrades rapidly when there is a mismatch between the testing and the training conditions, under which the recognizer parameters were estimated. It may not always be feasible to have consistent conditions in the testing and training phases. For example, in large-vocabulary dictation applications the speaker-independent performance degrades dramatically for outlier speakers, such as nonnative speakers of the recognizer language. Since modern large-vocabulary speech recognizers have millions of free parameters, it is not practical to collect large amounts of speaker-dependent data and retrain the recognizer models. Similarly, it is desirable to avoid the expense of collecting additional data when the recognizer is going to be used on speech transmitted through a different channel than the one used in training. Such problems may be solved by adapting the recognizer models, using much smaller amounts of adaptation data than those used in conventional training techniques. In this paper we focus on adapting the models to the speaker, although the techniques we describe can also be used at other levels [1].

One family of adaptation approaches attempts to match the new speaker's observations to the speaker-independent training data by transforming the new speaker's feature space [2][3][4]. The transformation approach has the advantage of simplicity. In addition, 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). A disadvantage of transformation methods is that they are usually text-dependent, that is, the new speaker must record sentences with the same text recorded previously by some reference speakers. Moreover, transformation methods may not take full advantage of large amounts of adaptation data.

A second 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 the nice property that 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 present a transformation-based method for continuous mixture-density hidden Markov models (HMMs) that was introduced in [7]. 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 based on the available amount of adaptation data for quick adaptation. We 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 and takes better advantage of large amounts of adaptation data than the transformation method.

2 TRANSFORMATION-BASED ADAPTATION

Transformation-based approaches to speaker adaptation are typically text-dependent, that is they require the new speaker to record some utterances with predetermined text. 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 [3][4][8].

In [7] we presented a novel transformation-based approach to speaker adaptation for continuous mixture-density HMMs. To eliminate mismatched training and testing conditions, transformations can be applied either directly to the features, or to the speech models [9]. We chose to apply the transformation at the distribution level, rather than transforming the feature vectors directly, since we can then use the Expectation-Maximization (EM) algorithm [10] to estimate the transformation parameters by maximizing the likelihood of the adaptation data (see Figure 1a). One advantage of this approach is that the need for time alignment between new and reference speaker data is eliminated, and the transformation parameters can be estimated using new-speaker data alone. The estimation of the transformation can also be viewed as a constrained estimation of Gaussian mixtures.

For continuous mixture-density HMMs with a large number of component mixtures, it is impractical to assume that enough adaptation data are available for independent reestimation of all the component densities. The constrained estimation we presented in [7] overcomes this problem by applying the same transformation to all components of a particular mixture (or a group of mixtures, if there is tying of transformations). Gaussians for which there were no observations in the training data are adapted based on data that were most likely generated by other Gaussians of the same or other neighboring mixtures.

To see how this method can be applied for adaptation, we assume that the speaker-independent (SI) HMM for the SI vector process $[y_t]$ has observation densities of the form

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

where $p_{SI}(y_t \mid s_t)$ is the observation density of the HMM state s_t, $p(\omega_i \mid s_t)$ is the probability of the i-th mixture component of state s_t, $N(y_t; \mu_{ig}, \Sigma_{ig})$ is the multivariate Gaussian density with mean μ_{ig} and covariance matrix Σ_{ig}, and 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. Specifically, we assume that, given the HMM state s_t, the speaker-dependent vector process $[x_t]$ can be obtained by an underlying process $[y_t]$ through the transformation

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

and that the underlying process $[y_t]$ was generated using the SI model of equation (1).

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

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

and only the parameters A_g, b_g, $g = 1, \ldots, N$ need to be estimated during adaptation, where N 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 for the transformation parameters 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

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

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*

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

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

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

3 COMBINING 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 single-Gaussian observation densities can be obtained using linear combinations of the speaker-dependent counts and some quantities that depend on the parameters of the prior distributions [5]. We use the term *counts* above to denote the sufficient statistics collected by performing one iteration of the forward-backward algorithm on the adaptation data. MAP estimates for the parameters of continuous mixture-density HMMs can be obtained in the same way, as shown in [6]. For example, the MAP estimate for the mean of the *i*th Gaussian in the HMM mixture density of the *g*th Gaussian codebook can be obtained using [6]

$$\mu_{ig}^{MAP} = \frac{\tau_{ig} m_{ig} + \sum_{t,\, s_t} \gamma_t(s_t) \phi_{it}(s_t) x_t}{\tau_{ig} + \sum_{t,\, s_t} \gamma_t(s_t) \phi_{it}(s_t)} \quad , \tag{8}$$

where $\gamma_t(s_t)$ is the probability of being at state s_t at time t given the current HMM parameters, and $\phi_{it}(s_t)$ is the posterior probability of the ith mixture component

$$\phi_{it}(s) = p(\omega_{ig} \mid x_t, s_t) = \frac{p(\omega_{ig} \mid s_t) N(x_t; \mu_{ig}, \Sigma_{ig})}{\sum_j p(\omega_{jg} \mid s_t) N(x_t; \mu_{jg}, \Sigma_{jg})} \quad . \tag{9}$$

The quantities τ_{ig}, m_{ig} are parameters of the joint prior density of the mixture parameters, which was chosen in [6] as a product of the Dirichlet and normal-Wishart densities. The parameter τ_{ig} is usually estimated empirically and can be used to control the adaptation

rate. Similar estimation formulae can be used for the covariances of the Gaussians. Based on (8) and the similar formulae for the second-order statistics, an approximate MAP (AMAP) estimation scheme can be implemented by linearly combining the speaker-independent and the speaker-dependent counts (see Figure 1b) for each component density

$$\begin{aligned} \langle x \rangle_{ig}^{AMAP} &= \lambda \langle x \rangle_{ig}^{SI} + (1-\lambda) \langle x \rangle_{ig}^{SD} \\ \langle x x^T \rangle_{ig}^{AMAP} &= \lambda \langle x x^T \rangle_{ig}^{SI} + (1-\lambda) \langle x x^T \rangle_{ig}^{SD}, \\ n_{ig}^{AMAP} &= \lambda n_{ig}^{SI} + (1-\lambda) n_{ig}^{SD} \end{aligned} \tag{10}$$

where the superscripts on the right-hand side denote the data over which the following statistics (counts) are collected during one iteration of the forward-backward algorithm

$$\begin{aligned} \langle x \rangle_{ig} &= \sum_{t,\, s_t} \gamma_t(s_t) \phi_{it}(s_t) x_t \\ \langle x x^T \rangle_{ig} &= \sum_{t,\, s_t} \gamma_t(s_t) \phi_{it}(s_t) x_t x_t^T \\ n_{ig} &= \sum_{t,\, s_t} \gamma_t(s_t) \phi_{it}(s_t) \end{aligned} . \tag{11}$$

The weight λ controls the adaptation rate. Using the combined counts, we can compute the AMAP estimates of the means and covariances of each Gaussian component density from

$$\begin{aligned} \mu_{ig}^{AMAP} &= \frac{\langle x \rangle_{ig}^{AMAP}}{n_{ig}^{AMAP}} \\ \Sigma_{ig}^{AMAP} &= \frac{\langle x x^T \rangle_{ig}^{AMAP}}{n_{ig}^{AMAP}} - \mu_{ig}^{AMAP} (\mu_{ig}^{AMAP})^T \end{aligned} . \tag{12}$$

Similar adaptation schemes have also appeared for discrete HMMs [11], 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 con-

tinuous 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. To combine the transformation and the approximate Bayesian methods, we can first transform the speaker-independent counts using the transformation parameters estimated with the constrained ML method described in Section 2,

$$\begin{aligned} \langle x \rangle_{ig}^{CML} &= A_g \langle x \rangle_{ig}^{SI} + b_g \\ \langle x x^T \rangle_{ig}^{CML} &= A_g \langle x x^T \rangle_{ig}^{SI} A_g^T + A_g \langle x \rangle_{ig}^{SI} b_g^T + b_g \langle x^T \rangle_{ig}^{SI} A_g^T + n_{ig}^{SI} b_g b_g^T \end{aligned} \quad . \tag{13}$$

The transformed counts can then be combined with the speaker-dependent counts collected using the adaptation data

$$\begin{aligned} \langle x \rangle_{ig}^{COM} &= \lambda \langle x \rangle_{ig}^{CML} + (1-\lambda) \langle x \rangle_{ig}^{SD} \\ \langle x x^T \rangle_{ig}^{COM} &= \lambda \langle x x^T \rangle_{ig}^{CML} + (1-\lambda) \langle x x^T \rangle_{ig}^{SD} \\ n_{ig}^{COM} &= \lambda n_{ig}^{CML} + (1-\lambda) n_{ig}^{SD} \end{aligned} \quad , \tag{14}$$

and the combined-method models can be estimated from these counts using

$$\begin{aligned} \mu_{ig}^{COM} &= \frac{\langle x \rangle_{ig}^{COM}}{n_{ig}^{COM}} \\ \Sigma_{ig}^{COM} &= \frac{\langle x x^T \rangle_{ig}^{COM}}{n_{ig}^{COM}} - \mu_{ig}^{COM} (\mu_{ig}^{COM})^T \end{aligned} \quad . \tag{15}$$

This procedure is shown schematically in Figure 1c.

4 EXPERIMENTAL RESULTS

We evaluated our adaptation algorithms on the Spoke 3 task of the phase-1, large-vocabulary Wall Street Journal (WSJ) corpus [12][13], trying to improve recognition performance for nonnative speakers of American English. Each test set used in this section consists of ten nonnative speakers of American English whose first languages are broadly distributed across the major languages. 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 [15]: 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[1] bigram and trigram language models provided by the MIT Lincoln Laboratory. The trigram language model was implemented using the N-best rescoring paradigm [16], 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 2, 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 perfor-

1. A closed-vocabulary language model is intended for recognizing speech that does not include words outside of the vocabulary.

mance 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 little as 5 adaptation sentences. When adapting using a single sentence, the performance is similar for different numbers of transformations, except for the case of two 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.

In Figure 3 we have plotted the word error rates of the combined scheme for the same numbers of transformations and adaptation sentences as in Figure 2. The systems used to obtain the results of Figure 2 are used as priors for the subsequent Bayesian estimation step, as explained in Section 3. We can see that the performance of the combined scheme becomes less sensitive to the number of transformations used, especially with larger numbers of adaptation sentences. This behavior should be expected, since Bayesian schemes will asymptotically converge to speaker-dependent performance as the amount of adaptation data increases. However, when the number of adaptation sentences is small, it is important to select the appropriate number of transformations and provide the Bayesian step with good prior information.

In Figure 4 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 on an independent test set 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. In the same figure, we can also see the success of the combined method, which outperforms significantly the first two methods over the whole range of adaptation sentences that we 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 [17][2]. 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 for the November 1994 test set is 8.2%. This number is comparable to the speaker-independent word error rate of the same recognition system on the 1993 development and 1994 evaluation sets of native speakers, which is 7.2% and 8.1%, respectively.

The improvement in performance is bigger for certain outlier speakers. The first author of this paper is a nonnative speaker of American English with a particularly heavy accent. His adaptation results for as many as 285 adaptation sentences (approximately 40 minutes of speech) are summarized in Table 2, where we see that his speaker-independent error rate decreases by a factor of 4 and 6 using 40 and 285 adaptation sentences, respectively. His speaker-adapted error rate of 7.1% is comparable to the state-of-the-art performance for native speakers on this task.

5 SUMMARY

We combined the transformation-based adaptation algorithm that we presented in [7] with Bayesian methods. We presented experiments that compare the performance of transfor-

2. The 1994 official ARPA benchmark results were not available when this paper was written.

mation and Bayesian adaptation for various amounts of adaptation data. Transformation-based adaptation performs well when only a limited amount of adaptation data is available, but Bayesian methods are better as the amount of adaptation data increases. The combined method retains the quick adaptation characteristics of transformation methods, and takes advantage of the nice asymptotic properties of Bayesian schemes as the amount of adaptation data increases. The combined scheme clearly outperforms both Bayesian and transformation methods over the whole range of various amounts of adaptation speech that we examined. Our baseline results are the best reported to date on the nonnative-speaker task of the Wall Street Journal corpus, and our nonnative speaker-adapted performance is comparable to the native speaker-independent performance with sufficient amounts of adaptation speech.

Acknowledgments

This research was supported by the Advanced Research Projects Agency through Office of Naval Research Contracts ONR N00014-92-C-0154 and ONR N00014-93-C-0142. The Government has certain rights in this material. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Project Agency.

REFERENCES

[1] L. Neumeyer and M. Weintraub, "Robust Speech Recognition in Noise Using Mapping and Adaptation Techniques," *Proceedings ICASSP,* pp. 141-144, Detroit, Michigan, May 1995.

[2] J. Bellegarda, P. V. de Souza, A. J. Nadas, D. Nahamoo, M. A. Picheny, and L. R. Bahl, "Robust Speaker Adaptation Using a Piecewise Linear Acoustic Mapping," *Proceedings ICASSP,* pp. I-445—I-448, San Fransisco, CA, 1992.

[3] K. Choukri, G. Chollet and Y. Grenier, "Spectral Transformations through Canonical Correlation Analysis for Speaker Adaptation in ASR," *Proceedings ICASSP,* pp. 2659—2662, Tokyo, Japan, 1986.

[4] S. Nakamura and K. Shikano, "A Comparative Study of Spectral Mapping for Speaker Adaptation," *Proceedings ICASSP,* pp. 157—160, Albuquerque, NM, 1990.

[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," *Proceedings ICASSP,* pp. II-558—II-561, Minneapolis, Minnesota, 1993.

[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] R. Schwartz, Y. L. Chow and F. Kubala, "Rapid Speaker Adaptation Using a Probabilistic Spectral Mapping," *Proceedings ICASSP,* pp. 633—636, Dallas, TX, 1987.

[9] A. Sankar and C.-H. Lee, "Stochastic Matching for Robust Speech Recognition," *IEEE Signal Processing Letters,* Vol.1, No.8, August 1994.

[10] 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.

[11] 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.

[12] F. Kubala *et al.,* “The Hub and Spoke Paradigm for CSR Evaluation,” *Proceedings of the HLT workshop, Princeton, NJ, March 1994.*

[13] D. Paul and J. Baker, “The Design for the Wall Street Journal-based CSR corpus,” *Proceedings of the DARPA Speech and Natural Language Workshop,* pp. 357—362, Feb. 1992.

[14] V. Digalakis and H. Murveit, “Genones: Optimizing the Degree of Tying in a Large Vocabulary HMM-based Speech Recognizer,” *Proceedings ICASSP,* Adelaide, Australia, 1994.

[15] H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub, “Large-Vocabulary Dictation Using SRI’s DECIPHER Speech Recognition System: Progressive Search Techniques,” *Proceedings ICASSP,* pp. II-319—II-322, Minneapolis, Minnesota, 1993.

[16] R. Schwartz and Y.-L. Chow, “A Comparison of Several Approximate Algorithms for Finding Multiple (N-Best) Sentence Hypotheses,” *Proc. ICASSP,* pp. 701-704, May 1991.

[17] D. Pallet, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, B. A. Lund, and M. A. Przybocki, “1993 Benchmark Tests for the ARPA Spoken Language Program,” *Proceedings of the HLT Workshop*, Princeton, NJ, March 1994.

TABLES

Test Set	# of Adaptation Sentences	Speaker-independent rate (%)	Speaker-adapted rate (%)
Development 93	40	23.5	10.3
Evaluation 93	40	16.5	10.0
Evaluation 94	40	23.2	11.3
	100		9.4
	200		8.2

TABLE 1. Speaker-independent and speaker-adapted word error rates on various test sets of nonnative speakers using different amounts of adaptation data.

System	# of Adaptation Sentences	Speaker-adapted rate (%)
Speaker Independent	0	42.7
Speaker Adapted	40	10.6
	285	7.1

TABLE 2. Word error rates for development speaker 4n0 and various amounts of adaptation data

FIGURES

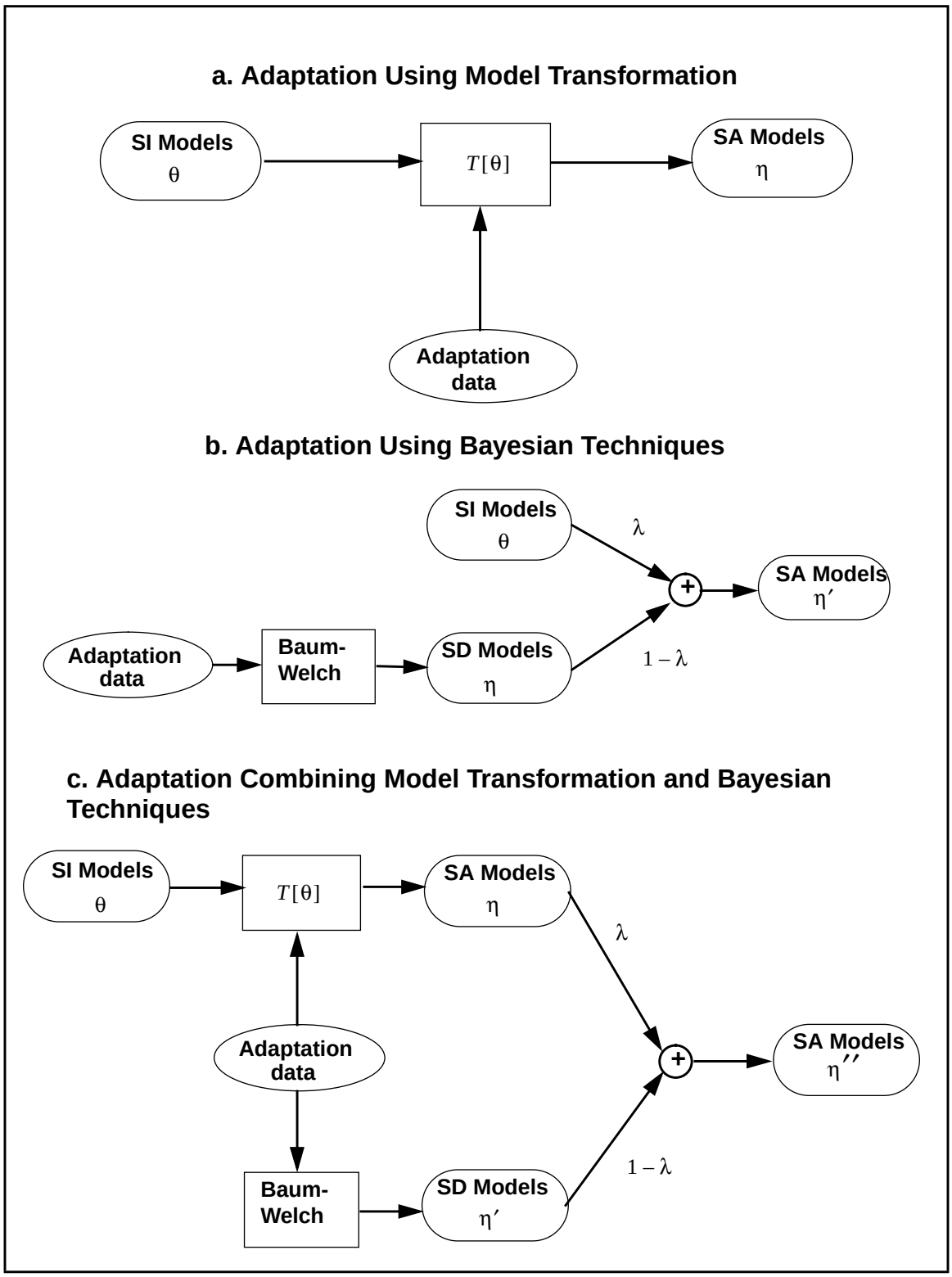

FIGURE 1. Hidden Markov model adaptation using transformation, Bayesian and combined techniques

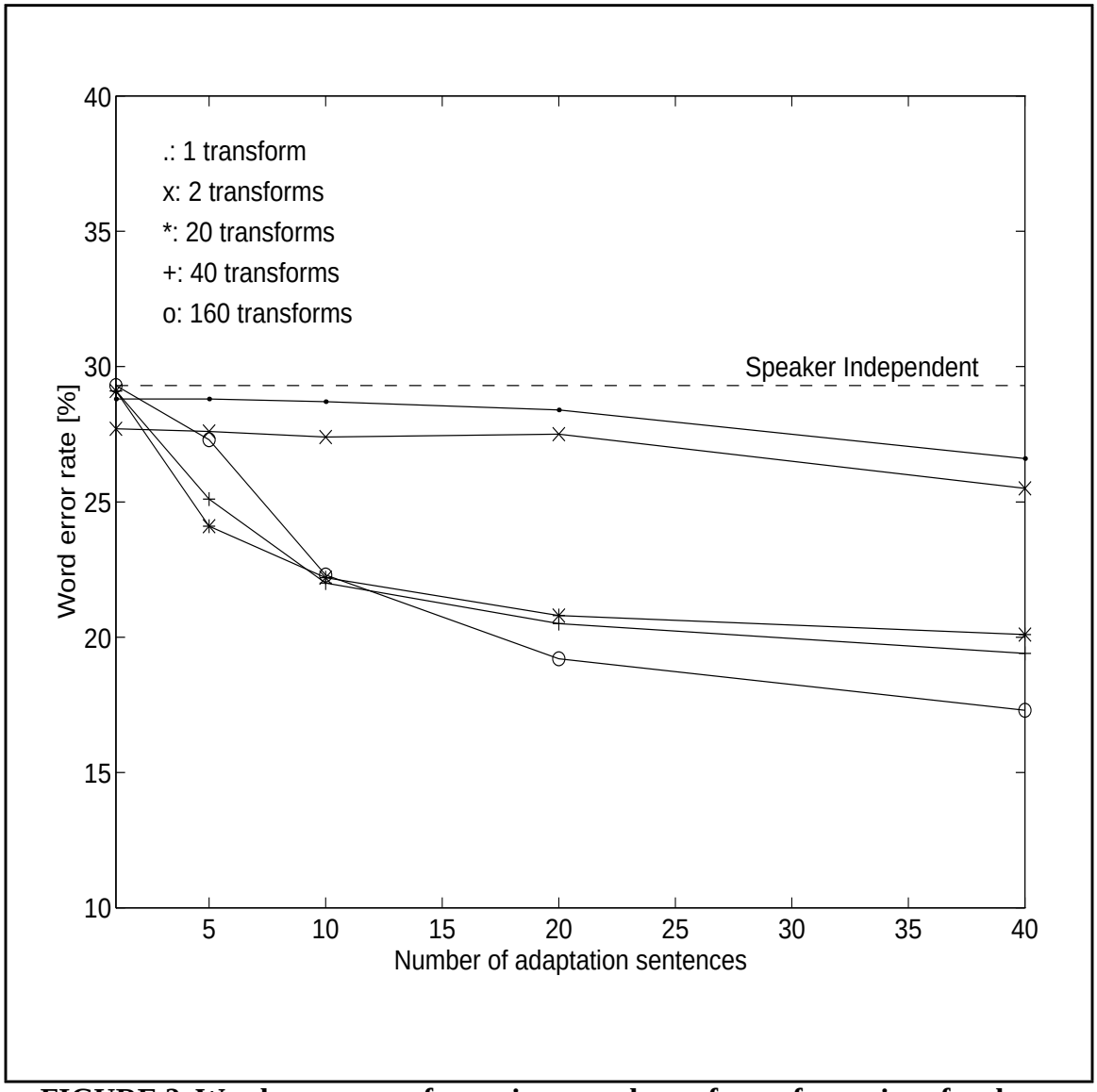

FIGURE 2. Word error rates for various numbers of transformations for the transformation-based adaptation

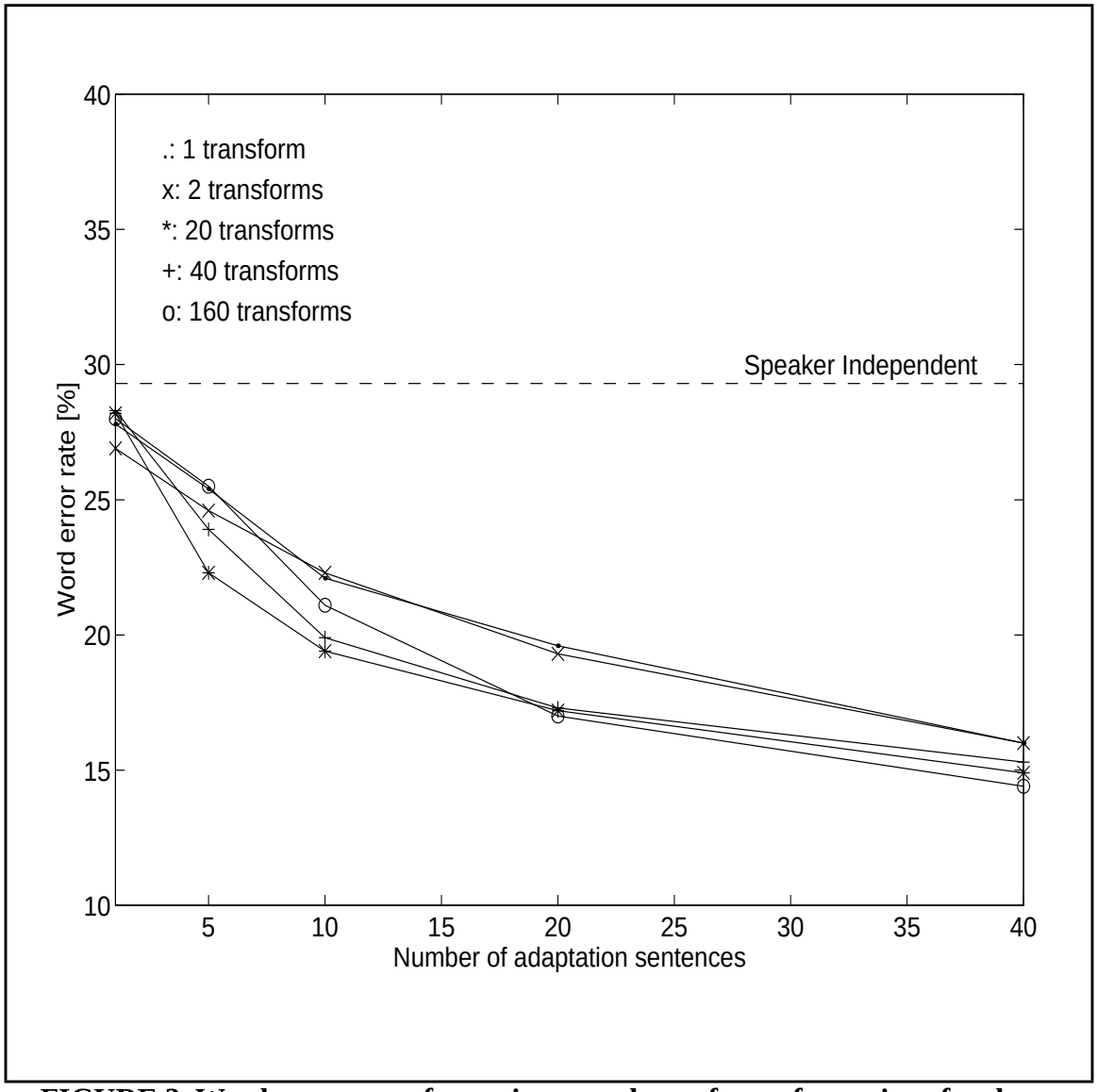

FIGURE 3. Word error rates for various numbers of transformations for the combined method

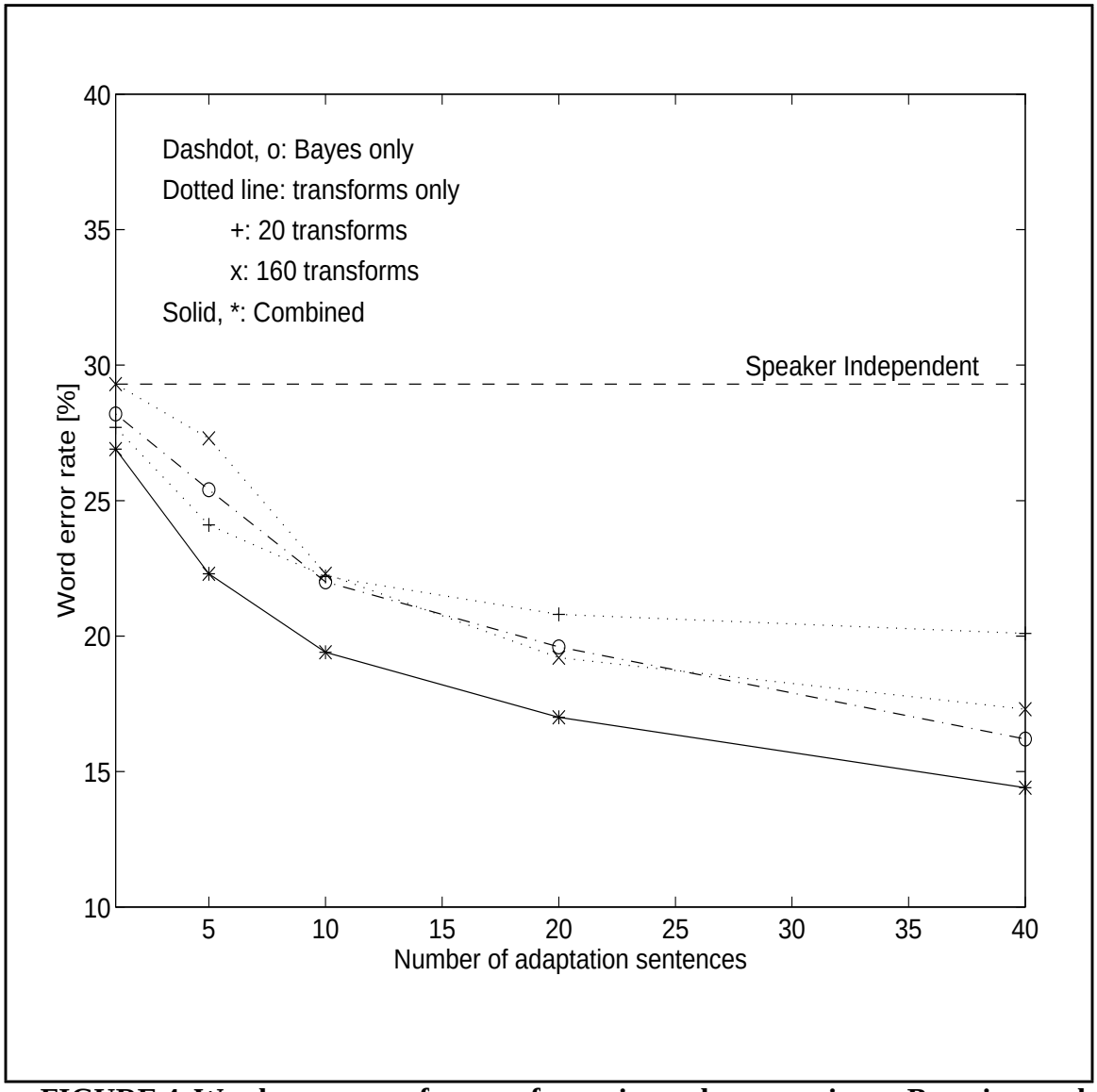

FIGURE 4. Word error rates for transformation-only, approximate Bayesian, and combined schemes