Chapter 7

Cross-language TIMIT - SUN Speech recognition

This chapter details cross-language, cross-database experiments performed using American English speech from the TIMIT [31] database in conjunction with Afrikaans speech from the SUN Speech [12] database to improve speech recognition performance on Afrikaans. Experiments compare word recognition performance when the set of cross-language adaptation strategies from Chapter 5 are applied. The results can be compared with results in Sections 6.4-6.8 from the previous chapter since the experiments described in this chapter also perform continuous word recognition as described in Section 6.2.3. Experiments in both chapters (Chapters 6 and 7) use reasonably large amounts of English source data in conjunction with smaller amounts of Afrikaans target data and test performance on the same speaker independent Afrikaans test set. Results should therefore give a good indication of the expected variation in performance of different techniques when multilingual data is used from the same database (i.e. the same recording conditions and labelling process) versus using data from different databases. It is expected that the performance achieved with cross-language use of the TIMIT database will be less than that achieved with using English speech from the SUN Speech database in recognising Afrikaans speech from SUN Speech, due to the fact that the characteristics and labelling of the databases differ, but

also because the acoustics of South African English may match the acoustics of Afrikaans more closely than American English. On the other hand, the fact that TIMIT contains approximately 80% more speech data than is contained in the English part of SUN Speech, 9 times more speech data than the Afrikaans training set and 45 times more speech data than the Afrikaans adaptation set, may positively influence performance.

The layout of the chapter is as follows. Some characteristics of the TIMIT database, as well as the mapping of the phoneme labels from TIMIT to SUN Speech are discussed first. Experiments then follow, discussing bilingual data pooling, Bayesian adaptation, transformation-based adaptation, combined Bayesian and transformation-based adaptation, discriminative adaptation and finally data augmentation experiments.

7.1 TIMIT - SUN Speech phonetic mapping

The TIMIT [31] database contains read speech in English from a large number of speakers from various dialect regions in the USA. Utterances are labelled phonetically and contain diverse phonetic content. TIMIT is easily available and has been used in previous research for seeding cross-lingual acoustic models [14, 16]. It is therefore well suited for use as a source language database, especially in our case since it allows some evaluation of the effect of database characteristics on cross-language use of acoustic information.

In order to use the TIMIT database with the SUN Speech database, it is necessary to determine a mapping from TIMIT phoneme labels to the SUN Speech phoneme labels. In Chapter 5 we discussed two methods of determining the phoneme mapping, namely a phonetic knowledge-based approach and an automatic approach to determining a phoneme mapping that uses the Bhattacharyya distance. A phonetic knowledge-based mapping from TIMIT phonemes to SUN Speech phonemes was performed by a phonetic expert, details of which are given in Appendix B. The two mappings agree (i.e. list the same TIMIT label for a given SUN Speech label) on 20 out of the 47 phoneme pairs that are used in

recognition experiments (see Tables 6.1 and 6.2 in Section 6.3.1 for the list of phonemes used in experiments). The automatically determined mapping assigns a smaller subset of the TIMIT phonemes in the mapping process, i.e. only 29 different TIMIT phonemes compared to the 38 different TIMIT phonemes listed as the first entry for the phonetically determined mapping.

Continuous word recognition experiments were performed to compare the performance achieved with the two techniques. Results for models trained on TIMIT data and tested on the Afrikaans test set deliver poor performance, achieving -2.6% word accuracy for the automatic approach and -5.9% accuracy for the phonetic approach. It is not surprising that the automatic approach delivers better performance for direct training, since it selects the "closest" source models, thereby reflecting to some extent the channel differences between the source and target data in its choice. In TIMIT/SUN Speech pooling experiments, however, pooling with models determined by the phonetic approach delivers 55.3% and 45.0% word accuracy, versus 50.9% and 32.7% for the automatic mapping approach, when pooling is done with the Afrikaans training set and subset respectively. Also, MAP adaptation of pooled data models indicates that the phonetically derived mapping produces better final results, with word accuracies of 67.7% and 57.0% achieved versus 66.8% and 54.7% for the automatic mapping approach, when adaptation is done on the Afrikaans training set and subset respectively.

The comparative results indicate that better performance is achieved by using the phonetic mapping approach and therefore results are reported only for the phonetically derived mapping in the rest of the chapter. The phonetically derived mapping associates a quality figure with each source/target phoneme pair, indicating qualitatively how accurate each mapping is expected to be, providing extra information which may be useful for seeding prior weight values for adaptation. We, however, did not experiment with using the quality figures.

7.2 Multilingual data pooling

This section evaluates the performance achieved by models trained on pooled speech data from more than one language and from different databases. Figure 7.1 shows word accuracy achieved on the Afrikaans test set when pooled data consisting of the entire TIMIT database in addition to the SUN Speech Afrikaans training set and training subset are used to train phoneme models. Performance is also shown for models trained on the data sets in isolation. Best performance of 69.0% is achieved by using the Afrikaans training set in isolation. Pooling of the Afrikaans training set with the TIMIT set degrades performance to 55.3% word accuracy. Peak performance of models trained on the Afrikaans training subset and the pooled TIMIT plus Afrikaans training subset both round off to 45.0%. Performance of models trained only on the TIMIT database perform poorly on the Afrikaans test set, achieving peak performance of only -3.7% word accuracy. The poor results indicate

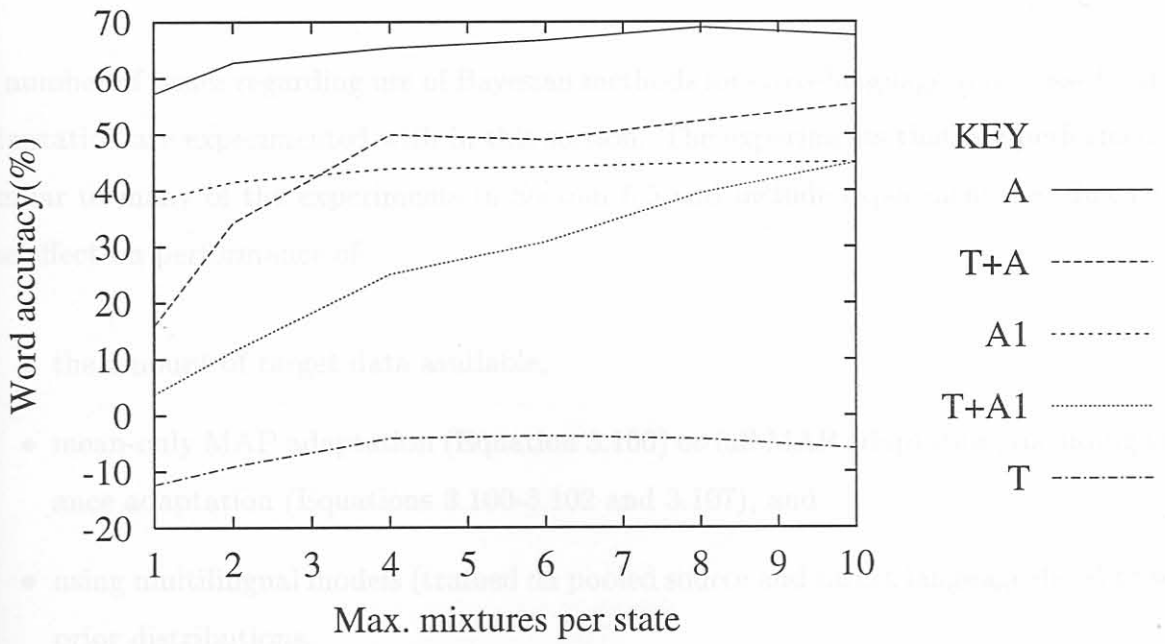

Figure 7.1: Word accuracy as a function of the maximum allowed number of mixtures per state for three state HMMs trained on various monolingual and pooled multilingual data sets using the TIMIT database (T), the Afrikaans training set (A) and the smaller Afrikaans training subset (A1) and tested on the Afrikaans test set

that a large mismatch exists between the TIMIT and SUN Speech databases, especially

if the results are compared to the results of same-database experiments in Section 6.4, where pooling of English and Afrikaans data from the SUN Speech database delivered better results than using Afrikaans data alone. In Section 7.7 we experiment with a data augmentation approach that attempts to improve upon the results achieved with simple data pooling.

For both the pooling approaches (TIMIT pooled with either the Afrikaans training set or training subset), we expect that the performance of pooled-data models will improve if more complex models are trained, i.e. if more than 10 mixtures per state are allowed, but we restrict our attention to using techniques that improve model performance without increasing model complexity.

7.3 Bayesian adaptation

A number of issues regarding use of Bayesian methods for cross-language and cross-database adaptation are experimented with in this section. The experiments that are performed are similar to many of the experiments in Section 6.5 and include experimental evaluation of the effect on performance of:

- the amount of target data available,
- mean-only MAP adaptation (Equation 3.100) or full MAP adaptation, including variance adaptation (Equations 3.100-3.102 and 3.107), and
- using multilingual models (trained on pooled source and target language data) to seed prior distributions.

All experiments also evaluate the influence on performance of the overall weight associated with the prior distribution as this value is determined empirically. The experiments all perform Bayesian adaptation, using the MAP estimation equations from Sections 3.2.3-3.2.5 and in particular Equation 3.107 for variance estimation.

7.3.1 Adaptation performance

Figure 7.2 shows the performance achieved as a function of the adaptation rate for TIMIT English prior models adapted on the SUN Speech Afrikaans training set and the Afrikaans training subset. Peak performance of 67.7% word accuracy is achieved when adapting on the full Afrikaans training set, which delivers an absolute 0.1% improvement over using only the Afrikaans training set (67.6% word accuracy for 3 state, 10 mixture models). Adaptation on the Afrikaans training subset achieves peak performance of 57.0% word accuracy, which is 12.0% better than that achieved by models trained on the Afrikaans training subset alone (45.0%) or by models trained on the pooled TIMIT/Afrikaans training subset (also 45.0% word accuracy).

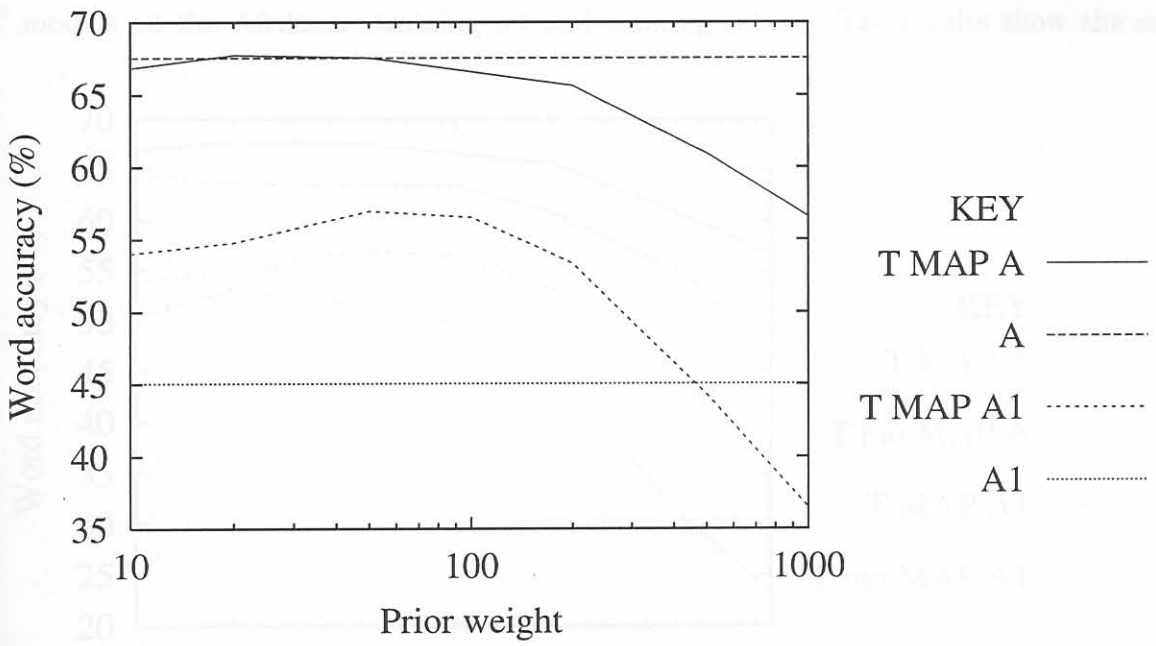

Figure 7.2: Word accuracy on the Afrikaans test set as a function of the adaptation rate for TIMIT models (T) adapted using MAP adaptation on the Afrikaans training set (A) and training subset (A1) with reference performance of monolingual models also shown

The results using English prior models trained on TIMIT are significantly poorer than corresponding results obtained using English priors trained on SUN Speech (67.7% versus 74.9% word accuracy for the Afrikaans training set and 57.0% versus 70.2% word accuracy

for the Afrikaans training subset). Peak performance for TIMIT priors is also achieved for smaller prior weighting ($20 < \varpi < 50$) than the weighting that delivers peak performance for the SUN Speech English prior models ($100 < \varpi < 200$), indicating that the TIMIT priors are less informative than the SUN Speech English priors. The disparity in performance between using TIMIT priors and SUN Speech English priors is expected since the pooling results (Sections 6.4 and 7.2) also show that the English SUN Speech data matches the SUN Speech Afrikaans data more closely than is the case for the TIMIT data.

7.3.2 Variance parameter adaptation

Mean-only and full MAP adaptation are compared in Figure 7.3 for the adaptation of TIMIT models on the Afrikaans training set and training subset. The results show the same

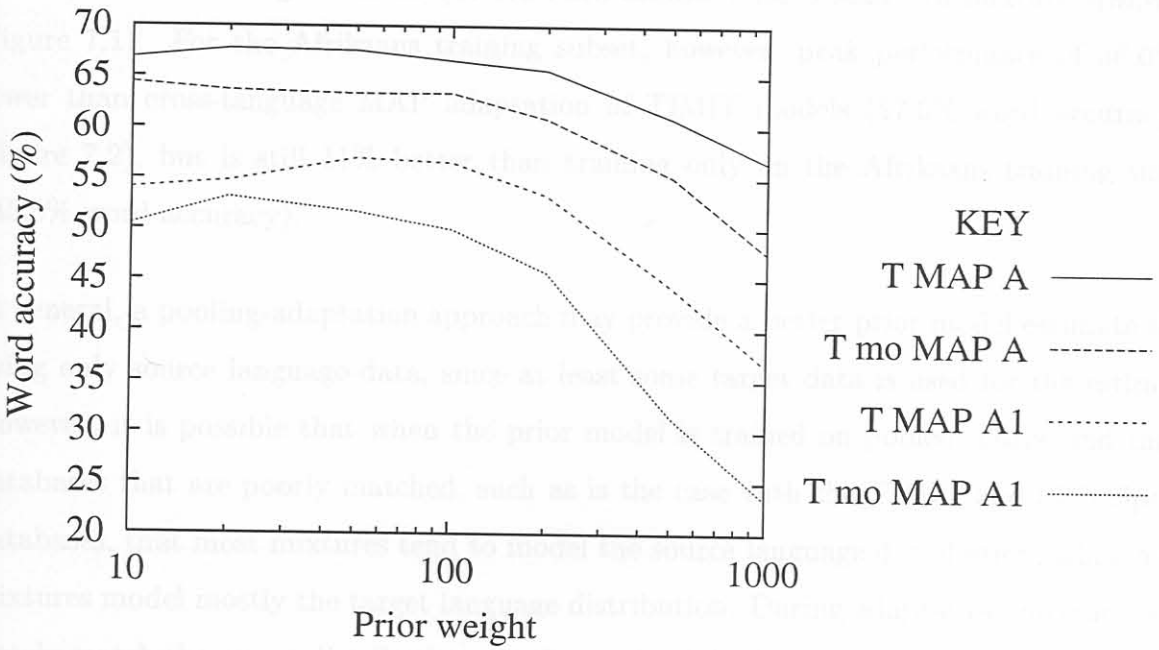

Figure 7.3: Comparison of word accuracy on the Afrikaans test set for mean-only (mo MAP) and full MAP (MAP) adaptation as a function of the adaptation rate for TIMIT models (T) adapted on the Afrikaans training set (A) and training subset (A1)

trend as was shown for adaptation of English prior models from SUN Speech, namely that better cross-language adaptation performance is achieved for full MAP adaptation than for

mean-only MAP adaptation. A 4.1% degradation in peak word accuracy (63.6% versus 67.7%) is attributable to mean-only adaptation versus full adaptation on the Afrikaans training set and a 3.9% degradation in peak word accuracy (53.1% versus 57.0%) is attributable to mean-only versus full adaptation on the Afrikaans training subset.

7.3.3 Pooling-adaptation performance

Figure 7.4 shows the performance achieved when models trained on pooled TIMIT and SUN Speech Afrikaans data set are adapted using full MAP adaptation on the respective Afrikaans data sets. Peak performance of 69.0% is achieved when adapting on the Afrikaans training set, which is 1.3% better than that achieved by cross-language MAP adaptation of TIMIT models (67.7% word accuracy in Figure 7.2) and 1.4% better than training on the Afrikaans training set alone (67.6% word accuracy for 3 state, 10 mixture HMMs in Figure 7.1). For the Afrikaans training subset, however, peak performance of 56.0% is lower than cross-language MAP adaptation of TIMIT models (57.0% word accuracy in Figure 7.2), but is still 11% better than training only on the Afrikaans training subset (45.0% word accuracy).

In general, a pooling-adaptation approach may provide a better prior model estimate than using only source language data, since at least some target data is used for the estimate. However, it is possible that when the prior model is trained on pooled source and target databases that are poorly matched, such as is the case with the TIMIT and SUN Speech databases, that most mixtures tend to model the source language distribution, while a few mixtures model mostly the target language distribution. During adaptation, mixtures that closely match the target distribution are observed, while the large fraction of mixtures that modelled the source language distribution in the initial model are not observed and are therefore also not adapted - negatively influencing performance. This can possibly explain why the pooling-adaptation approach does not necessarily deliver better performance than simply using source language priors for MAP adaptation.

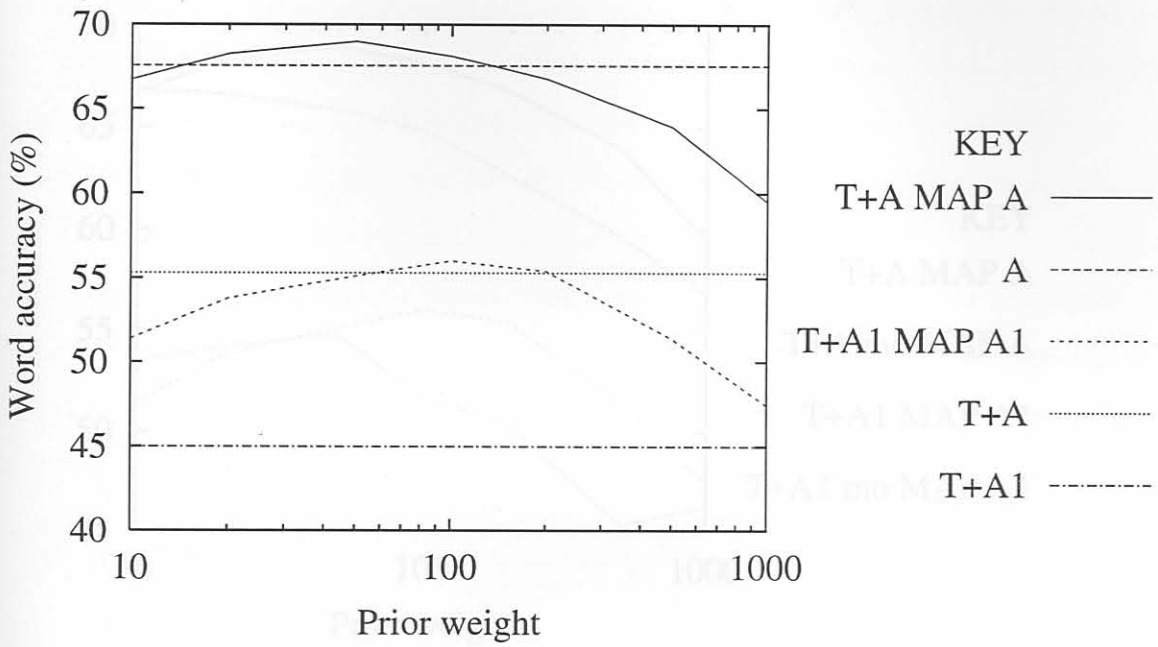

Figure 7.4: Word accuracy on the Afrikaans test set as a function of the adaptation rate (ϖ) for models trained on pooled TIMIT and Afrikaans training data (T+A) and pooled TIMIT and Afrikaans training subset data (T+A1) and adapted using MAP adaptation with reference performance of monolingual and multilingual models also shown

7.3.4 Pooling-variance parameter adaptation

A comparison between results achieved with mean-only and full MAP adaptation of models trained on the pooled TIMIT/Afrikaans data set is given in Figure 7.5. In contrast to the pooled-model MAP adaptation results of the previous chapter (see Figure 6.6), full adaptation outperforms mean-only adaptation for both Afrikaans sets, achieving 2.2% improvement for the Afrikaans training set (69.0% versus 66.8%) and 1.2% improvement for the Afrikaans training subset (56.0% versus 54.8% word accuracy). This may be due in part to the fact that the TIMIT set is even larger than the SUN Speech English set, thereby dominating the pooled model parameters to a larger extent and necessitating variance adaptation since the Afrikaans speech characteristics are not adequately represented in the pooled models.

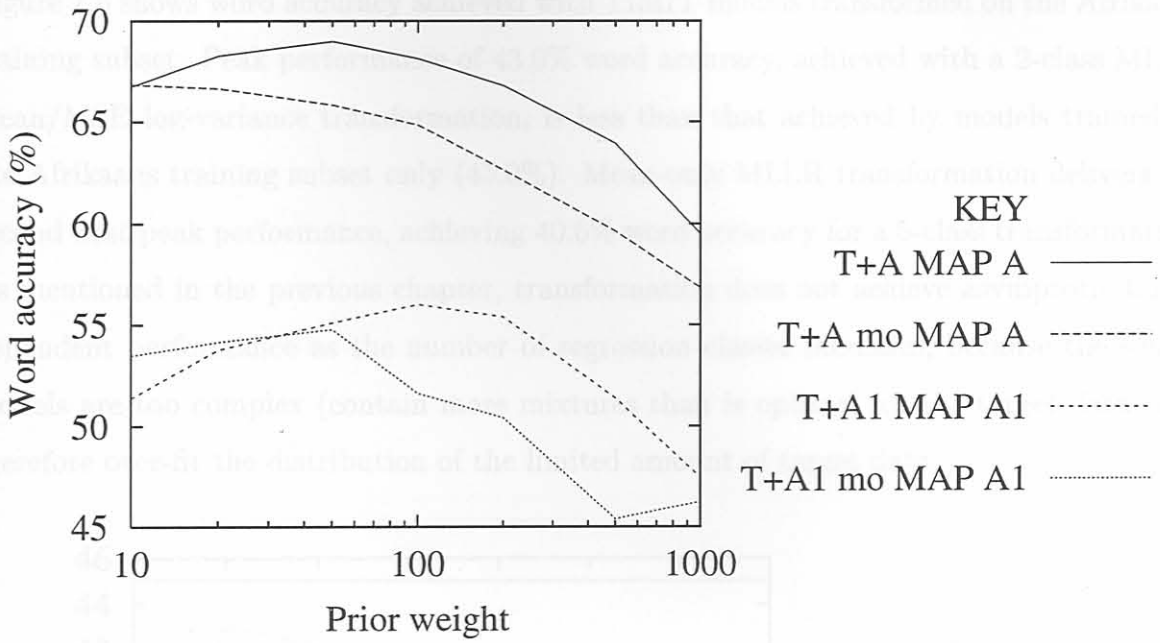

Figure 7.5: Comparison of word accuracy on the Afrikaans test set for mean-only (mo MAP) and full MAP (MAP) adaptation as a function of the adaptation rate for pooled TIMIT-Afrikaans models adapted on the Afrikaans training set (A) and training subset (A1)

7.4 Transformation-based adaptation

Experiments are performed to evaluate the performance of parameter transformation for cross-language and cross-database adaptation, as discussed in Section 5.3.2. Maximum likelihood linear regression (MLLR) transformation (Equation 3.121) is used to transform Gaussian mean parameters and various methods are experimented with for adaptation of Gaussian variance parameters, including:

- no adaptation,
- direct re-estimation (on only the target data),
- linear transformation with MSE criterion (Equation 3.129), and
- log-domain transformation with MSE criterion (Equation 3.136).

Figure 7.6 shows word accuracy achieved with TIMIT models transformed on the Afrikaans training subset. Peak performance of 43.0% word accuracy, achieved with a 2-class MLLR mean/MSE log-variance transformation, is less than that achieved by models trained on the Afrikaans training subset only (45.0%). Mean-only MLLR transformation delivers the second best peak performance, achieving 40.6% word accuracy for a 5-class transformation. As mentioned in the previous chapter, transformation does not achieve asymptotic target dependent performance as the number of regression classes increases, because the source models are too complex (contain more mixtures than is optimal for the target data) and therefore over-fit the distribution of the limited amount of target data.

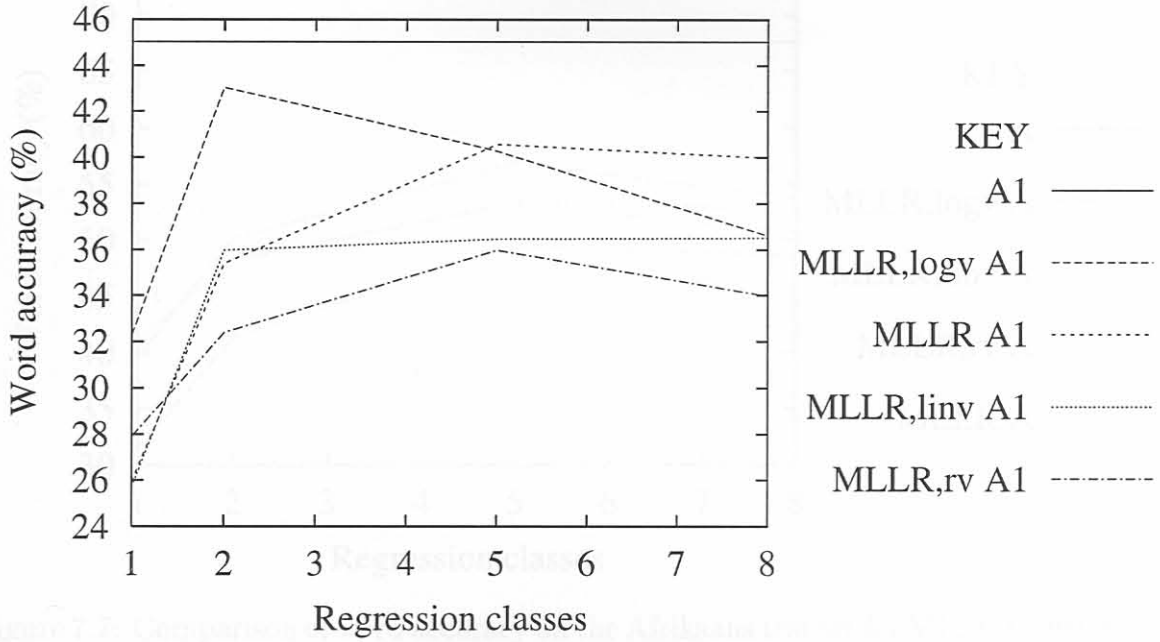

Figure 7.6: Comparison of word accuracy on the Afrikaans test set for MLLR transformation of Gaussian means combined with different variance transformation techniques: mean-only, log variance MSE (logv), linear variance MSE (linv) and variance re-estimation (rv) as a function of the number of regression classes for TIMIT models adapted on the Afrikaans training subset (A1)

Figure 7.7 shows word accuracy achieved with TIMIT models transformed on the Afrikaans training set. Peak performance of 56.9% word accuracy is achieved with a 5-class MLLR mean/MSE log-variance transformation, but is still less than the 67.6% accuracy achieved by models trained on the Afrikaans training set. Linear variance (53.5%) and variance re-

estimation (53.1%) deliver poorer performance, with poorest performance (49.6%) achieved with mean-only MLLR transformation. Other transformation approached were attempted, including block-diagonal transformation which computes separate transformations for cepstral, delta and delta-delta coefficients, as well as a diagonal transformation, which transforms each feature dimension independently. Use of these simpler transformations allows the use of a larger number of regression classes (up to 47 regression classes were used with the diagonal transformation), but did not improve upon the performance achieved with full transformation matrices (results not shown).

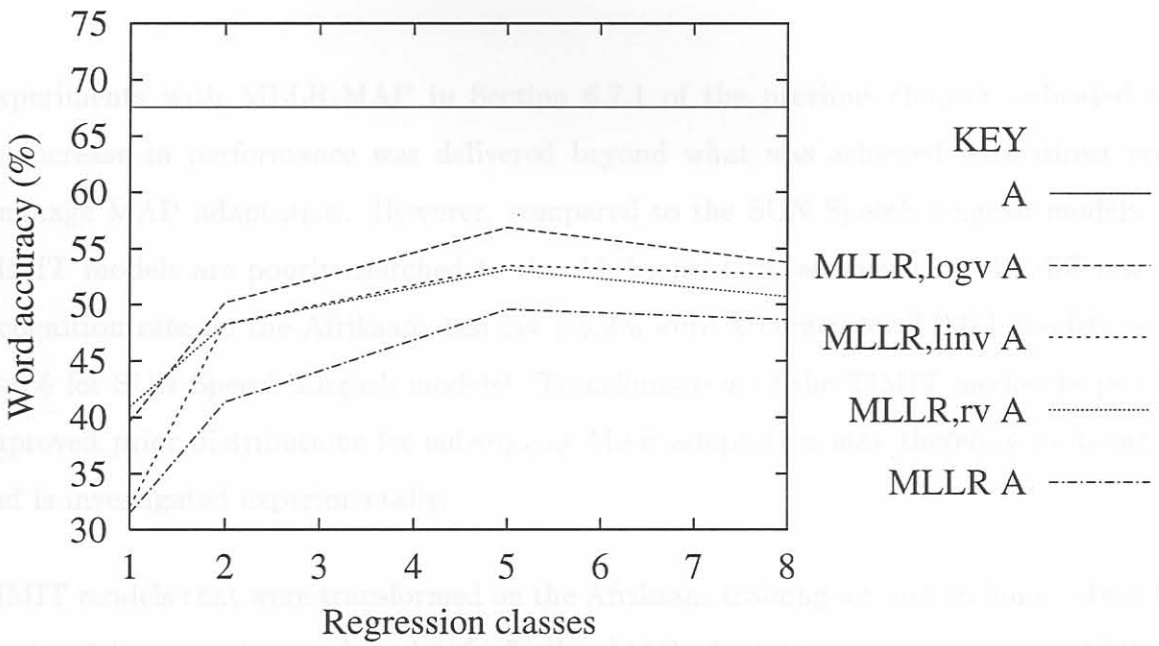

Figure 7.7: Comparison of word accuracy on the Afrikaans test set for MLLR transformation of Gaussian means combined with different variance transformation techniques: mean-only, log variance MSE (logv), linear variance MSE (linv) and variance re-estimation (rv) as a function of the number of regression classes for TIMIT models adapted on the Afrikaans training set (A)

It is, of course, meaningless to perform transformation-based adaptation if the performance achieved is less than that achieved by direct training on the target language data. However, in the next section we show that very good performance is obtained when these transformed models are used for further adaptation.

7.5 Combined transformation-Bayesian adaptation

Experiments are performed to evaluate the two ways of combining Bayesian and transformation-based techniques from Section 3.4 for cross-language and cross-database adaptation and show how MLLR-MAP in particular can lead to greatly improved performance over either MLLR or MAP approaches in isolation.

7.5.1 MLLR-MAP

Experiments with MLLR-MAP in Section 6.7.1 of the previous chapter indicated that no increase in performance was delivered beyond what was achieved with direct cross-language MAP adaptation. However, compared to the SUN Speech English models, the TIMIT models are poorly matched to the Afrikaans data, as shown by the difference in recognition rate on the Afrikaans test set (-5.9% word accuracy for TIMIT models versus 58.0% for SUN Speech English models). Transformation of the TIMIT models to produce improved prior distributions for subsequent MAP adaptation may therefore be beneficial and is investigated experimentally.

TIMIT models that were transformed on the Afrikaans training set and training subset (see Section 7.4) are used as seed models for further MAP adaptation on the respective Afrikaans sets. Figure 7.8 shows the word accuracy achieved as a function of the adaptation rate when the MLLR transformed models are adapted using MAP adaptation. Peak performance of 72.0% word accuracy is achieved when a 2-class (mean-only) MLLR transformation is followed by full MAP adaptation on the Afrikaans training set. This peak performance is 4.4% better than achieved with training on the Afrikaans set only (72.0% versus 67.6%) and 3.0% better than the best MAP adaptation results using TIMIT (69.0% word accuracy for adaptation of bilingual models in Figure 7.4). When the Afrikaans training subset is used for adaptation purposes, peak performance of 64.1% is achieved when a single class MLLR transformation is followed by MAP adaptation. This peak performance is

19.1% better than the 45.0% word accuracy achieved with models trained on the Afrikaans training subset only and 7.1% better than the 57.0% word accuracy achieved with MAP adaptation on the Afrikaans training subset in Figure 7.2. Use of the MSE log-variance

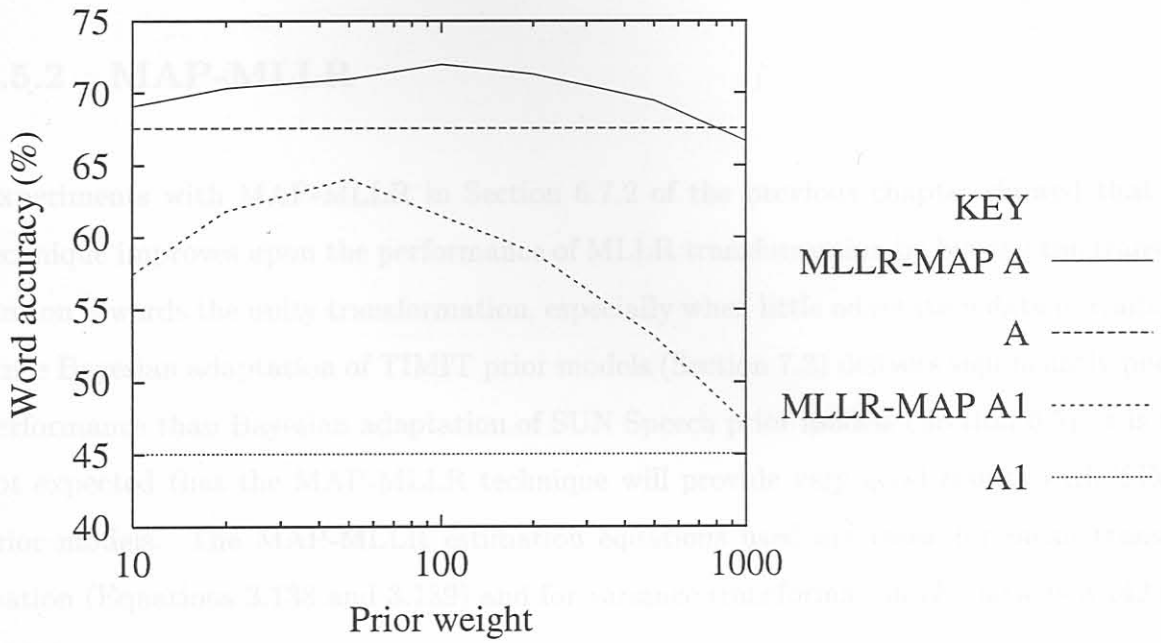

Figure 7.8: Word accuracy on the Afrikaans test set as a function of the adaptation rate for TIMIT models adapted using MLLR-MAP adaptation on the Afrikaans training set (A) and training subset (A1) with reference performance of monolingual models also shown

transformation in the first step of MLLR-MAP produces almost the same performance as using MLLR transformation (71.8% for a single regression class versus 72.0% for the 2-class MLLR transformation).

The results of Figure 7.8 show useful increases in performance by using the TIMIT database to improve the Afrikaans recogniser, indicating that the MLLR-MAP strategy is well suited for cross-database, cross-language adaptation. Best performance is not achieved by using the MLLR models that deliver the best performance (this would imply using 2-class and 5-class transformed models), but by using models transformed with simpler transformations (single and 2-class transformations). The transformation step acts to improve the priors by using correlation between the source language feature distribution and the target language feature distribution. The transformation step should therefore not necessarily be optimised

for transformed model performance because a too complex transformation may over-fit the target data, decreasing the usefulness of the transformed model in seeding the priors for subsequent MAP adaptation.

7.5.2 MAP-MLLR

Experiments with MAP-MLLR in Section 6.7.2 of the previous chapter showed that the technique improves upon the performance of MLLR transformation by biasing the transformation towards the unity transformation, especially when little adaptation data is available. Since Bayesian adaptation of TIMIT prior models (Section 7.3) delivers significantly poorer performance than Bayesian adaptation of SUN Speech prior models (Section 6.5), it is also not expected that the MAP-MLLR technique will provide very good results with TIMIT prior models. The MAP-MLLR estimation equations used are those for mean transformation (Equations 3.138 and 3.139) and for variance transformation (Equations 3.140 and 3.141).

Figure 7.9 shows the performance achieved as a function of the number of regression classes when MAP-MLLR transformation of the Gaussian mean parameters and, optionally, a MAP-like log-space transformation of the Gaussian variance parameters of TIMIT models are attempted on the Afrikaans training set and training subset. Peak performance of 56.1% word accuracy is achieved for mean and variance MAP-MLLR transformation on the Afrikaans training set. This performance is for a prior weight scaling factor of 10 ($\varpi = 10$), and is less than the peak word accuracy of 56.9% achieved with MLLR-mean/MSE log-variance transformation in Figure 7.7, i.e. when a prior weight scaling factor of zero is used ($\varpi = 0$), and is also significantly less than the performance achieved with models trained directly on the Afrikaans training set (67.6% word accuracy). The best performance on the Afrikaans training subset is 45.1% (also using $\varpi = 10$), at least slightly improving on direct training on the Afrikaans training subset (45.0% word accuracy) and also improving on using zero prior weighting (43.0% word accuracy).

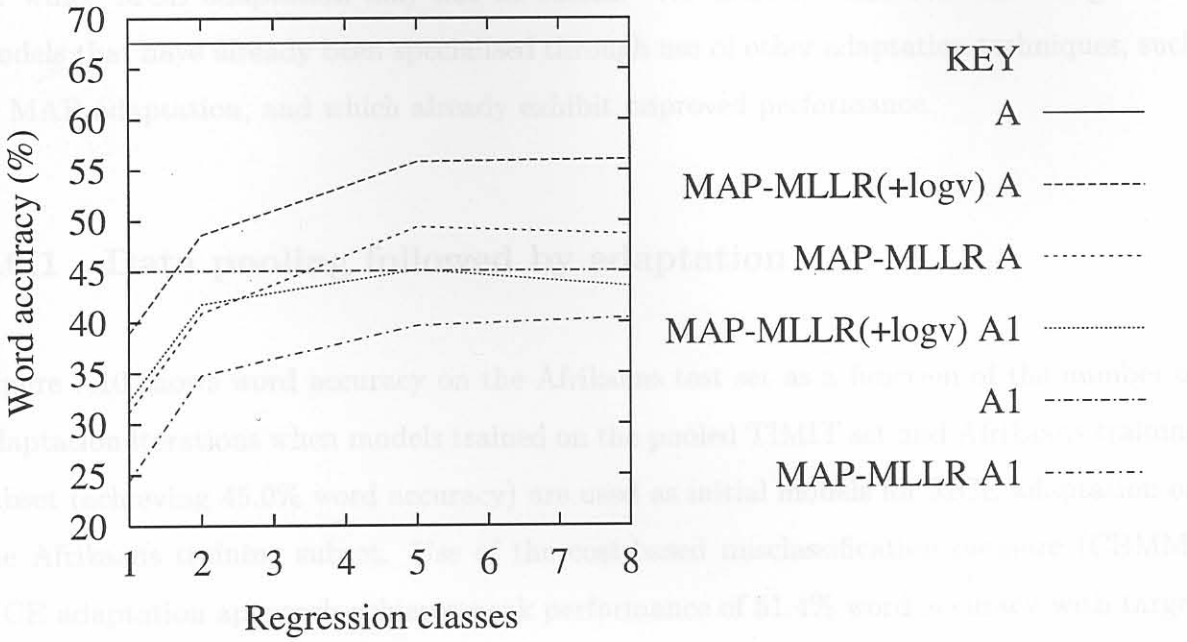

Figure 7.9: Comparison of word accuracy on the Afrikaans test set for MAP-MLLR transformation of Gaussian means, optionally combined with a MAP-like log-space (+logv) MSE transformation of Gaussian variance parameters, computed as a function of the number of regression classes for TIMIT models adapted on the Afrikaans training set (A) and training subset (A1)

7.6 Discriminative adaptation

Experiments are performed to evaluate the application of discriminative adaptation for cross-language and cross-database adaptation, as was discussed in Section 5.3.3. MCE adaptation experiments in Section 6.8 of the previous chapter delivered good performance when using initial models trained on pooled multilingual data and we therefore also attempt the same approach in experiments in this section. The reasoning behind using multilingual initial models is that they are robustly estimated from a large amount of data (pooled multilingual data) and may need only a degree of language specific "fine-tuning" to deliver good target language performance. It should, however, be taken into account that pooling Afrikaans data from SUN Speech with data from the TIMIT database did not improve on using the Afrikaans data alone (see Section 7.2) and may therefore produce initial models for MCE adaptation that need significant adaptation to reach a reasonable level of performance,

for which MCE adaptation may not be suited. We therefore also consider using initial models that have already been specialised through use of other adaptation techniques, such as MAP adaptation, and which already exhibit improved performance.

7.6.1 Data pooling followed by adaptation

Figure 7.10 shows word accuracy on the Afrikaans test set as a function of the number of adaptation iterations when models trained on the pooled TIMIT set and Afrikaans training subset (achieving 45.0% word accuracy) are used as initial models for MCE adaptation on the Afrikaans training subset. Use of the cost-based misclassification measure (CBMM) MCE adaptation approach achieves peak performance of 51.4% word accuracy with target context cost, compared to 49.4% word accuracy with target independent cost. Performance using MCE adaptation (without CBMM) is not as good, and only improves by 0.2% on the initial model performance (45.2% versus 45.0% word accuracy). The best performance of 51.4%, achieved with MCE adaptation using the Afrikaans training subset (obtained with CBMM MCE), is 6.4% better than that achieved with the Afrikaans training subset alone (45.0% word accuracy), delivering useful cross-language adaptation performance. The performance (51.4%) is, however, below the peak accuracy of 64.1% achieved with MLLR-MAP adaptation on the Afrikaans training subset.

For comparison purposes, MCE adaptation experiments were also performed using an initial model trained directly on the Afrikaans training set. Adaptation using target context CBMM achieves peak performance of 47.4%, which improves on the baseline 45.0% performance achieved with direct training on the Afrikaans training subset, but which is less than that achieved by adapting the multilingual (TIMIT and Afrikaans training subset) initial model.

Figure 7.11 shows word accuracy as a function of the number of adaptation iterations when models trained on the pooled TIMIT and Afrikaans training set (achieving 55.3% word accuracy) are used as initial models for MCE adaptation on the Afrikaans training set.

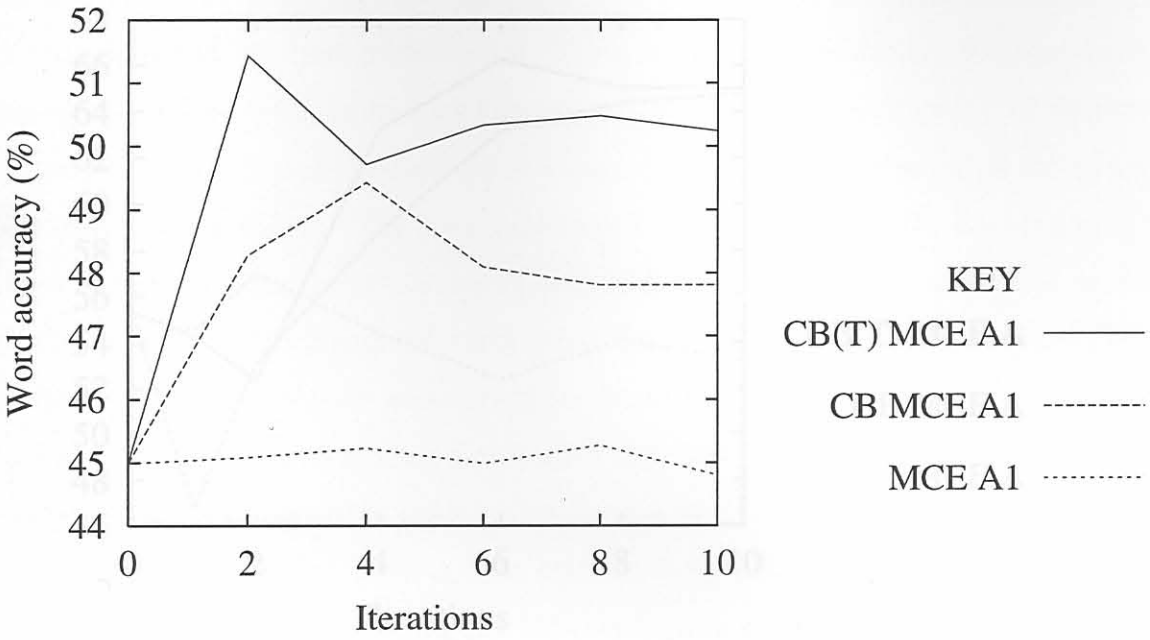

Figure 7.10: Comparison of word accuracy on the Afrikaans test set for MCE adaptation of pooled TIMIT-Afrikaans models on the Afrikaans training subset, also showing results using of a cost-based (CB) misclassification measure, optionally designed specifically for the target context (T)

CBMM MCE adaptation achieves peak performance of 66.2% word accuracy with target context cost and 64.7% word accuracy with target independent cost. As in the previous experiment, performance of MCE adaptation without CBMM is not as good as MCE with CBMM, and achieves only 57.0% peak word accuracy.

The best performance of 66.2% (achieved with target context CBMM MCE) is still below the baseline 67.6% achieved with direct training on the Afrikaans training set. The result illustrates a drawback of the pooling-adaptation approach. If the performance of the models trained on pooled multilingual data is far below that of models trained directly on the target language data only (e.g. in this case 55.3% for multilingual models versus 67.6% for target language only models), then the tendency of discriminative adaptation techniques to converge to local minima may result in a poorer final model than simple target language training.

The experiments with MCE adaptation show reasonable improvements in performance com-

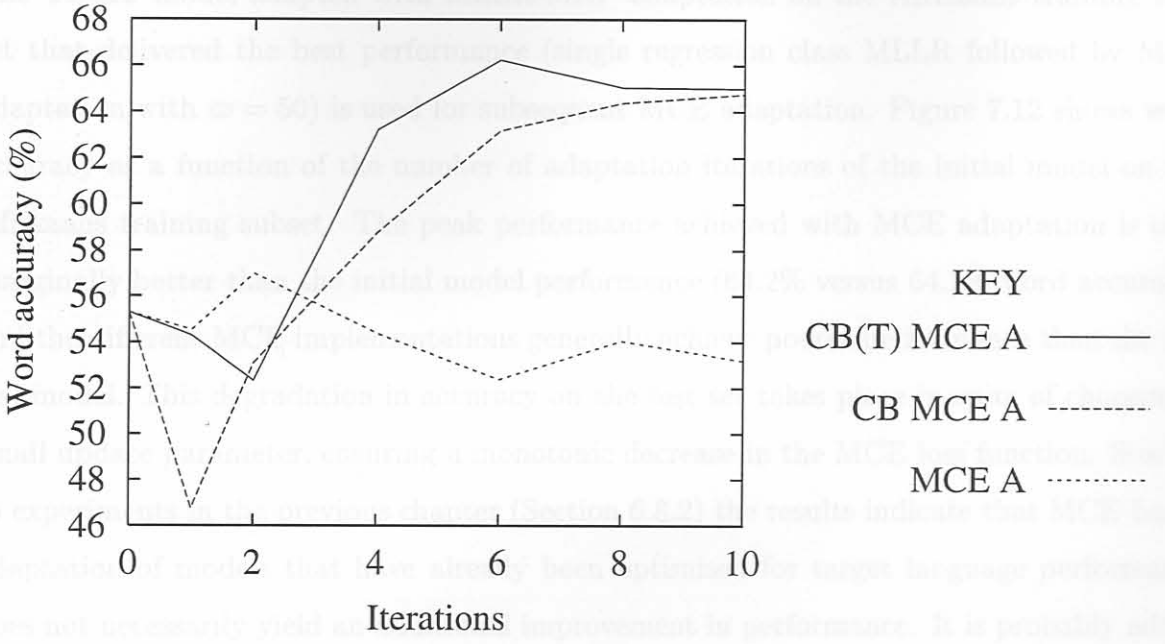

Figure 7.11: Comparison of word accuracy on the Afrikaans test set for the MCE adaptation of pooled TIMIT-Afrikaans models on the Afrikaans training set, also including use of a cost-based (CB) misclassification measure, optionally designed specifically for the target context (T)

pared to the performance of the initial models before adaptation. Of the MCE approaches, target context cost-based MCE adaptation delivers the best performance for multilingual model adaptation, increasing performance by 6.4% for Afrikaans training subset adaptation (51.4% versus 45.0% word accuracy). For Afrikaans training set adaptation, peak performance of 66.2% is achieved, which is still below that achieved with the Afrikaans training set in isolation (67.6%).

7.6.2 Improving best performing models

The performance achieved with MCE adaptation on both the Afrikaans training set and subset (Figures 7.10 and 7.11) is less than the best performance achieved with other adaptation techniques. The adapted models that delivered the best performance in previous experiments are now used as initial models in an attempt to improve performance with MCE adaptation.

The TIMIT model adapted with MLLR-MAP adaptation on the Afrikaans training subset that delivered the best performance (single regression class MLLR followed by MAP adaptation with $\varpi = 50$) is used for subsequent MCE adaptation. Figure 7.12 shows word accuracy as a function of the number of adaptation iterations of the initial model on the Afrikaans training subset. The peak performance achieved with MCE adaptation is only marginally better than the initial model performance (64.2% versus 64.1% word accuracy) and the different MCE implementations generally achieve poorer performance than the initial model. This degradation in accuracy on the test set takes place in spite of choosing a small update parameter, ensuring a monotonic decrease in the MCE loss function. Similar to experiments in the previous chapter (Section 6.8.2) the results indicate that MCE-based adaptation of models that have already been optimised for target language performance does not necessarily yield an additional improvement in performance. It is probably advisable to use a cross-validation set with such adaptation since a decrease in the MCE loss function does not guarantee improved performance on the test set.

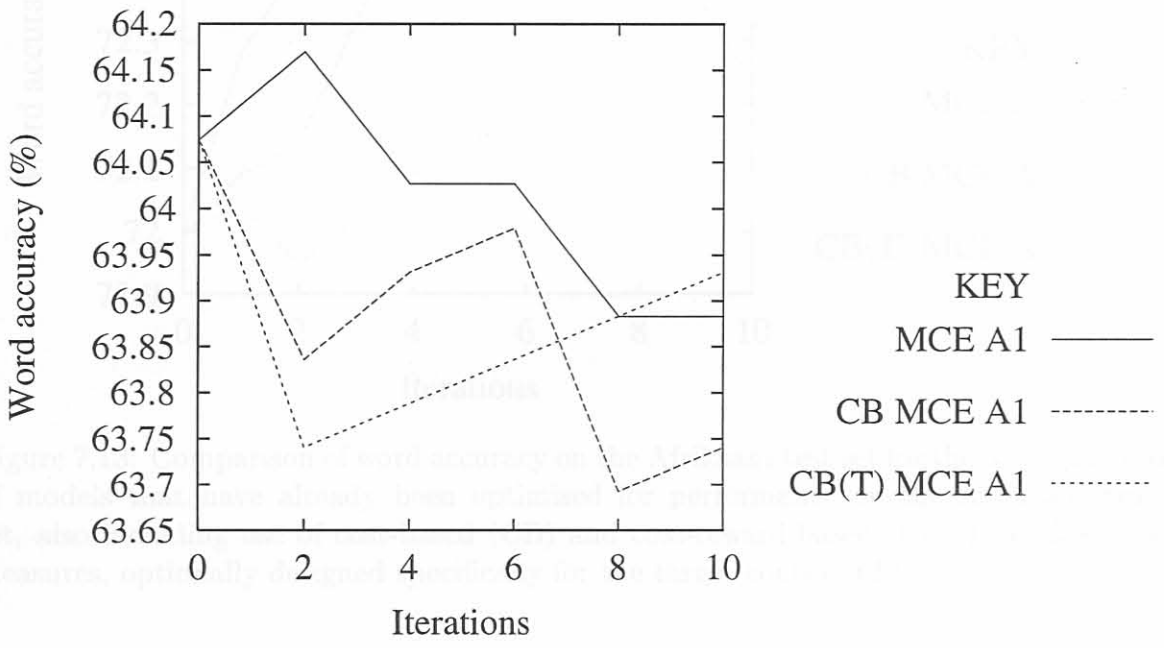

Figure 7.12: Comparison of word accuracy on the Afrikaans test set for the MCE adaptation of models that have already been optimised for performance on the Afrikaans training subset, also including use of cost-based (CB) and cost-reward-based (CRB) misclassification measures, optionally designed specifically for the target context (T)

The next experiment attempts to improve on MCE adaptation performance on the Afrikaans training set. The TIMIT model adapted with MLLR-MAP adaptation on the Afrikaans training set that delivered the best performance (2-class MLLR followed by MAP adaptation with $\varpi = 100$) is used for subsequent MCE adaptation. Figure 7.13 shows word accuracy as a function of the number of adaptation iterations when the initial model is adapted on the Afrikaans training set. Peak performance of 72.7% word accuracy is achieved with MCE adaptation (without using a cost function), which is 0.7% better than (unadapted) initial model performance (72.0% word accuracy). CBMM MCE adaptation delivers peak performance of 72.5% word accuracy and target dependent CBMM MCE delivers only 72.2% peak word accuracy.

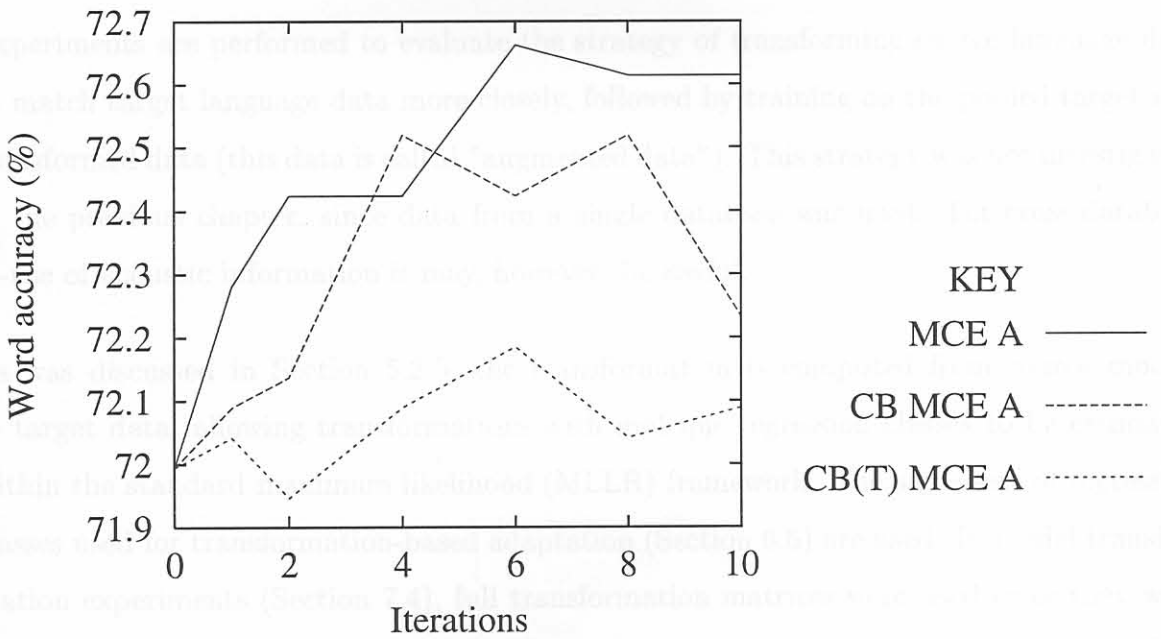

Figure 7.13: Comparison of word accuracy on the Afrikaans test set for the MCE adaptation of models that have already been optimised for performance on the Afrikaans training set, also including use of cost-based (CB) and cost-reward-based (CRB) misclassification measures, optionally designed specifically for the target context (T)

Similar to Section 6.8.2, experimental results for discriminative adaptation of models already optimised for target language dependent performance are not conclusive. When adapted initial models are further adapted with MCE, performance improves for Afrikaans training set adaptation, but does not improve for Afrikaans training subset adaptation. If

this approach is therefore followed, use of a cross-validation set is recommended to ensure that discriminatively adapted models are not used if their performance is less than that achieved with the initial models.

Throughout this chapter different methods of adapting model parameters were investigated. In the next section we investigate a technique for the transformation of source language data.

7.7 Data augmentation

Experiments are performed to evaluate the strategy of transforming source language data to match target language data more closely, followed by training on the pooled target and transformed data (this data is called "augmented data"). This strategy was not investigated in the previous chapter, since data from a single database was used. For cross-database re-use of acoustic information it may, however, be useful.

As was discussed in Section 5.2.5, the transformation is computed from source models to target data, allowing transformations with multiple regression classes to be estimated within the standard maximum likelihood (MLLR) framework. The same sets of regression classes used for transformation-based adaptation (Section 6.6) are used. In model transformation experiments (Section 7.4), full transformation matrices were used since they were found to deliver better performance than diagonal or block-diagonal matrices. For data augmentation, use of diagonal and block-diagonal transformations are reconsidered, especially if only the mel-cepstral coefficients are transformed (i.e. not including time derivative components). Single state HMMs with a maximum of 10 mixtures per state are trained on the TIMIT database and used as source models in the computation of the transformation. A single state source model is used in order to apply a single transformation to data from a particular class. Using multiple transformations per speech segment leads to discontinuities at the alignment points of the data (assuming Viterbi-alignment is done to segment source

data) and can degrade performance when time derivative feature components are calculated afterwards.

Table 7.1 summarises the performance achieved when training on the pooled transformed and target data. Results for zero regression classes indicate pooling with the (untransformed) source data. Diagonal transformation matrices (with offset) of all features, including time derivative features, were used as they were found to deliver better performance than using either block-diagonal or full transformation matrices. The results indicate that

Table 7.1: Word accuracy achieved on the Afrikaans test set for models trained on data from TIMIT that is transformed to better match the respective Afrikaans set and also pooled with the respective Afrikaans set

Set used for adaptation	Regression classes					
	0	1	2	5	8	15
Afr. training set	55.3%	53.3%	48.3%	49.3%	45.8%	40.3%
Afr. training subset	45.0%	31.0%	28.9%	29.4%	25.9%	22.8%

no gain in performance is achieved by first transforming the TIMIT data before pooling it with the SUN Speech data for the training of multilingual models. The transformation improves the likelihood of the source model (on the target data) when it is transformed, and therefore probably improves the match between the transformed data and the target data. However, it should be kept in mind that increased overlap between the (transformed) source data distribution and that of the target data does not necessarily imply improved model performance, since the degree of class confusability may also be increased.

The augmentation results of Table 7.1 show no reason to use the approach, but in the next section we put models trained on the augmented data to good use.

7.8 Augmentation followed by adaptation

Models trained on augmented data (source data transformed using single regression class transformations together with the target data) are used for subsequent MAP adaptation on target language data. Figure 7.14 shows word accuracy achieved on the Afrikaans test set as a function of the prior weight (ϖ) for MAP adaptation on the Afrikaans training set and training subset. Peak accuracy of 71.8% is achieved for MAP adaptation of models trained on augmented data and adapted on the Afrikaans training set, compared to 69.0% word accuracy (see Figure 7.4) achieved by adapting models trained on pooled data. For Afrikaans training subset adaptation, peak word accuracy of 61.8% is achieved when priors trained on augmented data are used, compared to 56.0% word accuracy for priors trained on pooled data. The results clearly show the benefit of using augmented data priors versus using priors trained simply on pooled data.

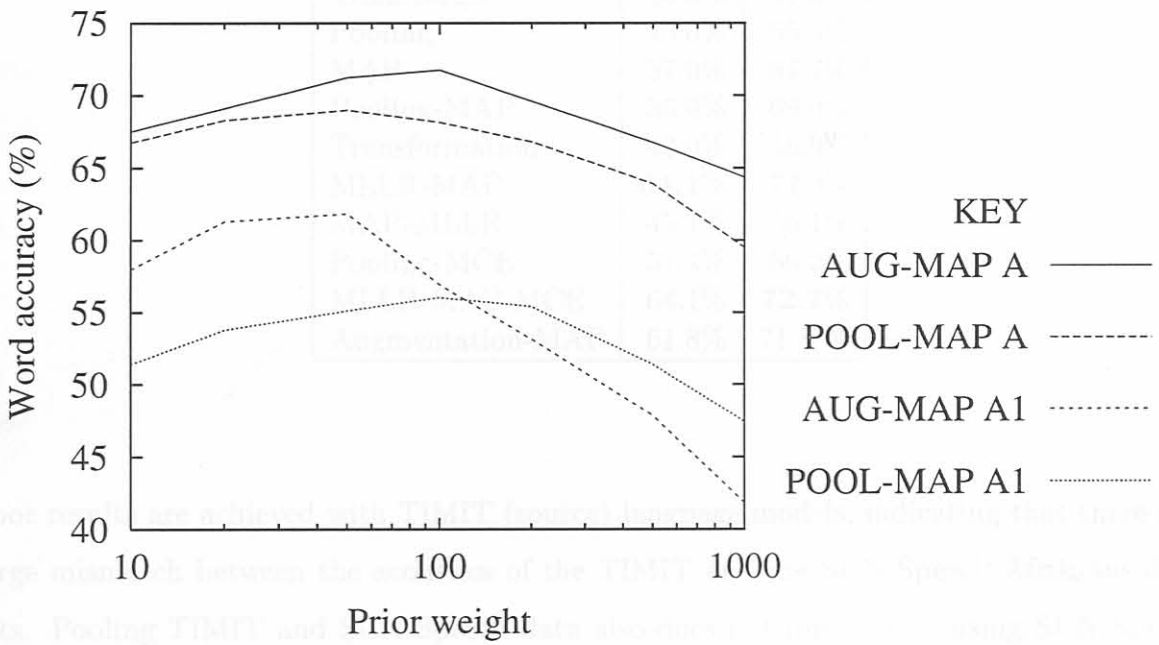

Figure 7.14: Comparison of the word accuracy on the Afrikaans test set as a function of the adaptation rate for models trained on transformed TIMIT data in addition to Afrikaans data (AUG), or on pooled TIMIT and Afrikaans data, when the Afrikaans training set (A) or training subset (A1) is used

7.9 Discussion of results

The experiments in this chapter covered application of the major categories of speaker adaptation techniques, as well as extensions and combinations of them, to cross-language and cross-database adaptation of acoustic parameters, combining data and models associated with both the TIMIT and SUN Speech databases. A number of approaches are shown to deliver useful cross-language adaptation performance. Table 7.2 summarises the methods that were experimented with and their results, which are briefly discussed next.

Table 7.2: Summary of peak word accuracy achieved on the Afrikaans test set in various experiments that evaluate different approaches to cross-language adaptation on the Afrikaans training set (A) and Afrikaans training subset (A1)

Method	Adapted on	
	A1	A
Train source	-5.9%	-5.9%
Train target	45.0%	67.6%
Pooling	45.0%	55.3%
MAP	57.0%	67.7%
Pooling-MAP	56.0%	69.0%
Transformation	43.0%	56.9%
MLLR-MAP	**64.1%**	72.0%
MAP-MLLR	45.1%	56.1%
Pooling-MCE	51.4%	66.2%
MLLR-MAP-MCE	64.1%	**72.7%**
Augmentation-MAP	61.8%	71.8 %

Poor results are achieved with TIMIT (source) language models, indicating that there is a large mismatch between the acoustics of the TIMIT and the SUN Speech Afrikaans data sets. Pooling TIMIT and SUN Speech data also does not improve on using SUN Speech data alone, due to the large mismatch between the databases. This is in contrast with results in the previous chapter (Section 6.4) where simple pooling of SUN Speech English and Afrikaans data improved on using only the Afrikaans data.

Cross-language MAP adaptation delivers reasonably good results, showing large improve-

ment over using the Afrikaans training subset only (57.0% versus 45.0%), but showing little improvement (0.1%) over using the Afrikaans training set only. Using multilingual priors for MAP adaptation (pooling-MAP) improves performance for Afrikaans training set adaptation, but degrades performance for the Afrikaans training subset.

Transformation-based adaptation (MLLR mean and log-variance MSE transformation) does not deliver useful performance by itself, achieving poorer performance than is achieved with direct training on target data. However, use of MLLR transformed models to seed prior distributions for subsequent MAP adaptation delivers the best overall results (except for subsequent MCE adaptation) achieving 64.1% word accuracy using the Afrikaans training subset and 72.0% when using the full Afrikaans training set. The results on the Afrikaans training subset are the best achieved in conjunction with the TIMIT database and represent an improvement of 19.1% (absolute) in word accuracy, or a 35% relative reduction of the word error rate, over using the Afrikaans training subset in isolation. The MLLR-MAP approach works well for relatively simple MLLR transformations (one or two regression classes), since simple transformations remove consistent bias between the source models and target data without over-specialising the models, thereby improving estimation of the prior distributions. Subsequent MAP adaptation efficiently uses the relatively large amounts of adaptation data to deliver robust parameter estimates. MAP-MLLR adaptation improves on the sensitivity of MLLR performance with respect to the number of regression classes (see Figure 7.9), but does not deliver consistently better peak performance than MLLR adaptation.

MCE adaptation of models trained on pooled multilingual data improves performance of the models, but performance for Afrikaans training set adaptation is still below that achieved with direct training on the training set. Performing MCE adaptation on models previously adapted with MLLR-MAP adaptation on the Afrikaans training set delivers the best performance of 72.7% on the Afrikaans training set. This represents an improvement of 5.1% in word accuracy, or a 16% relative reduction of the word error rate, over using the Afrikaans training set in isolation. MCE-based adaptation of best-performing Afrikaans training subset MLLR-MAP adapted models, however, does not deliver any further improvement in

performance.

The last approach for cross-language use of acoustic information that was investigated is the data augmentation approach. Models trained on target data augmented with transformed source data did not deliver improved performance. The approach is not without merit though, since models trained on the augmented data provide good prior models for MAP adaptation, achieving 61.8% word accuracy for Afrikaans training subset adaptation and 71.8% word accuracy for Afrikaans training set adaptation, which is better than that achieved with either cross-language MAP or pooling-MAP approaches.

Overall, use of the TIMIT database in addition to the Afrikaans data from SUN Speech delivers a significant improvement in performance, achieving peak improvement of between 16% and 35% reduction in relative word error rate, depending on the amount of target language data available.