Chapter 6

Cross-language recognition on SUN Speech

This chapter details the cross-language experiments performed using the English and Afrikaans speech from the bilingual SUN Speech database. The experiments compare the recognition performance of the set of cross-language adaptation strategies and algorithms discussed in the previous chapter. The use of English source data in developing models for recognising Afrikaans speech is investigated in particular. The results empirically support the proposed extensions to speaker adaptation and discriminative training algorithms and also support the newly proposed strategies for using multilingual data.

The SUN Speech database is discussed first, setting the environment for the experimental work in this chapter. The experimental protocol is discussed next, covering the selection of the parameters of the system, including various adaptation algorithm parameters. Parameter selection is a difficult task in speech recognition because there are many parameters that can influence the results. The influence of a number of the more important parameters on the recognition results are therefore shown in the experimental sections, rather than selecting just a single value as part of the experimental protocol. The experimental protocol also covers the process used to measure the results of experiments. The following sections discuss specific experiments that evaluate the various strategies for cross-language use of acoustic information. The chapter concludes with a comparison of the results from the different experiments.

6.1 The SUN Speech database

The SUN Speech database [12] contains phonetically labelled speech in both Afrikaans and English. Details of the database are given in Appendix A and only an brief overview is given here. The database contains read speech from 138 speakers totalling approximately 1500 utterances in English and 500 utterances in Afrikaans. The context of the database is limited since the English speech consists of only 40 different sentences and the Afrikaans speech of only 20 different sentences. The sentences were chosen to deliver a reasonable spread of the phonemes found in both languages. A total of 59 phonemes are used in the labelling of the database. They represent vowels, diphthongs, nasals, fricatives, affricates, glides, liquids, stops and an "other" category containing "silence" and "unknown" labels.

For the purpose of our experiments, the Afrikaans set is divided into a large training set, a smaller subset of the training set and a speaker and context independent test set, i.e. speech contained in the test set is from speakers not represented in the training set and the utterances for the test set differs from the utterances used in the training set. Details of the subdivision and composition of the database are given in Appendix A.2.

When all the available English speech data is used for training models, it amounts to 2 hours and 10 minutes of speech, which is approximately 5 times the amount of data contained in the Afrikaans training set (26 minutes of speech) and approximately 25 times the amount of data contained in the Afrikaans training subset (5 minutes of speech). For cross-language experiments it therefore makes sense to consider the English data as representing a source language with a relatively large amount of data and Afrikaans as the target language with a relatively small amount of adaptation data.

6.2 Experimental protocol

The goal of the experimental section is to evaluate the cross-language recognition performance of the various strategies and algorithms as fairly and accurately as possible, given the data that is available. For experimental purposes English is considered the source language and Afrikaans the target language, since a larger amount of labelled English speech is available in the SUN Speech database and because of the availability of other large English speech databases. Use of both the full Afrikaans training set and the training subset are evaluated to measure the effect of the amount of target language specific data on the recognition results. Initial experiments evaluate isolated phoneme recognition performance to focus on the performance of specific phonemes and classes of phonemes in the multilingual context. Later, more comprehensive experiments test continuous word recognition performance.

The results of the experiments are influenced by the parameters of the training, adaptation and recognition procedures. The selection of various parameters of the feature extraction, HMM modelling and training, duration modelling and adaptation processes is done so that system performance is nearly optimal, yet is not tuned to optimise results in favour of any particular approach.

6.2.1 General system setup

The system that is used for training and testing of the hidden Markov models was developed by the author and a colleague at the University of Pretoria. Details of the system are given in Chapter 2 and only a brief summary of the salient system parameters are given here.

Feature extraction computes 39 mel-scaled cepstral, delta and delta-delta features from 16 ms frames with a 10 ms frame advance. Continuous density hidden Markov models (HMMs) with Gaussian mixture distributions are used for modelling purposes. Strict left-to-right constraints are imposed on HMM transitions and three state HMMs are used to model each phoneme. Training proceeds using 3 stages, namely initialisation, segmental training and Baum-Welch training. Mixture splitting with stopping criteria is used to enable training of complex mixture distributions. The number of mixtures allowed is varied and up to 10 mixtures per state are allowed. A variance floor of 10^{-4} is imposed. State duration is modelled with a Gamma distribution and the duration parameters are estimated after model training is done through use of segmental training.

As far as recognition is concerned, two main categories of experiments, namely phoneme recognition experiments and word recognition experiments were performed. The experimental protocol of the two recognition approaches is discussed next.

6.2.2 Phoneme recognition experiments

Phoneme recognition is performed using a subset of 47 phonemes, including silence, from the total set of 59 phonemes. The 47 phonemes represent the labels most commonly used in labelling the Afrikaans speech, and exclude the "unknown" category as well as categories that represent less than 0.1% percent of the Afrikaans speech labels. Context independent phoneme modelling is used throughout because of the increased computational expense of context dependent modelling and also because the context of the SUN Speech database is relatively limited and differs significantly between the Afrikaans training and testing sets.

Experiments perform isolated phoneme recognition to allow a comparison to be made between the confusions that occur between the phoneme classes in the recognition process. These results indicate comparatively how well different phoneme models are seeded by their cross-language counterparts.

It is useful to consider some measure of the statistical significance of phoneme recognition results on the entire test set of 9413 labelled phones. Under the assumption of independence, for expected phoneme recognition rates in the range of 40% to 65%, the 95% confidence interval starts at between 1.4% and 1.6% in absolute phoneme recognition rate. We do not calculate confidence intervals for results from individual phonemes or phoneme groupings, as the experiments do not attempt to prove that the results differ (we are fairly sure that they should differ), but rather examine the type of differences encountered.

6.2.3 Word recognition experiments

The same 47 phoneme models that are reported on in the previous subsection are used to construct word models by connecting the phone HMMs according to a phonetic dictionary for all words occurring in utterances 11-20 of the speaker independent test set(see Appendix A for more detail). The phonetic dictionary is created by analysing the phoneme labels assigned to the speech of the 8 training subset speakers for utterances 11-20. Note that this speech does not form part of the Afrikaans test set. Multiple pronunciations of the same word are allowed, as long as at least two or more of the speakers used the given pronunciation. Using the pronunciation dictionary, in total 151 models for the 100 distinct words in the test utterances are created. In order to run a continuous speech recognition experiment, a small grammar was devised that allocates each word to one of 5 language categories comprising loosely verbs, nouns, adjectives, pronouns and conjunctives. A total of 18 transitions out of a possible 25 transitions between the 5 categories are allowed, limiting the possible sequences enough to deliver reasonable performance for continuous speech recognition in the absence of statistical language modelling.

Recognition results are obtained by aligning the output string from the recogniser with the true transcription and thereby identifying the insertions, deletions and substitutions that are needed to convert the transcription into the output string. Word *accuracy* is computed by subtracting the number of insertions, deletions, and substitutions from the number of words to be recognised and expressing this number as a fraction of the total number of words to be recognised.

It is useful to consider some measure of the statistical significance of word recognition results on the test set of 150 utterances, comprising 2096 distinct words. Under the assumption of independence of word recognition, for expected word error rates in the range of 25% to 40%, the 95% confidence interval starts at between 2.2% and 2.8% in absolute word error rate.

6.3 Initial phoneme recognition experiments

Initial experiments are performed to evaluate baseline same language (Afrikaans train, Afrikaans test) and different language (English train, Afrikaans test) recognition performance, as well as to examine some aspects of using the SUN Speech database in speech recognition experiments. Testing is done on the Afrikaans test set, consisting of 9413 labelled phonemes in continuous speech, or approximately 12.5 minutes of speech data. More details regarding the SUN Speech database and its subdivision into training and testing sets are given in Appendix A. Experiments perform isolated phoneme recognition to allow examination of how the recognition performance of individual phonemes and phoneme classes are affected in the cross-lingual scenario. When no training tokens are available for a model, the model is not used in recognition, and all test samples from the phoneme category are misclassified.

6.3.1 Overall phoneme recognition performance

Figure 6.1 shows isolated phoneme classification performance on the Afrikaans test set as a function of the model complexity allowed, for models trained on either the Afrikaans training set, training subset, or on the entire English set. As expected, using Afrikaans training data delivers models that more closely match the Afrikaans testing data and delivers a peak correct classification rate of 62.5% (10 mixtures), which is 13.5% better than the peak correct classification rate of 49.0% (10 mixtures) achieved with models trained on the 5 times larger English set. Models trained on the Afrikaans training subset achieve a peak correct classification rate of 53.1% (4 mixtures), which is 4.1% better than that achieved with models trained on the 25 times larger English set. The results indicate, at least as far as isolated phoneme recognition is concerned, that use of target language specific data may outperform using even a significantly larger amount of non-target language data.

Allowing a larger number of mixtures to be trained (allowing more mixture splitting in training), generally improves performance, as expected. Performance of models trained on the Afrikaans training set, and especially performance of models trained on the training subset, levels off at fewer mixtures than for models trained on the larger English set.

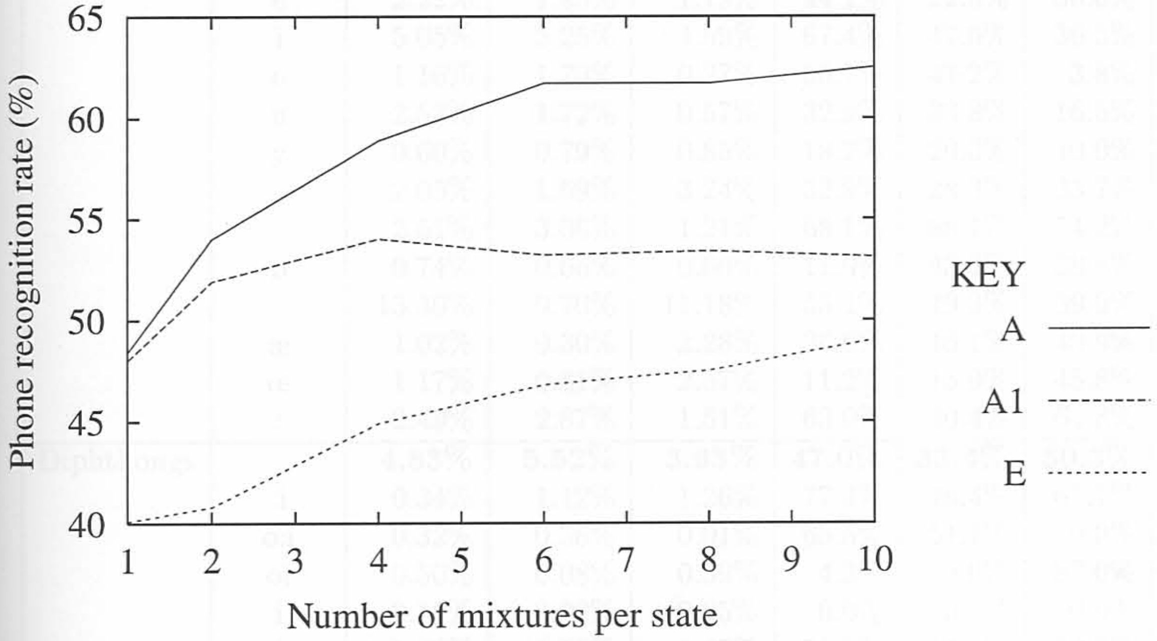

Figure 6.1: Isolated phone classification rate as a function of the number of HMM mixtures when training on the Afrikaans training set (A), the Afrikaans training subset (A1) and the entire English set (E) and testing on the Afrikaans test set

6.3.2 Individual phoneme recognition performance

The classification rates of Figure 6.1 give an overall view of relative phoneme classification rates, but it is of interest to study the classification rate of individual phoneme categories to compare their relative performance. Table 6.1 and 6.2 expands the 10 mixture per state results of Figure 6.1 by listing isolated phoneme recognition performance for each phoneme class when models trained on the Afrikaans training set, Afrikaans training subset and the English set, are tested on the Afrikaans test set.

Table 6.1: Phoneme classification rates achieved on the Afrikaans test set for models trained on the Afrikaans training set (A), training subset (A1), and the English set (E), also showing relative frequency of phonemes in the Afrikaans train (F_{Atrain}), Afrikaans test (F_{Atest}) and the English set (F_E)

Category	Symbol	F_{Atest}	F_{Atrain}	F_E	A	A1	E
Vowels		**38.37%**	**34.22%**	**32.90%**	**53.0%**	**44.3%**	**41.0%**
	a	3.53%	3.74%	2.84%	59.3%	65.5%	38.5%
	e	2.22%	1.93%	1.19%	44.1%	22.8%	50.0%
	i	5.05%	5.25%	4.59%	67.4%	47.6%	36.5%
	o	1.16%	1.70%	0.27%	55.7%	47.2%	3.8%
	u	2.53%	1.72%	0.57%	32.9%	33.8%	16.5%
	y	0.60%	0.79%	0.85%	18.2%	20.0%	40.0%
		2.05%	1.69%	3.24%	52.9%	28.3%	33.7%
		2.51%	3.06%	1.21%	68.1%	68.1%	74.2%
	ø	0.74%	0.66%	0.60%	11.9%	43.3%	38.8%
		13.30%	9.70%	11.18%	55.1%	49.3%	39.9%
	æ	1.02%	0.30%	2.28%	36.6%	16.1%	40.9%
	œ	1.17%	0.81%	2.57%	11.2%	15.0%	45.8%
	:	2.49%	2.87%	1.51%	63.0%	30.4%	64.8%
Diphthongs		**4.83%**	**5.52%**	**3.93%**	**47.0%**	**33.4%**	**30.3%**
	:i	0.34%	1.12%	1.26%	77.4%	48.4%	67.7%
	o:i	0.32%	0.36%	0.01%	65.5%	51.7%	0.0%
	oi	0.50%	0.08%	0.39%	4.3%	0.0%	87.0%
	i	0.16%	0.00%	0.05%	0.0%	0.0%	0.0%
	i	1.46%	1.50%	1.47%	51.9%	14.3%	38.3%
	ui	0.46%	0.65%	0.00%	64.3%	66.7%	0.0%
	iu:	0.30%	0.49%	0.05%	59.3%	59.3%	3.7%
	œu	0.45%	0.46%	0.60%	26.8%	26.8%	24.4%
	œy	0.84%	0.86%	0.10%	50.6%	55.8%	14.3%
Nasals		**9.36%**	**10.71%**	**11.96%**	**66.9%**	**59.9%**	**65.3%**
	m	2.11%	2.92%	2.77%	58.3%	69.8%	73.4%
	n	5.69%	6.56%	7.82%	75.5%	58.4%	67.1%
		1.56%	1.23%	1.37%	47.2%	52.1%	47.9%

Table 6.2: Phoneme classification rates achieved on the Afrikaans test set for models trained on the Afrikaans training set (A), training subset (A1), and the English set (E), also showing relative frequency of phonemes in the Afrikaans train (F_{Atrain}), Afrikaans test (F_{Atest}) and the English set (F_E)

Category	Symbol	F_{Atest}	F_{Atrain}	F_E	A	A1	E
Fricatives		**14.55%**	**15.71%**	**14.04%**	**81.4%**	**77.5%**	**62.8%**
	f	2.91%	3.39%	2.18%	86.4%	81.5%	94.3%
	h	0.11%	0.71%	0.80%	0.0%	0.0%	0.0%
	s	6.25%	6.10%	5.67%	87.2%	86.7%	73.2%
	v	1.72%	1.75%	1.90%	56.7%	48.4%	64.3%
	x	2.59%	2.65%	0.01%	94.5%	89.8%	0.8%
	z	0.53%	0.67%	1.64%	47.9%	31.2%	72.9%
		0.30%	0.44%	1.48%	66.7%	55.6%	74.1%
		0.14%	0.00%	0.36%	0.0%	0.0%	61.5%
Affricates		**1.00%**	**0.74%**	**1.77%**	**34.0%**	**25.2%**	**53.8%**
	ts^h	0.32%	0.06%	0.46%	10.3%	0.0%	31.0%
	t^h	0.68%	0.68%	1.31%	45.2%	37.1%	64.5%
Liquids		**8.75%**	**8.16%**	**6.45%**	**70.5%**	**60.4%**	**32.6%**
	r	5.20%	4.33%	2.95%	85.0%	81.2%	16.9%
	l	2.97%	3.50%	2.97%	59.0%	35.8%	60.1%
		0.58%	0.33%	0.53%	0.0%	0.0%	32.1%
Glides		**1.92%**	**1.25%**	**2.05%**	**41.1%**	**21.1%**	**56.6%**
	j	1.03%	1.15%	0.47%	67.0%	38.3%	50.0%
	w	0.89%	0.10%	1.58%	11.1%	1.2%	64.2%
Stops		**19.08%**	**14.33%**	**15.62%**	**65.9%**	**52.3%**	**53.4%**
	b	2.93%	1.54%	1.50%	63.7%	58.4%	61.4%
	d	3.51%	3.95%	2.36%	75.9%	51.9%	46.9%
	g	1.02%	0.47%	0.66%	16.1%	16.1%	52.7%
	k	3.66%	2.79%	3.21%	72.2%	59.6%	72.5%
	p	2.47%	0.95%	2.04%	47.1%	23.1%	81.3%
	t	5.49%	4.63%	5.85%	74.2%	64.4%	28.0%
Other		**2.12%**	**3.72%**	**3.42%**	**90.2%**	**85.0%**	**89.6%**
	sil	2.12%	3.72%	3.42%	90.2%	85.0%	89.6%

Same-language recognition performance

We first discuss the recognition performance of models trained on the Afrikaans training set. Classification performance achieved on vowels is only 53.0%, compared to the overall isolated phoneme classification rate of 62.5%. This differs from what is reported in literature

[26], i.e. that for English speech at least, classification performance on vowels is generally much better than on non-vowels. To some degree this is explained by the large number of vowel classes (13 in all) that are used for labelling in the SUN Speech database, as well as the specific choice of vowel classes. Some classes, especially the rounded vowels [y] and [ø] are often confused for their unrounded counterparts [i] and [e]. In English speech this distinction is not important, but in Afrikaans speech the distinction is important e.g. [mi:r] versus [my:r] and [le:n] versus [løn]. The presence of the central vowel [] in the labelling also causes confusion as both front and back vowels are often confused with it. Exact distinction of the central vowel may not be very important for word recognition, yet it has been assigned to more than 13% of the total number of labels in the test set. Classification performance on diphthongs (47.0%) is also not very good as they are often confused with vowels. Better performance (66.9%) is achieved with the class of nasals, with most of the misclassified examples also being classified as one of the other nasal categories. Somewhat surprisingly, excellent performance (81.4%) is achieved on fricatives. This can be due in part to the fact that there are only four frequently found fricatives in Afrikaans ([f], [s], [v] and [x]) which are all relatively distinct. The other categories do not deliver major surprises. We note that phoneme classes not well represented in the training set often perform very poorly in classification, with four classes even achieving 0% correct classification. This would seem to indicate over-fitting, but inspection reveals that these models contain few mixtures (between one and three), which should limit the degree of specialisation. Still, the 21 Afrikaans phoneme classes that each have less than 1% of the total training samples, together comprise 9.9% of the training set, 13% of the test set, and achieve a combined correct recognition rate of only 24%. The performance of models trained on the English data is discussed next.

Cross-language recognition performance

The average correct classification rate for English models is 49.0% compared to 62.5% achieved when training with the smaller Afrikaans training set and 53.1% achieved with the even smaller Afrikaans training subset. For phoneme classes that contain no data in the

English set, no models are trained and all test tokens are classified incorrectly. Compared with the Afrikaans training set model performance, most categories show a decrease in performance, except for affricates, which improve from 34.0% to 53.8% and glides, which improve from 41.1% to 56.6% correct classification rate. Nasals show only a 1.6% drop in classification performance from 66.9% to 65.3%. Overall performance is not poor, however, and the only phoneme categories that achieve less than 50% recognition rate are the vowels, diphthongs and liquids. Relatively poor performance of the vowel and diphthong categories is to be expected due to the differences between the language in these categories. Poor performance of the liquids is due to the English [r] model (16.9% correct) not exhibiting the diverse allophonic variations found in Afrikaans. The [x] model also achieves very poor performance (0.8% correct) since it is not a sound which occurs naturally in English. Somehow, two [x] labels were assigned to English speech, enabling training of a simple single mixture per state model.

An interesting phenomenon can be observed by comparing the results from selected classes of Afrikaans (training set) and English trained models. When one considers the phoneme classes for which the English frequency of occurrence is at least twice the Afrikaans training set frequency of occurrence, the resulting set of phonemes is [æ], [œ], [oi], [z], [], [], [ts^h], [w] and [p]. For these phoneme classes there are at least 10 times more samples in the English training set than in the Afrikaans training set. It is not too surprising then, to notice that the recognition performance for each of these phoneme classes is higher in the experiments with the English trained models than in the experiments with the Afrikaans trained models. This is indicative of the general problem of estimating distributions of such high dimensionality (39 feature dimensions plus the dimension of time), i.e. that a large amount of data is necessary to obtain robust performance.

A comparison between (small) Afrikaans training subset and English model results also delivers interesting insights. Although the overall performance of the English models is 4.1% lower than that obtained with Afrikaans training subset models (49.0% versus 53.1% correct), for 27 out of 44 phonemes (more than 60% of phonemes) the English models deliver better performance. If the arithmetic average is computed of the classification percentages of each phoneme (i.e. not taking into account the frequency of each phoneme in the test set), the English phoneme models average 45.5%, the Afrikaans trained models average 49.2% and the Afrikaans subset trained models average only 41.2%. The interpretation of isolated phoneme recognition results is problematic since the importance of individual phoneme misclassifications are not taken into account. It is therefore decided to perform continuous word recognition in subsequent experiments to represent the application of phoneme models for a useful purpose.

6.4 Multilingual data pooling

In this section we investigate multilingual data pooling in detail and perform experiments that measure word accuracy in continuous speech. Continuous word recognition experiments are performed as was discussed in Section 6.2.3 and evaluate the performance of monolingual and multilingual acoustic models for a real-world task.

Figure 6.2 shows the results achieved in continuous word recognition experiments on the Afrikaans test set of various monolingual and multilingual models. The best performance of 73.3% word accuracy is achieved by training on pooled English and Afrikaans data, followed by training on the Afrikaans training set (69.0% accuracy) and by training on the pooled English data and Afrikaans training subset (68.1% accuracy). Recognition using the English models peaks at 57.9% while training on only the Afrikaans training subset delivers an accuracy of only 45.0%. The results show a clear improvement in performance obtained by multilingual pooling versus using target language data only. Compared to using the Afrikaans training set, 4.3% absolute improvement in accuracy (73.3% versus 69.0%) is achieved by pooling with English data, and compared to using only the Afrikaans training subset, a large 23.1% absolute improvement in accuracy (68.1% versus 45.0%) is achieved by pooling with English data.

The baseline word accuracy results given in Figure 6.2 serve as reference for the results giv-

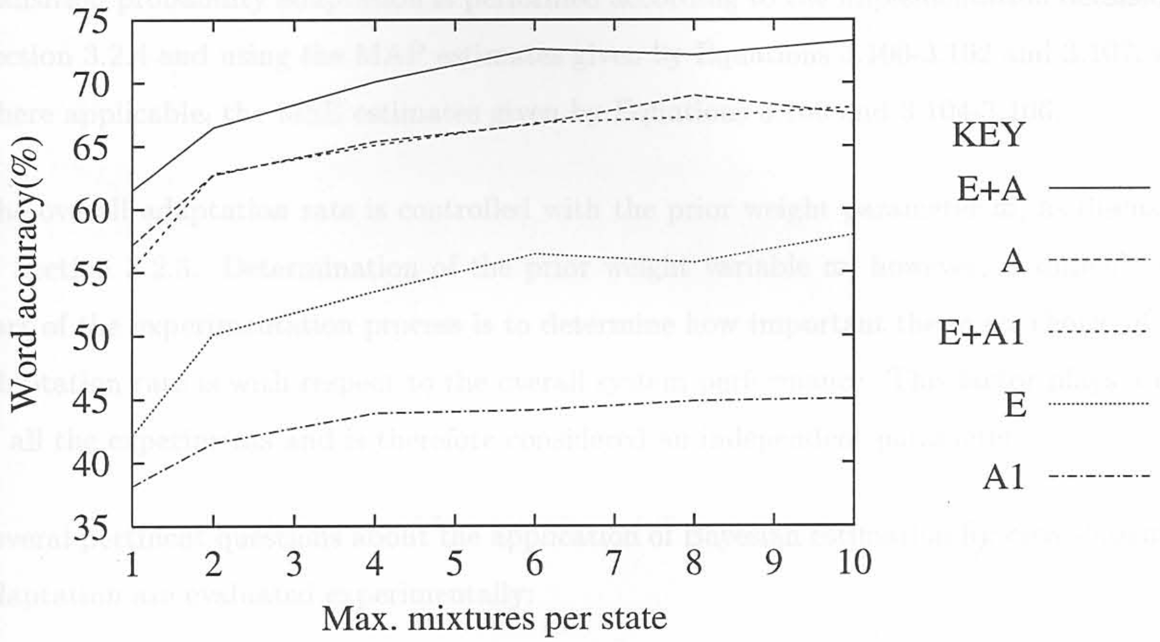

Figure 6.2: Word accuracy on the Afrikaans test set as a function of the maximum allowed number of mixtures per state for three-state HMMs trained on various monolingual and pooled multilingual data sets formed using the English set (E), the Afrikaans training set (A) and the smaller Afrikaans training subset (A1)

en in the following sections. It should be kept in mind that English and Afrikaans acoustics from the same database are used, explaining why even simple multilingual pooling delivers good results. The following sections detail experiments that evaluate the cross-language adaptation performance of techniques discussed in Section 5.3 for speaker independent speech recognition. Experiments using Bayesian, transformation-based and discriminative techniques are discussed. For all the experiments that follow, 3-state HMMs with a maximum of 10 mixtures per state are used as these represent the best performance for almost all models in Figure 6.2.

6.5 Bayesian adaptation

Experiments are performed to evaluate the application of Bayesian adaptation for cross-language adaptation as discussed in Section 5.3.1. Full mean, variance, mixture weight and transition probability adaptation is performed according to the implementation detailed in Section 3.2.4 and using the MAP estimates given by Equations 3.100-3.102 and 3.107, and where applicable, the MSE estimates given by Equations 3.100 and 3.104-3.106.

The overall adaptation rate is controlled with the prior weight parameter ϖ, as discussed in Section 3.2.5. Determination of the prior weight variable ϖ, however, is difficult and part of the experimentation process is to determine how important the exact choice of the adaptation rate is with respect to the overall system performance. This factor plays a role in all the experiments and is therefore considered an independent parameter.

Several pertinent questions about the application of Bayesian estimation for cross-language adaptation are evaluated experimentally:

- How does the amount of target language data influence the results ?
- How does the performance of cross-language model adaptation compare with adapting multilingual models using target language data ?
- How important is adaptation of variance parameters ?
- How does the performance achieved with MAP and MSE Bayesian adaptation compare?

These questions form the basis for experiments discussed next.

6.5.1 Cross-language model adaptation

Figure 6.3 shows the performance achieved as a function of the adaptation rate for English prior models adapted on the Afrikaans training set and the Afrikaans training subset. Peak performance of 74.9% word accuracy is achieved when adapting on the full Afrikaans training set, which delivers an absolute 7.3% improvement over using only the Afrikaans training set (67.6% for 3 state, 10 mixture models). This performance (74.9%) also delivers an absolute 1.6% improvement over using the pooled English and Afrikaans training set (73.3%). Even better relative performance is achieved by adaptation on the Afrikaans training subset, achieving peak performance of 70.2% word accuracy, which is 25.2% better than that achieved with the Afrikaans training subset alone (45.0%, not shown) and 2.1% better than that achieved with the pooled English and Afrikaans training subset (68.1%).

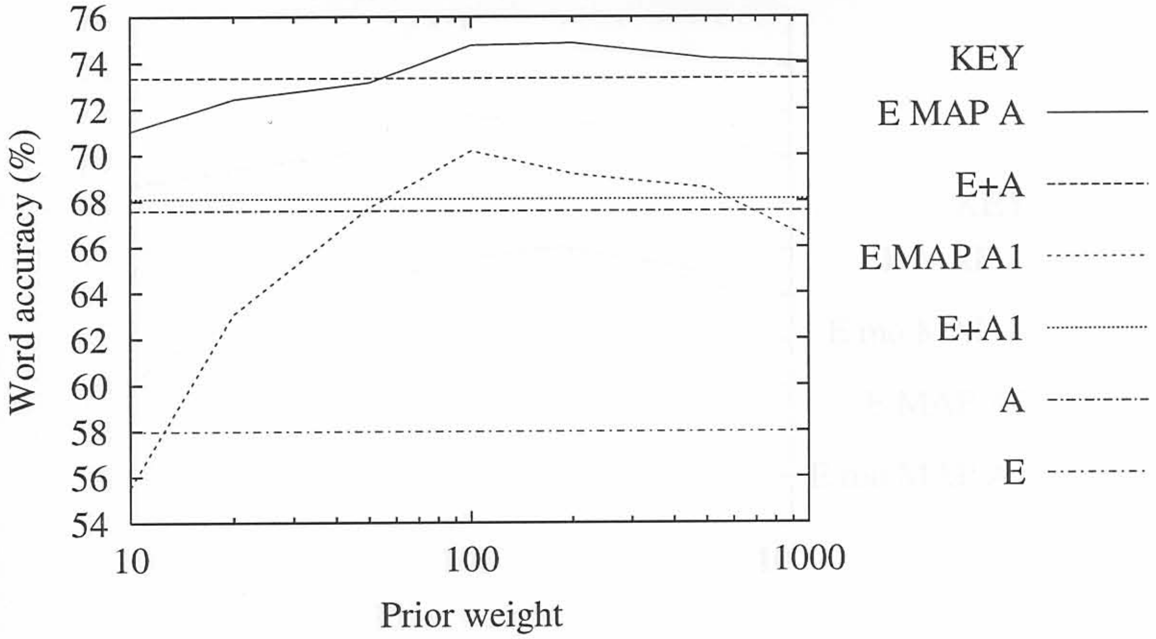

Figure 6.3: Word accuracy on the Afrikaans test set as a function of the adaptation rate for English models (E) adapted using MAP adaptation on the Afrikaans training set (A) and training subset (A1) with reference performance of monolingual and multilingual models also shown

The dependency between adaptation performance and the overall prior weight variable ϖ is apparent, with in particular, performance of adaptation with the small Afrikaans training subset suffering when ϖ is small. This is due to the fact that re-estimation on a small set delivers inaccurate estimates and therefore a larger weight should be associated with the prior to ensure good results.

6.5.2 Cross-language adaptation of variance

Figure 6.4 shows the effect of using mean-only MAP versus full MAP adaptation, which includes adaptation of the variance and mixture weight parameters when English prior models are adapted using the Afrikaans training set and training subset. It is apparent that significantly better performance is achieved when full adaptation is performed, with 3.1% degradation (74.9% versus 71.8%) in peak performance attributable to mean-only adaptation on the Afrikaans training set and 4.1% degradation (70.2% versus 66.1%) in peak performance attributable to mean-only adaptation on the Afrikaans training subset. The results indicate that adaptation of variance parameters is important to achieve good cross-language adaptation performance. However, for very small overall prior weight values ($\varpi < 20$) mean-only adaptation outperforms full adaptation on the Afrikaans training subset since variance re-estimation on little data is avoided.

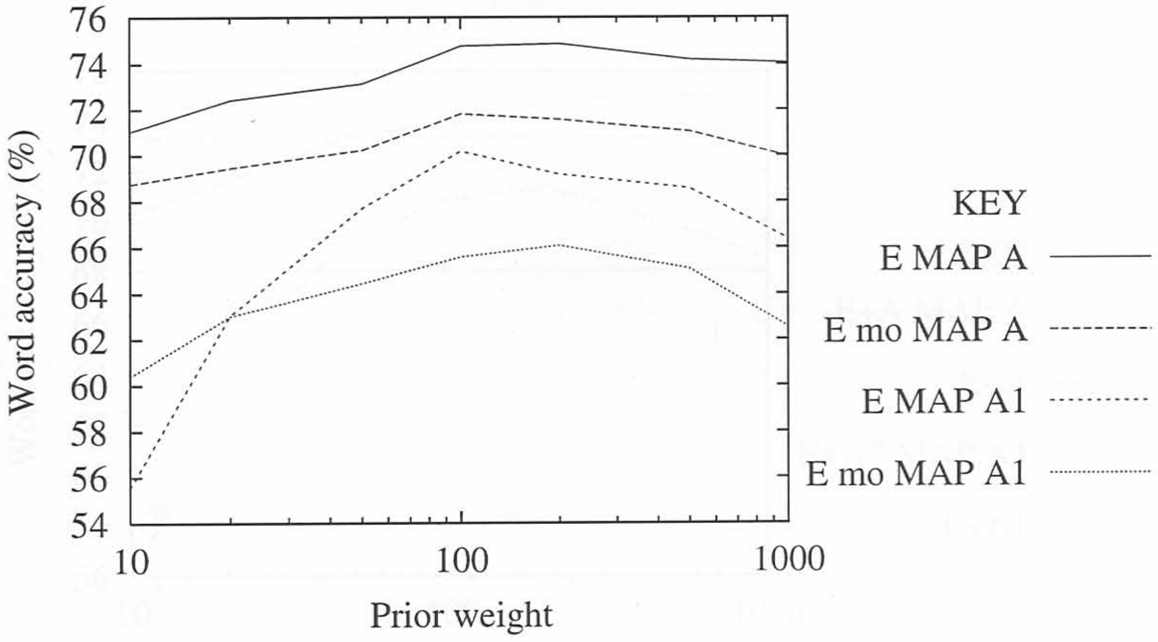

Figure 6.4: Comparison of word accuracy on the Afrikaans test set for mean-only (mo MAP) and full MAP (MAP) adaptation as a function of the adaptation rate for English models (E) adapted on the Afrikaans training set (A) and training subset (A1)

6.5.3 Data pooling followed by adaptation

Figure 6.5 shows the performance achieved as a function of the adaptation rate for bilingual prior models trained on the pooled English and Afrikaans training sets and on the pooled English and Afrikaans training subsets, when adapted on the Afrikaans training set and the Afrikaans training subset respectively. Peak performance of 75.1% word accuracy is achieved for pooling/adaptation on the full Afrikaans training set, which delivers an 1.8% improvement over data pooling (73.3%) and an additional 0.2% improvement over the direct cross-language adaptation of English source models (74.9%) in Section 6.5.1. Even better relative performance is achieved for pooling/adaptation on the Afrikaans training subset, delivering a 3.3% improvement in word accuracy (71.4% versus 68.1%) over data pooling and an additional 1.2% improvement over the direct cross-language adaptation of English source models (70.2%) in Section 6.5.1.

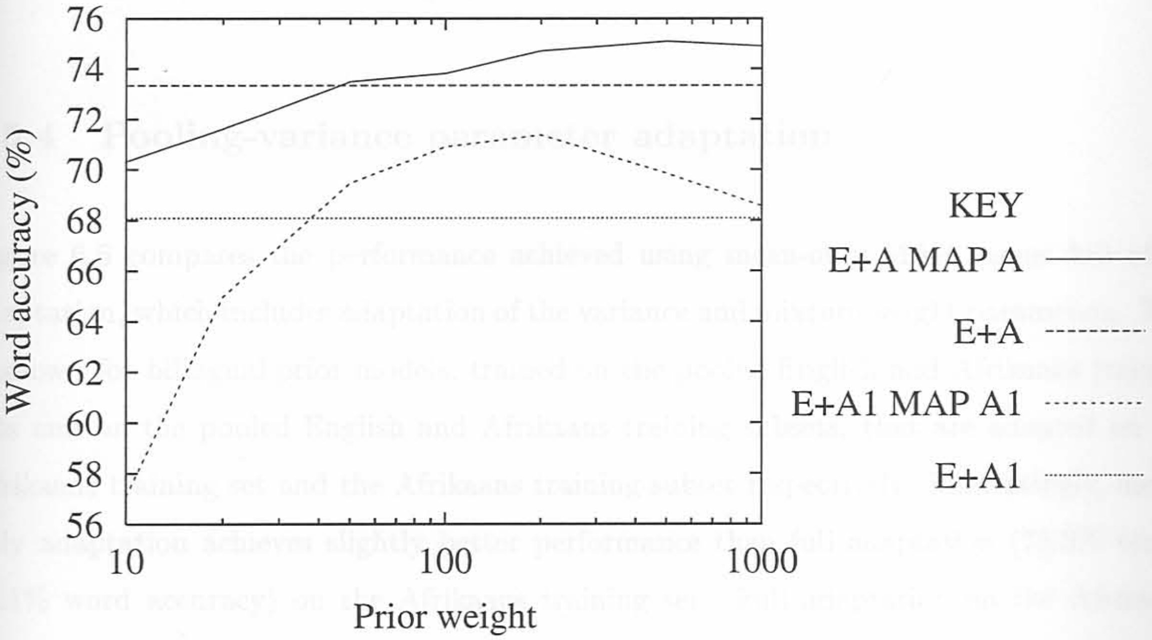

Figure 6.5: Word accuracy on the Afrikaans test set as a function of the adaptation rate for models trained on pooled English and Afrikaans training data (E+A) and pooled English and Afrikaans training subset data (E+A1) and adapted using MAP adaptation with reference performance of multilingual models also shown

The general trend that the smaller Afrikaans data set benefits more from the sharing of acoustic information with the English set than the larger Afrikaans set is to be expected, since with a sufficiently large Afrikaans data set we expect the benefit to asymptotically decrease to zero. It is interesting to observe that peak performance is achieved with larger prior weight values ($200 < \varpi < 500$) compared to direct cross-language models adaptation ($\varpi \approx 100$). This is indicative that less adaptation is required for peak performance when multilingual priors (which include the target language) are used compared to using priors from a single different source language.

6.5.4 Pooling-variance parameter adaptation

Figure 6.6 compares the performance achieved using mean-only MAP versus full MAP adaptation, which includes adaptation of the variance and mixture weight parameters. This is shown for bilingual prior models, trained on the pooled English and Afrikaans training sets and on the pooled English and Afrikaans training subsets, that are adapted on the Afrikaans training set and the Afrikaans training subset respectively. Interestingly, mean-only adaptation achieves slightly better performance than full adaptation (75.3% versus 75.1% word accuracy) on the Afrikaans training set. Full adaptation on the Afrikaans training subset, however, outperforms mean-only adaptation by 1.8% (71.4% versus 69.6%). This indicates that, when a reasonably large target language specific data set forms part of the pooled multilingual data set, variance adaptation may not be important. However, when a small amount of target specific data is used in pooling, variance adaptation may be necessary because training on pooled data may not represent the variance characteristics accurately enough.

6.5.5 MAP versus MSE estimation

So far in Section 6.5 we have been using MAP estimates, and in particular the proposed variance estimate of Equation 3.107. The next experiment compares the performance achieved with this method with the performance achieved using the biased MAP variance estimate

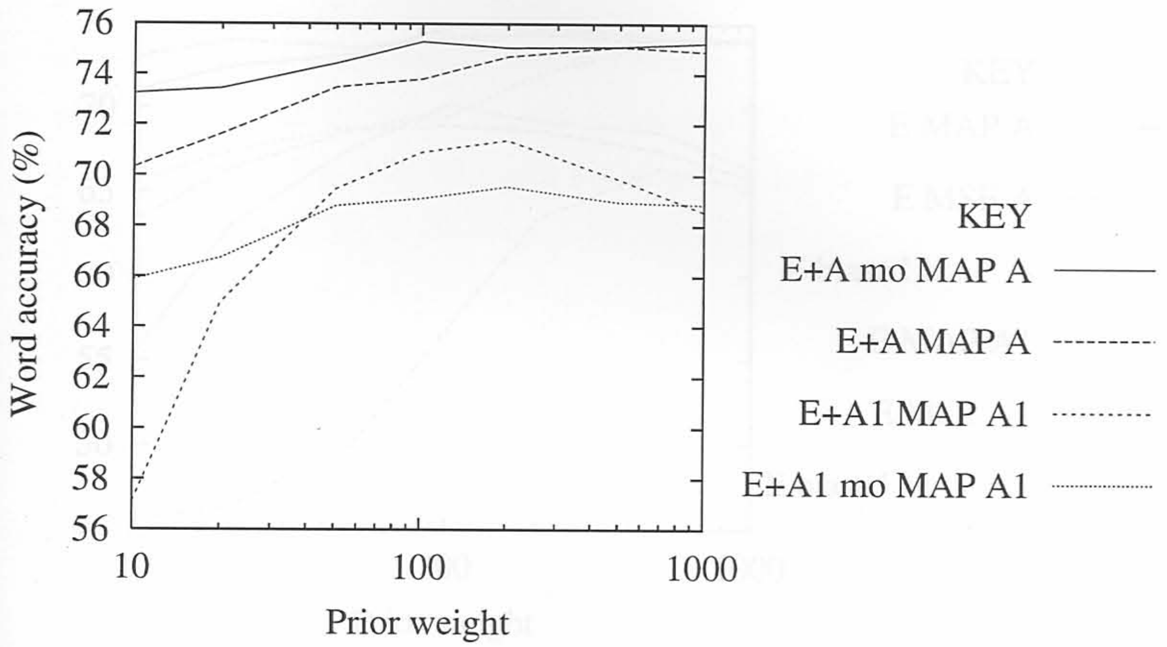

Figure 6.6: Comparison of word accuracy on the Afrikaans test set for mean-only (mo MAP) and full MAP (MAP) adaptation as a function of the adaptation rate for pooled English-Afrikaans models adapted on the Afrikaans training set (A) and training subset (A1)

of Equation 3.103 and the performance achieved with MSE variance estimation (Equations 3.100 and 3.104-3.106). Figure 6.7 shows word accuracy on the Afrikaans test set for the adaptation of English prior models on the Afrikaans training set and training subset. When adapting English priors on the Afrikaans training set, best performance of 74.9% is achieved with the proposed MAP estimate (Equation 3.107), 74.4% word accuracy is achieved with an MSE Bayes estimate, and a peak accuracy of 74.0% is achieved with the biased MAP estimate (Equation 3.103). For adaptation on the Afrikaans training subset, best performance of 70.2% is achieved with the proposed MAP estimate, 68.9% word accuracy is achieved with an MSE Bayes estimate and peak accuracy of 66.4% is achieved with the biased MAP estimate. For smaller prior weighting, the MSE estimator delivers the best performance (for $\varpi < 50$), while the performance achieved with the biased MAP estimate degrades significantly (for $\varpi < 200$).

To understand the better performance of the MSE estimate for small ϖ, we consider that the feature dimension D is basically a lower bound on the weight the MSE estimate attaches

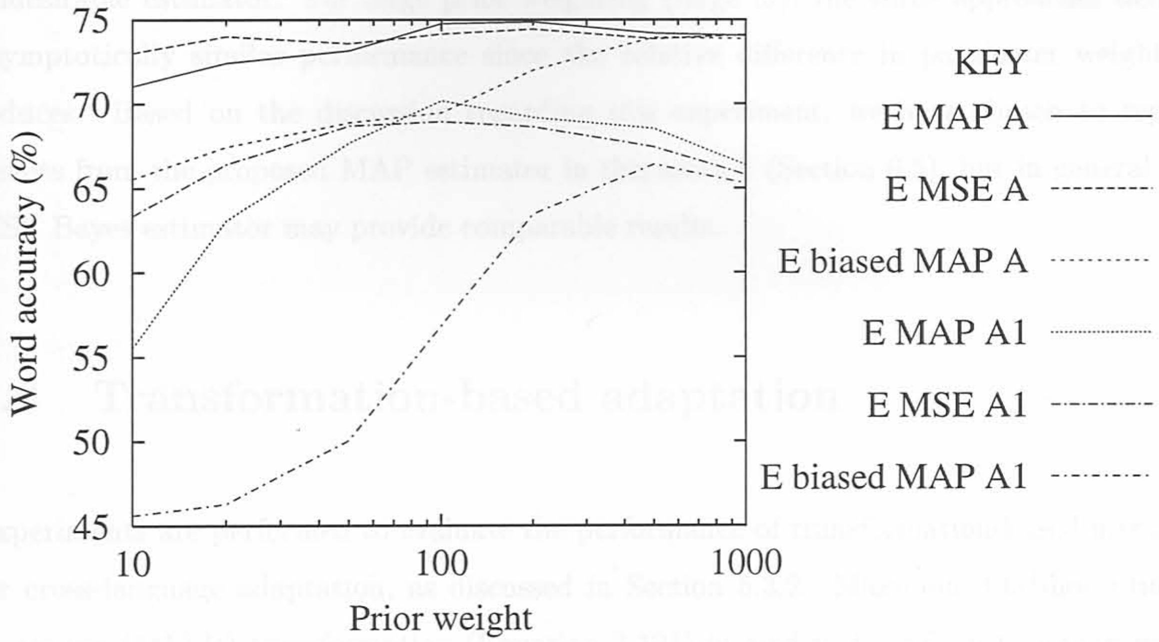

Figure 6.7: Comparison of word accuracy on the Afrikaans test set for MAP, biased MAP (using Equation 3.103 for variance estimation) and MSE Bayesian adaptation as a function of the adaptation rate for English models adapted on the Afrikaans training set (A) and training subset (A1)

to the prior variance $\hat{\Sigma}$ (see Equation 3.106), which in this case leads to improved performance since the degree to which target dependent variance re-estimation occurs is reduced. If the effective offset (of D) in ϖ is ignored, the only difference between the MAP and MSE approaches is that the MSE estimate effectively attaches less importance to the difference between the prior mean and the posterior mean estimate $(\mathbf{m}_{ik} - \hat{\mathbf{m}}_{ik})$. The reason for the proposed MAP estimate achieving better peak performance than the MSE estimate for this experiment must therefore be that it attaches greater weight to the difference between the prior mean and the posterior mean. This may have a positive influence on recognition performance because relatively larger displacements in mean position are likely to be incurred for poorly seeded distributions, in turn increasing posterior variance. Increased variance is applicable for poorly seeded models to the degree that limited target data does not allow for accurate estimation of the posterior mean.

The poor performance achieved with the standard MAP estimator for small ϖ can be attributed to its highly biased estimate of the variance for small ϖ, which makes it an undesirable estimator. For large prior weighting (large ϖ), the three approaches deliver asymptotically similar performance since the relative difference in parameter weighting reduces. Based on the discussion regarding this experiment, we have chosen to report results from the proposed MAP estimator in this section (Section 6.5), but in general the MSE Bayes estimator may provide comparable results.

6.6 Transformation-based adaptation

Experiments are performed to evaluate the performance of transformation-based methods for cross-language adaptation, as discussed in Section 5.3.2. Maximum likelihood linear regression (MLLR) transformation (Equation 3.121) is used to transform Gaussian mean parameters. In order to comprehensively adapt source model parameters, we also experiment with the adaptation of Gaussian variance parameters. The techniques used for the adaptation of the Gaussian variance parameters include:

- 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).

Relative to speaker adaptation, it is expected that cross-language adaptation will necessitate a more complex and comprehensive adaptation of source language models. In order to estimate a complex mapping, models are grouped into regression classes, with a separate transformation being calculated for each class. Grouping into classes is done according to broad phonetic groupings, i.e. for a two-class subdivision vowels/diphthongs are separated from the rest, a five-class subdivision separates vowels, diphthongs, fricatives/affricates, stops and nasals/glides/liquids and for a eight-class subdivision all mentioned categories are treated as distinct regression classes. Grouping transformations into classes has the advantage that each class of similar phonemes share a transformation, which is different from that used to transform the other classes. The assumption is that the distributions of the acoustic parameters for the target language exhibit correlation within each class. In experiments the effect of both the number of regression classes, as well as the method used for variance compensation are evaluated. We first experiment with cross-language transformation of English models using the Afrikaans sets and then evaluate the effect of transforming bilingual (pooled) models.

6.6.1 Cross-language model adaptation

Figure 6.8 shows word accuracy as a function of the number of regression classes when English models are transformed using the Afrikaans training subset only. Best performance of 65.7% is achieved with a 2-class MLLR mean/MSE log-variance transformation and second best performance is achieved with mean-only MLLR transformation, delivering peak word accuracy of 62.7%. The other techniques that adapt variance in addition to performing MLLR mean transformation perform significantly poorer and do not even improve on baseline (untransformed) English model performance. As expected, variance re-estimation performs poorly on the small Afrikaans training subset. The relatively poor results for the 5 and 8-class transformations indicate that there is not enough data to perform complex transformations contained in many regression classes. As the number of regression classes increases, performance degrades even below that achieved with direct training on the Afrikaans training subset. This happens because the English source models were trained on a large amount of data and therefore typically contain many mixtures per state (up to a maximum of 10). They are therefore easily over-fitted by the transformation on the limited amount of target data. In Section 6.7.2 we experiment with a technique that solves this problem to some degree by combining MAP with MLLR transformation.

Overall, the results in Figure 6.8 show that MLLR transformation of the Gaussian means, with or without log-variance transformation, at least delivers an improvement on baseline

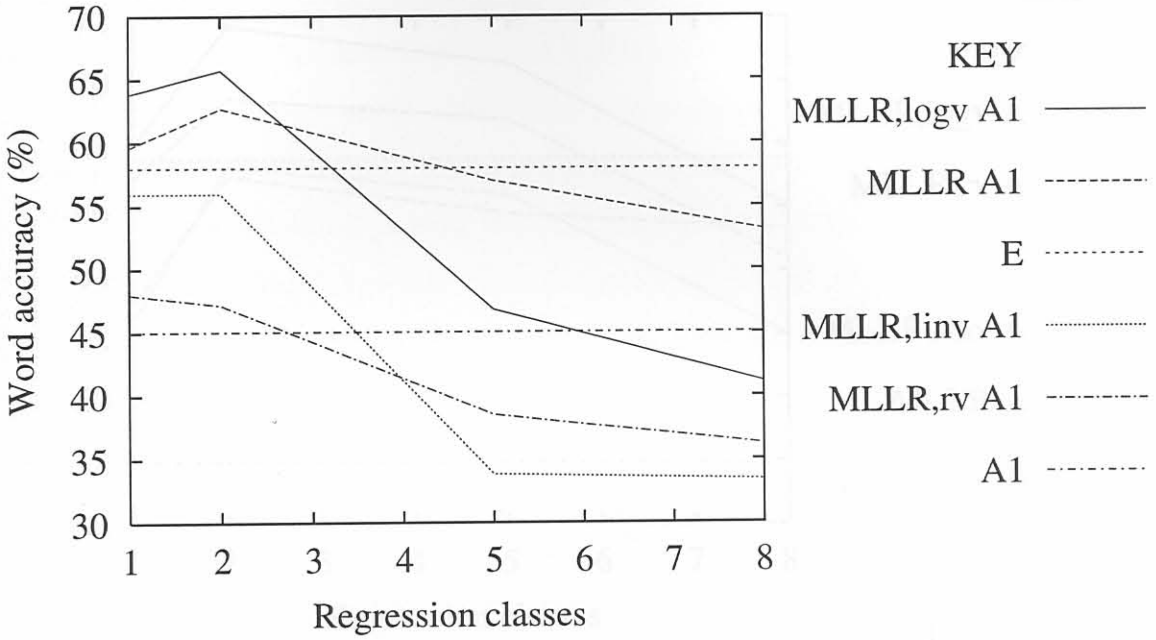

Figure 6.8: Comparison of word accuracy on the Afrikaans test set for MLLR transformation of Gaussian means combined with different variance transformation techniques: mean-only, log variance (logv), linear variance (linv) and variance re-estimation (rv) as a function of the number of regression classes for English models adapted on the Afrikaans training subset (A1)

English model performance (58.0% word accuracy) for few regression classes, but does not attain the performance achieved with bilingual pooling (68.1% word accuracy) in Section 6.4 or with cross-language MAP adaptation (peak word accuracy of 70.2%) in Section 6.5.1.

Figure 6.9 shows word accuracy as a function of the number of regression classes when English models are transformed using MLLR transformation of mean parameters combined with various techniques to adapt variance parameters on the (full) Afrikaans training set. Comparing the overall results with that of Figure 6.8, it is apparent that performance using more regression classes has improved since more target data is available.

Best performance of 71.8% is achieved with an MSE log-variance transformation when two regression classes are used. MLLR mean transformation with variance re-estimation also delivers good results (69.5% word accuracy), probably due to the fact that the Afrikaans training set is large enough for re-estimation to deliver reasonable estimates. Mean-only

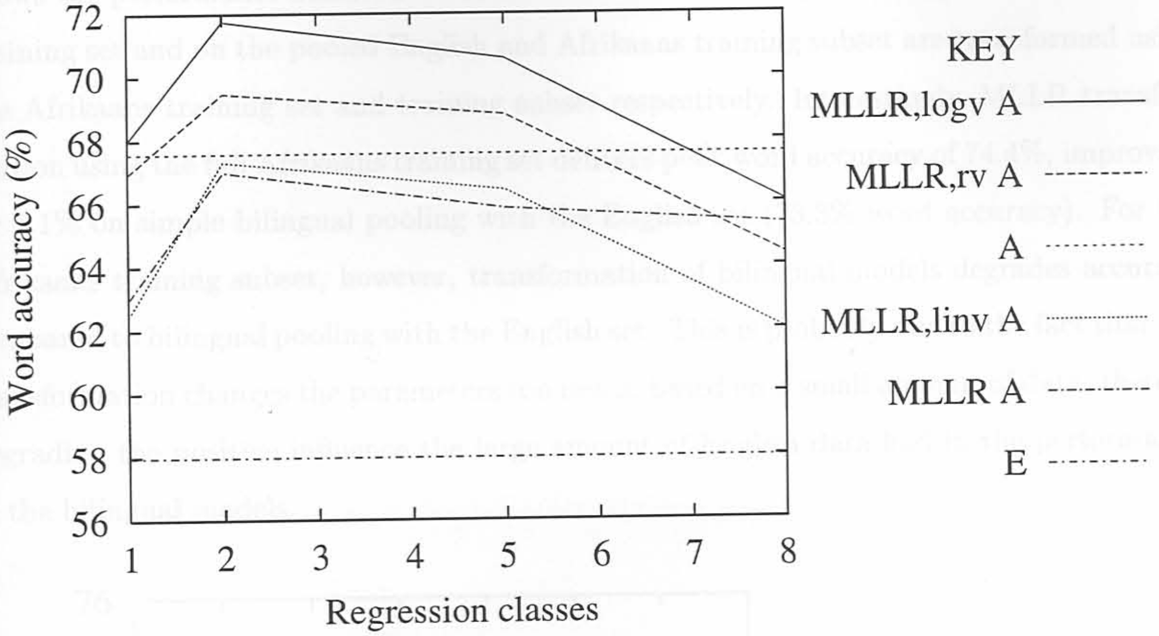

Figure 6.9: Comparison of word accuracy on the Afrikaans test set for MLLR transformation of Gaussian means combined with different variance transformation techniques: mean-only, log variance (logv), linear variance (linv) and variance re-estimation (rv) as a function of the number of regression classes for English models adapted on the Afrikaans training set (A)

(67.0%) and linear variance transformations (67.3%) show performance improvement over the baseline English models, but do not exceed the performance obtained with training directly on the Afrikaans training set. The peak word accuracy of 71.8% achieved with log-variance transformation is better than that achieved using English-only or Afrikaans-only training sets (58.0% and 69.0% respectively), but is still less than the word accuracy achieved with bilingual models (73.3%) in Section 6.4 or with cross-language MAP adaptation (74.9%) in Section 6.5.1.

6.6.2 Data pooling followed by adaptation

It was decided to evaluate the performance of transformation-based adaptation of bilingual models for the purpose of comparing the results with MAP adaptation under the same circumstances, even though the meaning of such a procedure is not intuitive. Figure 6.10

shows the performance achieved when models trained on the pooled English and Afrikaans training set and on the pooled English and Afrikaans training subset are transformed using the Afrikaans training set and training subset respectively. Interestingly, MLLR transformation using the full Afrikaans training set delivers peak word accuracy of 74.4%, improving by 1.1% on simple bilingual pooling with the English set (73.3% word accuracy). For the Afrikaans training subset, however, transformation of bilingual models degrades accuracy compared to bilingual pooling with the English set. This is probably due to the fact that the transformation changes the parameters too much, based on a small amount of data, thereby degrading the positive influence the large amount of English data had in the performance of the bilingual models.

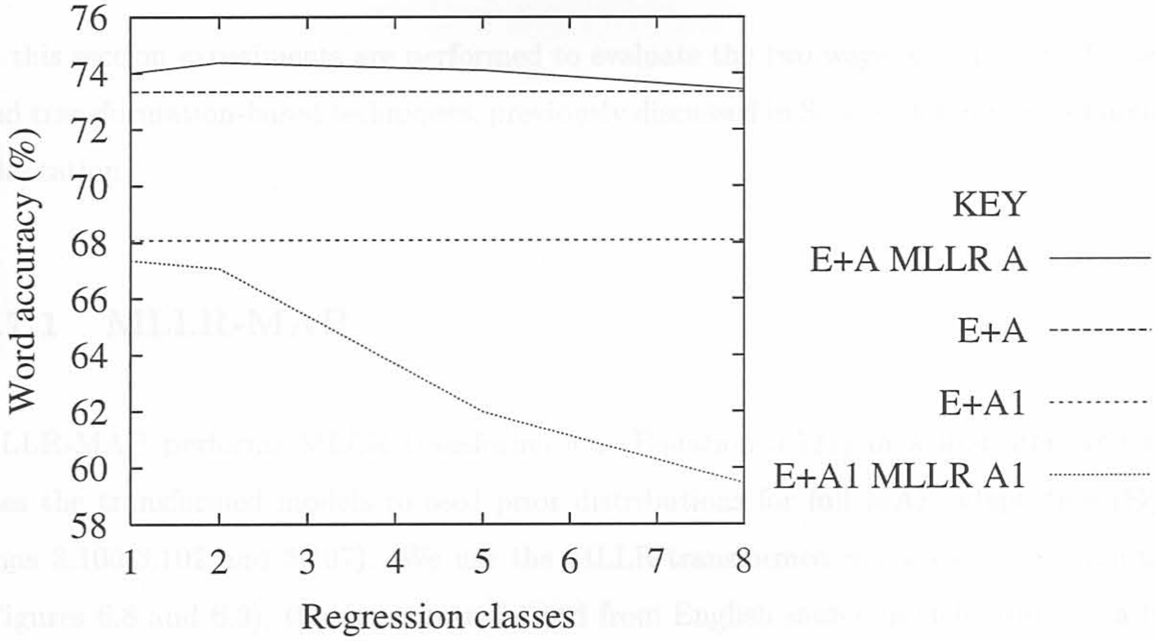

Figure 6.10: Word accuracy on the Afrikaans test set for (mean only) MLLR transformation of Gaussian means as a function of the number of regression classes for pooled English-Afrikaans models adapted on the Afrikaans training set (A) and Afrikaans training subset (A1)

Results for methods that perform variance compensation are not shown because they were found to degrade performance for bilingual model adaptation. This agrees with the results for MAP adaptation, where it was found that variance compensation of the bilingual models was less important than for the English models.

The peak word accuracy of 74.4% achieved with MLLR transformation is less than the 75.3% achieved with mean-only MAP adaptation of bilingual models in Section 6.5.4. Furthermore, transformation of the bilingual models presents a risk since it may degrade performance if too little target data is available, such as is the case for transformations calculated using the Afrikaans training subset. In the next section experiments are performed in an attempt to combine some of the advantages of both Bayesian and transformation-based adaptation.

6.7 Combined transformation-Bayesian adaptation

In this section experiments are performed to evaluate the two ways of combining Bayesian and transformation-based techniques, previously discussed in Section 3.4, for cross-language adaptation.

6.7.1 MLLR-MAP

MLLR-MAP performs MLLR transformation (Equation 3.121) in a first step and then uses the transformed models to seed prior distributions for full MAP adaptation (Equations 3.100-3.102 and 3.107). We use the MLLR-transformed models from Section 6.6.1 (Figures 6.8 and 6.9), that were transformed from English source models using data from the Afrikaans training set and training subset. These MLLR-transformed models are used as seed models for further MAP adaptation on the respective Afrikaans sets.

Figure 6.11 shows the word accuracy achieved on the Afrikaans test set as a function of the MAP prior weight when the MLLR transformed models are adapted using full MAP adaptation. Results are shown for single regression class mean-only MLLR transformations as this delivered the best performance, achieving peak performance of 74.8% word accuracy for Afrikaans training set adaptation and 69.9% word accuracy for Afrikaans training subset adaptation. The performance is slightly below that achieved with MAP adaptation of the English priors, indicating that the additional use of MLLR transformation does not improve performance in this case. This was expected since both the English and Afrikaans data are from the same database and the ability of MLLR to remove overall mismatch is not important. The results are of interest, though, for comparison with results in Chapter 7, where we show that MLLR-MAP is very useful for cross-database adaptation when significant differences exist with respect to the databases.

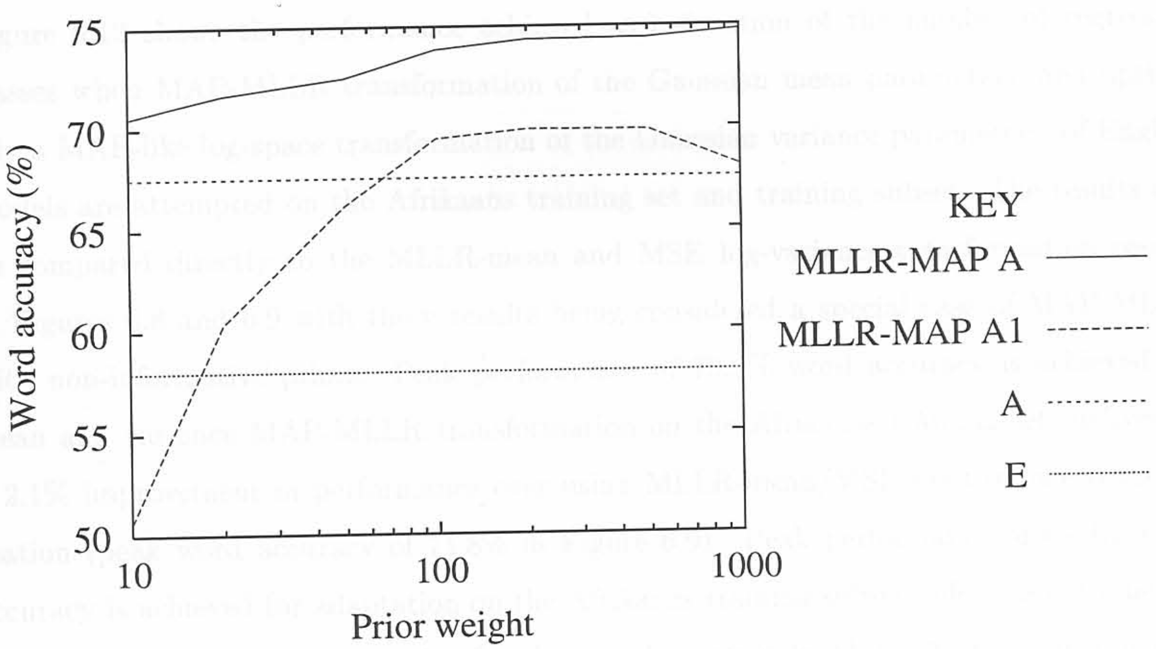

Figure 6.11: Word accuracy on the Afrikaans test set as a function of the adaptation rate for English models adapted using MLLR-MAP adaptation on the Afrikaans training set (A) and training subset (A1) with reference performance of monolingual models also shown

6.7.2 MAP-MLLR

MAPLR provides a second way of combining Bayesian and transformation-based adaptation and attempts to determine the linear regression parameters that deliver the maximum *a posteriori* probability model estimate. We have combined MAPLR (Equations 3.138 and 3.139) with a MAP-like variance adaptation technique (Equations 3.140 and 3.141) and group the combination of the techniques under the name MAP-MLLR. The MAP-MLLR transformations converge to unity transformations as the amount of adaptation data available decreases and converge to the MLLR (for Gaussian means) and log-variance MSE (for Gaussian variance) estimates as the amount of adaptation data available increases. The amount of adaptation that source models incur under the transformation can be controlled and performance can therefore be improved by decreasing the degree of over-fitting, especially for complex transformations that span many regression classes.

Figure 6.12 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 English models are attempted on the Afrikaans training set and training subset. The results can be compared directly to the MLLR-mean and MSE log-variance transformation results in Figures 6.8 and 6.9 with those results being considered a special case of MAP-MLLR with non-informative priors. Peak performance of 73.9% word accuracy is achieved for mean and variance MAP-MLLR transformation on the Afrikaans training set, delivering a 2.1% improvement in performance over using MLLR-mean/MSE log-variance transformation (peak word accuracy of 71.8% in Figure 6.9). Peak performance of 65.9% word accuracy is achieved for adaptation on the Afrikaans training subset, which is 0.2% better than using a non-informative prior for the transformation (peak 65.7% word accuracy in Figure 6.8). The results in Figure 6.12 also show that variance transformation (in addition to mean transformation) significantly outperforms mean-only transformation, by 5.9% on the Afrikaans training set (73.9% versus 68.0% word accuracy) and by 2.6% on the Afrikaans training subset (65.9% versus 63.3% word accuracy).

6.8 Discriminative adaptation

Experiments are performed to evaluate the application of discriminative adaptation for cross-language adaptation, as was discussed in Section 5.3.3. A major consideration when

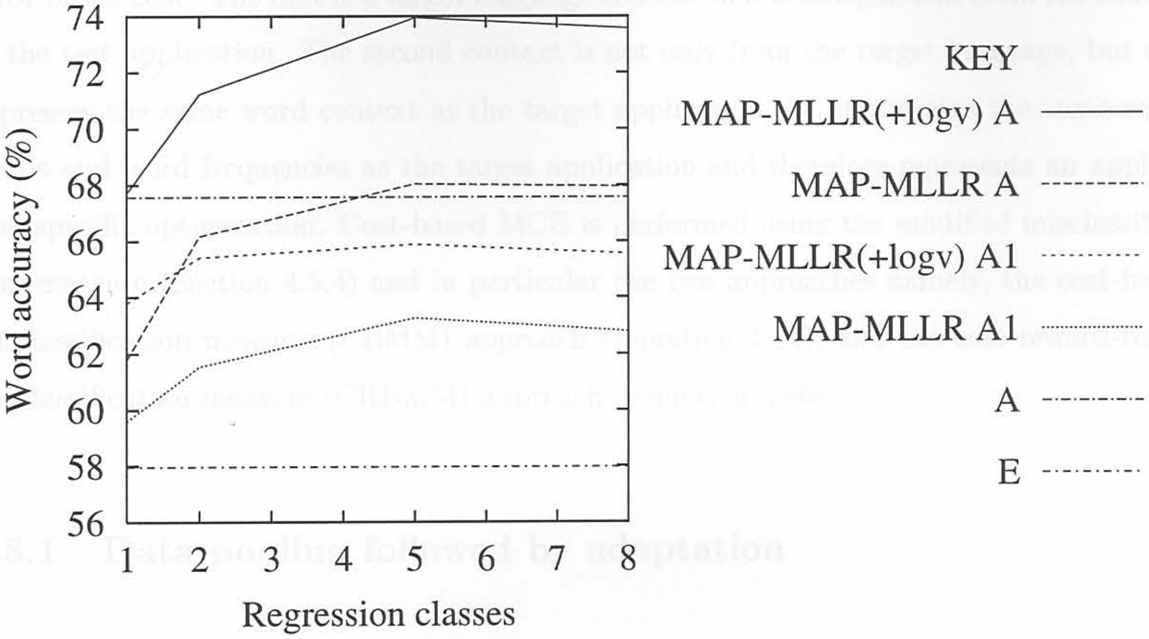

Figure 6.12: Comparison of word accuracy on the Afrikaans test set for MAP-MLLR transformation of Gaussian means, optionally combined with a MAP-like log-space (+logv) MSE transformation of Gaussian variance parameters, computed as a function of the number of regression classes for English models adapted on the Afrikaans training set (A) and training subset (A1)

applying discriminative adaptation is the selection of initial model parameters, as discriminative techniques are prone to converge to local minima (in terms of the loss function). The initial model must therefore be selected to exhibit desirable characteristics and discriminative optimisation is performed only to "fine-tune" the characteristics for the target language. We experiment with initial models that are trained on pooled multilingual data, as well as with models that are the product of other adaptation techniques, such as MAP adaptation, and therefore have already been specialised to some extent for improved target language performance.

The MCE framework for discriminative training from Chapter 4 is used and in particular experiments are performed to determine the performance of the cost-based extensions to MCE that we proposed in Section 4.5. A method from Section 4.5.3 is used to calculate the word error-based cost associated with each phoneme misclassification (in particular Equation 4.43). In experiments, two sets of word contexts were used to derive the word error-based cost. The first is a target language context that is independent from the context in the test application. The second context is not only from the target language, but also expresses the same word context as the target application, i.e. it contains the same set of words and word frequencies as the target application and therefore represents an application specific optimisation. Cost-based MCE is performed using the modified misclassification measure (Section 4.5.4) and in particular the two approaches namely, the cost-based misclassification measure (CBMM) approach (Equation 4.44), and the cost-reward-based misclassification measure (CRBMM) approach (Equation 4.46).

6.8.1 Data pooling followed by adaptation

MCE adaptation is performed on models trained on pooled English and Afrikaans data. The multilingual models are trained on a large amount of data, ensuring robust parameter estimation, although the models will be biased towards the source language since it represents most of the training data. Discriminative adaptation is performed, using target language data only, to adapt the multilingual models with the aim of improving performance specifically for the target language.

Figure 6.13 shows word accuracy on the Afrikaans test set as a function of the number of adaptation iterations when models trained on the pooled English and Afrikaans training subset (68.1% word accuracy) are used as initial models for MCE adaptation on the Afrikaans training subset. Peak performance of 71.3% word accuracy is achieved with target context CBMM MCE adaptation, which is 3.2% better than the performance of the baseline multilingual models. Similar peak performance is achieved with target context CRBMM MCE adaptation and (independent context) CBMM (both achieve 71.2% word accuracy). MCE adaptation (without a modified misclassification measure) achieves peak performance of 70.6%, which is still 2.5% better than the performance of the multilingual initial models. The best performance of 71.3% word accuracy is, however, 0.1% below the 71.4% word accuracy achieved with MAP adaptation of multilingual models in Figure 6.5. The results indicate that the CRBMM approach to MCE adaptation does not offer im-

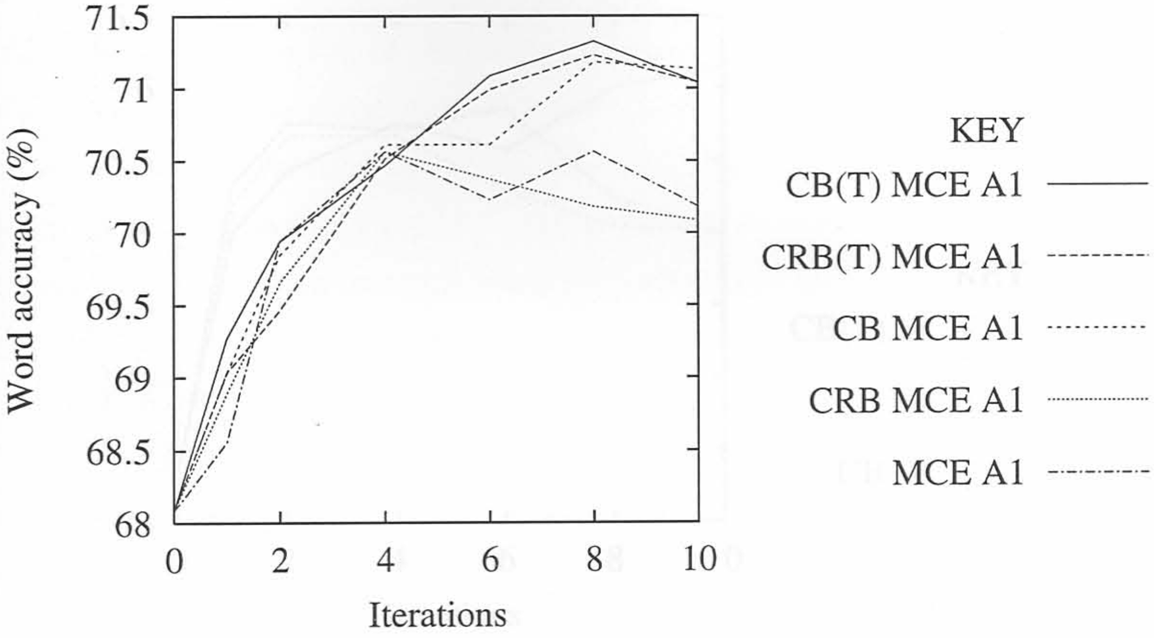

Figure 6.13: Comparison of word accuracy on the Afrikaans test set for the MCE adaptation of pooled English-Afrikaans models on the Afrikaans training subset, also including use of cost-based (CB) and cost-reward-based (CRB) misclassification measures, optionally designed specifically for the target context (T)

proved performance over the CBMM approach (this is also generally the case for the other experiments) and in following experiments we therefore discuss only the CBMM approach for incorporating cost into MCE adaptation.

Figure 6.14 shows word accuracy on the Afrikaans test set as a function of the number of adaptation iterations when models trained on the pooled English and Afrikaans training set (73.3% word accuracy) are used as initial models for MCE adaptation on the Afrikaans training set. Peak performance of 76.1% word accuracy is achieved with target context CBMM MCE adaptation, which is 2.8% better than the performance of the multilingual initial models and is also 0.8% better than the best performance previously reported in this chapter, namely the 75.3% word accuracy achieved with MAP adaptation of multilingual models in Figure 6.6. MCE adaptation (without CBMM) delivers peak performance of 75.9% and CBMM MCE delivers peak performance of 75.8% word accuracy. For Afrikaans training set adaptation, all of the MCE adaptation techniques therefore achieve better performance than the best performing non-MCE techniques that were evaluated.

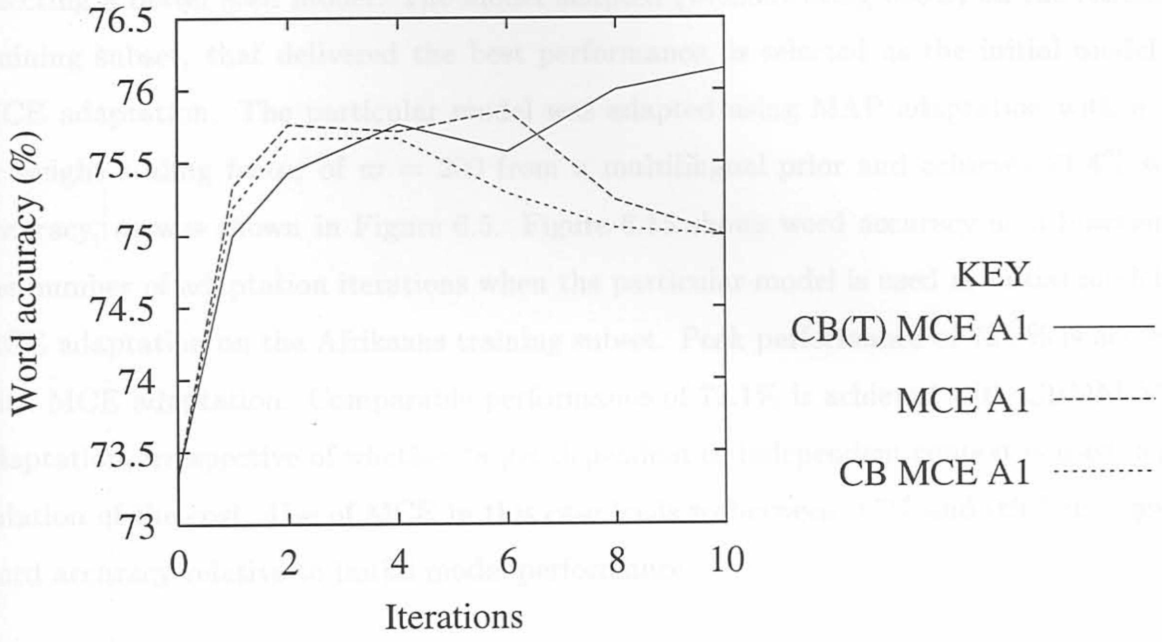

Figure 6.14: Comparison of word accuracy on the Afrikaans test set for the MCE adaptation of pooled English-Afrikaans models on the Afrikaans training set, also including use of a cost-based (CB) misclassification measure, optionally designed specifically for the target context (T)

The experiments with the different approaches to MCE adaptation show significant improvements in performance compared to the performance of multilingual initial models used for adaptation. Target context cost-based MCE adaptation delivers the best performance for multilingual model adaptation, increasing performance by 3.2% for Afrikaans training subset adaptation and by 2.8% for full Afrikaans training set adaptation. The results achieved on the Afrikaans training set are the best results that are reported. The results on the Afrikaans training subset are 0.1% lower than the best non-MCE adaptation method and an attempt is made in the next section to improve on this performance.

6.8.2 Improving best performing models

The performance achieved with MCE adaptation on the Afrikaans training subset (Figure 6.13) is less than the best performance achieved without using MCE. An experiment is performed to test whether performance of the MCE technique can be improved upon by selecting a better seed model. The model adapted (without using MCE) on the Afrikaans training subset, that delivered the best performance, is selected as the initial model for MCE adaptation. The particular model was adapted using MAP adaptation with a prior weight scaling factor of $\varpi = 200$ from a multilingual prior and achieves 71.4% word accuracy, as was shown in Figure 6.5. Figure 6.15 shows word accuracy as a function of the number of adaptation iterations when the particular model is used as initial model for MCE adaptation on the Afrikaans training subset. Peak performance of 72.3% is achieved with MCE adaptation. Comparable performance of 72.1% is achieved with CBMM MCE adaptation, irrespective of whether target dependent or independent context is used in calculation of the cost. Use of MCE in this case leads to between 0.7% and 0.9% increase in word accuracy relative to initial model performance.

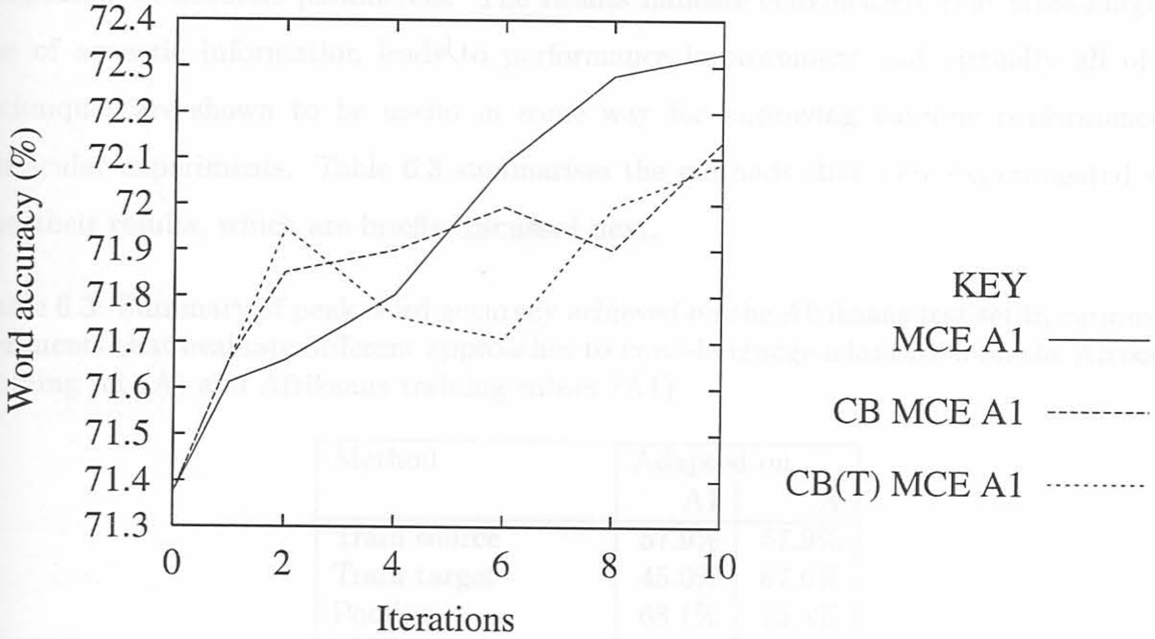

Figure 6.15: Comparison of word accuracy on the Afrikaans test set for the MCE adaptation of models that have already been optimised for performance on the Afrikaans training subset, also including use of cost-based (CB) and cost-reward-based (CRB) misclassification measures, optionally designed specifically for the target context (T)

The same approach, namely to use an initial model that delivers better performance than the multilingual models trained on pooled data, is attempted for the full Afrikaans training set adaptation. However, use of the MAP adapted model (from Figure 6.6) that produced the best results, as initial model for subsequent MCE adaptation does not deliver improved performance. Initial model performance of 75.3% word accuracy on the Afrikaans test set is only degraded by further MCE adaptation on the Afrikaans training set and is therefore not shown graphically. The result suggests that use of an initial model that achieves better word accuracy does not necessarily ensure that better final performance will be achieved.

6.9 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 adaptation of acoustic parameters. The results indicate convincingly that cross-language use of acoustic information leads to performance improvement and virtually all of the techniques are shown to be useful in some way for improving baseline performance in particular experiments. Table 6.3 summarises the methods that were experimented with and their results, which are briefly discussed next.

Table 6.3: 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	57.9%	57.9%
Train target	45.0%	67.6%
Pooling	68.1%	73.3%
MAP	70.2%	74.9%
Pooling-MAP	71.4%	75.3%
Transformation	65.7%	71.8%
MLLR-MAP	69.9%	74.8%
MAP-MLLR	65.9%	73.9%
Pooling-MCE	71.3%	**76.1%**
Pooling-MAP-MCE	**72.3%**	75.3%

The relatively good results achieved with English (source) language models, as well as the good results achieved with simple multilingual pooling should be seen in light of the "closeness" of the match between the SUN Speech English and Afrikaans data sets. Because multilingual data from a single database is used, there are no differences with respect to recording conditions between the data sets and also a consistent set of labels were used. This situation facilitates easy cross-language use of speech data. Cross-language MAP adaptation delivers good results, improving even further when adaptation is done from multilingual models, and achieves the best results of the non-discriminative adaptation approaches.

Cross-language transformation-based adaptation does not deliver very good performance and in isolation does not even achieve the level of performance achieved by the multilingual (pooling approach) models. MLLR-MAP delivers good performance, but performance is still less than that achieved with MAP adaptation in isolation - meaning that even the simplest transformation degrades the priors. MAP-MLLR improves upon using MLLR alone, allowing transformation-based adaptation to exceed pooling performance on the Afrikaans training set.

MCE adaptation delivers the best overall performance on both the Afrikaans training set and training subset, irrespective of whether the CBMM approach is used. Use of target context CBMM, in particular, achieves improved performance when adapting multilingual initial models and achieves the best overall performance of 76.1% word accuracy for Afrikaans training set adaptation. For adaptation on the Afrikaans training subset, MCE adaptation of multilingual initial models previously adapted with MAP adaptation (denoted pooling-MAP-MCE in Table 6.3) delivers the best performance of 72.3% word accuracy. MCE-based adaptation of best-performing Afrikaans training set adapted models does, however, not deliver any further improvement in performance.

This chapter discussed experiments performed to evaluate different strategies and techniques for cross-language use of acoustic information. In particular the use of English data from the SUN Speech database in addition to Afrikaans data, also from SUN Speech, was investigated for the purpose of improving recognition performance on an independent Afrikaans test set. The results indicate that significant performance improvement is attained by use of the English data in addition to the Afrikaans data, achieving an improvement of 27.3% (72.3% versus 45% word accuracy), or a 50% relative reduction in word error rate over using the Afrikaans training subset alone and an improvement of 8.5% (76.1% versus 67.6% word accuracy), or a 26% relative reduction in word error rate compared with using only the Afrikaans training set. Use of English data in addition to a small amount of Afrikaans data (the training subset) outperforms using five times more Afrikaans data (the full training set) by 3.3% (72.3% versus 67.6%). In the next chapter we investigate to what extent this gain in performance extends to use of acoustic information across different databases.