Chapter 5

Cross-language acoustic adaptation issues

In this chapter we discuss the framework for the experiments detailed in the following two chapters. Practical aspects regarding the cross-language use of algorithms detailed in the previous two chapters are covered as part of this framework. Language and database issues are also discussed to place the experiments that were performed in the proper perspective.

Cross-language use of acoustic information attempts to exploit the acoustic-phonetic similarities between languages. These similarities are evident from the use of international phonetic inventories, such as the International Phonetic Alphabet[1] (IPA) and Speech Assessment Methods Phonetic Alphabet[2] (SAMPA), that serve to classify the sounds of many languages. There are still, however, differences with respect to the acoustic properties of sounds from different languages that share the same labels. Also, often a target language may contain sounds that do not occur in languages for which large databases are available. Labelling conventions, recording conditions and the type of speech recorded may also differ between databases, making cross-language and cross-database use of acoustic information

[1] http://www2.arts.gla.ac.uk/IPA/fullchart.html

[2] http://www/phon.ucl.ac.uk/home/sample/home.html

a formidable task. Aspects regarding language and database specific issues are discussed in this chapter to facilitate cross-language use of acoustic information.

We consolidate the application of methods discussed in the previous three chapters for using acoustic information across language boundaries. The strategies that have been used in previous research are (i) training on pooled multilingual data and (ii) adapting models trained on one language using data from another. There are, however, as we shall discuss in this chapter, other strategies that can be followed such as (iii) training models on pooled source and target language data and then adapting the models using only target language data and (iv) cross-language transformation of source data followed by training on the pooled target and transformed source data.

In addition to the different strategies for cross-language adaptation that will be discussed, the algorithms used for adaptation or transformation should also be examined to use them efficiently for the specific purpose. In previous chapters we already proposed new techniques to improve cross-language adaptation performance, such as the MSE log-variance transformation, the discriminative adaptation of duration parameters and use of a word error-based cost function in discriminative phoneme adaptation. These techniques are not part of the published repertoire of speaker adaptation algorithms, perhaps because speaker adaptation differs essentially from cross-language adaptation. In this chapter we therefore reconsider how to apply the standard adaptation methods together with our extensions thereof for cross-language adaptation.

5.1 Language and database issues

The acoustic-phonetic similarities between languages are well documented in international phonetic inventories. The existence of such standards is central to our goal of using acoustic information across language boundaries. The international phonetic inventories were developed by phoneticians using expert phonetic knowledge, however, and not by using acoustic measurements or statistical techniques. There is thus no guarantee that these phonetic inventories represent an optimal classification of the acoustic properties of speech in different languages. Indeed, even for the same phonetic category, the acoustical realisation of the phoneme may differ between different languages for a number of reasons such as [48]:

- different phonetic context due to different phoneme sequence statistics and different phoneme inventories,
- different speaking styles,
- different prosodic features and
- different allophonic variations.

According to the principle of sufficient acoustic separation, the set of sounds in a language are kept acoustically distinct by its speakers to make it easy to distinguish between the sounds. Because the phonetic inventories of languages differ, the positions of the boundaries between phonemes are language dependent. In spite of the differences between languages with respect to the characteristics of speech of the same phonetic category, we still expect reasonable overlap between the phoneme feature distributions of different languages and that the phonetic categories give a good indication as to how the overlap occurs. This reasoning is supported by empirical evidence from systems with explicitly multilingual phone sets. In the next section we examine how differences with respect to phoneme inventories and context may influence the usefulness of speech data.

5.1.1 Phonetic inventories and context

Cross-language use of acoustic-phonetic information is limited to the overlap or junction of the phoneme inventories of the languages. When considering cross-language use of acoustic-phonetic information, it is therefore important to attempt to find languages that are as similar as possible. This ensures that maximal overlap of phonetic inventory as well as overlap of phonetic context between the languages occur. Significant overlap of phonetic context ensures that monophone models contain to a large degree the same built-in context information and also facilitates the cross-language use of data to train context dependent phoneme models. In this thesis exclusive use was made of two Germanic languages, namely English and Afrikaans, mainly because (i) suitable databases were available for these languages and (ii) research on South African languages are of particular interest to us. Two different databases were used, one containing both South African English and Afrikaans, and one containing American English only.

As far as the specific context of the source language databases are concerned, it is better if the source language databases are phonetically diverse and contain a large variety of contexts. This reduces the specialisation of the source language acoustics for specific contexts and may improve the performance of models for recognition tasks containing speech from an entirely different context [112], such as is typically expected for a cross-language task. In our case, both speech databases are phonetically rich and the American English database also has diverse context. The phonetic context of the bilingual database, however, is not very diverse as a large number of utterances of only 60 different sentences are used.

5.1.2 Labelling conventions

In order to exploit acoustic-phonetic information, phonetically labelled databases are necessary in both source and target languages. Consistent labelling conventions should be followed and the selection of the set of phoneme labels to use should be taken with care. Although an international phoneme inventory (such as IPA) may be used, a subset of the inventory is usually selected that covers the expected occurrence of phonemes in the speech of the database. Using a limited number of phoneme categories has the advantages of simplifying the labelling process and possibly reducing the number of incorrectly assigned labels. On the other hand, a small number of labels may group together phonetic categories, which separately could provide useful information. If a database is created for the explicit purpose of multilingual speech recognition, then use of a larger set of labels that suitably covers the phonetic variety over the combination of languages may facilitate the development of multilingual and cross-lingual application of the database. The bilingual database that is used in this thesis is of this nature, using a consistent set of labels for both languages. This is very convenient because it enables experiments to better quantify the effect of actual acoustical differences between the languages rather than possible artifacts of labelling differences.

An important aspect with respect to the accuracy of a database is the extent to which the database is labelled using phonetic or phonemic considerations. The purpose of labelling is usually to assign phonetic categories on an acoustical basis to the speech. This implies that phonetic labelling is attempted. Phonetic labelling is, however, a difficult and tedious task and it is often easier for the person performing the labelling to assign a label to a sound segment on a phonemic basis, i.e. on what was supposed to have been said, rather than on what was actually said. Perhaps the easiest way to assign labels to a speech database is by forced alignment, delivering a purely phonemic segmentation of the speech. For instance, in bootstrapping procedures this is the only solution because phonetic labelling of the target language database is not done. When source language models trained with forced alignment are used for a closed vocabulary task, or even for an open vocabulary same-language application this may not have a severe effect beyond the loss of acoustic resolution. It may even improve recognition performance in continuous speech when an imperfect pronunciation model is used. However, for a cross-language task, the loss of acoustic resolution, coupled with the incorporation of incorrect (source) language specific phonemic information, is likely to degrade performance for target language applications. For both databases used in this thesis, a phonetic labelling approach was used.

5.1.3 Phonetic mapping

Generally, if identical labelling conventions are used in the creation of the source and target databases, no work needs to be done at the phonetic level in determining how to implement cross-language use of the data. This is also the case for the bilingual database used in this thesis. Several multilingual speech recognition studies have demonstrated the efficient re-use of acoustic-phonetic information across multiple languages [19, 20]. It is when differences exist with respect to labelling, i.e. a one-to-one mapping of phonemes does not exist, that difficulties are encountered. Two approaches for determining how to use the acoustic-phonetic information across language boundaries are generally applied namely

- phonetic knowledge-based and
- distance measure-based

methods for pairing phonemes or groups of phonemes from multiple languages. In our case the process is somewhat simplified since only a one-way mapping from source language(s) to target language is desired. Both approaches were attempted in this thesis and it was found that the phonetic knowledge-based mapping approach delivered better performance than a distance measure-based mapping approach. The experiments and results are discussed in Section 7.1.

Phonetic knowledge-based mapping

Expert phonetic knowledge can be used to determine a mapping from the phonetic inventory of a source language database to the phonetic inventory of a target language database. As previously mentioned, this is the only viable approach for research on bootstrapping of target language models, since target language data is not labelled. Research on explicit multilingual phoneme-based recognition often also makes use of phonetically derived sharing of acoustic parameters e.g. [19, 49]. For this thesis a phonetic expert determined a mapping from the American English database to the bilingual database, details of which are given in Appendix B.

Many multilingual systems use phonetically derived categories with multilingual scope, but then use statistical procedures to select whether to consider the same phoneme in the different languages as one or whether to model them separately [30]. The procedures have also been extended to the creation of generalised triphone models of arbitrary complexity by using a decision tree clustering approach with both language and context questions in the splitting procedure. These approaches were not followed because the amount of data was deemed to be too limited.

Distance measure-based mapping

Research has shown the use of a metric such as the Bhattacharyya distance

$$D_{\mathrm{Bhat}} = \frac{1}{8}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \frac{1}{2}\log\frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}} \tag{5.1}$$

to measure the distance between Gaussian distributions representing phone classes of the same language for clustering purposes [113]. This metric has been used in a multilingual context to merge arbitrary phonemes from multiple languages [20] to reduce the complexity of the multilingual models. This entailed computing the distance between the phoneme models of the different languages and merging phonemes for which the distance measurement was below a pre-set threshold. Only a partial mapping of the phonemes was performed since some classes were not merged.

The question therefore still remains whether a distance measure can be efficiently used to perform a complete mapping of the phoneme set of a source language database to a target language database. An approach for automatic phoneme mapping is attempted in this thesis. Single state (single mixture) Gaussian distributions are estimated for each phoneme in both databases on CMS normalised data. For each target phoneme the list of closest source phonemes are then found using the Bhattacharyya distance (Equation 5.1).

A problem with an automatic approach to phoneme mapping is that it depends on the distance measure used, but more importantly, that the usefulness of the results depends on a close match between the acoustic properties of the speech in the relevant databases.

For example, say a significant (but mostly linear) bias in feature space exists between the databases. The knowledge-based phonetic approach to computing the mapping is independent of the recording properties of the database and thus should deliver a reasonable mapping, enabling the use of linear transformation to remove the bias. With the automatic approach, though, the mapping may be so poor that even iterative application of transformation and re-mapping may not converge to the optimal mapping. Iterative application of mapping and transformation is, however, not attempted in this thesis.

5.1.4 Database issues

The characteristics of databases used for speech recognition experiments influence to a large extent the expected results. Use of a multilingual database, or bilingual database, in our case, ensures that cross-language experiments using only the database can focus mainly on the acoustic differences between the languages. The characteristics of the database still determines to a large degree the type of experiments that can be performed and also the recognition performance expected when applying various techniques. Both databases used in this thesis contain read speech from many speakers. Read speech is easier to recognise than for example spontaneous speech. Both databases contain phonetically diverse speech, not limited to any particular topic or speaking style, thereby increasing the number of contexts each phoneme may occur in.

When attempting the transfer of acoustic information between databases, it may be important to compensate for differences between the recording conditions of the databases. For example, frequency range and even the frequency transfer functions imposed on recorded speech may differ. Linear frequency effects may be compensated for by cepstral mean subtraction (CMS), although it may be inaccurate to simply implement CMS over the complete databases if the phonetic contexts of the databases differ significantly. The approach that we follow to solve this problem is to train models on the source database and to then compute the maximum likelihood cepstral offset between the source models and the target data within the MLLR framework. The offset is applied to the source data to perform

CMS. Section 5.2.5 gives detail about a generalisation of this approach that performs a transformation of source data, as opposed to simply performing CMS. Experimental results are detailed in Section 7.7 of Chapter 7. Thus, while this thesis does not attempt to characterise the effect of using different databases in general, some experimentation is done to ascertain the relative influence of using the same source language (in this case English) from the same database and from a different database. In the next section we discuss various ways in which multilingual data sources can be used to create target language systems.

5.2 Strategies for using multilingual data sources

The typical position is that a large amount of data is available for one or more source languages and only a limited amount of data is available for the target language. The goal is to construct a recogniser that will achieve optimal performance on unseen data from the target language. We make the assumption that sensible use of all available data will lead to better recognition performance than using only the target language data. Baseline performance is thus set by training on target language data only and methods to improve on this performance are sought. The next sections discuss various ways of utilising the available data.

5.2.1 Data pooling

The simplest method of constructing a recogniser using all available data is to simply pool the data and train on the pooled data set. This technique is commonly used to construct explicitly multilingual systems. Previous studies [48, 19, 20, 30] have shown that for reasonably large amounts of data from each of the languages, a slight performance degradation is actually achieved by the multilingual system in comparison to the language specific recognisers because the accuracy of the models is decreased. If only a small amount of data for the specific language were available, then some improvement in performance is possible with the multilingual system, simply because robust models cannot be trained with too little data. It is difficult to predict what amount of language specific data is necessary before multilingual pooling will degrade performance. This depends of course also on how close the match between the languages is - the closer the match, the better the chance that simple pooling will lead to desirable results.

5.2.2 Model combination

A simple alternative to the multilingual pooling method is to first train models separately on both source and target language data and to then select the specific models that perform the best on a separate cross-validation set. The reasoning behind this is that when a limited amount of training data is available, there may be enough data to train models for the phonemes that occur most, but not enough to train models for phonemes with low *a priori* occurrence. For these models the source language models can be used. For certain phonemes that occur only in the target language, the target language models are used irrespective of the amount of training data. This method is very simple to implement because pre-trained source language models can be used and training on the limited amount of target language data is computationally inexpensive. The method, however, does not make optimal use of all the available data.

5.2.3 Model adaptation

Adaptation of source language models using limited amounts of target language data has been previously researched, mostly for bootstrapping, but also for directly constructing target language recognisers. The assumption is that too little target language data is available for direct training, but that this data may be enough to adapt source language parameters to sufficiently improve target language performance. The adaptation task is complex compared to a typical speaker adaptation task and it is therefore expected that at least a reasonable amount of data will be necessary for adaptation to achieve good performance. An argument in favour of the model adaptation approach is that complex source language models can be estimated by using the large amount of source language data. These models can then be adapted using only limited amounts of target language data. This can be achieved for example by transformation-based adaptation in which only the transformation parameters need to be estimated, which may comprise far fewer parameters than re-estimation of all model parameters. On the other hand, if a reasonable amount of target language data is available, a Bayesian or discriminative adaptation technique may utilise the available data more efficiently than a transformation-based technique.

5.2.4 Combined pooling and adaptation

Multilingual pooling and adaptation are considered separate approaches, but it may lead to more efficient use of the available data if training on pooled data were first done, followed by adaptation of the multilingual models using the target language data only. In this way robust models are trained to begin with and are "fine-tuned" using target language data. Adaptation on target language data may improve recognition compared to the multilingual models. This is because the accuracy of the models may be improved with the target language data without sacrificing the robustness of the multilingual models. The method may also outperform source language model adaptation with target language data because source language model adaptation only uses the target language data in the adaptation process, say in the estimation of a transformation. This process may not efficiently extract the available information, thereby leading to suboptimal performance. This may be especially apparent when a reasonable amount of data is available in the target language. Also, adaptation of source language models may "untrain" the acoustic characteristics of the source language in order to specialise the models for the target language - thereby degrading robustness of the models. On the other hand, combined pooling and adaptation is also prone to "untraining" of the source language acoustics, except that little adaptation is likely to be necessary, thereby decreasing that risk. Combined pooling and adaptation is likely to be most useful in conjunction with a technique that adapts the model parameters only where necessary to effect a target language specific "fine-tuning".

5.2.5 Data augmentation

Augmentation usually comprises the transformation of data from one or more speakers to the space of a new speaker to augment the data from that speaker. It is therefore a form of data pooling, but only of transformed data. When using multiple databases for cross-language adaptation this may be of specific interest because differences, other than language, may also be removed as part of the process. When large differences exist between databases, the data augmentation approach may deliver an improvement in performance over the simple multilingual pooling approach.

Computation of the transformation

With the augmentation approach, a transformation is applied to the source data to alter it to better reflect the characteristics of the target data. The transformation can be computed in a number of ways, namely

- from source data to target data,
- from source models to target data,
- from source models to target models and
- the inverse of the transformation from target models to source data.

A data to data estimation approach for the transformation has the disadvantage that it is inevitably very simplistic since the elements of the transformation are not accurately identifiable. A model to data approach is more powerful since the model can be used to compute occupancy statistics, thereby artificially making the source and target elements of the transformation identifiable, aiding in the estimation of multiple transformations. Model to data transformations can also be optimised easily for maximum likelihood or least square error criteria. A model to model transformation is again difficult to estimate since source and target model parameters are not identifiable, especially for mixture distribution models.

If a model to data transformation is thus preferred, the choice has to be made between using source models or target models. Since a larger amount of data is typically available for source models, its parameters can therefore be estimated more accurately and thus use of source models is preferable. Target data may not be fully representative and may make estimation of parameters such as variance inaccurate. A final reason why source model to target data is preferred has to do with the availability of a unidirectional phoneme mapping that is discussed in more detail in Section 5.3.2.

Application of the data transformation

The transformation that was computed from source models to target data is used to transform source data to more closely match the target data. Since the transform is applied to labelled speech tokens, the data can be grouped in a meaningful way and multiple regression classes can be identified for transformation. The transformations do not directly transform source data variance, except by its relationship to the scaling of the mean components. This may not present a problem since further adaptation is likely to be performed. The augmentative transformation of the data is mainly to remove the possibly large bias between source and target features and a single transformation may even be used for this purpose.

The transformed speech data is pooled with the target language speech data and used to train target language specific speech models. The ratio of transformed versus target language data may influence the results as too much transformed data may dominate the trained model parameters, degrading performance. A way to improve this situation is to again fine-tune models using target language data only and is discussed next.

5.2.6 Combined augmentation and adaptation

Models trained on the augmented data set consisting of the original target data along with transformed source data are still candidates for target language specific adaptation. This is due to the fact that the data transformation may be relatively simple, consisting of a single regression class and therefore perform mainly channel equalisation and frequency shifting between the databases and languages. Even a transformation with multiple regression classes will not compensate for the differences in variance between the phoneme data of the source and target languages. Another reason for performing further adaptation is that the amount of transformed data may be so much more than the original target data that the weight of the target data is not properly reflected by the pooling process.

Adaptation may therefore further improve performance by more efficiently utilising target language data in “fine-tuning” the models. At this stage, any adaptation method may be used of course, but it is expected that an approach that can adapt individual parameters efficiently may prove to be better than say transformation of a large number of tied parameters.

In this section we discussed various strategies for using the source and target data. A choice now exists between a number of algorithms to be used for the implementation of the strategies. Various implementation aspects of algorithms also bear discussion in order to ensure acceptable performance. These aspects are discussed in the next section.

5.3 Cross-language model adaptation issues

In this section we discuss how to apply various methods from the fields of speaker adaptation and discriminative training to cross-language acoustic adaptation. Cross-language adaptation of acoustic models is a more difficult and complex task than speaker adaptation for a number of reasons, including

- cross-language adaptation performs a SI to SI mapping, rather than an SI to SD mapping,
- acoustic variations across languages are expected to be far larger and more complex than same-language speaker variations, and
- for speaker adaptation the source models generally cover the acoustics of the target speaker well, but are just not very accurate - while for cross-language adaptation the source models may model the target acoustics poorly to start with.

According to the above criteria, dialect adaptation may be closely related to cross-language adaptation, except that the acoustic variations are not expected to be as large. A recent study on dialect adaptation found that more complex adaptation procedures delivered better performance than simpler procedures that perform well for speaker adaptation [114]. It is likely that the complex task of cross-language adaptation will benefit from even more complex adaptation procedures. In the previous two chapters we focussed on methods to utilise the available data as efficiently as possible and we examined adaptation of all HMM parameters including duration modelling parameters. In the following sections the specific application of Bayesian estimation, transformation-based adaptation and discriminative techniques for cross-language adaptation is discussed.

5.3.1 Bayesian adaptation

Bayesian methods exhibit the desirable property of asymptotic performance. This is especially applicable for cross-language adaptation since a reasonably large amount of data may be available in the target language - more than is typically available for speaker adaptation. The relatively slow adaptation performance of Bayesian techniques may also therefore not present a great problem, although it may necessitate a larger target language database than may be needed with alternative approaches. A problem that is related to the slow adaptation of Bayesian techniques, is the fact that with Bayesian techniques, only observed mixtures are adapted. This implies that if source and target language distributions overlap partially, only the distributions in the overlap region are adapted and the other source model mixtures remain unadapted. The effect is reduced by performing a large number of adaptation iterations, but it may still not completely solve the problem.

A partial solution to the problem is to first perform linear transformation-based adaptation to increase the overlap of the distributions, thereby reducing the number of unadapted mixtures. This may, however, have the unwanted side-effect of changing the priors (seeded by the transformed source models) so much that they do not represent useful prior parameter distributions anymore. In this sense, even though transformation may improve overlap between prior model and target data distributions, it may cause the Bayesian adaptation process to be meaningless. An alternative to first performing model transformation is to use augmentative transformation of source data to improve correlation with target data. This should increase the overlap between augmented data and target data distributions, but may, similar to source model transformation, also degrade the usefulness of the models used to seed prior distributions in successive application of Bayesian adaptation techniques.

An alternative approach for estimating the prior distributions is to use data from as many languages as possible. In this way the inter-language variability is represented in the source models and can probably best express the expected uncertainty with respect to the parameters of a new target language. Such an approach, however, requires the availability of data from a number of languages and is not attempted in this thesis.

5.3.2 Transformation-based adaptation

Transformation-based adaptation is very efficient for correlated source and target distributions and needs relatively little data for robust estimation. For speaker adaptation, a motivation for using transformation-based adaptation is its efficiency in coping with spectral differences between speakers, such as vocal tract length differences. It is not known to what extent this applies to the cross-language adaptation scenario, but the differences may be larger and more complex for the cross-language case. Transformation-based adaptation at least has the advantage over Bayesian techniques that a large level of mismatch between source and target distributions is not a problem by itself, the problem rather lies in whether a useful transformation exists and whether it can be estimated accurately.

In order to perform the complex adaptation expected to be necessary, relatively large amounts of data will typically be available in the target language. The question arises whether transformation-based techniques can efficiently use relatively large amounts of data since they do not guarantee asymptotic performance with respect to a target dependent system. An application in which transformation-based systems are expected to deliver an advantage over other approaches is when cross-database adaptation is performed as part of cross-language adaptation.

Effect of the mapping and transformation class grouping on the transformation

The fact that a phoneme mapping is used in the cross-language transformation can influence the transformation to a large degree. As we have discussed in Section 5.1, the mapping attempts to find the best source phoneme match for every target phoneme, but often there is no real counterpart and an approximate mapping may result. Source language phonemes may also be mapped to multiple target phonemes, of which some may present close matches, but others not. The transformation is computed from the statistics of a whole group of phonemes and the individual statistics from each phoneme mapping therefore influences the shared transformation. However, the shared transformation, as such, transforms each source model to only a single target model, and can therefore not discriminate at all between target classes seeded from the same source model. This is a serious disadvantage of the transformation-based approach, that, for example, does not present itself with Bayesian techniques where the parameters of each model are adapted independently. This problem may have to be addressed by post-transformation adaptation using Bayesian or discriminative adaptation methods.

A related problem caused by shared transformations is that inaccuracies in the mapping -

which are unavoidable - translate to bias in the adapted parameters. What this means is that "outlier" source models (mainly due to mapping inaccuracies) influence the shared transformations in an unpredictable and undesirable way. The method used to group phonemes into classes for separate transformations is therefore of importance since it determines which transformations should reasonably affect each other. The two grouping strategies that were discussed in Section 3.3.1 are phonetically motivated grouping and model clustering criteria. Grouping by phonetic category assumes that source-target correlation of distributions can be specified by phonetic category while clustering criteria assumes that it depends on position in feature space. Phonetically derived regression classes have been found to deliver better performance [80] than clustering procedures, perhaps because (speech production) information at a higher level than acoustics is used. The piece-wise linear feature space transformation implied by model clustering of regression classes is perhaps a too simplistic assumption for cross-language adaptation.

Adaptation of variance parameters

When the potentially large differences between the acoustic properties of languages are considered, the need for adaptation of variance parameters is obvious. The relationships between source model and target data variance may also be quite complex, necessitating the use of full transformation matrices. In this respect, the log variance transformation of Section 3.3.2 is applicable since constraints on variance parameters are maintained automatically and parameter accuracy is treated sensibly.

Implementation of variance adaptation entails first performing mean adaptation with MLLR, followed by a limited number of variance adaptation iterations. For speaker adaptation purposes often only a single iteration of MLLR is performed since the initial alignment is usually satisfactory. For cross-language adaptation, however, a reasonably large number of iterations may be necessary to achieve satisfactory alignment between the current model estimate and the target data. It is for this reason that estimation of the variance transform is preceded by mean transformation, otherwise very inaccurate initial variance estimates may cause poor convergence.

Full, diagonal or block-diagonal transformation

For speaker adaptation it has been found that use of a full transformation matrix delivers better performance than use of either diagonal or block-diagonal matrices [27, 65]. For our more complex application of cross-language adaptation we therefore also expect full transformation matrices to deliver better performance. When computing the transformation(s) to be used for data augmentation, the less complex approach of block diagonal transformation may be useful.

We have mentioned in this section that transformation-based adaptation may for various reasons not produce ideal target models. The transformation, though, may be very useful as a first adaptation stage to deal with large overall differences between source and target language distributions. These transformed models can then be used to compute prior distributions for Bayesian adaptation. Bayesian adaptation may function more efficiently on transformed models than on source language models (since the fraction observable mixtures should improve) and may deliver good performance in the complex "fine-tuning" of distributions. In place of Bayesian adaptation, discriminative training may also be used to adapt transformed models or pooled data models. This is the topic of the next section.

5.3.3 Discriminative adaptation using MCE

The main advantage of a discriminative training technique such as MCE over distribution estimation strategies is usually explained in terms of the improved goal of the technique - namely to directly improve expected classification performance. Another advantage that is only really apparent in the implementation of discriminative training is that it may use the available information more efficiently in certain respects. Both Bayesian and discriminative adaptation approaches suffer from the problem of updating only observed mixtures since every parameter is updated independently. However, with discriminative adaptation, both correct and false class tokens are used in computing the update for the parameters of a mixture, thereby greatly reducing the fraction of unobserved mixtures. This is especially applicable when cross-language adaptation is attempted, since the overlap between source and target distributions for some phoneme pairs may be poor.

Initial models for MCE training

The selection of the initial models to use is very important when MCE is applied because the approach is susceptible to local optima. Use of initial models that already achieve good performance, or that are expected to be robust under different testing circumstances is desirable. In Section 5.2, various strategies were discussed for using multilingual data and models. Some of these strategies produce models that are not necessarily optimal and can benefit from further MCE adaptation. Models that can serve as possible *initial models* for subsequent MCE adaptation can be produced by

- ML training on pooled multilingual data,
- source language models adapted using Bayesian and transformation-based techniques on target language data.
- multilingual models adapted using Bayesian techniques on target language data.

Explicitly multilingual models may be less accurate than target language specific models since model accuracy is decreased when data from multiple languages are used for training. However, these models may be more robust because more data is available for estimation and also because a larger set of contexts are represented. These models may therefore be suitable as initial models for MCE adaptation and performing MCE adaptation on these models may improve the accuracy of the models, while retaining some of the robustness achieved by the initial training that took place on a large, diverse training set.

Both source language models and multilingual models that have been adapted on target language data may produce models suitable for further MCE adaptation. Bayesian adaptation may deliver robust model estimates that already deliver good performance. Transformation-based adaptation may also deliver models that are good starting points for further adaptation if large, overall differences between source and target language distribution can be efficiently removed without severely impacting on the robust characteristics of these models. However, it should be taken into account that initial models that have already been adapted for improved target language performance may be specialised to the extent that they are not as robust as (unadapted) multilingual models.

Finally, the use of transformed source data to augment target data for model training may also produce good initial models for further MCE adaptation. The models should be more accurate than explicitly multilingual models since the transformation should at least partially compensate for differences between the languages.

MCE parameter optimisation

The use of the MCE method for the optimisation of the parameters of HMMs is relatively complicated since gradient-based optimisation has to be used. A further complicating factor is the existence of a number of parameters, namely η in the misclassification measure (Equation 4.6), γ and θ in the loss function (Equation 4.8) and ϵ in the parameter update (Equations 4.14-4.17). Since these parameters influence the results obtained with MCE, their importance is analysed at least in a qualitative manner.

The value of η determines the degree to which false classes contribute to the misclassification measure according to their likelihoods. We have elected to use $\eta = 4$ since this seems a reasonable trade-off between choosing the maximum incorrect class and averaging over the incorrect classes and was found empirically to deliver reasonable results. The value of γ scales the slope of the sigmoid and is important because it influences the size of the feature space region in which observations materially affect the update. Examination of

Equation 4.19 shows that for loss close to either 0 or 1, the derivative of the loss with respect to the misclassification measure becomes small. Smaller values of γ increase the region over which the derivate of the loss remains large, thereby taking into account more observations. High values of γ lead to the consideration of observations only in the immediate region of the decision boundary. In order to effectively use limited amounts of data and also to be able to significantly shift the current decision boundaries, for example when a mismatched seed model was used, it is expected that a reasonably small value of γ should be used. In implementation, we normalised the class conditional log-likelihood functions (the $g_i(\mathrm{X}; |\Lambda)$ of Equation 4.12) by dividing it by the number of frames in X. This amounts to expressing each class log-likelihood on a per-frame basis, which greatly reduces the range of likelihood values that are observed. A value of $\gamma = 1$ was found empirically to deliver good performance and was used in experiments.

The value of the update parameter ϵ should also be selected. An update value of $\epsilon = 0.1$ was found empirically to deliver good performance and was used in experiments. On-line training can also be used, but we have selected to use gradient descent with batch-mode updates for simplicity. An approach whereby ϵ is a linearly decreasing function of the iteration count is commonly used with MCE adaptation and we selected to also decrease the update value as a function of the iteration count through $\epsilon_n = \epsilon_0(N-n)/N$ for iterations $n = 0, .., N-1$, with N typically set to 10. If a cross-validation set is available, it may be used as a stopping criterion for adaptation.

MCE cost function application

It can be reasoned that if a suitably large amount of data is available, that string-level MCE will optimally achieve the goal of minimum word and string error rate recognition. For a small vocabulary task such as CDR, the amount of data needed will probably be relatively small. For vocabulary independent adaptation, however, the amount of data needed to properly represent a reasonable percentage of phonetic contexts may be quite large. The use of phoneme-level MCE that implements word error-based phoneme misclassification cost therefore presents an alternative. Estimation of the misclassification cost can be performed using pronunciation dictionaries as was detailed in Section 4.5.3 and can therefore be performed irrespective of the availability of speech databases. In fact, the cost-based MCE approaches can be applied directly on source language data, without using any target language data and may improve to some extent the class discrimination properties with respect to target language needs. However, we believe that the availability of at least some target language data is essential for the development of accurate speech recognition systems.

A side-effect of the cost-based methods, especially the reward-based method, is that the amount of adaption is decreased rather than increased. By reducing the loss associated with certain categories and even associating a reward with some categories, the overall loss and therefore indirectly the overall gradient in each iteration is reduced. This is desirable since over-specialisation can easily happen with MCE if too little target language data is available.

5.4 Discussion

In this chapter we discussed the issues involved with using data from multiple languages and databases in improving the recognition performance for a single target language. Strategies for cross-language use of data and models were proposed, as well as the implementation of these strategies via adaptation techniques. The suitability of implementing different strategies via specific adaptation techniques were discussed. Overall, the combination of the proposed strategies and their implementation in terms of adaptation techniques presents a framework for cross-language use of acoustic information.

Approaches from this framework are applied in the following two chapters on a multilingual database and on two different databases to empirically evaluate their performance for cross-language acoustic adaptation.