Chapter 8

Conclusion

In this thesis several general strategies for cross-language use of acoustic information were proposed. The strategies incorporate specific techniques from the field of speaker adaptation and attempt to use relatively large amounts of source language data to improve the performance of recognisers for a new target language in which a limited amount of speech data is available. Extensions to current speaker adaptation techniques were also presented, with the particular aim of improving the performance of these techniques for cross-language adaptation. Experimental results vindicate the new approaches for cross-language use of acoustic information, also indicating improved performance for the proposed extensions to speaker adaptation techniques when applied to cross-language adaptation.

A fundamental question answered by the research performed in this thesis is whether cross-language use of acoustic information can be useful for the purpose of improving target language specific recognition. Previous research on this subject has been inconclusive, with studies of multilingual speech recognition [48, 19, 30] generally reporting that sharing of acoustic information degrades performance, and a few studies [20, 21] reporting marginal improvements by either multilingual pooling or mean-only MAP adaptation. The results in Chapters 6 and 7 of this thesis show significant reductions in word error rate (between 16% and 50%) for continuous speech recognition in a target language by use of speech data from a source language, showing that cross-language use of acoustic information can significantly improve target language specific recognition.

Once it is ascertained that cross-language use of acoustic information may be useful, the question arises as to the approach that should be followed to deliver good results. The strategies presented in this thesis, allied with specific adaptation techniques, create a framework for research into cross-language acoustic adaptation. The strategies that are presented include:

- cross-language adaptation of hidden Markov acoustic models,
- multilingual speech data pooling followed by acoustic model adaptation, and
- cross-language speech data augmentation, followed by acoustic model adaptation.

All three general strategies were shown to deliver useful results, with the newly-proposed *multilingual pooling-adaptation* strategy (Section 5.2.4) being especially suited when source and target data are closely matched i.t.o. recording conditions and labelling conventions, while the newly-proposed *cross-language augmentation-adaptation* strategy (Sections 5.2.5-5.2.6) provides improved performance when bias exists between source and target databases, such as exist between the TIMIT and SUN Speech databases.

These general strategies, incorporating adaptation techniques, form different approaches for cross-language acoustic use of information. The most prominent methods from the major categories of HMM adaptation techniques were used, including

- Bayesian techniques, in particular maximum *a posteriori* (MAP) estimation (Section 3.2),
- transformation-based adaptation techniques, in particular maximum likelihood linear regression (MLLR) transformation (Section 3.3), and
- discriminative learning techniques, in particular minimum classification error (MCE) adaptation (Chapter 4).

Within the cross-language adaptation framework, a large number of experiments were performed, using two speech databases namely the SUN Speech and TIMIT databases, to evaluate the relative performance of different approaches. Interestingly, all three adaptation techniques were found to contribute to achieving the best performance achievable for particular cross-language adaptation tasks. For experimental purposes, English was considered a source language, with relatively large amounts of English speech data available in the TIMIT and SUN Speech databases. Afrikaans was considered the target language, with a relatively smaller amount of data available in the SUN Speech database. For the relatively closely matched English and Afrikaans data from the SUN Speech database, training on pooled models, followed by either MAP, MCE, or a combination of MAP and MCE adaptation delivered the best results. When the English SUN Speech data set was used in conjunction with a 25 times smaller Afrikaans data set, a pooling-MAP-MCE approach (Sections 6.5.3 and 6.8.2) achieved a 50% reduction in word error rate over using only the Afrikaans data. When the English speech data was used in conjunction with a 5 times smaller Afrikaans data set, a pooling-MCE approach (Section 6.8.1) achieved a 26% reduction in word error rate over using only the Afrikaans data. For experiments with the TIMIT and SUN Speech databases, cross-language MLLR-MAP transformation delivered the best results, with further MCE adaptation achieving additional improvement. When the TIMIT database was used in conjunction with a 45 times smaller Afrikaans data set, an MLLR-MAP approach (Section 7.5.1) achieved a 35% reduction in word error rate over using only the Afrikaans data. When the TIMIT database was used in conjunction with a 9 times smaller Afrikaans data set, an MLLR-MAP-MCE approach (Sections 7.5.1 and 7.6.2) achieved a 16% reduction in word error rate over using only the Afrikaans data.

Extensions to current speaker adaptation techniques were proposed, in an effort to improve the performance of these techniques for cross-language acoustic adaptation and are briefly reviewed next. MSE Bayesian estimation equations were derived (Sections 3.2.2-3.2.3), and both the MAP and MSE prior parameter estimation equations were specified in terms of seed model parameters (Section 3.2.5). The choice of prior parameter initialisation was found to materially influence recognition results, but the choice between MAP and MSE estimation was found to provide little difference in performance (Section 6.5.5).

A technique was proposed to perform full (i.e. considering parameter dependencies) MSE transformation of the diagonal Gaussian variance parameters, in addition to using MLLR for the transformation of Gaussian mean parameters (Section 3.3.2). The variance transformation is implemented in log-space to ensure that parameter constraints are maintained, as well as to ensure optimisation of the relative variance error. Use of the proposed variance transformation technique was found to outperform MLLR transformation for cross-language adaptation, achieving between 8% and 18% reduction in word error rate for same-database experiments (Section 6.6.1) and between 4% and 14% reduction in word error rate for cross-database experiments (Section 7.4) over using mean-only MLLR transformation. An extension of MAP-MLLR for variance adaptation was proposed, incorporating a MAP-like term into the variance transformation (Section 3.4.2). Experiments showed MAP-MLLR to improve sensitivity of the transformation with respect to the number of regression classes, but to not significantly improve peak performance (Sections 6.7.2 and 7.5.2).

The application of MCE adaptation for cross-language adaptation was improved, firstly by extending the MCE framework to include adaptation of all model parameters, including duration modelling parameters (Section 4.3), and secondly by defining a method for including the cost of phoneme errors into the misclassification measure (Section 4.5.4). Methods to estimate useful cost matrices were proposed (Section 4.5.3) and cost-based MCE adaptation was shown to outperform standard MCE on adaptation of multilingual prior models, delivering useful cross-language adaptation performance (Sections 6.8.1 and 7.6.1).

In conclusion, several strategies, including extensions of current speaker adaptation and discriminative learning techniques, were presented, providing a framework for cross-language use of acoustic information. Approaches from the framework were found to deliver good performance under a reasonable variety of conditions, exceeding the performance achieved with previously published approaches to cross-language adaptation and showing significant gain from cross-language use of acoustic information.

8.1 Future research

The research performed in this thesis evaluated a wide scope of adaptation techniques, but on a fairly limited set of languages and databases. Research should be performed to study in particular

- use of more data sets to study application of the concepts to other databases and languages than those used in this study,
- use of mappings from more than one source language to a target language as this may allow improved prior model estimation,
- cross-language adaptation of context dependent phoneme models for target language LVCSR system development,
- use of target data to estimate the optimal complexity for each phoneme model, followed by training of models with the specified complexity on either source data or on pooled multilingual data,
- automatic phoneme mapping procedures - iterative use of a distance metric-based source-target phoneme mapping in conjunction with estimation of a source-target acoustic transformation may provide improved automatic phoneme mapping.

Use of the cross-language techniques proposed here, for accent adaptation, or even for speaker adaptation may provide interesting results. Another, more general research topic that follows from discussions in this thesis is the application of the integral over HMM parameter space (Equations 3.1, 3.2) to perform Bayesian estimation. Research into optimisation of MCE training, as well as improving the generalisation achieved with MCE training, perhaps using bootstrapping or Bayesian techniques, also warrants further research. Research regarding the use of MCE for adaptation, rather than training, may also provide interesting results.