Chapter 1

Introduction

Speech is a natural and efficient way for humans to communicate. Automatic speech recognition for computers introduces a fundamental shift in the human-machine interface, leading to myriads of new applications and greatly improving the usability of many existing applications. Human to human communication will be significantly enhanced in the future through the co-development of speech recognition in multiple languages combined with automatic translation between languages.

For most languages of the world, however, no speech recognition systems exist. The standard methods used for constructing speech recognition systems have been shown to work well for a large number of languages. The methods, however, necessitate a large amount of training data to deliver acceptable performance. The collection of speech data and the subsequent labelling of that data is currently an expensive and labour-intensive process. For the majority of languages of the world the lack of sufficient databases is the barrier that limits the development of speech recognition technology.

An interesting field of research in speech recognition technology is the development of systems that can explicitly recognise speech in multiple languages. Multilingual systems generally necessitate the use of acoustic information from multiple languages in a single modelling environment to avoid the cost of keeping full models sets for each language, to facilitate improved systems integration and to enable recognition of words from multiple languages in the same utterance. Although the aim of multilingual systems as such is not the re-use of acoustic information across language boundaries, it does present an approach for the use of acoustic information from existing databases in the development of a speech recognition system for a new language for which a limited amount of data is available.

Another field of research that is of interest is the field of speaker adaptation. Speaker adaptation techniques change model parameters to improve recognition performance for a new target speaker based on a limited amount of data from the new speaker. We, however, propose using speaker adaptation techniques for the purpose of changing models that were trained on a source language or languages, to improve performance for a target language. We propose that in this way, a system for a new target language can be developed that uses acoustic information from existing source language databases, but the performance of which is optimised for the target language using whatever target language data is available. This thesis details how multilingual data should be used in conjunction with adaptation techniques to deliver optimal performance for a new target language in the absence of sufficient amounts of target language data for the development of a stand alone speech recognition system.

1.1 Speech recognition fundamentals

Current leading edge speech recognition systems are based firmly on statistical pattern recognition principles [1, 2]. As such, these systems are data driven, i.e. are the result of training models of suitable complexity on large amounts of data. To increase recognition performance, models of increasing complexity are used - which in turn need increased amounts of training data to train accurately. It is expected that this trend will continue for some time. Large projects have been launched to collect and label spoken data for many of the major language groupings of the world such as American English [3, 4, 5], Japanese

[6], French [7, 8], German [9], as well as for the collection of multilingual databases[10, 11]. The existence of a comprehensive collection of data is a prerequisite for the development of a successful speech recognition system using current algorithms and technology.

Large-vocabulary continuous-speech recognition (LVCSR) systems comprise of two main parts, firstly *acoustic modelling* of the basic sounds or phones of speech, and secondly *language modelling* which captures the statistics of sequences of words. Pre-processing or feature extraction forms an important part of acoustic modelling by transforming the raw speech signal into an acoustic vector sequence $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, ..\mathbf{x}_T$ that is more amenable to modelling. Bayes' rule expresses the probability $P(W|\mathbf{X})$ of a word sequence $W = w_1, w_2, ..w_n$, given an observed acoustic vector sequence $\mathbf{X}$ by

$$P(W|\mathbf{X}) = \frac{P(\mathbf{X}|W)P(W)}{P(\mathbf{X})}, \tag{1.1}$$

where $P(W)$ represents the *a priori* probability of observing the sequence of words W, independent of the observed signal, and $P(\mathbf{X}|W)$ represents the conditional probability of observing the vector sequence $\mathbf{X}$, given the word sequence W. The probability $P(W)$ is language specific and is determined by a language model, often in the form of a conceptually simple bigram or trigram that may contain millions of discrete probabilities. The estimation of the parameters of these language models are facilitated by the large amounts of text available electronically. The training of the acoustic model $P(\mathbf{X}|W)$, however, depends on the availability of speech databases that are phonetically labelled, or at least transcribed. To achieve good performance, the training data should also fit the expected use of the system as closely as possible and should include data from many speakers for speaker independent (SI) recognition.

A basic premise of acoustic modelling is that a speech signal consists of short periods exhibiting stationary behaviour. This leads to the simplification of subdividing a speech signal into frames of relatively short length (typically 10-25 ms) with respect to the periods over which speech is stationary. A further assumption is that words can be modelled as the concatenation of a sequence of basic sounds or phones. Hidden Markov models (HMMs)

[1] are used to model phones via a sequence of states with quasi-stationary behaviour in each state. If every phone is represented by an HMM, words and sentences can be modelled by a concatenation of HMMs. The distribution of acoustic parameters in each state of an HMM is typically modelled with parametric continuous-density output distributions such as multivariate Gaussian mixtures.

Context has a large influence on the way that phones are produced and thus also influences the acoustic properties of the phones. To obtain good phonetic discrimination it is desirable to train different HMMs for phones in different contexts if enough speech data is available. A solution is to use triphones, where there is a distinct model for each phone combined with a unique pair of left and right neighbours. In practice this leads to an extremely large number of model parameters, which is reduced by making use of state tieing. The idea is to tie together states that are acoustically indistinguishable or at least very similar. Data associated with each individual state are pooled, giving more robust estimates for the parameters of a tied state. Even if enough data is not available to train accurate context dependent models, context independent models should at least allow for relatively complex distributions (such as Gaussian mixture distributions) to be able to model different contexts of each phoneme model. In any event, the amount of parameters to be estimated is large and predicates the use of large databases for training.

For many languages, including 10 of the 11 official languages of South Africa (all but English), very little or no speech data is available for training acoustic models. Even for South African English, speech data from various local population groupings would be needed to develop a system with robust performance on the South African accents. As far as language modelling is concerned, the situation is somewhat better since moderate amounts of electronic text are available in at least some of the languages. The speech databases that are available for South African languages include a database for South African English and Afrikaans [12] and a Xhosa database [13]. It is foreseeable that some data may be collected for more of the local languages, but it is unlikely that the quantity of data collected will approach the amount of data routinely used in developing LVCSR systems for the major languages of the world.

It is apparent that techniques must be found to enable the training of robust acoustic models in the absence of large quantities of speech data. One possible way would be to attempt to use expert knowledge from phoneticians in the target language. Approaches based largely on phonetic knowledge have been superseded by the statistical modelling approach and do not present a feasible solution. The only other option available then is to find methods that can use available data from other languages. It is hoped that these methods can improve the performance of acoustic models for a target language in which little training data is available. The field of *multilingual* speech recognition, which investigates the sharing of phoneme sets across languages, is a starting point for this research. It should be noted that the main focus of multilingual research is the creation of systems that can explicitly recognise speech in multiple languages, which may be in conflict with our goal of optimising performance for any specific language.

Another set of techniques that may be of use in developing robust acoustic models for a new language are techniques used for speaker adaptation. Generally, speaker adaptation techniques have been applied and optimised to improve speaker dependent modelling performance given a certain limited amount of speaker specific data. Although they are called *speaker* adaptation techniques, they also adapt models to recording and transmission channel conditions they are exposed to. These techniques do not have to be directed at a specific speaker and can be performed in multispeaker or even speaker independent mode and in our case are investigated for their use in cross-language adaptation of acoustic models.

1.2 Multilingual speech recognition systems

Multilingual speech recognition has generally been researched for the development of systems that can handle speech input in multiple languages [14, 15], or for the bootstrapping of seed models for forced alignment of speech data in a new language [16, 17, 18]. Some studies have researched the explicit sharing of acoustic information between languages by constructing multilingual phone sets [19, 20], but have in most cases reported some recognition performance degradation in return for simplified modelling of acoustic parameters and easily integrated multilingual recognition. Few studies have considered using cross-language acoustic information for the explicit goal of improving the performance of a speech recogniser in a new target language. One study [20] pooled cross-language and target language data to improve recognition for a target language application. Another two studies [21, 22] performed mean-only Bayesian adaptation of source language models using target language data and showed improvements in recognition rate under certain conditions.

The re-use of acoustic information across language boundaries for improving recognition in a new target language is only partially addressed by current research. Especially the application of adaptation algorithms for this purpose needs further investigation. We next discuss the main categories of speaker adaptation algorithms that are relevant for this thesis before we continue with the discussion on their use for cross-language adaptation.

1.3 Speaker adaptation techniques

The field of speaker adaptation is usually of interest when considering the adaptation of acoustic model parameters to new speakers or new conditions. Speaker adaptation techniques generally attempt to adapt acoustic parameters from the speaker independent (SI) scenario to improve performance on the data from a specific speaker. Research in speaker adaptation, to a large degree, focuses on achieving good adaptation performance using as little data as possible from a new speaker, enabling faster enrolment for dictation systems and also enabling the use of speaker adaptation techniques for a wider range of applications. Our interest in speaker adaptation algorithms lies with their application for the adaptation of acoustic models from a source language using speech from a limited number of speakers in a target language. In this way we aim to train target language models that retain some of their original acoustic properties, rendering them more robust and leading to improved recognition performance in the target language.

Bayesian methods were amongst the first methods used for speaker adaptation [23, 24]. Bayesian methods are especially applicable if a sufficient amount of adaptation data is available and suitable prior distributions can be estimated for system parameters. Bayesian methods assume a prior distribution $P_o(\boldsymbol{\lambda})$ for the model parameters, usually determined from training with a large set of SI data and use observations from a new speaker to determine the *a posteriori* distribution of the model parameters. Using Bayes' theorem we may write the posterior distribution $P(\boldsymbol{\lambda}|\mathbf{X})$ as

$$P(\boldsymbol{\lambda}|\mathbf{X}) = \frac{P(\mathbf{X}|\boldsymbol{\lambda})P_o(\boldsymbol{\lambda})}{P(\mathbf{X})} \tag{1.2}$$

The prior distribution, $P_o(\boldsymbol{\lambda})$, effectively biases the parameter distribution with the statistics for the speaker independent (SI) scenario. Bayesian estimation in known to work well for the SI to speaker dependent (SD) mapping since the SI case is a generalisation of the SD case. This is not true for a cross-language mapping, i.e. observations from a new language are not expected to be distributed according to a subset of the distribution of a source language, and may thus limit the performance achievable with Bayesian adaptation. An advantage of using Bayesian estimation, though, is that it has the property that the parameters converge to the target dependent parameters if enough adaptation data is available. Since we expect at least reasonably large amounts of data to be available for adaptation to the target language, the asymptotic performance property of a Bayesian estimator is desirable.

In speech recognition literature the method most commonly used for Bayesian adaptation is that of maximum *a posteriori* (MAP) parameter estimation. MAP estimation [25] chooses the mode of the posterior parameter distribution (the mode of $P(\boldsymbol{\lambda}|\mathbf{X})$) to represent the estimate of the parameter and is thus related to *maximum likelihood* (ML) estimation, which chooses the mode of the likelihood function (the mode of $P(\mathbf{X}|\boldsymbol{\lambda})$). Bayesian estimation can also be based on the use of a loss function to ensure that in some sense the minimum risk is associated with the estimate. The use of loss function-based Bayes estimators is investigated in this thesis in addition to MAP estimation.

A second class of methods for speaker adaptation is based on the transformation of the acoustic model parameters. Speaker adaptation via transformation does not attempt to directly estimate the new SD parameters, but rather estimates a transformation of the acoustic parameters from the SI models to the SD models. As such, it is suitable even when the SI models do not represent a prior for the SD models. A transformation may have few parameters - far fewer than the models being transformed, allowing the method to work reasonably well even when very little data is available. Transformation-based approaches were originally used to perform spectral transformation for template adaptation, accounting for microphone and channel effects and also changing the spectrum to better match the spectral characteristics of a new speaker [26, p. 286]. More recently, linear transformations of model parameters such as implemented by the *maximum likelihood linear regression* (MLLR) technique [27], rather than feature space transformations, have been commonly used. By grouping phones into classes for transformation, multiple transformations can be estimated, increasing the ability of the approach to perform complex adaptation tasks. Transformation-based adaptation generally performs well when significant bias exists between source and destination parameters, but its performance for cross-language adaptation of acoustic parameters, which may entail managing a complex set of uncorrelated differences between source and target acoustics, has yet to be fully investigated.

A third class of methods, based on discriminative training, has only recently been applied to the problem of speaker adaptation. A particularly promising implementation of discriminative training, called the minimum classification error (MCE) [28] approach, has been shown for some applications to improve performance beyond that obtained with the traditional MAP approach [29]. The MCE approach differs from Bayesian and ML approaches in that it attempts to directly minimise the number of misclassification errors, rather than maximising the *a posteriori* model likelihood or the data likelihood. Because MCE is really an error-function optimisation approach, it has considerable flexibility, leading us to consider its use for the complex task of cross-language adaptation. Unfortunately, MCE also suffers from problems such as being prone to converge to local minima. Using MCE for cross-language adaptation has the advantages over Bayesian and transformation-based methods that MCE does not make the assumption that the parameters of a suitable prior distribution can be found, nor does it assume that a linear transformation of parameter space is applicable. MCE is not suited for the removal of consistent bias (such as transformation-based methods are well suited for), but can effect very complex 'tuning' of parameters. Similar to Bayesian adaptation, only observed parameters are adapted, predicating the availability of reasonably large amounts of adaptation data for good performance.

In the next section we proceed to discuss how we applied the methods from the research fields covered so far, namely multilingual speech recognition and speaker adaptation, to our principal problem of cross-language data re-use.

1.4 Cross-language re-use of acoustic information

Previous research has shown the feasibility of using acoustic information from languages for which large databases exist in aiding the development of speech recognition systems for new languages. Source language models have been shown to be useful for bootstrapping models in a new language. Most studies indicate, however, that the sharing of acoustic models in a multilingual context leads to some performance degradation in return for simplified modelling [19, 20, 30], because model accuracy is reduced when the same model is used across multiple languages. Research [20] shows that sharing phones can work well if the languages have large acoustic similarities e.g. Italian and Spanish. For some new target languages it may be possible to find an acoustically similar language in which large amounts of speech data are available, but there may still be some sounds that occur only in the target language and have no near counterpart in the source language. Even for phones that occur in both source and target languages, there are bound to be some systematic differences in pronunciation, as well as differences with respect to the context of the phones. Simple sharing of acoustic information across language boundaries thus does not present an optimal solution to the problem in general.

An alternative to the pooling of data is to train models on large amounts of source language data and to then adapt the acoustic parameters from the source language to the target language in the same way that acoustic parameters are adapted from speaker independent (SI) models to the speaker dependent (SD) models. Some issues have to be addressed, however, since cross-language adaptation entails an SI to SI mapping and not an SI to SD mapping. Our aim with cross-language adaptation is to retain the SI properties of the acoustic models from the source language while changing them to better reflect the overall distributions of feature parameters in the target language. Typically, more data is available for cross-language adaptation than is usually used for speaker dependent adaptation, since a more complex mapping is expected to be necessary and also since the process can be performed off-line. This implies that techniques which can efficiently use larger amounts of data, rather than techniques specialised for rapid adaptation, are expected to deliver better performance.

A problem with the application of speaker adaptation techniques for cross-language adaptation is the assumption that the same set of phonemes can be used, which is not true in general for different databases in different languages. To address this problem it is necessary to make use of phonetic experts, or to use distance metrics to determine which phone classes should be used in conjunction with which other phone classes in the different databases and languages. For models in the target language that have poor correspondence in the source language, cross-language use of data does not guarantee acceptable performance and adaptation has to be able to significantly alter the model parameters to achieve good performance.

Two main classes of methods have been employed for cross-language adaptation in previous research namely Bayesian methods such as MAP and transformation-based techniques such as MLLR. We also apply these two methods, albeit more comprehensively than previous studies, to cross-language adaptation. Our implementation of cross-language Bayesian estimation uses the first language models to provide *a priori* information on the expected distribution of the second language model parameters and we adapt Gaussian mean and variance parameters as well as the mixture weight and transition probability parameters.

We show that adapting all model parameters in a Bayesian framework leads to superior performance when compared to the mean-only adaptation approach reported in previous research [22, 21]. A further improvement is obtained when prior distributions for MAP adaptation are estimated from models trained on pooled source and target language data, especially when source and target language data present a close match. This strategy of first training on pooled multilingual data and then performing further target language specific adaptation is well suited to the Bayesian adaptation paradigm, because use of multilingual data is more likely to produce suitable prior distributions than use of source data alone.

We find that use of the MLLR technique does not achieve the same level of performance as that achieved with the MAP technique. We propose a method to also transform the Gaussian variance parameters, greatly improving performance, but still not achieving as good performance as with MAP. We find, however, that use of MLLR adaptation is especially applicable when the source and target databases differ in terms of the recording conditions so that there are spectral differences between the source and target signals. In such cases, MLLR is used to produce transformed models, which in turn are used to seed prior distributions for MAP adaptation, achieving the best performance on the independent test set for cross-database adaptation.

We find that models trained on pooled multilingual data present good initial models for discriminative adaptation, especially if the pooled data sets were closely matched. Adaptation of the multilingual models is done with MCE, using target language data only, thereby improving the performance of the models for the target language. Discriminative training at this stage allows the models to retain the multilingual acoustic distributions as far as possible, changing them only with respect to errors incurred on the target language data. We propose an extension to the MCE framework that modifies the MCE misclassification measure to associate a cost with each phoneme misclassification error. The cost is based on the probability of a phoneme error leading to a word error and is shown to deliver improved performance for cross-language MCE adaptation. We also apply discriminative adaptation to models that have already been optimised for target language performance using other approaches and find that the MCE approach can improve on the performance achieved with the MAP and combined MLLR and MAP approaches, but that improved performance is not guaranteed.

Finally, we propose a data augmentation strategy for cross-language use of acoustic information. Data augmentation comprises computing a relatively simple transformation of source language data to better match target language data and then a pooling of the transformed data and the target language data. This pooled data set is termed the *augmented* data set and is used for model training. Trained models can be subjected to further target language dependent training to improve performance, especially since the data transformation may not accurately capture all the differences between the acoustics of the respective languages.

Overall, we find that cross-language use of acoustic information can lead to greatly improved target language performance. We present a framework of strategies and techniques for cross-language adaptation and perform experiments to evaluate the performance of a variety of the approaches.

1.5 Organisation of thesis

The outline of the thesis is now given. Chapter 2 gives background on the hidden Markov modelling approach followed. A relatively comprehensive coverage of basic material is given for reference purposes from later chapters as well as to at least partially document the algorithms used in the development of the Hidden Markov Toolkit for Speech Recognition (HMTSR) software by Darryl Purnell and the author during their Ph.D. studies. Also as part of the background, Chapter 2 contains a discussion of previous research in the field of multilingual speech recognition, which sets the stage for the research undertaken in this thesis. Chapter 3 treats techniques commonly used for speaker adaptation as their use for cross-language adaptation is extensively evaluated in a later chapter. Improvements to current techniques are also proposed. Discriminative learning methods, especially the minimum classification error (MCE) technique, are discussed in depth in Chapter 4, as well as a cost-based extension of the MCE framework. Chapters 3 and 4 form the basis for the presentation in Chapter 5, which describes strategies for cross-language use of acoustic information, as well as factors to be considered in applying both speaker adaptation methods and discriminative training methods to cross-language adaptation of acoustic models.

Cross-language English-Afrikaans experiments on the SUN Speech database [12] are presented and discussed in Chapter 6, showing large improvements in recognition performance through cross-language re-use of acoustic information. Chapter 7 extends the results from Chapter 6 to include cross-language use of acoustic information between the TIMIT [31] and SUN Speech databases. Finally, the conclusion is presented in Chapter 8.

1.6 Contributions of thesis

The original contributions presented in this thesis include the following points.

- We present a framework of strategies and techniques for cross-language use of acoustic information [32]. New strategies are proposed, such as first training models on pooled source and target language data, followed by adaptation, as well as a cross-language data augmentation approach which transforms source language data for a better match with target language data. We use the strategies to apply specific techniques from the field of speaker adaptation and discriminative learning and show that our newly proposed approach of pooling-adaptation leads to superior performance for same-database experiments than source model adaptation.

- Our complete implementation, evaluation and comparison of Bayesian and transformation-based adaptation techniques (initial results published in [33, 34]), as applied to the task of cross-language adaptation, provides insights as to the conditions under which the algorithms perform well. Previous studies only adapted Gaussian mean parameters and we show that adaptation of Gaussian variance and other HMM parameters lead to large performance improvements. As part of describing Bayesian estimation, we note that although MAP is almost exclusively and sometimes interchangeably used for Bayesian adaptation in the speech recognition community, that alternative implementations defined by loss functions exist. Along with the well documented MAP estimators, we also provide Bayes estimators for a mean square error loss function and experimentally compare the approaches.

- We propose a technique that, in conjunction with MLLR transformation of the Gaussian means, performs a full matrix transformation of the (diagonal) Gaussian variance values based on the least squares estimation. The transformation is computed in log-space, maintaining constraints on the variance values and minimising relative error in the transformation. Our experimental results show that the proposed approach outperforms standard MLLR, linear variance transformation and variance re-estimation in all experiments.

- We implement the recently proposed MAPLR approach, which combines Bayesian and transformation-based adaptation of Gaussian mean parameters. We use the same concept to extend our log-space variance transformation technique to incorporate a MAP-like term, improving generalisation and especially improving sensitivity of the transformation with respect to the number of regression classes by reducing overfitting.

- We derive and implement a comprehensive version of the MCE algorithm, adapting all HMM parameters, including duration modelling parameters in a unified framework utilising both "true" class derivatives and the "false" class derivatives. We extend the MCE framework to include a cost associated with each misclassification into the misclassification measure. We derive equations to base the estimation of the cost of phoneme misclassification on word error rate. We show that the cost-based extension to MCE achieves superior performance for multilingual model adaptation than the standard approach in our experiments.

- We evaluate cross-language performance for a continuous speech recognition task and show that cross-language use of acoustic information from the same or a different database can greatly improve the performance of a continuous speech recogniser beyond that achievable using only target language data. We present a feasible and useful approach for the development of a speech recognition system in a new language when only a limited amount of data is available in the new language.