Grapheme-based Automatic Speech Recognition using Probabilistic Lexical Modeling

THÈSE N^{O} 6280 (2014)

PRÉSENTÉE LE 1ER OCTOBRE 2014

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

LABORATOIRE DE L'IDIAP

PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Ramya RASIPURAM

acceptée sur proposition du jury:

Prof. J.-Ph. Thiran, président du jury
Prof. H. Bourlard, Dr M. Magimai Doss, directeurs de thèse
Dr K. Knill, rapporteuse
Prof. S. Renals, rapporteur
Dr J.-M. Vesin, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2014

"They alone live,
who live for others"
— Swamy Vivekananda

To my parents, parents-in-law, loving husband Murali
and adorable son Nidhiravya

Acknowledgements

The four years of my PhD were enlightening, challenging, and gratifying. My PhD would not have been possible without the help, expertise and guidance from my advisors, committee, friends and family.

Firstly, I must acknowledge Mathew for providing me the opportunity to pursue PhD at Idiap, for his guidance, friendship and constant support throughout my PhD. He has been very supportive and understanding during tough times. Hervé's comments and constructive criticism on the thesis has helped me to improve my work great deal. Hervé's hard work has built a world renowned speech research group at Idiap and it is my pleasure to be a part of it.

I thank my committee members Prof. Steve, Dr. Kate, Dr. Jean-Marc, and Prof. J.P Thiran for their valuable suggestions. This work has also benefitted from the collaborations with David, Peter Bell, Marzieh and Guillermo. The feedback received from talented past and present speech team of Idiap has been valuable for my PhD. I have also benefited from the advice of Prof. Tanja in various conferences and during her visit to Idiap. I must also thank my master research advisors Prof. Hema and Prof. Ramalingam for introducing me to speech recognition research area and motivating me to go for PhD. My work was supported by the Swiss NSF through the grant "Flexible Grapheme-Based Automatic Speech Recognition (FlexASR)".

Thanks to Ms. Nadine for her support especially in finding suitable accommodation in Martigny multiple times during PhD and to Ms Sylvie for her help in navigating the requirements of living in Swiss. Thanks to Idiap system staff for their great help and support.

In my four years at Idiap, many friends have made the journey more memorable. I am grateful to many friends. To Lakshmi and Afsaneh for helping me many times to boost my optimism. To Lakshmi and Saheer who helped in so many ways to ease my transition into Swiss. To Afsaneh and Mohammad for being caring people I can count on. To David with whom I could work with ease. To Marzieh for proof reading this thesis and for her friendship. To Barbara for support and advice in all matters relating to navigating PhD with a kid and surviving as a single parent during weekdays. To Pierre-Edouard for the French abstract on a short notice. To my thesis writing buddies Samira, Alex, Laurent and Marco. To our great yoga teacher and yoga friends Norbert, Marc, Samira, Raphaël and Ronan. To Vincent and Alex for their rides to and from Idiap when I was having trouble walking. To Leibny for serving as my conference companion and for all the interesting discussions. To Phil whose wisdom always enlightens me.

I thank my Indian friends Abhilasha, Ashtosh, Chidhansh, Dinesh, Deepu, Francina, Gokul,

Harsha, Jagan, Kavitha, Parameshwari, Pranay, Sriram, Sucheta, and Venky who made me feel home. Thanks to Idiap colleagues and other friends Afroze, Alexandre, Alexandros, André, Anindya, Anirudha, Blaise, Cijo, Daira, Dimitri, Elham, Fabio, Fen, Gelareh, Guillem, Gulcan, Gwénolé, György, Hari, Holger, Hui, Ivana, James, Joan, Kenneth, Laurent, Leo, Marc, Marco, Maryam, Milos, Nesli, Novi, Oya, Paco, Petr, Phil, Pierre-Edouard, Pinto, Rammohan, Raphaël, Riwal, Ronan, Rui, Samira, Sara, Serena, Sharid, Srikanth, Tatiana, Thomas, Trinh-Minh, Xingyu, Youssef, and Zoltan.

Thank you to my friends back home Kalyan, Suneetha, Bama, Sadhana, Neeraja, and Padmanabhan for their love and affection. I should not forget thanking mama PhD friends who shared their knowledge of managing PhD with kids. I want to also acknowledge my sons caregivers for their loving care and particularly I would like to thank Chantal, Lily and Dunia for their friendship and kind words.

I thank for warm kindness from the families of my sister, brother-in-law and sister-in-law: Sowmya and Naveen, Jagan and Padma, and Sunita and Manju. Thanks to my lovely and awesome nephews and nieces: Srinidhi, Shamitha, Rakshitha, Vishwak, and Hethav for delighting me with many happy moments. Special thanks to my pinnama and late pinnaya for their love and blessings. I should also thank my other extended family members who have helped me at various stages.

I express my sincere gratitude to my parents Nagamani and Radhaswamy to whom I owe my life for their constant love, moral support and blessings. I would also like to express my gratitude to my parents-in-law Sridevi and Venu Gopal for their unfailing emotional support. I fondly recollect how refreshing it was to have my parents and in-laws in Swiss for few months. This thesis could not have been completed without my husband Murali's innumerable sacrifices and unconditional love. He always delicately pushed me to achieve more and taught me so much about love and life. I feel blessed to have him with me. PhD is a long journey, since my husband had to move to Germany for his job, it would have been even longer and lonelier without my son Nidhiravya. I thank my little son for bringing happiness into my life. Murali and Nidhiravya have given my life purpose and fulfilment. Words cannot express my gratitude and love for you both. This thesis is dedicated to my parents, in-laws, Murali and Nidhiravya.

Martigny, 13 July 2014 Ramya

Abstract

Automatic speech recognition (ASR) systems incorporate expert knowledge of language or the linguistic expertise through the use of phone pronunciation lexicon (or dictionary) where each word is associated with a sequence of phones. The creation of phone pronunciation lexicon for a new language or domain is costly as it requires linguistic expertise, and includes time and money. In this thesis, we focus on effective building of ASR systems in the absence of linguistic expertise for a new domain or language. Particularly, we consider graphemes as alternate subword units for speech recognition. In a grapheme lexicon, pronunciation of a word is derived from its orthography. However, modeling graphemes for speech recognition is a challenging task for two reasons. Firstly, grapheme-to-phoneme (G2P) relationship can be ambiguous as languages continue to evolve after their spelling has been standardized. Secondly, as elucidated in this thesis, typically ASR systems directly model the relationship between graphemes and acoustic features; and the acoustic features depict the envelope of speech, which is related to phones.

In this thesis, a grapheme-based ASR approach is proposed where the modeling of the relationship between graphemes and acoustic features is factored through a latent variable into two models, namely, *acoustic model* and *lexical model*. In the acoustic model the relationship between latent variables and acoustic features is modeled, while in the lexical model a probabilistic relationship between latent variables and graphemes is modeled. We refer to the proposed approach as probabilistic lexical modeling based ASR. In the thesis we show that the latent variables can be phones or multilingual phones or clustered context-dependent subword units; and an acoustic model can be trained on domain-independent or language-independent resources. The lexical model is trained on transcribed speech data from the target domain or language. In doing so, the parameters of the lexical model capture a probabilistic relationship between graphemes and phones. In the proposed grapheme-based ASR approach, lexicon learning is implicitly integrated as a phase in ASR system training as opposed to the conventional approach where first phone pronunciation lexicon is developed and then a phone-based ASR system is trained.

The potential and the efficacy of the proposed approach is demonstrated through experiments and comparisons with other standard approaches on ASR for resource rich languages, non-native and accented speech, under-resourced languages, and minority languages. The studies revealed that the proposed framework is particularly suitable when the task is challenged by the lack of both linguistic expertise and transcribed data. Furthermore, our investigations also showed that standard ASR approaches in which the lexical model is deterministic are more

suitable for phones than graphemes, while probabilistic lexical model based ASR approach is suitable for both. Finally, we show that the captured grapheme-to-phoneme relationship can be exploited to perform acoustic data-driven G2P conversion.

Keywords: Automatic speech recognition; Kullback-Leibler divergence based hidden Markov model; lexicon; grapheme subword units; phoneme subword units; probabilistic lexical modeling; grapheme-based automatic speech recognition; grapheme-to-phoneme conversion; under-resourced speech recognition.

Résumé

Les systèmes de reconnaissance automatique de la parole (RAP) intègrent des connaissances poussées de la langue ou une expertise linguistique à travers l'utilisation d'un dictionnaire de prononciation dans lequel chaque mot est associé à une séquence de phonèmes. La création de dictionnaires de prononciation phonétique pour une nouvelle langue ou un nouveau domaine est coûteuse car elle requiert une expertise linguistique, du temps et de l'argent. Dans cette thèse, nous concentrons sur la construction efficace de systèmes de RAP sans aucune expertise linguistique pour un nouveau domaine ou une nouvelle langue. En particulier, nous considérons les graphèmes en tant qu'unités sous-mots alternatives pour la reconnaissance vocale. Dans un dictionnaire de graphèmes, la prononciation d'un mot est obtenue depuis son orthographe. Toutefois, modéliser les graphèmes pour la reconnaissance de parole est une tâche difficile pour deux raisons. Premièrement, la relation graphème-à-phonème (GAP) peut être ambigüe car une langue continue d'évoluer une fois son orthographe standardisée. Deuxièmement, comme nous le montrons dans cette thèse, en général les systèmes de RAP modélisent directement la relation entre les graphèmes et les caractéristiques acoustiques, et ces caractéristiques acoustiques décrivent l'enveloppe de la parole, qui est liée aux phonèmes.

Dans cette thèse, une approche de RAP basée sur les graphèmes est proposée, où la relation entre les graphèmes et les caractéristiques acoustiques est prise en compte à travers une variable latente dans deux modèles, à savoir un *modèle acoustique* et un *modèle lexical*. Dans le modèle acoustique, on modélise la relation entre les variables latentes et les caractéristiques acoustiques, alors que pour le modèle lexical on modélise la relation probabiliste entre les variables latentes et les graphèmes. Nous référons à l'approche proposée comme RAP basée sur la modélisation lexicale probabiliste. Dans cette thèse, nous montrons que les variables latentes peuvent être des phonèmes ou phonèmes multilingues, ou unités de sous-mots groupées en fonction du contexte. Un modèle acoustique peut être entrainé sur des ressources indépendantes du domaine ou indépendantes de la langue. Le modèle lexical est entrainé sur de la parole transcrite de la langue ou du domaine cible. En procédant ainsi, les paramètres du modèle lexical capturent une relation probabiliste entre les graphèmes et les phonèmes. Dans l'approche de RAP basée sur des graphèmes, l'apprentissage du dictionnaire est intégré implicitement comme une phase dans l'apprentissage du système de RAP, alors que dans le cas d'une approche conventionnelle, le lexique de prononciation phonétique est d'abord développé puis un système de RAP basé sur des phonèmes est entrainé.

Le potentiel et l'efficacité de l'approche proposée sont démontrés à travers une série d'expériences et de comparaisons avec d'autres approches standard sur des langues pour lesquelles

les ressources sont importantes, de la parole non native et accentuée et sur des langues minoritaires ayant peu de ressources. Nos études révèlent que le système proposé est particulièrement adapté lorsque la tâche est difficile en raison d'un manque d'expertise linguistique et de données transcrites. De plus, nos recherches montrent également que les approches standard de RAP pour lesquelles le modèle lexical est déterministe sont plus adaptées pour les phonèmes que pour les graphèmes, alors qu'une approche de RAP basée sur un modèle lexical probabiliste est adaptée aux deux. Enfin, nous montrons que la relation graphème-à-phonème peut être exploitée pour convertir des graphèmes en phonèmes en se basant sur des données acoustiques.

Mots clés : Reconnaissance automatique de la parole ; modèle de Markov caché basé sur la divergence de Kullback-Leibler ; lexique ; graphème ; phonème ; modélisation lexicale probabiliste ; reconnaissance automatique de la parole basée sur des graphèmes ; conversion graphème-à-phonème ; reconnaissance de parole avec peu de ressources

Contents

List of Figures

List of Tables

List of Acronyms

ANN artificial neural network

ARPA advanced research projects agency

ASR automatic speech recognition

DARPA defense advanced research projects agency

EM expectation-maximization

G2P grapheme-to-phoneme

GMM Gaussian mixture model

HMM/ANN hidden Markov model system using artificial neural networks as acoustic models

HMM/GMM hidden Markov model system using Gaussian mixture models as acoustic models

HMM hidden Markov model

HMS-HMM hidden model sequences hidden Markov model

HTK hidden Markov model toolkit

IPA international phonetic alphabet

KL-HMM Kullback–Leibler divergence based hidden Markov model

KLT Karhunen–Loeve transform

KL Kullback–Leibler

LHN linear hidden network

MAP maximum a posteriori

MF-PLP Mel-frequency perceptual linear prediction coefficients

MFCC Mel-frequency cepstrum coefficients

MLLR maximum likelihood linear regression

MLP multilayer perceptron

MMF master macro file, used in HTK toolkit

PA phone accuracy

PC-HMM probabilistic classification of HMM states

PCA principal component analysis

PDTS polyphone decision tree specialization

PLP perceptual linear prediction coefficients

RM resource management

SAMPA speech assessment methods phonetic alphabet

SCHMM semi-continuous hidden Markov models

SGMM subspace Gaussian mixture model

SP-HMM scalar product based hidden Markov model

Tied-HMM tied posterior based hidden Markov model

TTS text-to-speech

WA word accuracy

WER word error rate

WSJ wall street journal

1 Introduction

The goal of automatic speech recognition (ASR) systems is to convert a speech signal into text output. Standard ASR technology relies on the pronunciation lexicon, transcribed speech corpora, and large collections of text data to achieve state-of-the-art performance. In a pronunciation lexicon each word is represented in terms of subword units. Typically, phones or phonemes, the basic sound units of a language, are used as subword units. Most often, the lexicon is constructed by linguistic experts or linguists who carefully craft pronunciations for each word in the vocabulary. In a transcribed speech corpora, each speech recording is associated with a parallel word-level transcription. The recording and processing of a speech corpora is a costly and time consuming task. The collection of text data for a language is typically addressed using large amount of textual resources available on the web.

The predominant statistical approach used to achieve ASR is hidden Markov models (HMM) [Rabiner, 1989]. Given the transcribed speech data, pronunciation lexicon and text resources from the language for which we are interested to build an ASR system, the development of HMM-based ASR system is often decomposed into two problems. First, the relationship between subword units or "*lexical units*" and the acoustic feature observations has to be learned. Second, the syntactic constraints of the language have to be incorporated. The performance of an ASR system depends on how well the above two problems are addressed. In this thesis, we shall be concerned with the first problem.

1.1 Motivation and Objectives

ASR technology has enjoyed decades of progress, including the successful introduction of commercial systems. However, most of the commercial systems are for resource-rich languages (like English, Chinese, French) where there are adequate resources. The conventional ASR approach is being challenged by "under-resourced domains or languages" where resources required to build ASR systems are not available. The term under-resourced according to [Besacier et al., 2014] refers to a language that lacks at least one of the following: unique writing system, stable orthography, presence on the web, linguistic expertise, transcribed speech data,

pronunciation dictionaries, vocabulary lists, electronic text resources, etc. However, minority languages (that are spoken by minority of population) are not the same as under-resourced languages.

The future of speech recognition technologies lies in their ability to deal with many languages. Furthermore, the development of ASR technologies for under-resourced and minority languages can help in reducing the "language divide" among languages of the world.

To develop ASR systems for under-resourced languages either resources required to train an ASR system could be developed, or approaches that leverage from the resources available in resource-rich languages could be developed. As already mentioned, the creation of resources to build ASR systems for a new language or a new domain is typically a costly task. In this thesis, we will focus on the latter approach.

In the literature, the lack of transcribed speech data (or acoustic data) has been typically addressed through multilingual and crosslingual ASR approaches [Schultz and Waibel, 2001b, Burget et al., 2010, Thomas et al., 2012, Swietojanski et al., 2012]. These approaches are based on the fact that the sounds produced across languages share a common acoustic space. The usual mechanism followed is to define a lexical unit set based on universal phones using either knowledge-based or data-driven approaches. Once the universal phone set is defined, the relationship between lexical units and acoustic feature observations is learned on language-independent data. To overcome the mismatch between sounds among different languages, typically, the learned relationship is adapted on the target language data.

If the linguistic expertise and pronunciation lexical resources in the target language are not available, then the issue of subword units and pronunciation lexicon must be addressed. One simple way to address is through the use of graphemes, the units of written language, as lexical units [Schukat-Talamazzini et al., 1993, Kanthak and Ney, 2002, Killer et al., 2003, Ko and Mak, 2014]. In a grapheme lexicon, pronunciations of words are derived from their spelling. However, modeling graphemes for ASR is not a trivial task, since the discrepancy between graphemes and phones is considerable in many languages; and ASR systems directly model the relationship between graphemes and acoustic feature observations that depict the short term envelop of speech (which is more related to phones). Therefore, ASR systems using grapheme lexicon generally perform worse compared to systems using phone lexicon.

The lack of both acoustic and lexical resources has been rarely studied in the past. In [Stüker, 2008a,b], multilingual acoustic modeling with graphemes as subword units was considered when the language lacked both acoustic and lexical resources. For multilingual acoustic modeling, phone lexical resources are indispensable, since, the approaches depend on phonetic similarities of sounds between languages. Therefore, unlike phone subword units, it is not trivial to share models of grapheme subword units, since the relationship between graphemes and phones may differ considerably across languages.

The main goal of the thesis is to tackle challenges related to the building of ASR systems

for languages and domains that lack proper pronunciation lexical resources and acoustic data. To mitigate the dependency on both acoustic and lexical resources, we will focus on exploiting acoustic and lexical resources available in resource rich languages and domains for grapheme-based ASR.

1.2 Contributions of the Thesis

In this thesis, we will show that the modeling of the relationship between lexical units and acoustic feature observations can be factored into two parts or models through a latent variable, referred to as an "acoustic unit", namely,

1. *acoustic model* where the relationship between acoustic units and acoustic feature observations is modeled;
2. *lexical model* where the relationship between acoustic units and lexical units is modeled.

In the thesis, we elucidate that in standard HMM-based ASR systems, the lexical model is deterministic (deterministic lexical modeling). This has two main implications: the lexical units and acoustic units are the same, and are based on a type of subword units (phones or graphemes, context-independent or context-dependent); the acoustic model directly models the relationship between acoustic feature observations and lexical units. As a result, to build an ASR system, acoustic and lexical resources from the target language or domain are required to train or adapt both the acoustic model and the lexical model.

We show that, there are approaches such as, Kullback-Leibler divergence based hidden Markov model [Aradilla et al., 2008], and tied posterior [Rottland and Rigoll, 2000] where the relationship between lexical units and acoustic units is probabilistic (probabilistic lexical modeling) [Rasipuram and Magimai.-Doss, 2013b]. Probabilistic lexical modeling relaxes certain constraints imposed by deterministic lexical modeling and as a consequence, the acoustic model and the lexical model can be independently trained on different set of resources; acoustic units and lexical units can be based on different kinds of subword units and different types of contextual units can be modeled in an ASR system.

Motivated by these findings, in this thesis we propose an approach for grapheme-based ASR in the framework of probabilistic lexical modeling, where the relationship between graphemes and acoustic feature observations is factored into two models using acoustic units. In this thesis, we will show that in the proposed approach,

- lexical units can be graphemes of the target language while the acoustic units can be *phones*, or *multilingual phones*, or clustered context-dependent states;
- the acoustic model can be trained on target domain or language resources if available; or on domain-independent or language-independent acoustic and lexical resources; and
- the lexical model, which captures a probabilistic relationship between graphemes and acoustic units, is trained on the target language-dependent acoustic data.

In this thesis, we investigate the potential of the proposed grapheme-based ASR approach in overcoming acoustic and lexical resource constraints in ASR system development. Progres-

sively, we show that:

1. The proposed approach can overcome lexical resource constraints by integrating lexicon learning as a phase in ASR system training. More precisely, in the proposed approach, with phones as acoustic units and graphemes as lexical units, the lexical model parameters capture a probabilistic grapheme-to-phoneme (G2P) relationship learned through acoustic data. In this regard, we show that the proposed grapheme-based ASR approach can perform better than the phone-based ASR approach where phone pronunciation lexicon is developed using automatic G2P conversion approaches.
2. The proposed approach is particularly suitable when the task is challenged by the lack of both acoustic and lexical resources. The approach exploits existing acoustic and phoneme lexical resources available in other languages to improve grapheme-based ASR in target domain or language.
3. The proposed framework can be extended to languages and domains where both acoustic and lexical resources are not available (zero-resourced ASR).
4. The G2P relationship captured in the lexical model parameters can be exploited to perform acoustic data-driven G2P conversion.
5. The set of acoustic units that capture phone-like information can be derived by modeling context-dependent graphemes on target language data. Through the use of acoustic units derived from context-dependent graphemes, the performance of grapheme-based ASR systems can be significantly improved.

The work presented in this thesis has been published in [Magimai.-Doss et al., 2011, Imseng et al., 2011, Rasipuram and Magimai.-Doss, 2012a,b, Rasipuram et al., 2013a, Rasipuram and Magimai.-Doss, 2013a, Rasipuram et al., 2013b]. Few parts of the work are in the form of publicly available research reports [Rasipuram and Magimai.-Doss, 2013b, 2014].

1.3 Organization of the Thesis

The reminder of the thesis is organized as follows:

- Chapter 2, Background, gives an overview of standard HMM-based ASR systems followed by the description of each of the components of an ASR system, namely, feature extraction, pronunciation lexicon, acoustic likelihood estimator, language model and decoder.
- Chapter 3, Probabilistic lexical modeling, is devoted to the description of the probabilistic lexical modeling framework. The chapter also describes and compares ASR approaches where the lexical model is probabilistic.
- In Chapter 4, we first show that the deterministic lexical modeling aspect of standard HMM-based ASR systems imposes the need for large amount of transcribed speech data and phone pronunciation lexical resources from the target language or domain; and the probabilistic lexical model based ASR approaches relax certain constraints of deterministic lexical model based ASR approaches. We propose a grapheme-based ASR approach in the framework of probabilistic lexical modeling in which the relationship between graphemes (lexical units)

and phones (acoustic units) is learned through acoustic data. The viability of the approach is demonstrated through a pilot study.

- Chapters 5 and 6 are devoted to studying the proposed grapheme-based ASR approach in lexical resource constrained ASR scenarios, and both acoustic and lexical resource constrained ASR scenarios, respectively. More specifically, we show the potential of the proposed approach in addressing both acoustic and lexical resource constraints.
- In Chapter 7, Zero-resourced ASR, we extend the proposed framework and show that ASR systems for a new language could be developed without using any acoustic and lexical resources from the language, i.e., zero-resourced ASR system. Furthermore, in the case where untranscribed speech data from the target language is available, we show that the lexical model parameters can be adapted in an unsupervised manner to improve the performance of an ASR system.
- In Chapter 8, we show that the G2P relationship captured in the lexical model parameters can be exploited to perform acoustic data-driven G2P conversion. The proposed G2P approach is experimentally evaluated and compared against conventional G2P approaches at pronunciation error level and ASR performance level.
- In Chapter 9, Improving phones-less grapheme-based ASR, we show that the clustered context-dependent graphemes model phone-like information and the poor performance of grapheme-based ASR systems proposed in the literature is primarily due to deterministic lexical modeling. We show that by incorporating probabilistic lexical modeling, the grapheme-based ASR system (without using any phone information or cross-domain resources) is more robust against possible pronunciation errors inherent in the grapheme lexicon.
- Finally, Chapter 10, concludes with possible directions for the future research.

2 Background

In this thesis we are concerned with the statistical ASR. In the statistical ASR approach, the goal is to find the best matching (most likely) word sequence $W^* = \{\mathbf{w}_1, \ldots, \mathbf{w}_m, \ldots, \mathbf{w}_M\}$ given the acoustic observation sequence $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T\}$ where M is the total number of words in the utterance and T represents the total number of frames in the speech signal. Formally,

$$
\begin{aligned}
W^* &= \arg\max_{W \in \mathscr{W}} P(W|X) && (2.1)\\
&= \arg\max_{W \in \mathscr{W}} \frac{p(X|W)P(W)}{p(X)} && (2.2)\\
&= \arg\max_{W \in \mathscr{W}} p(X|W)P(W) && (2.3)
\end{aligned}
$$

where $\mathscr{W}$ denotes the set of all possible word sequences and W denotes a word sequence. The Bayes rule is applied in Eqn (2.2). In Eqn (2.3), the denominator $p(X)$ is dropped because it is independent of word hypothesis and does not affect the maximization. The first term on the right hand side of Eqn (2.3) is the likelihood of the acoustic observation sequence X given a word sequence W and is referred to as the acoustic likelihood. The second term on the right hand side of Eqn (2.3) is the prior probability of a word sequence W or the language model probability.

2.1 Standard HMM-based ASR

Modeling the relationship between all acoustic observation sequences and all possible word sequences is practically infeasible. In general, speech recognition systems model words as a sequence of subword units, which are further modeled as a sequence of HMM states. The sequence of subword units for a word is given by its pronunciation model as specified in the pronunciation lexicon. Typically, the language model is an n-gram statistical model where the probability of the current word depends only on the previous $n-1$ words. ASR systems normally incorporate bi-gram ($n = 2$) or tri-gram ($n = 3$) language models. The most likely

word sequence W^* is obtained by,

$$W^* = \arg\max_{W \in \mathscr{W}} p(X|W,\Theta_A)P(W|\Theta_L) \tag{2.4}$$

$$= \arg\max_{W \in \mathscr{W}} \Big[\sum_{Q \in \mathscr{Q}} p(X,Q|W,\Theta_A) \Big] P(W|\Theta_L) \tag{2.5}$$

$$= \arg\max_{W \in \mathscr{W}} \Big[\sum_{Q \in \mathscr{Q}} p(X|Q,W,\Theta_A)P(Q|W,\Theta_A) \Big] P(W|\Theta_L) \tag{2.6}$$

$$= \arg\max_{W \in \mathscr{W}} \Big[\sum_{Q \in \mathscr{Q}} p(X|Q,\Theta_A)P(Q|W,\Theta_A) \Big] P(W|\Theta_L) \tag{2.7}$$

$$\approx \arg\max_{W \in \mathscr{W}} \Big[\max_{Q \in \mathscr{Q}} p(X|Q,\Theta_A)P(Q|W,\Theta_A) \Big] P(W|\Theta_L) \tag{2.8}$$

$$\approx \arg\max_{W \in \mathscr{W}} \Big\{ \Big[\max_{Q \in \mathscr{Q}} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t = l^i,\Theta_A)P(q_t = l^i|q_{t-1} = l^j,\Theta_A) \Big] \Big[P(\mathbf{w}_1|\Theta_L) \prod_{m=2}^{M} P(\mathbf{w}_m|\mathbf{w}_{m-1},\Theta_L) \Big] \Big\} \tag{2.9}$$

- The parameter set $\Theta = \{\Theta_A, \Theta_L\}$ includes the parameters of the acoustic likelihood estimator (Θ_A) and the parameters of the language model (Θ_L).
- In Eqn (2.5), the acoustic likelihood is obtained by summing over all possible state sequences $\mathscr{Q}$ where each $Q = \{q_1,\ldots,q_t,\ldots,q_T\}$ denotes a sequence of HMM states corresponding to a word sequence hypothesis.
- In Eqn (2.6), the Bayes rule is applied. $p(X|Q,W,\Theta_A)$ is the likelihood of the acoustic observation sequence given an HMM state sequence and a word sequence, $P(Q|W,\Theta_A)$ is the probability of the HMM state sequence given a word sequence (often termed as the pronunciation model and derived from the pronunciation lexicon), and $P(W)$ is the language model probability.
- Eqn (2.7) assumes that the acoustic likelihood is independent of words given the state sequence.
- In Eqn (2.8), a Viterbi approximation is employed where the sum over all possible state sequences is replaced with the most probable state sequence.
- In subword unit based ASR systems, HMM states represent lexical units i.e., $q_t \in \mathscr{L} = \{l^1,\ldots,l^i,\ldots,l^I\}$ and I is the number lexical units. If phones are used as subword units then the lexical unit l^i can represent a context-independent phone or a context-dependent phone and if graphemes are used as subword units then the lexical unit l^i can represent a context-independent grapheme or a context-dependent grapheme. Here, for the sake of simplicity we assume that a lexical unit is represented by an HMM state.
- Eqn (2.9) arises from HMM and language model assumptions. The two HMM assumptions are: (1) The output observation at time t is dependent only on the current state (*i.i.d*). (2) First order Markov assumption which states that the current state is dependent only on the previous state. Usually, $p(\mathbf{x}_t|q_t = l^i,\Theta_A)$ is referred to as the *local emission score* and $P(q_t = l^i|q_{t-1} = l^j,\Theta_A)$ is referred to as the *transition score*.

In other words, a sentence model consists of a sequence of word models constrained by the language model, a word model consists of a sequence of subword models constrained by the pronunciation lexicon and a subword model consists of concatenation of one or more HMM states. As a result, Eqn (2.9) can be simplified such that the most likely word sequence W^* is obtained by finding the most likely state sequence Q^*, i.e.,

$$Q^* = \arg\max_{Q \in \mathscr{Q}} p(X, Q|\Theta) \tag{2.10}$$

$$= \arg\max_{Q \in \mathscr{Q}} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t = l^i, \Theta_A) \cdot P(q_t = l^i|q_{t-1} = l^j, \Theta) \tag{2.11}$$

$$= \arg\max_{Q \in \mathscr{Q}} \sum_{t=1}^{T} [\log p(\mathbf{x}_t|q_t = l^i, \Theta_A) + \log P(q_t = l^i|q_{t-1} = l^j, \Theta)] \tag{2.12}$$

Similar to Eqn (2.9), Eqn (2.11) results after *i.i.d* and first order Markov assumptions. Eqn (2.12) is as a result of log transformation to Eqn (2.11). If l^j is the last lexical unit of a word and l^i is the first lexical unit of next word then $P(q_t = l^i|q_{t-1} = l^j, \Theta)$ incorporates the language model probability, otherwise it is the HMM state transition probability.

The parameters of the HMM-based ASR system include the parameters of acoustic likelihood estimator and the parameters of the language model. The parameters of the acoustic likelihood estimator include:

- The set of lexical units $\mathscr{L} = \{l^1, \ldots, l^i, \ldots, l^I\}$.
- State transition probabilities $\{a_{ij}\}_{i,j=1}^{I}$ where $a_{ij} = P(q_t = l^i|q_{t-1} = l^j)$ for $i, j = 1, \ldots, I$.
- The parameters of the local emission score $p(\mathbf{x}_t|q_t = l^i)$ estimator for each lexical unit.

The various components of an ASR system are illustrated in Figure 2.1. The reminder of this chapter will briefly elaborate on each of the components.

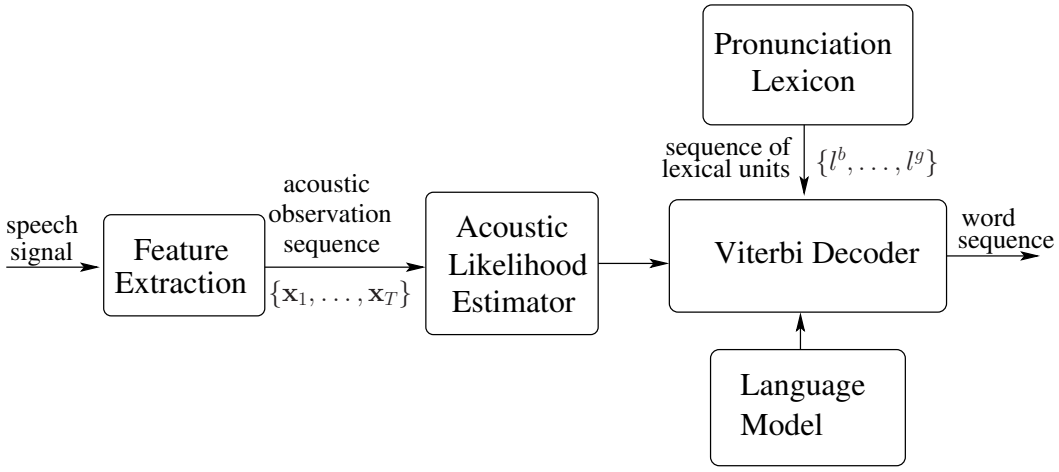

Figure 2.1 – Block diagram of an ASR system

2.2 Feature Extraction

The goals of feature extraction are to extract acoustic information from the speech signal that is relevant to the identification of the underlying sounds and suppress the non-linguistic information such as speaker and environmental variability. Feature extraction also provides a compact representation of the speech signal. The two common features that are used in most of the ASR systems are Mel-frequency cepstral coefficients (MFCC) [Davis and Mermelstein, 1980] and perceptual linear prediction (PLP) [Hermansky, 1990] features. These features are computed every 10 ms in overlapping analysis windows of 25 ms duration. This is based on the assumption that the speech signal is quasi-stationary in short-time intervals.

As described in [Gold and Morgan, 1999], for both MFCC and PLP, the first step is the computation of the power spectrum for each analysis window of the speech signal. This is done by windowing the analysis region (typically, using a Hamming window), calculating the FFT, and computing its squared magnitude. In the second step, the power spectrum is integrated within critical band filter responses. In the case of MFCCs this is done using the Mel scale which is roughly linear below 1kHz and logarithmic above 1kHz. In the case of PLPs, this is done using trapezoidal filters applied at roughly 1-Bark intervals. In the third step, the spectrum is pre-emphasized to account for the unequal sensitivity of human hearing at different frequencies. In the case of MFCCs, this is done before the spectral analysis, whereas in the case of PLPs this is implemented as explicit weighting of the critical band spectrum. In the fourth step, spectral amplitudes are compressed. In the case of MFCCs, log transformation is applied whereas for PLPs, the spectral amplitudes are compressed using cubic root compression. In the fifth step, decorrelation and dimensionality reduction is performed. In the case of MFCCs, typically DCT is applied and this step yields the cepstral coefficients. Dimensionality reduction is achieved by cepstral truncation where the first 12 or 13 components are retained. In the case of PLPs, an autoregressive model is used to smooth the compressed critical band spectrum. This step in PLP has been shown to achieve better noise robustness and speaker independence than cepstral truncation of MFCCs [Openshaw et al., 1993, Gold and Morgan, 1999].

In order to account for the dynamic behaviour of the speech signal, usually the MFCCs or PLPs are appended with first order and second order derivatives of static features computed across analysis frames [Furui, 1986]. In this thesis, we used MF-PLPs extracted using the hidden Markov model toolkit (HTK) as acoustic feature observations ($\mathbf{x}_t$) [Young et al., 2006]. The MF-PLP features are PLP features but use the Mel scale filter bank in place of the Bark scale filter bank.

2.3 Pronunciation Lexicon

In ASR systems, words are modeled in terms of subword units to address data sparsity issues and achieve generalization towards unseen words. The use of subword units in speech recognition presents two challenges. The first challenge is the choice of subword units and the second

is the transcription of each word in terms of a sequence of subword units. The collection of words and their pronunciations is usually referred to as the pronunciation lexicon or the pronunciation dictionary. The pronunciation lexicon acts as an interface between the words and the lexical model. All the components in an ASR system presume the availability of a subword unit set and a pronunciation lexicon. Therefore, in practice, ASR system development can be seen as a two stage process: development of pronunciation lexicon followed by ASR system training.

Ideally, it is good to have subword units that associate well with the acoustic signal (acoustic feature observations), are sufficiently frequent in the database used for ASR system training, are robust to the changes in context, allow easy generation of lexicon and provide flexibility for cross and multilingual portability. Unfortunately, there exists no single subword unit set which addresses all these concerns equally well.

2.3.1 Phone Subword Units

Typically, ASR systems use linguistically motivated *phones* or *phonemes* as subword units. Therefore, development of the subword unit set and the pronunciation lexicon is derived mostly from linguistic theory and incorporates linguistic expertise of a language. Linguists have categorized many of the sounds of the languages in the world into segments called phones [Gold and Morgan, 1999]. A phoneme is the smallest contrastive unit in the phonology of a language [O'Shaughnessy, 1987]. Therefore, a phone set is designed to cover the set of sounds in all languages where as a phoneme set is a set of sound categories of a particular language [Gold and Morgan, 1999]. Examples of phone sets include the international phonetic alphabet (IPA), the speech assessment methods phonetic alphabet (SAMPA), the ARPABET [1], and the CMUBET [2].

Phone pronunciations are typically obtained from a hand-built lexicon which the linguistic experts have prepared. During the preparation of the pronunciation lexicon by linguists, care is taken to minimize word level confusions and consistency is ensured across the lexicon. The hand crafted phone pronunciation lexicon will provide optimum performance for ASR. However, design of the phone pronunciation lexicon of significant size by linguistic experts is a tedious and costly task. Furthermore, a finite lexicon will always have limited coverage for ASR and text-to-speech (TTS) synthesis systems. For this reason, ASR and TTS systems use semi-automatic pronunciation generation methods when hand crafted pronunciations fail to cover vocabulary of a particular domain.

The phone pronunciation lexicon can also be built using phonological rules defined by linguists [Kaplan and Kay, 1994]. Phonological rules are often specific to a language, sometimes even to a dialect. Another commonly adopted way to generate or augment the phone pronunciation lexicon is through automatic G2P conversion systems [Pagel et al., 1998, Bisani and

1. http://en.wikipedia.org/wiki/Arpabet
2. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Ney, 2008]. The two main components of automatic G2P conversion systems are: a source lexicon, and a method to capture the G2P relationship observed in the source lexicon.

2.3.2 Grapheme Subword Units

Other alternatives for subword units are graphemes [Schukat-Talamazzini et al., 1993, Kanthak and Ney, 2002, Killer et al., 2003, Magimai.-Doss et al., 2011, Rasipuram et al., 2013a, Ko and Mak, 2014] which make the pronunciation lexicon development easy. Graphemes are the units of written language, e.g., alphabetic letters of English. With graphemes as subword units the pronunciation of words is derived from their orthography. According to Omniglot [3], as summarized in [Schultz and Kirchhoff, 2006, Chapter 4] [Stüker, 2009, Chapter 4], graphemes can be used as subword units for a wide number of languages that use the alphabetic, or Abugidas writing systems. However, modeling graphemes for ASR is not a trivial task because:

1. A sequence of graphemes may represent a single phoneme. In English, the grapheme sequence [S], [H] represents a single phoneme /sh/. In Scottish Gaelic, many vowels are present in orthography only to denote the nature of consonant next to it. These graphemes are never pronounced. Such graphemes in a word are usually referred to as silent letters.
2. A single grapheme at a particular instant may represent more than one phoneme. For example, in Arabic short vowels occurring next to a consonant are often not present in the written form.
3. The same letter at different instants can represent different phonemes. For example, in English, grapheme [C] maps to phoneme /k/ in *CAT*, to phoneme /s/ in word *CITE*, and to phoneme /ch/ in word *CHURCH*.
4. Different graphemes can represent the same phoneme. For example, in Polish, the graphemes [u] and [ó] represent the phoneme /u/.

The relationship between graphemes and phones is highly dependent on the language of interest. For languages such as Finnish and Spanish, the relationship is regular whereas for languages, such as English and French the relationship is irregular. As it is the case with phoneme subword units, deviations in the relationship between graphemes and acoustics may occur for different dialects of a language, non-native words, proper names, non-native speakers etc. Thus, it is challenging to model grapheme subword units for speech recognition. In Chapter 4 (see Section 4.1.2), we will present an overview of grapheme-based ASR approaches proposed in the literature.

2.3.3 Context Dependency

Depending on the subword context modeled, there are two types of ASR systems, namely, context-independent subword unit based ASR systems and context-dependent subword unit

3. http://www.omniglot.com/

based ASR systems. In context-dependent subword unit based ASR systems each context-independent subword unit in context is considered as a separate unit [Schwartz et al., 1985]. For example, the pronunciation of the word '*that*' would be represented as "/dh/ /ae/ /t/" in the case of context-independent subword units and as "/dh+ae/ /dh-ae+t/ /ae-t/" in the case of context-dependent subword units. Context-dependent subword modeling is primarily incorporated to model coarticulation i.e., each phone may be realized differently in different contexts. State-of-the-art ASR systems are typically based on context-dependent subword units.

2.4 Acoustic Likelihood Estimator

As given in Eqns (2.9) and (2.11), the estimation of acoustic likelihood $P(X|W,\Theta_A)$ involves the estimation of the local emission score $p(\mathbf{x}_t|q_t = l^i,\Theta_A)$ and the state transition score $P(q_t = l^i|q_{t-1} = l^j,\Theta_A)$. In this section we will briefly describe how these probabilities are estimated and how the parameters of the acoustic likelihood estimator Θ_A are learned.

Standard HMM-based ASR systems implicitly model the dependency between acoustic feature observation $\mathbf{x}_t$ and lexical unit l^i through an intermediate set of units (see Sections 3.1 and 3.2). In this thesis, these intermediate units are referred to as acoustic units. The set of acoustic units is denoted as $\mathscr{A} = \{a^1,\ldots a^d,\ldots a^D\}$ where D is the total number of acoustic units.

The two main approaches used in the literature to model the acoustic units are Gaussian mixture models (GMMs) and artificial neural networks (ANNs). The resulting ASR systems are usually referred to as HMM/GMM [Rabiner, 1989] and Hybrid HMM/ANN [Morgan and Bourlard, 1995] systems, respectively.

2.4.1 Acoustic Units

Let K be the number of context-independent subword units in the lexicon and M is the minimum duration constraint (typically, $M = 3$). There are two types of context-independent subword unit based ASR systems one can encounter:

1. The lexical units are context-independent subword units with minimum duration constraint, i.e., $I = K \times M$ and there is an acoustic unit for each lexical unit, i.e., $D = I$. This is a system where the relationship between acoustic feature vectors ($\mathbf{x}_t$) and lexical units (l^i) is directly modeled.
2. The lexical units are context-independent subword units with minimum duration constraint, i.e., $I = K \times M$, however, the acoustic units are context-independent subword units, i.e., $D = K$. In this case, the M lexical units associated with a context-independent subword unit share an acoustic model. The deterministic relationship between lexical and acoustic units is modeled by building a look-up table with I rows.

Context-independent subword unit based HMM/GMM systems are of the first kind. In the past, context-independent subword unit based hybrid HMM/ANN systems were typically of

the second kind [Morgan and Bourlard, 1995].

In the case of context-dependent subword unit based ASR systems, the number of lexical units $I = M \cdot K^{c_r+c_l+1}$ where c_l is the preceding context length, c_r is the following context length. Generally, not all context-dependent subword units will appear sufficiently often in the training data. Hence a sharing approach is used to enable multiple lexical units to share an acoustic model. This is done using decision-tree based state clustering and tying technique that uses the pronunciation lexicon, linguistic knowledge (phonetic question set) and acoustic data [Young et al., 1994].

The number of acoustic units D vary depending on hyper parameters such as the state occupancy count and the log-likelihood threshold that are used during decision-tree based state clustering. However, the number of acoustic units D is well below the number of lexical units I. The state tying process builds a look-up table with I rows that maps each lexical unit l^i to one of the D acoustic units. In toolkits such as HTK, this table is not explicitly seen. However, it is obtained from decision trees and is stored in the HMM definition file or the master macro file (MMF) and tied list after state clustering and tying [Young et al., 2006]. The resulting clustered units or the acoustic units are modeled with GMMs or with an ANN.

2.4.2 HMM/GMM Approach

In the HMM/GMM approach, each acoustic unit is represented by a GMM. The parameters of the acoustic likelihood estimator in the HMM/GMM system are learned by optimizing a maximum likelihood based objective function using the expectation maximization (EM) algorithm given the pronunciation lexicon and acoustic training data [Rabiner, 1989]. There are two EM-based approaches to estimate the model parameters, namely, Baum-Welch training or forward-backward training [Rabiner, 1989] and embedded Viterbi training [Juang and Rabiner, 1990].

The EM algorithm is an iterative algorithm consisting of an expectation step (E-step) and a maximization step (M-step). In the E-step, state occupancy estimates are computed from the training data given the initial model. This involves estimating the probability distributions of hidden variables and the expected log likelihood given an initial model. In forward-backward training, the state occupancy estimates are described by probability distributions. In Viterbi training, first the segmentation of data is performed in terms of states; and state occupancy estimates are Kronecker delta distributions computed based on an explicit segmentation and labeling of training data. In the M-step, the maximum likelihood method is applied to update the model parameters using the state occupancy statistics computed from the training data. The HTK toolkit was used to build all HMM/GMM systems used in this thesis.

2.4.3 Hybrid HMM/ANN Approach

In the Hybrid HMM/ANN approach, acoustic units are modeled using an ANN. An ANN is first trained to estimate $p(a^d|\mathbf{x}_t,\Theta_A)$ and then the scaled-likelihood $p_{sl}(\mathbf{x}_t|a^d,\Theta_A)$ is estimated as:

$$p_{sl}(\mathbf{x}_t|a^d,\Theta_A) = \frac{p(\mathbf{x}_t|a^d,\Theta_A)}{p(\mathbf{x}_t)} = \frac{P(a^d|\mathbf{x_t},\Theta_A)}{P(a^d)} \tag{2.13}$$

$P(a^d)$ is estimated on the training dataset through counting.

The most common neural network used in ASR is a multilayer perceptron (MLP) [Morgan and Bourlard, 1995]. An MLP is a layered feedforward neural network with an input layer, zero or more hidden layers, and an output layer. The inputs to the MLP are cepstral features. Typically, at each time frame t, a left and right context of four frames $(\mathbf{x}_{t-4},\ldots,\mathbf{x}_t,\ldots,\mathbf{x}_{t+4})$ is used as input to the MLP. The output categories of MLP are acoustic units. Each layer of the MLP computes a set of linear discriminant functions using a weight matrix followed by a nonlinear function, which is often a sigmoid function. Most often, in the output layer a softmax nonlinear function is used. MLP estimates the posterior probabilities of output classes conditioned on the input [Morgan and Bourlard, 1995]. That is, the output $\mathbf{z}_t$ from the MLP at time t can be written as,

$$\begin{aligned}\mathbf{z}_t &= [z_t^1,\ldots,z_t^d,\ldots,z_t^D]^{\mathrm{T}} \\ &= [P(a^1|\mathbf{x}_t),\ldots,P(a^d|\mathbf{x}_t),\ldots,P(a^D|\mathbf{x}_t)]^{\mathrm{T}}\end{aligned} \tag{2.14}$$

where $z_t^d = P(a^d|\mathbf{x}_t)$ is the posterior probability the acoustic unit a^d given an acoustic feature observation vector $\mathbf{x}_t$.

The parameters of the neural network (weights and biases) can be trained using the back-propagation algorithm [Rumelhart et al., 1986] either in online training mode or batch training mode. In on-line training, the weights are adjusted in the direction of the error gradient with respect to the weight vector estimated from an example. In the batch training, the weights are adjusted after each batch of training data. The training criteria that are used most often to train the ANN are relative entropy and cross entropy. MLPs are prone to overfitting the training data, therefore, performance on cross-validation data (that is independent of the training data) is used to stop the ANN training.

Training data labeled in terms of network outputs is required to train the MLP for classification. It has been shown that the embedded Viterbi training procedure can be used for this purpose [Morgan and Bourlard, 1995]. Initially, the MLP is trained with uniform segmentation of training data. In the second step, this MLP is used to estimate the probabilities of acoustic units which are converted into the scaled-likelihoods. The scaled-likelihoods are used in dynamic programming to determine the new labels for the next MLP training. Each MLP training is done using labels from the previous Viterbi alignment and the procedure is iterated until convergence. Alternatively, MLP training can be started from the labeling and segmentation

obtained from the HMM/GMM system.

The parameters of the acoustic likelihood estimator in hybrid HMM/ANN ASR systems [Morgan and Bourlard, 1995, Dahl et al., 2012] are transition probabilities a_{ij}, weights and biases of the trained ANN, the prior probabilities of acoustic units and the decision trees or the deterministic map between lexical and acoustic units. The posterior probability of the acoustic unit or the scaled likelihood as given in Eqn (2.13) is used directly as the local emission score. The state transition probabilities a_{ij} in hybrid HMM/ANN systems are usually fixed to 0.5 (unless i is the first state of the first word, i.e., $a_{01} = 1$) [Morgan and Bourlard, 1995].

Alternatively, the posterior probabilities of acoustic units can replace conventional cepstral features in HMM-based ASR systems via the Tandem technique [Hermansky et al., 2000]. In order to model the output of the MLP (that is typically non-Gaussian) with GMMs, the posterior probabilities of acoustic units are Gaussianized using the log function and then decorrelated using the Karhunen-Loeve transform (KLT). Optionally, dimensionality reduction can also be performed by retaining only the feature components that contribute most to the variance. In this case, ANN is used as a pre-processor for feature extraction.

In this thesis, we use three layer MLPs trained to classify context-independent subword units with the cross entropy error criteria. The input to MLPs is the 39-dimensional PLP feature vector with a context of four preceding frames and four following frames. The labels for the training data were always obtained from the HMM/GMM system. All the MLPs are trained using the Quicknet software [4].

2.5 Language Model

As given in Eqn (2.4), the language model estimates the prior probability $P(W|\Theta_L)$ of a word sequence W. The prior probability $P(W)$ of a word sequence W can be factored using the chain rule of probability as:

$$P(W) = \prod_{m=1}^{M} P(\mathbf{w}_m|\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{m-1}) \tag{2.15}$$

ASR systems, typically use n-gram statistical language models where it is assumed that given the previous $n-1$ words, the probability of a word is independent of remaining history i.e.,

$$P(W) = \prod_{m=1}^{M} P(\mathbf{w}_m|\mathbf{w}_{m-(n-1)}, \ldots, \mathbf{w}_{m-1}) \tag{2.16}$$

If $n = 2$, language model is referred to as bi-gram language model and if $n = 3$ it is referred to as tri-gram language model.

The parameters of the language model Θ_L or the probabilities $P(\mathbf{w}_m|\mathbf{w}_{m-(n-1)}, \ldots, \mathbf{w}_{m-1})$ are

4. http://www.icsi.berkeley.edu/Speech/qn.html

estimated through counting using large collection of text from the language. However, even for bi-gram and tri-gram language models, it can be difficult to make good probability estimates for infrequent or unseen combination of words in the training data. Since, this is undesirable and an ASR system should be able to recognize word combinations not seen in the training, backoff smoothing method is normally employed [Katz, 1987]. In backoff smoothing, if there are enough examples for a tri-gram then the probability obtained from relative frequencies is used directly; if there are not enough examples then a bi-gram probability is used; if there are not enough examples for bi-gram then unigram probability is used. Additionally, the probability estimates for lower order n-grams are sometimes modified using methods such as Kneser-Ney smoothing [Kneser and Ney, 1995]. In this thesis, we used bi-gram language models employing Katz backoff smoothing. The language models are trained using the *ngram-count* tool of the *SRILM* toolkit [Stolcke, 2002].

2.6 Viterbi Decoder

As shown in Figure 2.1, during decoding the acoustic likelihood estimates and the language model prior probabilities are combined using Bayes rule as in Eqn (2.3) to infer the optimal word sequence given the acoustic feature observation sequence. The Viterbi algorithm [Forney, 1973] is used to find the most possible state sequence, and thereby the most probable word sequence (see Eqn (2.12)). According to Eqn (2.12), the acoustic likelihood and the language model probability must be computed for all possible state sequences. Unfortunately, this full breadth search is time consuming even for small vocabularies. Therefore, speech recognition systems employ pruning via beam search techniques where the hypotheses with scores less than a given threshold are discarded [Greer et al., 1982].

The acoustic likelihood and language model probability are estimated independently using different models, model different knowledge resources (acoustic data and text data respectively) and have different dynamic ranges. In practice, the acoustic likelihood scores are much smaller than those of language model probabilities. To prevent language model probabilities being dominated by acoustic likelihood, the language model probability is scaled before combining with acoustic likelihood. Also, to reduce large number of errors due to insertion of many short words, ASR systems also penalise the probability of transitions between words using word insertion penalty. In this thesis, we have used the *HVite* decoder tool provided with the HTK toolkit. The language model scale factor and the word insertion penalty of all the systems in this thesis are tuned on the development set.

2.7 Evaluation

The performance of all the ASR systems in this thesis is evaluated in terms of word accuracy computed using the *HResults* tool provided with the HTK toolkit [Young et al., 2006]. The *HResults* tool matches the recognised and reference word label transcriptions by performing an

optimal string match using dynamic programming. Given the optimal alignment, the number of substitution (e_S), deletion (e_D) and insertion (e_I) errors are calculated. The performance in terms of word accuracy is defined as the following:

$$\text{Word accuracy} = \frac{N - e_D - e_S - e_I}{N} \times 100\% \tag{2.17}$$

where N denotes the total number of words in the reference transcriptions.

2.8 Summary

In this chapter, we gave a brief overview of standard HMM-based ASR systems with its components: feature extraction, pronunciation lexicon, acoustic model, language model and decoder. In the next chapter, we present the framework of probabilistic lexical modeling.

3 Probabilistic Lexical Modeling

In this chapter, we introduce the framework of probabilistic lexical modeling (Section 3.1) and elucidate that standard HMM-based ASR approaches use a deterministic lexical model (Section 3.2). We present three probabilistic lexical model based ASR approaches (Section 3.3) and contrast them with the standard ASR approach where the lexical model is deterministic (Section 3.5).

3.1 Probabilistic Lexical Modeling Framework

In HMM-based ASR systems, as seen the previous chapter (see Section 2.1 and Eqn (2.11)), the estimation of joint density for X and Q involves the estimation of the local emission score $p(\mathbf{x}_t|q_t = l^i,\Theta_A)$ and the transition score $P(q_t = l^i|q_{t-1} = l^j,\Theta)$. The local emission score $p(\mathbf{x}_t|q_t = l^i,\Theta_A)$ or the dependency between acoustic feature observation $\mathbf{x}_t$ and lexical unit l^i can be factored through a *latent* variable a^d as following:

$$p(\mathbf{x}_t|q_t = l^i,\Theta_A) = \sum_{d=1}^{D} p(\mathbf{x}_t, a^d|q_t = l^i,\Theta_A) \tag{3.1}$$

$$= \sum_{d=1}^{D} p(\mathbf{x}_t|a^d, q_t = l^i,\theta_a,\theta_l)\cdot P(a^d|q_t = l^i,\theta_l) \tag{3.2}$$

$$= \sum_{d=1}^{D} \underbrace{p(\mathbf{x}_t|a^d,\theta_a)}_{\textbf{acoustic model}}\cdot \underbrace{P(a^d|q_t = l^i,\theta_l)}_{\textbf{lexical model}} \tag{3.3}$$

We refer to the latent variable a^d as the acoustic unit and the set of acoustic units $\mathscr{A} = \{a^1,\dots a^d,\dots a^D\}$ where D is the total number of acoustic units. The relationship in Eqn (3.3) is a result of the assumption that given a^d, $p(\mathbf{x}_t|a^d, q_t = l^i,\theta_a,\theta_l)$ is independent of l^i. In Eqn (3.3), $p(\mathbf{x}_t|a^d,\theta_a)$ is the acoustic unit likelihood and $P(a^d|l^i,\theta_l)$ is the probability of the acoustic unit given the lexical unit and is given by the lexical model. In this thesis, we refer to $p(\mathbf{x}_t|a^d,\theta_a)$ as the acoustic model evidence and $P(a^d|l^i,\theta_l)$ as the lexical model evidence. The parameters of the acoustic likelihood estimator Θ_A now encompass the *acoustic*

model (θ_a), the *pronunciation lexicon* (θ_{pr}) and the *lexical model* (θ_l) parameters, therefore, $\Theta_A = \{\theta_a, \theta_{pr}, \theta_l\}$.

Figure 3.1 shows the Bayesian network of an ASR system that uses the factorization of Eqn (3.3). The lexical unit is given deterministically by the current word and its subword units. The lexical unit is mapped to all acoustic units probabilistically and the acoustic feature observation is conditioned on the acoustic units.

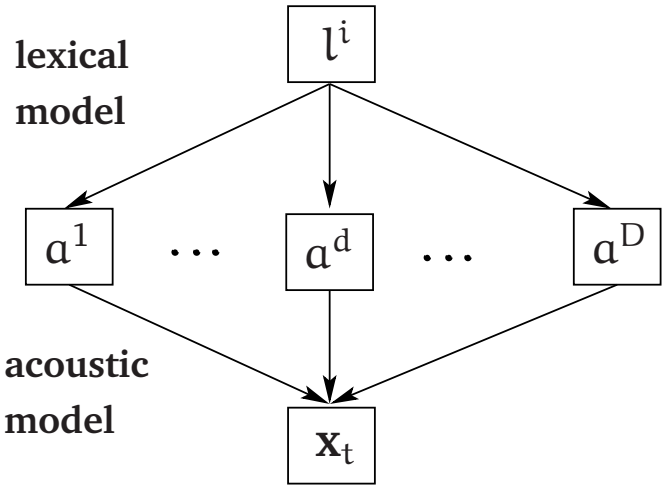

Figure 3.1 – The Bayesian network of an ASR system based on the Eqn (3.3)

For a lexical unit l^i, the lexical model evidence can be seen as a D dimensional categorical variable $\mathbf{y}_i = [y_i^1, \ldots, y_i^d, \ldots y_i^D]^{\mathrm{T}}$, $y_i^d = P(a^d|l^i, \theta_l)$ that models a probabilistic relationship between a lexical unit l^i and D acoustic units. Given the acoustic feature observation sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, the acoustic model computes a sequence of acoustic unit likelihood vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_T\}$, where $\mathbf{v}_t = [v_t^1, \ldots, v_t^d, \ldots, v_t^D]^{\mathrm{T}}$ and $v_t^d = p(\mathbf{x}_t|a^d, \theta_a)$. Having defined $\mathbf{y}_i$ and $\mathbf{v}_t$, Eqn (3.3) can be written as the following:

$$p(\mathbf{x}_t|q_t = l^i, \Theta_A) = \sum_{d=1}^{D} p(\mathbf{x}_t|a^d, \theta_a) \cdot P(a^d|q_t = l^i, \theta_l) \tag{3.4}$$

$$= \mathbf{y}_i^{\mathrm{T}} \mathbf{v}_t \tag{3.5}$$

Eqn (3.5) can be seen as a match between the acoustic and lexical model evidence which in this case turns out to be the scalar product of $\mathbf{y}_i$ and $\mathbf{v}_t$. The various components of a probabilistic lexical model based ASR system are illustrated in Figure 3.2.

The graphical model representation of a system based on Eqns (2.12) and (3.5) for the word sequence "IS IT" is illustrated in Figure 3.3. In the figure, I and F refer to the initial and final HMM states. The figure shows that the sequence of words constrained by the language model are represented by a sequence of lexical units (l^{ih} l^z l^{ih} l^t) as given by the pronunciation

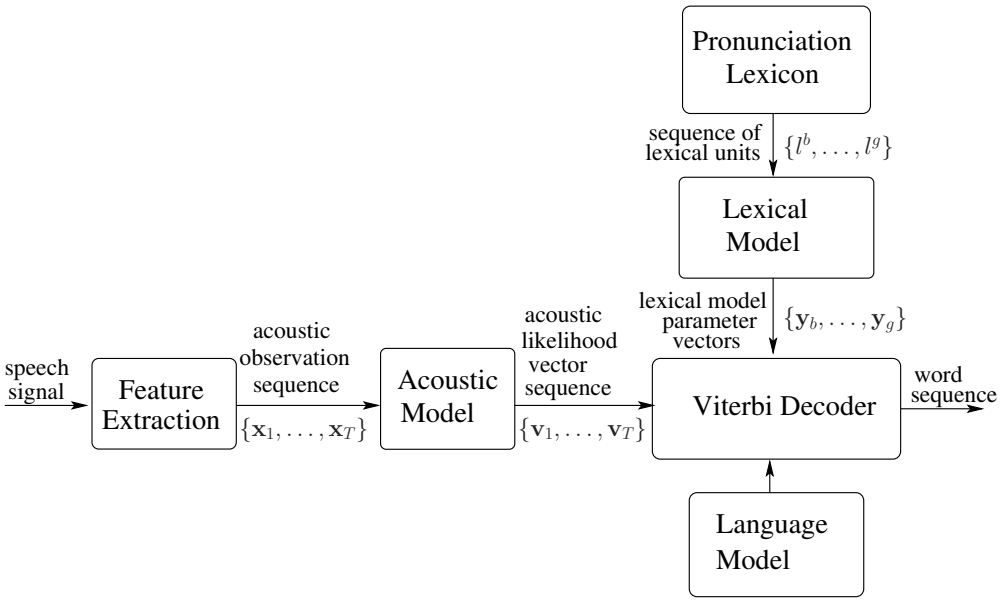

Figure 3.2 – Block diagram of a probabilistic lexical model based ASR system

lexicon. For each lexical unit l^i, the lexical model computes a D dimensional categorical variable $\mathbf{y}_i$. For each acoustic observation sequence $\mathbf{x}_t$, the acoustic model computes a D dimensional acoustic unit likelihood vector $\mathbf{v}_t$. The local emission score at time frame t is the match between the acoustic model evidence $\mathbf{v}_t$ and the lexical model evidence $\mathbf{y}_i$.

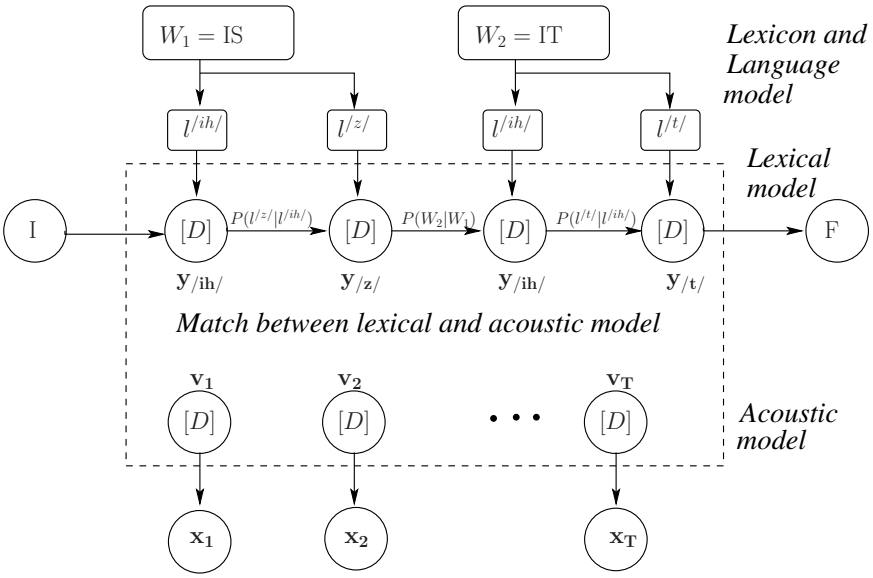

Figure 3.3 – The graphical model representation of a system based on Eqns (2.12) and (3.3)

3.2 Deterministic Lexical Model based ASR Approaches

In standard HMM-based ASR approaches like HMM/GMM and Hybrid HMM/ANN, the lexical model is deterministic, i.e., each lexical unit l^i is deterministically mapped to an acoustic unit a^j ($l^i \mapsto a^j$) as shown in Figure 3.4(a), i.e.,

$$y_i^d = P(a^d|q_t = l^i, \theta_l) = \begin{cases} 1, & \text{if d = j;} \\ 0, & \text{otherwise.} \end{cases} \tag{3.6}$$

As a result of the deterministic mapping, the only term contributing to the summation in Eqn (3.3) is the acoustic unit that is mapped to the lexical unit at time t. The Bayesian network of an ASR system at time frame t in which the lexical model is deterministic is illustrated in Figure 3.4(b). The lexical unit is given deterministically by the current word and its subword units. A lexical unit is mapped to an acoustic unit and the acoustic feature observation is conditioned on an acoustic unit.

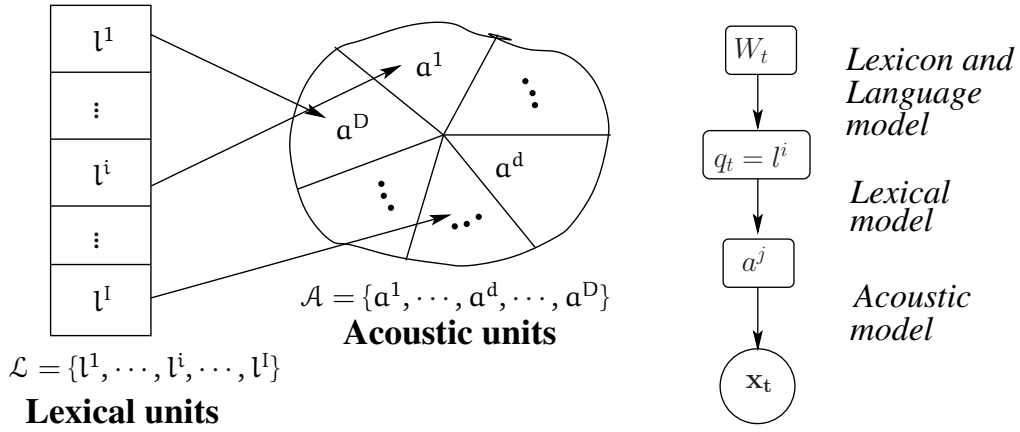

Figure 3.4 – Deterministic lexical modeling (a) deterministic mapping, and (b) graphical model representation

In context-independent subword unit based ASR systems, the deterministic relationship between lexical and acoustic units is knowledge driven. Thus, lexical model training is not involved. Therefore, in context-independent subword unit based ASR systems, the deterministic map between lexical and acoustic units is the lexical model and the GMMs (in the case of HMM/GMM) or the ANN (in the case of Hybrid HMM/ANN) is the acoustic model.

In context-dependent subword unit based ASR systems, lexical units are context-dependent subword units whereas acoustic units are clustered context-dependent subword units. As mentioned in Section 2.4.1, the decision trees i.e., the tree structure and the phonetic question set are used to deterministically relate a lexical unit to an acoustic unit. Therefore, in context-dependent subword unit based HMM/GMM systems, decision trees are the lexical model and the GMMs are the acoustic model. Similarly, in the case of Hybrid HMM/ANN systems, decision trees are the lexical model and the ANN is the acoustic model [Dahl et al., 2012, Hinton et al., 2012].

It is worth mentioning that in the HMM-based ASR literature, due to this deterministic relationship, typically no distinction is made between the acoustic and lexical units, or the acoustic and lexical models. Our main reason to refer to lexical unit l^i and acoustic unit a^j, or the acoustic and lexical models distinctively here is to bring out the contributions of the thesis clearly.

3.3 Probabilistic Lexical Model based ASR Approaches

The Eqn (3.3) with the two conditions, namely, $P(a^d|l^i,\theta_l) > 0$ and $\sum_{d=1}^{D} P(a^d|l^i,\theta_l) = 1$ characterizes an ASR approach where each lexical unit is probabilistically related to all acoustic units. We refer to them as probabilistic lexical model based ASR systems.

The probabilistic lexical modeling approaches presented in this chapter presume that an acoustic unit set $\mathscr{A}$ is defined and a trained acoustic model is available. Therefore, in the first step a standard HMM-based ASR system i.e., either an HMM/GMM system or a Hybrid HMM/ANN system is trained. The acoustic model (i.e., GMMs in the case of HMM/GMM or ANN in the case of Hybrid HMM/ANN) is used with the pronunciation lexicon and acoustic data to train the parameters of the probabilistic lexical model. More specifically, the parameters of the probabilistic lexical model are learned by training an HMM, whose states represent lexical units and each state l^i is parameterized by a categorical distribution $\mathbf{y}_i = [y_i^1,\ldots,y_i^d,\ldots y_i^D]^{\mathrm{T}}$ and $\sum_{d=1}^{D} y_i^d = 1$. The categorical distribution captures a probabilistic relationship between a lexical unit l^i and D acoustic units i.e., $y_i^d = P(a^d|l^i,\theta_l)$. In this case, the lexical model parameter set consists of $\theta_l = \{\mathbf{y}_i\}_{i=1}^{I}$.

We present these techniques from the perspective of hybrid HMM/ANN. That is, in this thesis we use an ANN, more precisely, an MLP as an acoustic model. ANNs are discriminative classifiers, and can provide invariance towards undesirable variabilities such as speaker and environment [Zhu et al., 2004, Ikbal, 2004]. Furthermore, ANNs have been shown to be effective for multilingual and crosslingual portability [Stolcke et al., 2006, Thomas et al., 2012, Lal and King, 2013]. As shown in Chapter 9, these approaches are equally applicable to the HMM/GMM framework.

3.3.1 Kullback-Leibler Divergence based HMM

In the Kullback-Leibler divergence based HMM (KL-HMM) approach [Aradilla et al., 2007, 2008], the parameters of the lexical model are learned through acoustic unit posterior probability estimates. That is the feature observations used to train the HMM are $\mathbf{z}_t = [z_t^1\ldots,z_t^d,\ldots,z_t^D)]^{\mathrm{T}}$ where $z_t^d = P(a^d|\mathbf{x}_t,\theta_a)$.

As both feature observations and state distributions are probability vectors, the local score or the match between acoustic and lexical model evidence at each HMM state can be the Kullback-Leibler (KL) divergence between the feature observation $\mathbf{z}_t$ and the categorical

distribution $\mathbf{y}_i$,

$$S_{KL}(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} y_i^d \log\Big(\frac{y_i^d}{z_t^d}\Big) \tag{3.7}$$

The above equation represents the case where the lexical model $\mathbf{y}_i$ is the reference distribution and the local score is denoted as S_{KL}. However, KL-divergence is an asymmetric measure. Thus, there are other possible ways to estimate the KL-divergence:

1. Reverse KL-divergence (S_{RKL}): In this case the acoustic unit probability vector $\mathbf{z}_t$ is the reference distribution.

$$S_{RKL}(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} z_t^d \log\Big(\frac{z_t^d}{y_i^d}\Big) \tag{3.8}$$

2. Symmetric KL-divergence (S_{SKL}): The local score S_{SKL} is the average of the local scores S_{KL} and S_{RKL}.

$$S_{SKL}(\mathbf{y}_i, \mathbf{z}_t) = \frac{1}{2} \cdot [S_{KL} + S_{RKL}] \tag{3.9}$$

Figure 3.5 illustrates the KL-HMM approach. The acoustic model or the ANN is trained to classify D acoustic units. Given the acoustic model, acoustic unit probability sequences of training data are estimated. The acoustic unit probability sequences are used as feature observations to train an HMM where states represent lexical units. The states of the HMM are parameterized by categorical distributions.

Training

Given a trained ANN and training set of N utterances $\{X(n), W(n)\}_{n=1}^{N}$, the set of acoustic unit probability vectors $\{Z(n), W(n)\}_{n=1}^{N}$ is estimated where for each training utterance n, $X(n)$ represents a sequence of cepstral features of length $T(n)$, $W(n)$ represents the sequence of underlying words, and $Z(n)$ represents a sequence of acoustic unit probability vectors of length $T(n)$.

The KL-HMM system is parameterized by $\Theta_{kull} = \{\{\mathbf{y}_i\}_{i=1}^{I}, \{a_{ij}\}_{i,j=1}^{I}\}$. The training data $\{Z(n), W(n)\}_{n=1}^{N}$ and the current parameter set Θ_{kull}, are used to estimate a new set of parameters $\hat{\Theta}_{kull}$ using the Viterbi expectation maximization algorithm which minimizes a cost function based on the local scores S_{KL} or S_{RKL} or S_{SKL}. The lexical model parameters $\{\mathbf{y}_i\}_{i=1}^{I}$ are initialized uniformly i.e., initially $y_i^d = \frac{1}{D}\ \forall i, d$. In the case of the local score S_{RKL} the cost function minimized is,

$$\hat{\Theta}_{kull} = \underset{\Theta_{kull}}{\arg\min}\Big[\sum_{n=1}^{N} \min_{Q \in \mathcal{Q}} \sum_{t=1}^{T(n)} \big[S_{RKL}(\mathbf{y}_{q_t}, \mathbf{z}_t(n)) - \log a_{q_{t-1}q_t}\big]\Big] \tag{3.10}$$

where $Q = \{q_1, \ldots, q_t, \ldots, q_{T(n)}\}$, $q_t \in \mathcal{L} = \{l^1, \ldots l^i \ldots l^I\}$ and $\mathcal{Q}$ denotes set of all possible HMM

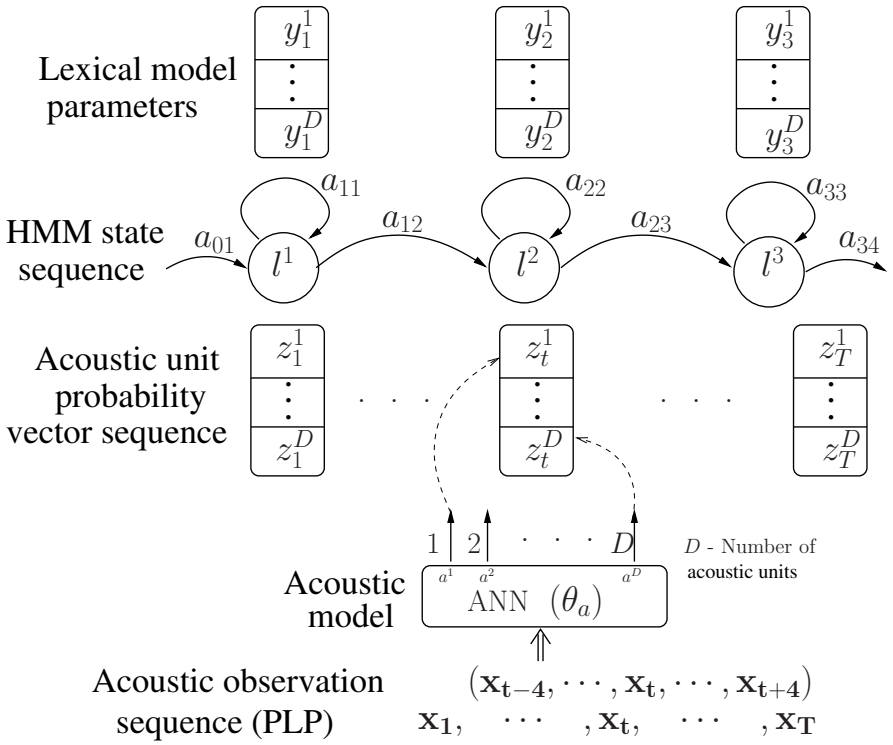

Figure 3.5 – Illustration of the KL-HMM approach

state sequences.

The training process involves iteration over the segmentation and the optimization steps until convergence. Given the current set of parameters, the segmentation step yields an optimal state sequence for each training utterance using the Viterbi algorithm. The optimization step estimates a new set of model parameters by minimizing Eqn (3.10) subject to the constraint that $\sum_{d=1}^{D} y_i^d = 1$, given optimal state sequences and acoustic unit posterior vectors. The state transition probabilities a_{ij} are fixed to 0.5 (unless i is the first state of the first word, i.e., $a_{01} = 1$), as it is usually done in hybrid HMM/ANN systems [Bourlard and Morgan, 1994]. The parameter estimation for local score S_{RKL} is illustrated in Figure 3.6.

Each of the KL-divergence based local scores lead to a different optimal state categorical distribution [Aradilla, 2008]. More precisely, if $Z(i)$ denotes the set of acoustic unit probability vectors assigned to state i (by the segmentation step) and $M(i)$ is the cardinality of $Z(i)$ then:

1. The optimal state distribution for the local score S_{KL} is the normalized geometric mean of the acoustic unit probability vectors assigned to the state, i.e.,

$$y_i^d = \frac{\bar{y}_i^d}{\sum_{d=1}^{D} \bar{y}_i^d} \text{ where } \bar{y}_i^d = \Big(\prod_{\mathbf{z}_t(n) \in Z(i)} z_t^d(n) \Big)^{\frac{1}{M(i)}} \quad \forall d \tag{3.11}$$

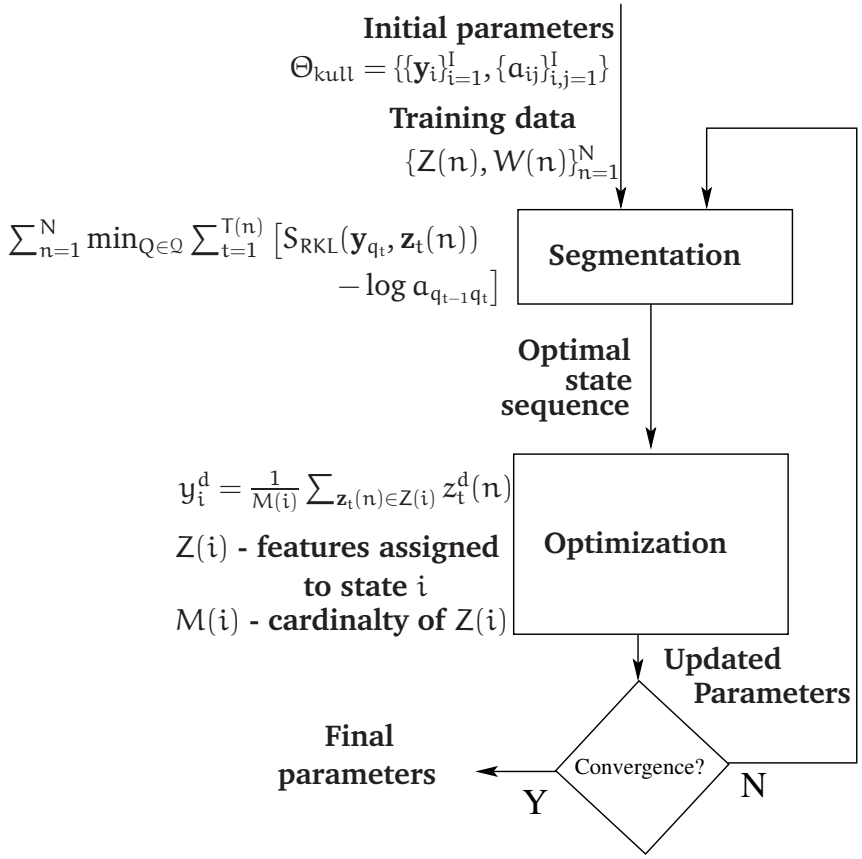

Figure 3.6 – Illustration of parameter estimation in the case of KL-HMM with local score S_{RKL}

where $\bar{y}_i^d$ represents the geometric mean of state i for dimension d.

2. The optimal state distribution for the local score S_{RKL} is the arithmetic mean of the acoustic unit probability vectors assigned to the state, i.e.,

$$y_i^d = \frac{1}{M(i)} \sum_{\mathbf{z}_t(n) \in Z(i)} z_t^d(n) \quad \forall d \tag{3.12}$$

3. For the local score S_{SKL}, there is no closed form solution to find the optimal state distribution. The optimal state distribution can be computed iteratively using the arithmetic and the normalized geometric mean of the acoustic unit probability vectors assigned to the state [Veldhuis, 2002].

Decoding

It is worth mentioning that KL-HMM was originally developed from the perspective of acoustic modeling [Aradilla et al., 2008], as an alternative to the Tandem approach [Hermansky et al., 2000]. As described in Section 2.4.3, in the Tandem approach, the outputs of ANN (that are

typically non-Gaussian) are first Gaussianized using the log function and then decorrelated using the KLT transformation. These transformed features are used as feature observations and modeled with GMMs. Rather than transforming the output of ANN as in the Tandem approach, in the KL-HMM approach the state distribution and the local score associated with each state of HMM are changed to categorical distribution and KL divergence, respectively [Aradilla et al., 2008].

However, the KL-HMM approach is a probabilistic lexical modeling approach where in Eqn (3.5) (see Chapter 2), the acoustic unit posterior probability vector ($\mathbf{z}_t$) is used in place of the acoustic unit likelihood vector ($\mathbf{v}_t$). Furthermore, given that both lexical evidence and acoustic evidence are in the form of posterior probabilities, the loglikelihood-based score in the standard Viterbi decoding of Eqn (2.12) is replaced with the negative of KL-divergence.

Given a sequence of acoustic unit probability vectors $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_t, \ldots, \mathbf{z}_T\}$ and the trained parameters $\Theta_{kull} = \{\{\mathbf{y}_i\}_{i=1}^{I}, \{a_{ij}\}_{i,j=1}^{I}\}$, decoding involves recognition of the underlying hypothesis W^*. The most likely word sequence W^* is obtained by finding the most likely state sequence Q^*, i.e.,

$$Q^* = \arg\max_{Q \in \mathcal{Q}} \sum_{t=1}^{T} [\log p(\mathbf{x}_t | q_t = l^i, \Theta_A) + \log P(q_t = l^i | q_{t-1} = l^j, \Theta)] \tag{3.13}$$

$$= \arg\max_{Q \in \mathcal{Q}} \sum_{t=1}^{T} [-S_{RKL}(\mathbf{y}_{q_t}, \mathbf{z}_t) + \log a_{q_{t-1} q_t}] \tag{3.14}$$

$$= \arg\min_{Q \in \mathcal{Q}} \sum_{t=1}^{T} [S_{RKL}(\mathbf{y}_{q_t}, \mathbf{z}_t) - \log a_{q_{t-1} q_t}] \tag{3.15}$$

3.3.2 Tied Posterior

In the second approach, the probabilistic lexical model is learned through scaled-likelihood estimates $p_{sl}(\mathbf{x}_t | a^d, \theta_a)$ (see Eqn (2.13)). The approach referred to as tied posterior approach [Rottland and Rigoll, 2000], was originally proposed in the framework of hybrid HMM/ANN to build context-dependent subword unit based ASR system using an ANN trained to classify context-independent subword units.

In the tied-posterior based HMM (Tied-HMM) approach, the emission likelihood at each context-dependent state $q_t = l_{cd}^i$ is estimated as,

$$p(\mathbf{x}_t | q_t = l_{cd}^i) = \sum_{d=1}^{D} w_i^d \cdot p_{sl}(\mathbf{x}_t | a_{ci}^d) \tag{3.16}$$

where a_{ci}^d is a context-independent phone, D here refers to the number of context-independent phones, $p_{sl}(\mathbf{x}_t | a_{ci}^d)$ is the scaled-likelihood (see Eqn (2.13)), $0 \leq w_i^d \leq 1$ is the weight corresponding to the context-dependent phone l_{cd}^i and $\sum_{d=1}^{D} w_i^d = 1$. The weights w_i^d are estimated by maximizing the log-likelihood using the EM algorithm. Comparison

between (3.16) and (3.3) shows that l^i_{cd} corresponds to the lexical unit l^i, a^d_{ci} corresponds to the acoustic unit a^d and w^d_i corresponds to $y^d_i = P(a^d|l^i,\theta_l)$. In other words, the Tied-HMM approach is an HMM-based ASR approach that incorporates probabilistic lexical modeling.

The Tied-HMM approach can be interpreted along lines similar to those of the KL-HMM approach where the states of the HMM are parameterized by $\mathbf{y}_i$. However, the feature observations used to train the HMM in the Tied-HMM approach are vectors of scaled-likelihood $\mathbf{v}_t = [v^1_t \ldots, v^d_t, \ldots, v^D_t]^{\mathrm{T}}$ where $v^d_t = p_{sl}(\mathbf{x}_t|a^d,\theta_a)$, and the local score is

$$S_{tied}(\mathbf{y}_i,\mathbf{v}_t) = \log\big(\sum_{d=1}^{D} y^d_i.v^d_t\big) = \log\big(\mathbf{y}_i^{\mathrm{T}}\mathbf{v}_t\big) \tag{3.17}$$

Training

Given a training set of N utterances $\{X(n),W(n)\}_{n=1}^N$, the set of likelihood vectors $\{V(n),W(n)\}_{n=1}^N$ is formed where $V(n) = \{\mathbf{v}_1(n),\ldots,\mathbf{v}_t(n),\ldots,\mathbf{v}_{T(n)}(n)\}$, $\mathbf{v}_t(n) = [v^1_t(n)\ldots,v^d_t(n),\ldots,v^D_t(n)]^{\mathrm{T(n)}}$ and $v^d_t(n)$ denotes scaled-likelihood.

The parameters of the model $\Theta_{tied} = \{\{\mathbf{y}_i\}_{i=1}^I,\{a_{ij}\}_{i,j=1}^I\}$ are estimated by the Viterbi expectation maximization algorithm that maximizes the cost function,

$$\hat{\Theta}_{tied} = \underset{\Theta_{tied}}{\arg\max}\Big[\sum_{n=1}^{N}\max_{Q\in\mathcal{Q}}\sum_{t=1}^{T(n)}\big[S_{tied}(\mathbf{y}_{q_t},\mathbf{v}_t(n)) + \log(a_{q_{t-1}q_t})\big]\Big] \tag{3.18}$$

where $Q = \{q_1,\ldots q_t,\ldots,q_{T(n)}\}$, $q_t \in \mathscr{L} = \{l^1,\ldots l^i \ldots l^I\}$ and $\mathcal{Q}$ denotes a set of all possible HMM state sequences.

Similar to the KL-HMM approach, the training process involves iteration over the segmentation and the optimization steps until convergence. The optimization step estimates a new set of model parameters $y^d_i \quad \forall d$ by setting the derivative of Eqn (3.18) with respect to y^d_i to zero, subject to the constraint that $\sum_{d=1}^D y^d_i = 1$, i.e.,

$$\frac{\partial}{\partial y^d_i}\Big[\sum_{n=1}^{N}\max_{Q\in\mathcal{Q}}\sum_{t=1}^{T(n)}\big[S_{tied}(\mathbf{y}_{q_t},\mathbf{v}_t(n))\big] + \lambda_i(\sum_{d=1}^{D} y^d_i - 1)\Big] = 0 \tag{3.19}$$

Replacing the local score from Eqn (3.17) and solving the above equation results in,

$$\sum_{\mathbf{v}_t(n)\in V(i)}\frac{v^d_t(n)}{\sum_{d=1}^D y^d_i.v^d_t(n)} + \lambda_i = 0 \tag{3.20}$$

where $V(i)$ denotes the set of acoustic unit probability vectors assigned to state l^i. Rearranging

the above equation results in,

$$\lambda_i = -\sum_{\mathbf{v}_t(n)\in V(i)} \frac{v_t^d(n)}{\sum_{d=1}^{D} y_i^d . v_t^d(n)} \tag{3.21}$$

By multiplying on both sides by y_i^d, summing over d and applying the sum to one constraint we obtain,

$$\begin{aligned}\lambda_i &= -\sum_{\mathbf{v}_t(n)\in V(i)} 1 \\ &= -M(i)\end{aligned} \tag{3.22}$$

where $M(i)$ is the cardinality of $V(i)$. Replacing the value of λ_i and then rearranging results in:

$$y_i^{d*} = \frac{1}{M(i)} \sum_{\mathbf{v}_t(n)\in V(i)} \frac{y_i^d . v_t^d(n)}{\sum_{k=1}^{D} y_i^k . v_t^k(n)} \tag{3.23}$$

where y_i^{d*} refers to the updated y_i^d. It can be observed from the above equation that there is no closed form solution to compute the lexical model parameters as the update equation also includes y_i^d.

The decoding is performed by replacing the log-likelihood based score in the standard Viterbi decoder of Eqn (2.12) with the local score given in Eqn (3.17).

The Tied-HMM approach was motivated by semi continuous HMMs [Huang and Jack, 1989, Bellegarda and Nahamoo, 1990]. In semi continuous HMMs, the emission likelihood is computed as the weighted sum of a pool of Gaussian densities. However, there is a key difference between Tied-HMM and semi-continuous HMM approaches. In the Tied-HMM approach, each acoustic unit is explicitly related to a context-independent subword unit and has its own density function (ANN), whereas in semi-continuous HMMs, the Gaussian densities do not have any link to context-independent or context-dependent subword units. In that sense, Tied-HMM is equivalent to probabilistic classification of HMM states (PC-HMM) proposed in [Luo and Jelinek, 1999] where acoustic units are clustered context-dependent phones and acoustic units are modeled with GMMs instead of an ANN. In both Tied-HMM and PC-HMM approaches, lexical units (context-dependent phones) are probabilistically related to all acoustic units (context-independent phones or clustered context-dependent phones).

3.3.3 Scalar Product HMM

In the KL-HMM system, the local score is based on KL-divergence. However, two posterior probability distributions can be compared with different cost functions such as scalar product [Asaei et al., 2010] and Bhattacharya distance [Soldo et al., 2011]. It is possible to envisage

an HMM where the local score is based on scalar product, i.e.,

$$S_{SP}(\mathbf{y}_i, \mathbf{z}_t) = \log\left(\mathbf{y}_i^{\mathrm{T}} \mathbf{z}_t\right) \tag{3.24}$$

We refer to this system as scalar product HMM (SP-HMM). Again, $\{\mathbf{y}_i\}_{i=1}^{I}$ can be estimated using the embedded Viterbi training algorithm, and the decoding can be performed by replacing the log-likelihood based score in the standard Viterbi decoder with $S_{SP}(\mathbf{y}_i, \mathbf{v}_t)$.

The SP-HMM is of particular interest for the following reasons:

1. It can be seen as a particular case of the Tied-HMM approach where the priors in the scaled-likelihood estimation are dropped or assumed to be equal. The optimal state distribution for the SP-HMM approach is,

$$y_i^{d*} = \frac{1}{M(i)} \sum_{\mathbf{z}_t(n) \in Z(i)} \frac{y_i^d . z_t^d(n)}{\sum_{k=1}^{D} y_i^k . z_t^k(n)} \quad \forall d \tag{3.25}$$

   where $Z(i)$ denotes the set of acoustic unit probability vectors assigned to state l^i and $M(i)$ is the cardinality of $Z(i)$.

2. SP-HMM and KL-HMM differ only in terms of the cost function used for parameter estimation and decoding.

3.4 Effect of Cost Functions on Lexical Model Parameter Estimation

As seen in the previous section, the optimal state distribution is different in various probabilistic lexical modeling approaches. The difference comes from the cost function optimized during parameter estimation. In the case of KL-HMM approaches, KL-divergence based local score between state distribution and feature observations belonging to a state is minimized, whereas in the case of Tied-HMM and SP-HMM approaches, the respective local score (S_{tied} or S_{SP}) between state distribution and feature observations belonging to a state is maximized.

The optimal state distributions for local scores S_{KL} and S_{RKL} are the geometric mean and arithmetic mean of probability vectors assigned to a state. More specifically, the optimal state distributions in the case of local scores S_{KL} and S_{RKL} are the combination of probability vectors. Depending on the training data, the probability vectors can be seen as coming from different lexical contexts, speakers, accents, dialects, environmental conditions etc. The optimal state distribution is obtained by combining different sources of variability to form an aggregate probability distribution. In that sense, the parameter estimation step in the KL-HMM approach is similar to classifier fusion using multiple probability distributions [Genest and Zidek, 1986, Kittler et al., 1998, Abbas, 2009]. In the classifier fusion literature,

- the geometric mean of probabilities is referred to as product combination or log-linear opinion pool. It has been shown that product combination or log-linear pool leads to a less

dispersive distribution, i.e., it captures the dominant decision; and
- the arithmetic mean of probabilities is referred to as sum combination or linear opinion pool. It has been shown that sum combination leads to a dispersive distribution compared to product combination, i.e., it captures competing decisions.

Therefore, from the point of view of lexical model parameters, the cost function based on local score S_{KL} leads to a less dispersive (or low entropy) distribution compared to the cost function based on local score S_{RKL}. Therefore, we hypothesize that the cost functions based on local scores S_{KL} and S_{RKL} can model better one-to-one and one-to-many lexical-to-acoustic unit relationships, respectively. The cost function based on local score S_{SKL} (which is the average of local scores S_{KL} and S_{RKL}) is hypothesized to model both one-to-one and one-to-many lexical-to-acoustic unit relationships. We validate this hypothesis later in Section 4.4.2.

To understand the update equation for the Tied-HMM approach, the right hand side of Eqn (3.23) is expanded in terms of the definition of individual variables i.e., $y_i^d = P(a^d|l^i)$, and $v_t^d = p(\mathbf{x}_t|a^d)$,

$$\frac{v_t^d(n).y_i^d}{\sum_{k=1}^{D} v_t^k(n).y_i^k} = \frac{p(\mathbf{x}_t(n)|a^d)P(a^d|l^i)}{\sum_{k=1}^{D} p(\mathbf{x}_t(n)|a^k)P(a^k|l^i)} \tag{3.26}$$

$$= P(a^d|\mathbf{x}_t(n), l^i) \tag{3.27}$$

The Eqn (3.27) represents the probability of acoustic unit a^d given the feature observation $\mathbf{x}_t$ and lexical unit l^i. In other words, the update for Tied-HMM is given by the average probability of the acoustic unit a^d given the lexical unit l^i and the feature vectors belonging to state l^i.

In the case of the KL-HMM KL and KL-HMM RKL approaches, the optimal state distribution is obtained by the classification of acoustic units (a^d) given the acoustic feature observations ($\mathbf{x}_t$) belonging to the state, while in the case of the Tied-HMM and SP-HMM approaches, it is obtained by the classification of acoustic units (a^d) given the lexical unit (l^i) and the acoustic feature observations ($\mathbf{x}_t$) belonging to the state.

3.5 Comparison between ASR Approaches

In the previous section, we have seen the effect of various local scores on lexical model parameter estimation. In this section, the similarities and differences between deterministic lexical model approaches and various probabilistic lexical modeling approaches are presented from decoding perspective.

3.5.1 Deterministic Lexical Model and Probabilistic Lexical Model based ASR Systems

Deterministic lexical modeling based ASR systems and probabilistic lexical modeling based ASR systems are similar in the sense that both achieve speech recognition by matching the sequence of reference 'latent' symbols obtained from the pronunciation lexicon and syntactic information with sequence of 'latent' symbols obtained from the acoustic speech signal. As discussed in [Razavi et al., 2014] and summarized in Table 3.1, deterministic lexical modeling and probabilistic lexical modeling based ASR systems can be compared using the following four components:

1. Latent symbols: The latent symbols or the acoustic units can be either context-independent subword units or clustered context-dependent subword units.
2. Acoustic model: The acoustic model models the relationship between latent symbols and acoustic features through a generative model like a GMM (as in HMM/GMM systems) or through a discriminative model like an ANN (as in hybrid HMM/ANN systems).
3. Lexical model: The lexical model models the relationship between lexical units and latent variables. In the case of deterministic lexical modeling based systems it is a deterministic relationship (as given in Eqn (3.6)) and in the case of probabilistic lexical model based systems it is a probabilistic relationship (as given in Eqn (3.3)).
4. Local score: The local score or the match between acoustic and lexical model evidence. In the case of deterministic lexical model based systems like HMM/GMM and hybrid HMM/ANN, the local score can be seen as the scalar product between the acoustic unit likelihood vector sequence and the lexical model parameter vector sequence as given in Chapter 2. Since the lexical model parameter vector is a Kronecker delta as given in Eqn (3.6), the only term contributing to the score is the latent symbol that is tied to a lexical unit. As given in this chapter, for probabilistic lexical modeling based systems, the local score can be the KL-divergence or the scalar product between the acoustic unit probability vector and the lexical model parameter vector (as in KL-HMM and SP-HMM approaches) or scalar product between the acoustic unit likelihood vector and the lexical model parameter vector (as in the Tied-HMM approach).

The language model component and the efficient search of the output word hypothesis using dynamic programming is common to deterministic lexical modeling and probabilistic lexical modeling based ASR systems.

Table 3.1 – Comparison between deterministic lexical modeling and probabilistic lexical modeling based ASR systems

	Deterministic Lexical Model	Probabilistic Lexical Model
Latent symbols	Context-independent or clustered context-dependent subword units	
Acoustic model	GMM or ANN	
Lexical model	Deterministic	Probabilistic
Local score	Scalar product	KL-divergence, scalar product

3.5.2 Probabilistic Lexical Model based ASR Systems

In ASR, the local score estimation at time frame t can be seen as a match between "bottom-up" acoustic information $\mathbf{z}_t$ or $\mathbf{v}_t$ and "top-down" lexical information $\mathbf{y}_i$ related to latent variable a^d, as shown in Figure 3.3. Yet another similarity between the three approaches is that they reduce to the standard hybrid HMM/ANN system described earlier in Chapter 2 when the lexical model is deterministic, i.e., $\mathbf{y}_i$ is the Kronecker delta distribution. However, the KL-HMM approach has additional advantages compared to the Tied-HMM and SP-HMM approaches. We will discuss them briefly in this section.

From the communication theory perspective [Bahl et al., 1983], the standard HMM-based ASR approach can be seen as a communication problem where the noisy output of acoustic channel (i.e., a sequence of acoustic unit likelihood vectors $\{\mathbf{v}_1,\ldots,\mathbf{v}_T\}$ or a sequence of acoustic unit posterior vectors $\{\mathbf{z}_1,\ldots,\mathbf{z}_T\}$) is decoded by a linguistic decoder, which implies comparison to possible sequences of lexical model parameter vectors (for e.g. $\{\mathbf{y}_i,\ldots,\mathbf{y}_g\}$ where $i,g \in \{1,\ldots,I\}$) with lexical transition constraints ($P(q_t = l^i | q_{t-1} = l^j)$). Thus, standard HMM-based ASR inherently gives more importance to the lexical model and consequently relies on purity or correctness of the lexical knowledge imparted into the system. This aspect has been particularly observed in the case of pronunciation variation modeling of conversational speech where one of the best approaches is to add pronunciation variants, i.e., improve the deterministic lexical model [Strik and Cucchiarini, 1999].

The KL-HMM approach using the local score $S_{KL}(\mathbf{y}_i,\mathbf{z}_t)$ where $\mathbf{y}_i$ is the reference distribution reflects the HMM-based ASR. More specifically,

$$\begin{aligned} S_{KL}(\mathbf{y}_i,\mathbf{z}_t) &= \sum_{d=1}^{D} y_i^d \log\Big(\frac{y_i^d}{z_t^d}\Big) \\ &= \sum_{d=1}^{D} y_i^d \log y_i^d - \sum_{d=1}^{D} y_i^d \log z_t^d \end{aligned} \tag{3.28}$$

The first part of Eqn (3.28), the entropy of probability distribution $\mathbf{y}_i$, takes into account the uncertainty in the lexical model, and the second part or the cross entropy compares the acoustic model against the lexical model. It is trivial to see the point made above about the purity of lexical knowledge by turning $\mathbf{y}_i$ into a deterministic lexical model i.e., Kronecker delta distribution. In such a case, the hybrid HMM/ANN approach [Bourlard and Morgan, 1994] where the acoustic model estimates $P(q_t = a^d|\mathbf{x}_t,\theta_a)$ rather than $p_{sl}(\mathbf{x}_t|q_t = a^d,\theta_a)$ can be seen as a special case of KL-HMM approach.

The KL-HMM approach, however, is capable of reversing the importance given to the acoustic

model and the lexical model by changing the local score to $S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$, i.e.,

$$\begin{aligned} S_{RKL}(\mathbf{y}_i, \mathbf{z}_t) &= \sum_{d=1}^{D} z_t^d \log\Big(\frac{z_t^d}{y_i^d}\Big) \\ &= \sum_{d=1}^{D} z_t^d \log z_t^d - \sum_{d=1}^{D} z_t^d \log y_i^d \end{aligned} \tag{3.29}$$

It can be observed from Eqn (3.29) that the first quantity, the entropy of probability distribution $\mathbf{z}_t$, is independent of lexical unit and the matching only takes place with the second quantity, i.e., the cross entropy between distributions $\mathbf{z}_t$ and $\mathbf{y}_i$, with $\mathbf{z}_t$ as the reference. The local score $S_{SKL}(\mathbf{y}_i, \mathbf{z}_t)$ is the case where equal importance is given to the acoustic model and the lexical model.

Yet another distinction between the KL-HMM and Tied-HMM/SP-HMM approaches is that, in the KL-HMM approach the local score is discriminative [Blahut, 1974], i.e., the acoustic model and lexical model evidence is matched discriminatively, irrespective of the type of local score used, i.e., $S_{KL}(\mathbf{y}_i, \mathbf{z}_t)$ or $S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$ or $S_{SKL}(\mathbf{y}_i, \mathbf{z}_t)$.

3.6 Summary

In this chapter, we introduced the framework of probabilistic lexical modeling. We elucidated that standard HMM-based ASR systems implicitly model the dependency between acoustic feature observations and lexical units through latent variables. The latent variables, also referred to as acoustic units, are either context-independent subword units or clustered context-dependent subword units. Furthermore, in standard HMM-based ASR systems each lexical unit is deterministically mapped to an acoustic unit i.e., the relationship is deterministic.

We presented three techniques in which a probabilistic relationship between lexical and acoustic units is learned. The KL-HMM approach was originally proposed as an acoustic modeling approach alternative to the Tandem approach. However, as shown in this chapter it is a probabilistic lexical modeling based ASR approach. The Tied-HMM approach was proposed in the framework of hybrid HMM/ANN systems to build context-dependent subword unit based systems using an acoustic model that classifies context-independent subword units. We showed that it is a probabilistic lexical modeling based ASR approach and equivalent to the PC-HMM approach in principle. We introduced another probabilistic lexical modeling approach referred to as SP-HMM, that is similar to the KL-HMM and Tied-HMM approaches. We contrasted the three probabilistic lexical modeling approaches in terms of the importance given to the acoustic model and the lexical model during parameter estimation and decoding.

4 Proposed Grapheme-based ASR Approach

This chapter motivates and proposes a novel grapheme-based ASR approach in the framework of probabilistic lexical modeling which forms the core of this thesis. In Section 4.1, we present the implications of the deterministic lexical model, the constraints arising from it and provide a survey on how these constraints are being addressed in the literature. In Section 4.2, we present the potential implications of probabilistic lexical modeling which relaxes the constraints imposed by deterministic lexical modeling and provide a literature survey to motivate the goals of this thesis.

The proposed grapheme-based ASR approach in the framework of probabilistic lexical modeling is presented in Section 4.3. With the help of a pilot study on the DARPA Resource Management (RM) corpus, we present a detailed analysis of the lexical model parameters learned using different approaches (KL-HMM, Tied-HMM and SP-HMM). The analysis elucidates that the lexical model parameters capture a probabilistic relationship between graphemes and phones. Furthermore, the influence of the cost function and the subword context on the G2P relationship captured by the lexical model parameters is studied.

4.1 Implications of Deterministic Lexical Model Systems

As described in Section 3.2, in standard HMM-based ASR systems the lexical model i.e., the relationship between lexical units $l^i \in \mathscr{L}$ and acoustic units $a^d \in \mathscr{A}$ is deterministic and the pronunciation lexicon (θ_{pr}) determines both the lexical unit set $\mathscr{L}$ and the acoustic unit set $\mathscr{A}$. As a consequence,

- if $\mathscr{L}$ is based on phone subword units (phone-based ASR system) or grapheme subword units (grapheme-based ASR system) then $\mathscr{A}$ is also based on phones or graphemes, respectively;
- if $\mathscr{L}$ is based on context-independent subword units (context-independent subword unit based ASR system) or context-dependent subword units (context-dependent subword unit based ASR system) then $\mathscr{A}$ is also based on context-independent subword units or context-dependent subword units, respectively.

The performance of deterministic lexical model based systems is strongly dependent on the accuracy of the deterministic mapping which in turn is determined by factors such as availability of acoustic data, availability of a well developed phone lexicon and portability of available resources. More specifically, deterministic lexical modeling imposes the following three constraints:

1. The first constraint is the availability of sufficient and well developed acoustic data in the target language or domain to train effectively both an acoustic model and a lexical model. Unfortunately, many languages do not have such well developed acoustic resources [Besacier et al., 2014].
2. The second constraint that arises as a result of deterministic lexical modeling is the availability of a well developed phonetic lexicon as most of the ASR systems use phones as lexical units. Again, many languages lack such well developed lexical resources [Davel and Martirosian, 2009, Besacier et al., 2014].
3. The third constraint that the deterministic lexical model introduces is that ASR system trained with one phone set can not be directly ported to or used as it is for a new domain which has a lexicon based on a different phone set [Imseng et al., 2013a]. For a language, it can happen that there are different phonetic lexicons based on different phone sets. For instance, in English there are phonetic lexicons based on ARPABET, CMUBET, SAMPA etc.

In the following subsections, we provide a survey of the literature on how the resource constraints are being addressed in the framework of deterministic lexical modeling.

4.1.1 Lack of Acoustic Resources

In the literature, the lack of acoustic resources has been typically addressed through the use of multilingual or crosslingual acoustic and lexical resources coupled with acoustic model adaptation techniques [Köhler, 1998, Beyerlein et al., 2000b, Schultz and Waibel, 2001b, Le and Besacier, 2009, Burget et al., 2010]. The first step in this process is the definition of a common or universal phone set across all out-of-domain languages and target language. This step ensures that the phone sets match across languages, thus addressing the third constraint mentioned earlier. The common or universal phone set can be defined either in a knowledge-based manner [Köhler, 1998, Beyerlein et al., 2000b, Schultz and Waibel, 2001b, Le and Besacier, 2009] or in a data-driven manner [Sim and Li, 2008, Sim, 2009]. Multilingual acoustic models (GMMs or ANNs) are first trained on language-independent data and then adapted on target language data.

In the framework of HMM/GMM systems, the parameters of multilingual acoustic model are adapted on target language data using techniques such as, bootstrapping, maximum a posteriori adaptation (MAP), maximum likelihood linear regression (MLLR) and subspace Gaussian mixture models (SGMM). The out-of-domain lexical model (decision trees) is either retained [Köhler, 1998, Beyerlein et al., 2000b, Le and Besacier, 2009] or redefined using target language data [Schultz and Waibel, 2001b, Burget et al., 2010].

- In the bootstrapping approach [Osterholtz et al., 1992], multilingual acoustic models are used as seed models for target language [Schultz and Waibel, 2001b]. The seed models are used only for initialization. After initialization, the acoustic and lexical model are trained on target language data.
- MAP [Gauvain and Lee, 1994] and MLLR [Leggetter and Woodland, 1995] were initially proposed for speaker adaptation and later shown to be useful for language adaptation [Beyerlein et al., 2000a]. In MLLR, the means and variances of Gaussians trained on language-independent data are adapted through linear transformations. The transformation matrix is usually tied over a number of Gaussians. Thus, only a relatively small amount of adaptation data is required. In MAP adaptation, the parameters are set using the prior distribution and the adaptation is performed at phone level. On one hand, adaptation is phone specific in MAP while on the other hand, adaptation can only be performed for acoustic models of phones that are observed. As a consequence, MAP needs more adaptation data than MLLR as we will see later in Chapter 6.
- In [Schultz and Waibel, 2001b], in addition to acoustic model adaptation, polyphone decision tree specialization (PDTS) was used to perform lexical model (decision tree) adaptation. In PDTS, the multilingual decision tree is adapted to the target language by restarting the decision tree growing process with a limited amount of adaptation data. It was found that PDTS based porting was beneficial compared to adaptation without using PDTS technique.
- In the SGMM approach [Povey et al., 2011], the state distributions are modeled as mixture of Gaussians whose parameters are constrained by a shared set of subspaces. The parameters of the SGMM system are divided into global parameters and state specific parameters. In the case of low acoustic resources [Burget et al., 2010], the global parameters can be trained on multilingual acoustic and lexical resources, while the state specific parameters (including decision trees) are trained on target language data.

In the framework of hybrid HMM/ANN, multilingual and crosslingual acoustic modeling has focussed mainly on Tandem approaches and neural network adaptation based methods [Stolcke et al., 2006, Thomas et al., 2012, Swietojanski et al., 2012].

- In the case of Tandem approaches, output of the multilingual neural network is transformed to be used as feature observations to train an acoustic model (GMMs) and a lexical model using target language data [Stolcke et al., 2006, Thomas et al., 2012]. However, as the GMMs are trained on target language data, minimal acoustic and lexical resources from the target language are necessary to robustly estimate the parameters.
- In other approaches, the multilingual neural network is adapted/retrained on target language using phoneset mapping [Thomas and Hermansky, 2010, Swietojanski et al., 2012].

4.1.2 Lack of Lexical Resources

As mentioned in Section 2.3, the use of hand a labeled phone pronunciation lexicon is optimal for ASR systems where the relationship between phone subword units and acoustic features is modeled directly. Most often, the existing hand labeled phone pronunciation lexicon (or seed lexicon) may not have complete coverage for a new domain (target domain) for which

we interested to build an ASR system. Therefore, typically given an initial lexicon, a G2P converter [Bisani and Ney, 2008, Novak, 2011] is first trained to extract pronunciations for new words. The augmented lexicon is then used to build an ASR system. G2P convertors are also used to augment the recognition vocabulary.

In [Gollan et al., 2005, Lööf et al., 2006], systems trained to transcribe parliamentary speeches used a G2P convertor to produce pronunciations for words that are not present in the cross-domain lexicon. In [Jouvet et al., 2012], the effect of using a G2P convertor based lexicon on ASR performance was studied in a more controlled scenario. The BDLex French pronunciation lexicon was used to generate pronunciations for ESTER2 French broadcast news transcription. It was observed that the performance of the ASR system using a G2P convertor based lexicon is worse than the ASR system using a baseline phone lexicon developed for the ESTER2 corpus. Furthermore, it was also reported that the gap in ASR performance between systems using a well developed lexicon and a G2P lexicon is bridged only when G2P pronunciations of most frequent words was manually verified by an expert.

Some languages may not even have the seed lexicon required to train a G2P converter. Therefore, alternate subword units like graphemes, which makes lexicon development easy, have been explored in the literature [Schukat-Talamazzini et al., 1993, Kanthak and Ney, 2002, Killer et al., 2003, Dines and Magimai-Doss, 2007, Ko and Mak, 2014]. The success of a grapheme-based ASR system primarily depends on the G2P relationship of the language. The reason for this is as follows. It can be seen in Eqn (3.3) that the acoustic model score $p(\mathbf{x}_t|a^d,\theta_a)$ models the relationship between the acoustic feature observation $\mathbf{x}_t$ and the acoustic unit a^d. Due to the deterministic lexical modeling in standard HMM-based ASR systems, both the acoustic unit a^d and the lexical unit l^i are the same and represent graphemes. However, the acoustic feature observations or the cepstral features depict the envelop of the short-term spectrum which is more related to phones.

Context-Dependent Grapheme Modeling

In the literature, to overcome the problem of the irregular G2P relationship, modeling of context-dependent graphemes has been explored [Kanthak and Ney, 2002, Killer et al., 2003, Mimer et al., 2004]. The implicit assumption here being that the relationship between context-independent graphemes and context-independent phones can be irregular, but relationship between context-dependent graphemes and context-independent phones could be regular. The same idea is exploited in G2P conversion systems.

As the context of grapheme subword units increases, the number of subword models to be trained also increases and can lead to data sparsity problems. Therefore, the grapheme-based ASR literature has focused on efficient and automatic state tying methods for grapheme subword units.

- In [Schukat-Talamazzini et al., 1993], the context-dependent grapheme-based ASR system was used for a train scheduling task. They employed backoff to handle unseen context-

dependent grapheme subword units.

- In [Kanthak and Ney, 2002], manual and automatic methods for decision tree based state tying of context-dependent grapheme subword units were compared. In manual decision tree based state tying, a grapheme was assigned to a phonetic question if grapheme is part of the phoneme, while in the case of automatic method, the questions were generated based on bottom-up clustering of context-independent grapheme HMM states using likelihood gain and observation count as merging criteria. It was found that the manually generated question set yielded better system.
- In [Killer et al., 2003, Killer, 2003], various question sets for decision tree based state tying were investigated. It was found that the singleton question set often yielded a better system compared to manually derived, bottom-up entropy, and entropy distance based question sets.
- In [Mimer et al., 2004, Stüker and Schultz, 2004], it has been shown that enhanced tree clustering where a single decision tree is constructed for all the sub-states of all graphemes improves grapheme-based ASR system performance. Through enhanced tree clustering it is possible to share the parameters across context-dependent graphemes with different central graphemes.
- More recently, in [Ko and Mak, 2014], the eigen trigrapheme approach was proposed to robustly estimate the parameters of context-dependent graphemes with few training samples. In the eigen trigrapheme approach, first eigenbases are derived over a set of clustered context-dependent graphemes. Later, each context-dependent grapheme is modeled as a distinct point in the space spanned by these basis vectors.

Context-dependent modeling of graphemes has been applied for a wide range of languages like English [Kanthak and Ney, 2002, Killer et al., 2003, Dines and Magimai-Doss, 2007, Ko and Mak, 2014], German [Kanthak and Ney, 2002, Killer et al., 2003], Dutch [Kanthak and Ney, 2002], Italian [Kanthak and Ney, 2002], Russian [Stüker and Schultz, 2004], Spanish [Killer et al., 2003], Vietnamese [Le and Besacier, 2009], Arabic [Biadsy et al., 2012], African languages [Schlippe et al., 2012, Ko and Mak, 2014], etc. These studies have shown that the use of grapheme as subword units has mainly succeeded for languages such as Spanish and Finnish for which the G2P relationship is regular. For languages such as English that have an irregular G2P relationship, it has been found that grapheme-based ASR systems perform worse compared to phone-based ASR systems. In [Sung et al., 2009], it was found that even with large training data (more than 1000 hours) grapheme subword units resulted in a poor system compared to phone subword units for English. In [Killer et al., 2003], it was also found that modeling context longer than the single preceding grapheme and single following grapheme did not always yield improved ASR performance.

Combining Grapheme and Phoneme Information

In addition to modeling only grapheme subword units, there have been studies where the ASR system uses both phone and grapheme subword units [Magimai-Doss et al., 2003, Magimai.-Doss et al., 2004, Dines and Magimai-Doss, 2007, Schlippe et al., 2012]. The aim of these

approaches was to improve the performance of ASR systems using graphemes in addition to using phones rather than addressing the lack of lexical resources.

In [Magimai-Doss et al., 2003, Magimai.-Doss et al., 2004], joint modeling of phone and grapheme information was investigated, where during training, grapheme and phone subword units are jointly modeled, and decoding is performed using either one subword unit or both. Experimental studies conducted on isolated word recognition tasks and small vocabulary speech recognition tasks in the framework of hybrid HMM/ANN ASR system showed that joint modeling of grapheme and phone information could be beneficial.

In [Dines and Magimai-Doss, 2007], in addition to context-dependent subword unit modeling, the use of Tandem features was investigated. The Tandem features can be seen as a data-driven projection of standard acoustic features along linguistic dimensions and thus could be expected to help in modeling grapheme subword units better when compared to standard cepstral features. ASR studies on English showed that (stand alone) Tandem features could help in bridging the performance gap between phone-based and grapheme-based ASR systems. Furthermore, combination of grapheme and phone information at lexical level was investigated. In this case, models for grapheme subword units and phone subword units were trained separately and then used/pooled together during decoding. The combined system was found to be beneficial.

In [Schlippe et al., 2012], combination of grapheme and phone information at the ASR hypothesis level was investigated. More precisely, the test speech was first decoded by both grapheme-based and phone-based ASR systems. The hypotheses resulting from the two systems were then combined using the confusion network combination technique. Experimental studies conducted on Hausa, an under-resourced language, showed this approach to be promising.

4.1.3 Lack of Acoustic and Lexical Resources

When the language lacks both acoustic and phone lexical resources, multilingual and crosslingual grapheme-based approaches that can leverage from resources available in other languages have been explored [Kanthak and Ney, 2003, Stüker, 2008a,b]. Similar to multilingual phone subword modeling, multilingual grapheme subword modeling is based on the *universal* or *multilingual* grapheme set formed by merging graphemes that are common across different languages. However, unlike multilingual phone sets, it is not trivial to port multilingual grapheme sets to new languages mainly because of two reasons. Firstly, grapheme sets of all languages may not match or overlap. To overcome this issues, either transliteration or data-driven mapping has been employed [Stüker, 2008b]. Secondly, sharing of acoustic models of graphemic subword units across languages is not evident, as the relationship between graphemes and phones may differ considerably across languages. Investigations until now have shown that multilingual grapheme-based ASR systems generally performed worse compared to monolingual grapheme-based ASR systems.

4.2 Potential of Probabilistic Lexical Modeling

In the case of probabilistic lexical modeling, each lexical unit l^i is related to all acoustic units $\{a^d\}_{d=1}^{D}$ in a probabilistic manner. As a consequence,

1. the parameters of the acoustic model θ_a and the lexical model θ_l can be trained on an independent set of resources. In this light, previous works on KL-HMM such as [Imseng et al., 2011, 2012b, Rasipuram et al., 2013a] suggest that ASR systems can be rapidly developed using domain-independent or language-independent acoustic model and by training only the lexical model on target language or domain data;
2. $\mathscr{L}$ and $\mathscr{A}$ can model different contextual units. For instance, as in previous works [Rottland and Rigoll, 2000, Magimai.-Doss et al., 2011, Imseng et al., 2011, 2012b, Rasipuram et al., 2013a, Razavi et al., 2014], $\mathscr{L}$ can be based on context-dependent subword units while $\mathscr{A}$ can be based on context-independent subword units. These ASR systems have been found to yield performance comparable to or better than standard context-dependent subword unit based HMM/GMM systems;
3. it is not necessary that the subword unit set used for defining acoustic units should be the same as the subword unit set used for defining lexical units. The lexical model can capture the relationship between the distinct subword unit sets through acoustics.

In the following section, we present a grapheme-based ASR approach that can exploit all the above advantages of probabilistic lexical modeling to address both acoustic and lexical resource constraints in ASR system development.

4.3 Proposed Grapheme-based ASR Approach

In the framework of probabilistic lexical modeling, the modeling of the relationship between graphemes and acoustic features can be factored into two parts through acoustic units:

1. *The acoustic model* where the relationship between acoustic units and acoustic features is modeled.
2. *The lexical model* where a probabilistic relationship between acoustic units and graphemes is modeled.

In this thesis, we show that:

- Acoustic units can be phones, multilingual phones or clustered context-dependent subword units.
- An acoustic model can be trained on either domain-independent or language-independent resources.
- Lexical units are the graphemes of the target domain or language and it is sufficient to train only the lexical model on the target domain or language data.

In the proposed approach with graphemes as lexical units and phones as acoustic units, the lexical model parameters capture a probabilistic relationship between graphemes and phones using acoustic data. Thus, the proposed grapheme-based ASR approach integrates lexicon

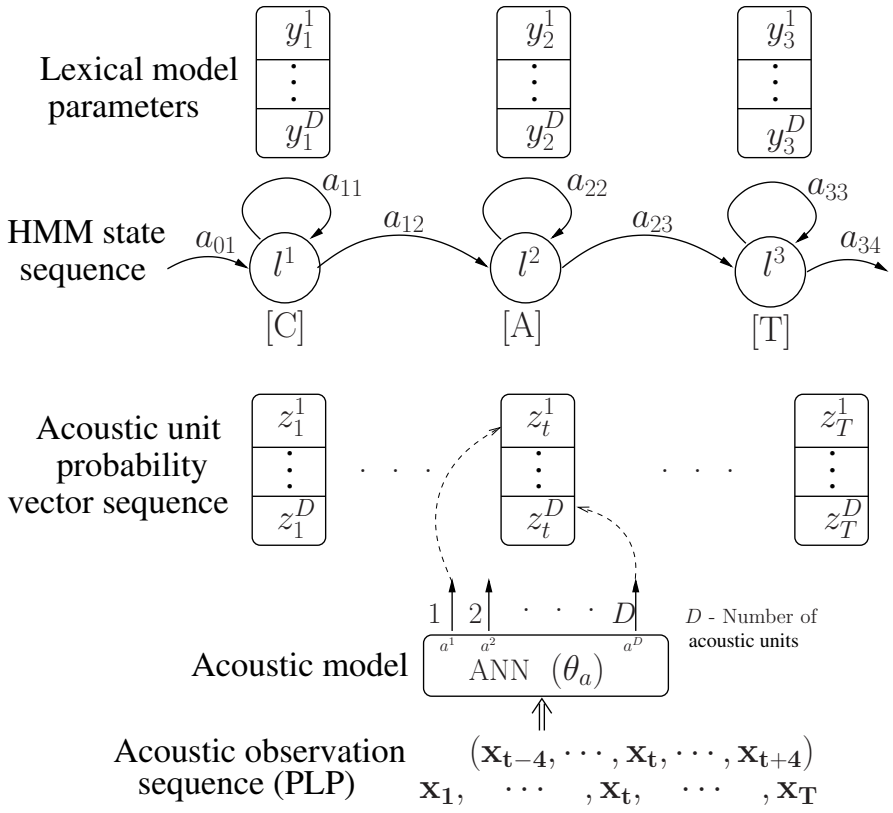

Figure 4.1 – Illustration of the proposed grapheme-based ASR approach using KL-HMM

learning as a phase in ASR system training and could potentially remove the necessity of training an explicit G2P converter.

Figure 4.1 illustrates the proposed grapheme-based ASR approach. The acoustic model (or an ANN) is trained to classify acoustic units. Given the acoustic model, acoustic unit probability sequences of training data are estimated. The acoustic unit probability sequences are used as feature observations to train an HMM with graphemes as lexical units using the KL-HMM approach. In this illustration, the HMM for the word "CAT" is composed of lexical units representing context-independent grapheme subword units "[C]", "[A]" and "[T]". However, the lexical units normally represent context-dependent subword units with a three-state minimum duration constraint.

4.4 Pilot Study on the RM Corpus

In this section, we present a pilot study on the RM corpus where both acoustic model parameters θ_a and lexical model parameters θ_l are trained using acoustic and lexical resources of target domain data. The main motivation of this pilot study is to validate our hypothesis that lexical model parameters indeed capture a probabilistic G2P relationship. Towards that we first

Table 4.1 – Overview of different systems. CI denotes context-independent subword units, CD denotes context-dependent subword units and cCD denotes clustered context-dependent subword units. P and G denote phone lexicon and grapheme lexicon, respectively. *Det* denotes lexical model is deterministic and *Prob* denotes lexical model is probabilistic.

System	Acoustic units $\mathscr{A}$	Lexicon	Lexical units $\mathscr{L}$	Approach
KL-HMM	CI	P or G	CI or CD	*Prob*
SP-HMM	CI	P or G	CI or CD	*Prob*
Tied-HMM	CI	P or G	CI or CD	*Prob*
HMM/GMM	cCD	P or G	CD	*Det*
Tandem	cCD	P or G	CD	*Det*

analyze the effect of the cost function and the subword context modeled on the captured G2P relationship. We then compare the proposed grapheme-based ASR system with the following systems:

- The phone-based ASR system using a well developed phone lexicon.
- The grapheme-based HMM/GMM system proposed in the literature [Kanthak and Ney, 2002, Killer et al., 2003].
- The phone-based Tandem system [Hermansky et al., 2000] and the grapheme-based Tandem system [Dines and Magimai-Doss, 2007].

4.4.1 Experimental Setup

In the pilot study, we followed the setup reported in [Dines and Magimai-Doss, 2007] for the RM corpus. In this setup a training set of 2,880 utterances was used to train the ASR systems as opposed to 3990 utterances (formed by combining the training and development sets) usually reported in the literature [Hain and Woodland, 1999]. The development set of 1110 utterances was used only to tune the word insertion penalty. However, the standard test set containing 1,200 utterances was used. We also used the MLP from the same study [Dines and Magimai-Doss, 2007]) trained on the RM corpus to classify 45 context-independent phones as an acoustic model. This was primarily done to have ASR systems that are comparable with ASR systems in [Dines and Magimai-Doss, 2007]. More details about the RM corpus such as subword units and lexicon are given in Appendix A.1. In this study we used grapheme lexicon transcribed with 29 graphemes.

As mentioned earlier, we compare the probabilistic lexical modeling systems, i.e., KL-HMM, SP-HMM and Tied-HMM with the deterministic lexical modeling systems i.e., HMM/GMM and Tandem systems. Three KL-HMM systems, i.e., KL-HMM KL, KL-HMM RKL and KL-HMM SKL that are based on three KL-divergence based local scores S_{KL}, S_{RKL} and S_{SKL} are built respectively. Table 4.1 summarizes the different systems and their capabilities.

As given in Section 3.2, in the HMM/GMM approach, the lexical model is deterministic and

thus both acoustic and lexical units should be of the same type and should model the same subword context. In context-independent HMM/GMM and Tandem systems both acoustic and lexical units are context-independent subword units. As given in Table 4.1, in context-dependent HMM/GMM and Tandem systems, acoustic units are clustered context-dependent subword units, lexical units are context-dependent subword units, and lexical and acoustic units are deterministically related. Standard context-independent and word internal context-dependent HMM/GMM systems are trained using either PLP features or Tandem features. Phone-based HMM/GMM and Tandem systems use a phonetic question set for state tying, whereas grapheme-based HMM/GMM and Tandem systems use a singleton question set. The phonetic question set is composed of phonetic attribute values such as vowel, consonant, and plosive. In the case of graphemes, the singleton question set is based only on the identity of the graphemes in the preceding and following context.

The training phase of KL-HMM, SP-HMM and Tied-HMM systems involves the estimation of lexical model parameters θ_l. We consider three different lexical unit sets,

- *mono* or context-independent subword units;
- *tri* or context-dependent subword units with single preceding and following context;
- *quint* or context-dependent subword units with two preceding and following contexts.

For context-dependent studies, we train word internal context models. Each subword unit is modeled by a three-state left-to-right HMM.

The total number of parameters in different systems (KL-HMM, HMM/GMM and Tandem) are compared in Tables 4.2 and 4.3 for context-independent and context-dependent subword units, respectively.

- The three KL-HMM systems modeling different lexical units (*mono*, *tri* and *quint*) use the same acoustic model. Therefore, the acoustic model complexity of these three systems is the same and the total number of acoustic model parameters is equal to the number of MLP parameters (≈0.5M).
- The total number of lexical model parameters for KL-HMM systems is given by the number of lexical units times the number of acoustic units ($I \times D$). The number of acoustic units is the same as the number of MLP outputs i.e., 45. The lexical model parameters in this study are not tied. Therefore, the number of lexical units is obtained from the lexicon, the context length and the minimum duration constraint. For example, in the case of *cd* phones, there are 2269 context-dependent phones in the RM lexicon, therefore $I = 2269 \times 3$.
  - For *mono* phones, $I = 45 \times 3$, $D = 45$ and the total number of lexical model parameters = $(45 \times 3) \times 45$.
  - For *cd* phones, $I = 2269 \times 3$, $D = 45$ and the total number of lexical model parameters = $(2269 \times 3) \times 45 \approx 0.3M$.
  - For *quint* phones, $I = 3942 \times 3$, $D = 45$ and the total number of lexical model parameters = $(3942 \times 3) \times 45 \approx 0.5M$.
  - For *mono*, *tri* and *quint* graphemes, $I = 29 \times 3$, $I = 1912 \times 3$ and $I = 4112 \times 3$, respectively. The lexical model parameters of grapheme-based ASR systems can be calculated similarly to phone-based systems.

- The total number of acoustic model parameters of the HMM/GMM system are computed based on the number of acoustic units and the parameters of each acoustic unit. In this chapter, we used eight mixture Gaussians with diagonal covariance to model each acoustic unit. For example,
  - for *mono* phones, $D = 45 \times 3$ and the total number of acoustic model parameters = $(45 \times 3) \times (8 \times (39 + 39 + 1)) \approx 0.1M$;
  - for *cd* phones, acoustic units are clustered context-dependent subwords $D = 1477$ and the total number of acoustic model parameters = $(1477) \times (8 \times (39 + 39 + 1)) \approx 1.0M$; and
  - for *mono* and *cd* graphemes, the number of acoustic units $D = 29 \times 3$ and $D = 1369$, respectively.
- The total number of acoustic model parameters of Tandem system include the parameters of ANN and the parameters of each of the acoustic units. For example,
  - for *mono* phones, the total number of acoustic model parameters = $0.5M + (45 \times 3) \times (8 \times (45 + 45 + 1)) \approx 0.6M$;
  - for *cd* phones, the total number of acoustic model parameters = $0.5M + (2013) \times (8 \times (45 + 45 + 1)) \approx 1.9M$; and
  - for *mono* and *cd* graphemes, the number of acoustic units $D = 29 \times 3$ and $D = 1985$, respectively.
- The lexical model in the case of HMM/GMM and Tandem systems is deterministic and the lexical model parameter set θ_l is a table which maps each lexical unit to a clustered state. Therefore, the total number of lexical model parameters is equal to the number of lexical units. For example,
  - for *mono* phones, the total number of lexical model parameters = 45×3;
  - for *cd* phones, the total number of lexical model parameters = 2269×3;
  - for *mono* graphemes, the total number of lexical model parameters = 29×3; and
  - for *cd* graphemes, the total number of lexical model parameters = 1912×3.

The SP-HMM and Tied-HMM systems have the same number of parameters as the KL-HMM system. The lexical model complexity of KL-HMM systems increases with context and the acoustic model complexity remains the same. Furthermore, for *tri* lexical units, it can be observed that the KL-HMM system has fewer acoustic model parameters and more lexical model parameters compared to the HMM/GMM system.

Table 4.2 – Number of parameters for systems modeling *mono* lexical units. θ_a denotes acoustic model parameters, θ_l denotes lexical model parameters.

System	grapheme			phone		
	θ_a	θ_l	Total	θ_a	θ_l	Total
KL-HMM	0.5M	4K	≈0.5M	0.5M	6K	≈0.5M
Tandem	0.5M+0.06M	87	≈0.56M	0.5M+0.1M	135	≈0.6M
HMM/GMM	0.05M	87	≈0.05M	0.1M	135	≈0.1M

Table 4.3 – Number of parameters for systems modeling context-dependent lexical units. θ_a denotes acoustic model parameters, θ_l denotes lexical model parameters.

System (context)	grapheme			phone		
	θ_a	θ_l	Total	θ_a	θ_l	Total
KL-HMM (*tri*)	0.5M	0.2M	≈0.7M	0.5M	0.3M	≈0.8M
Tandem (*tri*)	0.5M+1.4M	5.7K	≈1.9M	0.5M+1.4M	6.6K	≈1.9M
HMM/GMM (*tri*)	0.9M	5.7K	≈0.9M	0.9M	6.6K	≈0.9M
KL-HMM (*quint*)	0.5M	0.5M	≈1.0M	0.5M	0.5M	≈1.0M

4.4.2 Analysis of the Lexical Model Parameters

In this section, we analyze the effect of the cost function and the grapheme context on the G2P relationship being learned by the lexical model parameters. Part of the analysis, especially the effect of the grapheme context on the lexical model parameters has appeared in [Magimai.-Doss et al., 2011].

Effect of the Cost Function and the Local Score

In Section 3.4, we hypothesized that the local scores S_{KL} and S_{RKL} can model better one-to-one and one-to-many lexical-to-acoustic unit relationships, respectively whereas the local score S_{SKL} can model both one-to-one and one-to-many lexical-to-acoustic unit relationships. In order to analyze and visualize the effect of the cost function and the local score,

1. the lexical model parameters are trained using the KL-HMM (*KL*, *RKL*, *SKL*), SP-HMM and Tied-HMM approaches with context-independent graphemes as lexical units and context-independent phones as acoustic units. Each lexical unit is modeled by a single state HMM;
2. the lexical model parameters (or the categorical distribution) of each grapheme lexical unit are sorted according to their probability value and the coordinates with probability value greater than or equal to 0.1 are picked.

The G2P relationship captured by the lexical model parameters of different grapheme-based KL-HMM systems is presented in Table 4.4. From the table the effects of the local scores on the captured G2P relationship can be observed to be:

1. The lexical model parameters of KL-HMM *KL* system capture one-to-one G2P relationships (e.g., see [B], [L], [M], [P]) better than one-to-many G2P relationships (e.g., see vowel graphemes, [C], [H], [X]).
2. The lexical model parameters of KL-HMM *RKL*, in addition to relevant one-to-many G2P relationships (e.g., see vowel graphemes, [C], [G], [H]), also capture additional confusable and spurious relations. This can be particularly seen in the case of one-to-one G2P correspondence (e.g., see [B], [M]).

Table 4.4 – G2P relationship captured by the lexical model parameters of the KL-HMM *KL*, KL-HMM *RKL* and KL-HMM *SKL* approaches

Grapheme	phones		
	S_{KL}	S_{RKL}	S_{SKL}
A	ae(0.7) eh(0.2) ey(0.1)	ae(0.3) ey(0.3) eh(0.1) ax(0.1)	ae(0.5) eh(0.2) ey(0.1) ax(0.1)
B	b(1.0)	b(0.5) ah(0.2)	b(0.9)
C	k(1.0)	k(0.5) ch(0.2) s(0.1) t(0.1)	k(0.6) t(0.2) ch(0.1) s(0.1)
D	d(0.9) t(0.1)	d(0.5) t(0.2) sil(0.1)	d(0.7) t(0.1)
E	ax(0.4) ih(0.3) eh(0.1) iy(0.1)	iy(0.3) eh(0.1) ax(0.1) ih(0.1) ey(0.1)	iy(0.3) ax(0.2) ih(0.2) eh(0.1) ey(0.1)
F	f(1.0)	f(0.7) v(0.1) sil(0.1)	f(0.9)
G	g(0.9)	g(0.4) jh(0.2) sil(0.1) k(0.1) d(0.1)	g(0.7) d(0.1) k(0.1)
H	t(0.7) d(0.1) sil(0.1)	sh(0.3) dh(0.2) hh(0.1) th(0.1)	dh(0.2) sil(0.2) t(0.2) th(0.1) d(0.1) hh(0.1)
I	ih(0.8) ax(0.1)	ih(0.4) ay(0.2) ax(0.1) iy(0.1)	ih(0.5) ax(0.2) eh(0.1) ay(0.1)
J	jh(1.0)	jh(0.7) ch(0.1) d(0.1) t(0.1)	jh(0.9)
K	k(1.0)	k(0.7) sil(0.1) t(0.1)	k(0.9)
L	l(1.0)	l(0.5) el(0.1) ao(0.1) ow(0.1)	l(0.8)
M	m(1.0)	m(0.7) n(0.1)	m(0.9) n(0.1)
N	n(0.9)	n(0.5) en(0.1) ng(0.1)	n(0.8) en(0.1)
O	ao(0.4) aa(0.3) ow(0.1) ah(0.1)	ao(0.2) aa(0.2) ow(0.2) sh(0.1) ah(0.1) ax(0.1)	ao(0.2) aa(0.2) ow(0.2) ah(0.1) ax(0.1)
P	p(1.0)	p(0.8)	p(0.9)
Q	k(1.0)	k(0.5) w(0.2) uw(0.1) y(0.1)	k(0.9)
R	r(0.8) axr(0.2)	r(0.4) axr(0.3) aa(0.1) er(0.1)	r(0.6) axr(0.3) er(0.1)
S	s(0.9) z(0.1)	s(0.6) z(0.2)	s(0.8) z(0.2)
T	t(0.9)	t(0.5) sil(0.1) d(0.1) k(0.1)	t(0.8)
U	ax(0.4) uw(0.3) ih(0.1)	uw(0.3) y(0.2) ax(0.1) ah(0.1)	uw(0.3) ax(0.3) ih(0.1) ah(0.1)
V	ay(0.9)	v(0.5) ay(0.3)	v(0.9)
W	w(1.0)	w(0.6) aw(0.1) uw(0.1)	w(0.9)
X	k(0.9) t(0.1)	s(0.4) k(0.4)	k(0.5) s(0.3) t(0.1)
Y	iy(0.8) ey(0.1)	iy(0.4) ay(0.1) ey(0.1) oy(0.1)	iy(0.5) ey(0.3) ih(0.1)
Z	z(0.9)	ay(0.4) z(0.3) s(0.1)	z(0.8) s(0.1)
sil	sil(1.0)	sil(1.0)	sil(1.0)

3. The lexical model parameters of KL-HMM *SKL* tend to capture one-to-one G2P relationship similarly to the parameters of KL-HMM *KL*. They are also able to capture one-to-many G2P relationships better than local score S_{KL} but not to the same extent as local score S_{RKL} (e.g., see [G], [H], [N]).

The G2P relationship captured by the lexical model parameters of the Tied-HMM and SP-HMM approaches is given in Table 4.5. From the table it can be observed that the captured G2P relationship is similar to or better than the KL-HMM SKL and KL-HMM RKL approaches. For example, the G2P mapping for graphemes [A], [E] and [I] is better with the Tied-HMM and SP-HMM approaches than KL-HMM RKL or KL-HMM SKL approaches. There is no significant difference in the G2P relationship captured by the lexical model parameters of the Tied-HMM and SP-HMM approaches.

Table 4.5 – G2P relationship captured by the lexical model parameters of the Tied-HMM and SP-HMM approaches

Grapheme	phones	
	S_{tied}	S_{SP}
A	ae(0.5) ey(0.4) ax(0.1)	ae(0.5) ey(0.3) ax(0.1)
B	b(0.7) ah(0.3)	b(0.7) ah(0.3)
C	k(0.6) ch(0.2) s(0.2)	k(0.6) ch(0.2) s(0.2)
D	d(0.9) sil(0.1)	d(0.8) sil(0.1) t(0.1)
E	iy(0.4) eh(0.3) ax(0.2)	iy(0.4) eh(0.2) ax(0.2)
F	f(0.8) v(0.2)	f(0.8) v(0.1) sil(0.1)
G	g(0.4) ng(0.3) jh(0.2) ey(0.1)	g(0.5) jh(0.2) ey(0.1) ng(0.1) sil(0.1)
H	sh(0.3) dh(0.3) hh(0.2) th(0.2)	sh(0.3) dh(0.3) hh(0.1) sil(0.1) th(0.1)
I	ih(0.7) ay(0.2)	ih(0.7) ay(0.2) iy(0.1)
J	jh(1.0)	jh(1.0)
K	k(0.9) sil(0.1)	k(0.9) sil(0.1)
L	l(0.8) el(0.2)	l(0.9) el(0.1)
M	m(0.9) eh(0.1)	m(0.9) eh(0.1)
N	n(0.8) en(0.2)	n(0.9) en(0.1)
O	ow(0.2) ao(0.2) aa(0.2) sh(0.1) ah(0.1)	aa(0.3) ow(0.2) ao(0.2) sh(0.1) ax(0.1)
P	p(1.0)	p(1.0)
Q	k(0.6) w(0.2) y(0.1) uw(0.1)	k(0.6) w(0.2) uw(0.1) y(0.1)
R	r(0.5) axr(0.3) aa(0.1) er(0.1)	r(0.6) axr(0.3) aa(0.1)
S	s(0.7) z(0.3)	s(0.8) z(0.2)
T	t(0.9) sil(0.1)	t(0.9)
U	uw(0.5) y(0.2) ah(0.1) ao(0.1) zh(0.1)	uw(0.5) y(0.2) ah(0.1) ax(0.1) ao(0.1)
V	v(0.7) ay(0.3)	v(0.7) ay(0.3)
W	w(0.7) aw(0.2) uw(0.1)	w(0.7) aw(0.2) uw(0.1)
X	k(0.5) s(0.5)	s(0.5) k(0.5)
Y	iy(0.6) ay(0.2) oy(0.1) y(0.1)	iy(0.7) ay(0.2) oy(0.1)
Z	ay(0.5) z(0.4) w(0.1)	ay(0.5) z(0.5)
sil	sil(1.0)	sil(1.0)

In other words, for context *mono* the combination of one-to-one and one-to-many G2P relationships is better captured by lexical model parameters of the Tied-HMM, SP-HMM approaches followed by the KL-HMM RKL, KL-HMM SKL and KL-HMM KL approaches. The analysis also validates the hypothesis given in Section 3.4 that the local scores S_{KL} and S_{RKL} can model better one-to-one and one-to-many lexical-to-acoustic unit relationships, respectively, whereas the local score S_{SKL} can model both one-to-one and one-to-many lexical-to-acoustic unit relationships.

Effect of Increasing Grapheme Subword Unit Context

In this section, the lexical model parameters of consonant grapheme [C] and vowel grapheme [A] are analyzed with increasing subword context to gain insight into the effect of contextual modeling. The analysis is performed on the lexical model parameters of the KL-HMM *SKL* approach. In English, the G2P relationship for consonant grapheme [C] and vowel grapheme [A] is one-to-many.

Context-independent subword unit modeling: The top two components of the lexical model parameters of three-state grapheme models [C] and [A] with the corresponding phone label and probability values are given in Table 4.6. It can be observed that the parameters of different states capture relationship to different phones. The parameters of the consonant grapheme model [C] capture relationship to three phones /k/, /ch/ and /s/ in three different states and the parameters of the vowel grapheme model [A] capture /ae/, /ey/, /ax/, /eh/ in different states. In other words, the lexical model parameters capture crude phone information (also observed in the previous section). Similar trends were observed for other consonant graphemes (that have one-to-many G2P relationships) and vowel graphemes.

Table 4.6 – The first two components of the lexical model parameters arranged in descending order for grapheme models [C] and [A], shown with the corresponding phone label and the probability value

Model: [C]	State: 1	State: 2	State: 3
1st Max	/s/ (0.6)	/ch/ (0.3)	/k/ (0.9)
2nd Max	/z/ (0.1)	/t/ (0.3)	/t/ (0.02)
Model: [A]	State: 1	State: 2	State: 3
1st Max	/ae/ (0.64)	/ey/ (0.54)	/ax/ (0.32)
2nd Max	/eh/ (0.13)	/ax/ (0.08)	/ae/ (0.1)

Context-dependent subword unit modeling: The top two components of the lexical model parameters of grapheme models [C] and [A] in different contexts with the corresponding phone label and probability values are shown in Table 4.7. It can be observed from the table, that by modeling the single preceding and following context, ambiguity in the G2P relationship is resolved for three context-dependent graphemes [O-C+A], [R-C+E] and [I-C+H]. However, the parameters of the context-dependent vowel grapheme model [V-A+R] capture more than one phone (/ae/, /ey/ and /r/). Table 4.7 shows that the parameters of the vowel grapheme model [b~V-A+R*I] are able to resolve the ambiguity in the G2P relationship and dominantly capture the relationship to phone /ey/. Also, the lexical model parameters of the third state tend to model the transition information, i.e., transition to phone /r/.

In other words, similarly to G2P conversion systems [Taylor et al., 1998, Chen, 2003, Bisani and Ney, 2008, Novak, 2011], one-to-many G2P relationships captured by lexical model parameters tend to become more regular or close to one-to-one as a longer grapheme context is modeled.

Table 4.7 – The first two components of the lexical model parameters arranged in descending order for grapheme models [O-C+A], [R-C+E], [I-C+H], [V-A+R] and [*b*~V-A+R*I] ('*b*' refers to begin of the word tag), shown with the corresponding phone label and the probability value

Model: [O-C+A]	State: 1	State: 2	State: 3
1st Max	/ow/ (0.87)	/k/ (0.89)	/k/ (0.99)
2nd Max	/l/ (0.05)	/t/ (0.04)	/t/ (0.01)
Model: [R-C+E]	State: 1	State: 2	State: 3
1st Max	/s/ (0.95)	/s/ (0.82)	/s/ (0.90)
2nd Max	/z/ (0.04)	/z/ (0.12)	/z/ (0.08)
Model: [I-C+H]	State: 1	State: 2	State: 3
1st Max	/ch/ (0.80)	/ch/ (0.95)	/ch/ (0.76)
2nd Max	/t/ (0.14)	/t/ (0.03)	/t/ (0.15)
Model: [V-A+R]	State: 1	State: 2	State: 3
1st Max	/ae/ (0.35)	/ey/ (0.74)	/r/ (0.91)
2nd Max	/ey/ (0.34)	/eh/ (0.08)	/axr/ (0.04)
Model: [b~V-A+R*I]	State: 1	State: 2	State: 3
1st Max	/ey/ (0.35)	/ey/ (0.76)	/r/ (0.62)
2nd Max	/ae/ (0.34)	/eh/ (0.09)	/ey/ (0.20)

Global view: In order to get a global picture on the effect of context on the G2P relationship captured by lexical model parameters, we first trained single state grapheme models for three contexts (*mono*, *tri* and *quint*) using the KL-HMM *SKL* and Tied-HMM approaches. Then the entropy of the lexical model parameters of each grapheme lexical unit was computed. In the case of *tri* and *quint*, the average entropy of the lexical model parameters with the same center grapheme was computed. Figures 4.2 and 4.3 plot the entropy of the lexical model parameters for all the grapheme models with increasing context for the KL-HMM *SKL* and Tied-HMM approaches, respectively.

From Figure 4.2 the following observations can be made:

- The lexical model parameters of vowel graphemes ([A], [E], [I], [O], [U]) and some consonant graphemes ([C], [H], [R], [X]) have a high entropy for context *mono* signifying the fact that the parameters capture one-to-many G2P relationships. As the context increases, entropy decreases i.e., the lexical model parameters tend to capture one-to-one G2P relationship.
- The lexical model parameters of a few consonant graphemes like [B], [K], [P], [V] have low entropy for context *mono* which suggests that context-independent grapheme itself models one-to-one G2P relationship. However, the entropy slightly increases as the context increases. A closer look at the parameters revealed that this was due to the context information captured by the lexical model parameters.

For the Tied-HMM approach, as shown in Figure 4.3, entropy of the lexical model parameters of all the graphemes decreases rapidly with context compared to KL-HMM *SKL* lexical model

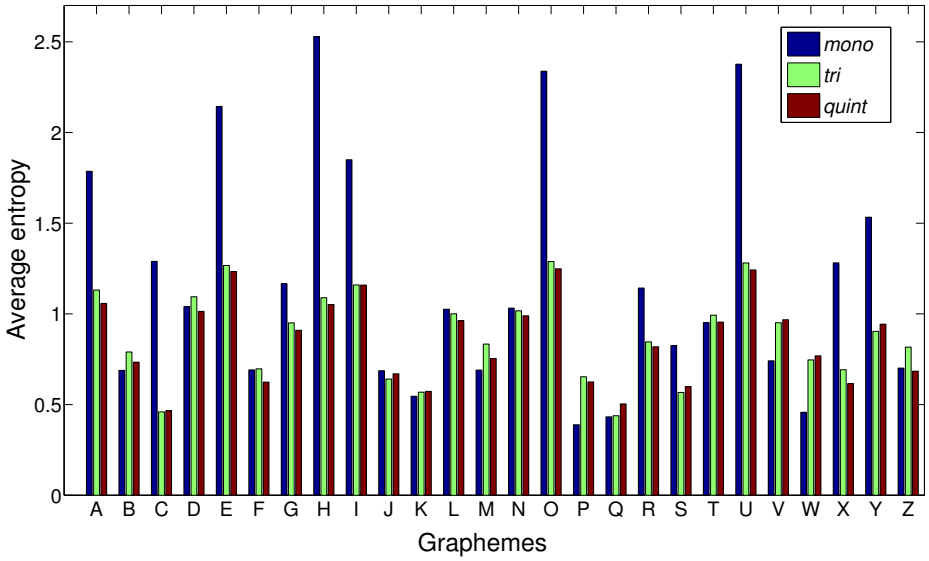

Figure 4.2 – Entropy of lexical model parameters of grapheme subword units trained using the KL-HMM *SKL* approach with increasing context. For contexts *tri* and *quint*, the average entropy of all the grapheme models with the same center grapheme is displayed.

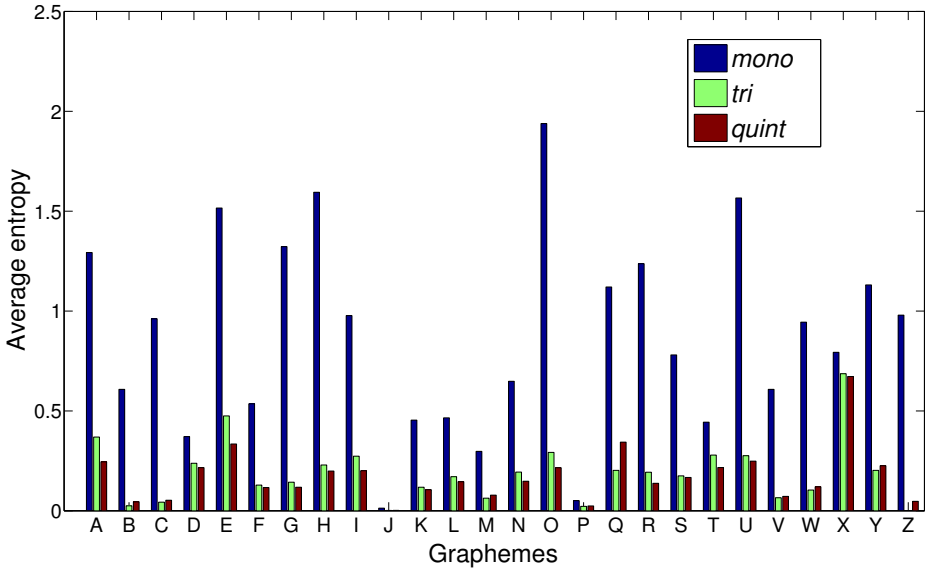

Figure 4.3 – Entropy of lexical model parameters of grapheme subword units trained using the Tied-HMM approach with increasing context. For contexts *tri* and *quint*, the average entropy of all the grapheme models with the same center grapheme is displayed.

parameters. This shows that with increasing context, the lexical-to-acoustic unit relationship modeled by the Tied-HMM approach is close to one-to-one or deterministic. However, as seen in Table 4.7 the relationship between context-dependent graphemes and context-independent phones can be one-to-many. This suggests that as subword context increases, the lexical model parameters of the Tied-HMM approach may not be able to capture one-to-many G2P relationships.

4.4.3 ASR Results

The word accuracy (WA) on the test set of the RM corpus for various systems is given in Table 4.8.

Table 4.8 – Word accuracies expressed in % on the test set of the RM corpus for various systems with phones and graphemes as subword units. The acoustic model of the probabilistic lexical model based systems is trained on the RM corpus. Boldface indicates the best system for each subword unit

System	Grapheme			Phone		
	mono	*tri*	*quint*	*mono*	*tri*	*quint*
KL-HMM *SKL*	67.1	**94.1**	**94.8**	92.9	**94.9**	**94.8**
KL-HMM *RKL*	74.2	93.5	94.3	92.0	94.1	94.2
KL-HMM *KL*	57.9	92.3	93.9	92.9	94.5	94.6
Tied-HMM	**78.8**	93.5	94.5	**93.2**	94.2	94.2
SP-HMM	77.1	92.9	93.7	**93.1**	94.0	94.0
Tandem	**78.7**	93.7	–	89.4	94.3	–
HMM/GMM	64.0	92.7	–	89.5	94.3	–

The key observations from the table are as follows:

- For context *mono*, all the grapheme-based systems yield significantly poor performance compared to their respective phone-based systems. As the context of grapheme lexical units is increased, performance of the systems improve. The grapheme-based KL-HMM *SKL* system modeling *quint* context performs comparable to the phone-based KL-HMM *SKL* system modeling *tri* context.
- HMM/GMM systems using graphemes as subword units performed significantly worse than systems using phones as subword units. For grapheme subword units, the performance of the Tandem systems is better than that of the HMM/GMM systems whereas for phone subword units, the performance of the HMM/GMM and Tandem systems were similar. The results consistent with the literature [Dines and Magimai-Doss, 2007] show that incorporating phone knowledge in grapheme systems could help in improving the performance.
- For context *mono*, among different probabilistic lexical modeling approaches, Tied-HMM and SP-HMM perform better than KL-HMM systems; and the KL-HMM *RKL* system performs better than the KL-HMM *SKL* and KL-HMM *KL* systems. The ASR results are consis-

tent with the analysis of lexical model parameters presented in the previous section where it was observed that for context *mono*, the G2P relationship is better captured by lexical model parameters trained using local scores S_{tied} and S_{SP} followed by local score S_{RKL}.
- For contexts *tri* and *quint*, the KL-HMM SKL system performs better than other probabilistic lexical modeling based ASR systems.
- The performance of the KL-HMM SKL system with phones as subword units is better than that of the HMM/GMM system.
- The performance of all KL-HMM systems modeling *tri* context increases compared to *mono* context for all local scores. However, modeling the *quint* context helps in the case of grapheme subword units and does not always improve over *tri* for phone subword units.

The results indicate that in the framework of probabilistic lexical modeling, grapheme-based system with *quint* context could yield performance comparable to that of the phone-based system modeling *tri* context. Furthermore, for both graphemes and phones, the performance of KL-HMM SKL systems modeling context-independent acoustic units is better than that of HMM/GMM and Tandem systems modeling context-dependent acoustic units.

4.5 Summary

This chapter provided a literature survey on how acoustic and lexical resource constraints are addressed in standard HMM-based ASR systems. It emerged that the deterministic lexical model imposes constraints such as, the acoustic units and the lexical units have to be of the same kind; the acoustic resources from the target language or domain are required to train or adapt both the acoustic model and lexical model. We then presented a brief literature survey to show the potential of probabilistic lexical modeling. Probabilistic lexical modeling relaxes certain constraints imposed by deterministic lexical modeling and, as a consequence the acoustic model and the lexical model can be independently trained on different set of resources; different kinds of subword units can be modeled in an ASR system and different types of contextual units can be modeled in an ASR system.

We proposed an approach to build grapheme-based ASR systems in the framework of probabilistic lexical modeling. In a pilot study, conducted on English, we showed that the parameters of the lexical model indeed capture a probabilistic G2P relationship. The analysis also validated our hypothesis that lexical model parameters learned with cost functions based on local scores S_{KL} and S_{RKL} model one-to-one and one-to-many G2P relationships, respectively, better than other cost functions based on other local scores.

The following chapters will progressively explore the potential of the proposed grapheme-based ASR approach to build ASR systems in various resource constrained scenarios. More specifically:

1. In Chapter 5, Lexical resource constrained ASR, we focus on ASR systems for a new domain with inadequate lexical resources.
2. In Chapter 6, Lexical and acoustic resource constrained ASR, we focus on rapid develop-

ment of ASR systems in both acoustic and lexical resource constraints.

3. In Chapter 7, Zero-resourced ASR, we focus on building an ASR system for a new language without any acoustic and lexical resources.

5 Lexical Resource Constrained ASR

In this chapter, we investigate the potential of the grapheme-based ASR approach proposed in Chapter 4 (Section 4.3) in addressing lexical resource constraints. More specifically, the target domain for which we are interested to build an ASR system has only acoustic data. Cross-domain acoustic and lexical resources are available, but they do not provide complete coverage on the target domain data. As presented in Section 4.1.2, in the literature the lack of lexical resources has typically been addressed using either a G2P convertor or alternate subword units like graphemes.

In the proposed grapheme-based ASR approach, first an acoustic model with phones as acoustic units and then a lexical model with graphemes as lexical units are trained. In the framework of probabilistic lexical modeling, the parameters of the acoustic model θ_a and the parameters of the lexical model θ_l can be trained on an independent set of resources (as given in Section 4.2). Therefore, it is possible to build an ASR system where the acoustic model is trained on cross-domain acoustic and lexical resources; and the lexical model is learned on target domain data. The resulting system is a grapheme-based ASR system that uses a cross-domain acoustic model and a target domain lexical model.

The proposed grapheme-based ASR approach can also be used to augment the recognition vocabulary. In this case, the grapheme-based ASR system can be trained where the acoustic unit set $\mathscr{A}$ is based on phones and the acoustic model is trained on acoustic and lexical resources of target domain data. The lexical unit set $\mathscr{L}$ is based on graphemes and the lexical model is also learned on target domain data. Since the lexical units are graphemes, the recognition vocabulary can be augmented easily.

Figure 5.1 illustrates the proposed grapheme-based ASR system. Following the analysis presented in Section 4.4.2, in the proposed approach with graphemes as lexical units and phones as acoustic units, the lexical model parameters capture a probabilistic relationship between graphemes and phones using acoustic data. Thus, the proposed grapheme-based ASR approach integrates lexicon learning as a phase in ASR system training.

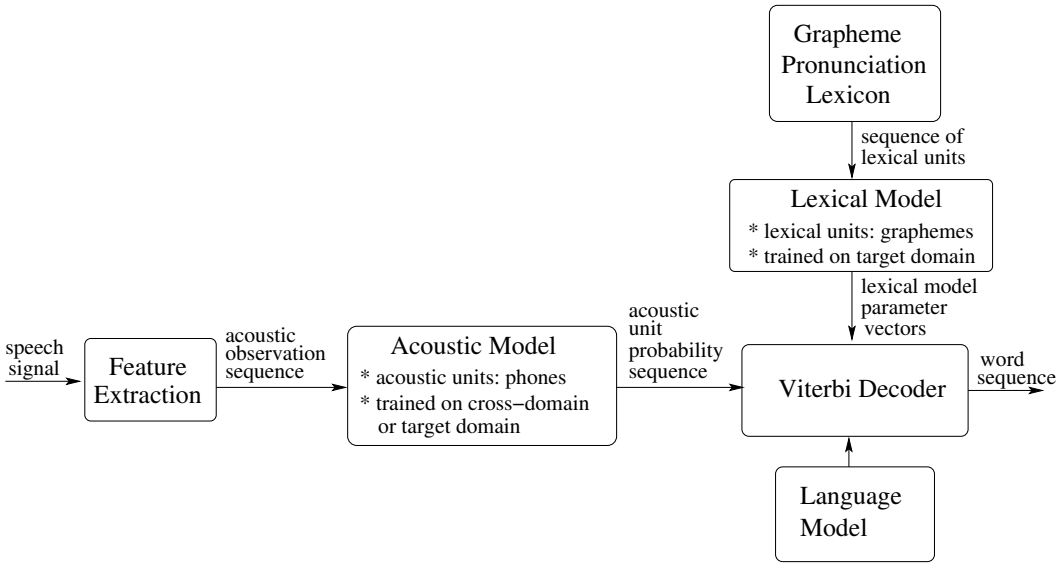

Figure 5.1 – Block diagram of the grapheme-based ASR system using probabilistic lexical modeling

In the experimental evaluation on three different resource constrained ASR tasks, the proposed approach is compared with the standard ASR approach where first a G2P convertor is trained on cross-domain lexical resources and then a phone-based ASR system is built on target domain resources. More specifically, we compare the following systems:

- The phone-based ASR system using well developed phone lexicon, as shown in Figure 5.2(a).
- The phone-based system using the phone lexicon obtained from a G2P convertor as shown in Figure 5.2(b). The G2P convertor is based on joint sequence models [Bisani and Ney, 2008].
- The grapheme-based ASR system using the HMM/GMM approach proposed in the literature [Kanthak and Ney, 2002, Killer et al., 2003] as shown in Figure 5.2(c).
- The grapheme-based ASR system using Tandem features [Dines and Magimai-Doss, 2007] as shown in Figure 5.2(d) and the phone-based ASR using Tandem features [Hermansky et al., 2000].
- The grapheme-based ASR system using the Tied-HMM approach as shown in Figure 5.2(e) and the phone-based ASR system using the Tied-HMM approach.
- The grapheme-based ASR system using the KL-HMM and SP-HMM approaches as shown in Figure 5.2(f), and the phone-based ASR system using the KL-HMM and SP-HMM approaches.

All the studies are on English where the G2P relationship is irregular.

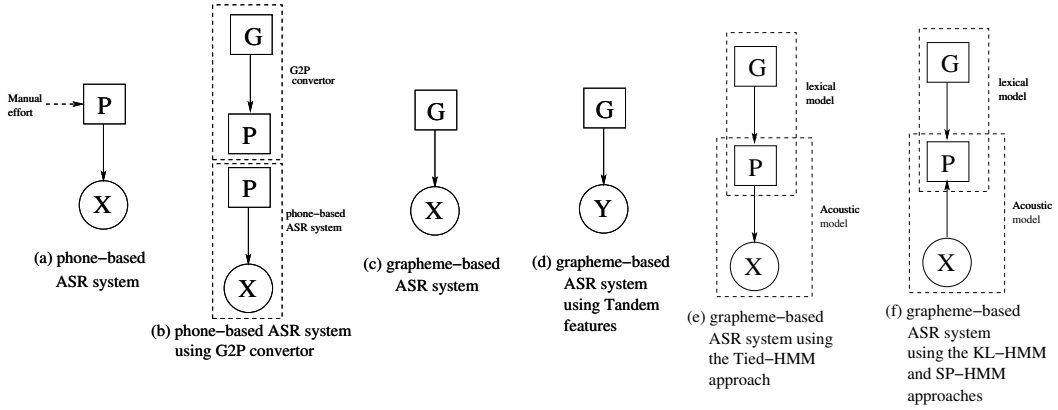

Figure 5.2 – Graphical model representation of various systems. In the figure, P refers to phone subword units, G refers to grapheme subword units, X refers to acoustic feature observations and Y refers to Tandem features

Table 5.1 – Overview of different systems. CI denotes context-independent subword units, CD denotes context-dependent subword units and cCD denotes clustered context-dependent subword units. P and G denote the phone lexicon and the grapheme lexicon, respectively. *Det* denotes the lexical model is deterministic and *Prob* denotes the lexical model is probabilistic.

System	Acoustic units $\mathscr{A}$	Lexicon	Lexical units $\mathscr{L}$	Approach
KL-HMM	CI	P or G	CD	*Prob*
SP-HMM	CI	P or G	CD	*Prob*
Tied-HMM	CI	P or G	CD	*Prob*
HMM/GMM	cCD	P or G	CD	*Det*
Tandem	cCD	P or G	CD	*Det*

5.1 Experimental Evaluation

In this chapter, the probabilistic lexical modeling systems, i.e., KL-HMM, SP-HMM and Tied-HMM are compared with the deterministic lexical modeling systems i.e., HMM/GMM and Tandem systems. The different systems and their capabilities are summarized in Table 5.1. ASR systems are built using the following three lexica:

1. *GRAPH* - grapheme lexicon transcribed using the orthography of words.
2. *G2P* - phone lexicon obtained by G2P conversion. We used the sequitur G2P toolkit [Bisani and Ney, 2008] for this purpose.
3. *PHONE* - well developed phone lexicon that serves as an optimistic case as it is manually built and verified.

In the pilot study all the probabilistic lexical model based systems modeled word-internal subword contexts. However, in this section and here after, all the systems model crossword context-dependent subword units (*tri* lexical units), as shown in Table 5.1. Probabilistic lexical

model based systems used the KL-divergence based decision tree approach for state tying and clustering [Imseng et al., 2012b]. This decision tree clustering approach optimizes a cost function based on the local score S_{KL}. Similarly to likelihood based decision tree criteria, the combination of minimum state occupancy count and the minimum decrease in cost-function threshold are used as stopping criteria.

We use MLPs trained to classify context-independent phones as the acoustic models for the KL-HMM, SP-HMM and Tied-HMM systems. Input to all the MLPs is the 39-dimensional PLP cepstral coefficient vector with a four-frame preceding and a four-frame following context. Following the previous work [Pinto et al., 2011], the size of the hidden layers for all the MLPs is determined by fixing the total number of parameters to 35% of the training data. The parameters of the lexical model for the KL-HMM, SP-HMM and Tied-HMM systems are trained on target domain acoustic data.

The 39-dimensional PLP feature vector used to train the MLP are also used to train the HMM/GMM systems. The Tandem features were extracted by transforming the output of the MLPs (same MLPs that are used as acoustic models in probabilistic lexical model based ASR systems) with log transformation followed by KLT. Similarly to probabilistic lexical model based systems, the Tandem system also exploits both target-domain and cross-domain resources. The HMM/GMM system is trained on target domain data alone. The number of mixture components for each of the tasks in the case of the HMM/GMM and Tandem systems are tuned on the development set. In this chapter, we do not perform acoustic model adaptation of HMM/GMM systems on cross-domain resources, as it is assumed that the tasks lack only lexical resources.

All the phone-subword based systems use a phonetic question set and grapheme subword based systems use a singleton question set for the decision tree state tying procedure. In the following subsections we will describe the three ASR studies investigated. The studies presented here reflect practical scenarios. For example, cross-domain porting where there is a need to adapt an existing ASR system to a new application domain or to accented speech; lexicon augmentation where there is a need to add new words to the recognition vocabulary because the language is under-resourced etc.

5.1.1 Cross-Domain ASR Study

The goal is to build an ASR system for a new domain with lexical resource constraints by exploiting cross-domain acoustic and lexical resources. The cross domain lexicon has a high out-of-vocabulary rate on the new domain. In that regard, we present an experimental study where the RM corpus (see Appendix A.1) is considered as the target domain and the WSJ1 corpus (see Appendix A.2) as the cross-domain. The standard RM setup with 3990 train utterances and 1200 test utterances is used in this study. Though RM and WSJ are similar domains, among the 1000 words present in the RM task, the WSJ task includes only 568 words. That is, the RM task has 432 words that are not seen in the WSJ pronunciation lexicon.

We use an *off-the-shelf* MLP [Aradilla et al., 2008] trained on the WSJ1 (to classify 45 context-independent phones) as the acoustic model for all the probabilistic lexical model based systems.

The *G2P* lexicon was built by training a joint n-gram based G2P convertor on the WSJ lexicon using the sequitur G2P toolkit [Bisani and Ney, 2008]. The width of the graphone (a grapheme phoneme pair) context was tuned on the development set (5% of the WSJ1 lexicon). The optimal n-gram context size was 5. The performance of the *G2P* lexicon compared to the *PHONE* lexicon given in the RM task was 92.2% phone accuracy. Furthermore, systems are also built with the *GRAPH* lexicon (transcribed using 79 graphemes, details are given in Appendix A.1) and the well developed *PHONE* lexicon given in the RM task.

5.1.2 Multi Accent Non-Native ASR Study

The goal is to build an ASR system for non-native speech including multiple accents in a lexical resource constrained scenario. In this study, cross-domain acoustic and lexical resources are from native language speakers. The spoken words in non-native speech are pronounced differently from native pronunciations. Capturing these variations through multiple pronunciations is not a trivial task [Strik and Cucchiarini, 1999]. Therefore, the approaches should implicitly handle lexical resource constraints and model the pronunciation variability.

We study multi-accent non-native speech recognition using the HIWIRE corpus (details are given in Appendix A.4). As cross-domain resources we use the SpeechDat(II) British English (see Appendix A.3) corpus that includes acoustic and lexical resources from native language speakers.

The acoustic model or the MLP for probabilistic lexical model based systems was trained on the SpeechDat(II) British English corpus to classify 45 context-dependent phones. SpeechDat(II) is a telephone speech corpus, hence, the HIWIRE speech was down sampled to 8kHZ before extracting PLP cepstral features and then forward passed through the SpeechDat(II) English MLP.

The *G2P* lexicon was built by training a joint n-gram based G2P convertor on the SpeechDat(II) British English lexicon using the sequitur G2P toolkit [Bisani and Ney, 2008]. To extract pronunciations of abbreviated words using the G2P convertor, the spelling of the word was modified according to the way the word is pronounced (similar to the pronunciations of abbreviated words in grapheme lexicon as given in Appendix A.4). For example, the word "S.I.D" is presented as "ES-EYE-DEE" to the G2P convertor. The optimal width of the graphone context was found to be 6. The performance of the *G2P* lexicon compared to the *PHONE* lexicon of the HIWIRE task was 89.4% phone accuracy.

5.1.3 Lexicon Augmentation Study

In this study our goal is to augment the test vocabulary with new words that are not present in the training vocabulary. The training data includes limited transcribed speech data with the phone pronunciations of words seen in the training data. The study is performed on the PhoneBook speaker-independent task-independent 600 word isolated word recognition corpus (details are given in Appendix A.5) where none of the words in the test vocabulary are present in the training vocabulary.

The acoustic model or the MLP for probabilistic lexical model based systems was trained on limited training data of the PhoneBook corpus to classify 42 context-independent phones. For MLP training, we followed the same setup as in [Dupont et al., 1997], where 19421 utterances are used for training and 7920 utterances for cross validation. Thus, the data used to train the MLP did not contain any of the test words.

Probabilistic lexical model systems and deterministic lexical model systems are built using both training and cross validation utterances consisting of 27341 utterances covering 2183 words. Phone-based probabilistic lexical model systems used the phone lexicon given in the PhoneBook corpus with acoustic units as context-independent phones and lexical units as context-dependent phones. Grapheme-based probabilistic lexical model systems are trained with acoustic units as context-independent phones and lexical units as context-dependent graphemes. In the case of the PhoneBook task word-internal context-dependent systems are built (as it is an isolated word recognition task).

The *G2P* lexicon for the test set was built by training a G2P convertor on the training and cross-validation pronunciation lexicon (consisting of pronunciations for 2183 words) using the sequitur toolkit [Bisani and Ney, 2008]. The performance of the *G2P* lexicon compared to the *PHONE* lexicon given in the PhoneBook task was 89.2% phone accuracy. Furthermore, systems are also decoded with the *PHONE* lexicon given in the PhoneBook corpus.

5.2 Results

5.2.1 Baselines

To study the effect of lexicon on ASR accuracy, in this section, we compare the performance of HMM/GMM systems using three different lexica, namely, *GRAPH*, *G2P* and *PHONE* on the RM, HIWIRE and PhoneBook tasks. Word accuracies of the HMM/GMM systems for the three tasks are given in Table 5.2. Results show that for all the three tasks, the system using *PHONE* lexicon performs better than the systems using the *GRAPH* or *G2P* lexicon. ASR results indicate that the standard HMM/GMM system that uses deterministic lexical modeling is not able to handle the pronunciation errors present in the *GRAPH* and *G2P* lexica.

The results also show that on the RM and HIWIRE tasks, systems using the *GRAPH* lexicon and

Table 5.2 – Word accuracies (expressed in %) of the crossword context-dependent HMM/GMM systems using the *GRAPH*, *G2P* and *PHONE* lexica on the RM, HIWIRE and PhoneBook tasks. Boldface indicates the best system for each task.

Task	*GRAPH*	*G2P*	*PHONE*
RM	94.8	95.1	**95.9**
HIWIRE	96.4	96.1	**97.2**
PhoneBook	91.0	87.1	**97.0**

the *G2P* lexicon perform similarly. However, on the PhoneBook task where the recognition vocabulary is entirely different from the train vocabulary, the system using the *GRAPH* lexicon performs significantly better than the system using the *G2P* lexicon.

5.2.2 Probabilistic Lexical Modeling based Systems

The performance in terms of word accuracy of the various systems using three different lexica on the RM, HIWIRE and PhoneBook tasks is given in Tables 5.3, 5.4 and 5.5, respectively.

Table 5.3 – Word accuracies (expressed in %) of the crossword context-dependent ASR systems on the test set of the RM corpus. Boldface indicates the best system for each lexicon.

System	*GRAPH*	*G2P*	*PHONE*
KL-HMM SKL	**95.5**	**95.6**	**95.9**
KL-HMM RKL	95.3	95.0	95.3
KL-HMM KL	94.5	95.0	95.5
Tied-HMM	94.0	94.3	94.5
SP-HMM	93.5	94.0	94.2
Tandem	94.5	94.6	95.4
HMM/GMM	94.8	95.1	**95.9**

Table 5.4 – Word accuracies (expressed in %) of the crossword context-dependent ASR systems on the test set of the HIWIRE corpus. Boldface indicates the best system for each lexicon.

System	*GRAPH*	*G2P*	*PHONE*
KL-HMM SKL	**97.5**	96.8	97.3
KL-HMM RKL	97.4	**97.2**	**97.4**
KL-HMM KL	97.3	96.1	96.8
Tied-HMM	95.9	94.6	95.9
SP-HMM	95.4	93.6	95.0
Tandem	96.6	96.2	97.0
HMM/GMM	96.4	96.1	97.2

The main observations from the three tasks are as follows:

Table 5.5 – Word accuracies (expressed in %) of the context-dependent ASR systems on the test set of the PhoneBook corpus. Boldface indicates the best system for each lexicon.

System	*GRAPH*	*G2P*	*PHONE*
KL-HMM *SKL*	**93.6**	**89.1**	97.8
KL-HMM *RKL*	93.3	88.3	97.7
KL-HMM *KL*	92.3	88.8	**97.9**
Tied-HMM	91.6	86.7	96.8
SP-HMM	90.5	86.7	96.6
Tandem	92.7	84.9	97.4
HMM/GMM	91.0	86.7	97.0

- In the framework of probabilistic lexical modeling, especially using the KL-HMM *SKL* approach, the ASR system using the *GRAPH* lexicon performs similar to or better than the ASR system using the *G2P* lexicon. Furthermore, in the case of the RM and HIWIRE tasks, the KL-HMM *SKL* system using the *GRAPH* lexicon achieves performance comparable to the system using the optimistic well developed *PHONE* lexicon.
- For the *GRAPH* and *G2P* lexica, among different probabilistic lexical modeling approaches, the KL-HMM *SKL* system generally performs better followed by the KL-HMM *RKL*, KL-HMM *KL*, SP-HMM and Tied-HMM systems. The best performing probabilistic lexical modeling approach for systems using the *PHONE* lexicon varied with the task.
- The Tied-HMM and SP-HMM systems perform worse compared to the KL-HMM or deterministic lexical model based systems. The analysis presented in Section 4.4.2 suggested that the lexical model parameters of the Tied-HMM system are not able to capture well the one-to-many G2P relationships with increasing context. Given that the acoustic model is trained on cross-domain data, and there exist inherent pronunciation errors in the *GRAPH* and *G2P* lexica, the lexical-to-acoustic unit relationship can be one-to-many. This could be the reason for the poor performance of the Tied-HMM systems compared to the KL-HMM systems.
- For the *GRAPH* and *G2P* lexica, performance of the KL-HMM *SKL* system is always better than that of the HMM/GMM systems. For the *PHONE* lexicon, the performance of the KL-HMM *SKL* system is similar to that of the HMM/GMM systems.
- The performance of the Tandem system, that also exploits cross-domain resources is worse (for all the three lexica) than the best performing probabilistic lexical model based ASR system. The performance of the Tandem and HMM/GMM systems are comparable.

The results confirm our hypothesis that the proposed grapheme-based system can perform better than or similarly to the phone-based system using the phone lexicon from a G2P convertor. Furthermore, the proposed grapheme-based system performs better than the grapheme-based HMM/GMM and Tandem systems proposed in the literature.

5.3 Summary

In this chapter, we evaluated the potential of the proposed grapheme-based ASR approach in addressing lexical resource constraints on three different ASR studies. In the first study, cross-domain acoustic and lexical resources are available but they do not provide a complete coverage for the target domain. In the second study, cross-domain resources are from native language speakers and include telephone speech whereas the target domain includes non-native speakers and clean speech. In the third study, the recognition vocabulary includes words that are not seen during training; therefore, there is a need to add new words to the system vocabulary.

Our studies have shown that the proposed grapheme-based ASR approach which implicitly integrates lexicon learning performs better than or comparably to the conventional two stage approach where G2P training is followed by ASR system development. Furthermore, the studies show that the proposed grapheme-based ASR approach that incorporates probabilistic lexical modeling outperformed the grapheme-based ASR approaches with deterministic lexical modeling.

6 Lexical and Acoustic Resource Constrained ASR

The previous chapter showed that it is possible to build a grapheme-based ASR system using probabilistic lexical modeling where the acoustic model is trained on cross-domain data and the lexical model is trained on target domain data with graphemes as lexical units. It has been observed that this grapheme-based ASR approach could perform better than or comparably to the two stage approach where phone lexicon development (through automatic G2P conversion) is followed by ASR system development, even for languages such as English.

In this chapter, we extend the proposed approach for rapid development of ASR systems for new domains or languages with both acoustic and lexical resource constraints. In the proposed approach:

- First, an acoustic model that models multilingual phones is trained on language-independent acoustic and lexical resources.
- Then, the lexical model which captures a probabilistic relationship between target language graphemes and multilingual phones is trained on a relatively small amount of target language-dependent acoustic data.

The first advantage of the proposed grapheme-based ASR approach is that it capitalizes on both acoustic and lexical resources of resource-rich languages other than the target language. As discussed earlier in Section 4.1.3, other multilingual and crosslingual grapheme-based ASR approaches proposed in the literature focussed on sharing grapheme models across languages [Kanthak and Ney, 2003, Stüker, 2008a,b]. The second advantage of the proposed approach is that lexicon learning is integrated as a phase in ASR system training.

We hypothesize that, compared to the conventional approach of rapid development of ASR system through acoustic model adaptation of deterministic lexical model based ASR systems, ASR systems can be rapidly and effectively built with the proposed grapheme-based probabilistic lexical modeling approach. To validate our hypothesis, the proposed approach is compared with standard deterministic lexical modeling based approaches such as, a) the HMM/GMM approach, where the acoustic and lexical model are trained on target language data and b) acoustic model adaptation based approaches (MAP and MLLR) that exploit language-independent resources and c) the Tandem approach that also exploits language-independent

resources.

The hypothesis is validated by training a single language-independent multilingual acoustic model and conducting ASR studies on the following three different resource-constrained tasks where only the lexical model is trained:

- Non-native accented speech recognition task that lacks both acoustic resources and "well developed" phonetic lexical resources (typically, the phone lexicon includes native speaker pronunciations). In the literature, non-native accented ASR research has mainly focused on acoustic model adaptation. We investigate it on English where the G2P relationship is irregular. The HIWIRE multi-accent non-native ASR corpus used in the previous chapter (see Section 5.1.2) is also used in this chapter.
- Rapid development of an ASR system for a new language that is not present in language-independent data using minimal acoustic and lexical resources. We demonstrate this aspect on a Greek ASR task.
- Development of an ASR system for a minority and under-resourced language, particularly, Scottish Gaelic which has only 60,000 speakers. The endangered status of Gaelic makes low-cost speech technology important for language conservation efforts. Gaelic also lacks sufficient acoustic resources and does not have any phonetic lexical resources.

6.1 Experimental Setup

In this section, we describe the different databases and the setup of the systems used.

6.1.1 Databases and Setup

The information about the various corpora used is summarized in Table 6.1.

Language-Independent Dataset

A part of the SpeechDat(II) corpus, particularly, British English, Italian, Spanish, Swiss French and Swiss German, is used as the language-independent dataset. Each language has approximately 12 hours of speech data, totally amounting to 63 hours. All the SpeechDat(II) lexica use SAMPA symbols. A multilingual phone set of 117 units obtained by merging phones that share the same symbols across the above mentioned five languages, serves as the acoustic (or the subword) unit set.

Non-native HIWIRE

To study non-native accented speech recognition we revisit the HIWIRE task (details are given in Appendix A.4) used in the previous chapter. Additionally, to simulate limited resources, the amount of adaptation data of HIWIRE is reduced from 150 min to 3 min (specifically,

Table 6.1 – Overview of the tasks and the respective corpora used in the study

Corpus (Description)	Language	# of Subword units		Training data	Test data
		Phones	Graphemes	(in min)	(in min)
SpeechDat(II)	English	45	27	744	n.a
(Native speech	French	42	43	810	n.a
sampled at 8K	German	59	42	846	n.a
used to train	Italian	52	34	690	n.a
the acoustic model)	Spanish	32	34	690	n.a
(language-independent data used to train *multilingual acoustic model*)		*117*	*47*	*3780*	n.a
HIWIRE (Non-native speech from natives of France, Spain, Italy and Greece)	English	42	27	0 to 150	150
SpeechDat(II) (Native Greek speech)	Greek	31	25	5 to 800	360
Scottish Gaelic (Broadcast news data)	Scottish Gaelic	n.a.	83 or 32	180	60

150 min, 120 min, 90 min, 64 min, 32 min, 16 min, 10 min and 3 min) by picking a subset of utterances as in [Imseng et al., 2011]. The standard adaptation set of HIWIRE consists of 50 sentences per speaker. The amount of adaptation data is decreased by considering 40, 30, 20, ten, five and three sentences per speaker. Furthermore, to ensure full coverage in terms of context-independent subword units, we picked different sentences for grapheme and phone based systems in the case of 32 min, 16 min and 10 min scenarios. For the three min case, utterances are randomly picked until all the context-independent phones or graphemes are covered.

As described in Appendix A.4, we use the phone lexicon based on the SAMPA phone set. With the SAMPA phone set, the HIWIRE and language-independent datasets have a shared subword unit set. This allowed the evaluation of acoustic model adaptation based systems (MAP and MLLR) discussed later in Section 6.1.2. Also, native English is present in out-of-domain resources. Therefore, in the case of the KL-HMM, SP-HMM and Tied-HMM approaches, the lexical model parameters trained on SpeechDat(II) English are adapted using HIWIRE adaptation data. Additionally, we could also investigate the case where no lexical model or acoustic model adaptation is performed.

Other details about the task such as grapheme lexicon, language model are given in Appendix A.4.

Greek SpeechDat(II)

Rapid development of an ASR system for a new language is studied using the Greek SpeechDat(II) corpus (details are given in Appendix A.3). The experimental setup is based on [Imseng

et al., 2012b]. Since the database does not include a standard language model, two optimistic bi-gram language models, one from the sentences in the development set and other from the sentences in the test set are built. To simulate limited resources, the amount of available training data is reduced from 13.5 hours to 5 minutes (specifically, 800 min, 300 min, 150 min, 75 min, 37 min, 18 min, 9 min and 5 min). All the systems were evaluated on the same test set. The test set contains 10k unique words. The performance of the phone-based KL-HMM, MAP, MLLR and HMM/GMM systems presented in [Imseng, 2013, Figures 4.3 and 4.4] is taken as reference in this thesis.

The acoustic model adaptation systems impose the constraint that subword unit sets of language-independent data and target language data match. As a result, grapheme-based acoustic model adaptation systems were not directly applicable to the Greek ASR task, as Greek graphemes are different from Roman graphemes. This necessitated transliteration of Greek alphabets in terms of English (Roman) alphabets, as given in Table 6.2, for grapheme-based acoustic model adaptation systems described later in Section 6.1.2.

Table 6.2 – Greek graphemes and their transliterated format (Trans.)

Grapheme	Trans.	Grapheme	Trans.
α	a	ν	n
β	b	ξ	x
γ	g	o	o
δ	d	π	p
ϵ	e	ρ	r
ζ	z	σ	s
η	h	τ	t
θ	th	υ	y
ι	i	ϕ	f
κ	k	χ	ch
λ	l	ψ	ps
μ	m	ω	w

More details about the database such as the standard phone lexicon and the grapheme lexicon are given in Appendix A.3.

Scottish Gaelic

To study the development of an ASR system for a minority and under-resourced language we use the Scottish Gaelic speech corpus collected by CSTR, University of Edinburgh [1]. The details about the corpus are given in Appendix A.6. We use two grapheme lexica in this study, namely, *orthography-based* and *knowledge-based* as given Appendix A.6. Since, the corpus does not include a language model, as done in the Greek ASR study, we trained two bi-gram language models, one from the sentences in the development set and other from the sentences in the test set.

1. http://forum.idea.ed.ac.uk/idea/gaelic-speech-recognition-and-scots-gaelic-sound -archive

Table 6.3 – Overview of different systems. CI denotes context-independent subword units, cCD denotes clustered states of the context-dependent subword-unit based HMM/GMM system and CD denotes context-dependent subword units. LI denotes language-independent data is used to train or adapt the model, LD denotes language-dependent data is used to train or adapt the model and LI+LD denotes both language-independent and language-dependent data is used to train the model. In Tandem, the ANN trained to classify context-independent acoustic units is used to extract features for the HMM/GMM system. This is indicated through (CI+), (ANN+) and (LI+) notation. *Det* denotes the lexical model is deterministic and *Prob* denotes the lexical model is probabilistic.

System	Acoustic model			Lexical Model		
	Acoustic units	Approach	Train/ Adapt	Lexical units	Approach	Train/ Adapt
KL-HMM	CI	ANN	LI	CD	Prob	LD
SP-HMM	CI	ANN	LI	CD	Prob	LD
Tied-HMM	CI	ANN	LI	CD	Prob	LD
Tandem	(CI+)cCD	(ANN+)GMM	(LI+)LD	CD	Det	LD
MAP	cCD	GMM	LI+LD	CD	Det	LI
MLLR	cCD	GMM	LI+LD	CD	Det	LI
HMM/GMM	cCD	GMM	LD	CD	Det	LD

6.1.2 Systems

We compare systems based on probabilistic lexical modeling approaches with standard HMM-based systems with different capabilities. Table 6.3 provides an overview of the systems that are investigated. The non-native and minority language ASR studies build on top of our preliminary investigations that focussed on KL-HMM and the use of word-internal context-dependent subword units [Imseng et al., 2011, Rasipuram et al., 2013a]. In this section, we provide details about the different systems given in Table 6.3 by grouping them into three categories.

Probabilistic Lexical Modeling based Systems

We use an off-the-shelf three layer MLP [Imseng et al., 2011] trained on the language-independent dataset to classify 117 context-independent multilingual phones as the acoustic model. The input to the MLP was 39-dimensional PLP feature vectors with nine frame temporal context as input. The total number of parameters was set to 10% of the number of available training frames. The lexical model is trained for each of the probabilistic lexical modeling systems, namely, KL-HMM, SP-HMM and Tied-HMM as described Chapter 3. The acoustic units are multilingual phones from the language-independent dataset and the lexical units are graphemes of the target language. Since, the acoustic units outnumber the lexical units, it is likely that the lexical-to-acoustic unit relationship is one-to-many. In Section 4.4.2, it was observed that the local score S_{RKL} captures one-to-many G2P relationship better than other

local scores. Therefore, in this chapter, we use the KL-HMM *RKL* systems.

Acoustic model adaptation based systems

We present ASR systems based on standard MAP and MLLR adaptation techniques. For this purpose, multilingual context-dependent phone-based and grapheme-based HMM/GMM systems were trained on the language-independent data set. The phone-based HMM/GMM system used multilingual phones as subword units.

All the five considered European languages use Roman alphabet. Therefore, the multilingual grapheme-based HMM/GMM system was developed by forming a multilingual grapheme set of 47 units by merging graphemes that are common across the languages in the language-independent data set. Accents and diacritics are treated as separate graphemes.

Each context-dependent subword unit was modeled using 3 HMM states and each HMM state was modeled using a mixture of 16 Gaussians. Then, MAP adaptation or MLLR adaptation [2] is performed using speech data from the target language or domain. As described earlier in Section 6.1.1, for the Greek task the transliterated grapheme-based lexicon was used while performing MAP or MLLR adaptation.

Standard language-dependent acoustic model and lexical model based ASR systems

These are HMM/GMM ASR systems where both the acoustic model and the lexical model are trained on the language-dependent data. We investigate two systems, the first system uses standard cepstral features as feature observations (HMM/GMM system) and the second system uses Tandem features as feature observations (Tandem system) [Hermansky et al., 2000]. As indicated in Table 6.3, the Tandem system exploits both language-dependent and language-independent resources similarly to probabilistic lexical model based systems and acoustic model adaptation based systems.

The Tandem features were extracted by transforming 117-dimensional outputs of the same multilingual MLP described earlier in Section 6.1.2, with log transformation followed by principal component analysis. The dimensionality of the output features is either kept the same or reduced to 39.

The HMM/GMM systems used 39-dimensional PLP cepstral feature vectors as acoustic features. All the phone subword based systems use a phonetic question set and grapheme subword based systems use a singleton question set for the decision tree state tying procedure. The number of mixture components for each of the tasks and the training conditions were tuned on the development set. Additionally, for tandem systems, the dimensionality of the feature observations (either 117 or 39 dimensions) was tuned on the development set.

2. In [Imseng et al., 2012a], it was observed that constrained MLLR [Digalakis et al., 1995] performed worse compared to MLLR for acoustic model adaptation. Therefore, in this work we investigated only MLLR adaptation.

6.2 Results

The results section is organized as follows. First, we present results on the rapid development of ASR with both acoustic and lexical resource constraints on the HIWIRE and Greek ASR tasks. Later, we present results on minority language speech recognition using the Scottish Gaelic task. Since, for the Scottish Gaelic task there are no phone-based ASR systems to compare with, we did not consider rapid development of ASR systems. The performance of all the systems is reported in terms of word accuracy.

6.2.1 Rapid ASR Development

The performance in terms of word accuracy on the HIWIRE and Greek tasks is summarized in Tables 6.4 and 6.5 for the KL-HMM, SP-HMM, Tied-HMM, Tandem, MAP, MLLR and HMM/GMM systems. The results are analyzed using Figures 6.1, 6.2, 6.3 and 6.4 along two aspects, namely, comparison of different probabilistic lexical model based systems, comparison of probabilistic lexical model based systems against acoustic model adaptation based systems and standard HMM/GMM systems.

Table 6.4 – Performance in terms of word accuracy on the HIWIRE test set for various crossword context-dependent ASR systems trained on varying amounts of the HIWIRE adaptation data.

System	3 min		10 min		120 min		150 min	
	Graph	Phone	Graph	Phone	Graph	Phone	Graph	Phone
KL-HMM	90.7	93.3	94.0	94.6	98.0	98.0	98.1	98.1
SP-HMM	91.4	93.3	92.1	94.2	95.0	95.6	95.0	95.6
Tied-HMM	86.4	92.5	88.6	93.2	94.3	95.3	94.4	95.4
MAP	86.7	91.6	88.9	92.6	96.7	97.9	96.9	98.0
MLLR	86.2	92.4	87.3	94.3	92.2	96.0	91.9	96.0
Tandem	39.5	55.3	68.9	85.4	95.4	96.2	95.9	96.5
HMM/GMM	26.7	48.3	64.8	82.6	95.8	96.6	96.4	96.8

Table 6.5 – Performance in terms of word accuracy on the Greek test set for various crossword context-dependent ASR systems trained on varying amounts of the Greek data.

System	5 min		37min		300 min		800 min	
	Graph	Phone	Graph	Phone	Graph	Phone	Graph	Phone
KL-HMM	78.0	80.3	81.4	83.0	83.8	84.4	84.5	84.8
SP-HMM	71.3	73.8	75.9	76.3	77.8	79.3	78.7	79.6
Tied-HMM	66.6	68.6	71.3	73.6	74.8	76.3	76.4	77.6
MAP	54.7	77.4	68.7	79.3	78.0	82.7	78.0	83.9
MLLR	50.0	77.3	52.6	78.7	52.8	79.1	52.8	78.7
Tandem	55.7	66.9	76.0	79.7	81.6	83.8	82.4	84.9
HMM/GMM	54.6	63.5	74.5	81.2	82.3	84.5	83.5	85.2

Probabilistic Lexical Modeling based Systems

The performance on the HIWIRE and Greek tasks is given in Figures 6.1 and 6.2, respectively, for the phone- and grapheme-based KL-HMM, SP-HMM and Tied-HMM systems with increasing amounts of training data. The figures show that the KL-HMM system consistently performs better compared to the SP-HMM and Tied-HMM systems for both phone and grapheme subword units. Furthermore, on the HIWIRE task the difference is more pronounced when the systems use graphemes as subword units.

Comparison of probabilistic lexical modeling based system with other Systems

The performance on the HIWIRE and Greek tasks is plotted in Figures 6.3 and 6.4, respectively, with varying amount of training data for the phone-based and grapheme-based KL-HMM, MAP, MLLR, Tandem and HMM/GMM systems. We can draw the following inferences from the figures:

1. KL-HMM based systems irrespective of the type of subword units used, phones or graphemes, tend to perform better than (when the training data is less) or comparable to (when training data is increased) phone-based or grapheme-based deterministic lexical model based systems. On both the HIWIRE and Greek tasks, the difference in performance between phone and grapheme-based systems is minimal for the KL-HMM approach compared to all other approaches.
2. On both the HIWIRE (where G2P relationship is irregular) and Greek (where G2P relationship is regular) tasks it can be been observed that deterministic lexical model based systems are more suitable for phones than graphemes.
    (a) On the HIWIRE task where lexical units and acoustic units match or have shared unit set, the acoustic model adaptation based systems perform better than the HMM/GMM or Tandem systems. However, the performance of acoustic model adaptation systems using graphemes is worse than with phones as subword units. On the Greek task where the transliterated grapheme-based lexicon was used, grapheme-based acoustic model adaptation systems perform significantly worse than phone-based acoustic model adaptation or HMM/GMM or Tandem systems. The results also show that in case of grapheme subword unit set mismatch, transliteration may not be the best possible alternative. In such cases, data-driven mapping of grapheme subword units could potentially be investigated [Stüker, 2008b].
    (b) When the available training data is larger, phone-based deterministic lexical model systems for both the HIWIRE and Greek tasks perform comparably to the phone-based KL-HMM system (though not using the same technique, on the HIWIRE task it is MAP and on the Greek task it is HMM/GMM and Tandem). However, in case of grapheme-based systems, this trend is not observed. The results, inline with the other multilingual grapheme-based ASR studies [Kanthak and Ney, 2003, Killer

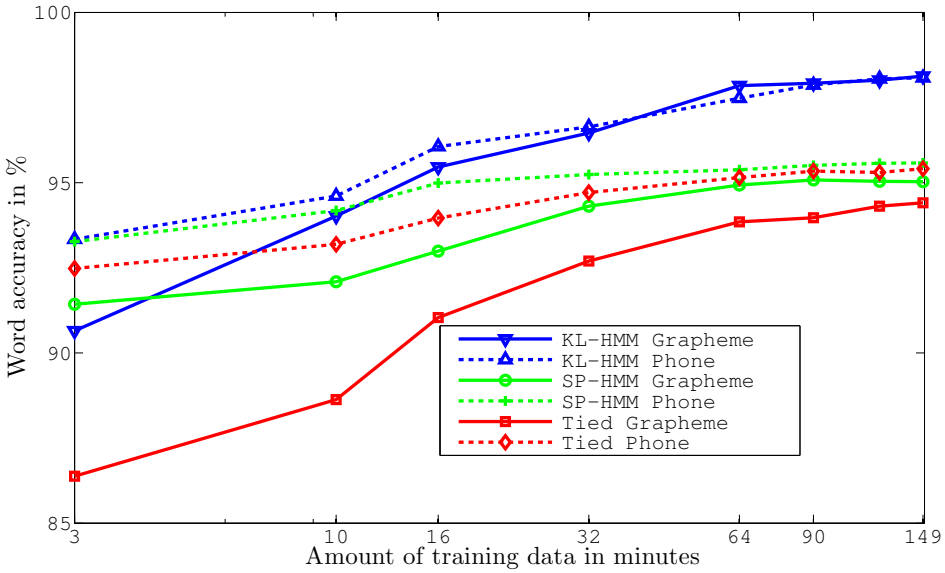

Figure 6.1 – Comparison of various probabilistic lexical modeling based systems with increasing amount of target domain training data on the HIWIRE non-native accented speech recognition task

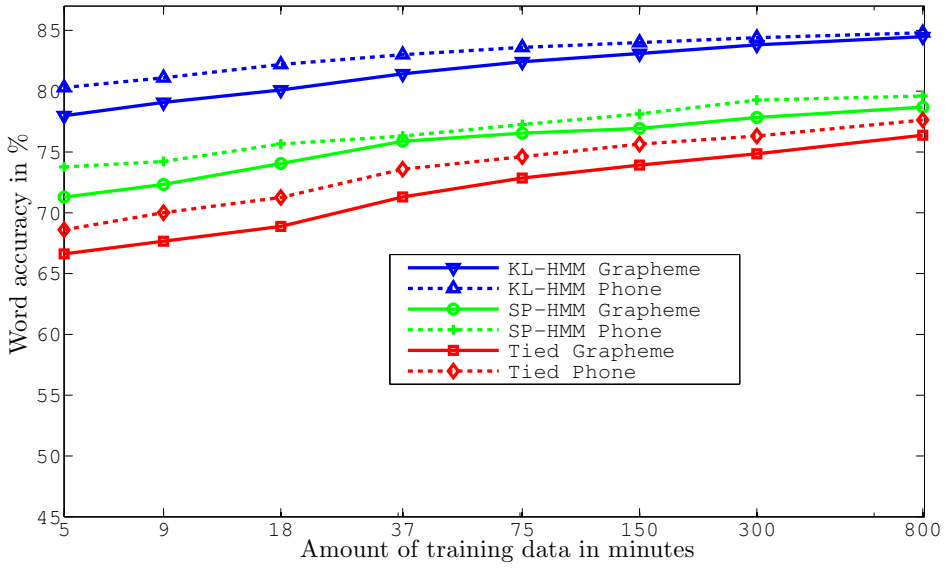

Figure 6.2 – Comparison of various probabilistic lexical modeling based systems with increasing amount of target language training data on the Greek ASR task

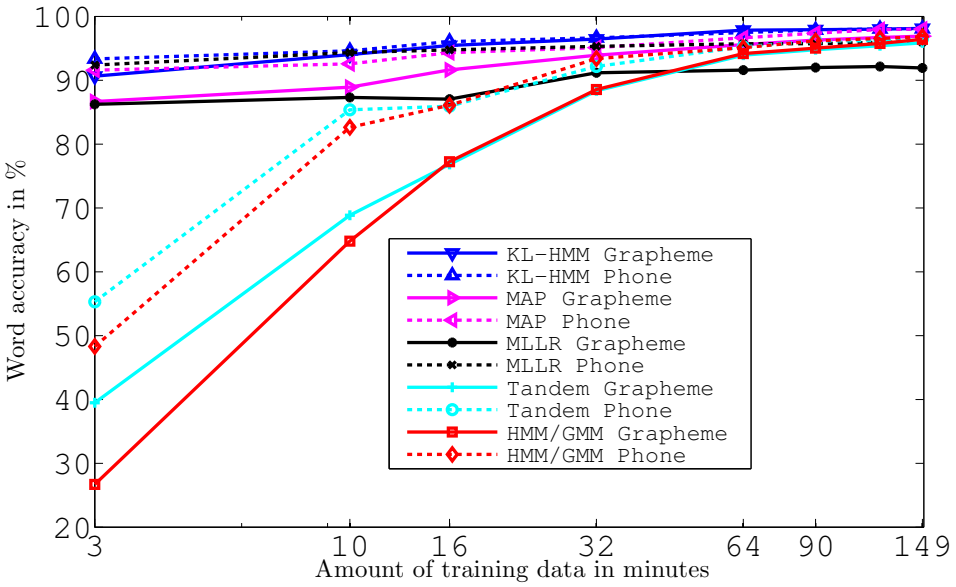

Figure 6.3 – Comparison of the phone-based and grapheme-based KL-HMM systems against the acoustic model adaptation based systems and the standard HMM/GMM system with increasing amount of target domain training data on the HIWIRE non-native accented speech recognition task

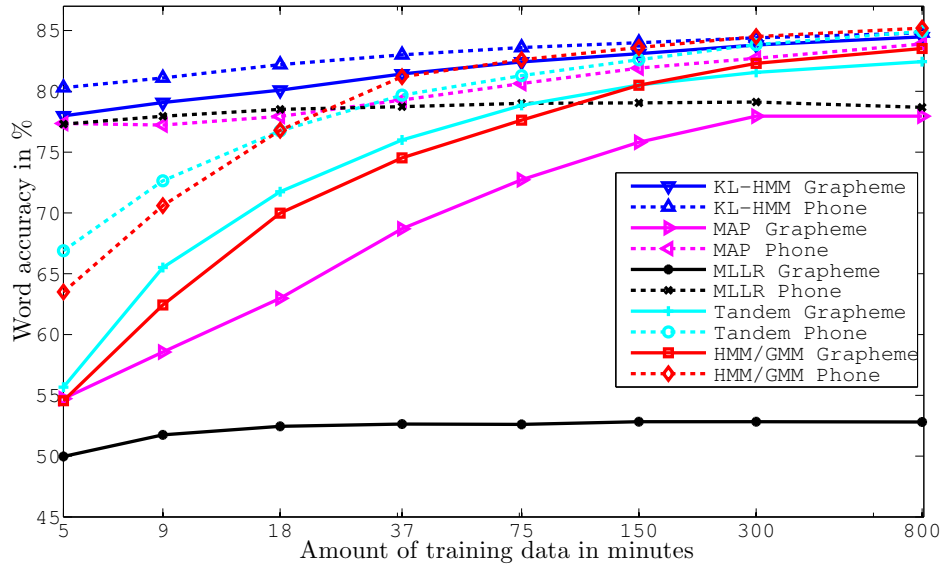

Figure 6.4 – Comparison of the phone-based and grapheme-based KL-HMM systems against the acoustic model adaptation based systems and the standard HMM/GMM system with increasing amount of target language training data on the Greek ASR task

et al., 2003, Stüker, 2008a] show that the use of multilingual grapheme models across languages does not appear evident.

3. Monolingual HMM/GMM systems and acoustic model adaptation based systems with the shared unit set (i.e., on HIWIRE task) that exploit multilingual speech tend to converge with the increase in acoustic resources.
4. Compared to the HMM/GMM approach, the Tandem approach is beneficial mainly in low acoustic resource conditions.
5. Comparing MAP and MLLR approaches, it can be observed that MLLR is better than MAP mainly in very low acoustic resource conditions.

As mentioned in Section 6.1.1, it is possible to directly decode the HIWIRE test set using language-independent acoustic and lexical models without any adaptation. The performance on the HIWIRE task for the KL-HMM, SP-HMM, Tied-HMM and the language-independent HMM/GMM systems is given in Table 6.6. The lexical model for the KL-HMM, SP-HMM and Tied-HMM systems is trained on the SpeechDat(II) English data. It can be observed that for both phone and grapheme subword units the KL-HMM system performs better than the SP-HMM, Tied-HMM and LI HMM/GMM systems. Also, it is interesting to note that irrespective of the subword units used, the performance of all the probabilistic lexical model based systems (that use context-independent phones as acoustic units) is better than that of the LI HMM/GMM system (that uses context-dependent phones as acoustic units).

Table 6.6 – Performance in terms of word accuracy on the HIWIRE test set using system trained on the SpeechDat(II) data. The LI HMM/GMM system refers to the multilingual HMM/GMM system trained on the language-independent (LI) data

System	Grapheme	Phone
KL-HMM	90.0	94.0
SP-HMM	87.3	93.2
Tied-HMM	86.0	91.6
LI HMM/GMM	84.2	91.3

6.2.2 Scottish Gaelic ASR

The performance on the test set of the Scottish Gaelic corpus for the KL-HMM, SP-HMM, Tied-HMM, Tandem and HMM/GMM systems for the *orthography-based* and *knowledge-based* grapheme lexica is given in Table 6.7. The MAP system was not investigated for the *knowledge-based* lexicon due to the mismatch between the acoustic unit set and the lexical unit set. It can be observed that the systems using the *knowledge-based* grapheme lexicon perform better than the systems using the *orthography-based* grapheme lexicon. This shows that integrating orthographic knowledge specific to the language in a grapheme lexicon can help in improving the performance of the grapheme-based ASR system. The KL-HMM systems perform better than all other systems. The Tandem system performs better than the HMM/GMM system. Furthermore, the MAP, SP-HMM and Tied-HMM systems perform worse than the Tandem and

HMM/GMM systems. Finally, in the case of the *orthography-based* lexicon, the MAP system is not able to capitalize on the language-independent data.

Table 6.7 – Performance in terms of word accuracy on the Gaelic test set for the various crossword context-dependent ASR systems.

System	*Orthography-based* lexicon	*Knowledge-based* lexicon
KL-HMM RKL	67.9	72.7
SP-HMM	52.0	56.7
Tied-HMM	54.5	59.7
MAP	55.1	–
Tandem	66.5	69.9
HMM/GMM	64.2	68.0

6.2.3 Analysis

From the experiments presented earlier in this section, it can be observed that despite using exactly the same acoustic model, the performance trends of the various probabilistic lexical modeling approaches KL-HMM, SP-HMM and Tied-HMM are different. The KL-HMM system performs better than the deterministic lexical model based systems in both under-resourced and well resourced conditions. The SP-HMM and Tied-HMM systems show gains over the deterministic lexical model based systems mainly in under-resourced conditions (see Tables 6.4 and 6.5).

In Section 3.5, we discussed the similarities and dissimilarities between the probabilistic lexical modeling approaches and contrasted them to the conventional HMM-approach. More specifically, we pointed out that the Tied-HMM, SP-HMM and KL-HMM KL systems are close to the conventional HMM-based ASR approach where the lexical model serves as the reference when matching the acoustic model and the lexical model or estimating the local score. While, in the KL-HMM RKL system, the acoustic model serves as the reference. In addition, KL-divergence is a discriminative local score. In this chapter, the KL-HMM system was based on local score $S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$. So, we attribute this superiority of the KL-HMM system to its ability to give more importance to the acoustic model evidence than the lexical model evidence through the use of the local score $S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$.

In order to ascertain the reason for difference in performance trends among the various probabilistic lexical modeling approaches, we conducted a study on the HIWIRE task with the 150 minute target data condition where the lexical model is trained using the KL-HMM RKL approach and decoding is performed with different local scores, namely, $S_{KL}(\mathbf{y}_i, \mathbf{z}_t)$, $S_{SKL}(\mathbf{y}_i, \mathbf{z}_t)$, $S_{tied}(\mathbf{y}_i, \mathbf{v}_t)$ and $S_{SP}(\mathbf{y}_i, \mathbf{z}_t)$. The study was conducted for both grapheme-based and phone-based systems. The results of this study are given in Table 6.8.

It can be observed that decoding with KL-divergence based local scores $S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$,

Table 6.8 – Comparison across different local scores used during decoding. The system trained with the KL-HMM RKL approach is decoded with all the other local scores.

Local score for decoding	grapheme	phone
$S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$	98.1	98.1
$S_{KL}(\mathbf{y}_i, \mathbf{z}_t)$	97.8	97.6
$S_{SKL}(\mathbf{y}_i, \mathbf{z}_t)$	98.1	98.1
$S_{SP}(\mathbf{y}_i, \mathbf{z}_t)$	96.5	96.7
$S_{tied}(\mathbf{y}_i, \mathbf{z}_t)$	97.3	97.1

$S_{SKL}(\mathbf{y}_i, \mathbf{z}_t)$ and $S_{KL}(\mathbf{y}_i, \mathbf{z}_t)$ results in better performance compared to decoding with $S_{SP}(\mathbf{y}_i, \mathbf{z}_t)$ and $S_{tied}(\mathbf{y}_i, \mathbf{v}_t)$ local score, ascertaining the fact that KL-divergence is a better local score compared to scalar product. Furthermore, decoding with $S_{KL}(\mathbf{y}_i, \mathbf{z}_t)$, $S_{SP}(\mathbf{y}_i, \mathbf{z}_t)$ and $S_{tied}(\mathbf{y}_i, \mathbf{v}_t)$ yields lower performance than decoding with $S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$. However, decoding with $S_{SKL}(\mathbf{y}_i, \mathbf{z}_t)$, that gives equal importance to the acoustic and lexical model yields performance similar to $S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$. It can also be noted that decoding the KL-HMM lexical model with $S_{SP}(\mathbf{y}_i, \mathbf{z}_t)$ and $S_{tied}(\mathbf{y}_i, \mathbf{v}_t)$ results in better performance compared to the SP-HMM trained and Tied-HMM trained lexical model, respectively (see Table 6.4). This indicates that the KL-HMM approach with the local score S_{RKL} is yielding a better lexical model compared to the SP-HMM or Tied-HMM approaches. The results are consistent with the analysis of lexical model parameters presented in Chapter 4, where it was observed that as the context of subword units is increased, the lexical model parameters of the KL-HMM approach captured the one-to-many lexical-to-acoustic units relationships better than the Tied-HMM approach.

6.2.4 Comparisons with the Literature

In the literature, there are studies that have been reported on the HIWIRE task [Segura et al., 2007, Gemello et al., 2007]. Despite using the same adaptation and test sets, the studies reported in this thesis and the literature differ in terms of the sampling frequency of speech data, type and amount of the out-of-domain data used. First, we compare with studies in which no kind of adaptation was performed.

- In [Segura et al., 2007], the TIMIT trained monophone HMM/GMM system without adaptation was found to achieve a performance of 91.4% word accuracy.
- In [Gemello et al., 2007], the monophone hybrid HMM/ANN system using an MLP trained on the TIMIT, WSJ0, WSJ1 and Vehiclus-ch0 corpora was found to achieve a performance of 90.5% word accuracy. Furthermore, the monophone hybrid HMM/ANN system using MLP trained on the LDC Macrophone and SpeechDat Mobile corpora on the HIWIRE speech downsampled to 8kHz was found to achieve performance of 88.4% word accuracy.

As shown in Table 6.9, the phone-based KL-HMM system performs better than the approaches reported in the literature and grapheme-based KL-HMM system performs comparable to the

approaches reported in the literature. It can also be observed from Tables 6.9 and 6.6 that the phone-based LI HMM/GMM system performs similarly to the above mentioned systems from the literature, whereas the grapheme-based LI HMM/GMM system performs worse.

Table 6.9 – Comparison of word accuracies on the HIWIRE test set without any adaptation.

System	Out-of-domain data	Sampling frequency	Performance
HMM/GMM	TIMIT	16kHz	91.4
Hybrid HMM/ANN	TIMIT, WSJ0, WSJ1, Vehiclus-ch0	16kHz	90.5
Hybrid HMM/ANN	LDC Macrophone, SpeechDat Mobile	8kHz	88.4
KL-HMM Grapheme	SpeechDat(II)	8kHz	90.0
KL-HMM Phone	SpeechDat(II)	8kHz	94.0

There are also studies on HIWIRE that report results with acoustic model adaptation where 150 min of HIWIRE adaptation data is used.

- In [Segura et al., 2007], it has been found that the TIMIT trained HMM/GMM system with MLLR adaptation achieves performance of 97.25% word accuracy.
- In [Gemello et al., 2007], linear hidden network (LHN) based adaptation in the hybrid HMM/ANN framework achieved performance of 98.2% on 16kHz sampled HIWIRE data. MLP trained on data from TIMIT, WSJ0, WSJ1 and Vehiclus-ch0 was adapted on HIWIRE data using LHN.

As shown in Table 6.10, the hybrid HMM/ANN system using LHN based adaptation performs similarly to the phone-based and grapheme-based KL-HMM systems.

Table 6.10 – Comparison of word accuracies on the HIWIRE test set with adaptation

System	Out-of-domain data	Sampling frequency	Performance
MLLR	TIMIT	16kHz	97.25
LHN	TIMIT, WSJ0, WSJ1, Vehiclus-ch0	16kHz	98.2
KL-HMM Grapheme	SpeechDat(II)	8kHz	98.1
KL-HMM Phone	SpeechDat(II)	8kHz	98.1

Furthermore, for both grapheme and phone subword units, the performance of ASR systems on the HIWIRE task (150 min adaptation data case) from this chapter is better than the performance in the previous chapter. The results indicate that using language-independent resources from multiple languages is more advantageous compared to using resources from only one language. In [Imseng et al., 2011], on the HIWIRE task, the performance of the grapheme-based KL-HMM system using low amounts of HIWIRE adaptation data (3min, 10min) was significantly worse than that of the phone-based KL-HMM system. In this work

the gap is significantly reduced as the lexical model parameters trained on SpeechDat(II) English are adapted using HIWIRE adaptation data whereas in [Imseng et al., 2011], lexical model parameters were directly trained on limited HIWIRE adaptation data.

In the case of the Greek task, as previously mentioned phone-based KL-HMM, MLLR, MAP and HMM/GMM systems reported in [Imseng et al., 2012b] and [Imseng, 2013, Figure 4.3 in Page 59 and Figure 4.4 in Page 60] have been used as reference. However, the phone-based Tandem systems reported in [Imseng, 2013] and this chapter differ. Unlike [Imseng, 2013], in our studies the dimensionality of the Tandem features was either 117 (all the dimensions) or 39 (same as the dimension of standard cepstral feature vector). The dimension of features was tuned on the development set for each of the training conditions. We found dimensionality reduction to be beneficial, especially in the low acoustic resource conditions. For example, on the 5 min acoustic resource case, performance of phone-based Tandem system reported in [Imseng, 2013] was 30.2% word accuracy, whereas in this chapter with reduced feature dimensionality we achieved 66.9% word accuracy.

In our previous study on Scottish Gaelic ASR [Rasipuram et al., 2013a], the knowledge-based grapheme lexicon that tagged word beginning and end graphemes was used and word-internal context-dependent graphemes were modeled. The KL-HMM and HMM/GMM systems achieved a word accuracy of 72.8% and 64.8%, respectively. In this work, the same knowledge-based grapheme lexicon was used but without any word begin and end tags. As a result, the total number of grapheme subword units is smaller. Furthermore, in this thesis we modeled crossword context-dependent subword based systems. As it can be seen from Table 6.7, the knowledge-based HMM/GMM system yields an absolute improvement of 3.2% WER compared to the previous work and the grapheme KL-HMM system achieves performance comparable to that of the previous study.

6.3 Summary

Our studies in this chapter showed that with probabilistic lexical modeling, especially using the KL-HMM approach, ASR systems can be rapidly developed for new languages and domains by training the language or domain independent acoustic model and learning the grapheme-to-phone relationship on small amount of target language or domain data. In doing so, we not only address the lack of adequate acoustic resource (speech data with transcription) problem but also the lack of lexical resource (phone pronunciation lexicon) problem.

7 Zero-Resourced ASR

As discussed earlier, standard ASR systems rely on acoustic resources (or transcribed speech), lexical resources (or the phone pronunciation lexicon), and text data to achieve state-of-the-art performance. In Chapter 5, we focussed on building ASR systems in lexical resource constrained scenarios whereas in Chapter 6, we focussed on building ASR systems in both acoustic and lexical resource constrained scenarios. More specifically, in the previous two chapters, we showed that in the framework of the proposed grapheme-based ASR approach, the acoustic model can be trained on domain-independent or language-independent resources and the lexical model alone on target domain or language resources.

In this chapter, we will show that the lexical model can be knowledge driven and ASR systems could be developed for a new language without using any acoustic and lexical resources from the language, i.e., (near) zero-resourced [1] ASR system. In the case where untranscribed speech data from the target language is available then an approach for unsupervised adaptation of the lexical model parameters is proposed. The potential of the proposed approach is studied on the Greek ASR task, that was also used in the previous chapter.

7.1 Related Work

There have been attempts in the past to build ASR systems without using any acoustic resources from the target language through cross language transfer of acoustic models [Schultz and Waibel, 2001a, Lööf et al., 2009, Vu et al., 2010]. In cross-language transfer, first a mapping between phones of source language(s) and target language is defined. In [Schultz and Waibel, 2001a], two techniques for cross-language mapping of phones were proposed. In the first method, mapping is developed manually and is based on knowledge of phones in source lan-

1. The use of the term zero-resourced is debatable. For example, in the JHU workshop [Jansen et al., 2013a], zero-resourced speech technologies referred to systems that operate without the expert provided linguistic knowledge and transcriptions. However, untranscribed speech data was assumed to be available. In [Besacier et al., 2014], any language that lacks one or more resources required to build an ASR system is referred to as under-resourced language. In this thesis, by zero-resourced we meant expert-provided linguistic knowledge, phone lexical resources, and transcription speech data are not available.

guage(s) and target language. In the second method, mapping is derived automatically using a small amount of target language acoustic data. It was shown that language-independent acoustic models trained on multiple languages perform better for cross-language transfer than acoustic models trained on a language [Schultz and Waibel, 2001a, Vu et al., 2010]. Given the mapping, the phone pronunciation lexicon, acoustic models of source language(s) and language models, the decoding of the target language speech data is possible, albeit with high error rate [Schultz and Waibel, 2001a, Lööf et al., 2009, Vu et al., 2010].

In the case where untranscribed speech data from the target language is available, then unsupervised acoustic model training/adaptation can be performed to improve the acoustic models. Typically, cross-language transfer is used as the starting point for unsupervised adaptation. That is, the cross-language acoustic models are used to recognize target language data. The recognized transcriptions with speech data are used in conventional acoustic model training or adaptation techniques like MAP or MLLR to retrain or update the models. Also, confidence measures are used to select or weight the utterances for effective use. Unsupervised training or adaptation of crosslingual or multilingual models can result in substantial performance improvements for ASR [Lööf et al., 2009, Vu et al., 2010].

Most of the ASR approaches that did not use any acoustic data from the target language assume that phone lexica in all source language(s) and target language are available. In addition to the acoustic data, if the target language has no available phone lexical resources, the above mentioned approaches are not directly applicable.

Other approaches that focussed on building speech applications from untranscribed speech data without any phone lexical resources are mostly based on the sound pattern structure of speech.

- In [Park and Glass, 2005], acoustic patterns in speech were discovered by matching subsequences between pairs of utterances.
- In [Jansen and Church, 2011, Jansen et al., 2013b], first the automatically discovered examples of word clusters are used to train whole word HMMs. Then, the word HMMs are clustered across HMM states to produce context-independent subword unit models.
- In [Gish et al., 2009, Siu et al., 2014], untranscribed speech is transcribed into self-organized units and the HMM training is optimized over both the parameter space and the transcription sequence space.

However, the approaches have been validated only on topic classification and spoken term detection tasks.

In this chapter, we show that the proposed grapheme-based ASR approach can be extended to zero-resourced conditions i.e., for languages where acoustic and lexical resources are not available.

7.2 Proposed Zero-Resourced ASR Approach

In this chapter, we present a zero-resourced grapheme-based ASR approach and assume that no transcribed speech data, and no phone lexical resources of the target language are available. However, we assume that we have knowledge of the possible words in the language and therefore its character or grapheme set is also known.

7.2.1 Knowledge-based Lexical Model Parameters

In the proposed zero-resourced approach various components i.e., the acoustic model, the lexical model and the pronunciation lexicon required to build a probabilistic lexical based system are obtained in the following way:

- Acoustic model: It is an ANN trained on language-independent data from multiple languages. The acoustic units or the outputs of ANN represent multilingual phones. Given the acoustic model, acoustic unit probability sequence for the test utterance is estimated.
- Pronunciation lexicon: The grapheme lexicon and the grapheme subword unit set are obtained from the list of possible words in the target language. The lexical units are context-independent graphemes of the target language.
- Lexical Model: An initial knowledge-based lexical model parameter set $\Theta_l^{kn} = \{\{\mathbf{y}_i\}_{i=1}^I\}$ is defined in the following way:
  1. Associate each grapheme lexical unit to one or more phone outputs of the ANN.
  2. The knowledge-based lexical model parameter set $\Theta_l^{kn} = \{\{\mathbf{y}_i\}_{i=1}^I\}$ is defined in the following way: if a lexical unit l^i is mapped to R of the D acoustic units where $R << D$ then

$$\forall d \in \{1,\ldots,D\} \quad y_i^d = P(a^d|l^i) = \begin{cases} \frac{s}{R}, & \text{if } l^i \mapsto a^d\,; \\ \frac{1-\frac{s}{R}}{I-R}, & \text{otherwise.} \end{cases} \tag{7.1}$$

     where I is the total number of acoustic units and s is chosen such that $s \geq 0.5$.
  3. Each context-independent grapheme is modeled as a three-state HMM.

The block diagram of the proposed zero-resourced ASR system is illustrated in Figure 7.1.

7.2.2 Unsupervised Adaptation of Lexical Model Parameters

In addition to the word list, if we assume that untranscribed speech data from the target language ($\{X(n)\}_{n=1}^N$) is available, then the knowledge-based lexical model parameter set can be updated in an unsupervised manner. More specifically, we replace the strong top-down constraints used during lexical model training in the form of transcribed speech data with weak top-down constraints obtained from grapheme sequence decoder. The grapheme sequence decoder is an ergodic HMM constructed with all graphemes of the target language and their knowledge-based lexical model parameters. The transition probabilities of the ergodic HMM are derived from an n-gram grapheme language model trained on the grapheme lexicon

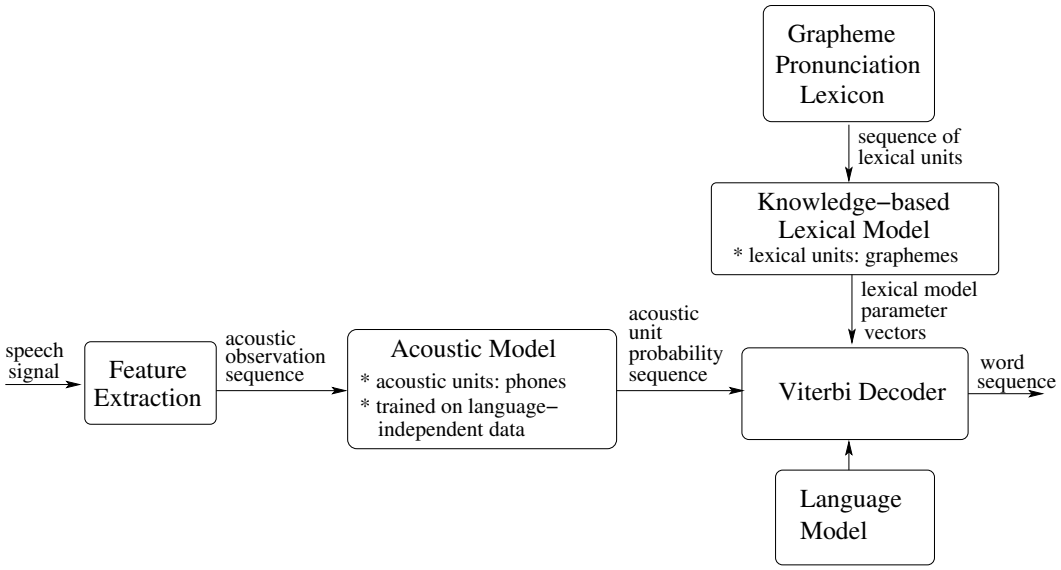

Figure 7.1 – Block diagram of the proposed zero-resourced ASR system

obtained from the word list. In this thesis, we used a bi-gram grapheme language model. Given the ergodic grapheme HMM, the unsupervised adaptation is performed as shown in Figure 7.2 and involves the following three steps:

1. The acoustic unit probability sequences $\{Z(n)\}_{n=1}^{N}$ are computed given the acoustic model and the acoustic feature observations $\{X(n)\}_{n=1}^{N}$ of the target language data.

2. The resulting acoustic unit probability sequences are decoded using the grapheme sequence decoder to generate grapheme level transcriptions of the speech data. In the first iteration, the ergodic HMM in the grapheme sequence decoder uses knowledge-based lexical model parameters.

3. The decoded grapheme transcriptions and their acoustic unit posterior probability estimates $\{Z(n)\}_{n=1}^{N}$ are used to update the lexical model parameters.

The lexical model parameter set can be updated iteratively by repeating the second and third steps.

The unsupervised lexical model parameter estimation can be extended to context-dependent subword units in the following way: use the decoded context-independent grapheme sequences to obtain context-dependent grapheme sequences for all the utterances in the training data; update/train the lexical model parameters of context-dependent graphemes using acoustic unit probability sequences $\{Z(n)\}_{n=1}^{N}$ and context-dependent grapheme sequences.

The grapheme-based zero-resourced ASR approach proposed in this chapter can also be extended to phone subword units if phone lexical resources are available. However, in this chapter we always consider that phone lexical resources are not available.

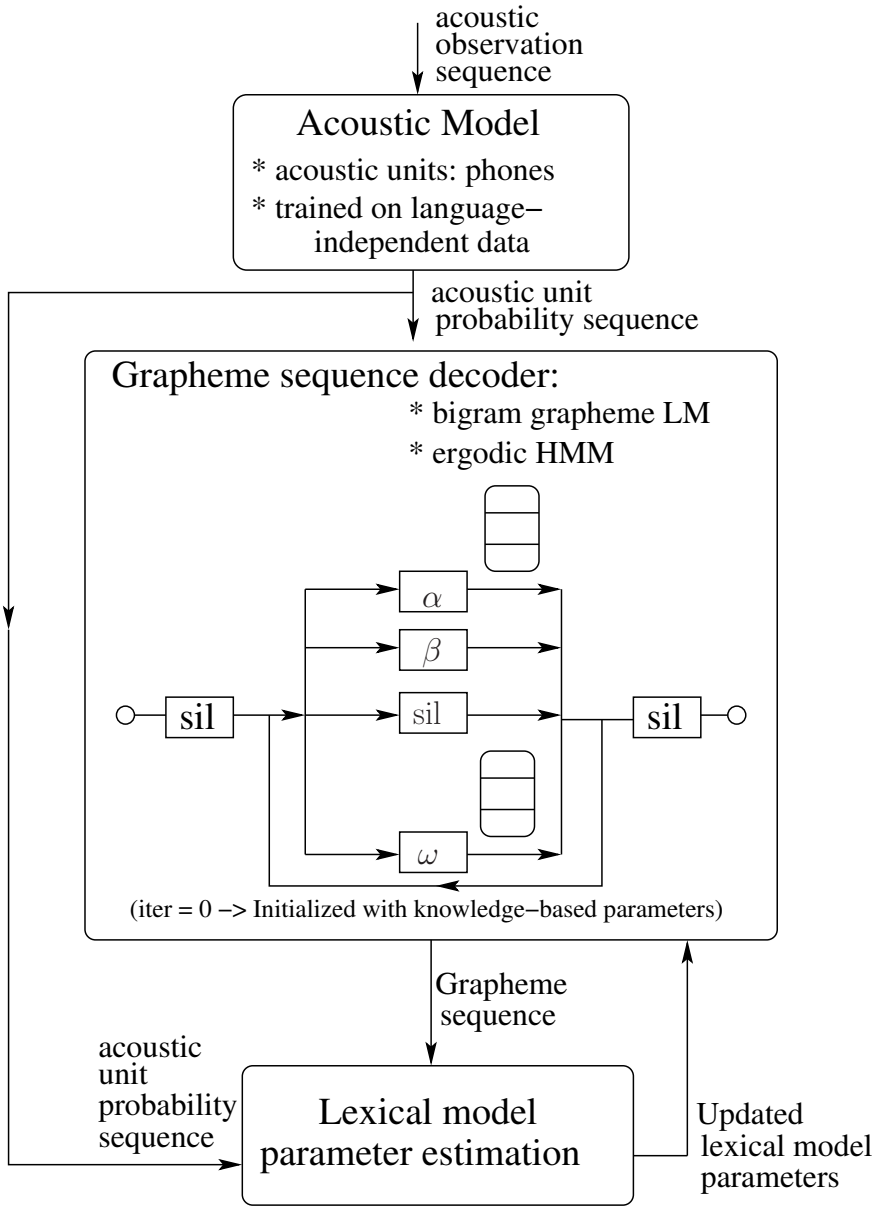

Figure 7.2 – Unsupervised lexical model parameter estimation

7.3 Experimental Setup and Results

To evaluate the proposed approach we consider Greek as the target language for which we are interested to build an ASR system, but we assume that acoustic and lexical resources are not available. We revisit the Greek SpeechDat(II) ASR task presented in the previous chapter and consider five other European languages from the SpeechDat(II) corpus namely British English (EN), Swiss French (SF), Swiss German (SZ), Italian (IT) and Spanish (ES) as language-independent resources.

Acoustic model: The three-layer multilingual MLP used in the previous chapter trained on the above mentioned five languages of the SpeechDat(II) corpus is used as an acoustic model. Additionally, we also trained another multilingual MLP but with five-layers using the data from the same five languages of the SpeechDat(II) corpus. The input features and output units of the five-layer MLP are the same as that of the three-layer MLP. The size of each hidden layer of the five-layer MLP was set to 2000 units.

Grapheme sequence decoder: The transition probabilities of the ergodic HMM in the grapheme sequence decoder are obtained from the bi-gram language model trained on the available word list. We use the list of 35146 words from the SpeechDat(II) Greek corpus for this purpose. The bi-gram grapheme language model built from the orthography of 35146 words has a perplexity of 9. Also, the grapheme lexicon built using 35146 words of the SpeechDat(II) Greek corpus is used as the pronunciation lexicon in all the experiments.

Lexical model: We present this study on KL-HMM RKL systems. More specifically, we evaluate the following two systems:

1. *RKL_MLP-3*: The KL-HMM RKL system using the three-layer multilingual MLP as the acoustic model
2. *RKL_MLP-5*: The KL-HMM RKL system using the five-layer multilingual MLP as the acoustic model

There are two main reasons to select the local score S_{RKL}. From a training perspective, it was chosen because one-to-many G2P relationships are better captured with the local score S_{RKL} than with other local scores (according to the analysis presented in Section 4.4.2). From a decoding perspective, as discussed in Section 3.5.2, the local score S_{RKL} has the capability to give more importance to acoustic model evidence than to lexical model evidence. Since in the zero-resource ASR approach the lexical model can be weak, it is preferable to give more importance to acoustic model evidence during decoding.

Evaluation: For evaluating the systems we report grapheme accuracy (GA) on the training set and word accuracy (WA) on the test set of the Greek SpeechDat(II) corpus. In a real world zero-resourced ASR scenario, it may not be possible to compute the GA on the training set as reference transcriptions are not available. However, since this is a simulated study (i.e., reference transcriptions are available), we report grapheme accuracy on the training set. To clarify, we did not tune the insertion penalty and the language scale factor while reporting the GA on the training set. As mentioned in the previous chapter, two optimistic language models trained from sentences in the development set and test set are used during decoding.

7.3.1 Evaluation of Knowledge-based Lexical Model Parameters for ASR

In this case only the list of possible words from the target language is assumed to be available. The knowledge-based lexical model parameter set is defined following the procedure given in Section 7.2.1. Columns 1 and 2 of Table 7.1 provide the grapheme-to-multilingual phone

map used in defining the knowledge-based lexical model parameter set. Empirically it was observed on the development data that the value of s above 0.7 did not significantly effect the grapheme sequence decoded using the ergodic HMM. So, we only present ASR results for the case with $s = 0.8$.

Table 7.1 – Greek graphemes with their transliterated format (Trans.), knowledge-based G2P map and automatic G2P map learned by unsupervised adaptation of lexical model parameters

Grapheme (Trans.)	Knowledge-based map	Unsupervised map	
		RKL_MLP-3	*RKL_MLP-5*
α (a)	a, a:	a	a
β (b)	b, v	b, v	b, v
χ (ch)	c, x	x, R, C	x, R, C
δ (d)	d, D	d, D	d, D
ϵ (e)	e	e, E	e
ϕ (f)	f	f, s	f
γ (g)	g, G, j	g, j	g, j
η (h)	E:, i	i	i
ι (i)	i, i:	i	i
κ (k)	k	k	k
λ (l)	l	l	l
μ (m)	m	m, n	m
ν (n)	n	n	n
o (o)	o	o	o
π (p)	p	p	p
ψ (ps)	s	s, S	s, S
ρ (r)	r	R, r	r, rr
σ (s)	s	s	s
θ (th)	T	T, s	T
τ (t)	t	t	t
ω (w)	o, O:	u	o
ξ (x)	x	k	k
υ (y)	i, y, y:	i	i
ζ (z)	dz, z	s, z, Z	z

The performance in terms of GA and WA of the two systems when the zero-resourced ASR approach uses knowledge-based lexical model parameters is given in Table 7.2. The systems using the five-layer MLP as the acoustic model perform better than the systems using the three-layer MLP as the acoustic model. The results indicate that the proposed approach can be used to build ASR systems in zero-resourced setup with minimal knowledge. This is interesting because the proposed zero-resourced ASR approach can serve as a practical starting point while building ASR systems for new languages without any acoustic and lexical resources.

Table 7.2 – Grapheme accuracy (GA) on the training set and word accuracy (WA) on the test set of the zero-resourced KL-HMM systems. Lexical units are context-independent graphemes

System	GA on training set	WA on test set
RKL_MLP-3	43.3	50.4
RKL_MLP-5	45.2	54.4

7.3.2 Evaluation of Unsupervised Adaptation of Lexical Model Parameters

In this case, untranscribed Greek speech data is also assumed to be available along with the list of words from the target language. The speech data corresponding to the training set of the Greek SpeechDat(II) corpus is used for this purpose. The speech data is forward passed through the acoustic model to obtain acoustic unit probability sequences. The knowledge-based lexical model parameter set is updated in the unsupervised training procedure described earlier in Section 7.2.2.

The performance in terms of GA and WA of the systems when the lexical model parameters are adapted in an unsupervised way is reported in Table 7.3. For the systems reported in this table lexical units are context-independent graphemes. The results show that the unsupervised adaptation of the knowledge-based lexical model parameters significantly improves the performance of the *RKL_MLP-3* system compared to the performance of the system using knowledge-based lexical model parameters given in Table 7.2. We updated the parameters of the lexical model iteratively. However, the unsupervised adaptation converged in just one iteration as GA and WA on train and test sets do not change significantly. This can be due to the relatively small number of lexical model parameters ($3 * 25 * 117$). Similarly to the case of knowledge-based lexical model parameters, systems using the five-layer MLP as acoustic model perform better than the respective systems using the three-layer MLP as acoustic model.

The columns 1, 3 and 4 of Table 7.1 provide the graphemes, and the G2P map obtained from the updated lexical model parameters of the systems *RKL_MLP-3* and *RKL_MLP-5*, respectively. There are differences in the G2P map captured by the lexical model parameters compared to the knowledge-based map (given in column 2). For example, the grapheme [α] is not mapped to phone /a:/ after the unsupervised adaptation procedure; the grapheme [ψ] is mapped to /S/ in addition to /s/ after the unsupervised adaptation procedure. The table also shows that the G2P relationship captured by the lexical model parameters of the system *RKL_MLP-5* is better than the G2P relationship captured by the lexical model parameters of the system *RKL_MLP-3*.

The performance in terms of GA and WA of the two KL-HMM systems with context-dependent graphemes as lexical units is given in Table 7.4. The context-independent grapheme transcriptions of the respective systems given in Table 7.3 are turned to context-dependent grapheme transcriptions, as done regularly when training context-dependent subword unit based ASR systems. Results show that the unsupervised adaptation of the lexical model parameters of context-dependent lexical units improves the performance of the *RKL_MLP-3* and *RKL_MLP-5*

Table 7.3 – Grapheme accuracy (GA) on the training set and word accuracy (WA) on the test set of the KL-HMM systems when the lexical model parameters are adapted in an unsupervised way. Lexical units are context-independent graphemes

System	GA on training set	WA on test set
RKL_MLP-3	45.2	66.3
RKL_MLP-5	47.8	70.5

systems compared to the respective systems using context-independent lexical units. It was again observed that only one iteration of unsupervised training was sufficient and subsequent iterations did not improve the ASR performance on the test set. Hence, we did not report those results here.

Table 7.4 – Grapheme accuracy (GA) on the training set and word accuracy (WA) on the test set of the KL-HMM systems when the lexical model parameters are adapted in an unsupervised way. Lexical units are context-dependent graphemes

System	GA on training set	WA on test set
RKL_MLP-3	48.2	68.1
RKL_MLP-5	50.2	72.0

7.4 Summary

In this chapter, we have shown that the lexical model can be knowledge driven in the proposed grapheme-based ASR approach. Therefore, ASR systems for a new language could be developed without using any acoustic and lexical resources from the language. More specifically, the acoustic model or an ANN is trained on language-independent resources. Furthermore, if untranscribed speech data from the target language is available, the knowledge-based lexical model parameters can be adapted in an unsupervised manner using graphemic constraints learned from the available word list.

8 Acoustic Data-Driven G2P Conversion

In the previous chapters, we focussed on grapheme-based ASR in the framework of probabilistic lexical modeling in various resource constrained ASR scenarios. In the proposed grapheme-based ASR approach, the acoustic model models the relationship between phones and acoustic features, while the lexical model models a probabilistic relationship between graphemes and phones. In this chapter, we show that the G2P relationship captured in the lexical model parameters can be exploited together with the sequence information in the orthographic transcription of the word to extract pronunciation models/variants.

In the following section, we will first present a brief overview of G2P approaches proposed in the literature. Then, the proposed G2P approach is presented in detail in Section 8.2 and the experimental studies are presented in Section 8.3.

8.1 Related Work

Automatic G2P conversion techniques can be broadly classified into rule-based approaches and data-driven approaches. Rule-based G2P conversion approaches are typically formulated in the framework of finite state automata[Kaplan and Kay, 1994]. The primary advantage of rule-based approaches is that they can provide complete coverage. However, the two main drawbacks of rule based approaches are: (1) Natural languages exhibit irregularities. Therefore, it is necessary to cross-check if the rules are applicable to all the entries. Often rule-based G2P systems also need an exception list. (2) Design of rules requires specific linguistic skills that may not be always available. The unavailability of linguistic skills in many languages rules out the use of rule-based G2P conversion for lexicon generation.

Lexical data-driven approaches for G2P conversion are based on the fact that given enough examples (or the seed lexicon) it should be possible to predict the pronunciation of an unseen word. The first step in most of the automatic G2P conversion approaches is the alignment of training data constituting sequences of graphemes and their corresponding sequences of phonemes. Given the alignments, a decision tree [Pagel et al., 1998] or a neural network [Se-

jnowski and Rosenberg, 1987] can be trained to learn the G2P relationship from the training data. In both decision-tree and neural-network based G2P conversion, the prediction of an output phoneme is based on the context of the current grapheme. Also, the decision for each grapheme takes place before proceeding to the next one. Therefore, these methods are based on local classification.

The G2P conversion problem has also been approached through probabilistic and sequence classification methods. Given a sequence of graphemes $G = \{g_1, g_2, \ldots, g_N\}$, the most likely sequence of phonemes $S^* = \{s_1, s_2, \ldots, s_M\}$ can be found by,

$$S^* = \arg\max_S P(S|G) \tag{8.1}$$

In [Taylor, 2005], the G2P conversion problem was formulated in the standard HMM way as:

$$S^* = \arg\max_S P(G|S)P(S) \tag{8.2}$$

where $P(S)$ is the prior probability of a sequence of phonemes, $P(G|S)$ is the likelihood of the grapheme sequence given the phoneme sequence. Each HMM represented one phoneme which can generate up to four graphemes in four different emitting states. After emitting an observation, the path moves to the next state or the exit state. The algorithm finds the most probable sequence of phonemes that could have generated the input grapheme sequence.

In joint multigram or joint n-gram approaches [Bisani and Ney, 2008], a model that represents the joint probability distribution over sequences of graphone (a grapheme-phoneme pair) units is used for G2P conversion,

$$S^* = \arg\max_S P(G, S) \tag{8.3}$$

The parameters of the joint n-gram model are estimated using the EM algorithm on an existing pronunciation lexicon.

In [Wang and King, 2011], the G2P conversion was achieved through conditional random fields that are discriminative models and are capable of global inference. A CRF directly models the conditional probability of phoneme sequence S given grapheme sequence G as in Eqn (8.1).

The pronunciations derived from automatic G2P convertors reflect the ambiguity and variation found in the lexical resources used to train the model. Therefore, their main drawback is that pronunciations or its variants may not reflect the natural phonological variation. For example, this can happen when a G2P convertor trained on native pronunciations is used to extend the vocabulary of a non-native ASR system; or when the new vocabulary has unusual words. To overcome this limitation acoustic samples of words were used to refine expert-provided or G2P-convertor based pronunciations.

- In [Xiao et al., 2007], the parameters of the G2P convertor were adapted using spoken

examples for a name recognition task.

- In [McGraw et al., 2013], the pronunciation variants of words given by the graphone-based G2P approach [Bisani and Ney, 2008] were given pronunciation weights using acoustic samples of words. The approach assumes that an expert provided pronunciation lexicon is available.
- In [Lu et al., 2013], an approach to enlarge the expert phonemic lexicon is proposed where the pronunciations of additional words are generated using their acoustic samples and a trained G2P convertor. More precisely, first a G2P convertor is trained using an expert lexicon. The G2P convertor is used to generate pronunciation variants for new words. The weights for these multiple pronunciations are estimated based on acoustic evidence using the WFST-based EM algorithm. Finally, the acoustic model is updated using the augmented lexicon. The process is repeated until convergence.

As shown in Figure 8.1, the above three G2P conversion approaches rely on a seed lexicon and a G2P convertor. The acoustic samples are used only to weigh or select the alternate pronunciations given by a G2P convertor.

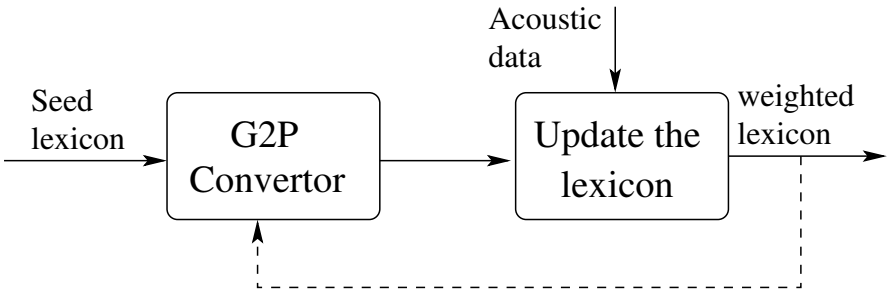

Figure 8.1 – Acoustic data-driven G2P conversion approaches proposed in the literature. The dotted line illustrates that some approaches iterate the G2P conversion process

In this chapter, we propose an acoustic data-driven G2P conversion algorithm that extracts pronunciations using acoustic data and word level transcriptions of the target domain. The approach is not constrained by the availability of acoustic samples of the words for which we are interested to generate phone pronunciations.

8.2 Proposed Approach

One of the key issue involved in the development of a G2P converter is to effectively capture the relationship between graphemes and phones. As discussed in Chapter 4, when using graphemes as lexical units in probabilistic lexical modeling based systems this relationship is captured in the lexical model parameters and is learned from acoustic data of the target domain.

The proposed G2P approach that builds upon this observation consists of two phases: a training phase and a decoding phase. In the training phase a probabilistic lexical model based

system with graphemes as lexical units is trained. Given the lexical model parameters of grapheme subword units and the orthographic transcription of a word, the decoding phase involves inferring a phone sequence. Figure 8.2 illustrates the block diagram of the proposed G2P approach.

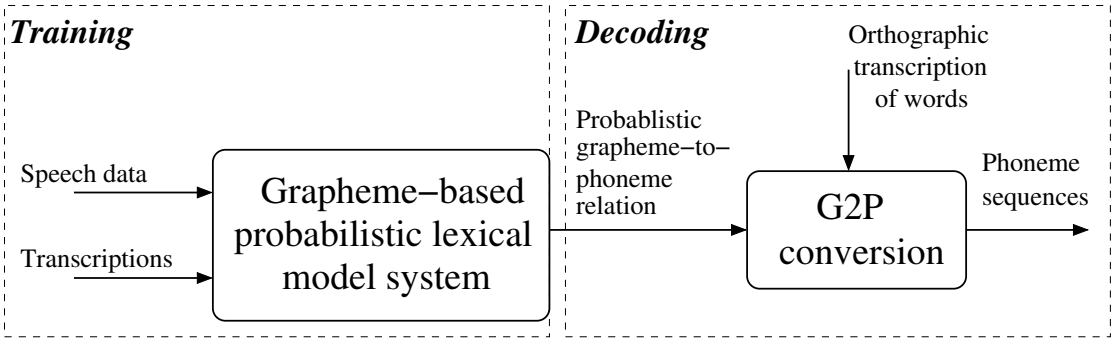

Figure 8.2 – Block diagram of the proposed acoustic data-driven G2P conversion approach.

8.2.1 Training Phase

In the training phase a grapheme-based probabilistic lexical model system is trained. Therefore this phase includes the training of an acoustic model and a lexical model as described in previous chapters.

Acoustic Model: Similar to previous chapters, the acoustic model is a well trained MLP with acoustic units or output classes as context-independent phones; and the acoustic model can be trained either on target-domain or target-language resources if available, or on domain-independent and language-independent acoustic and lexical resources.

Lexical Model: The speech data of the target language or domain is forward passed through the MLP to obtain acoustic unit probability sequences that are then used as feature observations to train an HMM (or the lexical model) with context-dependent graphemes as lexical units.

8.2.2 Decoding Phase

As illustrated in the block diagram of Figure 8.3, the decoding phase involves inference of a phone sequence given lexical model parameters of grapheme subword units and orthographic transcription of the word. More precisely, the decoding phase involves the following steps:

1. The orthographic transcription of a given word is parsed to extract the (context-independent) grapheme sequence. For example, the word AREA is expanded as [A] [R] [E] [A].
2. The context-independent grapheme sequence is then turned to the context-dependent grapheme sequence. For example, the sequence [A] [R] [E] [A] is expanded into the sequence [A+R] [A-R+E] [R-E+A] [E-A].

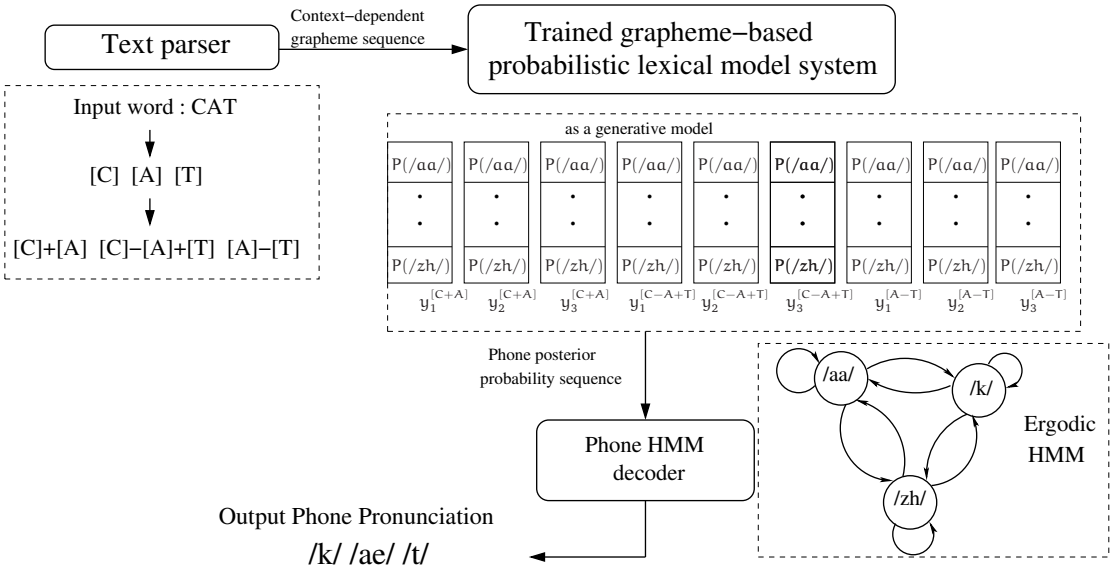

Figure 8.3 – Acoustic data-driven G2P conversion using lexical model parameters and orthographic transcription of words.

3. A word level HMM is created by concatenating the HMMs of context-dependent graphemes in the sequence. A sequence of acoustic unit probability vectors is then obtained by stacking the categorical distributions of the states in the (left-to-right) order in which the states are connected. In other words, the grapheme HMM sequence acts as a generative model where each state (in the left-to-right sequence) generates a single probability vector.

   For example, in the case of the context-dependent grapheme sequence [A+R] [A-R+E] [R-E+A] [E-A] the sequence of acoustic unit probability vectors starts with the lexical model parameters (categorical distribution) of the first HMM state of [A+R] followed by the lexical model parameters of the second HMM state of [A+R], and so on till the lexical model parameters of the final HMM state of [E-A].

4. Finally, the acoustic unit posterior probabilities in the sequence are used as *local scores*, exactly like in the case of the hybrid HMM/MLP system [Bourlard and Morgan, 1994], and decoded by a fully ergodic HMM system (that connects all D acoustic units with a uniform transition probability matrix) to infer the acoustic unit sequence. Since, the acoustic units in our case are context-independent phones, a phone sequence is inferred.

5. Multiple pronunciations for a word could be extracted using n-best decoding. However, in this thesis we only used 1-best decoding, i.e. single pronunciation model for each word.

8.2.3 Links to other G2P Approaches

Broadly, the proposed G2P approach can be considered to be similar to conventional G2P approaches that are based on decision trees and joint multigrams. In all the three of them, the phone sequence given the grapheme sequence is obtained based on the G2P relationship learned on the training data. The proposed approach learns the G2P relationship on acoustic data of the target language, while the conventional approaches learn the G2P relationship on a seed lexicon of the target language. However, the training and decoding algorithms are different in the conventional and the proposed G2P conversion approaches.

Specifically, the proposed G2P approach is similar to the decision-tree based G2P approach [Pagel et al., 1998], as they both learn the relationship between context-dependent graphemes and context-independent phones. The joint-multigram based G2P conversion approach [Bisani and Ney, 2008] jointly models context-dependent graphemes and context-dependent phones using graphones. Furthermore, the context of graphemes or phones considered in the proposed G2P approach is generally smaller than the context of graphemes and/or phones considered in decision-tree based or joint-multigram based G2P conversion approaches. On the other hand, unlike the decision-tree based G2P approach that is based on local classification, the proposed G2P approach is based on sequence classification like the joint-multigram based G2P conversion approach.

We hypothesize that,

- the ASR system using the phone lexicon generated from the proposed G2P approach should perform similarly to or better than the one generated from the decision-tree based G2P approach as it uses acoustic data and is a sequence classification approach; and
- the ASR system using the phone lexicon generated from the proposed G2P approach may perform similarly to or worse than the phone lexicon generated from joint-multigram based G2P approach as the context of subword units considered in the proposed approach is smaller than the joint-multigram approach.

8.2.4 Advantages of the Proposed G2P Approach

The advantages of the proposed G2P approach are:

- The proposed G2P approach is capable of generating phone lexical resources given only acoustic data from the target language. Furthermore, it can exploit the target language (as seen in Chapter 5) and/or language-independent (as seen in Chapter 6) resources for G2P conversion.

  The conventional G2P approaches [Taylor et al., 1998, Bisani and Ney, 2008, Novak, 2011, Chen, 2003] are only applicable if a seed lexicon from the target language is available. Therefore, they can only exploit phone lexical resources of the target language.
- Probabilistic lexical model based systems have a relatively small number of lexical model parameters (i.e., $I * D$) that can be learned on either a small amount of training data (Chapters 5 and 6) or can be knowledge driven (Chapter 7). The proposed G2P approach that is

based on lexical model parameters is of particular interest when there is less transcribed data.

- The approach uses acoustic data to generate phone lexical resources for the target domain. As a result it is expected to capture the natural phonological variation in the extracted pronunciations. However, unlike other acoustic G2P approaches [Xiao et al., 2007, McGraw et al., 2013, Lu et al., 2013], the proposed approach does not presume the availability of acoustic samples of words for which we are interested to generate phone pronunciations.
- The search involved during decoding to infer the phone sequence is relatively simple.
- In this thesis, we focus only on the generation of pronunciations with phones. However, the approach could be extended to other unit representations, such as syllables, automatically derived acoustic units and articulatory features.

8.3 Experimental Studies

To demonstrate the potential of the proposed G2P approach, we consider the following two cases:

1. In the first case, the goal is to build an ASR system for a domain that does not have any prior lexical resources, i.e., neither phone set nor pronunciation lexicon. However, we have access to a second domain that has acoustic and lexical resources. The lexicon of the second domain has a high out-of-vocabulary rate on the new domain for which we are interested to build ASR system. To study this case we consider the RM task that was also used in Chapter 5.
2. In the second case, we consider a scenario where limited transcribed speech data with its pronunciation lexicon constituting pronunciations of words seen in the speech data is available. The goal is to infer pronunciation models for words which are not seen in the training data (For example, to augment the system vocabulary with a new set of words). This scenario is likely to occur while developing ASR systems for under-resourced languages. To study this case we consider the PhoneBook task that was also used in Chapter 5.

The two cases presented here build on top of our preliminary investigations [Rasipuram and Magimai.-Doss, 2012a,b] where we focussed on extracting pronunciations using a grapheme-based KL-HMM system modeling word internal context-dependent subword units. In [Rasipuram and Magimai.-Doss, 2012a], pronunciations were generated using the grapheme-based KL-HMM system modeling context-dependent subword units either with the single preceding and the single following subword context (*tri*) or with the double preceding and the double following subword context (*quint*). In this chapter, we focus only on context-dependent subword units with *tri* context. Furthermore, in our previous work [Rasipuram and Magimai.-Doss, 2012b], we used back-off to handle unseen context-dependent graphemes, i.e., the context of unseen context-dependent graphemes encountered is decreased gradually until we encounter an observed grapheme. In the thesis, for the sake of consistency with the previous chapters, we use decision-tree based state clustering and tying approach to handle unseen context-

dependent graphemes. It is worth mentioning that there is no difference in performance either at pronunciation level or at ASR performance level between employing backoff and decision-tree method to handle unseen graphemes.

8.3.1 Experimental Setup

In the case of the RM and PhoneBook tasks we compare the following three lexica:

1. *acoustic-G2P*: Pronunciation lexicon generated using the proposed acoustic data-driven G2P approach.
2. *decision-G2P*: Pronunciation lexicon generated using a decision tree based G2P convertor [Pagel et al., 1998]. We use the G2P convertor in the festival toolkit [Taylor et al., 1998] for this purpose. The G2P convertor was trained on either the WSJ lexicon (RM task) or the PhoneBook train lexicon.
3. *graphone-G2P*: Pronunciation lexicon generated using the joint-multigram approach [Bisani and Ney, 2008]. We use the Sequitur G2P toolkit. The G2P convertor was trained on either the WSJ lexicon (RM task) or the PhoneBook train lexicon.

RM task: Similar to Section 5.1.1, we consider the DARPA RM corpus as the target domain for which we are interested to build a phone-based ASR system. However, we assume that phone lexical resources are not available. Wall Street Journal (WSJ) is used as the out-of-domain corpus where acoustic and lexical resources are available. Out of 1000 words in the RM lexicon only 568 words are seen in the WSJ pronunciation lexicon. In this case, the acoustic model is trained on the WSJ corpus and the grapheme-based probabilistic lexical modeling system is trained on the RM corpus. The phone-based pronunciation lexicon for the RM task is generated using the grapheme lexical model parameters and the orthography of the words.

PhoneBook task: To study the second case we use the PhoneBook task because the test vocabulary consists of words and speakers which are unseen during training. As done in Chapter 5, the acoustic model or the MLP for probabilistic lexical model based systems was trained on limited training data of the PhoneBook corpus to classify 42 context-independent phones. The phone-based pronunciation lexicon for the test vocabulary is generated using the grapheme lexical model parameters and the orthography of the words.

For both the RM and PhoneBook tasks, we use the context-dependent grapheme-based probabilistic lexical model systems trained using the local score S_{SKL} described in Section 5.1 as it resulted in minimum KL-divergence on the training data compared to other local scores.

In the case of the Sequitur G2P the width of graphone context was tuned on the development set (5% of WSJ1 lexicon or the development set of the PhoneBook task). The optimal graphone context size was 5 for both tasks. In the case of the festival G2P, the width of grapheme context was set to 5 in both cases. The pronunciation models of words generated with different methods are evaluated in terms of pronunciation errors and ASR performance. In the phone HMM decoder of the proposed G2P approach, each phone was modeled by a three-state

HMM.

8.3.2 Pronunciation Error Analysis

The pronunciation models of a few words generated using the proposed G2P approach with their respective pronunciation from the RM lexicon are given in Table 8.1.

Table 8.1 – Pronunciation models of a few words generated using the proposed acoustic data-driven G2P approach on the RM task. By actual pronunciation, we refer to the pronunciation given in the RM lexicon.

Word	Actual pronunciation	Extracted pronunciation
WHEN+S	/w//eh//n//z/	/w//eh//n//z/
ANCHORAGE	/ae/ /ng/ /k/ /er/ /ih/ /jh/	/ae/ /ng/ /k/ /ch/ /ao/ /r/ /ih/ /jh/
ANY	/eh/ /n/ /iy/	/ae/ /n/ /iy/
CHOPPING	/ch/ /aa/ /p/ /ih/ /ng/	/ch/ /aa/ /p/ /iy/ /ng/
ADDING	/ae/ /dx/ /ih/ /ng/	/ae/ /t/ /ih/ /ng/

The generated pronunciations were compared with the pronunciations given in the well developed lexicon of the RM or PhoneBook corpora. Tables 8.2 and 8.3 present this comparison in terms of phone accuracy (PA) and word accuracy (WA) on the RM and PhoneBook tasks respectively. It can be observed that the *decision-G2P* and *graphone-G2P* lexica perform better than the *acoustic-G2P* lexicon in terms of phone accuracy.

Table 8.2 – Evaluation of the extracted pronunciation models in terms of phone accuracy (PA) and word accuracy (WA) for three different approaches on the RM task.

Lexicon	PA	WA
acoustic-G2P	81.5%	34.6%
decision-G2P	86.6%	53.8%
graphone-G2P	92.2%	72.4%

Table 8.3 – Evaluation of the extracted pronunciation models in terms of phone accuracy (PA) and word accuracy (WA) for three different approaches on the PhoneBook task.

Lexicon	PA	WA
acoustic-G2P	72.4%	11.0%
decision-G2P	81.5%	31.0%
graphone-G2P	89.2%	50.6%

8.3.3 ASR Performance Analysis

We built context-dependent phone-based ASR systems using the pronunciation lexica generated with various G2P convertors. In the case of the RM task, crossword context-dependent systems are built whereas in the case of the PhoneBook task word-internal context-dependent systems are built (as it is an isolated word recognition task). Two types of ASR systems are trained, namely, HMM/GMM and KL-HMM *SKL* systems. In the case of the RM task, the KL-HMM systems use the MLP trained on the WSJ corpus as the acoustic model and in the case of the PhoneBook task, the MLP trained on the PhoneBook corpus is used as the acoustic model. The performance of the systems using G2P-convertor based lexica is compared with the performance of ASR systems using the *GRAPH* lexicon and the well developed *PHONE* lexicon from Chapter 5

RM Task

The ASR performance of different systems in terms of word accuracy on the RM task is presented in Table 8.4. It is interesting to note that the systems using the *acoustic-G2P* lexicon (that has correct pronunciations for only 34.6% words) perform better than the systems using the *decision-G2P* lexicon (that has correct pronunciations for 53.8% of words). It can be also observed that the ASR performance difference between systems using the G2P-convertor based lexicon and the *PHONE* lexicon is lower for the KL-HMM approach than the HMM/GMM approach.

Table 8.4 – The ASR performance in terms of word accuracy on the RM task for various crossword context-dependent systems using different lexica

Lexicon	System	
	HMM/GMM	KL-HMM
acoustic-G2P	94.7%	95.3%
decision-G2P	91.7%	94.0%
graphone-G2P	95.1%	95.6%
PHONE	95.9%	95.9%
GRAPH	94.8%	95.5%

Part of the words in the RM task are present in the WSJ lexicon. Therefore, we built three lexica (*Mixed-WSJ-acoustic-G2P*, *Mixed-WSJ-decision-G2P* and *Mixed-WSJ-graphone-G2P*) where the pronunciation for common words is obtained from the WSJ lexicon and the rest using the G2P-convertor based lexicon (*acoustic-G2P*, *decision-G2P* and *graphone-G2P*, respectively). Table 8.5 presents the ASR performance of these three systems in terms of word accuracy. It can be observed that the performance of systems using *Mixed-WSJ-acoustic-G2P* and *Mixed-WSJ-decision-G2P* lexica is better than the systems using *acoustic-G2P* and *decision-G2P* lexica, respectively. However, the performance of systems using *Mixed-WSJ-graphone-G2P* and *graphone-G2P* lexica system is the same. This could be because the multigram G2P

approach memorizes the pronunciations of words seen in the training data. Furthermore, the performance of systems using the *Mixed-WSJ-acoustic-G2P* and *Mixed-WSJ-graphone-G2P* lexica is the same.

Table 8.5 – The ASR performance in terms of word accuracy on the RM task for various crossword context-dependent systems using different lexica. The systems use a lexicon where the pronunciation of RM words present in the WSJ lexicon are retained and the pronunciations for rest of the RM words are generated using G2P conversion.

Lexicon	System	
	HMM/GMM	KL-HMM
Mixed-WSJ-acoustic-G2P	95.1%	95.6%
Mixed-WSJ-decision-G2P	92.8%	94.9%
Mixed-WSJ-graphone-G2P	95.1%	95.6%
PHONE	95.9%	95.9%
GRAPH	94.8%	95.5%

PhoneBook task

The performance of systems using *acoustic-G2P*, *decision-G2P* and *graphone-G2P* lexica in terms of ASR word accuracy on the test set of the PhoneBook task is presented in Table 8.6. Results show that the performance of the HMM/GMM system using the *acoustic-G2P* lexicon is worse than that of the HMM/GMM system using the *decision-G2P* lexicon. However, the performance of KL-HMM systems using *acoustic-G2P* and *decision-G2P* lexica is the same. The systems using the *graphone-G2P* lexicon performs better than the systems using *acoustic-G2P* and *decision-G2P* lexica. However, the performance of the system using any of the G2P-convertor based lexica is poor than that of the system using the optimistic *PHONE* lexicon. Unlike the RM task, the reason for the large ASR performance difference between systems using the *PHONE* and *acoustic-G2P* lexica on the PhoneBook task could be the following:

1. In the RM task, the MLP is trained on large amount of domain-independent data (about 66 hours of speech). While, in the PhoneBook task the MLP is trained on only 5 hours of speech. Moreover, in the RM task, the speech data is from a microphone, whereas in the PhoneBook task it is a telephone speech.
2. In the case of the RM task, the grapheme-based probabilistic lexical model system is trained on words for which the pronunciations are to be extracted. For the Phone-Book task the words are neither seen during MLP training nor during grapheme-based probabilistic lexical model system training.

Two more systems using *acoustic-decision-G2P* and *acoustic-graphone-G2P* lexica are built that use the pronunciation lexicon consisting of two pronunciations for each word, one from the *acoustic-G2P* lexicon and one from either the *decision-G2P* or the *graphone-G2P* lexicon. Results show that combination of pronunciations from different methods yields significant performance improvement over the systems using pronunciations from only one

Table 8.6 – The ASR performance in terms of word accuracy on the PhoneBook task for various context-dependent systems using different lexica

Lexicon	System	
	HMM/GMM	KL-HMM
acoustic-G2P	83.4%	86.7%
decision-G2P	85.4%	86.7%
graphone-G2P	87.1%	89.1%
acoustic-decision-G2P	89.3%	91.5%
acoustic-graphone-G2P	89.7%	91.7%
PHONE	97.0%	97.8%
GRAPH	91.0%	93.6%

of the method. This shows that the pronunciation models learned from *acoustic-G2P* and *decision-G2P* or *graphone-G2P* provide complementary information to ASR. It is encouraging to observe that combining pronunciations from different G2P approaches is beneficial for ASR performance. Also, by combining the two lexica, we combine a lexical knowledge driven approach (*decision-G2P* or *graphone-G2P*) and an acoustic data-driven approach (*acoustic-G2P*).

On both the RM and PhoneBook tasks, despite having high pronunciation error rate, the ASR system using the lexicon from the proposed approach performs better than or similarly to the systems using the lexicon from the decision-tree or joint n-gram based G2P convertors. In the acoustic G2P approach pronunciations are based on the probabilistic G2P relationship learned through both acoustic and lexical resources, whereas *decision-G2P* and *graphone-G2P* are based on the G2P relationship learned using lexical resources. Therefore, in the acoustic G2P approach errors at the pronunciation level could be due to substitution with acoustically similar phone (reflected in the target domain data) and thus are not affecting the ASR performance.

Furthermore, on both the RM and PhoneBook tasks, the performance of the proposed grapheme-based ASR system (i.e., KL-HMM system using the *GRAPH* lexicon) is similar to or better than that of the phone-based ASR systems using phone lexicon from either conventional G2P approaches or the proposed acoustic data-driven G2P approach.

8.4 Summary

In this chapter, we presented an acoustic data-driven G2P conversion approach that exploits the G2P relationship captured in the lexical model parameters of a grapheme-based probabilistic lexical model system. The approach has been investigated on two lexical resource constrained ASR tasks and compared with the decision-tree based G2P approach and the joint sequence model based G2P approach.

In terms of the pronunciation errors compared to the well developed lexicon, the proposed approach performs worse than the decision-tree based or joint n-gram based G2P approaches. On ASR tasks, as hypothesized, the system using the phone lexicon generated from the proposed G2P approach performed similarly to or better than the one generated from the decision-tree based G2P approach; and the ASR system using the phone lexicon generated from the proposed G2P approach performed similarly to or worse than the phone lexicon generated from the joint-multigram based G2P approach. Furthermore, it was also shown that the proposed G2P approach can complement decision-tree based or joint n-gram based G2P approaches for ASR.

9 Improving Phone-less Grapheme-based ASR

In the literature, as summarized in Section 4.1.2, research in the field of grapheme-based ASR has primarily focussed on context-dependent modeling and decision-tree based state tying within the framework of deterministic lexical modeling. The implicit assumption being that the relationship between context-independent graphemes and context-independent phones can be irregular, but the relationship between context-dependent graphemes and context-independent phones could be regular. In the previous chapters, we have shown that by modeling the relationship between context-dependent graphemes and context-independent phones, effective grapheme-based ASR systems could be built. Also, G2P conversion systems exploit this notion and model relationship between context-dependent graphemes and context-independent phones through decision trees or the joint n-gram model, for example.

In this chapter, we make the following two hypotheses:

- The clustered context-dependent graphemes model phone-like information, because, the decision-tree clustering in addition to the subword context is based on acoustic feature observations (which capture phone information).
- The effect of pronunciation errors in the grapheme lexicon on ASR performance could be mitigated through the use of probabilistic lexical modeling.

We validate our hypothesis by focussing on the grapheme-based HMM/GMM system incorporating probabilistic lexical modeling.

We will start the chapter with a brief motivation and related work. The details of the proposed approach are given in Section 9.2. In Section 9.3, we present experimental studies on two English ASR tasks.

9.1 Motivation and Related Work

As elucidated in Chapter 2, standard HMM-based ASR systems directly model the relationship between lexical units and acoustic feature observations. As a result HMM-based ASR systems rely on a well developed phone lexicon and subword units to handle the variability in the

acoustic training data. However, when the pronunciations in the lexicon do not reflect the underlying speech data then such a model may poorly represent the training data. For example, this can happen in the case of non-native speakers (where pronunciations normally reflect native speakers) or in the case of spontaneous and conversational speech (where spoken words are pronounced differently from lexicon pronunciations) or in the case of a grapheme lexicon (where pronunciations are based on the spelling of the word). To account for such variation, typically, phone-based ASR systems add pronunciation variants to the lexicon.

In the context of modeling pronunciation variability, the limitation of the standard HMM/GMM system imposed by deterministic mapping has been handled by modeling a probabilistic relationship between lexical and acoustic units [Luo and Jelinek, 1999, Saraclar et al., 2000, Hain and Woodland, 1999, Hain, 2005]. It is important to note that the notion of acoustic units and lexical units was not explicitly defined in these previous works. As described below, these approaches can be viewed from the point of view of probabilistic lexical modeling.

The PC-HMM approach [Luo and Jelinek, 1999], as described in Section 3.3.2, is a probabilistic lexical modeling approach where the decision-tree clustered context-dependent phones are the acoustic units and the lexical-to-acoustic unit probabilities are estimated using the EM algorithm along with the GMM parameters. In [Saraclar et al., 2000], this approach was applied to model pronunciation variability in spontaneous speech. The technique starts with standard GMMs trained using decision-tree based state tying, and then combines Gaussians from phones that are found to be frequent variants of each other in phonetic transcriptions. In [Hain and Woodland, 1999, Hain, 2005], hidden model sequences HMM (HMS-HMM) was proposed, where the deterministic mapping between phone-to-HMM or phone-to HMM-state was replaced with a stochastic model. More specifically, each phone was represented by a mixture of HMM state sequences corresponding to different variants. In [Hain, 2005], multiple pronunciations of a word in a lexicon were collapsed to a single pronunciation. It was shown that an HMM/GMM system using a pronunciation lexicon with single pronunciation for each word resulted in similar or better performance compared to an HMM/GMM system using a pronunciation lexicon with multiple pronunciations for words on both read and conversational ASR tasks. The use of pronunciation lexicon with single pronunciation for each word in the HSM-HMM system further improved the ASR performance.

In [Yu and Schultz, 2003], the limitation imposed by deterministic lexical modeling has been addressed through “enhanced tree clustering” that allows efficient parameter sharing across phones. In standard HMM/GMM systems, a decision-tree is trained for each phone, whereas in enhanced tree clustering a single decision-tree is constructed for all the sub-states of all phones. The clustering procedure starts with all polyphones at the root. The decision-tree can ask questions regarding the identity and phonetic properties of the center phone and the neighbouring phones plus the sub-state identity. Nevertheless, as discussed in Section 4.1.2, the enhanced tree clustering based ASR approach uses a deterministic map between lexical and acoustic units.

In [Mimer et al., 2004], it has been shown that enhanced tree clustering improves the performance of grapheme-based ASR systems. Through enhanced tree clustering it is possible to capture the fact that different graphemes may be pronounced in a similar manner depending on their context. On both German and English ASR tasks, the enhanced tree clustering procedure was able to improve the performance of grapheme-based ASR systems. However, even with enhanced tree clustering, for English the performance of the grapheme-based ASR system was significantly worse than that of the phone-based ASR system.

9.2 Proposed Approach

In this chapter, we show that the set of acoustic units can be based on context-dependent graphemes; and the performance of a grapheme-based ASR system can be improved by incorporating probabilistic lexical modeling. To be consistent with our previous work [Rasipuram and Magimai.-Doss, 2013a], the studies in this chapter used GMMs as acoustic models. However, the approach is not restricted to GMM acoustic models. For example, in [Rasipuram et al., 2013a] we used an ANN classifying context-independent graphemes as acoustic model.

The proposed approach is implemented in the following two stages:

- A standard context-dependent grapheme-based HMM/GMM system using decision tree based state tying is trained.
- As acoustic units $\mathscr{A}$, we use the decision-tree clustered states modeled using GMMs. The probabilistic relationship between lexical units and acoustic units is learned using the KL-HMM approach. The states of the KL-HMM system (or the lexical units) are context-dependent grapheme subword units. The acoustic unit probability sequence is estimated given the set of acoustic units and their corresponding GMMs as,

$$z_t^d = P(a^d|\mathbf{x}_t) = \frac{p(\mathbf{x}_t|a^d)}{\sum_{j=1}^{D} p(\mathbf{x}_t|a^j)} \tag{9.1}$$

  where $p(\mathbf{x}_t|a^d)$ is the likelihood of acoustic unit a^d. The above equation assumes equal priors for the acoustic units.

The resulting system is a grapheme-based ASR system that incorporates a probabilistic lexical model. Furthermore, the proposed grapheme-based ASR approach does not use any phonetic information or out-of-domain resources.

9.3 Experimental Setup and Results

In this section, we compare deterministic lexical modeling and probabilistic lexical modeling in the context of both grapheme-based and phone-based HMM/GMM systems. ASR studies are conducted on the DARPA RM and si-84 WSJ0 tasks. The details of the two tasks are given in Sections A.1 and A.2 of the Appendix.

For both tasks, the phone lexicon was obtained from the UNISYN lexicon [Fitt, 2000]. The grapheme lexicon was transcribed using 79 graphemes where the first and last graphemes of a word are treated as separate units.

In our initial study on the RM corpus [Rasipuram and Magimai.-Doss, 2013a] we have used the grapheme lexicon transcribed using 29 graphemes, i.e., first and last graphemes of words were not treated as separate graphemes. As a result, the performance presented in the paper [Rasipuram and Magimai.-Doss, 2013a] and this chapter for grapheme-based ASR systems differs.

9.3.1 Deterministic Lexical Model based ASR System

We build crossword context-dependent HMM/GMM systems with decision-tree based state tying using the HTK toolkit [Young et al., 2006]. Each context-dependent subword unit is modeled by three HMM states. The acoustic feature $\mathbf{x}_t$ is the 39-dimensional PLP cepstral feature vector. The phoneme-based HMM/GMM system uses a phonetic question set whereas the grapheme-based HMM/GMM system uses a singleton question set. For the RM task, state tying resulted in 1611 clustered/acoustic units for the phone-based system and 1536 clustered/acoustic units for the grapheme-based system. For SI-84 task, state tying resulted in 1900 clustered/acoustic units for the phone-based system and 2190 clustered/acoustic units for the grapheme-based system.

9.3.2 Probabilistic Lexical Model based ASR System

Given the GMMs of acoustic units, the training of a probabilistic lexical model based system involves,

1. the estimation of the acoustic unit posterior feature vector $\mathbf{z}_t = [z_t^1, \ldots, z_t^d, \ldots, z_t^D]^{\mathrm{T}}$ assuming equal priors for the acoustic units according to the Eqn (9.1); and
2. the estimation of the lexical model parameters using the KL-HMM RKL approach. As hypothesized in the beginning of the chapter, for grapheme lexical units, the lexical-to-acoustic unit relationship is expected to be one-to-many. Therefore, following the analysis in Chapter 3 and ASR results in previous chapters, S_{RKL} was chosen as the local score.

We train and test crossword context-dependent probabilistic lexical model systems where the lexical units impose three-state minimum duration constraint.

9.3.3 Systems

We built the following six systems:

1. BASE-PHONE: The phone-based ASR system with a deterministic lexical model, where lexical units are context-dependent phones and acoustic units are the clustered states.

2. BASE-GRAPH: The grapheme-based ASR system with a deterministic lexical model, where lexical units are context-dependent graphemes and acoustic units are the clustered states.
3. PROB-PHONE: The phone-based ASR system with a probabilistic lexical model, where lexical units are context-dependent phones and acoustic units are the clustered states of the system BASE-PHONE.
4. PROB-GRAPH: The grapheme-based ASR system with a probabilistic lexical model, where lexical units are context-dependent graphemes and acoustic units are the clustered states of the system BASE-GRAPH.
5. PROB-PHONE-CROSS: The phone-based ASR system with a probabilistic lexical model, where lexical units are context-dependent phones but the acoustic units are the clustered states of the system BASE-GRAPH.
6. PROB-GRAPH-CROSS: The grapheme-based ASR system with a probabilistic lexical model, where lexical units are context-dependent graphemes but the acoustic units are the clustered states of the system BASE-PHONE.

The system PROB-GRAPH-CROSS is somewhat similar to the grapheme-based ASR system presented in Section 5.1, in the sense that both use phone information. More precisely, in Section 5.1 acoustic units were context-independent phones and acoustic model was an MLP trained on cross-domain data. In the system PROB-GRAPH-CROSS, the acoustic units are clustered context-dependent phones and the acoustic model includes a set of GMMs learned on target-domain data. The system PROB-GRAPH that derives acoustic units from context-dependent graphemes does not use any phone information.

9.3.4 Results

The ASR performance of the various systems in terms of word accuracy is given in Table 9.1. The following observations are made:

- On both the RM and WSJ tasks, the system PROB-GRAPH performs significantly better than the system BASE-GRAPH. Furthermore, on the RM task, the system PROB-GRAPH performs better than the system BASE-PHONE. The results show that the probabilistic lexical model based ASR systems handled the errors in a grapheme pronunciation lexicon, as well as the mismatch between grapheme pronunciations and acoustic feature observations better than the deterministic lexical modeling based systems.
- The performance of the system PROB-PHONE is significantly better than that of the system BASE-PHONE on the RM task, while similar to the system BASE-PHONE on the WSJ task. This difference in improvement among the two tasks could be attributed to the size of the data and the implicit pronunciation variation modeling capability of GMM acoustic models. That is, given enough data and a well developed pronunciation lexicon, the GMM acoustic models should be capable of modelling and capturing the pronunciation variability implicitly [Hain, 2005].
- On the RM task, the performance of the system BASE-PHONE is same as the performance

Table 9.1 – The performance in terms of word accuracy of various crossword context-dependent systems on the RM and WSJ tasks

	System	Lexical Model	RM	WSJ0
1	BASE-PHONE	deterministic	95.9	91.1
2	PROB-PHONE	probabilistic	97.1	91.3
3	BASE-GRAPH	deterministic	94.8	85.4
4	PROB-GRAPH	probabilistic	96.5	88.0
5	PROB-PHONE-CROSS	probabilistic	97.1	88.8
6	PROB-GRAPH-CROSS	probabilistic	96.7	88.8

reported in the literature [Hain and Woodland, 1999, P et al., 2011]. Furthermore, on the RM task, the system PROB-PHONE performs better than the system based on HSM-HMM approach [Hain and Woodland, 1999]. The performance of the system based on the HSM-HMM approach was reported as 96.6% word accuracy [Hain and Woodland, 1999].

- On the WSJ task, the performance of the system BASE-PHONE is same as the performance of the HMM/GMM system using the same train and test sets as reported in the literature [Woodland et al., 1994].

The results validated our hypothesis that the performance of grapheme-based ASR systems can be significantly improved by incorporating probabilistic lexical modeling.

The systems PROB-PHONE-CROSS and PROB-GRAPH-CROSS were built to validate the hypothesis that context-dependent graphemes model phone-like information.

- On the RM task, it can be observed that the performance of the system PROB-PHONE-CROSS and the system PROB-PHONE are the same. Furthermore, the performance of the system PROB-GRAPH-CROSS and the system PROB-GRAPH are similar. This indicates that clustered states of the system BASE-PHONE and the system BASE-GRAPH are modeling a similar kind of acoustic information, and the poor performance of the system BASE-GRAPH is primarily due to the use of a deterministic lexical model.
- On the WSJ task, the performance of the system PROB-PHONE-CROSS is worse than that of the system PROB-PHONE while the performance of the system PROB-GRAPH-CROSS is better than that of the system PROB-GRAPH. The results indicate that to further improve the performance of PROB-GRAPH it may be necessary to improve the decision-tree clustering of graphemes and to model graphemes with context longer than the usual single preceding and single following one (to capture the irregular G2P relationship of English).

9.4 Summary

In this chapter, we showed that the performance of grapheme-based ASR systems can be significantly improved by incorporating probabilistic lexical modeling. The studies validated the hypothesis that the clustered context-dependent graphemes model phone-like information and the poor performance of grapheme-based ASR systems is primarily due to deterministic

lexical modeling. Furthermore, the studies indicated that the acoustic units, instead of being purely knowledge driven as in Chapters 5, 6 and 7, can also be derived using context-dependent graphemes and data-driven methods.

10 Conclusions and Future Directions

This thesis has focussed on addressing challenges related to the building of ASR systems for languages and domains that lack proper acoustic and lexical resources. In this thesis, the problem of modeling the relationship between lexical units and acoustic feature observations has been factored into two models using latent variables referred to as acoustic units: an acoustic model which models the relationship between acoustic feature observations and acoustic units, and a lexical model which models the relationship between lexical units and acoustic units. We have seen that in standard HMM-based ASR approaches like HMM/GMM and hybrid HMM/ANN, the relationship between lexical units and acoustic units is one-to-one and the lexical model is deterministic. We showed that in approaches like KL-HMM, Tied posterior proposed in the literature and SP-HMM proposed in this thesis, the lexical model models a probabilistic relationship between lexical units and acoustic units. The framework of probabilistic lexical modeling has been pivotal to the rest of the thesis.

Motivated by the three main advantages of probabilistic lexical modeling, i.e., 1) acoustic model and lexical model can be trained on independent set of resources, 2) lexical units and acoustic units can be different, and 3) lexical units and acoustic units can model different subword contexts, we proposed a novel grapheme-based ASR approach where the lexical units are graphemes and acoustic units can be phones or multilingual phones or clustered context-dependent subword units. In Chapter 4, we showed that the parameters of the lexical model capture a probabilistic G2P relationship. In Chapter 8, we proposed an acoustic data-driven G2P conversion approach in which the probabilistic G2P relationship captured in the lexical model parameters was exploited for G2P conversion. It has been observed that the performance of the proposed grapheme-based ASR system is similar to or better than that of the phone-based ASR system using phone lexicon from either conventional G2P approaches or proposed acoustic data driven G2P approach. The analysis in Chapter 4, and the studies in Chapters 5, and 8 revealed that the approach integrates lexicon learning as a phase in ASR system training and could potentially prevent the need for an explicit G2P convertor.

In Chapters 5, 6 and 7, we investigated the potential of the proposed grapheme-based ASR approach in addressing the lexical and acoustic resource constraints for ASR system devel-

opment. The studies showed that with probabilistic lexical modeling, especially using the KL-HMM approach, ASR systems can be rapidly developed for new languages and domains by training a language or domain independent acoustic model and learning the lexical model on a small amount of target language or domain data. Also, it was observed that irrespective of the type of subword units used, phones or graphemes, KL-HMM based systems performed better than (when training data is in-sufficient) or comparably to (when training data is sufficient) deterministic lexical model based systems. Our findings on phone-based ASR using the KL-HMM approach are inline with previous work on KL-HMM [Imseng, 2013]. Furthermore, in Chapter 7, we showed that in the proposed grapheme-based ASR approach, the lexical model parameters can be initialized based on the knowledge of the G2P relationship of the language. Therefore, the proposed framework can serve as a practical starting point while building ASR systems for new languages without any acoustic and lexical resources.

Among the various probabilistic lexical modeling approaches studied, it has been observed that the KL-HMM RKL approach is robust than the Tied-HMM and SP-HMM approaches. In Chapter 3, we showed that, from the parameter estimation point of view, the local score S_{RKL} has the capability to better model one-to-many G2P relationships than other local scores; from the decoding perspective it is capable of giving more importance to the acoustic model evidence than the lexical model evidence. In Chapter 4, where acoustic model and lexical model were trained on the same task, the KL-HMM RKL and Tied-HMM approaches performed similarly. However, in Chapters 5 and 6, where the acoustic model and the lexical model were trained on an independent set of resources, the KL-HMM RKL approach resulted in better performance. This suggests that the KL-HMM RKL approach can handle the mismatch between language-independent and target language resources better than other probabilistic lexical modeling approaches.

In Chapter 9, we showed that the acoustic units, instead of being purely knowledge driven as in Chapters 5, 6 and 7, can also be data-driven. In particular, our investigations in Chapter 9 indicated that a) the clustered context-dependent graphemes model phone-like information; b) the poor performance of grapheme-based ASR systems could be primarily due to deterministic lexical modeling; and c) the performance gap between grapheme-based and phone-based ASR systems can be significantly reduced by probabilistic lexical modeling.

In conclusion, our studies showed the following:

1. The demand for well-developed acoustic and phonetic lexical resources from the target language can be considerably reduced by replacing the deterministic lexical model with a probabilistic model learned on acoustic data.
2. The deterministic lexical model based ASR approaches are more suitable for phone-based ASR than grapheme-based ASR, while the probabilistic lexical model based ASR approach is suitable for both.
3. The proposed approach can effectively address the lack of both acoustic and lexical resources. More specifically, ASR systems can be rapidly developed for new languages and domains in the framework of probabilistic lexical modeling, especially using the

KL-HMM approach.

10.1 Directions for Future Research

In this thesis, we focussed mainly on the lexical model of an ASR system. Mostly, a three-layer MLP classifying context-independent phones was used as an acoustic model. The approach proposed in this thesis can be improved along the following directions.

Acoustic model: More recently, ANNs with deep architectures classifying context-dependent clustered phone units have gained lot of attention [Dahl et al., 2012, Hinton et al., 2012]. The proposed approach can be improved by:

- Improving the acoustic model using deep ANN architectures. In Chapter 7, we observed that the KL-HMM system using a five-layer MLP performs better than the KL-HMM system using a three-layer MLP as acoustic model. In recent works, it has been shown that KL-HMM retains its benefit over the standard hybrid HMM/ANN system even when deep neural networks are used [Imseng et al., 2013b, Razavi et al., 2014].
- Improving the acoustic unit set. In Chapter 9, where GMMs were used as acoustic model, we observed that the acoustic unit set can be clustered context-dependent subword units. The acoustic model could be further improved by using deep ANN architectures in place of GMMs.

Acoustic and lexical model adaptation: In this thesis, we compared probabilistic lexical model based systems (where only the lexical model is trained on target language data) with deterministic lexical model based systems (where either acoustic model is adapted on target language data or both acoustic model and lexical model are trained on target language data). In Chapter 6, we observed that with increase in target language acoustic data, the gap between KL-HMM system and acoustic model adaptation based systems reduces. This suggests that there may be benefits in combining acoustic model adaptation and probabilistic lexical modeling.

- When using ANN-based acoustic model, this can be achieved by training a hierarchical neural network [Pinto et al., 2011] or adapting a neural network with target language data [Swietojanski et al., 2012]. A study on Scottish Gaelic in the framework of KL-HMM has shown the potential of acoustic model adaptation using the hierarchical neural network approach [Rasipuram et al., 2013a].
- The KL-HMM approach is not restricted to ANN-based acoustic modeling alone as shown in Chapter 9. Therefore, using GMMs as acoustic model this can be achieved by adapting the GMMs through the MAP technique followed by KL-HMM training ; or the parameters of GMM and probabilistic lexical model can be jointly estimated using the PC-HMM approach [Luo and Jelinek, 1999],

As mentioned earlier in Section 4.1, in the framework of deterministic lexical modeling, acoustic model adaptation and lexical model adaptation can be combined in different ways. For instance, (a) by combining acoustic model adaptation with polyphone decision tree state tying

(PDTS) [Schultz and Waibel, 2001b], or (b) using the SGMM approach [Burget et al., 2010]. Comparing probabilistic lexical modeling and deterministic lexical modeling along these lines with graphemes as subword units would be interesting.

Acoustic data-driven G2P conversion: In Chapter 8, we discussed the potential of the proposed acoustic G2P approach. The proposed acoustic G2P approach could be further improved along the following directions:

- Acoustic model: as discussed above, the acoustic model (or the MLP) could be improved either using deep ANN architectures or by modeling clustered context-dependent phones. Given an acoustic model that classifies context-dependent phones, it is possible to learn the relationship between context-dependent graphemes and context-dependent phones through lexical model parameters. In such a case, the proposed G2P approach can be considered as similar to the joint-multigram based G2P approach and still carry the benefits of the proposed approach.
- Lexical model: following the conventional G2P approaches, it may be beneficial to model grapheme contexts longer than the single preceding and single following context.
- G2P conversion: it is possible to incorporate phonotactic constraints during the G2P conversion process if phone lexical resources from the language are available.
- Multiple pronunciations: it would be interesting to see the use of multiple pronunciations extracted with the help of N-best list. However, in that case it may be important to learn the weights for each pronunciation. Furthermore, the multiple pronunciations extracted could be weighted if the acoustic samples are also available. In this way, the proposed acoustic G2P approach could be combined with other acoustic G2P approaches (that combine conventional G2P approaches and acoustic samples, as discussed in Section 8.1) [McGraw et al., 2013, Lu et al., 2013].

Automatically derived subword units: In Chapter 9, we observed that acoustic models of clustered context-dependent graphemes model phone-like information. Therefore, clustered context-dependent graphemes can be used as acoustic units in the proposed acoustic G2P approach to derive automatic subword units and the corresponding lexicon using transcribed speech.

Unifying ASR and TTS: Statistical ASR and TTS systems have three components in common: pronunciation lexicon, lexical model and acoustic model. The advancements made in ASR technologies have shown to be effective for TTS systems [King et al., 2008, Dines et al., 2010, Saheer et al., 2012]. For example, speaker adaptation techniques developed for HMM-based ASR are effective in adapting HMM-based TTS to target speaker [King et al., 2008]. In a similar vein, the probabilistic lexical modeling techniques used in this thesis could be used to unify acoustic and lexical model components of ASR and TTS systems.

A Databases

In this appendix, we will summarize different databases used in the thesis. Brief details of all the databases used in thesis are given in Table A.1. The tasks studied are diverse in terms of complexity, lexicon size, recording conditions etc.

Table A.1 – Overview of the tasks and the respective corpora used in the thesis

Corpus (Description)	Language	Lexicon size (in words)	# of Subword units		Train data (in min)	Test data (in min)
			Phones	Graphemes		
RM (Read speech)	English	991	42	79 or 27	228	66
WSJ0 (Read speech)	English	10000	42	79 or 27	840	160
WSJ1 (Read speech)	English	13000	42	79 or 27	3960	160
SpeechDat(II) (Native speech sampled at 8K used to train the acoustic model)	English French German Italian Spanish Greek	11855 34867 48446 29936 24522 35148	45 42 59 52 32 31	27 43 42 34 34 25	744 810 846 690 690 800	270 290 318 258 258 360
HIWIRE (Non-native speech from natives of France, Spain, Italy and Greece)	English	130	42	27	150	150
PhoneBook (isolated words)	English	2783	42	27	300	72
Scottish Gaelic (Broadcast news data)	Scottish Gaelic	5082	n.a.	83 or 32	180	60

A.1 Resource Management

The DARPA Resource Management (RM) corpus consists of read queries on the status of Naval resources [Price et al., 1988]. The task is artificial in aspects such as speech type, range of

vocabulary, and grammatical constraint. The training set and development set consists of 3'990 utterances spoken by 109 speakers corresponding to approximately 3.8 hours of speech data.

There are four test sets provided by DARPA, namely, feb89, oct89, feb91, and sep92. Each of the test set contains 300 utterances spoken by 10 speakers. The test set used in this work is obtained by combining the four test sets and thus contains 1,200 utterances amounting to 1.1 hours in total. The test set is completely covered by a word pair grammar included in the task specification.

The lexicon consists of 991 words. The phone-based lexicon was obtained from UNISYN [1] lexicon. There are 42 context-independent phones including silence. About 35 words in phone lexicon have more than one pronunciation.

In the literature, performance of the standard crossword context-dependent HMM/GMM system using phones as subword units on this test set was reported as 95.9% word accuracy [Hain and Woodland, 1999, P et al., 2011].

We built two grapheme lexica. The first grapheme lexicon was transcribed using 29 context-independent graphemes (which includes silence, symbol hyphen and symbol single quotation mark). The second grapheme lexicon was transcribed using 79 graphemes. The first grapheme and the last grapheme of a word are treated as separate graphemes. Therefore, the grapheme set included 26 English graphemes ({[A],[B],...[Z]}), 26 English graphemes occurring at the begin of word ({[b_A],[b_B],...[b_Z]}), 26 English grapheme occurring at the end of word ({[e_A],[e_B],...[e_Z]}) and silence. The introduction of word begin and word end graphemes was motivated by reasons such as: the grapheme-to-phoneme relationship of few graphemes in English can differ based on the position of grapheme. For example, the grapheme [E] at the word end in words such as 'hope', 'drive' is not pronounced. Also,

A.2 Wall Street Journal

The DARPA wall street journal corpus (WSJ) corpus was designed to provide speech data with large vocabularies [Paul and Baker, 1992]. The WSJ corpus [Paul and Baker, 1992, Woodland et al., 1994] has two parts - WSJ0 with 14 hours of speech (7,193 utterances from 84 speakers) and WSJ1 with 66 hours of speech (29322 utterances from 200 speakers). Systems can be built using the WSJ0 (also referred to as SI-84 training material), or WSJ1, or WSJ0+WSJ1 (also referred to as the SI-284 training data) formed by combining data from both WSJ0 and WSJ1 training utterances which contains approximately 80 hours of speech data (or 36,515 utterances from 284 speakers).

As test set we used Nov' 93 Hub 2 5K speech material containing 215 sentences from 10 speakers. The 5K word closed vocabulary bigram language model was used for decoding. The

1. http://www.cstr.ed.ac.uk/projects/unisyn/

test set includes words which are not seen in the training set. In the literature, performance of the standard crossword context-dependent HMM/GMM system using phones as subword units on this test set was reported as 91.3% word accuracy [Woodland et al., 1994].

Phoneme based lexicon was obtained from UNISYN lexicon [Fitt, 2000]. Phoneme lexicon consists of 46 context-independent phones including silence.

Grapheme lexicon was transcribed using 79 graphemes where along with 26 English alphabets, the first grapheme and the last grapheme of a word are treated as separate graphemes.

A.3 SpeechDat

SpeechDat is a series of databases created for voice driven teleservices [2]. SpeechDat(II) which is one of the SpeechDat projects that includes speech recorded over telephone network for speech recognition and speaker verification tasks. The database is recorded at 8kHz and stored in uncompressed 8 bit μ-law. In this work, data of six languages, namely, British English, Swiss French, Swiss German, Greek, Italian and Spanish is used. We only use the part of the corpus which contains phonetically rich sentences (10 sentences per speaker). Furthermore, the data is gender-balanced, dialect-balanced according to the dialect distribution in a language region, and age-balanced.

The same number of speakers (2000) were chosen from all the languages to avoid any bias in terms of available data from each language. The data from 1350 speakers is chosen as training set, data from 150 speakers as development set and data from 500 speakers as test set. Each language has approximately 12 hours of training data and 1.5 hours of development data. British English, Swiss French, Swiss German, Italian and Spanish have about 4 hours of test data while Greek has about 7 hours of test data. All the SpeechDat(II) lexicons use SAMPA [3] symbols. The phone sets of different languages in the SpeechDat(II) corpus used in this thesis are given in Table A.2.

A.4 HIWIRE

HIWIRE is a non-native speech corpus that contains utterances spoken by natives of France (31 speakers), Greece (20 speakers), Italy (20 speakers) and Spain (10 speakers) [Segura et al., 2007]. The utterances contain spoken pilot orders made of 133 words. The database provides grammar with a perplexity of 14.9. The database contains both clean (recorded in a quite room using close talking microphone) and noisy speech (obtained by adding real cockpit noise to clean data) speech material. In this thesis we only use the clean part of the database.

The HIWIRE task does not have training data. It only includes adaptation data (50 utterances

2. http://www.speechdat.org/
3. http://www.phon.ucl.ac.uk/home/sampa/

Table A.2 – Phone sets in the SAMPA format of various languages in the SpeechDat(II) corpus. Table also gives the multilingual phoneset used in the thesis.

Language	Phone set	# of phones
English	{, @, 3:, A:, aI, aU, b, d, D, dZ, e, e@, eI, f, g, h, i:, I, I@, j, k, l, m, n, N, O:, OI, p, Q, r, s, S, t, T, tS, u:, U, @U, U@, v, V, w, z, Z, sil	45
Swiss French	&/, 2, 9, 9 , @, A, E, E/, H, J, N, O, O/, R S, Z, a, a , b, d, e, e , f, g, i, j, k, l, m, n o, o , p, r, s, t, u, v, w, y, z, sil	42
Swiss German	?, @, 2:, 2:6, 9, 96, a, a:, a6, a:6, aI, aU, b, C, d, e:, E, E:, e:6, E6, E:6, f, g, h, i:, I, i:6, I6, j, k, l, m, n, N, o:, O, o:6, O6, OY, p, pf, R, s, S, t, ts, tS, u:, U, u:6, U6, v, x, y:, Y, y:6, z, Z, sil	59
Italian	@, a, b, bb, d, dd, ddz, ddZ, dz, dZ, e, E, f, ff, g, gg, i, j, J, JJ, k, kk, l, L, ll, LL, m, mm, n, nn, o, O, p, pp, r, rr, s, S, ss, SS, t, ts, tS, tt, tts, ttS, u, v, vv, w, z, sil	52
Spanish	a, b, B, d, D, e, f, g, G, i, j, J, jj, k, l, L, m, n, N, o, p, r, rr, s, t, T, tS, u, w, x, z, sil	32
Greek	a, b, c, C, d, D, dz, e, f, g, G, gj, i, j, jj, k, l, m, n, o, p, r, s, t, T, ts, u, v, x, z, sil	31
Multilingual phone set	?, {, @, &/, 2, 2:, 2:6, 3:, 9, 9 , 96, a, a , a:, A, A:, a6, a:6, aI, aU, b, B, bb, C, d, D, dd, ddz, ddZ, dz, dZ, e, e , e:, e@, E, E:, E/, e:6, E6, E:6, eI, f, ff, g, G, gg, h, H, i, i:, I, I@, i:6, I6, j, J, jj, JJ, k, kk, l, L, ll, LL, m, mm, n, N, nn, o, o , o:, O, O:, O/, o:6, O6, OI, OY, p, pf, pp, Q, r, R, rr, s, S, ss, SS, t, T, ts, tS, tt, tts, ttS, u, u:, U, @U, U@, u:6, U6, v, V, vv, w, x, y, y:, Y, y:6, z, Z, sil	117

per speaker, approx. 150 min) and test data (50 utterances per speaker, approx 150 min). The distribution of the database according to the native language of the speaker is given in Table A.3.

The phone lexicon supplied with the HIWIRE corpus contains pronunciations based on ARPABET (US English).

A noticeable difference between other works on HIWIRE and this thesis is that, we use lexicon based on SAMPA phone set while in the previous studies lexicon based on ARPABET phone set supplied with HIWIRE corpus was used. The lexicon based on SAMPA phone set was created by borrowing pronunciations of 102 words that are in common from the SpeechDat(II) English lexicon. For the remaining 31 words, we obtained pronunciations by mapping ARPABET phones to SAMPA phones. The main reason to use SAMPA phone set based lexicon in this

Table A.3 – Speaker distribution in HIWIRE corpus by country and number of utterances.

Country	# of speakers	# of utterances
France	31	3100
Greece	20	2000
Italy	20	2000
Spain	10	999
Total	81	8099

thesis is to have a shared subword units set between HIWIRE and SpeechDat(II) corpora. SpeechDat(II) corpus is used as domain-independent resource (in Chapter 5) and language-independent resource (in Chapter 6) for HIWIRE ASR tasks conducted in this thesis.

We transcribed the grapheme lexicon using 27 graphemes (26 English alphabets, and silence). HIWIRE includes about 30 abbreviated words in the lexicon. The abbreviated words present in the lexicon were transcribed using a look up table given in Table A.4 specifying the way individual graphemes are pronounced [4]. For instance, the graphemic transcription of the word "S.I.D" according to the lookup table is "[E] [S] [E] [Y] [E] [D] [E] [E]".

A.5 PhoneBook

PhoneBook is speaker-independent task-independent isolated word recognition corpus [Pitrelli et al., 1995] for small size (75 words) and medium size (600 words) vocabularies. We use the medium size vocabulary task with 600 unique words [Dupont et al., 1997].

The overview of the PhoneBook corpus in terms of number of utterances, speakers and words present in train, cross-validation and test sets is given in Table A.5. Training set consists of 26,711 utterances (obtained by merging the small training set and cross-validation set as in [Dupont et al., 1997]), and test set consists of 6598 speech utterances. The test vocabulary consists of words and speakers which are unseen during training, i.e., training and test vocabulary/speakers are completely different. PhoneBook pronunciation lexicon is transcribed using 42 phones (including silence). The performance of Hybrid HMM/ANN system on this setup was reported as 96.0% word accuracy [Pinto et al., 2009].

The grapheme-based lexicon was transcribed using 27 context-independent graphemes including 26 alphabets and silence.

4. http://en.wikipedia.org/wiki/English_alphabet

Table A.4 – Lookup table entries used to transcribe graphemes in the abbreviated words

Letter	Grapheme pronunciation
A	A or A E
B	B E E
C	C E E
D	D E E
E	E E
F	E F
G	G E E
H	A I T C H or H A I T C H
I	E Y E
J	J A Y
K	K A Y
L	E L
M	E M
N	E N
O	O
P	P E E
Q	C U E
R	A R
S	E S
T	T E E
U	Y O U
V	V E E
W	D O U B L E U or D O U B L E Y O U
X	E X
Y	W Y or W Y E
Z	Z E D or Z E E

A.6 Scottish Gaelic

The Scottish Gaelic speech corpus was collected by CSTR, University of Edinburgh [5]. The database was first used in [Rasipuram et al., 2013a], in collaboration with Dr. Peter Bell of CSTR, University of Edinburgh. Since, the corpus is relatively new we first briefly describe the language characteristics, alphabet, orthography, G2P relationship of Scottish Gaelic.

A.6.1 Language Characteristics

Scottish Gaelic is one of three primary Goidelic languages. Classified within the Indo-European language family, it is contained within the group of Celtic languages, and as such is only distantly related to any of the well-resourced major European languages. Scottish Gaelic is derived from and is closely related to Irish Gaelic. It is considered as an endangered and minority language, spoken by only around 60,000 speakers, mainly from the remote islands of Scotland.

5. http://forum.idea.ed.ac.uk/idea/gaelic-speech-recognition-and-scots-gaelic-sound -archive

Table A.5 – Overview of the PhoneBook corpus in terms of number of utterances, speakers and words present in the train, cross-validation and test sets.

Number of	Train	Cross-validation	Test
Utterances	19421	7920	6598
Speakers	243	106	96
Words	1580	603	600

The Scottish Gaelic alphabet has 18 graphemes (A, B, C, D, E, F, G, H, I, L, M, N, O, P, R, S, T, U) and long vowels are marked with grave accents (À, È, Ì, Ò, Ù). The number of phones in Scottish Gaelic are approximately 51 (9 vowels, 10 dipthongs and 32 consonants) [Wolters, 1997]. The number of phones can vary depending on the dialect. The language lacks proper speech and linguistic resources (phone set and pronunciation lexicon).

A.6.2 Orthography

The number of graphemes in Gaelic words are typically greater than the number of phones in the word, for two primary reasons: Firstly, in Gaelic, consonants are either broad (velarized) or slender (palatalized). Broad consonants are surrounded by broad vowels A, O or U on both sides and slender consonants are surrounded by slender vowels I or E on both sides. This has the consequence that many vowels are present in orthography only to denote the broad or slender nature of consonant next to it. Secondly, consonants of Gaelic words may be changed because of a process called lenition. In the orthography, grapheme [H] is added next to the consonant to mark this change, which typically results in aspiration of the consonant.

Broadly, however, with the exception of some very common function words, the G2P relationship of Gaelic is regular, and many-to-one, making the task of pronunciation prediction straightforward, at least in principle.

A.6.3 Resources for ASR

The corpus consists of six hours of talk radio from the BBC's *Radio nan Gàidheal*, collected by the University of Edinburgh in 2010. The broadcasts are from the morning news and discussion programme, *Aithris na Maidne* recorded in clean studio conditions and sampled at 48kHz (any telephone speech from callers to the programme was removed). Speech is transcribed by fluent Gaelic speakers at utterance level. The speech data in the corpus can be categorized into three broad genres: read news, reports from correspondents and interviews. Due to the minority status of Gaelic within the UK, the corpus also has a high proportion of English words (853). English words present in the corpus are manually labelled. The corpus does not define a phone set, phone pronunciation lexicon or language model for ASR.

The corpus consists of speech from 46 speakers. This includes 4818 utterances and 5083 unique words. The corpus did not have train, and test set division for the purpose of ASR. Therefore, in [Rasipuram et al., 2013a] we divided the database in to train, development and test sets in a speaker independent way. The training set consists of 22 speakers, 2389 utterances amounting to 3 hours of speech, the development set consists of 12 speakers, 1112 utterances amounting to 1 hour of speech and the test set consists of 12 speakers, 1317 utterances amounting to 1 hour of speech. The test data consists of 2246 unique words which includes 772 words not seen during training.

A.6.4 Pronunciation Lexicon

In this work, the grapheme pronunciation lexicon was created for the words in the database. During the development of grapheme lexicon:

- Vowel graphemes (A, E, I, O, U) and long vowel graphemes or grave accents (À, È, Ì, Ò, Ù) were treated as separate graphemes.
- Lenited consonants (BH, CH, DH, FH, GH, MH, PH, SH and TH) were treated as separate graphemes.
- Consonant graphemes can be broad or slender. However, if the broad/slender assignment is ambiguous (i.e., they can be preceded by a broad vowel and followed by a slender vowel), the consonants are left as they are.
- Word initial and final graphemes were treated as separate graphemes.

Table A.6 presents the list of graphemes in the lexicon. The graphemes J, K, Q, V, W, X, Y and Z, though not present in Gaelic words are present in the grapheme set because of the English words in the corpus. For example, the grapheme pronunciation of Gaelic word "CIAMAR" is [s_C] [I] [A] [b_M] [A] [b_R]. Where 'b_*X*' represents [*X*] is a broad consonant and 's_*X*' represents [*X*] is a slender consonant. However, for English word "AIR" pronunciation is [A] [I] [Rl], i.e., there are no broad and slender consonants. This resulted in total 83 context-independent graphemes. We refer to this lexicon as *knowledge-based* grapheme lexicon. In [Rasipuram et al., 2013a], we used a grapheme lexicon where graphemes at the begin of word, end of word and were treated as separate units. This grapheme lexicon included 248 context-independent graphemes.

In the thesis, we also use another grapheme lexicon that does not use any knowledge, such as broad and slender consonants. We refer to it as *orthography-based* lexicon. This lexicon is transcribed in traditional way from the orthography of words and includes 32 Gaelic graphemes (25 alphabets, 5 accents and silence).

Table A.6 – Graphemes in Gaelic lexicon. 'b_*X*' represents [*X*] is a broad consonant and 's_*X*' represents [*X*] is a slender consonant

Type	Graphemes
Vowels	A, E, I, O, U
Long Vowels	À, È, Ì, Ò, Ù
Broad consonants	b_B, b_BH, b_C, b_CH, b_D, b_DH, b_F, b_FH, b_G, b_GH, b_H, b_L, b_M, b_MH, b_N, b_P, b_PH, b_R, b_RR, b_S, b_SH, b_T, b_TH
Slender consonants	s_B, s_BH, s_C, s_CH, s_D, s_DH, s_F, s_FH, s_G, s_GH, s_H s_L, s_M, s_MH, s_N, s_P, s_PH, s_R, s_RR, s_S, s_SH, s_T, s_TH
Consonants	B, BH, C, CH, D, DH, F, FH, G, GH, H, J, K, L, M, MH, N, P, Q, R, S, T, TH, V, W, X, Y, Z

Bibliography

A.E. Abbas. A Kullback-Leibler View of Linear and Log-Linear Pools. *Decision Analysis*, 6:25–37, 2009.

G. Aradilla. *Acoustic Models for Posterior Features in Speech Recognition*. PhD thesis, EPFL, Switzerland, 2008.

G. Aradilla, J. Vepa, and H. Bourlard. An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features. In *Proc. of ICASSP*, pages 657–660, 2007.

G. Aradilla, H. Bourlard, and M. Magimai Doss. Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task . In *Proc. of Interspeech*, pages 928–931, 2008.

A. Asaei, B. Picart, and H. Bourlard. Analysis of phone posterior feature space exploiting class-specific sparsity and MLP-based similarity measure. In *Proc. of ICASSP*, pages 4886–4889, 2010.

L.R. Bahl, F. Jelinek, and R. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2): 179–190, 1983.

J.R. Bellegarda and D. Nahamoo. Tied mixture continuous parameter modeling for speech recognition. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 38(12):2033–2045, 1990.

L. Besacier, E. Barnard, A. Karpov, and T. Schultz. Automatic Speech Recognition for Under-resourced Languages: A Survey. *Speech Communication*, 56:85–100, 2014.

P. Beyerlein, W. Byrne, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, and W. Wang. Towards language independent acoustic modeling. In *Proc. of ICASSP*, pages 1029–1032, 2000a.

P. Beyerlein, W. Byrne, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, and W. Wang. Towards Language Independent Acoustic Modeling. In *Proc. of ICASSP*, pages 1029–1032, 2000b.

F. Biadsy, P. Moreno, and M. Jansche. Google's Cross-Dialect Arabic Voice Search. In *Proc. of ICASSP*, pages 4441–4444, 2012.

M. Bisani and H. Ney. Joint-Sequence Models for Grapheme-to-Phoneme Conversion. *Speech Communication*, 50:434–451, 2008.

R. E. Blahut. Hypothesis testing and information theory. *IEEE Trans. on Information Theory*, IT-20(4), 1974.

H. Bourlard and N. Morgan. *Connectionist Speech Recognition - A Hybrid Approach*. Kluwer Academic Publishers, 1994.

L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, Kai Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, D. Povey, A. Rastrow, R.C. Rose, and S. Thomas. Multilingual Acoustic Modeling for Speech Recognition based on Subspace Gaussian Mixture Models. In *Proc. of ICASSP*, pages 4334–4337, 2010.

S. F. Chen. Conditional and Joint Models for Grapheme-to-Phoneme Conversion. In *Proc. of EUROSPEECH*, pages 933–936, 2003.

G.E. Dahl, D. Yu, L. Deng, and A. Acero. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition . *IEEE Trans. on Audio, Speech, and Language Processing*, 20:30–42, 2012.

M. Davel and O. Martirosian. Pronunciation dictionary development in resource-scarce environments. In *Proc. of Interspeech*, pages 2851–2854, 2009.

S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

V.V. Digalakis, D. Rtischev, and L.G. Neumeyer. Speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Trans. on Speech and Audio Processing*, 3(5):357–366, Sep 1995.

J. Dines and M. Magimai-Doss. A Study of Phoneme and Grapheme based Context-Dependent ASR Systems. In *Proc. of Machine Learning for Multimodal Interaction (MLMI)*, pages 215–226, 2007.

J. Dines, J. Yamagishi, and S. King. Measuring the Gap Between HMM-Based ASR and TTS. *Selected Topics in Signal Processing, IEEE Journal of*, 4(6):1046–1058, 2010.

S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J. M. Boite. Hybrid HMM/ANN Systems for Training Independent Tasks: Experiments on 'Phonebook' and Related Improvements. In *Proc. of ICASSP*, 1997.

S. Fitt. Documentation and User Guide to UNISYN Lexicon and Postlexical Rules. Technical report, Center for Speech Technology Research, University of Edinburgh, 2000.

G.D. Forney. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

S. Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 34(1):52–59, 1986.

J. Gauvain and Chin-Hui Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *Speech and Audio Processing, IEEE Transactions on*, 2(2):291–298, 1994.

R. Gemello, F. Mana, and S. Scanzio. Experiments on Hiwire Database using Denoising and Adaptation with a Hybrid HMM-ANN Model. In *Proc. of Interspeech*, pages 2429–2432, 2007.

C. Genest and J.V. Zidek. Combining Probability Distributions: A Critique and an Annotated Bibliography. *Statistical Science*, 1:114–148, 1986.

H. Gish, M.H. Siu, A. Chan, and W. Belfield. Unsupervised training of an HMM-based speech recognizer for topic classification. In *Proc. of Interspeech*, pages 1935–1938, 2009.

B. Gold and N. Morgan. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley & amp; Sons, Inc., 1999.

C. Gollan, M. Bisani, S. Kanthak, R. Schluter, and H. Ney. Cross domain automatic transcription on the tc-star epps corpus. In *Proc. of ICASSP*, volume 1, pages 825–828, 2005.

K. Greer, B. Lowerre, and L. Wilcox. Acoustic pattern matching and beam searching. In *Proc. of ICASSP*, volume 7, pages 1251–1254, 1982.

T. Hain. Implicit Modelling of Pronunciation Variation in Automatic Speech Recognition. *Speech Communication*, 46(2):171–188, 2005.

T. Hain and P. C. Woodland. Dynamic HMM Selection for Continuous Speech Recognition. In *Proc. of EUROSPEECH*, pages 1327–1330, 1999.

H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1751, 1990.

H. Hermansky, D. Ellis, and S. Sharma. Tandem Connectionist Feature Extraction for Conventional HMM Systems. In *Proc. of ICASSP*, volume 3, pages 1635–1638, 2000.

G. Hinton, L. Deng, D. Yu, G. Dahl, A-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6): 82–97, 2012.

X.D. Huang and M.A. Jack. Semi-continuous hidden Markov models for speech signal. *Computer Speech and Language*, 3(3):239–251, 1989.

S. Ikbal. *Nonlinear Feature Transformations for Noise Robust Speech Recognition*. PhD thesis, École Polytechnique Fédérale de Lausanne (EPFL), 6 2004.

D. Imseng. *Multilingual speech recognition A posterior based approach*. PhD thesis, École Polytechnique Fédérale de Lausanne (EPFL), June 2013.

D. Imseng, R. Rasipuram, and M. Magimai.-Doss. Fast and Flexible Kullback-Leibler Divergence based Acoustic Modeling for Non-native Speech Recognition. In *Proc. of Automatic Speech Recognition and Understanding (ASRU)*, pages 348–353, 2011.

D. Imseng, H. Bourlard, and P.N. Garner. Using kl-divergence and multilingual information to improve asr for under-resourced languages. In *Proc. of ICASSP*, pages 4869–4872, 2012a.

D. Imseng, J. Dines, P. Motlicek, P.N. Garner, , and H. Bourlard. Comparing different acoustic modeling techniques for multilingual boosting. In *Proc. of Interspeech*, 2012b.

D. Imseng, H. Bourlard, J. Dines, P.N. Garner, and M. Magimai-Doss. Applying Multi- and Cross-Lingual Stochastic Phone Space Transformations to Non-Native Speech Recognition. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 21(8):1713–1726, 2013a.

D. Imseng, P. Motlicek, P.N. Garner, and H. Bourlard. Impact of deep MLP architecture on different acoustic modeling techniques for under-resourced speech recognition. In *Proc. of ASRU*, 2013b.

A. Jansen and K. Church. Towards Unsupervised Training of Speaker Independent Acoustic Models. In *Proc. of Interspeech*, 2011.

A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudanpur, K. Church, N. Feldman, H. Hermansky, F. Metze, R. Rose, M. Seltzer, P. Clark, I. McGraw, B. Varadarajan, E. Bennett, B. Borschinger, J. Chiu, E. Dunbar, A. Fourtassi, D. Harwath, Chia-Ying Lee, K. Levin, A. Norouzian, V. Peddinti, R. Richardson, T. Schatz, and S. Thomas. A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition. In *Proc. of ICASSP*, pages 8111–8115, 2013a.

A. Jansen, S. Thomas, and H. Hermansky. Weak top-down constraints for unsupervised acoustic model training. In *Proc. of ICASSP*, pages 8091–8095, 2013b.

D. Jouvet, D. Fohr, and I. Illina. Evaluating grapheme-to-phoneme converters in automatic speech recognition context. In *Proc. of ICASSP*, pages 4821–4824, 2012.

B-H. Juang and L.R. Rabiner. The segmental K-means algorithm for estimating parameters of hidden Markov models. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 38(9): 1639–1641, 1990.

S. Kanthak and H. Ney. Context-Dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition. In *Proc. of ICASSP*, pages 845–848, 2002.

S. Kanthak and H. Ney. Multilingual Acoustic Modeling using Graphemes. In *Proc. of EUROSPEECH*, pages 1145–1148, 2003.

R.M. Kaplan and M. Kay. Regular models of phonological rule systems. *Computational Linguistics*, 20:331–378, 1994.

S. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 35(3):400–401, 1987.

M. Killer. Grapheme based speech recognition. Master's thesis, Eidgenössische Technische Hochschule Zürich, 2003.

M. Killer, S. Stüker, and T. Schultz. Grapheme based Speech Recognition. In *Proc. of EUROSPEECH*, 2003.

S. King, K. Tokuda, H. Zen, and J. Yamagishi. Unsupervised adaptation for HMM-based speech synthesis. In *Proc. of Interspeech*, 2008.

J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On Combining Classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

R. Kneser and H. Ney. Improved backing-off for M-gram language modeling. In *Proc. of ICASSP*, volume 1, pages 181–184, 1995.

T. Ko and B. Mak. Eigentrigraphemes for under-resourced languages. *Speech Communication*, 56:132–141, 2014.

J. Köhler. Language Adaptation of Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks. In *Proc. of ICASSP*, volume 1, pages 417–420, 1998.

P. Lal and S. King. Cross-lingual automatic speech recognition using tandem features. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(12):2506–2515, 2013.

V-B. Le and L. Besacier. Automatic Speech Recognition for Under-Resourced Languages: Application to Vietnamese Language. *IEEE Trans. on Audio, Speech, and Language Processing*, 17:1471–1482, 2009.

C.J. Leggetter and P.C. Woodland. Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression. In *Proc. ARPA Spoken Language Technology Workshop*, pages 110–115, 1995.

J. Lööf, M. Bisani, Ch, G. Heigold, Björn Hoffmeister, Ch, R. Schlüter, and H. Ney. The 2006 RWTH parliamentary speeches transcription system. In *Proceedings of Int. Conf. Spoken Language Processing*, 2006.

J. Lööf, C. Gollan, and H. Ney. Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a polish speech recognition system. In *Proc. of Interspeech*, 2009.

L. Lu, A. Ghoshal, and S. Renals. Acoustic Data-Driven Pronunciation Lexicon For Large Vocabulary Speech Recognition. In *Proc. of ASRU*, pages 374–379, 2013.

X. Luo and F. Jelinek. Probabilistic Classification of HMM States for Large Vocabulary Continuous Speech Recognition. In *Proc. of ICASSP*, pages 353–356, 1999.

M. Magimai-Doss, T.A. Stephenson, H. Bourlard, and S. Bengio. Phoneme-Grapheme Based Speech Recognition System. In *Proc. of ASRU*, pages 94–98, 2003.

M. Magimai.-Doss, S. Bengio, and H. Bourlard. Joint Decoding for Phoneme-Grapheme Continuous Speech Recognition. In *Proc. of ICASSP*, volume 1, pages 177–180, 2004.

M. Magimai.-Doss, R. Rasipuram, G. Aradilla, and H. Bourlard. Grapheme-based Automatic Speech Recognition using KL-HMM. In *Proc. of Interspeech*, pages 445–448, 2011.

I. McGraw, I. Badr, and J.R. Glass. Learning Lexicons From Speech Using a Pronunciation Mixture Model. *IEEE Trans. on Audio, Speech, and Language Processing*, 21(2):357–366, 2013.

B. Mimer, S. Stüker, and T. Schultz. Flexible Decision Trees for Grapheme based Speech Recognition. In *Elektronische Sprachsignalverarbeitung, Cottbus, Germany*, 2004.

N. Morgan and H. Bourlard. Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach. *IEEE Signal Processing Magazine*, pages 25–42, 1995.

J. Novak. Phonetisaurus: A WFST-driven Phoneticizer. http://code.google.com/p/phonetisaurus/, 2011.

J. P. Openshaw, Z. P. Sun, and J.S. Mason. A comparison of composite features under degraded speech in speaker recognition. In *Proc. of ICASSP*, volume 2, pages 371–374, 1993.

D. O'Shaughnessy. *Speech Communication - Human and Machine*. Addison-Wesley, 1987.

L. Osterholtz, C. Augustine, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, and Alex Waibel. Testing generality in JANUS: a multi-lingual speech translation system. In *Proc. of ICASSP*, pages 209–212, 1992.

Daniel P, Arnab G, Gilles B, Lukas B, Ondrej G, Nagendra G, Mirko H, Petr M, Yanmin Q, Petr S, Jan S, Georg S, and Karel V. The Kaldi Speech Recognition Toolkit. In *Proc. of ASRU*, 2011.

V. Pagel, K. Lenzo, and A.W. Black. Letter to Sound Rules for Accented Lexicon Compression. In *Proceedings of Int. Conf. Spoken Language Processing*, 1998.

A.S. Park and J.R. Glass. Towards Unsupervised Pattern Discovery in Speech. In *Proc. of ASRU*, pages 53–58, 2005.

D.B. Paul and J.M. Baker. The Design for the Wall Street Journal-based CSR Corpus. In *DARPA Speech and Language Workshop*. Morgan Kaufmann Publishers, 1992.

J. Pinto, Magimai-Doss, and H. Bourlard. MLP based hierarchical system for task adaptation in ASR. In *Proc. of ASRU*, pages 365–370, Nov 2009.

J.P. Pinto, G. S. V. S. Sivaram, M. Magimai.-Doss, H. Hermansky, and H. Bourlard. Analysis of MLP Based Hierarchical Phoneme Posterior Probability Estimator. *IEEE Trans. on Audio, Speech, and Language Processing*, 19:225–241, 2011.

J. Pitrelli, C. Fong, S.H. Wong, J.R. Spitz, and H.C. Leung. PhoneBook: a phonetically-rich isolated-word telephone-speech database. In *Proc. of ICASSP*, volume 1, pages 101–104, 1995.

D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas. The subspace Gaussian mixture model - A structured model for speech recognition. *Computer Speech and Language*, 2011.

P. J. Price, W. Fisher, and J. Bernstein. The DARPA 1000-word resource management database for continuous speech recognition. In *Proc. of ICASSP*, pages 651–654, 1988.

L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of IEEE*, 77(2):257–286, 1989.

R. Rasipuram and M. Magimai.-Doss. Acoustic Data-driven Grapheme-to-Phoneme Conversion using KL-HMM. In *Proc. of ICASSP*, pages 4841–4844, 2012a.

R. Rasipuram and M. Magimai.-Doss. Combining Acoustic Data Driven G2P and Letter-to-Sound Rules for Under Resource Lexicon Generation. In *Proc. of Interspeech*, 2012b.

R. Rasipuram and M. Magimai.-Doss. Improving Grapheme-based ASR by Probabilistic Lexical Modeling Approach. In *Proc. of Interspeech*, pages 505–509, 2013a.

R. Rasipuram and M. Magimai.-Doss. Probabilistic Lexical Modeling and Grapheme-based Automatic Speech Recognition. http://publications.idiap.ch/downloads/reports/2013/Rasipuram_Idiap-RR-15-2013.pdf, 2013b. Idiap Research Report.

R. Rasipuram and M. Magimai.-Doss. Acoustic and Lexical Resource Constrained ASR using Language-Independent Acoustic Model and Language-Dependent Probabilistic Lexical Model. http://publications.idiap.ch/downloads/reports/2014/Rasipuram_Idiap-RR-02-2014.pdf, 2014. Idiap Research Report.

R. Rasipuram, P. Bell, and M. Magimai.-Doss. Grapheme and Multilingual Posterior Features for Under-Resourced Speech Recognition: A Study on Scottish Gaelic. In *Proc. of ICASSP*, pages 7334–7338, 2013a.

R. Rasipuram, M. Razavi, and M. Magimai-Doss. Probabilistic Lexical Modeling and Unsupervised Training for Zero-Resourced ASR. In *Proc. of ASRU*, pages 446–451, 2013b.

M. Razavi, R. Rasipuram, and M. Magimai-Doss. On Modeling Context-dependent Clustered States: Comparing HMM/GMM, Hybrid HMM/ANN and KL-HMM Approaches. In *Proc. of ICASSP*, 2014.

J. Rottland and G. Rigoll. Tied Posteriors: An Approach for Effective Introduction of Context Dependency in Hybrid NN/HMM LVCSR. In *Proc. of ICASSP*, pages 1241–1244, 2000.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

L. Saheer, J. Dines, and P.N. Garner. Vocal Tract Length Normalization for Statistical Parametric Speech Synthesis. *IEEE Trans. on Audio, Speech, and Language Processing*, 20(7):2134–2148, 2012.

M. Saraclar, H. Nock, and S. Khudanpur. Pronunciation Modeling By Sharing Gaussian Densities Across Phonetic Models. In *Computer Speech and Language*, pages 137–160, 2000.

T. Schlippe, E.G.K. Djomgang, N.T Vu, S. Ochs, and T. Schultz. Hausa Large Vocabulary Continuous Speech Recognition. In *Proc.of the Spoken Languages Technologies for Under-resourced Languages (SLTU)*, 2012.

E.G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck. Automatic Speech Recognition Without Phonemes. In *Proc. of European Conf. on Speech Communication and Technology (EUROSPEECH)*, 1993.

T. Schultz and K. Kirchhoff. *Multilingual Speech Processing*. Academic Press, 2006.

T. Schultz and A. Waibel. Experiments on Cross-Language Acoustic Modeling. In *Proc. of EUROSPEECH*, pages 2721–2724, 2001a.

T. Schultz and A. Waibel. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35:31–51, 2001b.

R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul. Context-dependent modeling for acoustic-phonetic recognition of continuous speech. In *Proc. of ICASSP*, volume 10, pages 1205–1208, 1985.

J.C. Segura, T. Ehrette, A. Potamianos, D. Fohr, I. Illina, P-A. Breton, V. Clot, R. Gemello, M. Matassoni, and P. Maragos. The HIWIRE Database, a Noisy and Non-native English Speech Corpus for Cockpit Communication. http://cvsp.cs.ntua.gr/projects/pub/HIWIRE/WebHome/HIWIRE_db_description_paper.pdf, 2007.

T. J. Sejnowski and C. R. Rosenberg. Parallel networks that learn to pronounce English text. *Complex Systems*, 1:145–168, 1987.

K.C. Sim. Discriminative Product-of-Expert acoustic mapping for cross-lingual phone recognition. In *Proc. of ASRU*, pages 546–551, 2009.

K.C Sim and H. Li. Robust phone set mapping using decision tree clustering for cross-lingual phone recognition. In *Proc. of ICASSP*, pages 4309–4312, 2008.

M.H. Siu, H. Gish, A. Chan, W. Belfield, and S. Lowe. Unsupervised training of an hmm-based self-organizing unit recognizer with applications to topic classification and keyword discovery. *Computer Speech and Language*, 28(1):210–223, 2014.

S. Soldo, M. Magimai.-Doss, J.P Pinto, and H. Bourlard. Posterior Features for Template-based ASR. In *Proc. of ICASSP*, pages 4864–4867, 2011.

A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of Int. Conf. Spoken Language Processing*, pages 901–904, 2002.

A. Stolcke, M. Hwang, X. Lei, N. Morgan, and D. Vergyri. Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons. In *Proc. of ICASSP*, pages 321–324, 2006.

H. Strik and C. Cucchiarini. Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, 29:225–246, 1999.

S. Stüker. Modified Polyphone Decision Tree Specialization for Porting Multilingual Grapheme Based ASR Systems to New Languages. In *Proc. of ICASSP*, pages 4249–4252, 2008a.

S. Stüker. Integrating Thai Grapheme Based Acoustic Models into the ML-MIX Framework - For Language Independent and Cross-Language ASR. In *Proc. of SLTU*, 2008b.

S. Stüker. *Acoustic Modeling for Under-Resourced Languages*. PhD thesis, Karlsruhe Institute of Technology (KIT), July 2009.

S. Stüker and T. Schultz. A Grapheme Based Speech Recognition System for Russian. In *Proc. of Speech and Computer (SPECOM)*, 2004.

Y-H Sung, T. Hughes, F. Beaufays, and B. Strope. Revisiting Graphemes with Increasing Amounts of Data. In *Proc. of ICASSP*, pages 4449–4452, 2009.

P. Swietojanski, A. Ghoshal, and S. Renals. Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR. In *Proc. of ICASSP*, pages 246–251, 2012.

P. Taylor. Hidden Markov Models for Grapheme to Phoneme Conversion. In *Proc. of Interspeech*, 2005.

P. Taylor, A. Black, and R. Caley. The Architecture of the Festival Speech Synthesis System. In *Proc. of ESCA Workshop on Speech Synthesis*, 1998.

S. Thomas and H. Hermansky. Cross-lingual and multistream posterior features for low resource lvcsr systems. In *Proc. of Interspeech*, pages 877–880, 2010.

S. Thomas, S. Ganapathy, and H.Hermansky. Multilingual MLP features for low-resource LVCSR systems. In *Proc. of ICASSP*, pages 4269–4272, 2012.

R. Veldhuis. The Centroid of the Symmetrical Kullback-Leibler Distance. *IEEE Signal Processing Letters*, 9:96–99, 2002.

N.T. Vu, F. Kraus, and T. Schultz. Multilingual a-stabil: A new confidence score for multilingual unsupervised training. In *Proc. of IEEE Spoken Language Technology Workshop (SLT)*, pages 183–188, 2010.

D. Wang and S. King. Letter-to-Sound Pronunciation Prediction Using Conditional Random Fields. *Signal Processing Letters, IEEE*, 18(2):122–125, 2011.

M. Wolters. A Diphone-Based Text-to-Speech System for Scottish Gaelic. Master's thesis, University of Bonn, 1997.

P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young. Large Vocabulary Continuous Speech Recognition using HTK. In *Proc. of ICASSP*, volume 2, pages 125–128, 1994.

L. Xiao, A. Gunawardana, and A. Acero. Adapting grapheme-to-phoneme conversion for name recognition. In *Proc. of ASRU*, pages 130–135, 2007.

S. J. Young, J. J. Odell, and P. C. Woodland. Tree-based State Tying for High Accuracy Acoustic Modelling. In *Proceedings of the Workshop on Human Language Technology (HLT)*, pages 307–312, 1994.

S.J. Young, G. Evermann, M.J.F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C.Woodland. *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, UK, 2006.

H. Yu and T. Schultz. Enhanced tree clustering with single pronunciation dictionary for conversational speech recognition. In *in Proc. EuroSpeech*, pages 1869–1872, 2003.

Q. Zhu, B. Chen, N. Morgan, and A. Stolcke. On using MLP features in LVCSR. In *Proceedings of Int. Conf. Spoken Language Processing*, pages 921–924, 2004.

Ramya Rasipuram

Rue de la Maladiere, 8
1920, Martigny
00 41 78 749 00 60
FAX 00 41 27 721 77 12
ramya.rasipuram@idiap.ch
http://www.idiap.ch/ rramya/

Current

Since 2010	Research assistant, IDIAP Research Institute, Martigny,Switzerland.
Project	Flexible Grapheme based Automatic Speech Recognition
Research Interests	Automatic Speech Recognition (ASR), ASR for under-resourced languages, Non-native speech recognition, Articulatory features for ASR

Education

Since 2010	**Ph.D. Student**, *Electrical Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL)*, Lausanne, Switzerland.
2005-2008	**Master of Science**, *Electrical Engineering, Indian Institute of Technology Madras*, Chennai, India.
2000-2005	**Bachelor of Engineering**, *Electronics and Communications Engineering, Sri Krishnadevaraya University*, Andhra Pradesh, India.

Ph.D. Thesis

Title	*Grapheme-based automatic speech recognition using probabilistic lexical modeling*
Supervisors	Prof. Herve Bourlard & Dr. Mathew Magimai Doss
Description	My thesis investigates the use of graphemes as sub-word units for automatic speech recognition in a principled way by exploiting both grapheme representation and phoneme information in the framework of probabilistic lexical modeling. The outcome of my research provides means to build flexible automatic speech recognition systems for under-resourced and minority languages that lack proper resources.

Experience

2008–2010	**Research engineer**, *Applied Research Group, Satyam Computer Services Ltd*, Bangalore, India. Projects involved: Long term utility speaker recognition, Keyword spotting in action movies.
2005-2008	**Project associate**, *Indian Institute of Technology Madras*, Chennai, India. Projects involved: Multimodal interfaces to computer

Professional Activities

IEEE and IEEE SPS student member

ISCA student member

Reviewer for Speech Communication, Elsevier

Awards

One of the five finalists for best student paper award at ASRU 2011

Recipient of student travel grant by the EUSIPCO-2008

Recipient of Interspeech-2008 student grant

University gold medallist during under-graduation

Computer skills

Programming Languages: C, C++

Scripting Languages: PERL, SHELL

Tools: Hidden Markov Model Toolkit (HTK), Matlab, Octave , Latex

Operating System: Windows, Linux

Publications

Conference Proceedings

[1] Marzieh Razavi, **Ramya Rasipuram**, and Mathew Magimai.-Doss. On Modeling Context-Dependent Clustered States: Comparing HMM/GMM, Hybrid HMM/ANN and KL-HMM Approaches. In *International Conference on Acoustics, Speech, and Signal Processing*, 2014.

[2] **Ramya Rasipuram**, Marzieh Razavi, and Mathew Magimai.-Doss. Probabilistic Lexical Modeling and Unsupervised Training for Zero-Resourced ASR. In *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, pages 446–451, 2013.

[3] **Ramya Rasipuram** and Mathew Magimai.-Doss. Improving Grapheme-based ASR by Probabilistic Lexical Modeling Approach. In *Proceedings of Interspeech*, pages 505–509, 2013.

[4] **Ramya Rasipuram**, Peter Bell, and Mathew Magimai.-Doss. Grapheme and Multilingual Posterior Features For Under-Resource Speech Recognition: A Study on Scottish Gaelic. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7334–7338, 2013.

[5] **Ramya Rasipuram** and Mathew Magimai.-Doss. Combining Acoustic Data Driven G2P and Letter-to-Sound Rules for Under Resource Lexicon Generation. In *Proceedings of Interspeech*, 2012.

[6] David Imseng, **Ramya Rasipuram**, and Mathew Magimai.-Doss. Fast and Flexible Kullback-Leibler Divergence based Acoustic Modeling for Non-Native Speech Recog-

nition. In *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, pages 348–353, December 2011.

[7] Mathew Magimai.-Doss, **Ramya Rasipuram**, Guillermo Aradilla, and Hervé Bourlard. Grapheme-based Automatic Speech Recognition using KL-HMM. In *Proceedings of Interspeech*, pages 445–448, August 2011.

[8] **Ramya Rasipuram** and Mathew Magimai.-Doss. Improving Articulatory Feature and Phoneme Recognition using Multitask Learning. In *Artificial Neural Networks and Machine Learning - ICANN 2011*, volume 6791 of *Lecture Notes in Computer Science*, pages 299–306. Springer Berlin / Heidelberg, 2011.

[9] **Ramya Rasipuram** and Mathew Magimai.-Doss. Integrating Articulatory Features using Kullback-Leibler Divergence based Acoustic Model for Phoneme Recognition. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5192 – 5195, 2011.

[10] **Ramya Rasipuram**, Rajesh M Hegde, and Hema A Murthy. Significance of Group Delay based Acoustic Features in the Linguistic Search Space for Robust Speech Recognition. In *Proceedings of Interspeech*, 2008.

[11] **Ramya Rasipuram**, Rajesh M Hegde, and Hema A Murthy. Incorporating Acoustic Feature Diversity into the Linguistic Search Space for Syllable Based Speech Recognition. In *Proceedings of European Signal Processing Conference*, 2008.

Submitted Articles

[1] **Ramya Rasipuram** and Mathew Magimai.-Doss. Acoustic and Lexical Resource Constrained ASR using Language-Independent Acoustic Model and Language-Dependent Probabilistic Lexical Model. `http://publications.idiap.ch/downloads/reports/2014/Rasipuram_Idiap-RR-02-2014.pdf`, 2014. Submitted to Speech Communication.

[2] **Ramya Rasipuram** and Mathew Magimai.-Doss. Articulatory Feature based Continuous Speech Recognition using Probabilistic Lexical Modeling, 2014. Submitted to Computer Speech and Language.

Research Reports

[1] **Ramya Rasipuram** and Mathew Magimai.-Doss. Probabilistic Lexical Modeling and Grapheme-based Automatic Speech Recognition. `http://publications.idiap.ch/downloads/reports/2013/Rasipuram_Idiap-RR-15-2013.pdf`, 2013. Idiap Research Report.

[2] **Ramya Rasipuram** and Mathew Magimai.-Doss. Mulitask Learning to Improve Articulatory Feature Estimation and Phoneme Recognition. `http://publications.idiap.ch/downloads/reports/2011/Rasipuram_Idiap-RR-21-2011.pdf`, 2011. Idiap Research Report.