On Modeling the Synergy Between Acoustic and Lexical Information for Pronunciation Lexicon Development

THÈSE N^{O} 7851 (2017)

PRÉSENTÉE LE 10 AOÛT 2017
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

Marzieh RAZAVI

acceptée sur proposition du jury:

Prof. S. Süsstrunk, présidente du jury
Prof. H. Bourlard, Dr M. Magimai Doss, directeurs de thèse
Dr K. Knill, rapporteuse
Prof. M. Davel, rapporteuse
Prof. J.-Ph. Thiran, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2017

Two there are who are never satisfied –
the lover of the world
and the lover of knowledge.
— Rumi

To my family. . .

Acknowledgements

First and foremost, I would like to express my sincere gratitude toward Dr. Mathew Magimai.-Doss for his guidance, dedication and support during my PhD career. I learned a lot from him, not only from a professional standpoint through our discussions, but also from a personal standpoint to always stay positive. I am also very much grateful to Prof. Hervé Bourlard, for his constructive feedbacks and for providing an amazing academic atmosphere at Idiap Research Institute. I would like to sincerely thank the senior scientists at Idiap, in particular Dr. Phil Garner and Dr. Petr Motlicek for their constructive feedbacks. I would also like to express my profound gratitude to the committee members of my thesis defense, Prof. Marelie Davel, Dr. Kate Knill, Prof. Jean-Philippe Thiran and Prof. Sabine Süsstrunk. Their valuable comments and feedbacks helped me improve this thesis, of which I am very appreciative. This work was supported by the Hasler foundation through the grant "Flexible acoustic data driven grapheme to acoustic unit conversion".

I have been very fortunate to have the companionship of amazing colleagues at Idiap. In particular, I am very much indebted to Ramya for her guidance and incredible patience not only during our joint works, but also whenever I encountered a problem. I would also like to thank Ramya for her countless kindness and amazing friendship. I am also very much grateful to Afsaneh for her profound insights, positive energy and cheerful friendship. I would also like to sincerely thank David for always being available to answer questions, and Pierre-Edouard for helping out with the French abstract in a short notice. I extend my profound gratitude to Lakshmi, Marc, Srikanth, Raphael, Alexandros, Alexandre, Milos, Dimitri, Hannah, Harsha, Subhadeep, Pranay, Dhananjay, Sibo, Sucheta, Yang, Gulcan, Nikos, Mohammad, Oya, Ajay and Kenneth. I would also like to thank the administrative team at Idiap, in particular Mme. Nadine Rousseau and Mme. Sylvie Millius for all their help and support, and the Idiap IT team, especially Louis-Marie, Frank and Norbert for their prompt help with the technical issues.

I have shared numerous joyful moments with my Iranian friends in Lausanne and Martigny, that can hardly be acknowledged in words. I would like to especially thank Fatemeh Ghadamieh, Zahra Ejehi, Fereshteh, Hoda, Maliheh, Mahshid, Asieh, Mojgan, Niloofar, Fatemeh Rahimpour, Fatemeh Saleh, Maryam Habibi, Mahya, Maryam Gol, Zahra Moheb, Zeinab, Faezeh, Zohreh, Sara, Gelareh, Samira, Haleh, Noushin, Simin, Razieh, Mitra, Golnoush, Negar and Helma. Furthermore, I would like to deeply thank my friends from my Master's and

Acknowledgements

Bachelor's studies, especially, Neda, Mahshid, Parnian, Afra, Nastaran, Parisa, Zahra, Yasaman, Marjan, Hengameh, Maryam, Hoda, Shabnam, Naghmeh, Parastoo and Sara.

Finally, I would like to express my endless gratitude toward my family. I am extremely fortunate for having been blessed with my incredible parents who always supported me with their unconditional love and their blessings. My Sisters Mariam and Razieh, with all their care and love and my brother Mohsen with his unconditional support, as well as my caring mother-in-law Mina, my supportive brothers-in-law Afshin and Mehrdad, my lovely sisters-in-law Sanaz and Mona, my cheerful nephews Mohsen, Matin and Iliya and my adorable niece Nika are among the greatest treasures of my life. They have always hosted us with open arms during our holiday trips to Iran and UK, and the joyful moments of family gatherings and playing with my niece and nephews have always boosted my energy. Last but not least, I would like to express my heartfelt acknowledgment to my husband Alireza, for always supporting me, for encouraging me during the though times, and for his unwavering love. I cannot express in words how grateful and blessed I am for having him in my life, and I cannot thank him enough for his unconditional love and support. This dissertation is dedicated to my family for their constant love, support and encouragement.

Lausanne, 1 June 2017 Marzieh Razavi

Abstract

State-of-the-art automatic speech recognition (ASR) and text-to-speech systems require a pronunciation lexicon that maps each word to a sequence of phones. Manual development of lexicons is costly as it needs linguistic knowledge and human expertise. To facilitate this process, grapheme-to-phone (G2P) conversion approaches are used, in which given a *seed lexicon* provided by linguistic experts, the G2P relationship is learned by applying statistical techniques. Despite advances in these approaches, there are two challenges remaining: (1) the seed lexicon development through linguistic expertise incorporates limited acoustic information, which may not necessarily cover all natural phonological variations, and (2) the linguistic expertise required for the development of the seed lexicon may not be available for all languages, particularly under-resourced languages. The goal of this thesis is to address these challenges by developing a framework that effectively integrates linguistic information and acoustic data for pronunciation lexicon development.

To achieve that goal, we first study the problem of matching a word hypothesis to the acoustic signal, and show that the hidden Markov model-based ASR approach achieves that match via a latent symbol set. Building on that understanding, we develop a data-driven G2P conversion approach in which a probabilistic G2P relationship is learned by matching the acoustic signal with the word hypothesis represented by graphemes, using phones as the latent symbols. Through a theoretical development, we show that this acoustic G2P conversion approach is a particular case of an abstract posterior-based G2P conversion formalism, which requires estimation of phone class conditional probabilities. Through studies on two languages, we show that the acoustic G2P conversion approach yields lexicons that can perform comparable to state-of-the-art G2P conversion methods at the ASR level, despite performing relatively poorly at pronunciation level.

We build on the posterior-based formalism to show that different G2P conversion approaches in the literature can be regarded as different estimators of phone class conditional probabilities, and can be combined in a multi-stream fashion to yield better lexicons. We also demonstrate that the multi-stream formulation can be further extended to unify acoustic-to-phone conversion and G2P conversion. We validate the proposed multi-stream formulation on two challenging tasks on English.

Finally, we address the issue of developing lexical resources for under-resourced languages by proposing an acoustic subword unit (ASWU)-based lexicon development approach. In this approach, ASWU derivation is cast as the problem of determining a latent symbol space given the word hypothesis and acoustics, and the pronunciations are generated using the

proposed acoustic G2P conversion approach. Through experimental studies and analysis on well-resourced and under-resourced languages, we show that the derived ASWUs are "phone-like", and the ASWU-based lexicons yield better ASR systems compared to the alternative grapheme-based lexicons.

Keywords Phonetic lexicon development, grapheme-to-phone conversion, acoustic sub-word unit discovery, hidden Markov model, automatic speech recognition, under-resourced languages.

Résumé

L'état de l'art des systèmes de reconnaissance automatique de la parole (RAP) et de synthèse vocale repose sur l'utilisation d'un lexique de prononciation qui associe chaque mot à une séquence de phones correspondante. La création manuelle de tels lexiques est coûteuse car elle nécessite des connaissances linguistiques et une expertise humaine. Pour faciliter ce processus de création, des méthodes de conversion graphème-à-phone (GAP) existent : étant donné un *lexique de base* construit par des experts linguistes, la relation GAP est apprise en appliquant des techniques statistiques. Malgré les acancées dans ces domaines, il reste deux difficultés : (1) le développement du lexique de base en utilisant une expertise linguistique incorpore des informations acoustiques limitées, qui risquent de ne pas couvrir toutes les variations phonologiques naturelles, et (2) l'expertise linguistique nécessaire au développement de ces lexiques de base peut ne pas être disponible pour toutes les langues, en particulier les langues ayant peu de ressources. L'objectif de cette thèse est de s'attaquer à ces défis via le développement d'un système qui intègre à la fois des informations linguistiques et des données acoustiques pour la création de lexiques de prononciation.

Pour atteindre cet objectif, nous étudions dans un premier temps le problème d'association d'une hypothèse sur un mot avec le signal acoustique, et démontrons qu'une approche de RAP basée sur des modèles de Markov cachés permet d'obtenir cette correspondance à travers l'utilisation d'un ensemble de symboles latents. En se basant sur cette observation, nous développons une approche de conversion GAP axée sur les données : une relation probabiliste GAP est apprise en associant le signal acoustique avec l'hypothèse de mot représentée par des graphèmes, où les phones sont les symboles latents. À travers un développement théorique, nous démontrons que cette approche de conversion GAP acoustique est un cas particulier de formalisme abstrait de conversion GAP basée sur des probabilités postérieures et qui nécessite l'estimation de probabilités conditionnelles des classes de phones. En se basant sur des études dans deux langues, nous montrons que l'approche de conversion GAP acoustique produit des lexiques comparables à l'état de l'art des méthodes de conversion GAP pour la RAP, malgé des performances relativement faibles au niveau de la prononciation.

Nous nous appuyons sur ce formalisme pour montrer que différentes approches de conversion GAP dans la littérature peuvent être interprétées comme différents estimateurs de probabilités conditionnelles des classes de phones, et peuvent être combinées de manière multi-flux pour obtenir de meilleurs lexiques. Nous démontrons également que la formulation multi-flux peut être étendue pour unifier les conversions GAP et acoustique-à-phone. La formulation multi-flux proposée est validée sur deux tâches difficiles en anglais.

Enfin, nous nous attaquons au défi de la construction de ressources lexicales pour les langues ayant peu de données en proposant une approche de développement de lexique basée sur les unités acoustiques au niveau du sous-mot (UASM). Dans cette approche, la dérivation des UASMs est définie comme le problème qui consiste à déterminer un espace de symboles latents étant donné l'hypothèse de mot et le signal acoustique, et les prononciations sont générées en utilisant l'approche de conversion GAP acoustique proposée. À l'aide d'études expérimentales et d'analyses sur des langues ayant beaucoup de données et des langues ayant peu de données, nous montrons que les UASMs obtenues sont "semblables aux phone", et que les lexiques basés sur les UASMs permettent d'obtenir de meilleurs systèmes de RAP par rapport aux lexiques alternatifs basés sur les graphèmes.

Mots-clés Développement de lexiques phonétiques, conversion graphème-à-phone, découverte d'unités acoustiques au niveau du sous-mot, modèle de Markov caché, reconnaissance automatique de la parole, langues ayant peu de ressources.

Contents

List of Figures

List of Tables

List of Acronyms

A2P acoustic-to-phone
AF articulatory feature
AG2P acoustic grapheme-to-phone conversion
ANN artificial neural network
ASR automatic speech recognition
ASWU acoustic subword unit
CART classification and regression tree
CD context-dependent
cCD clustered context-dependent
CI context-independent
CNN convolutional neural network
CRF conditional random field
DCT discrete cosine transform
EM expectation maximization
FFT fast Fourier transform
G2P grapheme-to-phone
GMM Gaussian mixture model
HCRF hidden conditional random field
HMM/ANN hidden Markov model system using artificial neural networks
HMM/GMM hidden Markov model system using Gaussian mixture models
HMM hidden Markov model
HTK hidden Markov model toolkit
JMM joint multigram model
KL-HMM Kullback-Leibler divergence-based hidden Markov model
KL Kullback-Leibler
L2S letter-to-sound
LD Levenshtein distance
LSTM long short-term memory
MFCC Mel-frequency cepstrum coefficient
MLP multilayer perceptron
OOD out-of-domain
OOV out-of-vocabulary
PC-HMM probabilistic classification of HMM states

PLP perceptual linear prediction
PRR phone recognition rate
RM resource mangament
RNN recurrent neural network
SAMPA speech assessment methods phonetic alphabet
SMT statistical machine translation
TTS text-to-speech
Tied-HMM tied posterior-based hidden Markov model
WFST weighted finite state transducer
WPA word level pronunciation accuracy
WRR word recognition rate
WSJ wall street journal
i.i.d. independent and identically distributed

1 Introduction

Speech technologies such as automatic speech recognition (ASR) systems and text-to-speech (TTS) systems aim to link two modes of communication, namely the spoken form (i.e., speech signal) and the written form (i.e., text). In order to model the relation between the two forms, an intermediate unit space is commonly used. The intermediate units can be the whole words or, as shown in Figure 1.1, can be subword units. Subword units are preferred to words especially in large vocabulary tasks for two main reasons: (1) they are easily trainable compared to the whole words as the frequency of words in a text follows Zipf's law [Powers, 1998], and (2) they are generalizable for unseen words. On the other hand, using subword units in speech technologies brings two main questions: (1) how to decide on subword units for a specific language?, and (2) how to represent each word in terms of subword units?

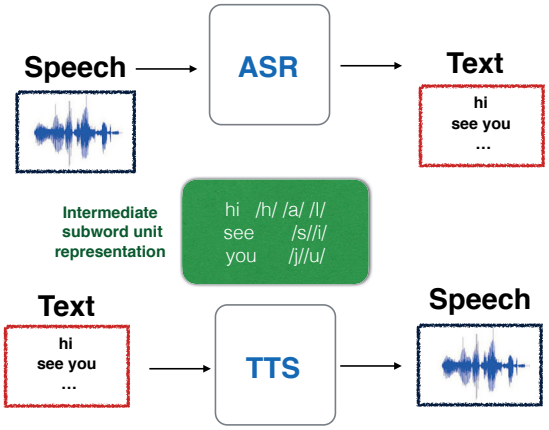


Figure 1.1 – Schematic view of ASR and TTS systems.

Answering these questions depends on the linguistic knowledge and resources available for the language of the interest. This thesis addresses these questions on both well-resourced languages, for which sufficient linguistic knowledge and resources are available, and under-resourced languages, for which limited linguistic knowledge and resources are available.

1.1 Motivation and objectives

State-of-the-art speech processing systems use phones as subword units. The popularity of phones comes from their relation to both spoken and written forms. The link between phones and the spoken form (i.e., the speech signal) comes from the fact that the envelope of magnitude spectrum of short-term speech signal typically depicts the characteristics of phones. The link between phones and graphemes originates from the alphabetic orthographies, which aim to present the phonetic structure of the spoken words in a graphic form [Frost, 1989].

Using phones as subword unit entails development of a pronunciation lexicon providing phonetic representation(s) for each word. A phonetic lexicon can be developed manually through use of linguistic knowledge. However, manual development of lexicons can be costly in terms of time and money [Davel and Barnard, 2003]. In addition, the developed lexicons are required to be constantly augmented with the evolution of languages and emergence of new words. Therefore, it is necessary to develop automatic pronunciation generation methods to reduce the amount of human effort. Toward that goal, grapheme-to-phone (G2P) conversion methods are applied in which given an initial phonetic lexicon called a *seed lexicon* provided by linguistic experts, typically data-driven and machine learning techniques such as decision trees [Black et al., 1998] or conditional random fields (CRFs) [Wang and King, 2011] are used to learn the G2P relationship. The learned G2P relationship is then used to infer pronunciations for the unseen words. The G2P conversion approaches have facilitated the development of phonetic lexicons and reduced the amount of human effort. However, they still encounter two main challenges:

1. *They rely on the availability of linguistic knowledge in the target language.* Data-driven G2P conversion approaches require a seed lexicon as the training data to learn the G2P relationship. The seed lexicon is obtained manually by employing linguistic knowledge and human expertise. Such a lexicon is readily available for well-resourced languages such as English, French and German. For under-resourced languages that lack proper lexical resources such as Scottish Gaelic and Haitian Creole, however, obtaining an initial phonetic lexicon is not trivial.[1] This issue makes the development of a phonetic lexicon for under-resourced languages very challenging.

2. *They do not incorporate the available acoustic information in the G2P relationship learning process.* Most of the proposed approaches in the literature for pronunciation generation rely only on the seed lexicon for learning the G2P relationship. During the process of development of a phone set and a seed lexicon by experts, both linguistic knowledge and acoustic information are incorporated. However, the acoustic information is based on limited acoustic examples, mainly used to identify minimal pairs. Therefore, a possibly large amount of acoustic data that is available during training

[1] There are approaches which employ bootstrapping techniques to accelerate pronunciation lexicon development for under-resourced languages [Davel and Barnard, 2006, Maskey et al., 2004]. These approaches, however, still require a human to verify the generated pronunciations from a G2P conversion approach.

speech technology systems is not exploited for pronunciation generation. Another issue with pronunciation generation using only the information in the seed lexicon is that the generated pronunciations may not capture the natural phonological variations. For example, this can happen in spontaneous speech when some of the sound units are dropped [Strik and Cucchiarini, 1999]; or when a G2P converter trained on baseform pronunciations is used to expand the lexicon for a non-native ASR system. As a result, the generated pronunciations using the existing G2P conversion approaches may not match well with the acoustic data at the application level.

The goal of the thesis is to develop a framework that effectively models the synergy between acoustic information and linguistic information, and addresses the aforementioned challenges for pronunciation lexicon development.

1.2 Contributions of the thesis

In this thesis, we first study the problem of matching an acoustic signal with a word hypothesis. We elucidate that this matching can be obtained through a latent symbol space that is shared between both the acoustic signal and the word hypothesis. We demonstrate that different ASR approaches such as hidden Markov model/Gaussian mixture model (HMM/GMM) [Rabiner, 1989], hybrid HMM/artificial neural network (ANN) [Morgan and Bourlard, 1995], and Kullback-Leibler divergence-based HMM (KL-HMM) [Aradilla, 2008] are variants of this problem, where the context-independent or clustered context-dependent phones serve as the latent symbols. Furthermore, we show that the KL-HMM approach has the inherent capability to achieve its best performance in a relatively small latent symbol space, compared to HMM/GMM and hybrid HMM/ANN approaches [Razavi et al., 2014, Razavi and Magimai.-Doss, 2014]. Building on that understanding, we then propose,

- *A posterior-based formalism for G2P conversion, enabling integrating acoustic information into the G2P relationship learning process*: We propose a posterior-based formalism for G2P conversion in an HMM framework, which requires estimation of the probability of phones given graphemes. We show that the phone class conditional probabilities given graphemes can be estimated through acoustics, by casting the problem of learning the G2P relationship as matching the acoustic signal represented by acoustic features with the word hypothesis represented with graphemes using phones as the latent symbols. Furthermore, we show that the probability of phones given graphemes can be estimated using the seed lexicon through the existing local classification-based G2P conversion approaches. Through experimental studies on two well-resourced languages with deep orthographies, namely English and French, we show that the acoustic data-driven G2P conversion approach can not only perform comparable to state-of-the-art G2P conversion approaches at ASR level, but can also provide complementary information to these approaches [Razavi et al., 2016].

- *Unifying multiple pronunciation extraction approaches*: G2P conversion is one approach for pronunciation generation, and there are different techniques to achieve that. Another approach for pronunciation extraction is to employ a phone recognition technique to generate a phonetic transcription of a word given its acoustic realization(s), referred to here as acoustic-to-phone (A2P) conversion approach. In this thesis, instead of viewing different pronunciation generation approaches as separate techniques, we regard them as different estimators of the phone class conditional probabilities. In that perspective, we build on the proposed posterior-based G2P conversion formalism and show how different G2P conversion approaches can be combined in a multi-stream fashion to enhance the phone class conditional probabilities, and consequently generate pronunciation lexicons that yield better ASR systems [Razavi and Magimai.-Doss, 2017]. Furthermore, we show how G2P conversion and A2P conversion can be unified in a similar multi-stream framework by extending the problem of matching an acoustic signal to a word hypothesis as the case where the acoustic example of a word represents the acoustic signal, the graphemic representation of the word represents the word hypothesis, and the phones represent the latent symbols. Through experimental studies on two challenging corpora on English, we show that the lexicons developed using the multi-stream combination approach lead to better ASR systems compared to the ones developed using individual G2P or A2P conversion approaches.
- *Acoustic subword unit-based lexicon development*: As explained in Section 1.1, one of the main challenges in the existing G2P conversion approaches is to develop lexicons for under-resourced languages with no phonetic resources available. To address this challenge, we propose an approach for automatic derivation of acoustic subword units (ASWUs) and development of an ASWU-based pronunciation lexicon [Razavi and Magimai.-Doss, 2015, Razavi et al., 2015b]. In this approach, ASWU derivation is cast as an extension of the matching problem where given the acoustic signal represented by acoustic features, and the word hypothesis represented by graphemes, the objective is to find a latent symbol space that relates to both information. Given the discovered ASWUs and the acoustic data, the proposed G2P conversion formalism briefly explained earlier can be extended to generate pronunciations for both seen and unseen words. We validate the proposed approach on English as a well-resourced language and on Scottish Gaelic as a genuinely under-resourced language. Our studies show that the ASWU-based lexicons lead to better ASR systems than the alternative grapheme-based lexicons. Furthermore, the ASWUs are "phone-like" and transferable across domains.

In the literature, the evaluation of G2P conversion approaches is typically limited to studies at the pronunciation level. In this thesis, we go one step further, where we evaluate the generated pronunciations through the proposed approaches at both pronunciation level (if feasible), and application level, which is in our case ASR. In our studies, we consistently find that the evaluation at the pronunciation level is not fully indicative of the performance at the application level [Razavi et al., 2016, Razavi and Magimai.-Doss, 2017]. Therefore, determining the best pronunciation lexicon purely based on the pronunciation level evaluation can be

suboptimal from the application perspective.

1.3 Organization of the thesis

The remainder of this thesis is organized as follows:

- Chapter 2 provides the related background that can be helpful in understanding the context of study in this thesis. It first defines common terminologies used in this thesis and provides an overview of the main components in a standard HMM-based ASR system. It then presents state-of-the-art approaches proposed in the literature for phonetic pronunciation lexicon development. Finally, it describes the evaluation metrics as well as the databases used in the thesis.

- Chapter 3 focuses on the problem of matching an acoustic signal with a word hypothesis through a latent symbol set in the context of ASR, and explains the fundamental issues in ASR systems in that perspective. It then investigates the space of latent symbols in different ASR systems namely, HMM/GMM, hybrid HMM/ANN and KL-HMM, and shows that in the framework of KL-HMM, the latent symbol space is relatively small.

- Chapter 4 focuses on integrating the acoustic information in the G2P relationship learning process. It first presents a posterior-based formalism for G2P conversion, akin to the hybrid HMM/ANN framework for ASR. It then shows how phone posterior probabilities given graphemes can be estimated through acoustics by building on the findings in Chapter 3 to match a speech signal with a word hypothesis. Finally, it validates the acoustic G2P conversion approach by benchmarking it against state-of-the-art G2P conversion approaches at both pronunciation level and ASR level.

- Chapter 5 focuses on unifying pronunciation extraction approaches. It first presents a posterior-based multi-stream formulation for G2P conversion, which enables unifying various G2P conversion approaches providing estimates of the probability of phones given graphemes during pronunciation inference. It then shows how such a multi-stream formulation can be extended to unify G2P conversion and A2P conversion approaches. Finally, it illustrates the validity of the proposed formalism through experimental studies on two challenging tasks on English.

- Chapter 6 focuses on the problem of lexicon development for under-resourced languages. It first proposes an HMM-based formalism for automatic derivation of ASWUs given only the word-level transcribed speech data. It then shows how the acoustic G2P conversion approach developed in Chapter 4 can be exploited to generate pronunciations based on ASWUs. Finally, it validates the proposed approach through experimental studies and analysis on English and Scottish Gaelic.

- Finally, Chapter 7 concludes the thesis and presents possible avenues for future research.

2 Background

The focus of this thesis is on data-driven methods for development of phonetic pronunciation lexicons, which has applications in both ASR and TTS systems.[1] In this thesis, ASR is considered as the end-level application for evaluating the generated pronunciation lexicons. This chapter first defines the mathematical notations and the specific terms used in the thesis. It then overviews the basic components of an ASR system, followed by presenting state-of-the-art methods in the literature for phonetic pronunciation lexicon development. Finally, it describes the evaluation metrics and the datasets used in the thesis.

2.1 Notations and terminology

In this thesis, we use boldface symbols to denote vectors. Subscripts are used for vector or time indices, while superscripts are used for class indices. The vector elements are enclosed in brackets [], the sequence terms are enclosed in parentheses (), and the set elements are enclosed in braces {}.

The important terminologies used in the thesis are defined hereafter:

- **Grapheme**: A grapheme is the smallest unit of a writing system of a language [Coulman, 1996] (e.g., alphabetic letters).
- **Phoneme**: Phonemes are "the smallest contrastive linguistic units which may bring about a change of meaning" [Chomsky and Halle, 1968] in a specific language.
- **Phone**: Phones are units of the speech sounds which can be designed to cover the set of sounds in all languages [Gold and Morgan, 1999, Ch. 23].
- **SAMPA**: The Speech Assessment Methods Phonetic Alphabet (SAMPA) is a machine

[1] There are some differences in the lexicon requirement for ASR and TTS systems though. For example, the lexicons for the TTS systems typically have a single pronunciation per word. In addition, in these lexicons, the lexical stress and the syllable are required to be marked.

readable phonetic alphabet for a vast number of languages.[2]

As implied from the definitions, phones and phonemes are two different terminologies. However, in the ASR community they are typically interchangeably used. Throughout this thesis, for the sake of clarity we use the term phones as it is more typical in speech recognition. The phones are enclosed in slashes //. The graphemes are enclosed in brackets [].

2.2 Automatic speech recognition

In the statistical ASR approach, given a sequence of acoustic feature observations $X = (\mathbf{x}_1, \cdots, \mathbf{x}_t, \cdots, \mathbf{x}_T)$ representing the speech signal obtained through a process called *feature extraction*, the goal is to obtain the most likely word sequence $W^* = (w_1, \cdots, w_m, \cdots, w_M)$:

$$W^* = \arg\max_{W \in \mathscr{W}} P(W|X, \Theta), \tag{2.1}$$

where $\mathscr{W}$ denotes the set of all possible word sequences, W represents a word sequence and Θ denotes the set of parameters of the system. For simplicity, in the remainder of this chapter Θ is dropped.

As direct estimation of $P(W|X)$ is a non-trivial task, typically Bayes' rule is applied, leading to,

$$W^* = \arg\max_{W \in \mathscr{W}} \frac{p(X|W)P(W)}{p(X)}, \tag{2.2}$$

$$= \arg\max_{W \in \mathscr{W}} p(X|W)P(W). \tag{2.3}$$

Eqn. (2.3) is obtained as a result of the assumption that $p(X)$ does not affect the optimization. Therefore, finding the most likely word sequence amounts to estimation of acoustic likelihood $p(X|W)$ and the word sequence probability $P(W)$. We refer to them as acoustic likelihood estimation and language model estimation respectively.

In state-of-the-art ASR systems, HMMs are used for acoustic likelihood estimation. More precisely, words are modeled as a sequence of phones, based on the information provided in the phonetic lexicon; and phones are further modeled as a sequence of HMM states. Language model estimation involves modeling the syntactic constraints of a language, typically through n-grams . The decoder searches through all possible word sequence hypotheses to infer the most likely word sequence. Figure 2.1 depicts the main components of an ASR system. The following sections explain each of these components in more detail.

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

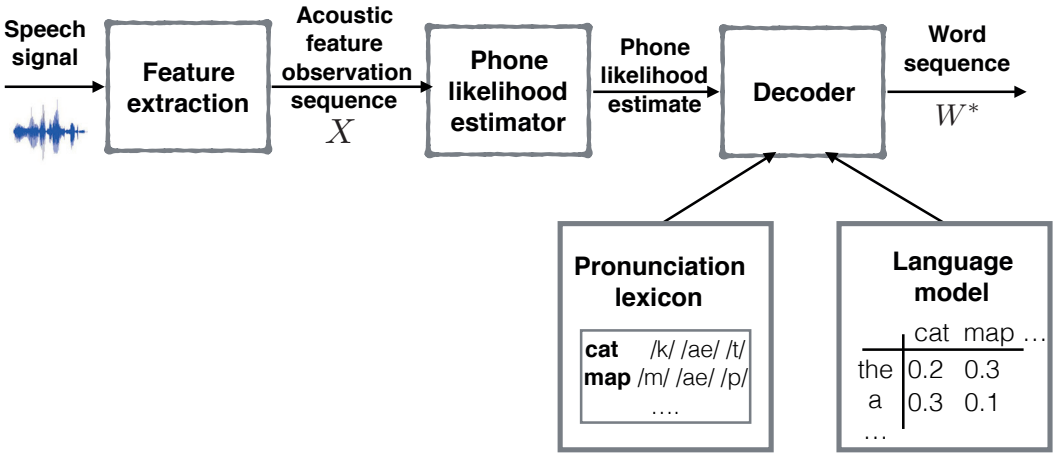


Figure 2.1 – The components of a general HMM-based ASR system.

2.2.1 Feature extraction

The goal of feature extraction is to obtain a compact representation of the speech signal that is relevant to sound unit identities and is robust to irrelevant variabilities such as speaker and environment characteristic. In the literature, these representations are typically computed every 10 ms over an analysis window of 25 ms. This is based on the assumption that speech signal is quasi-stationary in short time intervals. The two most commonly used representations are based on Mel-frequency cepstral coefficient (MFCC) [Davis and Mermelstein, 1980] and perceptual linear prediction (PLP) [Hermansky, 1990]. For a good description of MFCC and PLP features, the reader is referred to [Gold and Morgan, 1999, Ch. 22]. Briefly, the main steps involved in computation of MFCC and PLP features are as follows:

1. *Power spectrum estimation*: This is done by windowing the analysis region, computing fast Fourier transform (FFT) and its squared magnitude.

2. *Integration of power spectrum within critical band filter responses*: In order to approximate the response of human ear, a non-linear frequency scale is applied. For MFCCs, this is done using Mel scale, which is roughly linear below 1kHz and logarithmic above 1kHz. For PLPs, this is done using trapezoidal shaped filters at roughly 1-bark intervals.

3. *Spectrum pre-emphasis*: In order to account for the unequal sensitivity of human hearing in different frequencies, pre-emphasis is done. In the case of MFCCs, this is done before spectral analysis whereas for PLPs this is done through weighting of the critical band spectrum.

4. *Spectral amplitude compression*: In order to reduce the amplitude variations for spectral resonances, spectral amplitude is compressed. For MFCCs, this is done by applying log operator whereas for PLPs this is performed by applying cube root.

5. *Decorrelation and spectral smoothing*: For MFCCs, decorrelation is done using discrete

cosine transform (DCT) to obtain the cepstral coefficients, and spectral smoothing is achieved by cepstral truncation in which the first 12 or 13 coefficients ($c_0 - c_{12}$) are kept. For PLPs, spectral smoothing is achieved by using an autoregressive model.

The main difference between MFCC and PLP features therefore lies in spectral smoothing. It has been found that the use of an autoregressive model leads to better noise robustness [Openshaw et al., 1993] and speaker independence [Psutka et al., 2001] than cepstral truncation.

In addition to MFCCs or PLPs, in order to account for dynamic behavior of the speech signal, first order derivatives (Δ) and second order derivatives ($\Delta\Delta$) of static features computed over analysis frames are appended to the features [Furui, 1986]. This leads to a feature vector $\mathbf{x}_t$ (containing $c_0 - c_{12} + \Delta + \Delta\Delta$) with dimensionality of 39.

2.2.2 Pronunciation lexicon

As explained in Chapter 1, state-of-the-art ASR systems represent words in terms of subword units to resolve data sparsity issues and generalization toward unseen words. Using subword units brings two questions: (1) how to choose the subword units?, and (2) how to represent each word in terms of subword units?

Various types of subword units have been investigated in the literature to answer the first question [Livescu et al., 2012]. Two types of subword units commonly used in current ASR systems are phones and graphemes. Phones and graphemes in a language are related, however, depending on the language the relationship can be regular or irregular. In languages such as Finnish with shallow orthographies, the G2P correspondence is regular and one-to-one. In languages with deep orthographies, however, the correspondence between the graphemes and phones is not direct. More precisely, in languages such as English the G2P relationship is irregular, i.e., some prior knowledge about the word is required to accurately predict the relationship. In languages such as French on the other hand, the G2P relationship is regular, i.e., predictable given a set of linguistic rules, however, accurate prediction of the G2P relationship in these languages requires complex linguistic rules.

The main advantage of using graphemes as subword units is facilitating the development of a lexicon. More precisely, the graphemic representation of each word can be easily obtained from its orthography. This is particularly beneficial for under-resourced languages in which limited linguistic information is available. However, the success of grapheme-based ASR systems depends on the G2P relationship in the language [Kanthak and Ney, 2002, Rasipuram and Magimai.-Doss, 2015], which as explained above is not necessarily one-to-one. As a result, most of the state-of-the-art ASR systems are based on phones as subword units. In order to develop phone-based ASR systems, a phonetic lexicon is required in which each word is represented as a sequence of phones. The phonetic pronunciations are typically obtained from a hand-built lexicon. In order to augment the existing phonetic lexicons, usually G2P conversion approaches are used. We will overview the proposed methods for G2P conversion

in Section 2.3.

It is worth mentioning that in state-of-the-art ASR systems, each subword unit in the context is considered as a separate unit [Schwartz et al., 1985]. These units are referred to as context-dependent (CD) subword units. For example, the pronunciation of the word MAP = /m//ae//p/ is presented as /m/ /ae/ /p/ with context-independent (CI) subword unit representation and is presented as /m+ae/ /m-ae+p/ /ae-p/ with context-dependent subword unit representation. Context-dependent phone modeling was motivated from the coarticulation perspective as the same phone may be realized differently depending on the context [Livescu et al., 2012].

2.2.3 Acoustic likelihood estimation: Estimating $p(X|W)$

As explained in the previous section, standard ASR systems decompose words into a sequence of subword units, according to the representation provided in the pronunciation lexicon. As multiple pronunciations can exist for each word, the likelihood $p(X|W)$ is estimated as [Gales and Young, 2008],

$$p(X|W) = \sum_{\phi \in \Phi} p(X, \phi|W), \tag{2.4}$$

$$= \sum_{\phi \in \Phi} p(X|\phi, W) P(\phi|W), \tag{2.5}$$

$$= \sum_{\phi \in \Phi} p(X|\phi) P(\phi|W), \tag{2.6}$$

$$\approx \max_{\phi \in \Phi} p(X|\phi) P(\phi|W), \tag{2.7}$$

where Φ represents all valid pronunciation sequences for W, and ϕ is a particular pronunciation sequence. Eqn. (2.6) is obtained based on the assumption that given the pronunciation sequence ϕ, the acoustic observation sequence X is independent of the word sequence W. Eqn. (2.7) is obtained using Viterbi approximation, in which the summation over all possible pronunciations is replaced by maximization.[3] Therefore estimation of $p(X|W)$ amounts to estimation of $p(X|\phi)$ and $P(\phi|W)$.

Estimating $P(\phi|W)$

$P(\phi|W)$, the probability of a pronunciation sequence given the word sequence, is usually referred to as the pronunciation model, and is derived from the pronunciation lexicon. More precisely,

$$P(\phi|W) = \prod_{m=1}^{M} P(F^{(w_m)}|w_m), \tag{2.8}$$

[3] Applying Viterbi optimization is an algorithmic choice though.

where $F^{(w_m)}$ represents a pronunciation for the word w_m. Typically in speech recognition systems, the pronunciation lexicons are unweighted, which would be equivalent to setting $P(F^{(w_m)}|w_m)$ to one for all words. However, approaches exist that parameterize $P(\phi|W)$ [McGraw et al., 2013].

In this thesis, we used unweighted pronunciation lexicons, i.e., $P(F^{(w_m)}|w_m) = 1 \ \ \forall m$.

Estimating $p(X|\phi)$

A common way to model $p(X|\phi)$ in the literature is to use HMMs [Rabiner, 1989]. An HMM consists of two stochastic processes. One stochastic process generates the state sequence $Q = (q_1, \cdots, q_t, \cdots, q_T)$. The other stochastic process generates a sequence of observations according to the probability functions associated with each state. As for any observation sequence, the generating state sequence is hidden, it is referred to as hidden Markov model.

$p(X|\phi)$ in an HMM-based framework can be estimated by summing over all possible state sequences Q, i.e.,

$$p(X|\phi) = \sum_{Q \in \mathcal{Q}} p(X, Q|\phi), \tag{2.9}$$

$$= \sum_{Q \in \mathcal{Q}} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t = l^i) P(q_t = l^i|q_{t-1} = l^j), \tag{2.10}$$

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

where each HMM state q_t represents a lexical subword unit l^i in the pronunciation sequence ϕ, i.e., $q_t \in \mathcal{L} = \{l^i, \cdots, l^I\}$ with I being the number of lexical subword units[4]; $Q = (q_1, \cdots, q_t, \cdots, q_T)$ denotes a sequence of HMM states corresponding to the pronunciation sequence ϕ; and $\mathcal{Q}$ denotes the set of all possible HMM state sequences for the pronunciation sequence ϕ. Eqn. (2.10) is obtained by making two assumptions,

1. independent and identically distributed (i.i.d.) assumption, which states that the observations are conditionally independent of all other observations given the current state that generated them; and

2. first order Markov assumption, which states that the states are conditionally independent of all other states given the previous state.

Eqn. (2.11) is obtained by applying the Viterbi approximation, i.e., the sum of all possible state sequences is replaced with the most probable state sequence. $p(\mathbf{x}_t|q_t = l^i)$ is typically referred to as the *local emission score*, and $a_{ij} = P(q_t = l^i|q_{t-1} = l^j)$ is referred to as the *transition score*.

[4] In fact I is the number of lexical subword unit states.

As shown in [Rasipuram and Magimai.-Doss, 2015], standard HMM-based ASR systems implicitly model $p(\mathbf{x}_t|q_t = l^i)$ through a latent symbol set $\mathscr{A} = \{a^1, \cdots, a^d, \cdots, A^D\}$ referred to as *acoustic unit* set, i.e.,

$$p(\mathbf{x}_t|q_t = l^i) = \sum_{d=1}^{D} p(\mathbf{x}_t, a^d|q_t = l^i), \tag{2.12}$$

$$= \sum_{d=1}^{D} p(\mathbf{x}_t|a^d, q_t = l^i) \cdot P(a^d|q_t = l^i), \tag{2.13}$$

$$= \sum_{d=1}^{D} p(\mathbf{x}_t|a^d) \cdot P(a^d|q_t = l^i) (\text{assuming } \mathbf{x}_t \perp\!\!\!\perp q_t|a^d). \tag{2.14}$$

In the CI subword unit-based ASR systems, the acoustic units are directly defined from the pronunciation lexicon (i.e., they are obtained in a knowledge-driven manner). For the case of CD subword modeling, assuming there are U context-independent subword units in the lexicon and each subword unit is modeled with its c_l preceding and c_r following context, $U^{c_l+c_r+1}$ CD subword units must be modeled. Due to data sparsity issues, a parameter sharing mechanism is required to enable efficient modeling of the CD subword units. This is done using a decision tree clustering approach, in which the states of CD models are tied based on a maximum likelihood criteria [Young et al., 1994]. The number of obtained tied states depends on the hyper parameters such as minimum cluster occupancy and minimum increase in the log-likelihood threshold. The acoustic unit set $\mathscr{A}$ in the case of CD subword modeling is therefore derived by clustering the HMM states using decision tree methods in a data-driven manner.

In standard HMM-based ASR systems, the relation between the acoustic units and lexical subword units $P(a^d|q_t = l^i,)$ is a one-to-one deterministic map, i.e.,

$$p(\mathbf{x}_t|q_t = l^i) = p(\mathbf{x}_t|a^d), \text{ given } l^i \mapsto a^d,\ d \in \{1, \ldots, D\}. \tag{2.15}$$

In the case of CI subword modeling, the deterministic map is obtained through knowledge, while in the case of CD subword modeling the mapping is learned during state clustering and tying.

In the literature, two main approaches for estimating $p(\mathbf{x}_t|a^d)$ are GMMs and ANNs. The approach using GMMs is referred to HMM/GMM approach [Rabiner, 1989], and the approach using ANNs is referred to as hybrid HMM/ANN approach [Morgan and Bourlard, 1995]. In the remainder of this section we explain each of these approaches.

HMM/GMM approach

In the HMM/GMM approach, a GMM is used to model an acoustic unit [Rabiner, 1989], i.e.,

$$p(\mathbf{x}_t|a^d) = \sum_{n=1}^{N} c_n^d \mathcal{N}(\mathbf{x}_t; \mu_n^d, \Sigma_n^d), \tag{2.16}$$

where N denotes the number of Gaussian components per mixture for each acoustic unit; c_n^d, μ_n^d and Σ_n^i denote the mixture weight, mean and covariance for the n^{th} Gaussian modeling a^d respectively. The parameters of the HMM/GMM system to be estimated are therefore the transition probabilities, means, covariances and the mixture weights.

Hybrid HMM/ANN approach

In the hybrid HMM/ANN approach, the acoustic units are modeled using an ANN [Morgan and Bourlard, 1995]. One of the most commonly used neural networks are multi-layer perceptrons (MLPs). An MLP consists of an input layer, one or more hidden layers and an output layer, with each layer consisting of one or several nodes. Each layer is fully connected to the next layer. Except for the input nodes, each node computes a non-linear function of the weighted sum of its inputs. In order to learn the parameters of the MLP (i.e., weights and biases), the error backpropagation algorithm is used [Rumelhart et al., 1988]. The backpropagation algorithm requires a known label for each input in order to calculate a certain loss function gradient. The loss function used for MLP training is typically minimum squared error or cross-entropy.

In the hybrid HMM/ANN approach, the input nodes of the MLP are typically cepstral features with c preceding and c following frame context, i.e., $\mathbf{X}_t = [\mathbf{x}_{t-c}^{\mathrm{T}} \cdots \mathbf{x}_t^{\mathrm{T}} \cdots \mathbf{x}_{t+c}^{\mathrm{T}}]^{\mathrm{T}}$. For the hidden nodes, the non-linear function is typically a sigmoid function, while for the output nodes a softmax nonlinear function is usually used. The output nodes of the MLP are acoustic units, i.e., either CI subword units [Morgan and Bourlard, 1990] or clustered CD subword units [Dahl et al., 2012]. It has been shown that the MLP estimates the posterior probability of the output classes given the input [Bourlard and Morgan, 1994], i.e., the MLP estimates $\mathbf{z}_t = [z_{t,1} \cdots z_{t,d} \cdots z_{t,D}]^{\mathrm{T}}$ with $z_{t,d} = P(a^d|\mathbf{x}_t)$ where $P(a^d|\mathbf{x}_t)$ is the posterior probability of acoustic unit a^d given the acoustic observation vector $\mathbf{x}_t$. The posterior probability of the output classes given by the MLP is converted to a scaled-likelihood of an HMM state and is used as local emission score, i.e.,

$$p_{sl}(\mathbf{x}_t|a^d) = \frac{P(\mathbf{x}_t|a^d)}{p(\mathbf{x}_t)} = \frac{P(a^d|\mathbf{x}_t)}{P(a^d)}. \tag{2.17}$$

The state transition probabilities in the hybrid HMM/ANN framework are usually fixed to be 0.5 [Morgan and Bourlard, 1995].

Instead of fully connected feed forward networks, other architectures such as convolutional neural networks (CNNs) [Waibel et al., 1989, Sainath et al., 2013] and recurrent neural net-

works (RNNs) [Robinson et al., 1994, Vinyals et al., 2012] have also been studied for speech recognition. More recently, composite architectures have been proposed where the features and the local classifiers are jointly learned from the speech signal [Palaz et al., 2013, Tüske et al., 2014]

In both HMM/GMM and hybrid HMM/ANN approaches, the HMM parameters are learned using the EM algorithm with a cost function based on likelihood. Two common EM-based approaches in the HMM framework used for this purpose are Baum-Welch or forward-backward training [Rabiner, 1989, Hennebert et al., 1997] and embedded Viterbi training [Juang and Rabiner, 1990, Morgan and Bourlard, 1995]. In the embedded Viterbi training, which is more commonly used particularly in hybrid HMM/ANN framework, in the E-step, given the current model parameters, the optimal HMM state sequence is obtained. Then in the M-step, given the segmentation, the new set of parameters optimizing the cost function is trained.[5] In the forward-backward training, in the E-step, instead of obtaining a "hard" alignment, a *soft* alignment between the HMM states and the frames is estimated.

In this thesis, we trained MLPs with cross-entropy error criteria using the Quicknet software [Johnson et al., 2004]. The input to the MLPs was 39-dimensional PLP cepstral features with four preceding and four following frame context. The output to the MLP was either CI subword units or clustered CD subword units, and the output labels were obtained from the HMM/GMM system. In order to avoid overfitting, the early stopping method [Morgan and Bourlard, 1989] is employed in which the performance on the cross-validation set was used to stop MLP training.

2.2.4 Language model estimation: Estimating $P(W)$

The language model $P(W)$ estimates the prior probability of a word sequence W. Using the chain rule of probability $P(W)$ can be factorized as follows:

$$P(W) = \prod_{m=1}^{M} P(w_m | w_{m-1}, \cdots, w_1). \tag{2.18}$$

Estimation of $P(W)$ according to Eqn. (2.18) is not trivial, as the number of previous words is variable. Typically in the literature, $P(W)$ is estimated in the form of an *n-gram* language model based on the assumption that given the previous $n-1$ words, the probability of a word is independent of the rest of the history. Therefore,

$$P(W) = \prod_{m=1}^{M} P(w_m | w_{m-1}, \cdots, w_{m-n+1}). \tag{2.19}$$

The n-gram probabilities are estimated from a text corpora using maximum likelihood criteria, leading to estimates based on n-gram frequency counts. The major issue in such estimation

[5] For the hybrid HMM/ANN approach, alternately the segmentation can be obtained from a trained HMM/GMM system.

is data sparsity. This is usually resolved by using a smoothing method such as discounting, back-off or combination of these approaches [Katz, 1987, Kneser and Ney, 1995]. Recently with the advances in neural networks, recurrent neural networks have also been shown to lead to promising results for language modeling [Mikolov et al., 2010].

2.2.5 Decoding

Given the trained acoustic likelihood estimator and the language model, the most probable word sequence can be obtained. More precisely, this is obtained by finding the most probable state sequence Q representing W^* by incorporating lexical and syntactic knowledge:

$$W^* = \underset{W \in \mathcal{W}}{\arg\max}\, p(X|W)P(W), \tag{2.20}$$

$$\approx \underset{Q \in \mathcal{Q}}{\arg\max} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t = l^i)P(q_t = l^i|q_{t-1} = l^j), \tag{2.21}$$

$$\approx \underset{Q \in \mathcal{Q}}{\arg\max} \sum_{t=1}^{T} \left[\log p(\mathbf{x}_t|q_t = l^i) + \log P(q_t = l^i|q_{t-1} = l^j)\right], \tag{2.22}$$

where the local emission score $p(\mathbf{x}_t|q_t = l^i)$ is estimated either using GMMs or ANNs as explained in Section 2.2.3, $P(q_t = l^i|q_{t-1} = l^j)$ is obtained from the language model if l^j corresponds to the subword unit in a word and l^i corresponds to the subword unit in the next word, and otherwise is the HMM state transition probability. As the acoustic likelihood and language model probabilities have different dynamic ranges, in practice the language model probabilities are scaled before the combination with the acoustic likelihood scores. Furthermore, the word transitions are penalized in order to avoid insertion of many short words. Eqn. (2.22) is obtained as a result of applying log transformation to Eqn. (2.21). The most probable state sequence in Eqn. (2.22) can be obtained using the Viterbi algorithm [Forney, 1973]. However, a complete search may not be computationally feasible, as different factors such as the n-gram language model and the cross-word CD modeling can expand the search space. Therefore, in practice the search space is pruned to speed-up the search. A commonly used approach is beam search, in which only the paths whose likelihoods lie within a fixed beam width of the most likely path are kept for expansion [Greer et al., 1982].

2.3 Pronunciation lexicon development methods

One of the main components in both TTS and ASR systems is the phonetic pronunciation lexicon. The phonetic lexicon is typically prepared by linguistic experts. Pronunciation preparation is a tedious and time consuming task, as linguists must take into account different factors such as minimizing word-level confusions and ensuring pronunciation consistency across the lexicon. Furthermore, the hand-crafted lexicons must be constantly augmented with evolution of languages and emergence of new words. As a result, given an initial phonetic lexicon, ASR and TTS systems use G2P conversion methods to generate pronunciations for

the words not covered in the lexicon. In this section, we first elucidate two classes of G2P conversion methods, namely knowledge-based and data-driven approaches, which have been explored in the literature.

2.3.1 Knowledge-based approaches

Knowledge-based G2P conversion approaches exploit rules derived by humans or from linguistic studies to convert the sequence of graphemes in a word to a sequence of phones [Elovitz et al., 1976]. Commonly, the form of the rules is $A[B]C \mapsto D$, which states that the grapheme B with the left context A and the right context C maps to the phone or phone sequence D. Alternately, rule-based G2P conversion approaches are typically formulated in the framework of finite state automata [Kaplan and Kay, 1994]. While knowledge-based approaches exploiting rules can provide a complete coverage, they have two main drawbacks: (1) designing rules requires linguistic knowledge and expertise, which may not be always available, and (2) due to existence of irregularities in natural languages, exception rules or exception lists are required to be designed. Furthermore, the rules should be cross-checked to ensure that they are applicable to all the entries. Therefore, development of lexicons using knowledge-based approaches is a tedious task.

2.3.2 Data-driven G2P conversion approaches

In order to reduce the amount of human effort and linguistic knowledge, data-driven approaches are usually employed. Data-driven approaches for G2P conversion predict the pronunciation of an unseen word based on the examples in the training data (i.e., the seed lexicon). Typically the G2P conversion process in data-driven approaches can be viewed as a three-step process. The first step is the alignment of training data constituting sequences of graphemes and their corresponding sequences of phones [Damper et al., 2005, Jiampojamarn et al., 2007]. In the second step, a learning method is employed to capture the G2P relationship observed in the source lexicon. Finally as the third step, an inference algorithm is used to infer the best pronunciation.

The alignment step can be viewed as a common process in most of the G2P conversion approaches.[6] Therefore, what distinguishes different G2P conversion approaches from each other is the learning and inference methods utilized. Among various G2P conversion approaches proposed based on different techniques [Sejnowski and Rosenberg, 1987, Dedina and Nusbaum, 1991, Black et al., 1998, Pagel et al., 1998, Taylor, 2005, Bisani and Ney, 2008, Davel and Barnard, 2008, Wang and King, 2011], local classification-based and probabilistic sequence modeling-based approaches have gained wide attention, and are explained below.

[6] In some approaches, the alignment is done as a pre-processing step whereas in others the alignments are obtained while learning the G2P relationship.

Local classification-based approaches

In the local classification-based approaches, given the alignments, a decision tree [Black et al., 1998, Pagel et al., 1998] or a neural network [Sejnowski and Rosenberg, 1987] can be trained to learn the G2P relationship from the training data.[7] For the inference part, the sequence of input graphemes is processed sequentially in which for each grapheme, the corresponding phone (or phone sequence) is locally generated. Therefore, these methods are referred to as local classification-based approaches.

Probabilistic sequence modeling-based approaches

In probabilistic sequence modeling-based approaches, the G2P conversion task can be expressed formally as,

$$F^* = \arg\max_F P(F|G), \tag{2.23}$$

$$= \arg\max_F P(F,G), \tag{2.24}$$

where given a sequence of graphemes G, the goal is to find a sequence of phones F^* that maximizes the posterior probability $P(F|G)$. Eqn. (2.23) can also be expressed as finding a sequence of phones F^* maximizing the joint probability $P(F,G)$ using the Bayes' rule (Eqn. (2.24)). Various G2P conversion approaches based on above expressions are as follows:

1. *HMM-based approach*: In [Taylor, 2005], the G2P conversion problem is formulated in the standard HMM way by applying i.i.d. and first order Markov model assumptions as,

$$S^* = \arg\max_S P(S,G), \tag{2.25}$$

$$= \arg\max_S P(G|S)P(S), \tag{2.26}$$

$$= \arg\max_S \prod_n P(g_n|s_n)P(s_n|s_{n-1}), \tag{2.27}$$

   where $S = (s_1,\ldots,s_n,\ldots,s_N)$ represents the hidden sequence of phones and $G = (g_1,\ldots,g_n,\ldots,g_N)$ denotes the sequence of grapheme observations. In this framework, each HMM represents a phone that emits (up to four) grapheme symbols. As opposed to local classification approaches in which the alignments are obtained as a pre-processing step, in this framework the alignments can be derived during the Baum-Welch training. For the inference, the most probable sequence of phones that generated the input grapheme sequence is obtained using the Viterbi algorithm.

2. *Joint multigram approach*: In joint multigram or joint n-gram approaches, the joint probability $P(F,G)$ of a sequence of graphemes G and a sequence of phones F in Eqn. (2.24) is obtained based on the concept of graphones [Deligne et al., 1995]. A graphone is a pair

[7] A decision tree and a neural network are two (of many) examples of local classifiers.

of a sequence of graphemes and a sequence of phones. Figure 2.2 shows a sequence of graphones for the word *phone* along with its pronunciation.

ph	***o***	***n***	***e***
f	***ow***	***n***	***–***

Figure 2.2 – A possible sequence of graphones for the word *phone* and its associated pronunciation.

The joint probability $P(F,G)$ is obtained by summing over matching alignments which are derived from sequences of graphones Q in the space of all possible sequences of graphones for the (F,G) pair, i.e., S(F,G):

$$P(F,G) = \sum_{Q \in S(F,G)} p(Q). \tag{2.28}$$

The probability distribution over all matching alignments can be modeled using an n-gram approximation. In [Bisani and Ney, 2008], the parameters of the n-gram model are learned by maximizing the log-likelihood of the data using the expectation-maximization (EM) algorithm. There are other variants such as [Chen, 2003], in which the parameters of the maximum-entropy n-gram model are learned using the Viterbi EM algorithm. For the inference, the best sequence of phones can be derived by using the Viterbi algorithm. In [Novak et al., 2012], the best sequence of phones is obtained in the weighted finite state transducer (WFST) framework.

3. *CRF-based approach*: In CRF-based approaches, the conditional probability $P(F|G)$ in Eqn. (2.23) is modeled using a log-linear representation [Wang and King, 2011, Lehnen et al., 2011]. The CRF model is a discriminative model that can perform global inference. Therefore, it can exploit the advantages of both decision tree-based methods (which are discriminative) and joint multigram methods (which perform global inference). However, it can be computationally more expensive than the aforementioned approaches.

   The parameters of the log-linear CRF model are learned by maximizing the conditional log-likelihood. During decoding, the best phone sequence is inferred using the Viterbi algorithm. In [Hahn et al., 2013], hidden conditional random fields (HCRFs) are used for the G2P conversion task in which the alignment between the grapheme sequence and phone sequence is modeled via a hidden variable.

Recently, long short-term memory (LSTM)-based neural network architectures, which are a class of RNNs suitable for sequence modeling, have also been proposed for G2P conversion [Rao et al., 2015, Yao and Zweig, 2015]. In [Rao et al., 2015] it was shown that the pronunciations generated through the LSTM-based neural networks can provide complementary information to the pronunciations generated through the joint multigram approach.

2.3.3 Pronunciation extraction using acoustic data

As discussed earlier, conventional data-driven G2P conversion approaches learn the G2P relationship on the seed lexicon as the training data. As a result, the pronunciations obtained from such approaches reflect the information found in the seed lexicon, and may not capture the natural phonological variation. To overcome this limitation, in the context of pronunciation variation modeling, spoken examples of words are used to obtain pronunciation variants. Most often, automatic phone transcriptions of spoken examples obtained from a phone recognizer are used to determine possible alternative pronunciations of words [Mokbel and Jouvet, 1999]. For example, in the first stage, speech data transcribed at word level is passed through a phone recognizer to obtain phone transcriptions of words. The phone recognizers can impose phonotactic constraints [Mokbel and Jouvet, 1999, Magimai.-Doss and Bourlard, 2005] or exploit phone bigrams or trigrams [Fosler-Lussier, 2000]. Possible alternate phone sequences for words are then obtained by finding the best alignment between the output of the phone recognizer and pronunciations provided by the seed lexicon [Fosler-Lussier, 2000].

An issue with such techniques is that they often over-generate variants because of multiple acoustic samples for each word. Furthermore, this also increases the chance of confusion among words in the dictionary. Therefore, it is important to prune the pronunciation variants to produce a lexicon that results in an optimal recognition performance. Possible pruning options that have been explored are based on maximum number of pronunciations per word, removing pronunciation variants with a probability less than a threshold given the word [Riley, 1991]. Figure 2.3 illustrates a typical pronunciation variant extraction process.

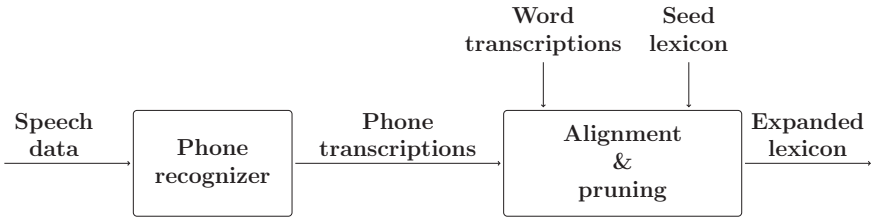


Figure 2.3 – Pronunciation lexicon expansion with possible pronunciation variants for words obtained using speech samples.

The pronunciations obtained from a phonemic decoder can be noisy [Fosler-Lussier, 2000]. Therefore, rather than obtaining variants from a phonemic decoder, recently there has been an interest to prune the pronunciation variants obtained through a G2P converter using spoken word examples. In [McGraw et al., 2013], a pronunciation mixture model approach was used to weigh the pronunciation variants of words obtained from a graphone-based G2P conversion approach, based on acoustic evidence using the EM algorithm. Lu et al. [2013] further build on the pronunciation mixture model approach and propose an approach to expand the expert phonetic lexicon using a trained G2P converter and acoustic examples. More precisely in this approach, given an initial phonetic lexicon, a G2P converter is trained to

generate pronunciation variants for new words. The pronunciation variants are then weighted based on acoustic evidence using the WFST-based EM algorithm. Given the new augmented lexicon, the acoustic likelihood estimator is then updated, and the process is iterated until convergence. Both of these G2P conversion approaches still require an initial seed lexicon. The acoustic samples are used only to weigh or select the alternate pronunciations provided by a G2P converter.

In addition to the aforementioned approaches, in [Xiao et al., 2007], two approaches, one based on maximum likelihood training and the other based on discriminative training, were presented to adapt the parameters of the graphone-based G2P converter using spoken examples for a name recognition task.

2.4 Evaluation

The performance of G2P conversion approaches is commonly evaluated at the pronunciation level using metrics such as phone recognition rate. However, such metrics may not be indicative of the performance of the system in real applications, in our case ASR. Therefore, it is important to evaluate the G2P conversion approaches at the application level as well. This section first explains the metrics used for evaluation at the pronunciation level and ASR level. It then explains how the difference in the performance of G2P conversion approaches (at pronunciation or ASR level) can be evaluated through statistical significance test.

2.4.1 Pronunciation level evaluation

To evaluate the performance of G2P conversion approaches commonly phone recognition rate (PRR) is used. PRR is obtained from the Levenshtein distance [Levenshtein, 1966] between the generated phonetic transcription of the word and its reference phonetic transcription. More precisely, PRR is obtained by finding the optimal alignment between the generated phone sequence and the reference phone sequence, and computing the number of phone substitutions (S), deletions (D) and insertions (I),

$$PRR = \frac{N-(S+D+I)}{N} \times 100, \tag{2.29}$$

where N denotes the number of phones in the reference.

The generated pronunciations can also be evaluated at the word level, by computing the proportion of words for which the generated phonetic transcription is the same as the reference phonetic transcription. We refer to it as word-level pronunciation accuracy (WPA).

2.4.2 ASR level evaluation

The performance of the ASR systems in this thesis is evaluated in terms of word recognition rate (WRR). Similar to PRR, WRR is obtained from the Levenshtein distance between the recognized and reference word sequences:

$$WRR = \frac{N-(S+D+I)}{N} \times 100 \tag{2.30}$$

where N denotes the number of words in the reference, and S, D, and I denote the number of word substitutions, deletions, insertions respectively.

2.4.3 Statistical significance test

In order to compare the performance of G2P conversion approaches, it is important to know whether the difference between the obtained systems (either at pronunciation level or ASR level) is statistically significant or not. In this thesis, we employed the bootstrap estimation method proposed in [Bisani and Ney, 2004]. The main idea in the bootstrap estimation method is to generate *bootstrap samples* through random sampling from the data set with replacement. When comparing two systems, it is important that difference in the number of errors in the two systems is calculated on identical bootstrap samples. Throughout this thesis we applied the statistical significant test presented in [Bisani and Ney, 2004] with the confidence level of 95%.

2.5 Databases

This section describes different databases used in the thesis.

2.5.1 MediaParl

MediaParl is a bilingual corpus containing recordings of Swiss parliamentary debates from Valais region in Swiss German and Swiss French. Valais is a state in Switzerland consisting of both French and German speakers with a variety of accents. In this thesis, we used both German part and French part of the corpus. The database is partitioned into training, development and test set according to the structure provided in [Imseng et al., 2012b]. Table 2.1 provides the overview of the MediaParl corpus. All the speakers in the training and development set are native speakers. In the test set, four speakers are German native speakers and for three speakers, French is the native language.

For the German part of MediaParl corpus, the preparation of the dictionary was started with the Phonolex pronunciation lexicon [Imseng et al., 2012b][8], and afterward the generated pronunciations were hand-corrected. For the words not found in the dictionary, a WFST-

[8] http://www.phonetik.uni-muenchen.de/forschung/Bas/BasPHONOLEXeng.html

Table 2.1 – Overview of the MediaParl corpus in terms of number of utterances, hours of speech data, speakers and words present in the train, development and test sets. For the test set, the amount of native and non-native data is shown as well.

(a) German Part of MediaParl

Number of	Train	Development	Test (native, non-native)
Utterances	5955	879	1692 (1605, 87)
Hours	14	2	4.5 (4.3,0.2)
Speakers	73	8	7 (4, 3)
Words	13485	3675	6148

(b) French Part of MediaParl

Number of	Train	Development	Test (native, non-native)
Utterances	5471	646	925 (474, 451)
Hours	16.1	2.2	3.2 (1.6, 1.6)
Speakers	110	8	7 (3, 4)
Words	10555	3376	4246

driven G2P conversion system[9] was used to generate the associated pronunciation. The manual dictionary of the German MediaParl corpus is in SAMPA format with a phone set of size 57 (including the phone sil) and contains all the words in the train, development and test sets. The vocabulary size is 16755. The training set consists of 13485 words. The test set contains 6148 words of which 2343 words are not seen during training.

For the French part of MediaParl corpus, the preparation of the dictionary was started with the BDLex pronunciation lexicon [Imseng et al., 2012b][10]. Similar to the German part of the corpus, for the words that were not found in the BDLex dictionary, a WFST-driven G2P system was employed to generate single-best pronunciations and the generated pronunciations were then hand-corrected. The manual dictionary of the French MediaParl corpus is in SAMPA format with a phone set of size 38 (including the phone *sil*) and consists of all the words in the train, development and test sets. The vocabulary size is 12362. The training set consists of 10555 words and 10709 pronunciations. The test set contains 4246 words of which 915 words are not seen during training. The unseen words did not occur frequently in the test set (the most frequent unseen word occurred only 7 times). The average number of pronunciations per word was 1.01, which implies that the pronunciation variants are provided only for a few words in the dictionary. It is also worth mentioning that during the database preparation by Imseng et al. [2012b], liaison handling was not considered.

For the language model, a bigram model was trained on transcriptions of the training set for each language as well as EuroParl corpus (which consists of about 50 million words for each

[9] http://code.google.com/p/phonetisaurus/
[10] http://www.irit.fr/~Martine.deCalmes/IHMPT/ress_ling.v1/rbdlex_en.php

language).

2.5.2 PhoneBook

PhoneBook is a phonetically-rich isolated-word telephone-speech corpus [Pitrelli et al., 1995]. In [Dupont et al., 1997], the corpus was partitioned into small size (75 words) and medium size (602 words) vocabulary tasks. In this thesis, we use the medium size vocabulary task with 602 unique words according to the setup provided in [Dupont et al., 1997]. The overview of the PhoneBook corpus in that setup is given in Table 2.2.

Table 2.2 – Overview of the PhoneBook corpus in terms of number of utterances, hours of speech data, speakers and words present in the train, development and test sets.

Number of	Train	Development	Test
Utterances	19421	7290	6598
Hours	7.7	2.9	2.6
Speakers	243	106	96
Words	1580	603	602

The training set consists of 26,711 utterances (obtained by merging the small training set and development set as in [Dupont et al., 1997]), and test set consists of 6598 speech utterances. The test vocabulary consists of words and speakers that are unseen during training. PhoneBook pronunciation lexicon is manually transcribed using 42 phones (including the phone *sil*). The manual lexicon contains only a single pronunciation per word.

2.5.3 NameDat

The NameDat corpus [Adde and Svendsen, 2010] is a database containing English proper names spoken by native Norwegians. The English proper names appear within a Norwegian sentence. The speakers were asked to pronounce the proper names in a way they would actually do in an everyday speech. Therefore each proper name can be pronounced differently depending on the speaker. The database contains 669 words.[11] Due to the limited size of the corpus, a three-fold training and testing strategy similar to the approach in [Adde and Svendsen, 2011] was applied where the dataset was divided into training and test set three times such that there is no overlap between the speakers in the three test sets. In our experiments, we randomly selected 10% of the training data and used it as the development set. Table 2.3 provides an overview of the dataset. For the NameDat corpus, no canonical phonetic pronunciation lexicon is available. However, auditory verified phonetic transcription for each utterance containing the proper name has been provided. We extracted the auditory verified pronunciation for each proper name in the utterance, and created an auditory verified pronunciation lexicon on each training set. The average number of pronunciations per word

[11] Note that in [Adde and Svendsen, 2011], the number of words are 619, as some of the words have been removed.

in the auditory verified lexicons was 2.5, 2.7 and 2.7 on set-1, set-2 and set-3 respectively.

Table 2.3 – Overview of the NameDat corpus in terms of number of utterances, minutes of speech data and speakers in train, development and test sets.

Number of	Train			Development			Test		
	set-1	set-2	set-3	set-1	set-2	set-3	set-1	set-2	set-3
Utterances	2362	2534	2564	262	281	284	1521	1329	1295
Minutes	51	55	55	6	6	6	33	28	28
Speakers	12	13	13	12	13	13	7	6	6

2.5.4 WSJ0

The WSJ corpus has been originally designed for large vocabulary speech recognition and natural language processing, and it contains wide range of vocabulary size [Paul and Baker, 1992]. The WSJ corpus [Woodland et al., 1994] has two parts - WSJ0 with 14 hours of speech and WSJ1 with 66 hours of speech. In this thesis, we use the WSJ0 corpus for training, which contains 7106 utterances (about 14 hours of speech) and 83 speakers. We report recognition studies on Nov92 test set, which contains 330 utterances from 8 speakers unseen during training. The training set contains 10k unique words. The recognition vocabulary size is 5k words. The language model consists of a bigram model. The grapheme lexicon is obtained from the orthography of the words and contains 27 subword units including silence. The phone lexicon was based on UNISYN dictionary, and contains 46 phones (including the phone *sil*).

2.5.5 DARPA resource management

The DARPA Resource Management (RM) task is a 1000 word continuous speech recognition task based on naval queries [Price et al., 1988]. The training set consists of 3990 utterances spoken by 109 speakers amounting to approximately 3.8 hours speech data. The test set, formed by combining Feb89, Oct89, Feb91 and Sep92 test sets, contains 1200 utterances amounting to 1.1 hours of speech data. The word-pair grammar supplied with the RM corpus is used as the language model for decoding. The grapheme lexicon is obtained from the orthography of the words. In addition to the English characters, silence, symbol hyphen and symbol single quotation mark are considered as separate graphemes. Therefore, the lexicon contains 29 subword units including silence. The phone lexicon is based on UNISYN dictionary, and contains 42 phones (including the phone *sil*).

2.5.6 Scottish Gaelic

The Scottish Gaelic corpus was collected by the University of Edinburgh in 2010 and contains recordings from broadcast news and discussion programs.[12] The database is partitioned into training, development and test sets according to the structure provided in [Rasipuram et al., 2013a]. The overview of the Scottish Gaelic corpus is given in Table 2.4.

Table 2.4 – Overview of the Scottish Gaelic corpus in terms of number of utterances, hours of speech data and speakers in the train, development and test sets.

Number of	Train	Development	Test
Utterances	2389	1112	1317
Hours	3	1	1
Speakers	22	12	12

The database does not provide any phonetic lexicon. The graphemic lexicon can be simply obtained from the orthography of the words. As the corpus also contains borrowed English words, the graphemes J, K, Q, V, W, X, Y and Z are also present in the lexicon. Therefore the graphemic lexicon consists of 32 graphemes including silence as shown in Table 2.5.

As the corpus does not provide a language model, a bigram language model trained on the sentences from the test set, as done in [Rasipuram et al., 2013a] is used.

Table 2.5 – Graphemes used in the Scottish Gaelic corpus.

Vowels	A, E, I, O, U, À, È, Ì, Ò, Ù
Consonants	B, C, D, F, G , H , L, M, N, P, R, S, T
English Graphemes	J, K, Q, V, W, X, Y, Z

2.6 Summary

In this chapter, we briefly explained the main components of an ASR system: feature extraction, pronunciation lexicon, acoustic likelihood estimator, language model and decoder. We then described the proposed methods in the literature for pronunciation lexicon development, which is the focus of this thesis. Finally, we described the databases used in the thesis.

[12] http://forum.idea.ed.ac.uk/tag/scots-gaelic

3 Matching a speech signal with a word hypothesis through latent symbols

Pronunciation lexicon development, as discussed in the earlier chapters, is a semi-automatic process. This process involves developing a seed lexicon by linguistic experts who infer and refine a sequence of phones given the acoustic knowledge and the linguistic knowledge. Given the seed lexicon, the pronunciations for the new words are then generated through G2P conversion approaches. In this thesis, our interest lies in developing a framework that can integrate and exploit the abundantly available acoustic information for pronunciation lexicon development, such that it not only enables modeling phonological variations, but can also handle lack of linguistic expertise in the target language. For that purpose, we first focus on the problem of matching an acoustic signal with a word hypothesis in a data-driven fashion, given prior linguistic knowledge, as this matching process is one of the fundamental steps done by humans (linguistic experts) to obtain a phonetic transcription of the word.

Toward that, in this chapter we re-visit the estimation of $P(X, W)$ in ASR systems, which can be regarded as the matching of an acoustic feature sequence X representing the acoustic signal with the word hypothesis W, via a latent symbol set. We show that this matching problem can be cast into four sub-problems: (1) determining the latent symbol set (acoustic unit set), (2) modeling the relationship between the speech signal and latent symbols (acoustic model), (3) modeling the relationship between the lexical subword units representing the word hypothesis and the latent symbols (lexical model), and (4) choice of the cost function to locally match the evidences about the latent symbols provided by the acoustic model and lexical model. We study different ASR systems that can be recognized based on their approaches to address these sub-problems (Section 3.1).

We hypothesize that depending on the acoustic model, lexical model and the cost function, the latent symbol space can vary (Section 3.3). We validate our hypothesis by comparing different ASR approaches, namely HMM/GMM, hybrid HMM/ANN and KL-HMM using varying number of latent symbols. We show that in the KL-HMM approach, the latent symbol space is relatively small compared to HMM/GMM and hybrid HMM/ANN approaches (Section 3.4).

It is worth mentioning that part of the material in this chapter has been presented in [Razavi

et al., 2014, Razavi and Magimai.-Doss, 2014]. The idea explored in this chapter and in [Razavi et al., 2014, Razavi and Magimai.-Doss, 2014] is the same. The main difference lies in the number of latent symbols and the MLPs trained for classifying the latent symbols.

3.1 ASR as a latent symbol matching problem

As explained in Section 2.2.5, in a standard HMM-based ASR framework, the most probable word sequence is inferred by finding the most probable state sequence Q, assuming an i.i.d. distribution and first order Markov model:

$$W^* = \arg\max_{Q \in \mathscr{Q}} \sum_{t=1}^{T} \log p(\mathbf{x}_t | q_t = l^i) + \log P(q_t = l^i | q_{t-1} = l^j). \tag{3.1}$$

Assuming that language modeling and pronunciation modeling are common aspects across HMM-based approaches, the main component that is of interest to be estimated is $S = \log p(\mathbf{x}_t | q_t = l^i)$. It was noted in Section 2.2.3 that estimation of $p(\mathbf{x}_t | q_t = l^i)$ can be factored through a latent symbol set $\mathscr{A} = \{a^1, \cdots, a^d, \cdots, a^D\}$ as,

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

In that perspective, we can view the acoustic unit space as an intermediate shared space that relates to both acoustic information ($\mathbf{x}_t$) and lexical information (l^i). With that understanding, as depicted in Figure 3.1, four main components for HMM-based ASR systems can be realized [Razavi et al., 2014, Razavi and Magimai.-Doss, 2014]:

1. **Latent symbols (acoustic units):** The acoustic units $\{a^d\}_{d=1}^{D}$ can be based on either CI subword units, or clustered CD (cCD) subword units.

2. **Acoustic model:** The relationship between the acoustic feature $\mathbf{x}_t$ and the acoustic units is modeled through an *acoustic model*.

3. **Lexical model:** The relationship between the acoustic units and lexical subword unit l^i is given by a *lexical model*.

4. **Cost function:** The acoustic model evidence and the lexical model evidence are locally matched based on the cost function.

The HMM-based ASR approaches can be classified into two categories based on the choice for the cost function. In the remainder of this section, we present each category along with the choices for the acoustic model and lexical model architectures.

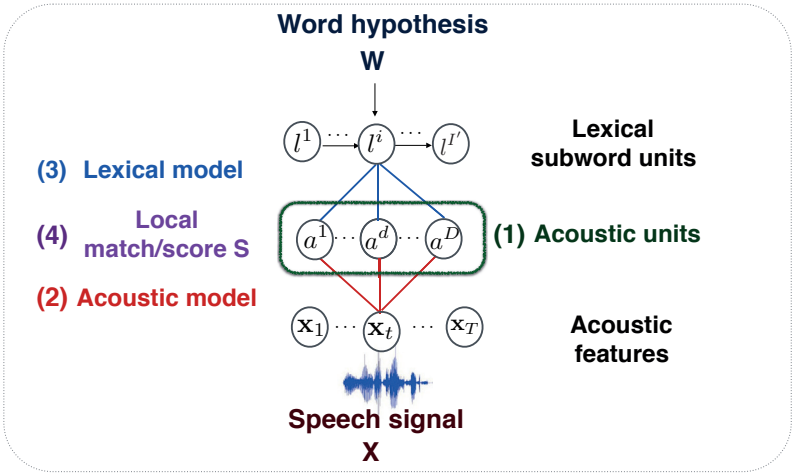


Figure 3.1 – Schematic view of an HMM-based ASR approach as a matching problem.

3.1.1 Likelihood-based matching of acoustic model evidence and lexical model evidence

As explained in Section 2.2.3, in standard HMM-based ASR systems the relationship between the acoustic observation $\mathbf{x}_t$ and acoustic units $\{a^d\}_{d=1}^D$ is modeled through either GMMs or ANNs. The GMMs estimate a likelihood probability vector $\mathbf{v}_t = [v_{t,1} \cdots v_{t,d} \cdots v_{t,D}]^{\mathrm{T}}$ with $v_{t,d} = p(\mathbf{x}_t|a^d)$. The ANNs first estimate an acoustic unit posterior probability vector $\mathbf{z}_t = [z_{t,1} \cdots z_{t,d} \cdots z_{t,D}]^{\mathrm{T}}$ with $z_{t,d} = P(a^d|\mathbf{x}_t)$. Then the scale-likelihood vector $\mathbf{v}_t$ with $v_{t,d} = p_{sl}(\mathbf{x}_t|a^d) = \frac{P(a^d|\mathbf{x}_t)}{P(a^d)}$ is estimated.

The relationship between the acoustic units $\{a^d\}_{d=1}^D$ and the lexical subword unit l^i is either deterministic or probabilistic, leading to deterministic or probabilistic lexical modeling approaches respectively. In deterministic lexical modeling approaches, as noted in Section 2.2.3, there exists a one-to-one deterministic map between acoustic units and lexical subword units. If the lexical unit l^i is deterministically mapped to the acoustic unit a^k, then the relationship is modeled through the Kronecker delta distribution $\mathbf{y}^i = [y_1^i \cdots y_d^i \cdots y_D^i]^{\mathrm{T}}$ with $y_d^i = P(a^d|l^i)$ in which,

$$y_d^i = \begin{cases} 1, & \text{if } d = k; \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

The deterministic mapping is obtained either through knowledge (for CI lexical subword units) or learned during clustering and tying of states (for CD lexical subword units). HMM/GMM systems [Rabiner, 1989] and hybrid HMM/ANN systems [Bourlard and Morgan, 1994] are examples of ASR approaches with a deterministic lexical model.

In probabilistic lexical modeling approaches, on the other hand, the relationship between acoustic units and lexical units is probabilistic. More precisely, a probabilistic lexical model

is parameterized by a categorical distribution $\mathbf{y}^i = [y_1^i \cdots y_d^i \cdots y_D^i]^{\mathrm{T}}$ with $y_d^i = P(a^d|l^i)$, which learns a probabilistic relationship between the lexical subword unit l^i and the acoustic units $\{a^d\}_{d=1}^D$. The lexical model parameters $\{\mathbf{y}^i\}_{i=1}^I$ are estimated based on the acoustic unit evidence obtained from the acoustic model. More precisely, the parameter estimation is done using the Viterbi EM algorithm with a cost function based on likelihood. In the expectation (segmentation) step, an optimal lexical unit state sequence is obtained for each training utterance using the Viterbi algorithm. Then in the maximization step, given the optimal lexical unit state sequences and the acoustic unit evidence, i.e., $\mathbf{v}_t$ belonging to each of these states, the new set of parameters $\{\mathbf{y}^i\}_{i=1}^I$ is estimated by maximizing a cost function based on likelihood with the constraint that $\sum_{d=1}^D y_d^i = 1$. More details about estimation of the parameters of the probabilistic lexical modeling approach can be found in [Rasipuram and Magimai.-Doss, 2015]. Probabilistic classification of HMM states (PC-HMM) [Luo and Jelinek, 1999] and tied posterior-based HMMs (tied-HMM) [Rottland and Rigoll, 2000] are examples of approaches with probabilistic lexical models. In the case of PC-HMMs, the likelihood vectors v_t are estimated from GMMs, whereas in the case of tied-HMMs, the scaled likelihood vectors v_t are estimated from ANNs.

Irrespective of the acoustic model or lexical model used, the local score S in the aforementioned HMM-based ASR approaches is the log of dot product between acoustic model likelihood vector $\mathbf{v}_t$ (obtained from GMM or ANN) and lexical model posterior probability vector $\mathbf{y}^i$ (obtained from a deterministic or probabilistic lexical model), i.e., $S = \log(\mathbf{v}_t^{\mathrm{T}}\mathbf{y}^i)$.

3.1.2 Posterior-based matching of acoustic model evidence and lexical model evidence

In the previous section, we observed that in standard HMM-based ASR systems the match between the acoustic model and lexical model evidence is the scalar product between the acoustic unit likelihood vector ($\mathbf{v}_t$) and the lexical model parameter probability vector ($\mathbf{y}^i$). Instead of estimation of an acoustic unit likelihood vector $\mathbf{v}_t$, the posterior probability of acoustic units $\mathbf{z}_t = [z_{t,1} \cdots z_{t,d} \cdots z_{t,D}]^{\mathrm{T}}$, $z_{t,d} = P(a^d|\mathbf{x}_t)$ can be estimated. In this case, both $\mathbf{z}_t$ and $\mathbf{y}^i$ are probability vectors, and therefore can be matched using different measures such as Bhattacharyya distance [Bhattacharyya, 1943] or Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951, Kullback, 1987]. In the case of using KL-divergence measure, the local score would be,

$$S_{KL}(\mathbf{y}^i, \mathbf{z}_t) = \sum_{d=1}^{D} y_d^i \log(\frac{y_d^i}{z_{t,d}}). \tag{3.4}$$

As KL-divergence is not a symmetric measure, the local score can be estimated in other ways such as,

$$S_{RKL}(\mathbf{y}^i, \mathbf{z}_t) = \sum_{d=1}^{D} z_{t,d} \log(\frac{z_{t,d}}{y_d^i}), \tag{3.5}$$

or

$$S_{SKL}(\mathbf{y}^i, \mathbf{z}_t) = \frac{1}{2}(S_{KL} + S_{RKL}). \tag{3.6}$$

The relationship between the acoustic units and lexical subword units, similar to the approaches explained in Section 3.1.1 can be deterministic or probabilistic. In the case of deterministic lexical modeling, $\mathbf{y}^i$ is a Kronecker delta distribution, which as explained earlier can be obtained either through knowledge or can be learned decision tree clustering and tying. With S_{KL} as local score, this case would be equivalent to hybrid HMM/ANN system in which the prior probability of acoustic units are assumed to be equal [Aradilla, 2008, Sec. 6.4.1].

In the case of probabilistic lexical modeling, the relationship between acoustic units and lexical subword units is probabilistically learned through the categorical distribution $\mathbf{y}^i$. The KL-HMM approach proposed in [Aradilla, 2008], is such an approach. The parameters of KL-HMM can be estimated by a Viterbi EM procedure similar to the probabilistic lexical modeling approaches explained in Section 3.1.1, except that (1) instead of acoustic unit likelihood vector $\mathbf{v}_t$, posterior probability of acoustic units $\mathbf{z}_t$ is estimated from the acoustic model, and (2) instead of maximizing a cost function based on likelihood, a cost function based on KL-divergence is minimized. More details about KL-HMM training and decoding are provided in Appendix A. It is worth mentioning that in the KL-HMM approach originally proposed in [Aradilla, 2008], ANNs were used to estimate the acoustic unit posterior probabilities $\mathbf{z}_t$. However, as shown in [Rasipuram and Magimai.-Doss, 2013a], the posterior probability of acoustic units $\mathbf{z}_t$ can also be estimated using GMMs.

3.2 Implications of the choices for components of ASR systems

The distinctive factors of different approaches explained in Sections 3.1.1 and 3.1.2 are summarized in Table 3.1.

Table 3.1 – Comparison of distinctive factors of different approaches based on acoustic units, lexical subword units, acoustic model, lexical model and local score.

Systems	Acoustic unit	Lexical unit	Acoustic model	Lexical model	Local score
HMM/GMM	CI cCD	CI CD	GMM (Generative)	Deterministic	$\log(\mathbf{v}_t^{\mathrm{T}}\mathbf{y}^i)$
Hybrid HMM/ANN	CI cCD	CI CD	ANN (Discriminative)	Deterministic	$\log(\mathbf{v}_t^{\mathrm{T}}\mathbf{y}^i)$
PC-HMM	CI/cCD	CI/CD	GMM (Generative)	Probabilistic	$\log(\mathbf{v}_t^{\mathrm{T}}\mathbf{y}^i)$
Tied-HMM	CI/cCD	CI/CD	ANN (Discriminative)	Probabilistic	$\log(\mathbf{v}_t^{\mathrm{T}}\mathbf{y}^i)$
KL-HMM	CI/cCD	CI/CD	ANN or GMM	Probabilistic	$S_{KL}(\mathbf{y}^i, \mathbf{z}_t)$ or $S_{RKL}(\mathbf{y}^i, \mathbf{z}_t)$ or $S_{SKL}(\mathbf{y}^i, \mathbf{z}_t)$

In that sense, we can view matching of an acoustic signal with a word hypothesis based on different aspects:

- *Generative versus discriminative acoustic modeling*: In the maximum likelihood trained HMM/GMM and PC-HMM systems the acoustic model is a generative model (GMM), whereas in hybrid HMM/ANN, tied-HMM and KL-HMM the acoustic model is a discriminative model (ANN).[1] The ANNs are directly trained to minimize acoustic unit classification error at the frame level, which could be expected to be correlated with minimizing the word error rate [Shire, 2001]. Furthermore, the non-linear functions used in the ANNs can enable better modeling of non-linear decision boundaries, compared to GMMs.

- *Deterministic versus probabilistic lexical modeling:* In HMM/GMM and hybrid HMM/ANN systems the lexical model provides a one-to-one deterministic map between acoustic units and lexical subword units. On the other hand, in tied-HMM, PC-HMM and KL-HMM the lexical model provides a soft mapping between the lexical subword unit and the acoustic units. Deterministic lexical modeling imposes certain constraints. For example, the acoustic units and lexical subword units should be of the same type. i.e., if the lexical subword units are CI or CD phones (or graphemes), then the acoustic units are also constrained to be CI or CD phones (or graphemes) respectively [Rasipuram and Magimai.-Doss, 2015]. Probabilistic lexical modeling, on the other hand, removes such constraints. For example, the acoustic units $\{a^d\}_{d=1}^{D}$ can represent phones while lexical subword units $\{l^i\}_{i=1}^{I}$ represent graphemes. We will further discuss the advantages of probabilistic lexical modeling in more detail in Chapter 4.

- *Likelihood-based versus posterior-based matching of acoustic model evidence and lexical model evidence:* In HMM/GMM, PC-HMM, hybrid HMM/ANN and tied-HMM the local score is the scalar product between acoustic unit likelihood vector and the lexical model parameter probability vector. In the KL-HMM, the local score is based on KL-divergence between the lexical model parameter vector and the acoustic unit posterior probability vector. The KL-divergence local score is discriminative, in the sense that $S_{KL}(\mathbf{y}^i, \mathbf{z}_t)$ is the expected log likelihood ratio between the lexical model parameter vector $\mathbf{y}^i$ and the acoustic unit posterior probability vector $\mathbf{z}_t$ with respect to $\mathbf{y}^i$, which is known as *discrimination* function [Blahut, 1974]. In addition to being discriminative, the KL-divergence-based local score enables giving different importance to the acoustic model and lexical model. With S_{KL} as the local score, more importance is given to the lexical model, as $\mathbf{y}^i$ is the reference distribution; with S_{RKL} as the local score, more importance is given to the acoustic model; and with S_{SKL} equal importance is given to the acoustic model and lexical model [Rasipuram and Magimai.-Doss, 2015]. The advantage of KL-divergence-based local score for parameter estimation was observed in the studies in [Rasipuram and Magimai.-Doss, 2015], in which the performance of a tied-HMM system was improved by using the parameters $\{\mathbf{y}^i\}_{i=1}^{I}$ estimated from a KL-HMM approach.

[1] Note that GMMs can also be trained discriminatively using criteria such as maximum mutual information [Bahl et al., 1986], minimum classification error [Juang and Katagiri, 1992] or minimum Bayes' risk [Kaiser et al., 2000]. Throughout this thesis, we have trained GMMs using the maximum likelihood criteria.

3.3 Research question

ASR systems have evolved from modeling of CI phones to CD phones [Schwartz et al., 1985]. The original motivation for using CD phones was the coarticulation phenomenon, i.e., phones in the preceding and the following context tend to influence the realization of the current phone. Modeling of CD phones leads to two issues: (a) data sparsity issue, i.e., not all CD phones have sufficient observations for parameter estimation. Alternatively, phones in a language do not have equal priors, and (b) unseen contexts, i.e., not all CD phones are observed during the training. HMM state clustering and tying approach was developed to handle these two issues in the HMM/GMM framework [Young, 1992, Ljolje, 1994]. As explained earlier, this leads to determination of clustered CD units (acoustic units), which is often in the order of thousands, and learning of a decision tree that maps the CD phones to cCDs (deterministic lexical model). This practice has continued with the emergence of CD phone-based hybrid HMM/ANN systems [Hinton et al., 2012].

Given the implications of different approaches to match a word hypothesis with a speech signal, we question the role of the acoustic units $\{a^d\}_{d=1}^{D}$ with respect to the acoustic model, the lexical model and the local matching cost function used. More precisely, from the perspective of achieving the match between word hypothesis and speech signal through a latent symbol space we hypothesize that D can be relatively small when using a discriminative acoustic model, a probabilistic lexical model and a discriminative local cost function, as done in the KL-HMM framework. Specifically, one of the key factors that influence performance of the ASR systems is their ability to discriminate between the words. In the deterministic lexical modeling approaches, as a one-to-one mapping between the lexical subword units and the acoustic units exists, in order to increase the discrimination between the words, the acoustic unit space must increase. On the other hand, in the probabilistic lexical modeling approaches, the soft mapping between the acoustic units and lexical subword units could potentially increase the discrimination between the words without requiring to increase the acoustic unit space. We illustrate this aspect through the example depicted in Figure 3.2.

Consider the two words *BET* and *PET* with the pronunciations /b/ /E/ /t/ and /p/ /E/ /t/ in the SAMPA format respectively. In the CI lexical subword unit representation, the Levenshtein distance (LD) between the two words is one. If we expand the CI lexical subword unit space to CD lexical subword unit space, the discrimination (LD) between the two words at the lexical level increases to two. However, if the acoustic unit space is fixed to be the CI phone space, in the deterministic lexical modeling framework, the expansion from CI to CD in lexical subword units space does not lead to increase in the model discrimination, as both /b-E+t/ and /p-E+t/ will be mapped to the central phone /E/, and therefore the LD reduces to one again. In the probabilistic lexical modeling framework, on the other hand, as a soft mapping between each CD subword unit and the acoustic units is learned, the model level discrimination can still be improved, even with an acoustic unit space based on CI phones.

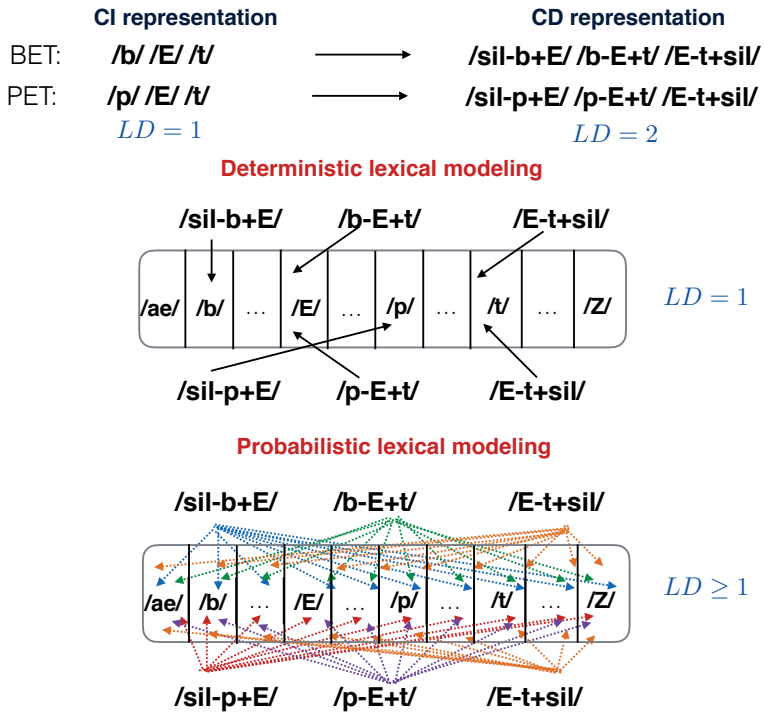


Figure 3.2 – The effect of deterministic and probabilistic lexical modeling on discrimination between lexical subword units in the same acoustic unit space. The solid lines represent a deterministic one-one relationship, while the dotted lines represent a soft relationship.

3.4 Experimental studies

In order to validate our hypothesis, we compared three systems: (1) HMM/GMM system, which uses a generative acoustic model and deterministic lexical model, (2) hybrid HMM/ANN system, which uses a discriminative acoustic model and a deterministic lexical model, and (3) KL-HMM system, which uses a discriminative acoustic model, a probabilistic lexical model and a discriminative local score.

3.4.1 Experimental setup

We conducted experimental studies on German and French part of the MediaParl corpus, described in Section 2.5.1. In this section, we explain the setup for HMM/GMM, hybrid HMM/ANN and KL-HMM systems along with the MLPs used.

HMM/GMM systems

We trained standard CI and cross-word CD HMM/GMM systems with 39 dimensional PLP cepstral features extracted using HTK toolkit [Young et al., 2006]. Each subword unit was modeled with three HMM states. In the CD HMM/GMM systems, the acoustic units were

derived by clustering CD phones in HMM/GMM framework using decision tree state tying. Different number of acoustic units D was obtained by adjusting the threshold on log-likelihood increase. In our experiments $D \in \{200, 400, 600, 800, 1000, 3000\}$. For both CI and cross-word CD HMM/GMM systems the number of Gaussians was tuned on the cross-validation set.

MLPs

For hybrid HMM/ANN and KL-HMM systems, we studied various ANNs, more precisely, MLPs that vary in terms of number of output units. We used 39-dimensional PLP cepstral features with four frames preceding context and four frames following context as MLP input. All the MLPs were trained with output non-linearity of softmax and minimum cross-entropy error criterion, using Quicknet software [Johnson et al., 2004]. We investigated the following MLPs:

- *MLP-CI-M*: a five-layer MLP modeling CI phones as output units. The number of hidden units in each layer was set to 2000. For the studies on German part of MediaParl corpus $M = 57$, and for the studies on French part of MediaParl $M = 38$.
- *MLP-CD-D*: a five-layer MLP modeling $D \in \{200, 400, 600, 800, 1000, 3000\}$ clustered CD phones obtained from HMM/GMM systems as outputs. The number of hidden units in each layer was set to 2000. In [Razavi et al., 2014, Razavi and Magimai.-Doss, 2014], all the MLP weights were randomly initialized. In this chapter, we took a different approach where we first trained a five-layer MLP classifying CI phones as the output units. We then striped off the output layer, replaced it with the clustered CD phones, and randomly initialized the weights between the last hidden layer, and the output layer. Given this initialization for the MLP weights, we then retrained the MLP. We refer to this approach as MLP pre-training. The procedure for MLP pre-training is illustrated in Figure 3.3.

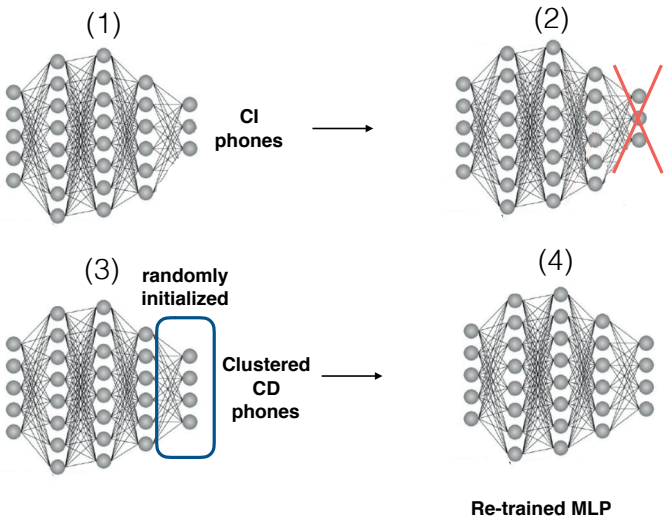


Figure 3.3 – Pre-training procedure for MLPs classifying clustered CD subword units.

Hybrid HMM/ANN systems

We estimated the scaled likelihoods in hybrid HMM/ANN system by dividing the posterior probabilities $P(a^d|\mathbf{x}_t)$ derived from MLP with the prior probability of acoustic unit $P(a^d)$ estimated from relative frequencies in the training data. These scaled likelihoods were used as emission probabilities for HMM states.

KL-HMM systems

KL-HMM systems used acoustic unit posterior probabilities as feature observations and modeled CD (tri) phones. The KL-HMM parameters were trained by minimizing the cost functions based on local scores S_{KL}, S_{SKL} and S_{RKL} and the local score that resulted in minimum KL-divergence on training data was selected. In most of the cases, this resulted in selection of S_{RKL} as the local score. For tying KL-HMM (lexical) states we applied KL-divergence-based decision tree state tying method proposed in [Imseng et al., 2012a].

3.4.2 ASR results

Figure 3.4 presents the results in terms of WRR for HMM/GMM, hybrid HMM/ANN and KL-HMM systems with varying number of acoustic units for German part of MediaParl corpus. It can be observed that for the HMM/GMM system, as the number of acoustic units increases, the WRR improves. Similar trend exists for HMM/ANN system. However, when $D \geq 800$ the increase in WRR is not statistically significant. As hypothesized, the WRR of the KL-HMM system is less sensitive to the number of acoustic units D. The system achieves the best WRR with fewer number of acoustic units ($D = 800$) compared to hybrid HMM/ANN and HMM/GMM frameworks ($D = 3000$).

With the same number of acoustic units, the hybrid HMM/ANN system performs better than the HMM/GMM system. This can be attributed to the use of ANN, which is not only discriminatively trained, but is also exploiting the acoustic contextual information. Furthermore, the KL-HMM system, which uses a probabilistic lexical model in addition to the discriminative acoustic model, performs better than hybrid HMM/ANN system. A probabilistic lexical model can better handle pronunciation variations, as a result of providing a soft mapping between acoustic units and lexical subword units. This can be particularly useful for MediaParl corpus, which contains debates which are a type of spontaneous speech. Overall, as presumed, it can be observed that as the number of acoustic units increases, the gap between the systems reduces.

It is also interesting to note that performance of the hybrid HMM/ANN system using the acoustic unit set of cardinality $D = 600$ is comparable to the KL-HMM system using CI phones as the acoustic units. Furthermore, such a KL-HMM system is able to outperform the best-performing HMM/GMM system. This trend can be attributed to the ability of probabilistic lexical model to increase the discrimination between the words in a relatively small acous-

tic unit space, as argued in Section 3.3. We will investigate this aspect in more detail in Section 3.4.3.

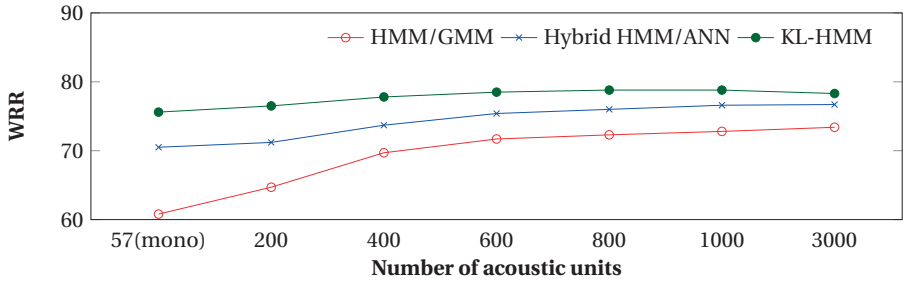


Figure 3.4 – ASR results in terms of WRR for HMM/GMM, hybrid HMM and KL-HMM systems with varying number of acoustic units on German part of MediParl corpus.

Figure 3.5 illustrates the results in terms of WRR for HMM/GMM, hybrid HMM/ANN and KL-HMM systems with varying number of acoustic units for French part of MediaParl corpus. Similar to the observations on the German part of the corpus, it can be seen that the KL-HMM system can achieve its optimal WRR with fewer number of acoustic units ($D = 600$) compared to the hybrid HMM/ANN and HMM/GMM frameworks ($D = 3000$). However, compared to Figure 3.4, it can be seen that the performance of all systems is less sensitive to the increase in the acoustic units, particularly when $D \geq 800$. This could be due to the fact that in the French part of MediaParl, about 50% of the utterances in the test set are spoken by non-native speakers, while all the speakers in training set and cross-validation set are native speakers.[2] Therefore, increasing the acoustic unit space may not be helpful in non-native speech recognition [Razavi and Magimai.-Doss, 2014].

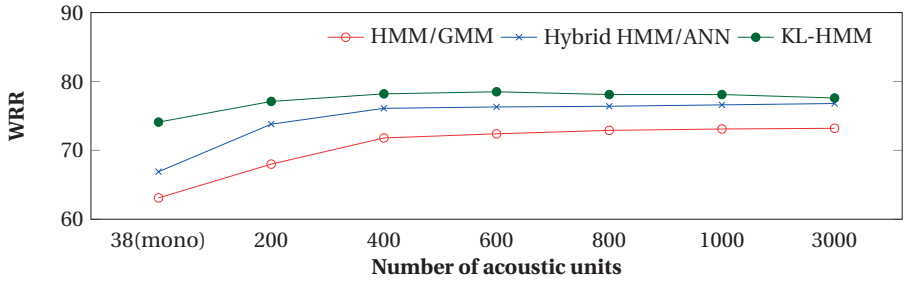


Figure 3.5 – ASR results in terms of WRR for HMM/GMM, hybrid HMM and KL-HMM systems with varying number of acoustic units on French part of MediParl corpus.

Table 3.2 summarizes the best results for hybrid HMM/ANN and KL-HMM systems in terms of WRR in German and French part of MediaParl corpus. We have also provided the best results reported in [Razavi et al., 2014, Razavi and Magimai.-Doss, 2014] in which the MLP weights were randomly initialized. It can be seen that the hybrid HMM/ANN system and KL-HMM

[2] In the German part of MediaParl corpus, only 5% of the utterances in the test set are spoken by non-native speakers.

system trained in this chapter perform better than the systems trained in [Razavi et al., 2014, Razavi and Magimai.-Doss, 2014], which indicates that the MLP pre-training scheme used in this chapter is indeed helpful. Nevertheless, the trends observed in this chapter and in [Razavi et al., 2014, Razavi and Magimai.-Doss, 2014] regarding the difference in the performance of ASR approaches remains the same.

Table 3.2 – The best performance of different ASR systems in terms of WRR on German and French part of MediaParl corpus, when using randomly initialized MLPs (as done in [Razavi et al., 2014, Razavi and Magimai.-Doss, 2014]) and pre-trained MLPs (as done in this chapter).

Systems	German		French	
	Randomly-initialized MLP	Pre-trained MLP	Randomly-initialized MLP	Pre-trained MLP
Hybrid HMM/ANN	74.5	76.7	74.5	76.8
KL-HMM	77.4	78.8	77.7	78.5

3.4.3 Analysis

The ASR results in Section 3.4.2 showed that the hybrid HMM/ANN systems perform better than the HMM/GMM systems. As the lexical models (decision trees) are the same in both systems, the improvements in the case of hybrid HMM/ANN system can be attributed to the discriminative acoustic model used, i.e., the ANN. For the hybrid HMM/ANN and KL-HMM systems on the other hand, the acoustic model is the same, and therefore, the difference in the performance of the systems can (partly) be attributed to the lexical model.[3]

In Section 3.3, we argued that with the same number of acoustic units, a probabilistic lexical modeling approach can enable better discrimination between the words compared to a deterministic lexical modeling approach. In order to validate this argument, we have estimated the discrimination between the words obtained from the KL-HMM system, which uses a probabilistic lexical model, with the hybrid HMM/ANN system, which uses a deterministic lexical model.[4] This was done by randomly selecting 100 words from the lexicon, and computing the Levenshtein distance (LD) between the CD tied state representation of each word and CD tied state representation of each of the remaining words in the dictionary according to the lexical model. Toward deriving the tied state representation of the words, the following steps were performed:

- First the CD phonetic representation of the words were obtained from the pronunciation of the word.
- Then for each CD phone unit modeled with three HMM states, the corresponding tied states were derived.

[3] Note that another different factor that can affect the performance of the two systems is the cost function.

[4] The discrimination obtained from the HMM/GMM system would be the same as the hybrid HMM/ANN system, as the lexical models are the same in both systems.

- In the case of the hybrid HMM/ANN system, the tied states are the acoustic units found through decision tree clustering in the HMM/GMM framework.
- In the case of the KL-HMM system, the tied states are obtained through the KL-divergence-based decision tree clustering method, as explained in Section 3.4.1.

Table 3.3 shows examples of the CD phonetic representation and CD tied state representation for the two words "aber" and "abord" when using a KL-HMM system and hybrid HMM/ANN system with the acoustic unit set of cardinality $D = 200$.

Table 3.3 – CD phonetic representation and CD tied state representation of the words "aber" and "abord" obtained from the KL-HMM system and hybrid HMM/ANN system using acoustic unit sets of cardinality $D = 200$, together with the Levenshtein distance (LD) between the tied state representation of the two words.

Approach		CD phonetic representation/ tied state representation	LD
KL-HMM	aber	/sil-a+b/ /a-b+E/ /b-E+R/ /e-R+sil/ ST_a_246 ST_a_390 ST_a_4108 ST_b_226 ST_b_320 ST_b_413 ST_E_270 ST_E_3123 ST_E_440 ST_R_23 ST_R_31 ST_R_456	9
	abord	/sil-a+b/ /a-b+O/ /b-o+R/ /o-R+sil/ ST_a_246 ST_a_390 ST_a_4108 ST_b_227 ST_b_319 ST_b_422 ST_O_218 ST_O_353 ST_O_455 ST_R_24 ST_R_314 ST_R_46	
Hybrid HMM/ANN	aber	/sil-a+b/ /a-b+E/ /b-E+R/ /e-R+sil/ ST_a_21 ST_a_33 ST_a_44 ST_b_21 ST_b_31 ST_b_41 ST_E_24 ST_E_31 ST_E_41 ST_R_25 ST_R_33 ST_R_41	4
	abord	/sil-a+b/ /a-b+O/ /b-o+R/ /o-R+sil/ ST_a_21 ST_a_33 ST_a_44 ST_b_21 ST_b_31 ST_b_41 ST_O_22 ST_O_31 ST_O_41 ST_R_22 ST_R_33 ST_R_41	

Given the LD between each pair of words, we computed the difference betweeen the LD obtained from the KL-HMM system and the LD obtained from the hybrid HMM/ANN system. For example, in Table 3.3 the difference in the LD between the words "aber" and "abord" is 5. Table 3.4 presents the average difference in the LD between the pair of words, when using KL-HMM and hybrid HMM/ANN systems with different number of acoustic units. It can be seen

Table 3.4 – The average difference in the LD between the pair of words, when using KL-HMM and hybrid HMM/ANN systems with different number of acoustic units in the German and French parts of the MediaParl corpus.

# of acoustic units	German	French
mono	3.9	2.1
200	2.9	1.2
400	1.6	0.6
600	1.1	0.5
800	0.8	0.4
1000	0.7	0.3
3000	0.2	0.03

that with the same number acoustic units, the probabilistic lexical model (i.e., KL-HMM) is able to better discriminate between the words compared to the deterministic lexical model (i.e., hybrid HMM/ANN). Furthermore, as the number of acoustic units increases, the difference

between the LD obtained from the KL-HMM system and the hybrid HMM/ANN system decreases. This trend is consistent with the reduction in the gap between the performance of the KL-HMM system and the hybrid HMM/ANN system as the number of acoustic units increases.

3.5 Summary

In this chapter, we studied the problem of matching an acoustic signal with a word hypothesis in the context of ASR. We showed that different ASR systems can be explained through one-and-same principle i.e., ASR by matching acoustic information obtained from the speech signal with the lexical and syntactic information obtained from the word hypothesis and pronunciation lexical through a latent symbol set. In that sense, we explained four fundamental issues in an ASR system namely, choosing the latent symbol set (acoustic unit set), modeling relationship between latent symbols and acoustic signal (acoustic model), modeling of relationship between latent symbols and lexical subword units (lexical model), and cost function to (locally) match the acoustic model and lexical model evidences.

We argued that based on the acoustic model, lexical model and the cost function used in the ASR systems, the required acoustic unit space varies. More precisely, we hypothesized that in the KL-HMM framework, in which the acoustic model is discriminative, the lexical model is probabilistic, and the local score is a measure of discrimination, the acoustic unit space can be relatively small. To validate our hypothesis, we studied different ASR systems, namely, a standard HMM/GMM system, a hybrid HMM/ANN system and a KL-HMM system using various number of acoustic units. Through experimental studies on German and French part of MediaParl corpus, we showed that the KL-HMM approach can achieve its best ASR performance using a relatively smaller acoustic unit space compared to the HMM/GMM and hybrid HMM/ANN approaches, which use a deterministic lexical model.

4 Acoustic data-driven G2P conversion using probabilistic lexical modeling

In this chapter we address the challenge of incorporating the available acoustic information in the G2P relationship learning process. Toward that, we first propose a posterior-based formalism for G2P conversion in an HMM framework, which requires estimation of the posterior probability of phones given graphemes (Section 4.1). We then build on the findings in Chapter 3 and show how phone posterior probabilities can be estimated through acoustics by formulating the problem as matching the acoustic information with the word hypothesis represented by graphemes in the probabilistic lexical modeling framework, where phones are the acoustic units (Section 4.2.1). We show that the recently proposed acoustic data-driven G2P conversion approach [Rasipuram and Magimai.-Doss, 2012a] is a particular case of this formalism where a KL-HMM is used as the probabilistic lexical model. Furthermore, we draw similarities between various G2P conversion approaches and show that local classification approaches can be seen as a particular case of the proposed posterior-based G2P conversion formalism. We validate the proposed formalism by benchmarking it against two G2P conversion approaches, namely decision tree-based approach and joint multigram approach (Section 4.3) and evaluating the generated pronunciations at both pronunciation level (Section 4.4) and ASR level (Section 4.5). We show that despite performing poorly at pronunciation level, the proposed approach can perform comparable to the state-of-the-art G2P conversion approaches at the ASR level.

It is worth mentioning that most of the material in this chapter has appeared in [Razavi et al., 2016]. The ASR studies in [Razavi et al., 2016] were conducted in the HMM/GMM framework as well as KL-HMM framework, as an indicator of the performance in the HMM/ANN framework. In this chapter, we present studies in the hybrid HMM/ANN framework, which is currently the state-of-the-art ASR framework. Thus, the ASR results reported in this chapter are consistently improved over the ASR results published in [Razavi et al., 2016], whilst the trend with respect to other G2P conversion approaches investigated remain similar.

4.1 Posterior-based G2P conversion formalism

Given a sequence of graphemes $G = (g_1, \ldots, g_n, \ldots, g_N)$, the G2P conversion problem in an HMM-based framework can be expressed as finding the most probable phone sequence F^* that can be achieved by finding the most likely state sequence S^*:

$$S^* = \underset{S \in \mathscr{S}}{\arg\max}\, P(G, S|\Theta), \tag{4.1}$$

$$= \underset{S \in \mathscr{S}}{\arg\max}\, P(G|S, \Theta) P(S|\Theta), \tag{4.2}$$

where Θ denotes the parameters of the system, $\mathscr{S}$ denotes the set of possible HMM state sequences and $S = (s_1, \cdots, s_n, \cdots, s_N)$ denotes a sequence of HMM states that corresponds to a phone sequence hypothesis with $s_n \in \mathscr{F} = \{f^1, \ldots, f^k, \ldots, f^K\}$ where K is the number of phone units. By applying i.i.d. and first order Markov assumptions, Eqn. (4.2) can be simplified as,

$$S^* = \underset{S \in \mathscr{S}}{\arg\max} \prod_{n=1}^{N} P(g_n|s_n = f^k, \Theta) P(s_n = f^k|s_{n-1} = f^{k'}, \Theta). \tag{4.3}$$

By applying Bayes' rule to Eqn. (4.3) we obtain,

$$S^* = \underset{S \in \mathscr{S}}{\arg\max} \prod_{n=1}^{N} \frac{P(s_n = f^k|g_n, \Theta) P(g_n|\Theta)}{P(s_n = f^k|\Theta)} P(s_n = f^k|s_{n-1} = f^{k'}, \Theta). \tag{4.4}$$

As $P(g_n|\Theta)$ does not affect the maximization, Eqn. (4.4) can be simplified as,

$$S^* = \underset{S \in \mathscr{S}}{\arg\max} \prod_{n=1}^{N} \frac{\overbrace{P(s_n = f^k|g_n, \Theta)}^{\text{Posterior probability}}}{\underbrace{P(s_n = f^k|\Theta)}_{\text{Prior probability}}} \underbrace{P(s_n = f^k|s_{n-1} = f^{k'}, \Theta)}_{\text{transition probability}}. \tag{4.5}$$

In Eqn. (4.5), assuming a uniform transition probability distribution and a uniform prior probability distribution, the estimation of the parameters would be restricted to learning the relationship between graphemes and phones, i.e., $P(s_n = f^k|g_n, \Theta)$. In this chapter, we will see that $P(s_n = f^k|g_n, \Theta)$ can be estimated either using a seed lexicon through local classification methods (as discussed later in Section 4.2.4) or as presented in the following section, it can be estimated by exploiting acoustic data. We refer to the latter approach as the acoustic G2P conversion approach.

4.2 Acoustic G2P conversion approach

In this section, we first explain the training phase in the acoustic G2P conversion approach in which the posterior probability of phones given graphemes are estimated using acoustic information. We then explain the pronunciation inference phase together with the implemen-

tation details. Finally, we compare the acoustic G2P conversion approach with other existing approaches in the literature.

4.2.1 Estimating $P(s_n = f^k|g_n)$ using acoustic data

As explained in Section 3.2, the probabilistic lexical modeling approaches can model different types of subword units. In that sense, the probabilistic lexical modeling framework brings certain advantages over the deterministic lexical modeling framework, which can be useful for learning the G2P relationship using acoustic information, as described below.

1. *The acoustic units and lexical subword units can represent different types of subword units*: In the deterministic lexical modeling framework, as the acoustic units and lexical subword units are deterministically related, they are constrained to be of the same type. For example, if the set of lexical subword units $\mathscr{L}$ is based on the phones (or graphemes), then the acoustic unit set $\mathscr{A}$ is also constrained to be based on phones (or graphemes). However, in the probabilistic lexical modeling framework, as a result of the probabilistic relationship between the acoustic and lexical units, the constraint is relaxed. Therefore, the acoustic units can represent phones while the lexical subword units can represent graphemes [Rasipuram and Magimai.-Doss, 2015, Magimai.-Doss et al., 2011b]. In this case, the parameters of the lexical model $\{\mathbf{y}^i\}_{i=1}^I$ capture a probabilistic G2P relationship, which is of our interest.

2. *The acoustic and lexical units can represent subword units with different context lengths*: In the deterministic lexical modeling-based ASR approaches, due to the deterministic mapping, the units are restricted to be of the same context length. For example, if $\mathscr{L}$ is based on CI or CD subword units, then $\mathscr{A}$ is also based on CI or CD subword units respectively. In the probabilistic lexical modeling-based framework, however, such a constraint is relaxed. For example, the acoustic units can represent CI subword units while the lexical units can denote CD subword units [Razavi et al., 2014, Imseng et al., 2011]. This could be beneficial for languages with complex G2P correspondence, which require modeling of longer grapheme contexts to correctly capture the relationship between graphemes and phones.

3. *The acoustic model and the lexical model can be trained on different sets of data*: In the probabilistic lexical modeling framework, the acoustic model and lexical model can be trained independently (one after another) and can exploit different sources of data during training. In [Rasipuram and Magimai.-Doss, 2015], it was shown that grapheme-based ASR systems can be effectively built by (a) training a multilingual ANN that learns the relationship between acoustic features and multilingual phones using acoustic and lexical resources from auxiliary languages, and then (b) learning a probabilistic relationship between graphemes of the target language and the multilingual phones using the target language acoustic data. Examples of similar work with the use of cross-domain acoustic and lexical resources for G2P relationship learning can be found in [Magimai.-Doss et al., 2011b, Rasipuram and Magimai.-Doss, 2012a]. Alternately, such a framework relaxes the need for a phonetic seed lexicon in the target language or domain for learning the G2P relationship. Thus, it can have potential implications for

lexicon development for under-resourced languages and domains.

In this chapter, we exploit the advantages of the probabilistic lexical modeling framework to learn the G2P relationship through acoustic data. More precisely, we cast the parameter estimation problem for the HMM explained in Section 4.1 as learning the parameters $\{\mathbf{y}^i\}_{i=1}^I$ in the probabilistic lexical modeling framework in which the acoustic unit set $\mathscr{A}$ is equal to the set of phones $\mathscr{F} = \{f^1,\ldots,f^k,\ldots,f^K\}$ (in Section 4.1) and the lexical subword unit set $\mathscr{L}$ contains the possible graphemes in the target language (i.e., $\forall G_n = g_n : g_n \in \mathscr{L}$). This is depicted in Figure 4.1.

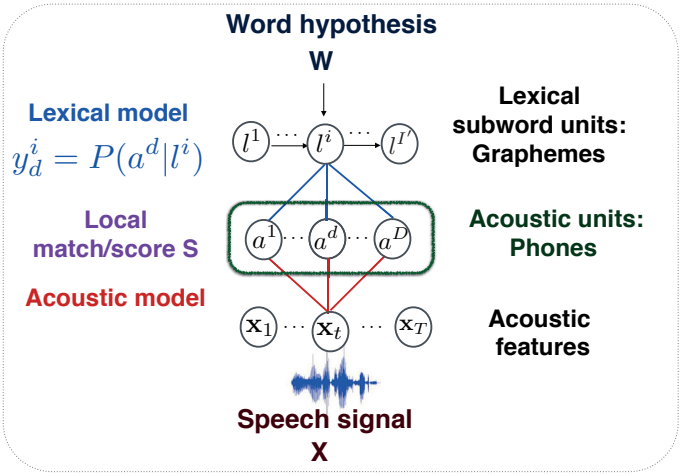


Figure 4.1 – Casting the G2P relationship learning through acoustics as learning the lexical model parameters in a probabilistic lexical modeling framework with acoustic units representing phones and lexical subword units representing graphemes.

4.2.2 Pronunciation inference

Given the orthographic transcription of the word and the estimated parameters of the probabilistic lexical model, the lexical model can be used to obtain a sequence of phone posterior probabilities. The most probable phone sequence is then inferred by decoding the sequence of phone posterior probabilities using the ergodic HMM presented in Section 4.1. Multiple pronunciations for a word can be extracted within this framework using N-best decoding. The pronunciation variants can also be generated in other ways, such as using different cost functions at the parameter estimation stage to possibly capture different G2P relationships [Razavi et al., 2015a]. However, selecting the best method for generating pronunciation variants is beyond the scope of this chapter.

4.2.3 Summary and implementation

Figure 4.2 provides a summary of the acoustic G2P conversion approach using the probabilistic lexical modeling framework as a three-step process, which is also described below.

1. *Acoustic model training*: An acoustic model (ANN or GMM) is trained to estimate phone posterior probabilities $\mathbf{z}_t$ or phone likelihoods $\mathbf{v}_t$, given the transcribed speech data and the phonetic lexicon.

2. *Grapheme-based probabilistic lexical model training*: A grapheme-based probabilistic lexical model is trained to learn the relationship between graphemes and phones, given the word-level transcribed speech data and the estimates $\mathbf{z}_t$ or $\mathbf{v}_t$ from the acoustic model.

3. *Inference*: Given the trained lexical model and the orthographic transcription of the word, the most probable sequence of phones is inferred using the HMM framework in Section 4.1. The ergodic HMM is implemented using the HTK toolkit [Young et al., 2006].

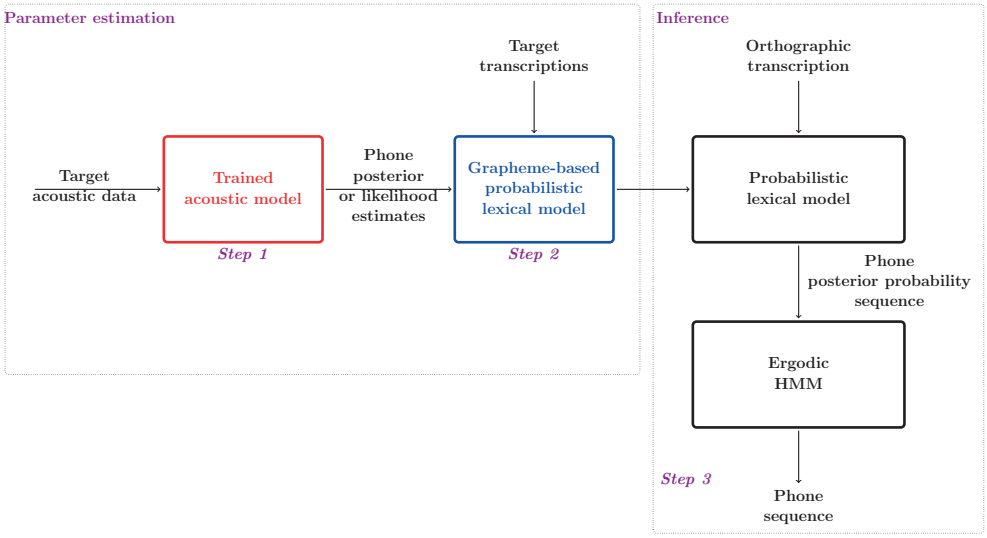


Figure 4.2 – Block diagram of the acoustic G2P conversion approach.

It can be seen that the recently proposed acoustic data-driven G2P conversion approach [Rasipuram and Magimai.-Doss, 2012a] in the KL-HMM framework is a particular case of this formalism where the acoustic model is estimating posterior probabilities $\mathbf{z}_t$ and the G2P relationship is captured through the parameters of the KL-HMM, i.e., a probabilistic lexical model. The KL-HMM approach in this case is illustrated in Figure 4.3.

In this thesis, we focus on the KL-HMM as the probabilistic lexical model. This is motivated from the previous observations in which the KL-HMM framework was found to be consistently leading to a better system compared to other probabilistic lexical modeling-based ASR approaches [Rasipuram and Magimai.-Doss, 2015].

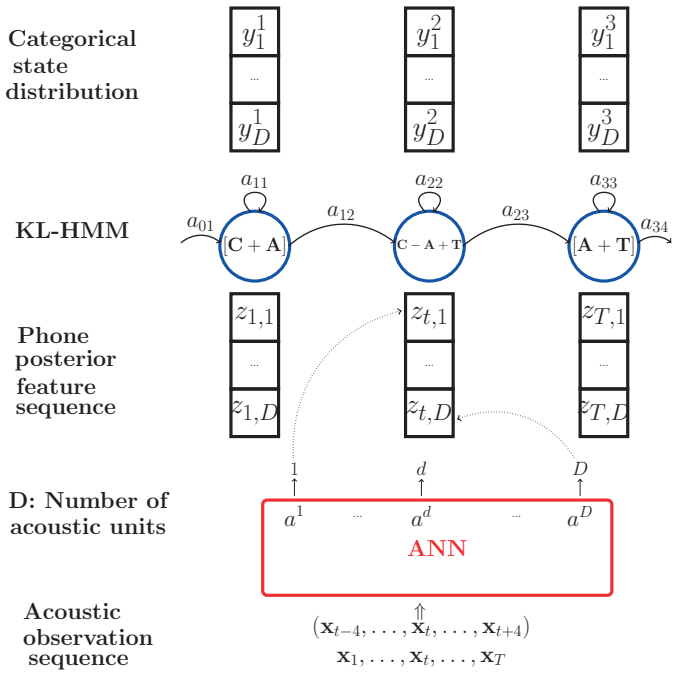


Figure 4.3 – Illustration of KL-HMM approach in which graphemes are used as lexical units and the acoustic model is an ANN.

4.2.4 Comparison to existing approaches

The parameters of the probabilistic lexical model in the acoustic G2P conversion approach are estimated using the Viterbi EM algorithm as shown in Figure 4.4. Similar to the acoustic G2P conversion approach, data-driven G2P conversion approaches can be considered to consist of an *E-step* and an *M-step*:

- The *E-step*, which provides an alignment between the grapheme sequence and the phone sequence, is common to most of the G2P conversion approaches.
- The *M-step*, which captures the relationship between graphemes and phones, is performed through different learning methods such as decision trees, neural networks, n-gram models or CRFs.

Table 4.1 further compares the acoustic G2P conversion approach with the G2P conversion approaches explained in Section 2.3.2 based on optimization criteria and required training data. The table also includes distinctive remarks on each approach.

The key distinctive factor in the acoustic G2P conversion approach is exploiting acoustic data to learn the G2P relationship, in contrast to conventional data-driven G2P conversion approaches, which use only the seed lexicon. The proposed acoustic G2P conversion approach is similar to the local classification-based approaches, as they can be both seen as a particular case of the formalism in Section 4.1 where the transition and prior probability distributions

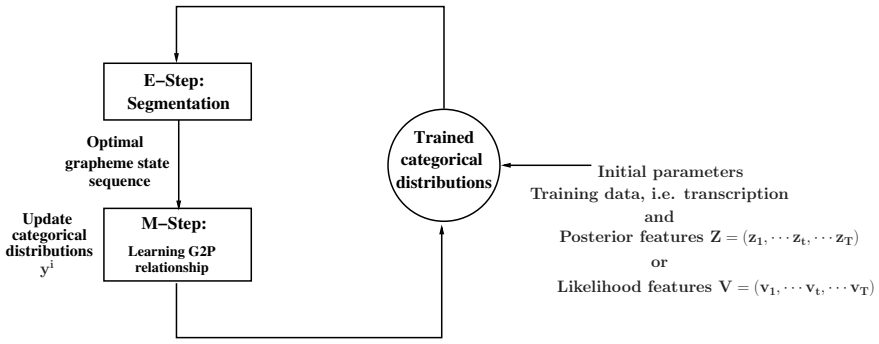


Figure 4.4 – Illustration of parameter estimation in the probabilistic lexical modeling framework, where the acoustic units represent phones and lexical units represent graphemes.

are uniform. In the local classification-based approaches, the phone posterior probabilities $P(s_n = f^k|g_n)$ are estimated either through decision trees or ANNs. For the decision tree-based approach, as the output of the decision tree is deterministic, the phone posterior probabilities would be zero or one. For the ANN-based approach, however, the output of the neural network directly provides phone posterior probability estimates.

Table 4.1 – Summary of different G2P conversion approaches based on optimization criteria, required data and distinctive remarks.

Approach	Optimization criteria	Required data	Distinctive remarks
Local classification	Discriminative	Seed lexicon	Variation of the posterior-based approach in Eqn. (4.5) where $P(s_n = f^k\|g_n)$ is estimated using decision trees/ANNs.
HMM	Generative	Seed lexicon	Models the likelihood $P(g_n\|s_n)$ unlike the posterior-based approach in Eqn. (4.5) which models $P(s_n = f^k\|g_n)$.
Joint multigram	Generative	Seed lexicon	Exploits the concept of graphones.
CRF	Discriminative	Seed lexicon	Exploits both discriminative training and global inference.
Acoustic G2P conversion	Generative	Seed lexicon & acoustic data	Exploits acoustic information to estimate $P(s_n = f^k\|g_n)$ in Eqn. (4.5).

In this chapter, we benchmark the acoustic G2P conversion approach against two conventional G2P conversion approaches: (1) decision tree-based G2P conversion approach, which like the acoustic G2P conversion approach is a particular case of the HMM-based formalism in Section 4.1, and (2) the state-of-the-art joint multigram G2P conversion approach. We evaluate the G2P conversion approaches on English and French as two languages with deep orthographies.

4.3 Experimental setup

The performance of G2P conversion approaches depends on various factors, some of which are stated below.

- *Language*: As discussed earlier, alphabetic orthographies can be deep or shallow depending on the language. The G2P conversion task for languages with deep orthographies is more challenging.
- *Seed lexicon size*: The size of the initial seed lexicon can be different depending on the amount of linguistic resources available in a language. Different G2P conversion approaches may perform differently according to the amount of training data available.
- *Variations in speech*: Depending on the type of speech data (being read or conversational, isolated or continuous, etc.) used for ASR level evaluation, the quality of generated pronunciations using G2P conversion approaches can have marginal or major effects on the performance of ASR systems.

In this chapter, we considered the aforementioned factors thoroughly to design efficient experimental studies.

4.3.1 Datasets

We conducted our studies on two databases: (1) PhoneBook, as a small-vocabulary isolated word recognition English corpus (explained in Section 2.5.2), and (2) French part of MediaParl, as a large-vocabulary continuous speech recognition (LVCSR) corpus (explained in Section 2.5.1).

PhoneBook: Isolated word recognition English corpus

The G2P conversion task on the PhoneBook corpus is challenging for several reasons: (1) the G2P relationship in English is highly irregular, (2) the training and test vocabulary sets are totally different, (3) the corpus contains uncommon English words and proper names (e.g., Witherington, Gargantuan, etc.), and (4) it can be seen as a resource-limited scenario as there are only about 2000 training words and 10 hours of transcribed speech data available. Furthermore, the reader is pointed to an existing literature [McGraw et al., 2013] that also shows the difficulty of G2P conversion on PhoneBook.

MediaParl: LVCSR bilingual corpus

The G2P conversion study on MediaParl corpus is different from the PhoneBook corpus for the following reasons: (1) in French, the G2P relationship is regular (though the conversion

rules can be complex), while in English the relationship is irregular, (2) the amount of training data is greater than for the PhoneBook corpus, (3) the number of unseen words in the test set is relatively small (20% of the words in the test set), and (4) the MediaParl corpus contains not only spontaneous speech and debates but also non-native speech.

4.3.2 Evaluation

We used the G2P conversion approaches to generate pronunciations for the words that were not seen during training. We refer to them as "G2P-generated" pronunciations. Therefore, the "G2P-based" lexicons in this chapter contain pronunciations from the manual dictionary for the words seen during training and the G2P-generated pronunciations for the unseen words. Toward pronunciation generation, we considered two scenarios: (a) *single-best pronunciation* scenario where only a single-best pronunciation per word is generated, and (b) *multiple pronunciation* scenario where pronunciation variants for the words are generated. We evaluated the G2P-based lexicons at the pronunciation level by computing PRR and WPA (explained in Section 2.4.1) and analyzing the pronunciations using a confusion matrix. The pronunciation level studies are presented in Section 4.4. As the pronunciation level evaluation may not be indicative of the performance of the systems in real applications [Hahn et al., 2013, Rasipuram and Magimai.-Doss, 2012a], we further evaluated the G2P-based lexicons through ASR tasks. The ASR level studies are presented in Section 4.5.

4.4 Pronunciation level studies

In this section, we first present the pronunciation generation setup using different G2P conversion approaches. We then compare the acoustic G2P conversion approach with the joint multigram and the decision tree-based approaches at the pronunciation level. Furthermore, we provide pronunciation level analysis for the G2P conversion approaches.

4.4.1 Pronunciation generation setup

We exploit the following G2P conversion approaches to generate both single-best pronunciations and pronunciation variants for the words unseen during training. The number of pronunciation variants were optimized, if feasible, for each approach separately to have a fair comparison between the G2P conversion approaches.[1] The hyper-parameters in each of the G2P conversion approaches were tuned on the cross-validation set. The tuning on the cross-validation set could possibly help in better generalization toward unseen contexts.

[1] Note that there is a trade-off between the coverage of alternative pronunciations and increasing the confusion between the words when adding pronunciation variants [Livescu et al., 2012]. As the generated pronunciations through each approach can be different, using the same number of pronunciation variants for all G2P conversion approaches could be suboptimal.

Decision tree-based approach

We used the Festival toolkit [Taylor et al., 1998] which is based on classification and regression trees (CART). The width of grapheme context was optimized based on the PRR on the cross-validation set. For the PhoneBook corpus, the optimal grapheme context length was 7 (three preceding and three following grapheme context). For the MediaParl corpus, the best performing grapheme context length was 9.

Predicting reliable N-best pronunciations in the decision tree-based approach is not trivial, because in CART the inference is based on individual phones and hence smoothing the confidence scores (posterior probabilities) could be difficult [Wang and King, 2011]. In this chapter, we generated multiple pronunciations by training CART trees using different grapheme context lengths. More precisely, we generated up to three pronunciations for each unseen word using the CART trees trained with grapheme contexts of length 5, 7 and 9. The average number of pronunciations for each unseen word in the PhoneBook and MediaParl corpora was 1.4 and 1.1 respectively.

Joint multigram approach

We used the Sequitur software developed at RWTH Aachen University[2]. The maximum width of the graphone used was one in both PhoneBook and MediaParl corpora. The n-gram context size was tuned on the cross-validation set and the optimal n-gram context size was 4 and 6 for the PhoneBook and MediaParl corpora respectively.

The Sequitur software enables generating pronunciation variants. the number of variants can be pre-determined or can be optimized for each word based on a threshold on the overall posterior probability mass of the generated variants. In our experiments the threshold was set to 0.7, similar to the setup provided in [Hahn et al., 2012]. The average number of pronunciations for each unseen word in the PhoneBook and MediaParl corpora was 4.9 and 2.7 respectively.

Acoustic G2P conversion approach

The acoustic G2P conversion approach includes three steps. In the first step, ANNs, more specifically MLPs, were trained. We used 39-dimensional PLP cepstral features with four preceding and four following frame context as the MLP input. All the MLPs were trained with output non-linearity of softmax and minimum cross-entropy error criterion, using the Quicknet software [Johnson et al., 2004].

In the previous studies, only three-layer MLPs were used as the posterior feature estimators [Rasipuram and Magimai.-Doss, 2012a,b]. However, recent advances in speech technology have shown that ANNs with deep architectures can improve the performance of the ASR

[2] http://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html

systems [Hinton et al., 2012]. In order to investigate the effect of different MLP architectures on the performance of the acoustic G2P conversion approach, we built the following MLPs with various number of layers and output units:

- *MLP-3-CI-M*: a three-layer MLP classifying M CI phones. For the PhoneBook corpus $M = 42$ and for the MediaParl corpus $M = 38$.
- *MLP-5-CI-M*: a five-layer MLP classifying CI phones.
- *MLP-5-CD-M*: a five-layer MLP modeling M clustered CD phones as outputs. The output units were derived by clustering CD phones in the HMM/GMM framework using decision tree state tying. Various numbers of acoustic units were derived by adjusting the log-likelihood difference, considering the observations in Chapter 3, which state that the acoustic unit space in the KL-HMM framework can be relatively small. For the PhoneBook corpus $M \in \{212, 321, 441, 642\}$ and for the MediaParl corpus $M \in \{266, 437, 626, 817\}$.

In order to determine the optimal number of units in the output layer of the MLP, first the posterior probabilities of output units belonging to the same CI unit were marginalized together. Then using the marginalized posterior probabilities, the MLP architecture with the highest frame accuracy on the cross-validation set (without considering silence) was selected. In our experiments, *MLP-5-CD-321* and *MLP-5-CD-437* led to the highest frame accuracy for the PhoneBook and MediaParl corpora respectively.

In the second step in pronunciation generation, a KL-HMM system modeling tri-graphemes (single preceding and single following contexts[3]) was trained. The choice of local score to learn the KL-HMM parameters is important as previously shown in [Rasipuram and Magimai.-Doss, 2013b]. By using the local score S_{KL}, the system is better capable of capturing one-to-one G2P relationships. On the other hand, when using S_{RKL} as the local score, the system can better handle one-to-many relationships. For the case when using S_{SKL} as local score, the system is able to capture both one-to-one and one-to-many relations. In this chapter, the KL-HMM parameters were trained by minimizing the cost function based on the local score S_{RKL} as it is suitable for the scenarios where the G2P relationship is irregular. For tying KL-HMM states we applied the KL-divergence-based decision tree state tying method proposed by Imseng et al. [2012a].

In the inference step, each MLP output unit was modeled with three left-to-right HMM states. For the case of PhoneBook, silence was removed in the ergodic HMM as it could lead to deletion of some phones when generating pronunciations. However, for MediaParl, as many of the word endings are not pronounced, silence was used in the ergodic HMM together with insertion penalties to control the amount of insertion. The insertion penalties were tuned on the cross-validation set. The inference step is demonstrated through the example word “MAP” in Figure 4.5.

[3] This is mainly due to the limitations of the HTK in tying longer contexts.

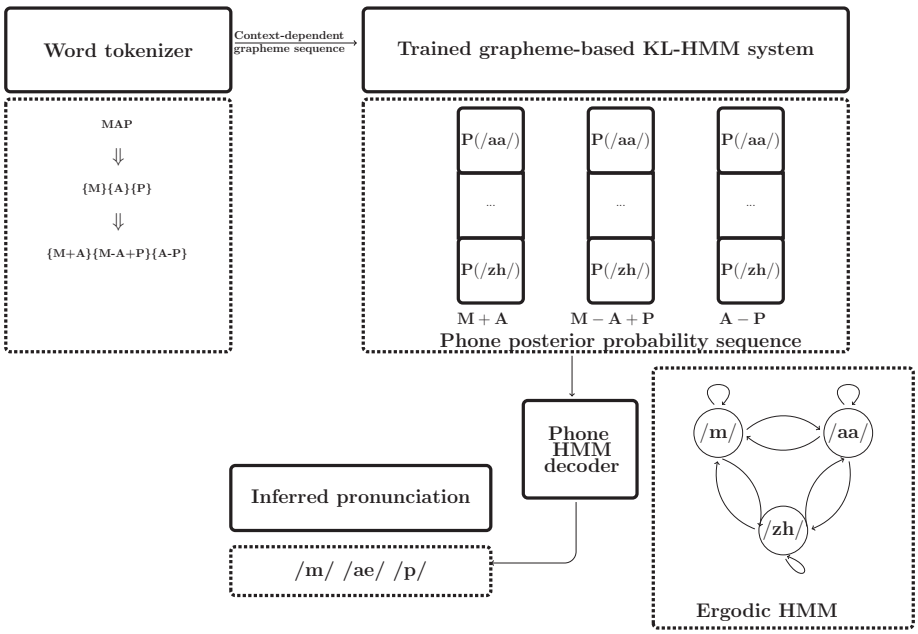
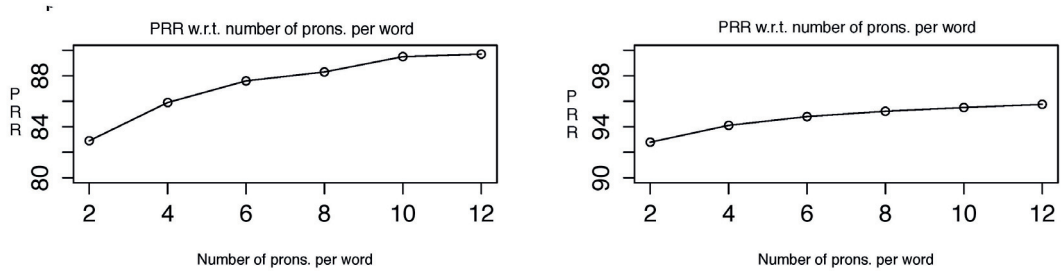


Figure 4.5 – Block diagram of the inference phase in acoustic data-driven G2P conversion task. For the sake of clarity, the figure is depicted for the case where each CD grapheme in the KL-HMM is modeled with a single HMM state.

Note that the use of clustered CD phones as MLP output units could possibly help to better model the relationship between the phones and the graphemes (similar to the effect of graphones in the joint multigram approach). However, in the inference we are interested in inferring CI phone sequences. To resolve this issue, after training the KL-HMM, for each lexical unit l^i, the parameters $\{y_d^i = P(a^d|l^i)\}_{d=1}^D$ were marginalized, i.e., the posterior probabilities of the acoustic units $P(a^d|l^i)$ belonging to the same central phone were summed together.

We generated multiple pronunciations at the inference stage through N-best decoding. Among the N-best hypotheses, the pronunciation level accuracy was calculated for the pronunciation which had the lowest Levenshtein distance to the manual pronunciation. The optimal N was then determined based on the PRR on the training words. Figure 4.6 shows the pronunciation level performance on the training words in terms of PRR. For the PhoneBook corpus, it can be seen that when $N \geq 10$ the increase in the PRR is not significant. For MediaParl, on the other hand, when $N \geq 6$ the pronunciation level performance does not change significantly. As a result, the number of pronunciations per word was selected to be 10 and 6 in the PhoneBook and MediaParl corpora respectively.

We pruned the generated N-best pronunciations by removing the silence phone and the spurious phones (consecutive appearance of the same phone) from the pronunciations. As a result of pruning, the number of unique pronunciations for each word was lower than N. The average number of unique pronunciations for each unseen word in the PhoneBook and MediaParl corpora was 7.1 and 3.7 respectively.



(a) PhoneBook corpus

(b) MediaParl corpus

Figure 4.6 – Pronunciation level performance on the training words in terms of PRR when using multiple pronunciations per word. The horizontal axis corresponds to different number of pronunciation variants N, where $N \in \{2,4,6,8,10,12\}$.

4.4.2 Pronunciation level results

Table 4.2 provides pronunciation level evaluation results in terms of PRR and WPA for different G2P conversion approaches. To better analyze different G2P conversion approaches, we have presented the results when generating pronunciations for the training words as well. For the acoustic G2P conversion approach, it can be observed that deep MLP architectures generally perform better than three-layer MLP architectures. More precisely, for PhoneBook, through use of more layers and more outputs in the MLP, the performance of the acoustic G2P conversion approach at pronunciation level constantly improves (in both *single-best pronunciation* and *multiple pronunciation* scenarios). Similar trends can be seen for the MediaParl corpus when using multiple pronunciations. However, in the *single-best pronunciation* case, exploiting a five-layer MLP alone does not lead to improvements; and the improvements are achieved when using more outputs and marginalizing the posterior probabilities in the KL-HMM.

Additionally, it can be seen that for the PhoneBook corpus, the joint multigram approach is able to generate exact pronunciations for about 94 % and 97% of the training words in the *single-best pronunciation* and *multiple pronunciation* scenarios respectively. This shows that the joint multigram approach can memorize the pronunciations. Similarly for the MediaParl corpus, the pronunciations generated by the joint multigram and decision tree-based methods are more consistent with the pronunciations in the manual dictionary compared to the acoustic G2P conversion approach.

The overall comparison of the results for different G2P conversion approaches shows that conventional G2P conversion approaches perform better than the acoustic G2P conversion approach at the pronunciation level. This can be attributed to the fact that in conventional approaches, the G2P relationship is learned through direct use of the manually-generated train lexicon, while the acoustic G2P conversion approach learns this relationship using acoustic information. Furthermore, the acoustic G2P conversion approach uses only single preceding

Table 4.2 – Pronunciation level evaluations in terms of phone recognition rate (PRR) and word-level pronunciation accuracy (WPA) using different G2P conversion approaches in the *single-best pronunciation* and *multiple pronunciation* scenarios. AG2P, JMM-G2P and DT-G2P represent acoustic G2P conversion approach, joint multigram G2P conversion approach and decision tree-based G2P conversion approach respectively.

(a) PhoneBook corpus

Approach	*Single-best pronunciation*		*Multiple pronunciation*	
	PRR (WPA) on train	PRR (WPA) on unseen	PRR(WPA) on train	PRR (WPA) on unseen
AG2P-*MLP-3-CI-42*	76.4 (16.1)	71.6 (9.8)	86.5 (39.3)	81.4 (25.2)
AG2P-*MLP-5-CI-42*	77.2 (17.9)	72.4 (10.8)	87.3 (43.1)	82.3 (29.2)
AG2P-*MLP-5-CD-321*	80.0 (23.4)	75.2 (15.4)	89.5 (50.2)	84.1 (32.6)
JMM-G2P	98.8 (93.9)	89.2 (50.5)	99.5 (97.2)	94.4 (70.1)
DT-G2P	89.3 (53.0)	85.0 (38.7)	90.9 (59.2)	87.1 (43.9)

(b) MediaParl corpus

Approach	*Single-best pronunciation*		*Multiple pronunciation*	
	PRR (WPA) on train	PRR (WPA) on unseen	PRR (WPA) on train	PRR (WPA) on unseen
AG2P-*MLP-3-CI-38*	89.9 (54.8)	88.0 (49.6)	94.1 (71.3)	92.6 (64.9)
AG2P-*MLP-5-CI-38*	89.9 (54.5)	87.8 (49.5)	94.5 (72.7)	93.1 (67.0)
AG2P-*MLP-5-CD-437*	91.4 (59.6)	89.6 (54.0)	94.8 (74.1)	93.4 (67.9)
JMM-G2P	99.8 (99.3)	97.4 (89.0)	99.9 (99.4)	98.4 (92.5)
DT-G2P	98.4 (92.8)	96.6 (85.6)	98.8 (94.5)	97.3 (88.5)

and single following grapheme contexts while conventional G2P conversion approaches exploit longer grapheme contexts. The pronunciation level results also show that through use of multiple pronunciations, the gap between the acoustic G2P conversion approach and conventional G2P conversion approaches reduces.

Finally, it is worth mentioning that the gap between the pronunciation level accuracy on the training and unseen words is significantly greater in the PhoneBook corpus compared to the MediaParl corpus. This can be due to the language difference (English versus French), existence of uncommon words and availability of fewer amount of training data in the PhoneBook corpus, which makes generalizability of the G2P conversion approaches toward unseen grapheme contexts more difficult.

4.4.3 Analysis

In this section, we provide the pronunciation level analysis for the joint multigram approach (as the state-of-the-art G2P conversion approach) and the acoustic G2P conversion approach using single-best pronunciations.[4]

[4]The comparison is provided only for the single-best pronunciations, as the main goal in this section is to compare the potential of different G2P conversion approaches, rather than investigating the effect of adding pronunciation variants.

Table 4.3 shows examples of the phone confusions according to the confusion matrix of the generated pronunciations through acoustic G2P and joint multigram G2P conversion approaches for the PhoneBook corpus. It can be observed that most of the confusions come from vowel phones such as /E/ (as in the word “aber”: /a/ /b/ /E/ /R/) which are confused with similar phones such as /x/ (as in the word “allow”: /x/ /l/ /W/) in both G2P conversion approaches. Confusions can also occur for consonant phones. For instance, the consonant phone /Z/ is confused with the phone /z/ and /S/ in the joint multigram and acoustic G2P conversion approaches respectively. For the case of acoustic G2P conversion approach, in fact the phone set size reduces as the phone /Z/ is replaced with the unvoiced phone /S/ which can be due to the confusion present at the output of MLP. It is interesting to note that the phone confusions in the two approaches can be different. For instance, in the acoustic G2P conversion approach the phone /@/ is mostly confused with /e/, while in the joint multigram approach it is confused with /x/. This indicates that the two approaches could possibly provide complementary information to each other.

Table 4.3 – Examples of the phone confusions in the generated pronunciations through acoustic G2P conversion (AG2P) and joint multigram (JMM-G2P) approaches for the PhoneBook corpus. The table presents phones together with their most confusable phones according to the confusion matrix.

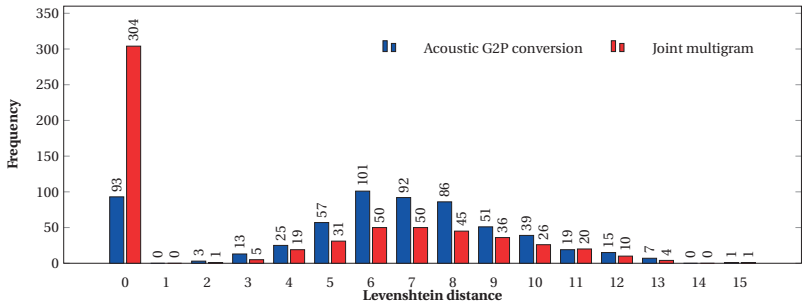
Actual phone		@	a	x	Y	E	R	X	e	I	i	o	c	u	D	Z
Confused phone	AG2P	e	o	@	x	x	X	r	@	x	x	a	a	^	T	S
	JMM-G2P	x	x, o	@,a	I	x	X	R	@	x	E	a	a	^	T	z

Similarly for MediaParl, as shown in Table 4.4, it can be seen that the confusions are mostly related to vowel phones. For example, the phone /o/ (as in the word “ausse”: /o/ /s/) is confused with the phone /O/ (as in the word “aussi”: /O/ /s/ /i/) in both G2P conversion approaches. Similar to the PhoneBook corpus, in the acoustic G2P conversion approach the phone set size is reduced since the phones /_6_/ and /_9_^/ are replaced with similar vowel phones. Furthermore, the phone confusions in the two approaches are different, similar to the observations in PhoneBook corpus. For instance, the phone /g/ is confused with the phones /Z/ and /k/ in the acoustic G2P conversion approach and joint multigram approach respectively.

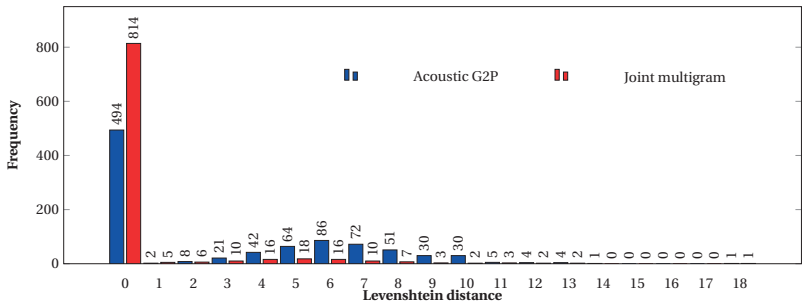
Table 4.4 – Examples of the phone confusions in the generated pronunciations through acoustic G2P conversion (AG2P) and joint multigram (JMM-G2P) approaches for the MediaParl corpus. The table presents phones together with their most confusable phones according to the confusion matrix.

Actual phone		J	g	e^	o	_6_	_9_^
Confused phone	AG2P	n	Z	n	O	@	e^
	JMM-G2P	n	k	a^	O	E	-

To further analyze the performance of the acoustic G2P conversion and joint multigram approaches at pronunciation level, we calculated the frequency of the unseen words in the test set based on Levenshtein distance between the generated pronunciation and the manual pronunciation. Figure 4.7 depicts the results when using pronunciations derived from the acoustic G2P conversion and joint multigram approaches.



(a) PhoneBook corpus



(b) MediaParl corpus

Figure 4.7 – Frequency of the words in terms of Levenshtein distance between the generated pronunciation and the manual pronunciation for PhoneBook and MediaParl databases using acoustic G2P conversion and joint multigram approaches.

For the acoustic G2P conversion approach, about 15.9% and 55.1% of the words lie within the Levenshtein distance of two in PhoneBook and MediaParl databases respectively. For the joint multigram approach, however, most of the words (50.7% and 90.2%) are within the Levenshtein distance of two in PhoneBook and MediaParl databases.

To have a better sense about the quality of the pronunciations generated by acoustic G2P conversion and joint multigram approaches, Tables 4.5 and 4.6 present examples of the generated pronunciations for the unseen words in the PhoneBook and MediaParl corpora respectively.

It can be observed from both tables that the joint multigram and acoustic G2P conversion approaches show different kinds of capabilities in generating correct pronunciations. More precisely, in the English words “yowler”, “uncharted” and “uninspired”, the acoustic G2P conversion approach is providing better pronunciations than the joint multigram approach. Similarly for the French words “anodin” and “tes”, the acoustic G2P conversion approach is

Table 4.5 – Sample unseen words from the PhoneBook corpus along with their joint multigram-based (JMM-based), acoustic G2P conversion-based (AG2P-based) and manual pronunciations.

Word	JMM-based pronunciation	AG2P-based pronunciation	Manual pronunciation
yowler	/y/ /o/ /l/ /X/	/y/ /W/ /l/ /X/	/y/ /W/ /l/ /X/
uncharted	/^/ /n/ /k/ /a/ /r/ /t/ /x/ /d/	/^/ /n/ /C/ /a/ /r/ /t/ /x/ /d/	/^/ /n/ /C/ /a/ /r/ /t/ /x/ /d/
uninspired	/^/ /n/ /I/ /n/ /s/ /p/ /Y/ /r/ /d/	/^/ /n/ /x/ /n/ /s/ /p/ /Y/ /X/ /d/	/^/ /n/ /x/ /n/ /s/ /p/ /Y/ /X/ /d/
activist	/@/ /k/ /t/ /x/ /v/ /I/ /s/ /t/	/@/ /k/ /x/ /v/ /I/ /s/ /t/	/@/ /k/ /t/ /x/ /v/ /x/ /s/ /t/
amputate	/@/ /m/ /p/ /y/ /u/ /t/ /e/ /t/	/@/ /m/ /p/ /U/ /t/ /e/ /t/	/@/ /m/ /p/ /y/ /x/ /t/ /e/ /t/
bearskin	/b/ /i/ /r/ /s/ /k/ /I/ /n/	/b/ /i/ /r/ /s/ /k/ /x/ /n/	/b/ /e/ /r/ /s/ /k/ /I/ /n/

Table 4.6 – Sample unseen words from the MediaParl corpus along with their joint multigram-based (JMM-based), acoustic G2P conversion-based (AG2P-based) and manual pronunciations.

Word	JMM-based pronunciation	AG2P-based pronunciation	Manual pronunciation
bourlard	/b/ /u/ /R/ /a/ /R/	/b/ /u/ /R/ /l/ /a/ /R/	/b/ /u/ /R/ /l/ /a/ /R/
tes	/t/	/t/ /E/	/t/ /E/
anodin	/a/ /n/ /O/ /d/ /i/ /n/	/a/ /n/ /O/ /d/ /e/ /^/	/a/ /n/ /O/ /d/ /e^/
examinerons	/E/ /g/ /z/ /a/ /m/ /i/ /n/ /@/ /R/ /o^/	/E/ /z/ /a/ /m/ /i/ /n/ /E/ /R/ /o^/	/E/ /g/ /z/ /a/ /m/ /i/ /n/ /@/ /R/ /o^/
rèadaptation	/R/ /E/ /a/ /d/ /a/ /p/ /t/ /a/ /s/ /j/ /o^/	/R/ /E/ /a/ /d/ /a/ /t/ /a/ /s/ /j/ /o^/	/R/ /E/ /a/ /d/ /a/ /p/ /t/ /a/ /s/ /j/ /o^/
banale	/b/ /a/ /n/ /a/ /l/	/b/ /a^/ /n/ /a/ /l/	/b/ /a/ /n/ /a/ /l/

able to generate correct pronunciations, while the joint multigram approach fails. On the other hand, the joint multigram approach is able to provide better pronunciations for the English words “activist” and “amputate” and for the French words “examinerons” and “banale” compared to the acoustic G2P conversion approach. As the joint multigram and acoustic G2P conversion approaches generate different types of errors, it can be hypothesized that combination of the two approaches can help in improving the ASR accuracy. We will see the effect of combination of G2P conversion approaches on the ASR performance in Section 4.5.2.

4.5 ASR level studies

We evaluated the G2P-based lexicons at the ASR level by building hybrid HMM/ANN systems considering using (a) individual G2P conversion approaches, and (b) combination of different G2P conversion approaches. We also compared the phone-based ASR system using the G2P-based lexicons with the alternative grapheme-based ASR system in the KL-HMM framework. This section presents the ASR evaluation setup and results for each of these aspects.

Note that as explained in Section 4.3.2, the G2P-based lexicon contains pronunciations from the manual dictionary for the words seen during training, and G2P-generated pronunciations for the unseen words.[5] As the pronunciations for the unseen words are added to the lexicon before decoding, the ASR systems do not have any out-of-vocabulary words. Furthermore, there is no bias in any of the ASR systems due to missing pronunciation variants for the high

[5]The rationale for this scenario is that the G2P conversion approaches are commonly used to generate pronunciations for the words that are not seen during training.

frequency words since for the PhoneBook corpus, the manual dictionary does not include any pronunciation variants for the unseen words; and for the MediaParl corpus, the unseen words occur rarely in the test set.

4.5.1 Individual G2P conversion approaches

Toward building hybrid HMM/ANN systems, for the isolated word recognition task on the PhoneBook corpus, we first trained a CD phone-based HMM/GMM system using the manual dictionary. The acoustic feature was 39 dimensional PLP cepstral features ($c_0 - c_{12} + \Delta + \Delta\Delta$) extracted using HTK [Young et al., 2006]. The number of tied states in the HMM/GMM system was 2177. We then trained a five-layer MLP classifying the tied states obtained from HMM/GMM system. The input to the MLP was 39-dimensional PLP cepstral features with four preceding and four following frame context. The number of hidden units in each hidden layer was 1000. The MLP was trained with output non-linearity of softmax and minimum cross-entropy error criterion, using the Quicknet software [Johnson et al., 2004]. For the MediaParl corpus, we used the five-layer MLP classifying 3000 tied states from HMM/GMM system explained in Section 3.4.1 as the acoustic model.

We then estimated the scaled likelihoods in the hybrid HMM/ANN systems by dividing the posterior probabilities estimated from MLPs with the prior probability of tied states estimated from relative frequencies in the training data. These scaled likelihoods were used as emission probabilities for HMM states. During decoding, the G2P-based lexicons were used.

Table 4.7 presents the performance of hybrid HMM/ANN systems in terms of WRR using single-best and multiple pronunciations from different G2P conversion approaches for the unseen words. For the sake of clarity, we have investigated the ASR experimental results in the *single-best pronunciation* and *multiple pronunciation* scenarios separately.

ASR results using single-best pronunciations

For the acoustic G2P conversion approach, it can be observed from Table 4.7 that similar to the pronunciation level results in Table 4.2, with improvements in the ANN architecture, the performance of hybrid HMM/ANN systems also improves in most of the cases.

The performance of the acoustic G2P conversion approach is not significantly different than the joint multigram and the decision tree-based G2P methods in the MediaParl corpus. However, for the PhoneBook task, the joint multigram and decision tree-based G2P conversion approaches perform significantly better than the acoustic G2P method. The difference in the behavior of the acoustic G2P conversion approach in the two databases could be due to the following factors:

- Language. Since the G2P relationship in English is irregular compared to French, it may require modeling of more than single preceding and single following grapheme context.

Table 4.7 – Performance of hybrid HMM/ANN systems in terms of WRR using different G2P conversion approaches. AG2P, JMM-G2P and DT-G2P represent acoustic G2P conversion approach, joint multigram G2P conversion approach and decision tree-based G2P conversion approach respectively.

(a) PhoneBook corpus

G2P conversion approach	*Single-best pronunciation*	*Multiple pronunciation*
AG2P-*MLP-3-CI-42*	86.3	90.8
AG2P-*MLP-5-CI-42*	87.5	91.5
AG2P-*MLP-5-CD-321*	87.5	92.1
JMM-G2P	92.0	94.9
DT-G2P	89.2	90.5
Manual dictionary	98.9	98.9

(b) MediaParl corpus

G2P conversion approach	*Single-best pronunciation*	*Multiple pronunciation*
AG2P-*MLP-3-CI-38*	76.1	76.7
AG2P-*MLP-5-CI-38*	76.0	76.8
AG2P-*MLP-5-CD-437*	76.3	76.8
JMM-G2P	76.7	76.8
DT-G2P	76.5	76.7
Manual dictionary	76.8	76.8

- Discrepancy between the manually-generated and G2P-generated pronunciations. As ican be seen from Table 4.2, the WPA for the acoustic G2P conversion approach is poor (in particular in the PhoneBook corpus). This is partly due to replacement of vowel phones with similar vowels as observed in Tables 4.3 and 4.4. As a consequence, the phone contexts seen in the manual lexicon, which are used for ASR system training, are different from the phone contexts obtained from the generated pronunciations at decoding. This effect could lead to pronunciation model mismatch at the ASR system level when training is done using manual dictionary and decoding is performed using the G2P-based pronunciations for the unseen words. The pronunciation model mismatch could particularly affect the ASR performance in the case of PhoneBook task where the words are uncommon and the words in the test data are entirely different than training data, i.e., the test set vocabulary is completely unseen. For the MediaParl corpus, however, as mentioned earlier the unseen words are 20% of the overall words in the test vocabulary, which do not appear frequently in the test set. As a result, the possible discrepancies between the existing and G2P-generated pronunciations for the unseen words may not affect the performance of the system.

In order to ascertain the effect of inconsistencies, we generated lexicons for the PhoneBook corpus, in which G2P-generated pronunciations were exploited for the seen words in addition to the unseen words (no pronunciation from the manual lexicon was used). We then trained

the ASR system using the new lexicon. Table 4.8 presents the ASR performance in terms of WRR.

Table 4.8 – Performance of ASR systems in terms of WRR when using single-best G2P-generated pronunciations at both train and test lexicons for the PhoneBook corpus. AG2P, JMM-G2P and DT-G2P represent acoustic G2P conversion approach, joint multigram G2P conversion approach and decision tree-based G2P conversion approach respectively.

G2P conversion approach	Using G2P-generated pronunciations at test lexicon	Using G2P-generated pronunciations at both train and test lexicons
AG2P-*MLP-5-CD-321*	87.5	92.7
JMM-G2P	92.0	92.9
DT-G2P	89.2	91.9
Manual dictionary	98.9	98.9

It can be observed that in all cases, the ASR systems using G2P-generated pronunciations in both train and test lexicons perform better than the systems using G2P-generated pronunciations only for unseen words. These improvements can be attributed to reducing the inconsistencies between the train and test dictionary by using G2P-generated pronunciations in both lexicons. Such observations have also been made in a previous study [Jouvet et al., 2012]. The difference between the ASR performance of the acoustic G2P conversion approach and the joint multigram approach is not statistically significant when using G2P-generated pronunciations in both train and test lexicons.

ASR results using multiple pronunciations

As can be observed from Table 4.7, for the PhoneBook corpus, using multiple pronunciations leads to significant improvements in WRR over single-best pronunciations for all the G2P conversion approaches. Furthermore, through use of multiple pronunciations, the gap between the acoustic G2P conversion approach and conventional G2P conversion approaches decreases. In the case of MediaParl, the systems using manual lexicon and G2P-based lexicon with multiple pronunciations perform similar. Similar to the studies in the *single-best pronunciation* scenario, to overcome the pronunciation inconsistency issue, we conducted experiments on the PhoneBook corpus by training an ASR system using the single-best G2P-generated pronunciations in the train lexicon, and then decoding using the multiple G2P-based pronunciations in the test lexicon. Table 4.9 presents the ASR performance in terms of WRR. It can be seen that the G2P conversion approaches can benefit from using G2P-generated pronunciations in both train and test lexicons.

4.5.2 Combination of G2P conversion approaches

As discussed earlier in Section 4.2.4, different G2P conversion approaches exploit different resources and techniques to learn the G2P relationship and infer pronunciations. It would be

Table 4.9 – Performance of ASR systems in terms of WRR when using single-best G2P-generated pronunciations at the train lexicon and multiple G2P-generated pronunciations at test lexicon for the PhoneBook corpus. AG2P, JMM-G2P and DT-G2P represent acoustic G2P conversion approach, joint multigram G2P conversion approach and decision tree-based G2P conversion approach respectively.

G2P conversion approach	Using G2P-generated pronunciations at test lexicon	Using G2P-generated pronunciations at both train and test lexicons
AG2P-*MLP-5-CD-321*	92.1	94.6
JMM-G2P	94.9	95.1
DT-G2P	90.5	93.0
Manual dictionary	98.9	98.9

interesting to investigate whether a combination of pronunciation lexicons obtained through various G2P conversion approaches can bring any benefits for the ASR systems. Table 4.10 presents the average number of unique pronunciations for each unseen word for the PhoneBook and MediaParl corpora when combining G2P-based lexicons. The results show that combining the acoustic G2P conversion approach with a conventional G2P conversion approach leads to more diverse pronunciations than combination of conventional G2P conversion approaches.

Table 4.10 – Average number of pronunciations per unseen word obtained through combining different G2P conversion approaches. The first column in each database represents the average number of pronunciations per unseen word when combining single-best pronunciations from each of the G2P conversion approaches. The second column shows the average number of pronunciations when combining pronunciation variants generated from each of the G2P conversion approaches. AG2P, DT-G2P and JMM-G2P represent acoustic G2P conversion approach, decision tree-based G2P conversion approach and joint multigram G2P conversion approach respectively.

G2P conversion approach Combinations	PhoneBook		MediaParl	
	Comb. of single-best G2P-based prons.	Comb. of multiple G2P-based prons.	Comb. of single-best G2P-based prons.	Comb. of multiple G2P-based prons.
AG2P + DT-G2P	1.9	8.2	1.4	4.7
AG2P + JMM-G2P	1.8	11.4	1.4	6.2
JMM-G2P + DT-G2P	1.6	5.7	1.1	2.8
AG2P+ JMM-G2P+ DT-G2P	2.4	12.1	1.6	6.4

Table 4.11 reports the ASR performance of hybrid HMM/ANN systems in terms of WRR when combining pronunciations from different G2P conversion approaches for the unseen words. Similar to experimental studies in Section 4.5.1, we present the ASR results using a combination of single-best pronunciations and multiple pronunciations from each of the G2P conversion approaches separately.[6]

[6]In both cases, the manual dictionary is used for training, and the generated pronunciations are used for decoding.

Table 4.11 – ASR performance in terms of WRR when combining pronunciations from different G2P conversion approaches. AG2P, JMM-G2P and DT-G2P represent acoustic G2P conversion approach, joint multigram G2P conversion approach and decision tree-based G2P conversion approach respectively.

(a) PhoneBook

G2P conversion approach	Combination of single-best G2P-based pronunciations	Combination of multiple G2P-based pronunciations
AG2P+JMM-G2P	94.2	96.4
AG2P+DT-G2P	93.1	94.7
JMM-G2P + DT-G2P	94.8	96.1
AG2P + JMM-G2P +DT-G2P	95.1	96.4
Manual dictionary	98.9	98.9

(b) MediaParl

G2P conversion approach	Combination of single-best G2P-based pronunciations	Combination of multiple G2P-based pronunciations
AG2P + JMM-G2P	76.7	76.8
AG2P + DT-G2P	76.7	76.7
JMM-G2P + DT-G2P	76.8	76.8
AG2P + JMM-G2P + DT-G2P	76.8	76.7
Manual dictionary	76.8	76.8

ASR results using combination of single-best pronunciations from each of the G2P conversion approaches

For the PhoneBook corpus, significant improvements in terms of WRR are achieved through combination of the G2P conversion approaches compared to the case using single-best pronunciations from a G2P conversion approach presented in Table 4.7 (95.1% WRR compared to 92.0% WRR).

For the MediaParl corpus, it can be seen that the systems using the lexicon obtained from combination of G2P conversion approaches yield a comparable or even the same performance as the system using the manual dictionary. However, compared to the PhoneBook corpus, the improvements in WRR through combination of G2P conversion approaches are less noticeable. This can be due to availability of larger amount of training data in the MediaParl corpus which reduces the effect of adding pronunciation variants. Furthermore, as the unseen words are only about 20% of the words in the test set, the possible improvements at the pronunciation level may not affect the performance at the ASR level significantly.

As it can be seen from Table 4.7, the performance of the systems using multiple pronunciations from the joint multigram approach (with 4.9 and 2.7 pronunciations per unseen word in PhoneBook and MediaParl respectively) is comparable to the performance of the systems using multiple pronunciations through combination of single-best G2P-based pronunciations

from various G2P conversion approaches (with 2.4 and 1.6 pronunciations per unseen word in the PhoneBook and MediaParl respectively).[7] This indicates that by obtaining multiple pronunciations through combination of single-best G2P-based pronunciations from various approaches, it is possible to achieve a similar performance to the case using multiple pronunciations from a single G2P conversion approach, but with a fewer number of pronunciation variants.

ASR results using combination of multiple pronunciations from each of the G2P conversion approaches

It can be seen from Table 4.11 that for the PhoneBook corpus, a combination of pronunciation variants from each of the G2P conversion approaches leads to improvements over the combination of single-best G2P-based pronunciations. Moreover, it brings further improvements over the case using multiple pronunciations from a single G2P conversion approach (Table 4.7).[8] This indicates that different G2P conversion approaches bring complementary information to one another. For the MediaParl corpus, similar to the observations in the previous section, the combination of G2P conversion approaches does not lead to significant changes in the ASR performance. In fact, the ASR performance in some cases slightly degrades, which could suggest that in large vocabulary continuous speech recognition tasks, adding pronunciation variants without any pruning can lead to confusions between the words.

4.5.3 Comparison with grapheme-based ASR using KL-HMM

The grapheme-based KL-HMM system was originally developed for ASR [Magimai.-Doss et al., 2011b] and was later exploited for pronunciation generation. As grapheme-based approaches can avoid the need for a phonetic lexicon, it would be interesting to investigate whether doing lexicon development and ASR training in two separate stages as done in current phone-based ASR systems can bring any benefits over grapheme-based KL-HMM systems. For this purpose, we used the grapheme-based KL-HMM systems explained in Section 4.4.1 directly for decoding, and compared them with the phone-based KL-HMM systems. More precisely, for the PhoneBook corpus we compared the grapheme-based KL-HMM system using *MLP-5-CD-321* as the acoustic model with a CD phone-based KL-HMM system that only differs in the lexicon used, i.e., instead of a graphemic lexicon it uses the manual phonetic lexicon during training, and the lexicon obtained from combination of G2P conversion approaches during decoding. Similarly for the MediaParl corpus, we compared the grapheme-based KL-HMM system using *MLP-5-CD-437* as the acoustic model with a phone-based KL-HMM system

[7] The systems using multiple pronunciations from the joint multigram approach yielded 94.9% and 76.8% WRR, and the systems using multiple pronunciations through combination of single-best G2P-based pronunciations from various G2P conversion approaches yielded 95.1% and 76.8% WRR for the PhoneBook and MediaParl corpora respectively.

[8] The system using multiple pronunciations from the joint multigram approach yielded 94.9% WRR, and the systems using multiple pronunciations through combination of pronunciation variants from each of the G2P conversion approaches yielded 96.4% WRR for the PhoneBook corpus.

using the manual lexicon during training and the lexicon obtained from the combination of G2P conversion approaches during decoding.

Table 4.12 presents the ASR results in terms of WRR. The results show that building an ASR system as a two stage process helps, since it not only enables exploiting phonetic pronunciations, but also facilitates using pronunciation variants obtained either through combination of different G2P conversion approaches or from a single G2P conversion approach.

Table 4.12 – Comparing the performance of the grapheme-based KL-HMM system with the phone-based KL-HMM systems using the pronunciations derived from the combination of G2P conversion approaches during decoding.

Database	Grapheme-based KL-HMM	phone (G2P)-based KL-HMM
PhoneBook	95.2	96.4
MediaParl	75.2	76.8

4.6 Summary

In this chapter, we presented a novel HMM-based G2P conversion formalism in which the G2P relationship is locally modeled as a distribution of phone probabilities given a grapheme input. We showed that the formalism together with recent developments in grapheme-based ASR using probabilistic lexical modeling naturally leads to a G2P conversion approach where the G2P relationship is learned through acoustics. Furthermore, the existing local classification-based G2P conversion approaches based on decision trees and ANNs can be seen as a particular case of this formulation.

We compared the proposed acoustic G2P conversion approach against the conventional G2P approaches on two different languages with deep orthographies and considered using both single-best pronunciations and multiple pronunciations per word. The studies showed that the acoustic G2P-based lexicon performs poorly at the pronunciation level compared to conventional G2P conversion approaches when using a single-best pronunciation per word. However, through use of pronunciation variants, the gap in performance between the proposed approach and conventional G2P conversion approaches reduces. Despite the relatively poor performance at the pronunciation level, the studies showed that the ASR system using the acoustic G2P-based lexicon can perform comparable to the system using a lexicon from conventional G2P conversion approaches. Furthermore, the acoustic G2P conversion approach can bring complementary information to the state-of-the-art G2P conversion approaches. i.e., combination of lexicons from the acoustic G2P conversion approach and conventional approaches can yield better ASR systems. The ASR system using the manual dictionary for both training and decoding still achieves the best performance in terms of WRR.

5 Posterior-based multi-stream formulation for pronunciation generation

In Chapter 4, we proposed a G2P conversion formalism, which requires estimation of the posterior probability of phones given graphemes. The posterior probabilities bring in certain advantages: (1) they are automatically discriminative, (2) they can be used as confidence scores [Williams and Renals, 1999, Bernardis and Bourlard, 1998], (3) they minimize the error in a Bayesian classification framework [Duda et al., 2001], and (4) they can be enhanced or refined by combining multiple complementary estimates [Genest and Zidek, 1986, Tax et al., 2000]. In this chapter, we build on the idea of combining posterior probabilities and propose a multi-stream formulation for pronunciation generation to (a) unify various G2P relationship learning techniques providing estimates of the probability of phones given graphemes, and (b) unify the orthography-based approach for pronunciation extraction (i.e., G2P conversion approach) and the acoustic exemplar-based approach for pronunciation extraction (Section 5.1). We validate the proposed multi-stream formulation on two challenging tasks on English. We show that the multi-stream formulation leads to development of lexicons that can significantly improve the performance of ASR systems (Sections 5.3 and 5.4).

It is worth mentioning that part of the work on multi-stream formulation presented in this chapter was originally published in [Razavi and Magimai.-Doss, 2017].

5.1 Multi-stream combination approach for pronunciation generation

In Section 4.1, we presented a posterior based G2P conversion formalism that requires estimation of phone posterior probabilities given graphemes $P(s_n = f^k|g_n)$ to obtain the most probable phone sequence S^*,

$$S^* = \underset{S \in \mathscr{S}}{\arg\max} \prod_{n=1}^{N} \frac{\overbrace{P(s_n = f^k|g_n)}^{\text{Posterior probability}}}{\underbrace{P(s_n = f^k)}_{\text{Prior probability}}} \cdot \underbrace{P(s_n = f^k|s_{n-1} = f^{k'})}_{\text{Transition probability}}. \tag{5.1}$$

Such a formalism is abstract in the sense that $P(s_n = f^k|g_n)$ can not only be estimated by using different techniques but also by combining multiple estimates. More precisely, in Section 4.2.4, we elucidated that $P(s_n = f^k|g_n)$ can be estimated from the seed lexicon using local classifiers such as decision trees and ANNs, or from the seed lexicon and acoustic data using a probabilistic lexical modeling approach. However, $P(s_n = f^k|g_n)$ can also be estimated as a combination of multiple estimates. In statistics such an approach can be interpreted as opinion pooling [Genest and Zidek, 1986]. This has been the underlying idea behind multiple classifier fusions in statistical pattern recognition literature and multi-stream approaches in the automatic speech recognition literature [Janin et al., 1999, Misra et al., 2003, Valente, 2010, Sun et al., 2012, Variani et al., 2013].

Given two estimates of K-class conditional probability distributions $[P(c^1|u^1)\cdots P(c^k|u^1)\cdots P(c^K|u^1)]$ and $[P(c^1|u^2)\cdots P(c^k|u^2)\cdots P(c^K|u^2)]$ for the input streams u^1 and u^2 , a refined estimate can be obtained through the product combination rule as,

$$P(c^k|u^1,u^2) = \frac{1}{Z_{prod}} \cdot \prod_{i=1}^{2} P(c^k|u^i)^{w^i_{prod}} \; \forall k, \tag{5.2}$$

and through the sum combination rule as,

$$P(c^k|u^1,u^2) = \frac{1}{Z_{sum}} \cdot \sum_{i=1}^{2} P(c^k|u^i) \cdot w^i_{sum} \; \forall k, \tag{5.3}$$

where $0 \leq w^i_{prod}, w^i_{sum} \leq 1$ are the weights, $\sum_{i=1}^{2} w^i_{prod} = 1$, $\sum_{i=1}^{2} w^i_{sum} = 1$, and Z_{prod} and Z_{sum} are normalization constants [Tax et al., 2000, Misra et al., 2003, Sun et al., 2012]. Naturally, this can be extended to the case where the number of estimates is more than two.[1]

By building on the idea that class conditional probability estimates can be refined by combining multiple estimates, we extend the posterior-based formulation to improve pronunciation lexicon development by unifying,

1. different G2P conversion approaches. More precisely, the phone class conditional probability $P(s_n = f^k|g_n)$ is estimated as a combination of estimates from different G2P conversion approaches (Section 5.1.1); and

2. the orthography-based approach and the acoustic exemplar-based approach for pronunciation generation. Alternately, the phone class conditional probability $P(s_n = f^k|g_n)$ is estimated as a combination of the phone class conditional probability estimates obtained through the orthography-based approach and the acoustic exemplar-based approach (Section 5.1.2).

[1] Note that the inputs u^i can be the same, while the classifiers can be different, which is the case for various G2P conversion techniques.

5.1.1 Unifying G2P relationship modeling techniques

In Chapter 4, we elucidated that G2P conversion involves two steps: (1) learning the G2P relationship and (2) inference of a phone sequence given the orthography and the learned G2P relationship. Furthermore, learning the G2P relationship can be further visualized in the EM framework, where the E-step is about getting the alignment between the grapheme sequence and the phone sequence, and the M-step is about learning the G2P relationship given the alignment between a grapheme sequence and a phone sequence. Given these insights, it can be observed that (a) different G2P conversion approaches mainly differ in the M-step, i.e., in learning the G2P relationship. For instance, given the alignment, (a) in the letter-to-sound (L2S) conversion approach a decision tree is trained; (b) in the joint multigram approach a joint n-gram model of graphones is estimated; (c) in the CRF-based approach a global classifier is trained; and (d) in the acoustic data-driven G2P conversion approach a categorical distribution of phone probabilities conditioned on the CD grapheme state is estimated.

Estimation of $P(s_n = f^k|g_n)$ is modeling of the G2P relationship in a statistical sense. So, as opposed to visualizing different G2P conversion approaches as separate techniques, we could envisage different approaches as means to obtain different estimates of $P(s_n = f^k|g_n)$, which could be complementary as each approach can make a different modeling assumption. Such an interpretation, as elucidated in Chapter 4, is straightforward in the case of local classifier-based approaches such as the decision tree-based, ANN-based approach and acoustic data-driven G2P conversion using KL-HMM. In the CRF-based approach, such an estimate can be obtained through the forward-backward algorithm [Lafferty et al., 2001], except that the estimate of phone probabilities is conditioned on the whole input grapheme sequence G, i.e., $P(s_n = f^k|G)$.[2]

These different estimates can be combined by employing the probability combination rules, and the phone sequence can be inferred to enhance pronunciation lexicon development. For example, when combining estimates obtained by the CRF-based approach with the acoustic data-driven approach using KL-HMM, $P(s_n = f^k|g_n)$ is estimated as,

$$\text{G2P-Comb-Prod: } P(f^k|g_n, G) = \frac{1}{Z_{prod}(n)} \cdot \left[P(f^k|g_n)^{w^{ag2p}} \cdot P(f^k|G)^{w^{crf}} \right] \tag{5.4}$$

$$\text{G2P-Comb-Sum: } P(f^k|g_n, G) = \frac{1}{Z_{sum}(n)} \cdot \left[w^{ag2p} \cdot P(f^k|g_n) + w^{crf} \cdot P(f^k|G) \right], \tag{5.5}$$

where w^{crf} is the weight given to CRF G2P relationship stream and w^{ag2p} is the weight given to acoustic data driven G2P relationship stream, $0 \leq w^{crf}, w^{ag2p} \leq 1$ and $w^{crf} + w^{ag2p} = 1$.

Figure 5.3 illustrates the proposed approach for the case when unifying the CRF-based ap-

[2] In the case of the joint multigram approach, estimation of such local phone class conditional probabilities is not straight forward due to modeling of graphone units with arbitrary context in both grapheme and phone space. However, it may be possible to estimate it by generating multiple phone sequence hypotheses by setting a threshold on $P(F|G)$ and using their respective alignment information with the grapheme sequence during the inference step.

proach and acoustic data-driven approach under the posterior-based formulation by, (a) estimating two streams or sequences of phone class conditional probabilities; (b) combining them locally using probability combination rules; and (c) inferring the phone sequence by decoding the resulting sequence of phone probability distributions through an ergodic HMM.

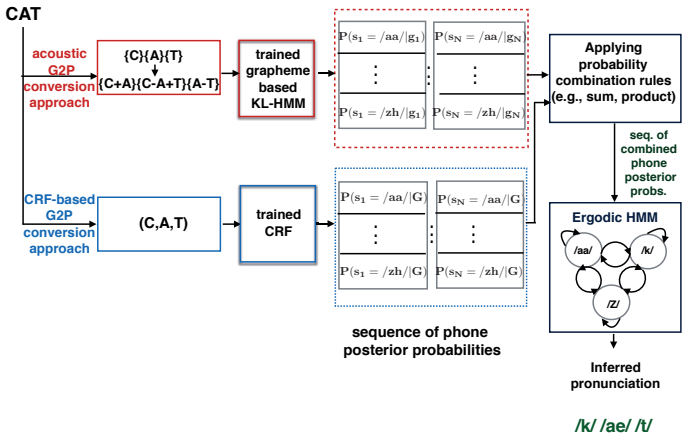


Figure 5.1 – Illustration of pronunciation inference using the multi-stream combination of CRF-based phone posterior probabilities sequence and acoustic data-driven G2P conversion-based phone posterior probabilities sequence.

5.1.2 Unifying G2P conversion and A2P conversion

A2P conversion and G2P conversion are both sequence-to-sequence conversion problems. Specifically, in the A2P conversion task, the grapheme input sequence G is replaced by the acoustic feature sequence $X = (\mathbf{x}_1, \cdots, \mathbf{x}_t, \cdots, \mathbf{x}_T)$, and the most probable phone state sequence Q^* can again be obtained by a posterior-based formulation [Morgan and Bourlard, 1995],

$$Q^* = \arg\max_{Q \in \mathscr{Q}} \prod_{t=1}^{T} \frac{\overbrace{P(q_t = f^k|\mathbf{x}_t)}^{\text{Posterior probability}}}{\underbrace{P(q_t = f^k)}_{\text{Prior probability}}} \cdot \underbrace{P(q_t = f^k|q_{t-1} = f^{k'})}_{\text{Transition probability}}, \tag{5.6}$$

where $Q = (q_1, \cdots, q_t, \cdots, q_T)$ denotes a sequence of HMM states that corresponds to a phone sequence hypothesis with $q_t \in \mathscr{F}$.

Alternately, under the posterior-based formulation, G2P conversion and A2P conversion bear a striking similarity. In both tasks, the relationship between the observations (graphemes or acoustic features) and the phones is not deterministic. Thus, there is a need for statistical techniques to learn the relationship between the observations and phones or estimate phone posterior probabilities. As discussed in the previous section, in the case of G2P relationship modeling we can envisage different techniques to estimate phone posterior probabilities. Similarly, in the case of A2P relationship modeling, phone posterior probabilities can be

estimated via local classifiers such as ANNs [Morgan and Bourlard, 1995], Gaussian mixture models [Rabiner, 1989] using Bayes' rule or global classifiers such as CRFs [Fosler-Lussier and Morris, 2008].

This understanding automatically leads to a multi-modal multi-stream approach that unifies G2P conversion and A2P conversion for pronunciation generation, where $P(f^k|g_n)$ or $P(s_n = f^k|G)$ estimated by modeling G2P relationship and $P(f^k|\mathbf{x}_t)$ estimated A2P relationship are combined using probability combination rules for each phone k,

$$\text{A2P-G2P-Comb-Prod: } P(f^k|g_n,\mathbf{x}_t) = \frac{[P(f^k|g_n)^{w^{g2p}} \cdot P(f^k|\mathbf{x}_t)^{w^{a2p}}]}{Z_{prod}(t)}, \tag{5.7}$$

$$\text{A2P-G2P-Comb-Sum: } P(f^k|g_n,\mathbf{x}_t) = \frac{[w^{g2p} \cdot P(f^k|g_n) + w^{a2p} \cdot P(f^k|\mathbf{x}_t)]}{Z_{sum}(t)}, \tag{5.8}$$

and then decoded to infer the phone sequence. $Z_{prod}(t)$ and $Z_{sum}(t)$ are normalization factors at time instance t, w^{g2p} is the weight given to the G2P relationship stream and w^{a2p} is the weight given to the A2P relationship stream, $0 \leq w^{g2p}, w^{a2p} \leq 1$ and $w^{g2p} + w^{a2p} = 1$.

Such an approach, as depicted in Figure 5.2, can be seen as a natural extension of the process to match a word hypothesis and an acoustic signal presented in Chapter 3, where a sequence of phone posterior probability vectors Y obtained from a G2P conversion approach, i.e., $Y = (\mathbf{y}_1, \cdots, \mathbf{y}_n, \cdots, \mathbf{y}_N)$, $\mathbf{y}_n = [P(f^1|g_n) \cdots P(f^k|g_n) \cdots P(f^K|g_n)]^{\mathrm{T}}$ and a sequence of phone posterior probability vectors Z obtained from an A2P conversion approach, i.e., $Z = (\mathbf{z}_1, \cdots, \mathbf{z}_t, \cdots, \mathbf{z}_T)$, $\mathbf{z}_t = [P(f^1|\mathbf{x}_t) \cdots P(f^k|\mathbf{x}_t) \cdots P(f^K|\mathbf{x}_t)]^{\mathrm{T}}$ are matched by using a local score based on KL-divergence. Then at each tuple (n,t) on the best path, $P(f^k|g_n)$ and $P(f^k|\mathbf{x}_t)$ are combined and finally decoded through an ergodic HMM.

Figure 5.3 depicts the approach to unify G2P relationship and A2P relationship under the posterior-based formulation. In comparison to the approach to unify G2P relationship modeling techniques, there is mainly one difference: the alignment step, which is needed to relate the phone information provided at different rates by A2P relationship modeling and G2P relationship modeling. Otherwise, the combination mechanism and the pronunciation inference mechanism remain the same.

5.1.3 Relation to existing literature

The proposed method takes a unified approach toward pronunciation generation. In the context of G2P conversion, it can be regarded as combination of G2P conversion approaches. Such approaches have been investigated in the literature. In [Jouvet et al., 2012, Rasipuram and Magimai.-Doss, 2012b], combination was performed at the lexicon level, in which pronunciations obtained from different G2P conversion approaches were used to develop the lexicon. There are also approaches that have investigated hypotheses level combination. For instance, in [Hahn et al., 2012] combination of various joint n-gram model based systems was

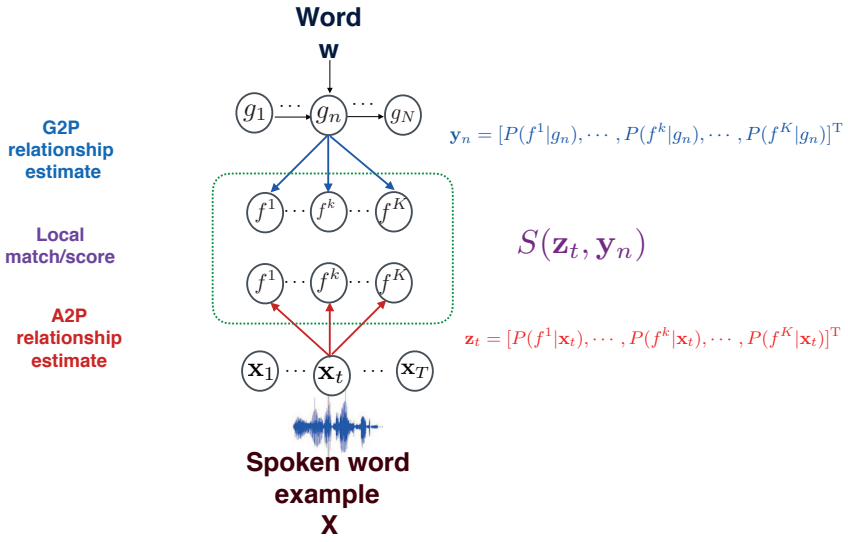


Figure 5.2 – Schematic view of the match between the phone posterior probability vector given graphemes $\mathbf{y}_n$ (from the sequence $Y = (\mathbf{y}_1, \cdots, \mathbf{y}_n, \cdots, \mathbf{y}_N)$) with the phone posterior probability vector given acoustics $\mathbf{z}_t$ (from the sequence $Z = (\mathbf{z}_1, \cdots, \mathbf{z}_t, \cdots, \mathbf{z}_T)$).

performed using ROVER [Fiscus, 1997]. Similarly, in [Schlippe et al., 2014] combination of statistical machine translation-based joint n-gram and decision tree-based G2P conversion approaches was investigated by generating an N-best lattice from the first best hypothesis from each of the approaches. Other works also exist that investigate combination of G2P approaches at the hypothesis level by representing the output of each approach by a finite state transducer (FST) and then considering the intersection of FSTs to obtain the best pronunciation [Rao et al., 2015, Wu et al., 2014]. In comparison to these approaches, a distinctive aspect of our approach is that it focuses on G2P relationship modeling, where estimate of $P(s_n = f^k|g_n)$ is refined through multiple estimators.

As discussed earlier in Section 2.3.3, in the literature typically the acoustic examples of words are exploited to select or weigh the pronunciation variants generated by the G2P converter [McGraw et al., 2013, Lu et al., 2013]. The proposed formulation for unifying A2P conversion and G2P conversion is different from these approaches, as the acoustic examples are used to refine the phone posterior probability estimation, and consequently the refined posterior probabilities are used for pronunciation generation.

5.2 Design of the validation study

For validating the proposed multi-stream formulation, we set two main criteria for choosing the test corpora: (1) difficulty of the G2P conversion task, and (2) existence of natural phonological variations. These criteria led to selection of the PhoneBook corpus and the NameDat corpus, which are both challenging tasks for pronunciation generation. For the

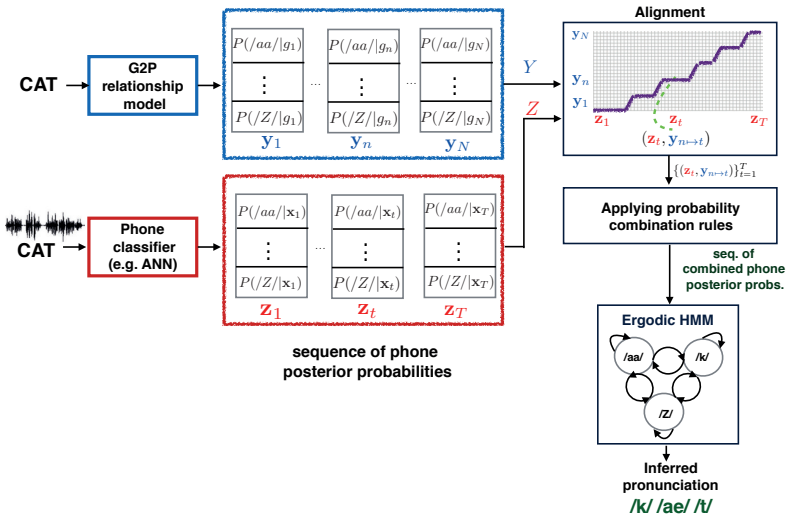


Figure 5.3 – Illustration of multi-stream combination of G2P relationship and A2P relationship.

PhoneBook corpus, at noted in Section 4.3.1, the task is difficult since the train and test words are completely different, and the corpus contains unusual words. For the NameDat corpus, the task is challenging as Norwegian speakers can pronounce English names differently, despite sharing the alphabets. This could be due to various factors such as the existence of the word in their native language with a different pronunciation. For example, David is pronounced as /deIvId/ in English while it is pronounced as /da:vIt/ in German.

As noted in the previous sections, various approaches for estimating phone class conditional probabilities exist, and as a consequence the space of possible combinations of these approaches can be large. For the sake of clarity, in this chapter, we limit our studies to use of the CRF-based G2P conversion approach and the acoustic data-driven G2P conversion approach to investigate the multi-stream formulation for unifying G2P relationship learning techniques (Section 5.3); and we use the CRF-based G2P conversion approach and the ANN-based A2P conversion approach to investigate the multi-stream formulation for unifying G2P conversion approach and A2P conversion approach (Section 5.4).[3]

5.3 Investigations on the unification of G2P relationship learning techniques

In this section, we first explain the setup for lexicon generation based on individual G2P conversion approaches and the multi-stream combination of G2P conversion approaches (Section 5.3.1). We then present the pronunciation level evaluation results (if applicable)

[3] It is worth noting that we have also studied the multi-stream combination of the acoustic G2P conversion and the decision tree-based G2P conversion approach as well as unification of the A2P conversion approach and the acoustic data-driven G2P conversion approach on the PhoneBook corpus. In both cases, we observed similar trends to the presented results in this chapter.

(Section 5.3.2) followed by the ASR level evaluation results (Section 5.3.3). Furthermore, we compare the multi-stream combination approach with the alternative approach of combining pronunciations at the lexicon level (Section 5.3.4). Finally, we provide a brief analysis on the generated pronunciations through each of the approaches (Section 5.3.5).

5.3.1 Lexicon generation setup

This section describes the setup for generating baseline lexicons based on acoustic G2P conversion approach and CRF-based G2P conversion approach, along with the setup for generating lexicons based on the multi-stream combination of the baseline approaches.

Acoustic data-driven G2P conversion approach

As a first step toward learning the probabilistic G2P relationship, following the observations in Chapter 4 regarding the MLP architecture, we trained a five-layer MLP classifying clustered CD phones using the Quicknet software [Johnson et al., 2004]:

- For the PhoneBook corpus, the input to the MLP was 39-dimensional PLP cepstral features with four preceding and four following frame context. The number of hidden units in each hidden layer was 1000. The MLP output units were 313 clustered CD phones derived by clustering CD phones in the HMM/GMM framework.[4]
- For the NameDat corpus, as the amount of training data is relatively small, we used 15 hours of data from AMI corpus [McCowan et al., 2005] to train a five-layer MLP classifying CI phones. Each hidden layer had 2000 hidden units. The labels for training the MLP were obtained from an HMM/GMM system trained on the AMI corpus using the CMU dictionary[5]. In order to adapt the ANN to the NameDat data, we first trained a phone-based HMM/GMM system on the NameDat corpus using auditory verified pronunciations in the lexicon. The labels for ANN adaptation were then obtained by force aligning the NameDat acoustic data to clustered CD phone states in the trained HMM/GMM. The ANN trained on the AMI corpus was then adapted by re-initializing the weights between last hidden layer and the output layer, which now models the clustered CD phone units from the HMM/GMM system trained on the NameDat corpus, and then training the ANN on this corpus. The number of output units for set-1, set-2 and set-3 was 369, 388 and 390 respectively.[6]

As the second step, we trained a single preceding and following CD grapheme-based KL-HMM system. In the cost function based on the KL-divergence, the output of MLP was used as the

[4] The HMM/GMM system was trained on the manual lexicon.

[5] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[6] We found that using the adapted ANN ultimately leads to a better ASR system than using the ANN trained on the AMI corpus.

reference distribution (i.e., S_{RKL} was used as the local score). To handle unseen contexts, we used the KL-divergence based decision tree state tying method proposed in [Imseng et al., 2012a]. After the KL-HMM training, as we are interested in inferring CI phone sequence, the clustered CD phone categorical distribution estimated for each state was marginalized based on the central phone information, similar to the studies in Chapter 4.

CRF-based G2P conversion approach

The CRF-based G2P conversion approach [Wang and King, 2011] is a probabilistic sequence modeling approach that enables global inference, discriminative training and relaxing the independence assumption existing in HMMs [Lafferty et al., 2001]. In the case of G2P conversion, the input to the CRF is the grapheme sequence obtained from the orthography of the word, and the CRF output is the predicted phone sequence. In this approach, the posterior probability for each phone f^k given the entire grapheme sequence G denoted as $P_{crf}(s_n = f^k|G)$ can be efficiently estimated using the well-known forward-backward algorithm [Lafferty et al., 2001]. In other words, each time instance n will yield a probability vector $[P_{crf}(s_n = f^1|G) \cdots P_{crf}(s_n = f^K|G)]^{\mathrm{T}}$.

In order to train the CRF, an initial preliminary alignment between the graphemes and phones in the training lexicon is required. In this thesis, we used the m2m-aligner [Jiampojamarn et al., 2007] to determine the G2P alignment. We treated the inserted epsilons during alignment at the phone side (e.g., "APE" $\mapsto$ "EY P EPSILON") as the silence. To train and decode the CRF, we used the publicly available CRF++ software[7]. We used bigram features and set the grapheme context to 9, i.e., four preceding and following graphemes as done in [Jouvet et al., 2012]. Note that for the NameDat corpus, we used the CMU dictionary as the training lexicon, as the amount of data in the NameDat lexicon was relatively small.[8]

Multi-Stream combination of G2P relationship learning techniques

For the PhoneBook corpus, the weights w^{crf} and w^{ag2p} were estimated by running the multi-stream combination based pronunciation inference on the training data and selecting the one yielding the highest percentage of correct phones. In our studies, for the product rule (Eqn. (5.4)) $w^{crf} = 0.8$, and for the sum rule (Eqn. (5.5)) $w^{crf} = 0.9$. For the NameDat corpus, as no canonical pronunciation is available, the weights w^{crf} and w^{ag2p} in Eqns. (5.4) and (5.5) were not tuned and were set to be 0.5, i.e., $w^{crf} = w^{ag2p} = 0.5$.

Inference

For the pronunciation inference, estimation of the prior probability $P(s_n = f^k)$ and the transition probability $P(s_n = f^k|s_{n-1} = f^{k'})$ from the seed lexicon may lead to bias, since in

[7] https://taku910.github.io/crfpp/

[8] We found that training the CRF-based approach on the CMU dictionary instead of the auditory-verified lexicon leads to development of a lexicon that yields a better ASR system.

the PhoneBook corpus, the train and test lexicons are very different and contain uncommon words, and in the NameDat corpus, the auditory verified lexicon is relatively small. Therefore, rather than estimating the prior and transition probabilities, we consider the probability distributions to be uniform. With these assumptions, Eqn. (5.1) can be rewritten as,

$$S^* = \underset{S \in \mathscr{S}}{\arg\max} \prod_{n=1}^{N} \underbrace{P(s_n = f^k | g_n)}_{\text{Posterior probability}} \ . \tag{5.9}$$

Similar to the studies in Chapter 4, for the PhoneBook corpus silence was removed in the ergodic HMM as it could lead to deletion of some phones when generating pronunciations

5.3.2 Pronunciation level evaluation

In order to evaluate the generated pronunciations at the pronunciation level, a canonical pronunciation lexicon is required. For the PhoneBook corpus, such a pronunciation lexicon is available while for the NameDat corpus this is not the case. Therefore, in this section, we present the pronunciation level results only for the PhoneBook corpus.

Table 5.1 provides the pronunciation level evaluation results in terms of the number of deletions, substitutions, insertions and PRR when combining G2P conversion approaches in the PhoneBook corpus. It can be observed that the proposed multi-stream combination method leads to significant improvements at the pronunciation level compared to the acoustic G2P conversion approach. However, it performs worse than the CRF-based approach. As can be noticed, the difference is mainly due to insertions. We will investigate the effect of insertions later in Section 5.3.5.

Table 5.1 – Pronunciation level results on the PhoneBook corpus in terms of the number of deletions (D), substitutions (S), insertions (I) and PRR for the baseline CRF-based G2P conversion approach and acoustic G2P conversion approach together with the multi-stream combination of the two approaches.

Approach	D	S	I	PRR
CRF	78	364	56	88.5
Acoustic G2P (AG2P)	111	644	245	76.9
G2P-Comb-Sum	49	379	201	85.5
G2P-Comb-Prod	52	377	127	87.1

5.3.3 ASR level evaluation

This section presents the ASR experimental setup and results on the PhoneBook corpus and the NameDat corpus respectively.

ASR studies on the PhoneBook corpus

To evaluate the proposed approach at the application level, in our case ASR, we built a CD phone-based HMM/GMM system and a hybrid HMM/ANN system. The acoustic feature was 39 dimensional PLP cepstral features ($c_0 - c_{12} + \Delta + \Delta\Delta$) extracted using HTK [Young et al., 2006]. Following the observations in Chapter 4, we used the G2P-generated lexicons to train the ASR system, as it yields better systems than the case when trained with the manual lexicon and tested with the G2P-generated lexicon. The number of tied states were between 2174 and 2270. Each tied state in the HMM/GMM system was modeled by 8 Gaussians. In the case of hybrid HMM/ANN, we trained a five-layer MLP to classify the tied states using Quicknet [Johnson et al., 2004]. We then estimated the scaled likelihoods in hybrid HMM/ANN system by dividing the posterior probabilities estimated from MLP with the prior probability of tied state estimated from relative frequencies in the training data.

Table 5.2 presents the ASR level evaluation results in terms of WRR when unifying the CRF-based G2P conversion approach and acoustic G2P conversion approach through the multi-stream combination to generate a single pronunciation per word. It can be observed that irrespective of the ASR framework used, the lexicon based on the proposed multi-stream combination approach leads to the best system. The difference between systems using lexicons based on *Comb-G2P-sum* and *Comb-G2P-prod* rules is not statistically significant. Interestingly, despite performing poorly at the pronunciation level, the acoustic G2P conversion approach yields a better system in the frameworks of hybrid HMM/ANN, and inferior system in the framework of HMM/GMM when compared to CRF-based approach. In all cases though the performance of the systems based on CRF and acoustic G2P conversion approaches is statistically comparable. This trend is more attributed to the fact that acoustic G2P conversion approach typically leads to acoustically confusable substitutions (as seen in Section 4.4.3), which a discriminative acoustic model (ANN) seems to handle better than a generative acoustic model (GMM). Finally, the best performance of 93.1% is considerably lower than manual dictionary-based best system performance of 98.9%. This is indicative of the difficulty of the G2P conversion task on the PhoneBook corpus.

Table 5.2 – ASR level evaluations on the PhoneBook corpus in terms of WRR when using different lexicons based on individual G2P conversion approaches (Acoustic-G2P and CRF-G2P) and the multi-stream combination of G2P conversion approaches. † denotes that the performance gain is statistically significant compared to the best performing individual G2P conversion approach.

	Acoustic-G2P	CRF-G2P	*G2P-Comb-Sum*	*G2P-Comb-Prod*
HMM/GMM	88.5	89.2	90.4†	89.9
Hybrid HMM/ANN	92.7	92.1	93.1	93.1

ASR studies on the NameDat corpus

For the NameDat corpus, only one hour of data for each training set is available, which may not be sufficient for effective training of hybrid HMM/ANN systems. As it has been shown that the KL-HMM approach can effectively handle the acoustic data scarcity problem [Imseng et al., 2013, Rasipuram and Magimai.-Doss, 2015], we conducted ASR studies in the KL-HMM framework, using the three-fold training and testing strategy explained in Section 2.5.3. Toward that, for each set, we first used the corresponding five-layer MLP classifying clustered CD phones (explained in Section 5.3.1) to obtain posterior feature observations. We then trained single preceding and following CD phone based KL-HMM systems using the generated lexicons.

Table 5.3 presents the ASR results when using lexicons obtained from the CRF-based approach, the acoustic G2P conversion approach and the multi-stream combination of these approaches to generate a single pronunciation per word. Furthermore, the performance of the KL-HMM system trained using the auditory verified pronunciation lexicon is also provided as a strong baseline. The ASR results are provided in terms of WRR on each test set along with the average performance. It can be observed that the two baselines perform comparable to each other on average. It is also interesting to note that the system using auditory verified pronunciations performs only slightly better than the two CRF-based and acoustic G2P conversion based systems on average. This could be due to the use of multiple pronunciations for each word in the auditory verified lexicon, which can lead to confusion between the words. It can be seen that the systems using the pronunciations from the multi-stream combination perform better than the baseline systems in almost all cases. However, compared to the PhoneBook corpus, the improvements obtained through multi-stream combination are less significant in some cases. This could be due to the following reasons:

1. In the NameDat corpus, all the words are seen during training while in the PhoneBook corpus, the words in the test set are not seen during training.

2. In the NameDat corpus, the weights for the CRF-based G2P relationship stream and the acoustic G2P relationship stream are not tuned, while for the PhoneBook corpus the weights are tuned.

3. In the NameDat corpus, the words are pronounced by non-native speakers while in the PhoneBook corpus this is not the case. As the words in the NameDat corpus can be pronounced differently depending on the non-native speaker, a single pronunciation for each word obtained through the multi-stream combination may not capture all the possible variants.

It is worth mentioning that the proper name recognition task on the NameDat corpus was also studied in [Adde and Svendsen, 2011], where the pronunciation variants for the words were selected through a discriminative tree search. However, a fair comparison between the

Table 5.3 – ASR level evaluations in terms of WRR using pronunciations obtained from the multi-stream combination of the CRF-based approach and the acoustic data-driven G2P conversion approach on the NameDat corpus.

	Auditory verified	Acoustic G2P	CRF G2P	*G2P-Comb-Sum*	*G2P-Comb-Prod*
KL-HMM-set-1	94.1	94.0	94.5	94.5	94.7
KL-HMM-set-2	94.4	94.9	94.1	95.3	95.6
KL-HMM-set-3	94.2	93.6	93.7	93.9	93.2
Average	94.2	94.2	94.1	94.6	94.5

proposed approach in [Adde and Svendsen, 2011] and the proposed multi-stream formulation cannot be drawn, as the ASR systems presented in this chapter using the baseline G2P conversion approaches already perform better than the ASR systems in [Adde and Svendsen, 2011] using the selected pronunciation variants.[9]

5.3.4 Comparison to combination of lexicons

An alternative approach for exploiting different G2P conversion approaches would be to obtain pronunciation lexicons by combining the lexicons developed using the individual G2P conversion approaches, as also studied in Section 4.5.2. Table 5.4 presents the results of the ASR study on the PhoneBook corpus, comparing the lexical level combination of the CRF-based approach and acoustic G2P conversion approach, i.e., simply merging the lexicons (Acoustic G2P+CRF) against the multi-stream approach with two-best pronunciations. It can be seen that ASR systems using the multi-stream combination lexicon perform better than the systems using merged lexicon. This indicates that combination at the pronunciation inference level can be more fruitful than combination at the lexical level.

Table 5.4 – The performance of ASR systems in terms of WRR on the PhoneBook corpus, when using lexicons obtained through the lexical level combination of G2P conversion approaches versus the multi-stream combination of G2P conversion approaches. ‡ denotes that the performance gain is statistically significant

	Acoustic G2P +CRF	*G2P-Comb-sum*	*G2P-Comb-prod*
HMM/GMM	91.7	93.0‡	92.4
Hybrid HMM/ANN	94.2	94.9‡	94.4

Table 5.5 compares the combination of the acoustic G2P conversion approach and the CRF approach at the lexicon level with the multi-stream combination of the approaches using two-best pronunciations on the NameDat corpus. It can be seen that the system using the

[9] The best ASR system reported in [Adde and Svendsen, 2011] achieves the WRR of 88% on average.

multi-stream combination approach with the sum rule performs better than the system using the combined lexicons. It is interesting to note that adding pronunciation variants does not lead to improvement in all cases. This can be clearly seen in the case of using the multi-stream combination approach with the product rule. This is due to the fact that adding pronunciation variants can lead to confusion between the words. In fact in the NameDat corpus, several words with the same or similar pronunciations exist, which makes the task quite challenging. Examples of such word pairs are (*Berwin*, *Berwyn*), (*Windgate*, *Wingate*), and (*Worthen*, *Worden*). Overall, it can be seen that the multi-stream combination approach with a single pronunciation per word can already perform better than the combination of approaches at the lexical level.

Table 5.5 – The performance of ASR systems in terms of WRR on the NameDat corpus, when using lexicons obtained through the lexical level combination of G2P conversion approaches versus the multi-stream combination of G2P conversion approaches.

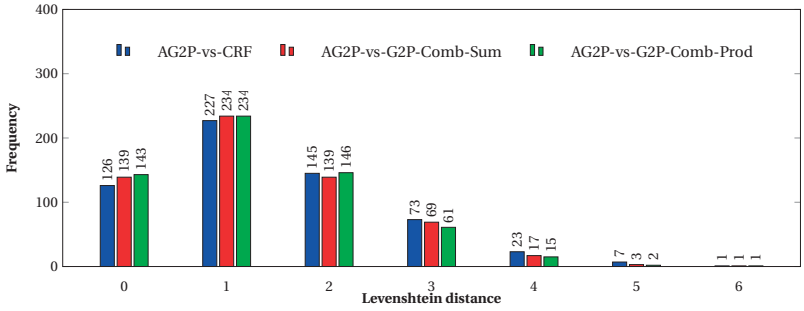
	Acoustic G2P +CRF	G2P-Comb-Sum	G2P-Comb-Prod
KL-HMM-set-1	94.9	94.9	94.5
KL-HMM-set-2	94.2	95.3	95.3
KL-HMM-set-3	93.9	94.2	93.1
Average	94.3	94.8	94.3

5.3.5 Analysis

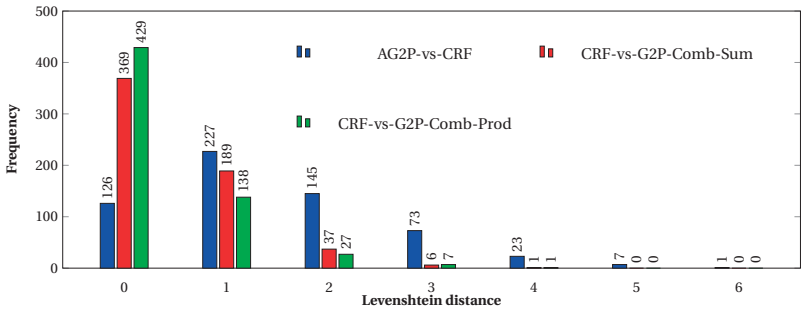
In order to understand if the multi-stream approach is indeed effective, we conducted analysis studies on the PhoneBook corpus, which has canonical pronunciations. We first analyzed the generated pronunciations by investigating how many pronunciations are different across the generated lexicons and how different they are. This was done by computing the Levenshtein distance between the generated pronunciations for the words in the test set. More precisely, we computed the Levenshtein distance between the two pronunciations for each word: one obtained through an individual G2P conversion approach and the other obtained through the multi-stream combination approach, as depicted in Figure 5.4. The figure also provides the Levenshtein distance between the pronunciation obtained through the CRF-based approach and the pronunciation obtained through the acoustic G2P conversion approach for each word in the test set.

It can be observed that for the majority of the words, the generated pronunciations using the acoustic G2P conversion approach are different from the generated pronunciations using the CRF-based approach or the multi-stream combination approach. Among the words with different pronunciations, however, the Levenshtein distance in most of the cases is less than or equal to two. The generated pronunciations through the multi-stream combination approach are more similar to the pronunciations obtained using the CRF-based approach than the

acoustic G2P conversion approach. This could be expected, as for the product rule (Eqn. (5.4)) $w^{crf} = 0.8$, and for the sum rule (Eqn. (5.5)) $w^{crf} = 0.9$. Despite that, we can observe that for about 40% and 30% of the words, the generated pronunciations using the CRF-based approach are different than the generated pronunciations using the *G2P-Comb-Sum* rule and the *G2P-Comb-Prod* rule respectively.



(a) Acoustic G2P conversion



(b) CRF-based approach

Figure 5.4 – Frequency of the words in terms of Levenshtein distance between the generated pronunciations for the test set, either through individual G2P conversion approaches (i.e., acoustic G2P conversion (AG2P) approach/CRF-based approach) or through an individual G2P conversion approach and the multi-stream combination approach.

We further analyzed the generated pronunciations by computing the confusion matrix for the generated pronunciations through each of the approaches. Figure 5.5 illustrates the percentage correctly labeled for a few example phones when using the multi-stream combination of G2P conversion approaches. It can be seen that, in most cases, the CRF-based G2P conversion approach is the best individual model. However, there are cases where the acoustic G2P conversion approach performs better, despite its overall poor PRR. Nevertheless, the proposed multi-stream approach is able to perform better than or equal to the best individual models.

Table 5.6 presents a few example pronunciations inferred using the multi-stream combination of G2P relationship estimates. It can be observed that the multi-stream combination is able to leverage from the individual models to generate a better pronunciation. For example, for the word *EXORBITANT*, the CRF-based approach is able to correctly predict the /aa/ sound, the acoustic G2P conversion approach is able to correctly predict the /g/ and /z/ sounds, and

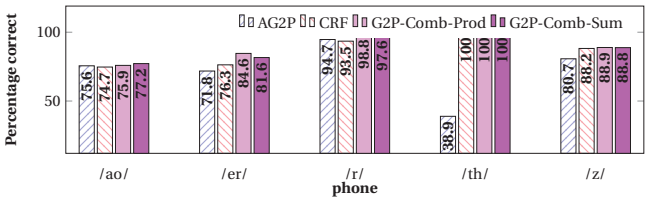


Figure 5.5 – Percentage correct for a selected few phones according to the confusion matrix for individual G2P conversion approaches together with the multi-stream combination of the approaches on the PhoneBook corpus.

the multi-stream combination approach is able to exploit the merits of both approaches to correctly predict the whole pronunciation.

Table 5.6 – Pronunciations generated by individual G2P conversion approaches along with the multi-stream combination of the approaches on the PhoneBook corpus.

Pronunciation	ATTRIBUTION	ORION	EXORBITANT
CRF-based	ae t r aa b uw sh aa n	ao r aa n	aa k s ao r b aa t aa n t
Acoustic G2P	ae t r ay b ah sh aa n	ao r iy aa n	aa g z ao r b aa t ae n t
Combination	ae t r aa b y uw sh aa n	ao r ay aa n	aa g z ao r b aa t aa n t
Manual	ae t r aa b y uw sh aa n	ao r ay aa n	aa g z ao r b aa t aa n t

These analyses show that indeed the multi-stream combination is exploiting the complementarities of the individual G2P relationship learning techniques. However, it does not explain the difference in the trend observed at PRR level and ASR level, i.e., at the pronunciation level the CRF-based lexicon yields a better PRR than the multi-stream combination based lexicons, but at the ASR level it yields inferior performance. One plausible reason could be that PRR is measured with a single manual pronunciation as a reference, while uncommon English words and proper names can exhibit more pronunciation variability. Another reason could also be that the multi-stream G2P conversion is making systematic errors that an ASR system is able to compensate. To further understand that aspect, we examined the pronunciation level errors closely. It can be observed in Tables 5.1 that low PRR for the multi-stream combination when compared with the CRF-based approach is mainly due to insertions. Therefore, we examined the generated pronunciations to investigate the type of insertions. We found that several of the insertions were due to systematic insertion of acoustically close phones, such as /axr/ → /axr/ /r/ or /ey/ → /ey/ /iy/. We speculate that the ASR level trend is a combination of two factors: fewer deletions, and the ability of the ASR system to handle the systematic errors present at the output of multi-stream pronunciation generation process.

5.4 Investigations on the unification of G2P and A2P conversion approaches

In order to validate the proposed approach for the unification of A2P and G2P conversion methods, as mentioned earlier, we investigated using an ANN to estimate the A2P relationship, and using a CRF model to estimate the G2P relationship. The G2P relationship estimates based on the CRF model were obtained in the same setup explained in Section 5.3.1. In this section, we first explain the setup for lexicon generation based on the ANN-based A2P conversion approach and the multi-stream combination of the CRF-based G2P conversion approach and the ANN-based A2P conversion approach (Section 5.4.1). We then evaluate the generated pronunciations at the pronunciation level (Section 5.4.2) and ASR level (Section 5.4.3). Furthermore, we evaluate the approach by comparing it against the alternative approach of pronunciation variant selection using acoustics (Section 5.4.4). Finally, we provide a brief analysis on the generated pronunciations (Section 5.4.5).

5.4.1 Lexicon generation setup

This section explains the setup for generating the baseline lexicon based on the A2P conversion approach, together with the setup for generating lexicons based on the multi-stream combination of ANN-based A2P conversion and CRF-based G2P conversion approaches.

A2P conversion approach

In order to generate pronunciations using the A2P conversion approach, first five-layer MLPs similar to the setup in 5.3.1 were trained, except that instead of clustered CD phones, CI phones were used in the output layer of the MLP. The trained MLPs were used to estimate a sequence of phone posterior probability vectors $Z = (\mathbf{z}_1, \cdots, \mathbf{z}_t, \cdots, \mathbf{z}_T)$, given a spoken word example represented as a sequence of cepstral features $X = (\mathbf{x}_1, \cdots, \mathbf{x}_t, \cdots, \mathbf{x}_T)$. The phone posterior probability sequence Z was then decoded using an ergodic HMM, in which each state represents a phone. Each phone in the ergodic HMM was modeled with three left-to-right HMM states.

In the case of the PhoneBook corpus, we randomly selected one acoustic example for each word in the training data. As the words in the test set do not appear in the training set, we adopted the setup used in [Aradilla et al., 2009, Soldo et al., 2012] for template-based ASR where one acoustic example per word was randomly selected from the test set. We removed those acoustic examples during ASR evaluation.[10]

In the case of the NameDat corpus, all the words in the test set are seen during training. The average number of acoustic examples for each word in training sets was four. Among the

[10] As in later studies we also use two acoustic examples for each word in the test set, we have reported all the ASR results by removing the two examples for each word.

pronunciation hypotheses generated from acoustic examples of a given word, we selected the best hypothesis based on a double-normalization posterior based confidence measure as defined in [Bernardis and Bourlard, 1998],

$$NPCM(w) = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{e_j - b_j + 1}\sum_{t=b_j}^{e_j}\log P(q_t = ph_j|\mathbf{x}_t), \tag{5.10}$$

where J denotes the number of phones in the word, b_j and e_j are the begin and end frames of the j^{th} phone ph_j in the phonetic representation of the word w, $ph_j \in \mathscr{F}$. During ASR system training, we used the single-best pronunciations obtained based on the above confidence measure, and during decoding we used all the generated pronunciation variants.[11]

During the inference phase, for the same reasons explained in Section 5.3.1, we assumed uniform prior probability and transition probability distributions. Therefore, Eqn. (5.6) can be written as,

$$Q^* = \underset{Q\in\mathscr{Q}}{\arg\max}\prod_{t=1}^{T}\underbrace{P(q_t = f^k|\mathbf{x}_t)}_{\text{Posterior probability}}\ . \tag{5.11}$$

Unification of A2P conversion and G2P conversion

Toward unification of ANN-based A2P conversion and CRF-based G2P conversion approaches, we aligned the sequences of phones posterior probability vectors Y and Z explained in Section 5.1.2, using dynamic time warping with the symmetric Kullback-Leibler divergence as the local score. The weights w^{a2p} and w^{g2p} in Eqns. (5.7) and (5.8) were not tuned and were set to be 0.5, i.e., $w^{a2p} = w^{g2p} = 0.5$.

As there are multiple acoustic examples available for a given word in the training set of the NameDat corpus, we selected the best pronunciation for each word based on the double-normalization posterior based confidence measure explained in the previous section (Eqn. (5.10)). Similarly, the ASR systems were trained by using single-best pronunciations during training, and using all the pronunciation variants during decoding. The average number of pronunciations per word was 2.7.

5.4.2 Pronunciation level evaluation

Table 5.7 presents the pronunciation level evaluation results when unifying A2P and G2P conversion approaches on the PhoneBook corpus.[12] It can be observed that the number of substitutions is significantly reduced in the multi-stream combination approach compared to

[11] We found that using single-best pronunciations during training leads to better ASR systems than using all the pronunciation variants.

[12] As explained in Section 5.3.2, pronunciation-level evaluation for the NameDat corpus was not possible as there are no canonical pronunciations available for the corpus.

the CRF-based and A2P conversion approaches. Similar to observations in Table 5.1, the main reason for lower PRR in the multi-stream combination approach compared to the CRF-based approach is the number of insertions.

Table 5.7 – Pronunciation level results on PhoneBook corpus in terms of the number of deletions (D), substitutions (S), insertions (I) and PRR for the baseline CRF-based G2P conversion approach and ANN-based A2P conversion approach together with the multi-stream combination of the two approaches.

Approach	D	S	I	PRR
CRF	78	364	56	88.5
A2P	232	868	427	64.7
A2P-G2P-Comb-Sum	69	231	272	86.8
A2P-G2P-Comb-Prod	84	213	241	87.6

5.4.3 ASR level evaluation

This section presents the ASR experimental setup and results on the PhoneBook corpus and the NameDat corpus respectively.

ASR studies on the PhoneBook corpus

Similar to the studies conducted in Section 5.3.3, we used HMM/GMM and hybrid HMM/ANN frameworks for building ASR systems. We trained the ASR systems in a similar setup explained in Section 5.3.3, using the lexicons obtained through the A2P conversion approach or the unification of G2P conversion and A2P conversion approaches.

Table 5.8 presents the ASR level results when unifying A2P conversion and G2P conversion. It can be seen that the ASR system using A2P conversion-based pronunciations performs poorly compared to the ASR system using CRF G2P conversion-based pronunciations. This could be expected, as the A2P conversion approach uses only one acoustic example per word, and the words in the test set are not seen during training. Despite the poor performance of A2P conversion approach, it can be observed that unification of the CRF-based G2P conversion approach and A2P conversion approach leads to a significantly better ASR system. The improvements can be seen throughout both HMM/GMM and hybrid HMM/ANN frameworks. This indicates that the proposed multi-stream approach can lead to improvements irrespective of the ASR framework used.

Table 5.8 – ASR level evaluations on PhoneBook corpus in terms of WRR when using individual lexicons based on A2P conversion approach and CRF-based G2P conversion approach together with the multi-stream combination of the two approaches. † denotes that the performance gain is statistically significant compared to the best performing individual G2P/A2P conversion approach.

	CRF G2P	A2P Conversion	*A2P-G2P-Comb Sum*	*A2P-G2P-Comb-Prod*
HMM/GMM	89.1	85.9	93.1†	94.3†
Hybrid HMM/ANN	92.0	88.4	95.2†	95.6†

ASR studies on the NameDat corpus

We used the KL-HMM framework for building ASR systems on the NameDat corpus, in a similar setup explained in Section 5.3.3, using the lexicons obtained through the A2P conversion approach or the unification of G2P conversion and A2P conversion approaches. Table 5.9 presents the ASR results in terms of WRR when using lexicons obtained from the CRF-based approach, A2P conversion approach and the multi-stream combination of these approaches. Compared to the PhoneBook corpus, it can be observed that the gap between the systems using the pronunciations from the CRF-based G2P conversion approach and the A2P conversion approach is reduced. This can be due to the fact that in the NameDat corpus, multiple acoustic examples for each word is used. Furthermore, all the words have been seen during training.

Table 5.9 – ASR level evaluations in terms of WRR using pronunciations obtained from the multi-stream combination of CRF-based approach and A2P conversion approach on the NameDat corpus. † denotes that the performance gain is statistically significant compared to the best performing individual G2P/A2P conversion approach.

	CRF G2P	A2P Conversion	*A2P-G2P-Comb-Sum*	*A2P-G2P-Comb-Prod*
KL-HMM-set-1	94.5	93.0	94.5	94.6
KL-HMM-set-2	94.1	93.4	95.0†	95.2†
KL-HMM-set-3	93.7	93.4	93.8	94.5
Average	94.1	93.3	94.4	94.8

The multi-stream combination of the G2P relationship estimate and A2P relationship estimate leads to improvements over the baselines, particularly when using the product rule. The superiority of product rule over sum rule when unifying G2P conversion and A2P conversion has also been observed in the PhoneBook studies (Table 5.8). The rationale for such observations lies in the underlying assumptions made to obtain the sum rule and the product rule. In the case of the sum rule, as noted in [Tax et al., 2000], the feature spaces are assumed to be identical. Therefore, the sum rule is expected to perform well in the case of having highly correlated feature spaces where the classifiers make independent errors, which is the case

when unifying G2P conversion approaches. In the case of the product rule, on the other hand, the underlying assumption is that the feature spaces are different and class-conditionally independent. Therefore the product rule is expected to perform better when the feature spaces are different. As in the case of unifying A2P conversion and G2P conversion the feature spaces (one based on acoustic information and the other based on grapheme information) are from different modalities, the product rule should be a better choice for combining phone posterior probabilities, as it is also evident from the experimental results.

5.4.4 Comparison to a pronunciation variation selection approach using spoken word examples

An alternative approach to the unification of A2P conversion approach and the G2P conversion approach would be to use the spoken word examples to select from pronunciation variants generated by a G2P conversion approach. Toward that, we used a likelihood-based approach similar to [McGraw et al., 2013, Anumanchipalli et al., 2007] to select the pronunciation variant obtained from a G2P converter that has the highest acoustic likelihood. More precisely, for the PhoneBook corpus, we used the CRF-based G2P conversion approach to generate five-best pronunciations for each word. Each hypothesis in the list of five-best pronunciations was then force-aligned to the spoken sample of the word using the HMM/GMM system trained on the manual lexicon, producing an acoustic likelihood for each pronunciation variant. The pronunciation variant leading to the highest acoustic likelihood was then selected as the best hypothesis.

Table 5.10 shows the performance of ASR systems trained through pronunciation variation selection approach and the multi-stream combination of G2P conversion and A2P conversion approaches on the PhoneBook corpus. We have presented the results in the case of using only one acoustic example per word and using two acoustic examples per word in the test set, i.e., using each of the two spoken word examples to generate or select a pronunciation for each word.

It can be observed that the multi-stream combination approach for unifying A2P conversion and G2P conversion leads to a significantly better ASR system compared to the pronunciation variant selection approach using acoustic examples. The improvements are retained across both ASR frameworks. It is also interesting to note that the performance of the ASR system using *A2P-G2P-Comb-Prod* combination approach with two pronunciations is not far from the ASR system using the manual pronunciation lexicon (98.9% WRR).

Similar to the strategy explained for the PhoneBook corpus, for the NameDat corpus, we used the CRF-based G2P conversion approach to generate five-best pronunciations per word. The pronunciation variant leading to the highest acoustic likelihood according to the HMM/GMM system trained on auditory verified lexicons was chosen as the best hypothesis. Table 5.11 compares the unification of A2P conversion approach and G2P conversion approach with pronunciation variant selection method using spoken word examples for the NameDat corpus.

Table 5.10 – The performance of ASR systems in terms of WRR on the PhoneBook corpus, when using lexicons obtained through the pronunciation variant selection approach versus the multi-stream combination of A2P conversion and G2P conversion approaches. † and ‡ denote that the performance gain against the pronunciation variant selection approach is statistically significant when using single pronunciation and two pronunciations per word respectively.

	Pronunciation variant selection		*A2P-G2P-Comb Sum*		*A2P-G2P-Comb-Prod*	
	single-pron.	two-pron.	single-pron.	two-pron.	single-pron.	two-pron.
HMM/GMM	92.7	93.9	93.1	95.4‡	94.3 †	96.6 ‡
Hybrid HMM/ANN	93.8	95.5	95.2†	96.8‡	95.6†	98.1‡

As in the NameDat corpus multiple spoken word examples for each word exists, we selected the pronunciation variant chosen by the majority of the spoken word examples. For building ASR systems, we adopted the setup used for the unification of G2P conversion and A2P conversion, where during training we used the selected pronunciation variant for each word according to the acoustic likelihood, and during decoding we used all the pronunciation variants. It can be seen that the multi-stream combination of G2P conversion and A2P conversion performs better than pronunciation variant selection approach in almost all cases.

Table 5.11 – The performance of ASR systems in terms of WRR on the NameDat corpus, when using lexicons obtained through the pronunciation variant selection approach versus the multi-stream combination of A2P conversion and G2P conversion approaches.

	Pronunciation variant selection	A2P-G2P-Comb -Sum	A2P-G2P-Comb-Prod
KL-HMM-set-1	94.1	94.5	94.6
KL-HMM-set-2	94.8	95.0	95.2
KL-HMM-set-3	93.9	93.8	94.5
Average	94.3	94.4	94.8

5.4.5 Analysis

Similar to the studies in Section 5.3.5, we computed the confusion matrix for the generated pronunciations through each of the approaches on the PhoneBook corpus, which has canonical pronunciations.[13] Figure 5.6 illustrates the percentage correctly labeled for a few example phones in the case of employing the multi-stream combination of G2P conversion approach and A2P conversion approach. It can be seen that the CRF-based G2P conversion approach is the best individual model in most of the cases. However, in some cases, the A2P conversion approach performs better, despite its overall poor PRR. In spite of that, the proposed unification approach is able to perform better than the best G2P conversion or A2P conversion approach.

[13]The average number of pronunciations per word is one.

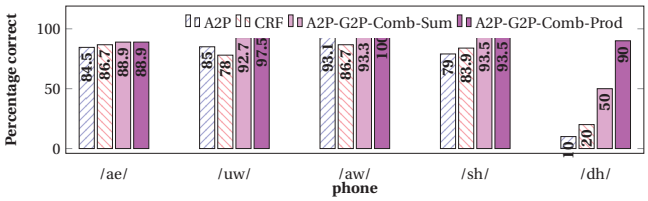


Figure 5.6 – Percentage correct for a selected few phones according to the confusion matrix for individual G2P and A2P conversion approaches together with the multi-stream combination of the approaches on the PhoneBook corpus.

Table 5.12 presents a few example pronunciations inferred using unification of G2P conversion and A2P conversion on the PhoneBook corpus. It can be observed that the proposed unification approach enables leveraging from the individual G2P conversion and A2P conversion approaches to infer a better pronunciation. For example, for the word *WOUNDS*, the CRF-based approach is able to correctly predict the /d/ sound, the A2P conversion approach is able to correctly predict the /uw/ sounds, and the unification approach is able to exploit the strengths of each approach to correctly predict the whole pronunciation.

Table 5.12 – Pronunciations generated by individual G2P and A2P conversion approaches along with the multi-stream combination of the approaches on the PhoneBook corpus.

Pronunciation	FRIDAYS	BEIRUT	WOUNDS
CRF-based	f r ih d ey z	b iy r ah t	w aw n d z
A2P-based	f r ay g ey z	b ey uw r uw t	w uw n z
Combination	f r ay d ey iy z	b ey r uw t	w uw n d z
Manual	f r ay d ey z	b ey r uw t	w uw n d z

As it can be seen from Table 5.7, the lower PRR in the multi-stream combination approach, despite significantly reducing the phone substitutions, is mainly due to the phone insertions. Similar to the analysis studies in Section 5.3.5, we looked into the type of insertions and found similar trends, i.e., we found that several of the insertions were due to systematic insertion of acoustically close phones, such as /axr/ → /axr/ /r/ or /ng/ → /ng/ /g/. Therefore, in this case, the improvements at the ASR level obtained through unification of G2P conversion and A2P conversion approaches despite the relatively poor performance at the pronunciation level can possibly be attributed to (a) reduction in the phone substitutions, and (b) the ability of the ASR system to handle the systematic insertion errors present at the output of multi-stream pronunciation generation process.

5.5 Summary

G2P conversion can be achieved using different techniques. These techniques primarily differ in the manner the G2P relationship is learned and in the sequential modeling approach em-

ployed. The central premise of this chapter was that we can exploit various G2P relationship modeling techniques in order to estimate complementary multiple streams of $P(f|g)$. These streams can then be combined, in a manner analogous to multi-stream speech recognition approach, to improve G2P conversion. We validated the proposed approach by investigating the combination of $P(f|g)$ estimates obtained from the CRF-based approach and acoustic data-driven G2P conversion approach. We further showed how the multi-stream combination approach can be extended for the unification of A2P conversion and G2P conversion approaches, when the acoustic example of a given word is available.

Our studies on PhoneBook and NameDat, as two challenging corpora for G2P conversion, showed that the unification approaches lead to development of lexicons that can yield better ASR systems, compared to the lexicons obtained from individual G2P/A2P conversion approaches.

6 Acoustic subword unit discovery and lexicon development

This chapter addresses the challenge of pronunciation lexicon development for under-resourced languages that lack phonetic lexical resources. In the absence of a phonetic lexicon, alternatively grapheme subword units based on writing system have been explored in the literature [Kanthak and Ney, 2002, Killer et al., 2003, Dines and Magimai.-Doss, 2007, Magimai.-Doss et al., 2011a, Ko and Mak, 2014, Rasipuram and Magimai.-Doss, 2015, Gales et al., 2015] (Section 6.1.1). However, as discussed earlier, the success of grapheme-based ASR systems commonly depends on the G2P relationship of the language. Another way to handle a lack of a phonetic lexicon is to derive subword units automatically from the speech signal and build the associated lexicon. In the literature, interest in acoustic subword unit (ASWU)-based lexicon development emerged from the pronunciation variation modeling perspective, specifically with the idea of overcoming the limitations of linguistically motivated subword units, i.e., phones [Lee et al., 1988, Svendsen et al., 1989, Paliwal, 1990, Lee et al., 1988, Bacchiani and Ostendorf, 1998, Holter and Svendsen, 1997]. However, recently, there has been a renewed interest from the perspective of handling lexical resource constraints [Singh et al., 2000, Lee et al., 2013, Hartmann et al., 2013] (Section 6.1.2). A limitation of most of the existing methods for acoustic subword unit-based lexicon development is that they are not able to handle unseen words.

In this chapter, we propose an approach for ASWU-based lexicon development where the ASWU derivation is cast as a problem of determining a latent symbol space given the acoustic data and its word level transcription (Section 6.2). In this approach, first a set of ASWUs is derived by modeling the relationship between the graphemes and the acoustic speech signal in an HMM framework based on two assumptions,

1. writing systems carry information regarding the spoken system. Alternately, a written text embeds information about how it should be spoken. Though this embedding can be deep or shallow depending on the language; and

2. the envelope of the short-term spectrum tends to carry information related to phones.

Given the derived ASWUs, a graphemes-to-ASWU (G2ASWU) relationship is learned through the acoustic signal, and finally a lexicon is developed by G2ASWU conversion analogous to the acoustic G2P conversion approach, explained in Chapter 4.

In this chapter, we first establish the proposed framework on a well-resourced language by comparing it against related approaches in the literature and investigating the transferability of the derived subword units to other domains (Section 6.3). We then show the scalability of the proposed approach on real under-resourced scenarios by conducting studies on Scottish Gaelic, a genuinely under-resourced language (Section 6.4). Finally, we provide a mechanism to relate the ASWUs to the phonetic identities (Section 6.5).

It is worth mentioning that the ASWU-based lexicon development approach was originally published in [Razavi and Magimai.-Doss, 2015] and further studied in [Razavi et al., 2015b].

6.1 Relative literature

In this section, we first briefly explain the grapheme-based ASR approach as an alternative approach in the absence of a phonetic lexicon for a language. We then present a survey on the existing approaches for derivation of ASWUs and lexicon development.

6.1.1 Grapheme-based ASR

In the literature, the issue of lack of well developed phonetic lexicons has been addressed by using graphemes as subword units. Most of the studies in this direction have been conducted in the framework of deterministic lexical modeling, where $\{l^i\}_{i=1}^{I}$ model context-dependent graphemes, $\{a^d\}_{d=1}^{D}$ are clustered context-dependent grapheme units and $\mathbf{y}^i$ is a decision tree learned while state tying based on either singleton question set or phonetic question set [Kanthak and Ney, 2002, Killer et al., 2003]. In [Gales et al., 2015], the question set was based on the attributes derived from the information available in the unicode character description. It was shown that such an approach yields an ASR system that can perform comparable to the phone-based ASR system.

In the framework of probabilistic lexical modeling, it has been shown that grapheme-based ASR systems can be built with $\{a^d\}_{d=1}^{D}$ based on phones of auxiliary languages or domains, and $\{l^i\}_{i=1}^{I}$ based on target language graphemes. More precisely, a phone class conditional probability $\mathbf{z}_t$ estimator is trained with acoustic and lexical resources from auxiliary languages or domains, and $\mathbf{y}^i$, which captures a probabilistic G2P relationship, is trained on the target language or domain acoustic data [Magimai.-Doss et al., 2011b, Rasipuram and Magimai.-Doss, 2015]. It has been shown that this approach can effectively address both acoustic resource and lexical resource constraints [Rasipuram and Magimai.-Doss, 2015, Rasipuram et al., 2013b].

6.1.2 Literature survey on ASWU derivation and pronunciation generation

The idea of using lexicons based on ASWUs instead of linguistically motivated units has been appealing to the ASR community for three main reasons: (1) ASWUs tend to rather be data-dependent than linguistic knowledge-dependent, as they are typically obtained through optimization of an objective function using training speech data [Lee et al., 1988, Bacchiani and Ostendorf, 1998], (2) they could possibly help in handling pronunciation variations [Livescu et al., 2012], and (3) they can avoid the need for explicit phonetic knowledge [Lee et al., 2013].

Typically, the ASWU-based lexicon development process, in addition to the speech signal, requires the corresponding transcription in terms of words. Alternately, the lexicon development process is weakly-supervised similar to acoustic model development in an ASR system. More recently, in the context of “zero-resourced" ASR system development, there are efforts toward developing methods that are fully unsupervised [Chung et al., 2013, Lee et al., 2015]. Such methods are at very early stages and are out of the scope of this chapter. In the reminder of this section, we provide a brief literature survey on weakly-supervised ASWU-based lexicon development. ASWU-based lexicon development involves two key challenges: (a) derivation of ASWUs, and (b) pronunciation generation based on the derived ASWUs. The approaches proposed in the literature can be grouped into two categories based on how these two challenges are addressed. More precisely, there are approaches that decouple these two challenges and address them separately (Section 6.1.2), and there are approaches that address these two challenges in an unified manner with a common objective function (Section 6.1.2).

Automatic subword unit discovery followed by pronunciation generation approaches

The very first efforts approached the ASWU derivation problem as a segmentation of *isolated word* speech signals into acoustic segments and clustering acoustic segments into groups each representing a subword unit [Lee et al., 1988, Svendsen et al., 1989, Paliwal, 1990]. More precisely, as shown in Figure 6.1, in the segmentation step, the speech utterance $X = (\mathbf{x}_1, \cdots, \mathbf{x}_t, \cdots, \mathbf{x}_T)$ is partitioned into I consecutive segments (with boundaries $B = (b_1, \cdots, b_i, \cdots, b_I)$) such that the frames in a segment are acoustically similar. Then in the clustering step, the acoustic segments are clustered into groups of subword units.

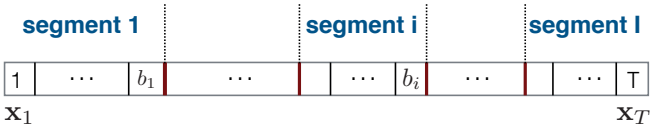


Figure 6.1 – Segmentation of speech utterance **x** into I segments.

In [Lee et al., 1988, Svendsen et al., 1989], the segmentation step was approached by applying dynamic programming techniques and finding the segment boundaries b_i such that the likelihood ratio distortion between the speech frames in segment i and the generalized spectral centroid of segment i (i.e., the centroid LPC vector) is minimized. The obtained acoustic

segments were then clustered using the K-means algorithm in which each acoustic segment was represented by its centroid. Once a *pre-set* number of subword units was determined, a set of pronunciations for each word was found from its occurrences in the training data and were clustered to select representative pronunciations [Paliwal, 1990, Svendsen et al., 1995]. The studies on an isolated word recognition task on English demonstrated the potential of the approach. A limitation of these approaches is that they can generate pronunciations only for the words which are seen during training. Furthermore, these approaches need to know the word boundaries explicitly.

In [Jansen and Church, 2011], an approach was proposed in which the need for transcribed speech is limited. Specifically, given an acoustic example of each word, a spoken term discovery algorithm [Park and Glass, 2008] is exploited to search and cluster the acoustic realizations of the words from untranscribed speech. Then for each word cluster, a whole word HMM is trained in which each HMM state represents a subword unit. The number of subword units for each word is determined based on the duration of acoustic examples and the expected duration of a phone. The subword unit states are then finally clustered based on the pairwise similarities between their emission scores using a spectral clustering algorithm [Shi and Malik, 2000]. The viability of the approach was limited to a spoken term detection task. A limitation of the approach is that an acoustic example of each word in the dictionary is required.

Hartmann et al. [2013] proposed an approach based on the assumption that the orthography of the words and their pronunciations are related. In this approach, the subword units are obtained by clustering CD grapheme models. This is achieved through a spectral-based clustering approach [Ng et al., 2001], similar to [Jansen and Church, 2011]. The main difference is that in this case the pairwise similarities are computed between the CD grapheme models (instead of the HMM states). The pronunciations for seen and unseen words are finally generated by employing a statistical machine translation (SMT) framework. On the Wall Street Journal task, it was found that the resulting ASWU-based lexicon yields a better ASR system than the grapheme-based lexicon.

Joint approaches for ASWU derivation and pronunciation generation

As opposed to decoupling the ASWU derivation and pronunciation generation problems, there are also approaches that aim to jointly determine the subword units and pronunciations using a common objective function. In [Holter and Svendsen, 1997], this was done through an iterative process of acoustic model estimation and pronunciation generation. In [Bacchiani and Ostendorf, 1999, 1998], a segmentation and clustering approach was exploited for derivation of subword units, with two main differences compared to the approaches explained in Section 6.1.2: (1) in the segmentation step, pronunciation related constraints are applied such that a given word has the same number of segments across the acoustic training data, and (2) a maximum-likelihood criteria that is consistent for both segmentation and clustering is utilized. On the RM task, it was shown that the proposed approach leads to improvements over a phone-based ASR system.

In [Singh et al., 2000, 2002], a maximum likelihood strategy was presented that decomposed the ASWU-based ASR system development as the joint estimation of the pronunciation lexicon (including determination of ASWU set size) and acoustic model parameters. More precisely, with an initial pronunciation lexicon based on CI graphemes, the acoustic model parameters and the pronunciation lexicon are updated iteratively. The lexicon update step is an iterative process within itself consisting of word segmentation estimation given the acoustic model and update of the lexicon based on the segmentation. After each iteration of lexicon update and acoustic model update, convergence is determined by evaluating the ASR system on cross-validation data. If not converged, the ASWU set size is increased and the process is repeated. A proof of concept was demonstrated on the RM corpus.

Recently, in [Lee et al., 2013] a hierarchical Bayesian model approach was proposed to jointly learn the subword units and pronunciations. This is done by modeling two latent structures: (1) the latent phone sequence, and (2) the latent L2S mapping rules, using an HMM-based mixture model in which each component represents a phone unit and the weights over HMMs are indicative of the L2S mappings. It was shown that the proposed approach together with the pronunciation mixture model retraining leads to improvements over the grapheme-based ASR system on a weather query task.

6.2 Proposed approach

This section presents an HMM-based formulation to derive phone-like ASWUs and develop an associated pronunciation lexicon. Essentially, the formulation builds on grapheme-based ASR in a deterministic lexical modeling framework as well as a probabilistic lexical modeling framework. More specifically, we show that,

1. the problem of derivation of ASWUs can be cast as a problem of finding phone-like acoustic units $\{a^d\}_{d=1}^{D}$ given transcribed speech, i.e., speech signal and orthographic transcription, in the grapheme-based ASR framework. Section 6.2.1 dwells on this aspect; and

2. given the derived ASWUs $\{a^d\}_{d=1}^{D}$ and the transcribed speech, the pronunciation lexicon development problem can be cast as a problem akin to acoustic data-driven G2P conversion explained in Chapter 4. Section 6.2.2 deals with this aspect.

6.2.1 Automatic subword unit derivation

State clustering and tying methods in the HMM-based ASR have emerged from the perspective of addressing the data sparsity issue and handling unseen contexts [Young, 1992, Ljolje, 1994]. However, this methodology can be adopted, as it is, to derive acoustic subword units in the framework of grapheme-based ASR. More precisely, we hypothesize and show that the clustered CD grapheme units $\{a^d\}_{d=1}^{D}$ obtained in a CD grapheme-based ASR system can serve

as phone-like subword units.

The reasoning behind our hypothesis is as follows. Recall from Section 3.1 that in statistical ASR, the most probable sequence of words W^* given the acoustic observation sequence $X = (\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T)$ is obtained by finding the most probable sequence of states Q^* representing W^*:

$$W^* = \arg\max_{Q \in \mathcal{Q}} \sum_{t=1}^{T} \{\log p(\mathbf{x}_t | q_t = l^i) + \log P(q_t = l^i | q_{t-1} = l^j)\}. \tag{6.1}$$

Estimation of $p(\mathbf{x}_t | q_t = l^i)$ is typically factored through acoustic units $\{a^d\}_{d=1}^D$ as:

$$p(\mathbf{x}_t | q_t = l^i) = \sum_{d=1}^{D} p(\mathbf{x}_t, a^d | q_t = l^i), \tag{6.2}$$

$$= \sum_{d=1}^{D} p(\mathbf{x}_t | a^d) \cdot P(a^d | q_t = l^i) (\text{assuming } \mathbf{x}_t \perp\!\!\!\perp q_t | a^d), \tag{6.3}$$

$$= \mathbf{v}_t^{\mathrm{T}} \mathbf{y}^i. \tag{6.4}$$

The acoustic units $\{a^d\}_{d=1}^D$ are obtained by maximizing the likelihood of the training data, which is essentially determined by estimation of $p(\mathbf{x}_t | q_t = l^i)$, as during training the sequence model for each utterance is fixed given the associated transcription and lexicon. As observed earlier in Eqn. (6.4), $p(\mathbf{x}_t | q_t = l^i)$ estimation involves the matching of acoustic information $\mathbf{v}_t$ with lexical information $\mathbf{y}^i$. We know that standard features such as cepstral features have been designed to model the envelope of the short-term spectrum, which carry information related to phones. In other words, standard features such as MFCCs or PLPs for ASR primarily target modeling the spectral characteristics of the vocal tract system while incorporating speech perception knowledge.

Similarly it is very well known that CD graphemes capture information related to phones. This is one of the central assumptions in most of G2P conversion approaches, i.e., the relationship between CI graphemes and phones can be irregular but the relationship can become regular when contextual graphemes are considered. For example, as illustrated in Figure 6.2, in the decision tree-based G2P conversion approach [Pagel et al., 1998], given the grapheme context a decision tree is learned to map the central grapheme to a phone.

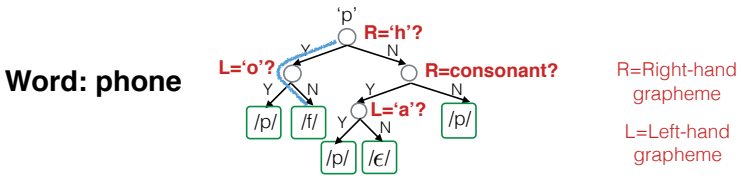


Figure 6.2 – Example of the decision tree-based G2P conversion.

Therefore, as illustrated in Figure 6.3, for the likelihood of the training data to be maximized, clustered CD grapheme units $\{a^d\}_{d=1}^D$ should model an information space that is common to

both the short-term spectrum-based feature $\mathbf{x}_t$ space and the CD grapheme-based lexical unit l^i space, which we hypothesize to be a phone-like subword unit space.

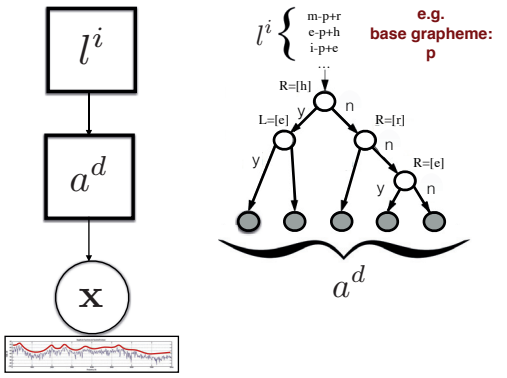


Figure 6.3 – The clustered states a^d of a grapheme-based CD HMM/GMM system obtained through decision tree-based clustering are exploited as ASWUs. As a^d should be related to both CD graphemes l^i and cepstral features $\mathbf{x}$, they are expected to be phone-like.

Our argument is further supported by an ASR study that demonstrated the interchangeability of clustered CD phone units space and clustered CD grapheme units space in the framework of probabilistic lexical modeling [Rasipuram and Magimai.-Doss, 2013a], as well as by earlier works on grapheme-based ASR that have explored integration of phonetic information in clustering CD grapheme units and state tying [Killer et al., 2003].

6.2.2 Lexicon development through grapheme-to-ASWU conversion

In order to build speech technologies with the derived ASWUs, we need a mechanism to map the orthographic transcription of words to sequences of ASWUs for both seen and unseen words. For that purpose, an approach similar to G2P conversion is desirable. However, conventional G2P approaches are not directly applicable, as they necessitate a seed lexicon, which maps a few word orthographies into sequence of phones (in our case ASWUs). As shown in Chapter 4, G2P conversion can be achieved by learning the G2P relationship through acoustics using HMMs. Such an approach has the inherent ability to alleviate the necessity for a seed lexicon, and thus can be exploited to develop a G2ASWU converter for lexicon development. This approach can be essentially considered as an extension of the grapheme-based ASR approach, where either a deterministic lexical model or a probabilistic lexical model $\{\mathbf{y}^i\}_{i=1}^I$ that captures G2ASWU relationship is learned and ASWU-based pronunciations are inferred. We present below these two frameworks.

Deterministic lexical modeling-based G2ASWU conversion

This method of lexicon development is a straightforward extension of the ASWU derivation. More precisely, in the process of ASWU derivation a deterministic one-to-one map between CD graphemes ($\{l^i\}_{i=1}^{I}$) and ASWUs ($\{a^d\}_{d=1}^{D}$) is learned. The pronunciations can be inferred using this information similar to the decision tree based G2P conversion approach [Pagel et al., 1998], discussed briefly earlier in Section 6.2.1 (Figure 6.2), and in Section 4.2.4.

Probabilistic lexical modeling-based G2ASWU conversion

Another possibility is to learn a probabilistic relationship between graphemes and ASWUs and infer pronunciations in terms of ASWUs following the acoustic data-driven G2P conversion approach using KL-HMM explained in Chapter 4. More precisely, this approach of G2ASWU conversion would involve,

1. training of an ANN-based $\mathbf{z}_t$ estimator given the alignment of the training data in terms of $\{a^d\}_{d=1}^{D}$. This step is the same as training a CD neural network for an ASR system;[1] then
2. training of a CD grapheme-based KL-HMM using $\mathbf{z}_t$ as feature observations [Magimai.-Doss et al., 2011a]; and finally
3. inferring the pronunciations given the KL-HMM parameters $\{\mathbf{y}^i\}_{i=1}^{I}$ and the orthographies of the words in the lexicon. More precisely, first a sequence of ASWU posterior probability vectors is obtained from the KL-HMM given the orthography of the target word. The sequence is then decoded by an ergodic HMM in which each state represents an ASWU to infer the pronunciation.

6.2.3 Summary of the proposed approach

Figure 6.4 summarizes our approach. As illustrated, the approach consists of three phases. *Phase I* involves derivation of ASWUs. *Phase II* involves learning G2ASWU relationship given transcription and acoustic data. *Phase III* deals with lexicon development given the G2ASWU relationship and the word orthographies. *Phase II* is explicitly needed for learning the probabilistic G2ASWU relationship. In the case of deterministic G2ASWU conversion, it is implicit in *Phase I*. *Phase III* can be seen as decoding a sequence of ASWU posterior probability vectors $\mathbf{y}^i$. It is worth mentioning that the pronunciation inference step, i.e. *Phase III*, for both deterministic and probabilistic lexical modeling-based approaches is the same. More precisely, in the case of deterministic lexical modeling-based approach, the inference step is equivalent to decoding a sequence of Kronecker delta distributions resulting from the one-to-one mapping of CD graphemes (in the word orthography) to ASWU units using the decision tree, as

[1] If the $\mathbf{z}_t$ estimator is based on Gaussians then it would amount to going from single Gaussian to GMMs (mixture increment step) of ASR system training.

explained in Section 4.2.4.

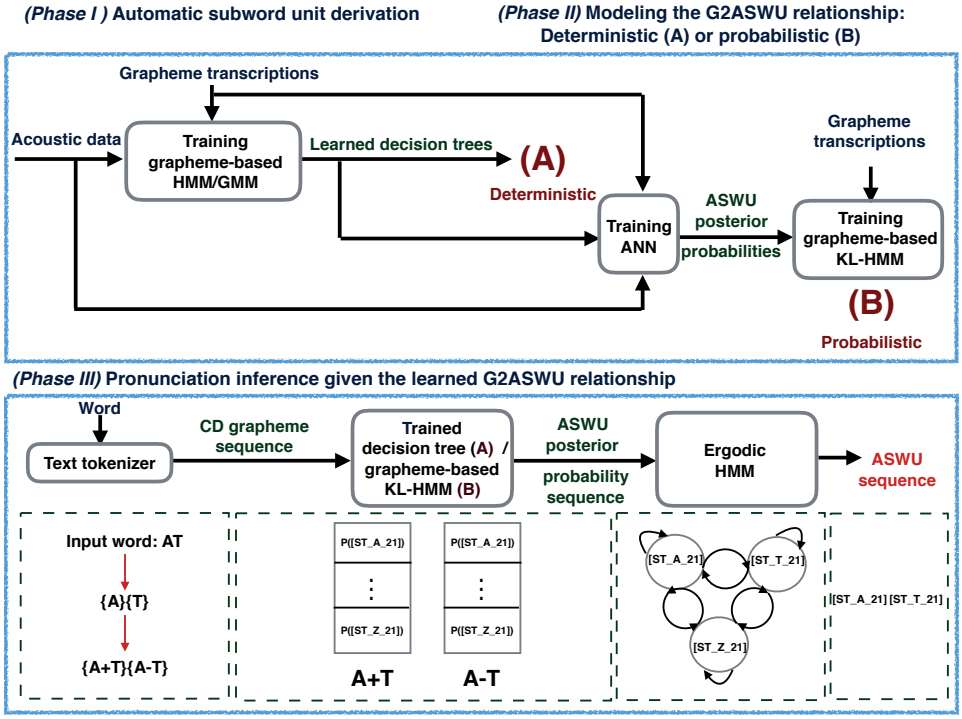


Figure 6.4 – Block diagram of the HMM formalism for subword unit derivation and pronunciation generation. *Phase III* is shown for the case where only a single posterior probability vector for each CD grapheme is generated.

A central challenge in the proposed approach is how to determine the cardinality of the ASWU set $\{a^d\}_{d=1}^{D}$. In the studies validating the proposed approach, presented in the remainder of the paper, we show that this can be achieved via cross-validation. Specifically, a range of values for acoustic units set cardinality D can be considered based on the knowledge that the ratio of number of phones to number of graphemes is not an extremely high value, and can be selected via cross-validation at the ASR level. For instance in English, if one considers the CMU dictionary, then the ratio is $\frac{38}{26}$ or $\frac{84}{26}$ (when lexical stress is considered). Alternately, D can be chosen relative to the number of graphemes, and is much lower than the number of acoustic units considered for building CD grapheme-based ASR systems, which is typically in the order of thousands.

6.3 In-domain and cross-domain studies on well-resourced languages

In this section, we establish the proposed framework for subword unit derivation and lexicon development through experimental studies on a well-resourced language using only its word-level transcribed speech data. The rationale for studying on a well-resourced language is to

enable analyzing the discovered subword units and relating them to phonetic identities. We selected English as the well-resourced language, as it is a challenging language for automatic pronunciation generation due to its irregular G2P relationship, and has been the focus of many previous works on ASWU derivation and lexicon development. Our investigations are organized as explained below.

1. *Evaluation of the proposed approach through in-domain studies*: We investigate the proposed approach for derivation of ASWUs and corresponding pronunciations on two English corpora, namely WSJ0 and RM (explained in Sections 2.5.4 and 2.5.5). We evaluate the ASWU-based lexicons through in-domain ASR studies where the performance of the ASWU-based ASR systems is compared against grapheme-based and phone-based ASR systems (Section 6.3.1).

2. *Investigating the transferability of the ASWUs through cross-domain studies*: A central challenge in ASWU-based lexicon development and its adoption for wider use is ascertaining whether the ASWUs derived from limited amount of acoustic resources generalize across domains, similar to linguistically motivated subword units, i.e. phones and graphemes. To the best of our knowledge, none of the previous works have tried to ascertain that aspect. In that sense, we go a step further to conduct cross-domain studies where the ASWUs are derived from the WSJ0 corpus and the lexicon is developed for the RM corpus. We present three methods for development of lexicons in such a scenario, and investigate the transferability of the ASWUs by building and evaluating ASR systems using the developed lexicons (Section 6.3.2).

3. *Comparison to related approaches in the literature*: in Section 6.1.2, we discussed a few prominent approaches proposed in the literature for derivation of ASWUs and pronunciation generation. We compare the performance of the our approach with two of the related approaches in the literature studied on WSJ0 and RM corpora (Section 6.3.3). Indeed, one of the main reasons for selecting these two corpora is to enable comparing to these related works in the literature.

6.3.1 In-domain ASR studies

In this section we first explain the setup for derivation of ASWUs and development of ASWU-based lexicons. We then present the in-domain ASR studies for evaluation of the ASWU-based lexicons.

ASWU derivation and lexicon development setup

The setup for subword unit derivation and lexicon development through G2ASWU conversion is described below.

Acoustic subword unit derivation: Toward automatic discovery of subword units, cross-word single preceding and single following CD grapheme-based HMM/GMM systems were trained with 39 dimensional PLP cepstral features extracted using HTK toolkit [Young et al., 2006]. Each CD grapheme was modeled with a single HMM state. The subword units were derived through likelihood-based decision tree clustering using singleton questions. Different number of ASWUs were obtained by adjusting the log-likelihood increase during decision tree-based state tying. The numbers of clustered units were obtained such that they are within the range of two to four times the number of graphemes, based on the general idea explained in Section 6.2.3. Therefore, for the WSJ0 corpus, ASWUs of size 60, 78 and 90 were investigated, and for the RM corpus, ASWUs of size 79, 92 and 109 were studied.

Deterministic lexical modeling-based G2ASWU conversion: Given the learned decision trees for each ASWU set, the pronunciation for each word was inferred by mapping each grapheme in the word orthography to an ASWU by considering its neighboring (i.e., single preceding and single following) grapheme context. We denote the lexicons in the form of Lex-*DB*-Det-ASWU-*M* where *DB* and *M* correspond to the database and the number of ASWUs respectively. For example, the lexicon generated on WSJ0 corpus using 78 ASWUs is denoted as Lex-*WSJ*-Det-ASWU-78.

Probabilistic lexical modeling-based G2ASWU conversion: In this case, given the obtained ASWUs,

1. first a five-layer MLP was trained to classify the ASWUs. The input to the MLP was 39-dimensional PLP cepstral features with four preceding and four following frame context. The hyper parameters such as the number of hidden units per hidden layer were decided based on the frame accuracy on the development set. Each hidden layer had 2000 and 1000 hidden units in the WSJ0 and RM corpora respectively. The MLP was trained with output non-linearity of softmax and minimum cross-entropy error criterion using Quicknet software [Johnson et al., 2004];

2. then using the posterior probabilities of ASWUs as feature observations, a grapheme-based KL-HMM system modeling single preceding and single following grapheme context was trained. Each CD grapheme was modeled with three HMM states. The parameters of the KL-HMM were estimated by minimizing a cost function based on the S_{RKL} local score [Aradilla, 2008], i.e., the MLP output distribution is the reference distribution, as previous studies had shown that training KL-HMM with S_{RKL} local score enables capturing one-to-many G2P relationships. Unseen grapheme contexts were handled by applying the KL-divergence-based decision tree state tying method proposed in [Imseng et al., 2012a];

3. finally, given the orthography of the word and the KL-HMM parameters, the pronunciations were inferred by using an ergodic HMM in which each ASWU was modeled with three left-to-right HMM states.

During pronunciation inference, some of the ASWUs with less probable G2ASWU relationships were automatically pruned or filtered out. This can be observed from Table 6.1, which shows the properties of the ASWU-based lexicons together with the MLPs used for the WSJ0 and RM corpora respectively. The MLPs are denoted as MLP-*DB*-*N*, with *DB* and *N* denoting the database and the size of the ASWU set respectively. Similarly, the lexicons are shown as Lex-*DB*-Prob-ASWU-*M*, with *M* denoting the actual number of ASWUs used in the lexicon. As an example, it can be seen that in Lex-*RM*-Prob-ASWU-101, from the original ASWU set of cardinality 109, only 101 remained after G2ASWU conversion.

Table 6.1 – Summary of the ASWU-based lexicons obtained through probabilistic lexical modeling-based G2ASWU conversion for WSJ0 and RM corpora.

(a) WSJ0 corpus

Lexicon	MLP
Lex-*WSJ*-Prob-ASWU-58	MLP-*WSJ*-60
Lex-*WSJ*-Prob-ASWU-74	MLP-*WSJ*-78
Lex-*WSJ*-Prob-ASWU-88	MLP-*WSJ*-90

(b) RM corpus

Lexicon	MLP
Lex-*RM*-Prob-ASWU-77	MLP-*RM*-79
Lex-*RM*-Prob-ASWU-90	MLP-*RM*-92
Lex-*RM*-Prob-ASWU-101	MLP-*RM*-109

Selection of optimal ASWU-based lexicon

Given different lexicons obtained through deterministic and probabilistic G2ASWU conversion, the optimal lexicon was determined based on the ASR WRR on the development set. More precisely, first HMM/GMM systems using different ASWU-based lexicons were trained with 39 dimensional PLP cepstral features. Then, the ASWU-based lexicon that led to the best performing HMM/GMM ASR system on the development set was selected.[2] The difference in the performance of ASR systems using different numbers of ASWUs was marginal (it was not statistically significant). In our experiments, in case of using the deterministic G2ASWU conversion for pronunciation generation, Lex-Det-*WSJ*-ASWU-90 and Lex-Det-*RM*-ASWU-92; and in case of using the probabilistic approach, Lex-Prob-*WSJ*-ASWU-88 and Lex-Prob-*RM*-ASWU-90 were selected as the optimal lexicons and are therefore used in the rest of the chapter. Table 6.2 presents the number of ASWUs per grapheme in the WSJ0 corpus and the RM corpus when using the ASWU sets with the cardinality of 90 and 92 respectively. It can be observed that the number of ASWUs per vowel grapheme is generally more than the number of ASWUs per consonant grapheme.

[2] It is worth mentioning that for WSJ0 and RM corpora there are no explicit development sets defined. To be more precise, in the case of RM, the development set (1110 utterances) was merged with the training set (2880) to create training set of 3990 utterances in literature. So, we used the part of the data that was used for early stopping through cross validation in MLP training as the development data, and trained ASWU-based HMM/GMM systems on the remaining part of the training data. For instance, in the case of RM three HMM/GMM systems corresponding to the lexicons *Lex-RM-Prob-ASWU-77*, *Lex-RM-Prob-ASWU-90*, and *Lex-RM-Prob-ASWU-101* were trained on 2880 utterances and the best lexicon was selected using the 1110 utterances. We followed a similar procedure for WSJ0.

Table 6.2 – The number of ASWUs per grapheme in the WSJ0 corpus and the RM corpus when using the ASWU set with the cardinality 90 and 92 respectively.

(a) WSJ0 corpus

Central grapheme	# of ASWUs	Central grapheme	# of ASWUs
A	8	N	4
B	1	O	9
C	3	P	1
D	3	Q	1
E	9	R	6
F	2	S	5
G	1	T	5
H	3	U	4
I	7	V	1
J	1	W	1
K	1	X	1
L	4	Y	3
M	4	Z	1

(b) RM corpus

Central grapheme	# of ASWUs	Central grapheme	# of ASWUs
A	12	N	2
B	1	O	6
C	3	P	2
D	2	Q	1
E	11	R	6
F	1	S	5
G	1	T	5
H	4	U	4
I	7	V	2
J	1	W	2
K	1	X	1
L	4	Y	2
M	1	Z	1

Evaluation

To evaluate the generated ASWU-based lexicons, we compared the performance of ASWU-based ASR systems with the grapheme-based and phone-based ASR systems. Toward that, we trained both CI and cross-word CD HMM/GMM systems with 39 dimensional PLP cepstral features. Each subword unit was modeled with three HMM states. For the CI grapheme-based systems, the number of Gaussian mixtures for each HMM state was decided based on the ASR WRR on the cross-validation set, resulting in 256 and 128 Gaussian mixtures for WSJ0 and RM corpora respectively. In case of using ASWUs, in order to have a comparable number of parameters to the grapheme-based ASR system, each HMM state was modeled with 64 and 32 Gaussian mixtures in the WSJ0 and RM corpora respectively. Similarly, for phone subword units, the number of Gaussian mixtures for each HMM state was 128 and 64 in the WSJ0 and RM corpora. In the CD case, for tying the HMM states, only singleton questions were used. Each tied state was modeled by a mixture of 16 and 8 Gaussians on WSJ0 and RM corpora respectively. The number of tied states in all the systems trained on a corpus was roughly the same to ensure that possible improvements in ASR WRR are not due to the increase in complexity.

Table 6.3 presents the performance of ASR systems based on different lexicons. We refer to the grapheme-based lexicons on WSJ0 and RM corpora as *Lex-WSJ-Gr*-26 and *Lex-RM-Gr*-29 respectively. Similarly, the phone-based lexicons on WSJ0 and RM corpora are referred to as *Lex-WSJ-Ph*-46 and *Lex-RM-Ph*-42 respectively. In the case of using CI units, the ASWU-based ASR systems perform significantly better than the grapheme-based ASR systems in both WSJ0 and RM corpora. In the case of CD units, it can be seen that for the WSJ0 corpus, the HMM/GMM system using ASWUs performs significantly better than the baseline grapheme-based ASR system. For the case of RM corpus, however, the improvements are not statistically significant. This could be due to the fact that in the RM task almost all the words are seen

during both training and evaluation. In all cases, the ASWU-based lexicon yields a system that lies between phone-based ASR system and grapheme-based ASR system.

When using CI subword units, it can be seen that the performance of the system using probabilistic lexical modeling-based G2ASWU conversion is comparable or even better than the system using deterministic lexical modeling G2ASWU conversion, whereas when using CD subword units, this is not the case. A plausible reasoning for such a trend is that CI subword unit-based systems using deterministic lexical modeling-based G2ASWU conversion may require more parameters. We tested that by building CI ASWU-based ASR systems using deterministic and probabilistic lexical modeling-based pronunciations with varying number of Gaussian mixtures (from 8 to 256). We observed that the difference between the best performing CI ASR systems using deterministic and lexical modeling-based G2ASWU conversion is not statistically significant,[3]thus indicating that the deterministic lexical modeling-based G2ASWU conversion approach leads to a better ASR system compared to the probabilistic approach. A potential explanation for this difference could be that unlike the probabilistic lexical modeling-based G2ASWU conversion approach, deterministic lexical modeling-based G2ASWU conversion approach avoids ASWU deletions and could therefore generate a more consistent pronunciation lexicon for English.

Table 6.3 – HMM/GMM ASR system performances in terms of WRR using CI and CD subword units. The number of tied states in all the systems trained on a corpus was roughly the same to ensure that possible improvements in the ASR WRR are not due to the increase in complexity. In the cases where increasing the number of parameters has led to improvement in the performance of the system, we have presented the results within the brackets.

(a) WSJ0 corpus.

Lexicon	CI	CD
Lex-*WSJ*-Gr-26	68.9	85.8
Lex-*WSJ*-Det-ASWU-90	78.6 [80.1]	88.7 [89.1]
Lex-*WSJ*-Prob-ASWU-88	78.7 [79.7]	87.3 [87.9]
Lex-*WSJ*-Ph-46	88.6	93.5

(b) RM corpus.

Lexicon	CI	CD
Lex-*RM*-Gr-29	84.2	94.0
Lex-*RM*-Det-ASWU-92	89.1 [90.2]	94.5
Lex-*RM*-Prob-ASWU-90	90.7	94.2
Lex-*RM*-Ph-42	93.5	95.9

6.3.2 Cross-domain ASR studies

This section presents a study that investigates the transferability of the ASWUs to a condition or domain unobserved during derivation of ASWUs. As noted earlier, for ASWUs to be adopted for mainstream speech technology, this characteristic is highly desirable. Toward that we present a cross-database study where the ASWU derivation is carried out on out-of-domain (OOD) WSJ0 corpus and the lexicon is developed for the target domain RM corpus. Similar to the G2P conversion as elucidated in Section 2.3.2, G2ASWU conversion (presented earlier in Section 6.2.2) can seen as a two step process: (1) learning the relationship between the

[3]The results for the best performing ASR systems are shown within the brackets in Table 6.3.

graphemes and the derived ASWUs, and (2) inferring the ASWU sequence (pronunciation) given the word orthography and the learned G2ASWU relationship. We present three methods for cross-domain ASWU-based lexicon development based on that understanding.

Method-I: Applying standard G2P conversion approach on the seed lexicon obtained from the OOD corpus

One possible way to generate pronunciations for the in-domain RM corpus is to use the ASWU-based lexicon from the WSJ0 corpus as the seed lexicon and train a G2ASWU converter. For this purpose, we investigated the state-of-the-art joint multigram approach [Bisani and Ney, 2008] for G2ASWU conversion. This was done by using the Sequitur software developed at RWTH Aachen University[4]. In our experiment, the maximum width of the graphone used was one, and the n-gram context size was 6.[5] As shown in Figure 6.5, first the G2ASWU relationship is learned on the ASWU-based lexicon for the WSJ0 corpus by training the G2ASWU converter. Then given the words in the RM corpus and the learned G2ASWU relationship, the pronunciations are inferred.[6]

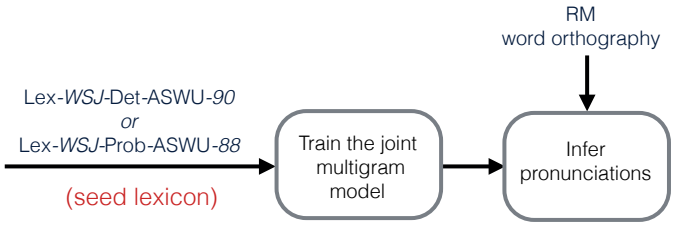


Figure 6.5 – Diagram of joint multigram-based pronunciation generation for RM corpus using the seed lexicon trained on WSJ0 corpus (*Method-I*).

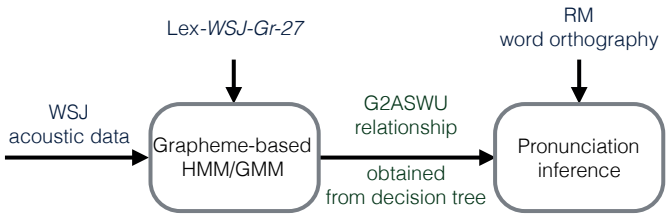
Method-II: Using the learned G2ASWU relationship on the OOD corpus for pronunciation inference on the in-domain corpus

Instead of using the ASWU-based lexicon from the WSJ0 corpus, only the learned G2ASWU relationships can be exploited for inferring pronunciations on the RM corpus. More precisely, we investigate using the deterministic and probabilistic G2ASWU relationships obtained from (a) the decision trees learned on WSJ0, and (b) the KL-HMM trained on WSJ0 respectively to generate pronunciations for the RM corpus, as illustrated in Figure 6.6.

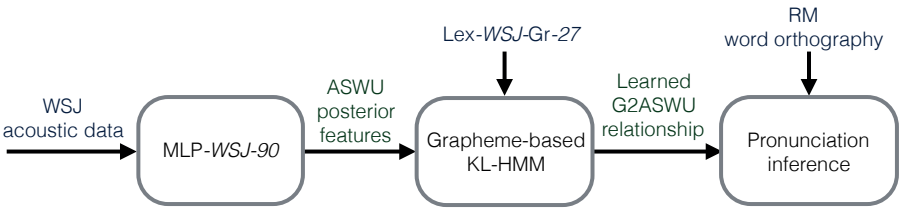
[4] http://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html

[5] As there are no canonical pronunciations in case of using ASWUs are available, we decided on the optimal n-gram context size based on the ASR WRR.

[6] The grapheme symbols such as single hyphen that appear in the RM word orthographies and have not been observed in the WSJ0 word orthographies were removed for the inference.



(a) Using a deterministic G2ASWU relationship learned on WSJ0 (*Method-II-a*).



(b) Using a probabilistic G2ASWU relationship learned on WSJ0 (*Method-II-b*).

Figure 6.6 – Illustration of pronunciation generation for RM corpus in *Method-II*.

Method-III: Learning the G2ASWU relationship on the in-domain corpus through acoustics

Instead of using the learned G2ASWU relationship on the WSJ0 corpus, we can use the trained MLP on WSJ0 corpus to estimate ASWU posterior probabilities for the RM speech data. Given the ASWU posterior probabilities as feature observations, a grapheme-based KL-HMM system can be trained on the RM corpus data. The pronunciation inference can then be done given the trained KL-HMM and the word orthographies, as shown in Figure 6.7.

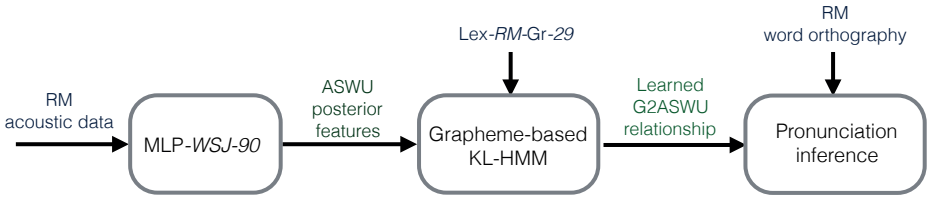


Figure 6.7 – Illustration of pronunciation generation for RM corpus using *Method-III*.

We generated ASWU-based lexicons for the RM corpus based on the three methods presented above. It is worth mentioning that, in addition to acoustic differences between the two corpora, there are also differences at lexicon level, i.e. 507 out of the 990 words in the RM lexicon do not appear in WSJ0 lexicon. For each of the lexicons developed, we trained CI and cross-word CD ASWU-based HMM/GMM systems with 39 dimensional PLP cepstral features extracted using the HTK toolkit. Each subword unit was modeled with three HMM states. Each CI HMM state was modeled by 32 Gaussian mixtures similar to in-domain studies in Section 6.3.2. Each tied HMM state was modeled by a mixture of 8 Gaussians. The HMM states were tied using a

singleton question set.

Table 6.4 presents the results in terms of WRR. For comparison purpose, we have reproduced the results for the lexicon Lex-*RM*-Gr-29, presented earlier in Table 6.3. It can be observed that the CI ASR systems, regardless of the method used for pronunciation generation, perform better than the grapheme-based CI ASR system. The performance of the CD ASR systems using the pronunciations generated through *Method-I* is inferior to the grapheme-based ASR system (Table 6.3). The performance of the ASR systems using *Method-II* for pronunciation generation is comparable with the ASR systems obtained through in-domain studies (Table 6.3). Generating pronunciations using *Method-III* also leads to a comparable system to the in-domain ASWU-based ASR systems. Comparing the performance of the systems using *Method-I* for pronunciation generation with the systems using *Method-II* and *Method-III* shows that it is better to transfer the learned G2ASWU relationship or learn the G2ASWU relationship on target domain speech. A potential reason for this trend is that *Method-I* relies on availability of ground truths, like availability of seed lexicon obtained through linguistic expertise in G2P conversion, which in the present scenario is not available. Overall, *Method-II* leads to the best ASR performance. It may be possible to improve *Method-III* by acoustic model adaptation techniques to adapt the MLP trained on the out-of-domain data. This is open for further research. Together these studies show that, in the proposed approach, the derived ASWUs and the G2ASWU relationship learned from one domain are transferable to another or target domain. Alternately, the proposed approach inherently enables such transfer.

Table 6.4 – ASR system performances in terms of WRR on RM corpus using different cross-domain pronunciation generation methods.

Method	G2ASWU relationship	CI	CD
Method-I	Deterministic	87.5	92.3
	Probabilistic	85.2	91.3
Method-II	Deterministic	89.0	94.4
	Probabilistic	88.8	94.0
Method-III	Probabilistic	89.0	94.0
Lex-*RM*-Gr-29	-	84.2	94.0

6.3.3 Comparison to existing approaches

In this section, we compare the present work with two of the existing approaches in the literature that have reported studies on the WSJ0 and RM corpora with the same setup as that used in our studies. More precisely, we first compare our approach to the spectral clustering-based approach proposed in [Hartmann et al., 2013] on the WSJ0 corpus . We then study the proposed approach in comparison to the approach proposed by Bacchiani and Ostendorf in [Bacchiani and Ostendorf, 1999] and tested on the RM corpus.

Comparison to Hartmann et al. [2013] approach

In essence, the proposed approach is similar to the spectral-based clustering approach proposed in [Hartmann et al., 2013], as they both discover the ASWUs from the grapheme-based HMM/GMM system. However, there are two key differences between these approaches:

1. In our approach the ASWUs are discovered through decision tree-based clustering of the HMM states, while in [Hartmann et al., 2013], the subword units are derived through spectral based-clustering, which requires computation of similarity matrix between HMMs.

2. In our approach, the pronunciations are generated using the KL-HMM framework, while in [Hartmann et al., 2013], the pronunciations are transformed using a statistical machine translation approach.

As the experimental setup in this chapter on WSJ0 corpus and the work in [Hartmann et al., 2013] are the same, we provide a comparison between the baseline and the results in both works in Table 6.5. In [Hartmann et al., 2013] there are two grapheme baselines. One based on the standard orthography (denoted as grapheme-direct) and the other based on grapheme-to-grapheme (G2G) conversion (denoted as grapheme-transformed) employing an approach similar to machine translation. Similarly, in the ASWU-based study as well they have two systems: One where the pronunciations are generated directly by mapping the graphemes to ASWUs based on the spectral clustering (denoted as ASWU-direct), and the other where ASWU-to-ASWU conversion is performed like G2G case mentioned above (denoted as ASWU-transformed). We ensured that our systems have comparable number of parameters in the case of both using CI subword unit and CD subword unit-based systems. It can be observed that the ASWU-based lexicon developed by our approach leads to a better ASR system. Furthermore, when comparing the best systems there is an absolute difference of 2.5% WRR, which indicates that the proposed approach in this chapter leads to a better ASR system.

Table 6.5 – Comparison with the related work in [Hartmann et al., 2013].

Approach	Lexicon	CI	CD
Approach proposed in [Hartmann et al., 2013]	Grapheme-direct	60.1	84.2
	Grapheme-transformed	68.6	85.5
	ASWU-direct	70.7	85.6
	ASWU-transformed	76.7	86.2
Present work	Grapheme	68.9	85.8
	Lex-*WSJ*-Det-ASWU-90	78.6	88.7
	Lex-*WSJ*-Prob-ASWU-88	78.7	87.3

Comparison to Bacchiani and Ostendorf [1999] approach

In a broad sense, the proposed approach and the joint subword unit derivation and pronunciation generation method proposed in [Bacchiani and Ostendorf, 1999] can be considered to be similar as,

1. both approaches consist of segmentation and clustering steps, except that in our approach the segmentation and clustering is guided through graphemes during the HMM/GMM training; and

2. both approaches apply the pronunciation length constraint which ensures uniformity in the number of segments for training tokens of a word. In our approach this is automatically achieved through use of a unique grapheme sequence representation for each word.

In our studies, we have used the RM corpus, which was also used in [Bacchiani and Ostendorf, 1999]. However there are a few distinctions. In [Bacchiani and Ostendorf, 1999], the states of the HMMs were modeled by a single Gaussian as opposed to a mixture of Gaussians and the evaluation was carried out only on *Feb89* test set. So we also trained a single Gaussian HMM/GMM system using the ASWU lexicon developed by our approach and evaluated on the *Feb89* test set. Table 6.6 presents the results in the case where the two approaches are similar in terms of number of ASWUs and clustered states. Table 6.7 provides a comparison between the best performance reported in [Bacchiani and Ostendorf, 1999] and the performance achieved with the lexicon based on our approach on the *Feb89* test set with 2937 clustered states. These results indicate that the ASWU lexicon developed by the proposed approach can yield ASR systems comparable to the ASWU lexicon developed by Bacchiani and Ostendorf [1999] approach, which needs additional heuristics to constrain the ASWU derivation and pronunciation generation process and necessitates all the words to be observed.

Table 6.6 – Comparison with the related work in [Bacchiani and Ostendorf, 1999] on *Feb89* test set using single Gaussian distributions.

	# of base units	# of clustered states	WRR
Approach proposed in [Bacchiani and Ostendorf, 1999]	124	1519	86.3
Present work	92	1559	86.9

Before concluding this section, it is worth mentioning that the approach proposed in [Singh et al., 2002] was also investigated on the RM corpus. Furthermore, there are also similarities with respect to our approach, as it also exploits transcribed speech data and it uses a grapheme-based dictionary as the initial lexicon. However, the results presented in [Singh et al., 2002] can not be fairly compared against our results for the following reasons: (1) the training and test

Table 6.7 – Comparison of the best result reported in [Bacchiani and Ostendorf, 1999] on *Feb89* test set with the result using the present work on the same test set using single Gaussian distributions.

	WRR
Approach proposed in [Bacchiani and Ostendorf, 1999]	91.2
Present work	91.1

sets are different. In particular, in their studies the test set contains 1600 utterances as opposed to the standard test of 1200 utterances, and (2) their ASR system is based on semi-continuous HMMs while in the present work the ASR system is based on continuous density HMMs. Informally, it can be stated that the proposed approach in this chapter has been investigated against stronger grapheme-based and phone-based baselines than the investigations reported in [Singh et al., 2002].

6.4 Application to an under-resourced language

In the previous section, we demonstrated the potential of the proposed framework for subword unit derivation and pronunciation generation on the well-resourced English language. Most of the state-of-the-art speech recognition approaches have emerged through investigations on English. So it can be argued that the proposed approach of deriving ASWUs using grapheme-based HMM/GMM system may be well-suited just for English. Furthermore, the G2P relationship varies across languages. Therefore, a question that arises is that whether the proposed approach is scalable to other languages or not.

In this section, our goal is two-fold. More precisely, our goal is to show the transferability of the approach to a new language as well as its utility to under-resourced languages, specifically languages that do not have well developed phonetic resources. In that direction, we present investigations on a genuinely under-resourced language, Scottish Gaelic. Unlike English, which belongs to family of Germanic languages, Scottish Gaelic belongs to family of Celtic languages. Our investigations are organized along two lines:

1. *Monolingual ASR studies*: We investigate the potential of the ASWU-based lexicons through monolingual ASR studies where we compare the performance of the ASWU-based ASR system with the alternative grapheme-based ASR system, as done in the studies on English.

2. *Multilingual ASR studies*: In [Rasipuram and Magimai.-Doss, 2015], it has been shown that performance of an under-resourced ASR system can be significantly improved by (a) training a multilingual acoustic model that estimates multilingual phone posterior probabilities using resources of well-resourced languages, and then (b) learning a probabilistic lexical model that captures the grapheme-to-multilingual phone relationship on the target language speech. So we also investigate if the ASWU-based lexicons hold their benefit

in such a multilingual ASR system scenario as well. As a product of the study, later in Section 6.5, we show how phonetic identities of the derived ASWUs could be discovered.

The remainder of the section is organized as follows. Section 6.4.1 briefly describes the characteristics of Scottish Gaelic. Section 6.4.2 presents the details of the ASWU-based lexicon development. Finally, Section 6.4.3 and Section 6.4.4 present the monolingual ASR and multilingual ASR studies, respectively.

6.4.1 Characteristics of the Scottish Gaelic language

Scottish Gaelic belongs to the class of Celtic languages. There are six Celtic languages that are still spoken. These languages are divided into two groups of Goidelic languages and Brythonic languages. Scottish Gaelic belongs to Goidelic languages along with Irish and Manx. It can be considered as a truly endangered language as it is spoken only by about 60,000 people. There are about 51 phones in the language [Rasipuram et al., 2013a]. However, the number of phones can change depending on the dialect. The language lacks a proper phonetic lexicon and the available transcribed speech data is also limited.

Scottish Gaelic alphabet has 18 letters, consisting of five vowels and thirteen consonants. The long vowels are represented with grave accents (À, È, Ì, Ò, Ù). There are twelve basic consonant types in Scottish Gaelic (B, C, D, F, G , I , L, M, N, P, R, S, T):

- Each consonant is either fortis or lenis (i.e., they are produced with greater or less energy). The lenited consonants are presented in the orthography with a grapheme [H] next to them.

- Each consonant is either broad (velarized) or slender (palatalized). Broad consonants are surrounded by broad vowels (A, O or U), while slender consonants are surrounded by slender vowels (E or I).

Scottish Gaelic orthography is less complicated than English. The complications partly arise due to the reason that modern orthography is based on Classical Irish orthography and the L2S rule may depend on the dialect [Rasipuram et al., 2013a]. The number of graphemes in Gaelic words is typically greater than the number of phones in the word due to the effect of lenited and broad/slender graphemes on the pronunciation. The G2P relationship in Scottish Gaelic can therefore be many-to-one. For example, the ratio of the number of graphemes to phones in the Gaelic word *SUIDHEACHADH* with pronunciation "sMj@x@G" (in the SAMPA format) is 1.7.

We conduct the studies on the Scottish Gaelic corpus explained in Section 2.5.6.

6.4.2 ASWU derivation and pronunciation generation setup

The setup for subword unit derivation and pronunciation generation for Scottish Gaelic is explained below.

Acoustic subword unit derivation: For automatic discovery of subword units, cross-word CD grapheme-based HMM/GMM systems were trained using 39-dimensional PLP cepstral features. Each CD grapheme was modeled with a single HMM state. Different numbers of ASWUs were obtained by adjusting the log-likelihood increase during decision tree clustering. The range for the number of ASWUs was decided to be similar to the range investigated in the studies on English, resulting in 85, 91 and 97 units.

Deterministic lexical modeling-based G2ASWU conversion: For deterministic lexical modeling-based G2ASWU conversion, the learned decision trees during ASWU derivation were exploited to map each grapheme in the word to an ASWU. We denote the lexicons generated using the deterministic lexical modeling-based G2ASWU conversion as Lex-*SG*-Det-ASWU-*M* where *M* denotes the number of ASWUs.

Probabilistic lexical modeling-based G2ASWU conversion: For probabilistic lexical modeling-based G2ASWU conversion, first a five-layer MLP classifying ASWUs was trained in which each hidden layer had 1000 hidden units. Then given the ASWU posterior probabilities from the ANN as feature observations, a CD grapheme-based KL-HMM was trained. For the pronunciation inference, the ASWU posterior probabilities were decoded through the ergodic HMM in which each ASWU was modeled with three left-to-right HMM states.

Table 6.8 shows the properties of the ASWU-based lexicons generated using a probabilistic lexical modeling-based G2ASWU conversion. Similar to the studies on English, it can be observed that some of the ASWUs are pruned out during the pronunciation generation given the probabilistic G2ASWU mapping.

Table 6.8 – Summary of the ASWU-based lexicons obtained through probabilistic lexical modeling-based G2ASWU conversion for Scottish Gaelic corpus.

Lexicon	MLP
Lex-*SG*-Prob-ASWU-76	MLP-*SG*-85
Lex-*SG*-Prob-ASWU-82	MLP-*SG*-91
Lex-*SG*-Prob-ASWU-86	MLP-*SG*-97

We selected the optimal number of ASWUs and the corresponding lexicon based on the WRR on the development set. Lex-*SG*-Det-ASWU-85 and Lex-*SG*-Prob-ASWU-82 yielded the best ASR systems and are therefore used in the ASR studies presented below.

6.4.3 Monolingual ASR system studies

As mentioned earlier, there is no well developed phonetic lexicon for Scottish Gaelic. So we evaluate the utility of the developed ASWU-based lexicon against a grapheme-based lexicon by conducting monolingual ASR studies. Specifically, we compare them across two frameworks, namely, HMM/GMM framework and KL-HMM framework, which has shown to be useful in under-resourced scenarios [Vu et al., 2014, Rasipuram and Magimai.-Doss, 2015].

HMM/GMM framework: We trained CI and cross-word CD HMM/GMM systems with 39 dimensional PLP cepstral features extracted using the HTK toolkit. Each subword unit was modeled with three HMM states. In the case of using CI subword units, the optimal number of Gaussian mixtures for the grapheme-based ASR system was 64 based on the best WRR obtained on the development set. For the ASWU-based ASR systems, the number of Gaussian mixtures was set to 16 so as to have a comparable number of parameters to the grapheme-based system. In the case of using CD subword units, for tying the HMM states singleton questions were used. Each HMM state was modeled by a mixture 8 Gaussians. The number of tied states in all the systems was roughly the same.

KL-HMM framework: This is done by using the posterior-based framework of KL-HMM directly for speech recognition. More precisely, instead of using the KL-HMM parameters capturing a probabilistic G2ASWU relation for pronunciation inference, they are used in the KL-HMM ASR framework. In this case, we can visualize it as an approach that integrates pronunciation learning implicitly as a phase in ASR system training [Rasipuram et al., 2015]. Our main motivation for performing this study was to ascertain whether doing lexicon development and ASR training as two separate stages can bring any advantage over doing direct speech recognition using grapheme-based KL-HMM system. For this purpose, we compared the KL-HMM system corresponding to the grapheme-based lexicon, i.e., Lex-*SG*-Gr-32, with two KL-HMM systems corresponding to lexicons Lex-*SG*-Det-ASWU-85 and Lex-*SG*-Prob-ASWU-82 as illustrated in Figure 6.8. All the systems use the same MLP, which is *MLP-SG-91*, as the acoustic model to estimate posterior feature observations.

Table 6.9 presents the performance of the HMM/GMM and KL-HMM systems in terms of WRR. It can be observed that Lex-*SG*-Prob-ASWU-82 yields significantly better CI and CD systems than Lex-*SG*-Gr-32 in both the HMM/GMM framework and the KL-HMM framework. Lex-*SG*-Det-ASWU-85 yields a better system in KL-HMM framework but a worse system in the HMM/GMM framework against Lex-*SG*-Gr-32. A possible reason for such a trend could be that, as discussed earlier, in Scottish Gaelic the G2P relationship is many-to-one due to lenition and broad and slender consonants. So, when inferring pronunciations using the deterministic G2ASWU mappings, each grapheme in the word is invariably mapped into an ASWU. This can result in systematically erroneous pronunciations, leading to mismatch between acoustics and the pronunciation model, as is the case for pronunciation variation. In the literature, it

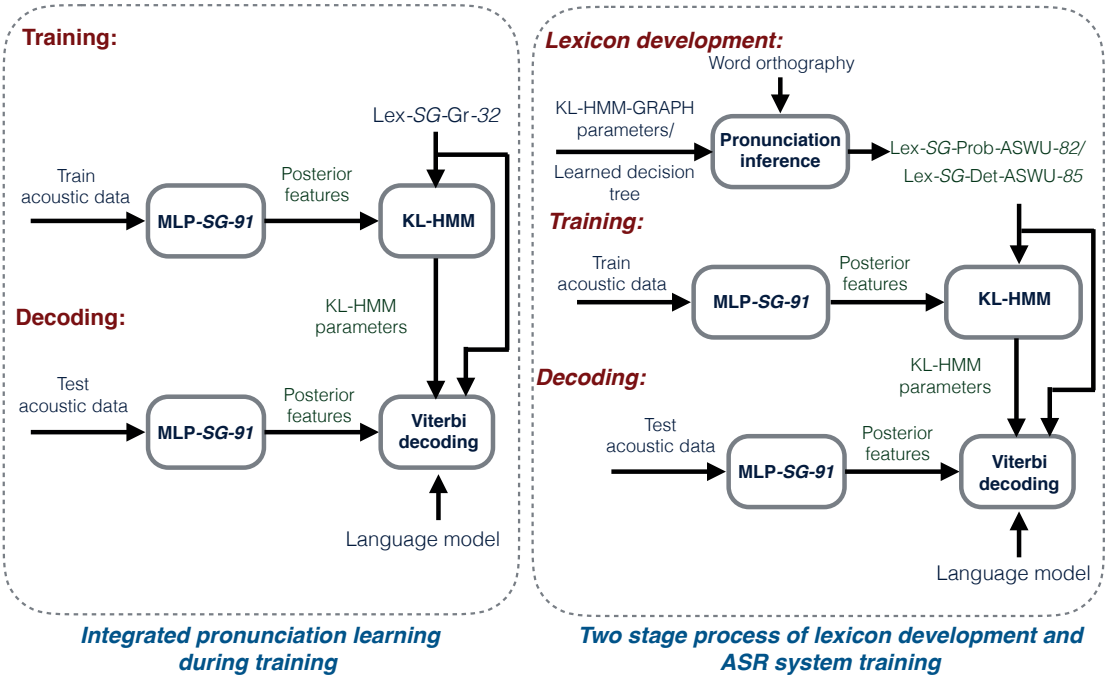


Figure 6.8 – Illustration of KL-HMM-based ASR systems based on Lex-*SG*-Gr-32, Lex-*SG*-Det-ASWU-85 and Lex-*SG*-Prob-ASWU-82.

has been observed that the KL-HMM approach is capable of handling pronunciation variation [Imseng et al., 2011, Razavi and Magimai.-Doss, 2014]. As a consequence, unlike the HMM/GMM framework, we observe that Lex-*SG*-Det-ASWU-85 yields a better system than Lex-*SG*-Gr-32 in KL-HMM framework.

Table 6.9 – Performance of HMM/GMM and KL-HMM systems in terms of WRR using context-independent (CI) and context-dependent (CD) subword units. For the KL-HMM systems, MLP-*SG*-91 is used as the acoustic model.

Lexicon	HMM-GMM		KL-HMM	
	CI	CD	CI	CD
Lex-*SG*-Gr-32	46.0	64.6	35.6	66.8
Lex-*SG*-Det-ASWU-85	54.5	63.3	52.2	69.1
Lex-*SG*-Prob-ASWU-82	59.6	66.4	57.5	69.5

6.4.4 Multilingual ASR system studies

As mentioned earlier, the performance of the under-resourced ASR system can be improved by using an acoustic model or ANN that classifies multilingual phones, and learning a probabilistic relationship between the graphemes and multilingual phones using KL-HMM. We compared the grapheme-based lexicon and the ASWU-based lexicon in that framework by,

1. first training a five-layer multilingual MLP on five auxiliary languages from SpeechDat(II) corpus namely British English, Swiss French, Swiss German, Italian and Spanish to estimate posterior probabilities of multilingual phones. The multilingual phoneset was formed by merging the phones that are shared across the aforementioned languages, leading to 117 phone units. We refer to this MLP as MLP-*MULTI*-117; and then

2. training a KL-HMM-based ASR system corresponding to each of the lexicons Lex-*SG*-Gr-32, Lex-*SG*-Det-ASWU-85 and Lex-*SG*-Prob-ASWU-82, as illustrated in Figure 6.9.

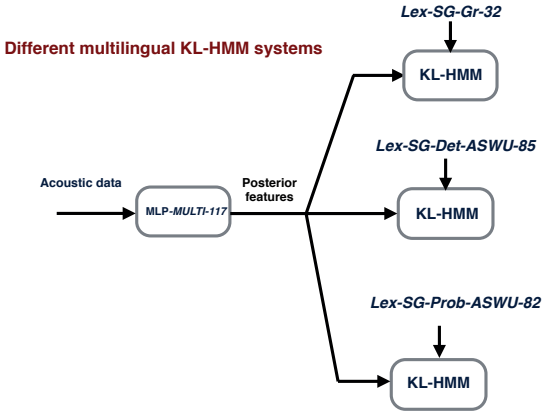


Figure 6.9 – Illustration of KL-HMM-based ASR systems using Lex-*SG*-Gr-32, Lex-*SG*-Det-ASWU-85 and Lex-*SG*-Prob-ASWU-82, and exploiting auxiliary multilingual resources.

Table 6.10 presents the performance of the different KL-HMM-based systems in terms of WRR. It can be observed that the ASWU-based lexicon yields a significantly better system than grapheme-based lexicon. Thus, showing that the proposed approach of ASWU-based lexicon development generalizes to multilingual resource sharing scenarios.

Table 6.10 – Performance of KL-HMM-based ASR systems exploiting auxiliary resources from well-resourced languages in terms of WRR. In these systems, MLP-*MULTI*-117 is used as the acoustic model.

Lexicon	CI	CD
Lex-*SG*-Gr-32	36.7	69.1
Lex-*SG*-Det-ASWU-85	52.1	70.7
Lex-*SG*-Prob-ASWU-82	57.7	**72.6**

6.5 Analysis

The ASR studies validated the proposed ASWU-based lexicon from a speech technology perspective. As explained in Section 6.2.1, one of our hypotheses in this chapter was that the ASWUs obtained from the clustered CD grapheme units are "phone-like". This section focuses on that aspect through an analysis of the derived ASWUs (Section 6.5.1) and the generated

pronunciations (Section 6.5.2). It is worth mentioning that a fully fledged quantitative analysis and concretely linking the derived ASWUs and lexicon to existing linguistic knowledge would need a separate investigation, and is thus out of the scope of the chapter. In this section, our main goal is to provide a qualitative analysis and demonstrate how links to existing linguistic knowledge can be established to gain better understanding. We notate the derived ASWUs with the notation used by HTK to represent clustered CD units. For example, ASWU [ST_A_26] means a clustered CD unit with the center grapheme [A] as the root node in the decision tree.

6.5.1 Relating the derived ASWUs to phonetic units

This section analyzes the relationship between the derived ASWUs and phonetic identities for English and Scottish Gaelic. In the case of English, the analysis uses the acoustic models of the phone-based system, while in the case of Scottish Gaelic there are no phone-based lexicons available. So the analysis leverages from the ASWU-to-multlinugual phone relationship learned by the KL-HMM system presented in Section 6.4.4.

Studies on English

For both WSJ0 and RM corpora, we computed the KL-divergence between the Gaussian distribution modeling a mono-phone unit and the Gaussian distribution modeling an ASWU in the HMM/GMM setup. We computed the KL-divergence between single Gaussians, as this is the step at which ASWU is derived by clustering CD graphemes. The KL-divergence between the Gaussian $\mathcal{N}_0(\mu_0, \Sigma_0)$ modeling a CI phone unit as the reference distribution and the Gaussian $\mathcal{N}_1(\mu_1, \Sigma_1)$ modeling an ASWU as the measured distribution is computed as [Duchi, 2007]:

$$0.5\{\mathrm{T}(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - K - \ln\frac{|\Sigma_0|}{|\Sigma_1|}\},$$

where μ, Σ and K are the mean vector, the covariance matrix and dimension of the vector space respectively.

Table 6.11 provides a few ASWUs along with the three most related phones according to the KL-divergence matrix. Furthermore, the table also provides example English words that contain the ASWUs within their pronunciations. In each example, the grapheme that has been mapped to the ASWU in the pronunciation is highlighted.

It can be observed from the table that a consistent relationship between the ASWUs and phones exists. This relationship can be clearly observed in the case of consonant graphemes (such as [L], [M], [N] and [R]). For example, the ASWUs belonging to grapheme [L] (such as [ST_L_22] and [ST_L_24] in the WSJ0 corpus) are more related to /el/ and /l/ sounds and the ASWUs belonging to grapheme [R] (such as [ST_R_25] and [ST_R_26] in the RM corpus) are more related to /r/, /axr/, and /er/ sounds. The observations here are also consistent with the

Table 6.11 – Relation between example automatically derived subword units and phone units based on the KL-divergence matrix. The example pronunciations are obtained from *Lex-WSJ-Det-ASWU-90* and *Lex-RM-Prob-ASWU-90* for the WSJ0 and RM corpora respectively.

(a) WSJ0 corpus

ASWU	mapped phone	example word	ASWU	mapped phone	example word
[ST_A_26]	/eh/,/ae/,/ey/	DECELERATION	[ST_L_24]	/l/,/el/,/ao/	INCLINED
[ST_A_28]	/eh/,/ih/,/ae/	AHEAD	[ST_M_22]	/m/,/em/,/n/	CRAMMING
[ST_C_21]	/z/,/s/,/zh/	DEVICE	[ST_N_22]	/ng/,/en/,/n/	RACING
[ST_C_22]	/t/,/dx/,/k/	FORTHCOMING	[ST_N_23]	/n/,/en/,/ng/	REMAINS
[ST_D_23]	/dx/,/d/,/g/	FOUNDATION	[ST_O_22]	/ow/,/ao/,/aa/	QUOTAS
[ST_E_27]	/ih/,/eh/,/uh/	SEND	[ST_R_21]	/r/,/er/,/axr/	AMERICA
[ST_E_28]	/iy/,/y/,/uw/	SEEN	[ST_R_25]	/axr/,/r/,/uh/	ADVERTISERS
[ST_F_22]	/th/,/f/,/t/	SHIFTED	[ST_S_21]	/s/,/z/,/f/	ACCOUNTS
[ST_H_23]	/hh/,/dx/,/th/	HAD	[ST_T_21]	/t/,/th/,/dx/	AUSTRIA
[ST_I_24]	/iy/,/ey/,/y/	INVENTORIES	[ST_U_24]	/uh/,/ax/,/ih/	ACTUAL
[ST_I_27]	/ih/,/uh/,/ax/	JIMMY	[ST_V_21]	/v/,/d/,/dh/	ACHIEVED
[ST_J_21]	/dx/,/jh/,/t/	JOIN	[ST_W_21]	/w/,/l/,/dx/	ALWAYS
[ST_K_21]	/t/,/dx/,/k/	LOCKED	[ST_Y_23]	/iy/,/y/,/ih/	ANYBODY
[ST_L_22]	/el/,/l/,/w/	IMPOSSIBLE	[ST_Z_21]	/z/,/s/,/dx/	ZEUS

(b) RM corpus.

ASWU	mapped phone	example word	ASWU	mapped phone	example word
[ST_A_211]	/aa/,/aw/,/ay/	CHART	[ST_N_21]	/n/,/en/,/ng/	CAMDEN
[ST_A_25]	/ae/,/ey/,/ay/	TRACK	[ST_O_21]	/ow/,/ao/,/ah/	LOCATED
[ST_A_26]	/ey/,/eh/,/ae/	DEGRADE	[ST_O_26]	/ah/,/ow/,/uh/	MONDAY
[ST_B_21]	/d/,/b/,/t/	BAD	[ST_R_25]	/er/,/axr/,/r/	SUMERRIZE
[ST_C_21]	/z/,/s/,/hh/	GARCIA	[ST_R_26]	/r/,/axr/,/er/	THREAT
[ST_D_22]	/dx/,/em/,/d/	ADDING	[ST_S_21]	/sh/,/ch/,/s/	WABASH
[ST_E_21]	/iy/,/ey/,/uw/	SPEED	[ST_S_24]	/z/,/s/,/ch/	WADSWORTH
[ST_E_25]	/axr/,/er/,/r/	SURFACE	[ST_T_21]	/t/,/th/,/dx/	WESTERN
[ST_F_22]	/f/,/th/,/hh/	VANDERGRIFT	[ST_T_24]	/dx/,/em/,/t/	BETTER
[ST_H_22]	/hh/,/dx/,/em/	HAD	[ST_U_21]	/ah/,/uh/,/ax/	DOUBLE
[ST_H_24]	/dh/,/hx/,/em/	NORTHERN	[ST_U_22]	/uw/,/ey/,/iy/	TWO
[ST_I_24]	/ih/,/eh/,/uh/	BAINBRIDGE	[ST_W_21]	/w/,/dx/,/em/	WEDNESDAY
[ST_M_21]	/m/,/n/,/ng/	BISMARK	[ST_Y_22]	/ih/,/y/,/uw/	ANYBODY

empirical observations made in an earlier grapheme-based ASR study on English [Rasipuram and Magimai.-Doss, 2013b], where the G2P relationship is also learned through acoustics.

Studies on Scottish Gaelic

As mentioned earlier, in the case of Scottish Gaelic there are no phonetic lexicons available. So we analyzed the parameters or categorical distributions of the CI KL-HMM system trained with the lexicon Lex-*SG*-Prob-ASWU-82 in the multilingual ASR studies. Table 6.12 provides examples of mappings between the ASWUs and multilingual phones obtained by selecting the multilingual phone with the maximum probability in the categorical distribution corresponding to the ASWU. The mapped phones are shown in the SAMPA format along with the

probability of the multilingual phone within the brackets. Similar to the analysis on English, we have presented example Gaelic words that contain the ASWUs within their pronunciations.

Table 6.12 – Some of the ASWUs together with their mapped phones in SAMPA format and some example words.

ASWU	Mapped phone	Example word	ASWU	Mapped phone	Example word
[ST_C_21]	/x/ [0.7]	CA**C**H			
[ST_C_22]	/C/ [0.7]	SMAOINI**C**H	[ST_T_21]	/h/ [0.6]	**TH**OG
[ST_C_23]	/k/ [0.9]	**C**ADAL	[ST_T_24]	/t/ [0.7]	MO**T**A
[ST_S_21]	/S/ [0.8]	RI**S**	[ST_G_22]	/g/ [0.5]	**G**AD
[ST_S_23]	/s/ [0.8]	THU**S**A	[ST_G_23]	/k/ [0.5]	LA**G**
[ST_F_21]	/f/ [0.7]	**PH**ÀIRT	[ST_R_22]	/r/ [0.4]	MA**R**
[ST_B_21]	/b/ [0.5]	**B**RIS	[ST_L_21]	/l/ [0.8]	SAOI**L**
[ST_B_22]	/v/ [0.4]	A-**BH**OS	[ST_L_23]	/l/ [0.5]	SGEU**L**
[ST_À_21]	/a/ [0.5]	MH**À**L	[ST_Ò_21]	/o/ [0.3]	SP**Ò**RS
[ST_A_212]	/@/ [0.4]	AG**A**D	[ST_O_23]	/o/ [0.3]	ST**O**C
[ST_E_21]	/@/ [0.4]	S**E**	[ST_I_23]	/I/ [0.7]	TR**I**C
[ST_E_23]	/l/ [0.3]	WHAL**E**S	[ST_I_28]	/i/ [0.2]	TR**Ì**

It can be observed from Table 6.12 that the ASWUs indeed relate to phonetic units in a consistent manner. For example, the ASWU [ST_S_21] is mapped to the phone /S/ (as found in the pronunciation of the English word *SHIP*: /S/ /I/ /p/) and is used in the pronunciation of the Scottish Gaelic word *RIS*, which has the slender consonant grapheme [S]. On the other hand, the ASWU [ST_S_23] is mapped to the sound /s/ (as used in the pronunciation of the English word *SKY*: /s/ /k/ /a/ /I/) and is found in the pronunciation of the Gaelic word *THUSA*, which contains the broad consonant [*S*].[7] Similarly the consonant ASWUs [ST_F_21] and [ST_R_22] are related to sound units /f/ and /r/. For the vowel ASWUs such as [ST_I_28] and [ST_E_21], the ASWUs are related to the phonetic units, however with a relatively low probability. In our approach, the ASWUs are derived by clustering CD graphemes. So the low probability can be due to the reason that a CD vowel grapheme unit can get mapped to more than one phone, whereas a CD consonant grapheme can have a one-to-one relationship to a phone.

6.5.2 Generated pronunciations

This section provides a brief analysis on the generated pronunciations through deterministic and probabilistic G2ASWU modeling for English and Scottish Gaelic to get an understanding about the generated pronunciations along with the relation to phonetic identities inferred in the previous section.

[7] Note that in Scottish Gaelic, the broad consonant grapheme [S] is pronounced as the English sound /s/ while the slender [*S*] is pronounced as the English sound /S/.

English

Table 6.13 presents a few words selected from ASWU-based lexicons generated for WSJ0 and RM corpora. For each word, the first pronunciation is based on the deterministic G2ASWU conversion and the second pronunciation is based on the probabilistic G2ASWU conversion.

Table 6.13 – Few example words together with their generated pronunciations based on a deterministic or a probabilistic lexical modeling-based G2ASWU conversion on WSJ0 and RM corpora.

(a) WSJ0 corpus.

Word	Lex-*WSJ*-Det-ASWU-90 Lex-*WSJ*-Prob-ASWU-88
ACCENT	[ST_A_22] [ST_C_23] [ST_C_21] [ST_E_27] [ST_N_24] [ST_T_24] [ST_A_22] [ST_C_23] [ST_S_25] [ST_E_27] [ST_N_24] [ST_T_24]
ACCORD	[ST_A_22] [ST_C_23] [ST_C_22] [ST_O_21] [ST_R_23] [ST_D_21] [ST_A_22] [ST_C_23] [ST_C_22] [ST_O_21] [ST_R_23] [ST_D_21]
ALAN	[ST_A_22] [ST_L_24] [ST_A_27] [ST_N_21] [ST_A_22] [ST_L_24] [ST_A_25] [ST_N_21]
ALARM	[ST_A_22] [ST_L_24] [ST_A_24] [ST_R_26] [ST_M_24] [ST_A_22] [ST_L_24] [ST_A_24] [ST_R_26] [ST_M_24]
PHONE	[ST_P_21] [ST_H_23] [ST_O_29] [ST_N_24] [ST_E_21] [ST_F_22] [ST_O_29] [ST_N_21]
UPHELD	[ST_U_24] [ST_P_21] [ST_H_23] [ST_E_29] [ST_L_24] [ST_D_21] [ST_O_27] [ST_P_21] [ST_H_23] [ST_L_24] [ST_D_21]

(b) RM corpus.

Word	Lex-*RM*-Det-ASWU-92 Lex-*RM*-Prob-ASWU-90
CHOP	[ST_C_22] [ST_H_22] [ST_O_26] [ST_P_22] [ST_C_22] [ST_H_22] [ST_O_26] [ST_P_22]
CODE	[ST_C_23] [ST_O_26] [ST_D_22] [ST_E_24] [ST_C_23] [ST_O_26] [ST_D_22]
FLASHER	[ST_F_22] [ST_L_23] [ST_A_21] [ST_S_21] [ST_H_22] [ST_E_25] [ST_R_21] [ST_F_22] [ST_L_23] [ST_A_21] [ST_S_21] [ST_H_22] [ST_E_25] [ST_R_21]
PRESENT	[ST_P_22] [ST_R_26] [ST_E_28] [ST_S_24] [ST_E_6] [ST_N_22] [ST_T_25] [ST_P_22] [ST_R_26] [ST_E_28] [ST_S_24] [ST_I_27] [ST_N_22] [ST_T_25]

With the information provided in Table 6.11a and Table 6.11b, it can be observed that the G2ASWU conversion approach is able to recognize different sounds of the same grapheme to provide a pronunciation similar to what is seen in a phone-based lexicon. For example, in the case of the word *ACCENT*, the first grapheme [C] in the word is mapped to [ST_C_23], which in the earlier analysis was found to map to phone /k/. The second grapheme [C] is mapped to [ST_C_21] in the case of deterministic G2ASWU conversion and is mapped to [ST_S_25] in the case of probabilistic G2ASWU conversion, and in both cases the ASWUs map to /s/. Similar trends can be observed in the example pronunciations provided for the RM corpus. For example, the grapheme [S] is mapped to [ST_S_21] when it corresponds to /sh/ (*FLASHER*)

and is mapped to [ST_S_24] when it is related to /z/ (*PRESENT*). The distinction between the deterministic and probabilistic G2ASWU conversion can be very well observed through words *PHONE* and *UPHELD*. In the case of the word *PHONE*, the deterministic G2ASWU conversion maps each grapheme to an ASWU unit while the probabilistic G2ASWU conversion is able to map a group of graphemes to an ASWU, i.e. *PH* to /f/ and *NE* to /n/. In the case of the word *UPHELD*, it can be observed that the probabilistic G2ASWU conversion leads to deletion of a unit while the deterministic G2ASWU preserves the unit. We speculate that the inferior performance of the probabilistic G2ASWU conversion in the ASR studies on English is mainly due to such deletions.

Scottish Gaelic

Table 6.14 presents a few words selected from the ASWU-based pronunciations in the case of using deterministic and probabilistic G2ASWU conversion. In order to help in interpreting the generated pronunciations in terms of known sound units, each ASWU in the pronunciation has been mapped to a multilingual phone with the highest probability, as explained in Section 6.5.1. Furthermore, we have provided the 'perceived' pronunciations for each word through informal hearing of the Gaelic words. This was done by using an online community-driven dictionary for Gaelic in which for most of the words an audio file pronouncing the word was available.[8]

Table 6.14 – Example words from Scottish Gaelic together with their pronunciations obtained from *Lex-SG-Det-ASWU-91* and *Lex-SG-Prob-ASWU-82*. For each word, we have also provided the mapped pronunciation based on the sequence of multilingual phone units together with its perceived pronunciations.

Word	Lex-*SG*-Det-ASWU-85 Lex-*SG*-Prob-ASWU-82	Mapped pron.	Perceived pron.
MHÀL	[ST_M_21] [ST_H_27] [ST_À_21] [S_L_22] [ST_B_22] [ST_À_21] [S_L_23]	/v/ /h/ /a/ /l/ /v/ /a/ /l/	/v/ /a/ /l/
THOG	[ST_T_21] [ST_H_27] [ST_O_23] [ST_G_23] [ST_T_21] [ST_O_23] [ST_G_23]	/h/ /h/ /o/ /k/ /h/ /o/ /k/	/h/ /O/ /g/
PHÒS	[ST_P_21] [ST_H_27] [ST_Ò_21] [ST_S_23] [ST_F_21] [ST_Ò_21] [ST_S_23]	/p/ /h/ /e/ /s/ /f/ /o/ /s/	/f/ /o/ /s/
VOTE	[ST_V_21] [ST_O_23] [ST_T_24] [ST_E_21] [ST_B_22] [ST_O_23] [ST_T_24] [ST_E_21]	/v/ /o/ /t/ /@/ /v/ /o/ /t/ /@/	/v/ /@U/ /t/
YOU	[ST_Y_21] [ST_O_23] [ST_U_22] [ST_I_28] [ST_O_23]	/j/ /o/ /u/ /i/ /o/	/j/ /u:/
KATY	[ST_K_21] [ST_A_212] [ST_T_24] [ST_Y_21] [ST_G_23] [ST_A_212] [ST_T_24] [ST_I_28]	/k/ /@/ /t/ /j/ /k/ /@/ /t/ /i/	/k/ /eI/ /t/ /i/

To better understand the generated pronunciations, we first note that in Scottish Gaelic, broad consonants *MH* and *PH* are pronounced as /v/ and /f/, respectively; and the broad consonant *TH* is pronounced as /h/.[9] It can be seen that the pronunciations obtained through

[8] http://www.learngaelic.net/dictionary/index.jsp
[9] https://en.wikipedia.org/wiki/Scottish_Gaelic_orthography

probabilistic lexical modeling-based G2ASWU conversion can better capture the linguistic rules compared to the pronunciations obtained through a deterministic lexical modeling-based G2ASWU conversion. For instance, in the word *PHOS* the broad consonant PH is mapped to /f/ in the probabilistic lexical modeling-based G2ASWU conversion, while in the deterministic approach, it is mapped to /p/ and /h/. Similarly, in the word *MHÀL*, the broad consonant *MH* corresponds to [ST_B_22], which is mapped to the /v/ in the pronunciation obtained from the probabilistic G2ASWU relationship modeling, whereas it is mapped to the /v/ and /h/ sounds in the pronunciations generated through the deterministic G2ASWU relationship modeling. Indeed, it can be observed that the mapped pronunciations obtained from the probabilistic G2ASWU modeling corroborate well with the perceived pronunciations in several cases.

For some of the borrowed English words (e.g., *YOU* and *KATY*), on the other hand, the generated pronunciations using the ASWUs seem to be influenced by Gaelic pronunciations. This could be due to a combination of factors such as accented English and limited number of English words in the training data.

6.6 Summary

This Chapter presented a novel approach for subword unit derivation and pronunciation generation using only word level transcribed speech data. In this approach, the subword units are first derived by clustering CD graphemes in an HMM-based ASR framework using maximum likelihood criteria; followed by modeling of the relationship between the graphemes and the derived units in a deterministic or probabilistic manner using acoustic data; and finally inferring pronunciations given the learned relationships and the word orthographies using an ergodic HMM. In comparison to existing approaches in the literature, a distinguishing aspect of the proposed approach is that it fits within the well-known HMM framework for ASR and speech synthesis, and is therefore fairly straight-forward to implement given the available toolkits such as HTK [Young et al., 2006] and KALDI [Povey et al., 2011].

Our experimental studies on two languages showed that the ASWU-based lexicon can be developed in a fully data-driven manner, i.e. the set of ASWUs and the corresponding lexicon can be selected through cross-validation. The ASR studies on both the languages showed that the ASWU-based lexicons consistently yield significantly better ASR systems compared to the grapheme-based lexicons. For G2ASWU conversion, we investigated two approaches, namely, decision tree-based approach and KL-HMM based acoustic G2P conversion. Our experimental studies also showed that both G2ASWU approaches are equally applicable, with the acoustic G2P conversion approach holding advantage for languages with many-to-one G2P relationship. Also, in one of the first efforts, we showed that the discovered ASWUs and the learned G2ASWU relationship can be transferred across domains in a language and the G2ASWU conversion mechanism inherently enables such transfer. Furthermore, the analysis of the learned models and the generated pronunciations showed that the derived ASWUs to a good extent are

systematically related to phonetic identities. In particular, studies on Scottish Gaelic showed that the multilingual ASR approach not only aids in development of a lexicon that yields a better ASR system, but also enables discovery of the phonetic identities of the derived ASWUs through the use of multilingual resources. This opens potential venues for further research and development to improve phonetic and lexical resources and technologies for under-resourced languages through transfer of linguistic knowledge and data across languages.

7 Conclusions and future directions

The goal of this thesis was to overcome the limitations of current methodologies for pronunciation lexicon development in terms of their ability to model natural phonological variation and dependency on availability of linguistic expertise. Toward that, we first focused on the problem of matching an acoustic signal with a word hypothesis, which is inherent in development of pronunciation lexicons through humans. We showed that the HMM-based ASR approaches achieve that match in an automatic manner via a latent symbol space, with the latent symbols being CI phones or cCD phones. Furthermore, we showed that the posterior-based matching approach like the KL-HMM approach is capable of achieving a performance comparable or better than the HMM/GMM approach and the hybrid HMM/ANN approach with a relatively small latent symbol set.

We then developed an abstract posterior-based formulation for pronunciation generation in an HMM framework, akin to hybrid HMM/ANN framework for ASR, and showed that the acoustic data-driven G2P conversion approach using KL-HMM is a particular case of this formulation. More specifically, we elucidated that the approach of using KL-HMM to learn the G2P relationship is an approach for learning a phone class conditional probability estimator by matching a word hypothesis represented in terms of graphemes with the speech signal using phones as the latent symbols. We incorporated the recent advances in neural network based acoustic modeling, i.e., use of deep architecture MLPs and modeling of cCD phones, into the acoustic data-driven G2P conversion approach and benchmarked it on two languages with deep orthographies, namely, English and French. Our studies showed that, despite the inferior PRR, the lexicon resulting from the acoustic data-driven G2P conversion approach yields ASR systems that are comparable to the ones using lexicons resulting from state-of-the-art G2P conversion approaches.

We further built on the posterior-based formulation to develop a multi-stream framework to: (a) unify G2P conversion approaches by utilizing them as multiple phone class conditional probability estimators and (b) unify G2P conversion and A2P conversion seamlessly through the aforementioned matching paradigm . We validated the multi-stream framework on the challenging task of developing pronunciation lexicons for uncommon words and proper

names, and demonstrated its utility by comparing it against other approaches commonly used in the literature to combine G2P conversion approaches and to incorporate acoustics along with G2P conversion for pronunciation variant selection.

Finally, we developed a novel approach for ASWU-based lexicon development. The proposed approach casts the problem of ASWU derivation as a problem of determining a latent symbol space given the acoustic signal and the corresponding word hypothesis, and exploits the capability of the acoustic data-driven G2P conversion approach to alleviate the need for a seed lexicon in the target domain for pronunciation generation. Our investigations on a well-resourced language English and a truly under-resourced language Scottish Gaelic showed that the derived ASWUs are phone-like and the ASWU-based lexicon yields better ASR systems than the grapheme-based lexicon.

In conclusion, this thesis developed a framework that can effectively exploit the available acoustic information and linguistic knowledge toward automatic pronunciation lexicon development. The framework essentially achieves that by integrating a novel posterior-based formulation for pronunciation generation with a posterior-based approach to match a word hypothesis with an acoustic signal through a latent symbol set. In doing so, the framework brings the pronunciation generation task and the pronunciation variation modeling task closer, and enables exploitation of tools and techniques developed for ASR to jointly address the challenges related to these tasks.

The work in this thesis could be further developed along the following directions:

1. **Extension to non-alphabetic languages:** The methods developed in this thesis for pronunciation lexicon development presume that the writing system is an alphabetic writing system, which encodes phone information and time sequence information. Not all languages have such writing systems. For example, Devanagri script is syllabic, where the script encodes consonant-vowels, not necessarily in a time linear fashion. Similarly Chinese script is logographic, where the symbol may represent both morpheme and meaning. Extending the proposed approaches to such writing systems is open for further research. One possible way would be to combine the proposed approaches with transliteration and transcription methods.

2. **Advancing the posterior-based formulation for G2P conversion:** Throughout our experimental studies using the posterior-based formulation for pronunciation generation, we assumed a uniform prior probability distribution and transition probability distribution. These assumptions were mainly made due to the limitation of data or lack of a canonical pronunciation lexicon. In some of our preliminary studies, we investigated incorporating phone transition probabilities by training phone n-grams, however the obtained pronunciation lexicons did not lead to better ASR systems. This could be due to the reason that the MLPs trained for estimating the posterior probability of acoustic units exploit the acoustic contextual information, and consequently the states of KL-HMM, which model the MLP outputs when learning the G2P relationship, could also

capture the contextual information [Rasipuram and Magimai.-Doss, 2016]. So further research is needed to ascertain the role of transition probabilities.

In this thesis, we mainly focussed on generation of single pronunciation or N-best pronunciations. While we observed improvements at the ASR level with N-best pronunciations, it is well understood that N-best pronunciations may not necessarily be optimal, especially due to the possibility of increasing confusion between the words. So there is a need to develop a pronunciation variant selection method, in conjunction with the use of approaches that can implicitly handle pronunciation variation [Luo and Jelinek, 1999, Hain, 2005, Imseng et al., 2011, Razavi and Magimai.-Doss, 2014].

3. **Use of articulatory features for ASWU derivation:** In the approach proposed for ASWU-based lexicon development, the problem of ASWU derivation was as posed as a problem of finding a latent symbol space that can be related to acoustic data and associated transcriptions (or graphemes). In this thesis, we used standard cepstral features, which tend to carry information related to phones to find the latent symbol space. However, there are alternative features or representations that carry phone related information and could be exploited to find a phone-like latent symbol space. For instance using linguistically motivated articulatory features (AFs) [Jakobson et al., 1992, Ladefoged, 1993], which may be a more robust representation when compared to spectral-based features and could help in reducing the gap between ASWU-based approach and phone-based approach. This could be achieved without deviating from the HMM framework through the recently proposed AF-based ASR framework using KL-HMMs [Rasipuram and Magimai.-Doss, 2016], where it has been show that ASR systems can be developed by learning the grapheme-to-AF relationship through acoustics. Alternately, we could cast the ASWU-based lexicon development as a three step process, where first acoustic-to-AF relationship is learned on the available multilingual resources; then grapheme-to-AF relationship is learned from the target language transcribed speech and clustered to derive ASWUs using KL-HMMs; and finally G2ASWU conversion is performed, as done in this thesis.

4. **Validating the developed framework on TTS:** In this thesis, we considered ASR as the end application to validate the proposed approaches for pronunciation lexicon development. It would be interesting to validate the framework with TTS as the end application. In particular, in our studies we found that the pronunciation level evaluation may not determine the best pronunciation lexicon for ASR. A question arising is that whether the same trend holds for TTS as well. Furthermore, it would be also interesting to investigate the potential of the ASWU-based lexicon development approach for development of TTS systems for under-resourced languages.

A KL-HMM

This appendix explains the KL-HMM training and decoding procedure.

A.1 KL-HMM training

Given a training set of N utterances $\{Z(n), W(n)\}_{n=1}^{N}$, where for each training utterance n, $Z(n)$ represents sequence of acoustic unit probability vectors $Z(n) = (\mathbf{z}_1(n), \cdots, \mathbf{z}_t(n), \cdots, \mathbf{z}_{T(n)}(n))$ of length $T(n)$ and $W(n)$ represents the sequence of underlying words, the KL-HMM parameters are estimated by a Viterbi EM procedure that minimizes the cost function,

$$C = \sum_{n=1}^{N} \min_{Q \in \mathcal{Q}} \sum_{t=1}^{T(n)} [S_{(R/S)KL}(\mathbf{y}^{q_t}, \mathbf{z}_t(n)) - \log a_{q_{t-1}q_t}] \tag{A.1}$$

where $Q = (q_1, \cdots, q_t, \cdots, q_{T(n)})$ denotes a sequence of HMM states, $q_t \in \{1, \cdots, I\}$, $\mathcal{Q}$ denotes the set of all possible HMM state sequences, and $a_{q_{t-1}q_t}$ corresponds to transition probabilities.

In practice, the transition probabilities $a_{q_{t-1}q_t}$ are assumed to be constant (0.5), similar to the hybrid HMM/ANN approach. Therefore parameter estimation amounts to estimating $\{\mathbf{y}^i\}_{i=1}^{I}$. Given a uniformly initialized set of parameters $\{\mathbf{y}^i\}_{i=1}^{I}$ (i.e., $y_d^i = \frac{1}{D} \ \ \forall i, D$) the segmentation step yields an optimal state sequence for each training utterance using Viterbi algorithm. Given the optimal state sequences, i.e., alignment and $\mathbf{z}_t$ belonging to each of these states, the optimization step then estimates a new set of model parameters by minimizing the cost function based on KL-divergence (Eqn. (A.1)) with the constraint that $\sum_{d=1}^{D} y_d^i = 1$. This process of segmentation and the optimization is iteratively done until convergence.

With S_{RKL} as the local score, the optimal state distribution is the arithmetic mean of the training acoustic state probability vectors assigned to the state, i.e.,

$$y_d^i = \frac{1}{M(i)} \sum_{\mathbf{z}_t(n) \in Z(i)} z_{t,d}(n) \quad \forall n, t \tag{A.2}$$

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

With S_{KL} as the local score, the optimal state distribution is the normalized geometric mean of the training acoustic state probability vectors assigned to the state, i.e.,

$$y_d^i = \frac{\hat{y}_d^i}{\sum_{d=1}^{D} \hat{y}_d^i} \quad where \quad \hat{y}_d^i = \Big(\prod_{\mathbf{z}_t(n) \in Z(i)} z_{t,d}(n) \Big)^{\frac{1}{M(i)}} \quad \forall n, t \tag{A.3}$$

where $\hat{y}_d^i$ represents the geometric mean of state i for dimension d, $Z(i)$ denotes the set of acoustic state probability vectors assigned to state i and $M(i)$ is the cardinality of $Z(i)$.

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

A.2 KL-HMM decoding

As defined by [Aradilla, 2008, Ch. 6.2.3], given the sequence of acoustic unit posterior probability vectors $Z = (\mathbf{z}_1, \cdots, \mathbf{z}_t, \cdots, \mathbf{z}_T)$ and the KL-HMM parameters, the best matching word sequence is obtained by minimizing the cost function,

$$W^* = \arg\min_{Q} \sum_{t=1}^{T} \left\{ S(\mathbf{y}^{q_t}, \mathbf{z}_t) - log a_{q_{t-1} q_t} \right\} \tag{A.4}$$

where $Q = (q_1, \cdots, q_T)$ denotes a sequence of HMM states. It can be observed that Eqn. (A.4) is similar to Eqn. (2.22), except that maximizing the log-likelihood $p(\mathbf{x}_t | q_t = l^i)$ is replaced with minimizing a KL-divergence based score $S(\mathbf{y}^{q_t}, \mathbf{z}_t)$.

Bibliography

L. Adde and T. Svendsen. NameDat: A Database of English Proper Names Spoken by Native Norwegians. *Higher education*, 17:16, 2010.

L. Adde and T. Svendsen. Pronunciation Variation Modeling of Non-Native Proper Names by Discriminative Tree Search. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4928–4931. IEEE, 2011.

G. K. Anumanchipalli, M. Ravishankar, and R. Reddy. Improving Pronunciation Inference using N-best List, Acoustics and Orthography. In *Proceedings of ICASSP*, volume 4, pages IV–925. IEEE, 2007.

G. Aradilla. *Acoustic Models for Posterior Features in Speech Recognition*. PhD thesis, Ph. D. dissertation, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2008.

G. Aradilla, H. Bourlard, and M. Magimai-Doss. Posterior Features Applied to Speech Recognition Tasks with User-Defined Vocabulary. In *Proceedings of ICASSP*, pages 3809–3812. IEEE, 2009.

M. Bacchiani and M. Ostendorf. Using Automatically-Derived Acoustic Sub-word Units in Large Vocabulary Speech Recognition. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 1998.

M. Bacchiani and M. Ostendorf. Joint Lexicon, Acoustic Unit Inventory and Model Design. *Speech Communication*, 29(2):99–114, 1999.

L. Bahl, P. Brown, P. De Souza, and R. Mercer. Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition. In *Proceedings of ICASSP*, volume 11, pages 49–52. IEEE, 1986.

G. Bernardis and H. Bourlard. Improving Posterior Based Confidence Measures in Hybrid HMM/ANN Speech Recognition Systems. In *Proceedings of ICSLP*, pages 775–778, 1998.

A. Bhattacharyya. On a Measure of Divergence Between Two Statistical Populations Defined by their Probability Distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

M. Bisani and H. Ney. Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation. In *Proceedings of ICASSP*, volume 1, pages 409–412, May 2004.

M. Bisani and H. Ney. Joint-sequence Models for Grapheme-to-phoneme Conversion. *Speech Communication*, 50(5):434–451, 2008.

A. W. Black, K. Lenzo, and V. Pagel. Issues in Building General Letter to Sound Rules. *ESCA Workshop on Speech Synthesis*, pages 77–80, 1998.

R. Blahut. Hypothesis Testing and Information Theory. *IEEE Transactions on Information Theory*, IT-20(4), 1974.

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

S. F. Chen. Conditional and Joint Models for Grapheme-to-Phoneme Conversion. In *Proceedings of Interspeech*, 2003.

N. Chomsky and M. Halle. *The Sound Pattern of English*. Harper & Row, New York, NY, 1968.

C-T Chung, C-A Chan, and L-S Lee. Unsupervised Discovery of Linguistic Structure Including Two-Level Acoustic Patterns Using Three Cascaded Stages of Iterative Optimization. In *Proceedings of ICASSP*, pages 8081–8085, 2013.

F. Coulman. The Blackwell Encyclopedia of Writing Systems, 1996.

G.E. Dahl, D. Yu, L. Deng, and A. Acero. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42, 2012.

R. I. Damper, Y. Marchand, J-D. S. Marsters, and A. I. Bazin. Aligning Text and Phonemes for Speech Technology Applications Using an EM-Like Algorithm. *International Journal of Speech Technology*, 8(2):149–162, 2005.

M. Davel and E. Barnard. Bootstrapping for Language Resource Generation. In *Proceedings of the 14th Symposium of the Pattern Recognition Association of South Africa, South Africa*, pages 97–100, 2003.

M. Davel and E. Barnard. Bootstrapping Pronunciation Models. 2006.

M. Davel and E. Barnard. Pronunciation Prediction with Default&Refine. *Computer Speech & Language*, 22(4):374–393, 2008.

S. B. Davis and P. Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pages 357–366, 1980.

M. J. Dedina and H. C. Nusbaum. PRONOUNCE: a Program for Pronunciation by Analogy. *Computer Speech & Language*, 5(1):55–64, 1991.

S. Deligne, F. Yvon, and F. Bimbot. Variable-Length Sequence Matching for Phonetic Transcription Using Joint Multigrams. In *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*. EUROSPEECH, 1995.

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

J. Duchi. Derivations for linear algebra and optimization. http://www.cs.berkeley.edu/~jduchi/projects/general_notes.pdf, 2007.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2001.

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 *Proceedings of ICASSP*, 1997.

H. Elovitz, R. Johnson, A. McHugh, and J. Shore. Letter-to-Sound Rules for Automatic Translation of English Text to Phonetics. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(6):446–459, 1976.

J. G. Fiscus. A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER). In *Proceedings of ASRU*, pages 347–354. IEEE, 1997.

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

E. Fosler-Lussier. A Tutorial on Pronunciation Modeling for Large Vocabulary Speech Recognition. volume 2705, pages 38–77. Springer, 2000.

E. Fosler-Lussier and J. Morris. Crandem Systems: Conditional Random Field Acoustic Models for Hidden Markov Models. In *Proceedings of ICASSP*, pages 4049–4052. IEEE, 2008.

R. Frost. Orthography and Phonology: The Psychological Reality of Orthographic Depth. Technical Report SR-99/100, 162-171, Haskins Laboratories, 1989.

S. Furui. Speaker-Independent Isolated Word Recognition using Dynamic Features of Speech Spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(1):52–59, February 1986.

M. Gales and S. Young. The Application of Hidden Markov Models in Speech Recognition. *Foundations and trends in signal processing*, 1(3):195–304, 2008.

M. Gales, K. Knill, and A. Ragni. Unicode-Based Graphemic Systems for Limited Resource Languages. In *Proceedings of ICASSP*, pages 5186–5190, 2015.

C. Genest and J. V. Zidek. Combining Probability Distributions: A Critique and an Annotated Bibliography. *Statist. Sci.*, 1(1):114–135, 02 1986.

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

K. Greer, B. Lowerre, and L. Wilcox. Acoustic Pattern Matching and Beam Searching. In *Proceedings of ICASSP*, volume 7, pages 1251–1254. IEEE, 1982.

S. Hahn, P. Vozila, and M. Bisani. Comparison of Grapheme-to-Phoneme Methods on Large Pronunciation Dictionaries and LVCSR Tasks. In *Proceedings of Interspeech*, pages 2538–2541, 2012.

S. Hahn, P. Lehnen, S. Wiesler, R. Schlüter, and H. Ney. Improving LVCSR with Hidden Conditional Random Fields for Grapheme-to-Phoneme Conversion. In *Proceedings of Interspeech*, pages 495–499, 2013.

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

W. Hartmann, A. Roy, L. Lamel, and J. Gauvain. Acoustic Unit Discovery and Pronunciation Generation from a Grapheme-Based Lexicon. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 380–385, 2013.

J. Hennebert, C. Ris, H. Bourlard, S. Renals, and N. Morgan. Estimation of Global Posteriors and Forward-Backward Training of Hybrid HMM/ANN Systems. In *Proceedings of EUROSPEECH*. International Speech Communication Association, 1997.

H. Hermansky. Perceptual Linear Predictive (PLP) Analysis of Speech. *Journal of the Acoustical Society of America*, 57(4):1738–52, April 1990.

G. Hinton et al. 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.

T. Holter and T. Svendsen. Combined Optimisation of Baseforms and Model Parameters in Speech Recognition Based on Acoustic Subword Units. In *Proceedings of ASRU*, pages 199–206, Dec 1997.

D. Imseng, R. Rasipuram, and M. Magimai.-Doss. Fast and Flexible Kullback-Leibler Divergence based Acoustic Modeling for Non-native Speech Recognition. In *Proceedings of ASRU*, December 2011.

D. Imseng, J. Dines, P. Motlicek, P. N. Garner, and H. Bourlard. Comparing Different Acoustic Modeling Techniques for Multilingual Boosting. In *Proceedings of Interspeech*, September 2012a.

D. Imseng, P. Motlicek, P. Garner, and H. Bourlard. Impact of Deep MLP Architecture on Different Acoustic Modeling Techniques for Under-Resourced Speech Recognition. In *Proceedings of ASRU*, December 2013.

D. Imseng et al. MediaParl: Bilingual Mixed Language Accented Speech Database. In *Proceedings of IEEE Workshop on Spoken Language Technology*, pages 263–268, December 2012b.

R. Jakobson, G. Fant, and M. Halle. *Preliminaries to Speech Analysis: the Distinctive Features and their Correlates*. MIT Press, 1992.

A. Janin, D. Ellis, and N. Morgan. Multi-Stream Speech Recognition: Ready for Prime Time? In *Proceedings of EUROSPEECH*. ISCA, 1999.

A. Jansen and K. Church. Towards Unsupervised Training of Speaker Independent Acoustic Models. In *Proceedings of Interspeech*, pages 1693–1692, 2011.

S. Jiampojamarn, G. Kondrak, and T. Sherif. Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 372–379, April 2007.

D. Johnson et al. ICSI Quicknet Software Package. http://www.icsi.berkeley.edu/Speech/qn.html, 2004.

D. Jouvet, D. Fohr, and I. Illina. Evaluating Grapheme-to-Phoneme Converters in Automatic Speech Recognition Context. In *Proceedings of ICASSP*, pages 4821–4824, 2012.

B-H Juang and Sh. Katagiri. Discriminative learning for minimum error classification (pattern recognition). *IEEE Transactions on Signal Processing*, 40(12):3043–3054, 1992.

B-H. Juang and L. R. Rabiner. The Segmental K-means Algorithm for Estimating Parameters of Hidden Markov Models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38: 1639–1641, 1990.

J. Kaiser, B. Horvat, and Z. Kacic. A Novel Loss Function for the Overall Risk Criterion based Discriminative Training of HMM Models. In *Proceedings of ICSLP*, 2000.

S. Kanthak and H. Ney. Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition. In *Proceedings of ICASSP*, pages 845–848, 2002.

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 Transactions on Acoustics, Speech, and Signal Processing*, 35(3): 400–401, 1987.

M. Killer, S. Stüker, and T. Schultz. Grapheme-Based Speech Recognition. In *Proceedings of Eurospeech*, pages 3141–3144, 2003.

R. Kneser and H. Ney. Improved Backing-Off for M-Gram Language Modeling. In *Proceedings of ICASSP*, volume 1, pages 181–184. IEEE, 1995.

T. Ko and B. Mak. Eigentrigraphemes for Under-Resourced Languages. *Speech Communication*, 56:132–141, 2014.

S. Kullback. Letter to the Editor: The Kullback-Leibler Distance. 1987.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 03 1951.

P. Ladefoged. *A Course in Phonetics*. Harcourt Brace College Publishers, 1993.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML*, pages 282–289, 2001.

C. Lee, Y. Zhang, and J. R. Glass. Joint learning of phonetic units and word pronunciations for ASR. In *Proceedings of EMNLP*, pages 182–192, 2013.

C-H Lee, F. K. Soong, and B-H Juang. A Segment Model-Based Approach to Speech Recognition. In *Proceedings of ICASSP*, 1988.

C-Y Lee, T. J. O'Donnell, and J. Glass. Unsupervised Lexicon Discovery From Acoustic Input. *Transactions of the Association for Computational Linguistics*, 3:389–403, 2015.

P. Lehnen, S. Hahn, A. Guta, and H. Ney. Incorporating Alignments Into Conditional Random Fields for Grapheme to Phoneme Conversion. In *Proceedings of ICASSP*, pages 4916–4919, 2011.

V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

K. Livescu, E. Fosler-Lussier, and F. Metze. Subword Modeling for Automatic Speech Recognition: Past, Present, and Emerging Approaches. *IEEE Signal Processing Magazine*, 29(6): 44–57, 2012.

A. Ljolje. High Accuracy Phone Recognition using Context Clustering and Quasi-Triphonic Models. *Computer Speech & Language*, 8(2):129–151, 1994.

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

X. Luo and F. Jelinek. Probabilistic Classification of HMM States for Large Vocabulary Continuous Speech Recognition. In *Proceedings of ICASSP*, volume 1, pages 353–356. IEEE, 1999.

M. Magimai.-Doss and H. Bourlard. On the Adequacy of Baseform Pronunciations and Pronunciation Variants. In *Proceedings of the First International Conference on Machine Learning for Multimodal Interaction*, MLMI'04, pages 209–222, 2005.

M. Magimai.-Doss, R. Rasipuram, G. Aradilla, and H. Bourlard. Grapheme-Based Automatic Speech Recognition using KL-HMM. In *Proceedings of Interspeech*, August 2011a.

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

S. Maskey, A. W. Black, and L. Tomokiya. Boostrapping Phonetic Lexicons for New Languages. In *Proceedings of Interspeech*, 2004.

I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, et al. The AMI Meeting Corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, 2005.

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

T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur. Recurrent Neural Network Based Language Model. In *Proceedings of Interspeech*, volume 2, page 3, 2010.

H. Misra, H. Bourlard, and V. Tyagi. New Entropy Based Combination Rules in HMM/ANN Multi-Stream ASR. In *Proceedings of ICASSP*, 2003.

H. Mokbel and D. Jouvet. Derivation of the Optimal Set of Phonetic Transcriptions for a Word from its Acoustic Realizations. *Speech Communication*, 29(1):49 – 64, 1999.

N. Morgan and H. Bourlard. Generalization and Parameter Estimation in Feedforward Nets: Some Experiments. In *Proceedings of NIPS*, pages 630–637. MIT Press, 1989.

N. Morgan and H. Bourlard. Continuous Speech Recognition using Multilayer Perceptrons with Hidden Markov Models. In *Proceedings of ICASSP*, pages 413–416, 1990.

N. Morgan and H. Bourlard. Continuous Speech Recognition: An Introduction to the Hybrid HMM/Connectionist Approach. *IEEE Signal Processing Magazine*, pages 25–42, 1995.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an Algorithm. In *Proceedings of NIPS*, pages 849–856, 2001.

J. R. Novak, N. Minematsu, and K. Hirose. WFST-Based Grapheme-to-Phoneme Conversion: Open Source tools for Alignment, Model-Building and Decoding. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pages 45–49, July 2012.

J. P. Openshaw, Z. P. Sun, and J. S. Mason. A Comparison of Composite Features under Degraded Speech in Speaker Recognition. In *Proceedings of ICASSP*, pages 371–374, 1993.

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

D. Palaz, R. Collobert, and M. Magimai.-Doss. Estimating Phoneme Class Conditional Probabilities from Raw Speech Signal using Convolutional Neural Networks. In *Proceedings of Interspeech*, August 2013.

KK. Paliwal. Lexicon-building methods for an acoustic sub-word based speech recognizer. In *Proceedings of ICASSP*, pages 729–732, 1990.

A. S. Park and J. R. Glass. Unsupervised Pattern Discovery in Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):186–197, 2008.

D. B. Paul and J. M. Baker. The Design for the Wall Street Journal-Based CSR Corpus. In *Proceedings of WSNL*, pages 357–362, 1992.

J. Pitrelli, C. Fong, S.H. Wong, J.R. Spitz, and H.C. Leung. PhoneBook: a Phonetically-Rich Isolated-Word Telephone-Speech Database. In *Proceedings of ICASSP*, volume 1, pages 101–104, 1995.

D. Povey et al. The Kaldi Speech Recognition Toolkit. In *Proceedings of ASRU*, 2011.

D. M. W. Powers. Applications and Explanations of Zipf's Law. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, pages 151–160, 1998.

P. Price, W. M Fisher, J. Bernstein, and D. S. Pallett. The DARPA 1000-word Resource Management Database for Continuous Speech Recognition. In *Proceedings of ICASSP*, pages 651–654. IEEE, 1988.

J. Psutka, L. Müller, and J. V. Psutka. Comparison of MFCC and PLP Parameterizations in the Speaker Independent Continuous Speech Recognition Task. In *Proceedings of EUROSPEECH*, pages 1813–1816, 2001.

L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of IEEE*, 77(2):257–286, 1989.

K. Rao, F. Peng, H. Sak, and F. Beaufays. Grapheme-to-Phoneme Conversion using Long Short-Term Memory Recurrent Neural Networks. In *Proceedings of ICASSP*, pages 4225–4229, 2015.

R. Rasipuram and M. Magimai.-Doss. Acoustic Data-Driven Grapheme-to-Phoneme Conversion Using KL-HMM. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (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 *Proceedings of Interspeech*, 2012b.

R. Rasipuram and M. Magimai.-Doss. Improving Grapheme-Based ASR by Probabilistic Lexical Modeling Approach. In *Proceedings of Interspeech*, 2013a.

R. Rasipuram and M. Magimai.-Doss. Probabilistic Lexical Modeling and Grapheme-Based Automatic Speech Recognition. Idiap-RR Idiap-RR-15-2013, Idiap, 4 2013b.

R. Rasipuram and M. Magimai.-Doss. Acoustic and lexical resource constrained asr using language-independent acoustic model and language-dependent probabilistic lexical model. *Speech Communication*, 68:23–40, April 2015.

R. Rasipuram and M. Magimai.-Doss. Articulatory Feature Based Continuous Speech Recognition using Probabilistic Lexical Modeling. *Computer Speech and Language*, 36:233–259, 2016.

R. Rasipuram, P. Bell, and M. Magimai.-Doss. Grapheme and Multilingual Posterior Features for Under-Resourced Speech Recognition: A Study on Scottish Gaelic. In *Proceedings 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 *Proceedings of ASRU*, 2013b.

R. Rasipuram, M. Razavi, and M. Magimai.-Doss. Integrated Pronunciation Learning for Automatic Speech Recognition Using Probabilistic Lexical Modeling. In *Proceedings of ICASSP*, pages 5176–5180, 2015.

M. Razavi and M. Magimai.-Doss. On Recognition of Non-Native Speech Using Probabilistic Lexical Model. In *Proceedings of Interspeech*, 2014.

M. Razavi and M. Magimai.-Doss. An HMM-Based Formalism for Automatic Subword Unit Derivation and Pronunciation Generation. In *Proceedings of ICASSP*, 2015.

M. Razavi and M. Magimai.-Doss. A Posterior-Based Multi-Stream Formulation for G2P Conversion. *IEEE Signal Processing Letters*, 24(4):475–479, 2017.

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 *Proceedings of ICASSP*, 2014.

M. Razavi, R. Rasipuram, and M. Magimai.-Doss. Towards Multiple Pronunciation Generation in Acoustic G2P Conversion Framework. Idiap-RR Idiap-RR-34-2015, Idiap, 10 2015a.

M. Razavi, R. Rasipuram, and M. Magimai.-Doss. Pronunciation Lexicon Development for Under-Resourced Languages Using Automatically Derived Subword Units: A Case Study on Scottish Gaelic. In *Proceedings of 4th Biennial Workshop on Less-Resourced Languages*, 2015b.

M. Razavi, R. Rasipuram, and M. Magimai.-Doss. Acoustic Data-Driven Grapheme-to-Phoneme Conversion in the Probabilistic Lexical Modeling Framework. *Speech Communication*, 80, 2016.

M. Riley. A statistical model for generating pronunciation networks. In *Proceedings of ICASSP*, volume 2, pages 737–740, 1991.

T. Robinson, M. Hochberg, and S. Renals. IPA: Improved Phone Modelling with Recurrent Neural Networks. In *Proceedings of ICASSP*, volume 1, pages I–37. IEEE, 1994.

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

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Neurocomputing: Foundations of research. chapter Learning Representations by Back-propagating Errors, pages 696–699. 1988.

T. N Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran. Deep convolutional neural networks for LVCSR. In *Proceedings of ICASSP*, pages 8614–8618. IEEE, 2013.

T. Schlippe, W. Quaschningk, and T. Schultz. Combining Grapheme-to-Phoneme Converter Outputs for Enhanced Pronunciation Generation in Low-Resource Scenarios. In *Proceedings of SLTU*, pages 139–145, 2014.

R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul. Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech. In *Proceedings of ICASSP*, pages 1205–1208, 1985.

T. J. Sejnowski and C. R. Rosenberg. Parallel Networks that Learn to Pronounce English Text. *Complex Systems*, 1:145–168, 1987.

J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

M. L. Shire. Relating Frame Accuracy with Word Error in Hybrid ANN-HMM ASR. In *Proceedings of Interspeech*, pages 1797–1800, 2001.

R. Singh, B. Raj, and R. M. Stern. Automatic Generation of Phone Sets and Lexical Transcriptions. In *Proceedings of ICASSP*, pages 1691–1694, 2000.

R. Singh, B. Raj, and R. M. Stern. Automatic Generation of Subword Units for Speech Recognition Systems. *IEEE Transactions on Speech and Audio Processing*, 10(2):89–99, 2002.

S. Soldo, M. Magimai.-Doss, and H. Bourlard. Synthetic References for Template-based ASR using Posterior Features. In *Proceedings of Interspeech*, 2012.

H. Strik and C. Cucchiarini. Modeling Pronunciation Variation for ASR: A Survey of the Literature. *Speech Communication*, 29(2-4):225–246, 1999.

Y. Sun et al. Combination of Sparse Classification and Multilayer Perceptron for Noise Robust ASR. In *Proceedings of Interspeech*, 2012.

T. Svendsen, KK. Paliwal, E. Harborg, and P. Husoy. An Improved Subword-Based Speech Recognizer. In *Proceedings of ICASSP*, pages 108–111, 1989.

T. Svendsen, F. K. Soong, and H. Purnhagen. Optimizing Baseforms for HMM-Based Speech Recognition. In *Proceedings of EUROSPEECH*, 1995.

D. M.J. Tax, M. van Breukelen, R. P.W. Duin, and J. Kittler. Combining Multiple Classifiers by Averaging or by Multiplying? *Pattern Recognition*, 33(9):1475 – 1485, 2000. ISSN 0031-3203.

P. Taylor. Hidden Markov Models for Grapheme to Phoneme Conversion. In *Proceedings of Interspeech*, pages 1973–1976, 2005.

P. Taylor, A. Black, and R. Caley. The Architecture of the Festival Speech Synthesis System. In *Proceedings of ESCA Workshop on Speech Synthesis*, 1998.

Z. Tüske, P. Golik, R. Schlüter, and H. Ney. Acoustic Modeling with Deep Neural Networks using Raw Time Signal for LVCSR. In *Proceedings of Interspeech*, pages 890–894, 2014.

F. Valente. Multi-Stream Speech Recognition Based on Dempster-Shafer Combination Rule. *Speech Communication*, 52(3):213–222, 2010.

E. Variani, F. Li, and H. Hermansky. Multi-Stream Recognition of Noisy Speech with Performance Monitoring. In *Proceedings of Interspeech*, 2013.

R. Veldhuis. The Centroid of the Symmetrical Kullback-Leibler Distance. *IEEE Signal Processing Letters*, 9(3):96–99, 2002.

O. Vinyals, S. V Ravuri, and D. Povey. Revisiting Recurrent Neural Networks for Robust ASR. In *Proceedings of ICASSP*, pages 4085–4088. IEEE, 2012.

N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard. Multilingual Deep Neural Network based Acoustic Modeling For Rapid Language Adaptation. In *Proceedings of ICASSP*, pages 7639 – 7643. IEEE, May 2014.

A. Waibel et al. Phoneme Recognition Using Time-Delay Neural Networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339, 1989.

D. Wang and S. King. Letter-to-Sound Pronunciation Prediction Using Conditional Random Fields. *IEEE Signal Processing Letters*, 18(2):122–125, 2011.

G. Williams and S. Renals. Confidence Measures from Local Posterior Probability Estimates. *Computer Speech & Language*, 13(4):395–411, 1999.

P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young. Large Vocabulary Continuous Speech Recognition Using HTK. In *Proceedings of ICASSP*, pages 125–128, 1994.

K. Wu, C. Allauzen, K. B Hall, M. Riley, and B. Roark. Encoding Linear Models as Weighted Finite-State Transducers. In *Proceedings of Interspeech*, pages 1258–1262, 2014.

L. Xiao, A. Gunawardana, and A. Acero. Adapting Grapheme-to-Phoneme Conversion for Name Recognition. In *Proceedings of ASRU*, pages 130–135, 2007.

K. Yao and G. Zweig. Sequence-to-Sequence Neural Net Models for Grapheme-to-Phoneme Conversion. *arXiv preprint arXiv:1506.00196*, 2015.

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*, pages 307–312, 1994.

S.J. Young. The General Use of Tying in Phoneme-Based HMM Speech Recognisers. In *Proceedings of ICASSP*, volume 01, pages 569–572, 1992.

S.J. Young et al. *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, UK, 2006.

Avenue de l'Eglise-Anglaise 14
1006 Lausanne, Switzerland
+41 (78) 975 0084
+41 (27) 721 7711
marzieh.razavi@idiap.ch
marziehrazavi.github.io
marziehrazavi

Marzieh RAZAVI

Education

Mar. 2013 – Present **PhD**, *École Polytechnique Fédérale de Lausanne*, Lausanne, Switzerland, Electrical Engineering.
Supervised by Professor Hervé Bourlard and Dr. Mathew Magimai.-Doss

Aug. 2012 **M.Sc.**, *Simon Fraser University*, Vancouver, Canada, *Computing Science*.
Supervised by Professor Anoop Sarkar

Sep. 2009 **B.Sc.**, *Sharif University of Technology*, Tehran, Iran, *Computer Engineering*.
Supervised by Dr. Hossein Sameti

Professional Experience

Mar. 2013 – Present **Doctoral Researcher**, *Idiap Research Institute*, Martigny, Switzerland.

Sep. 2010 – Dec. 2012 **Research Assistant**, *Natural Language Lab - Simon Fraser University*, Vancouver, Canada.
- Worked on ensembling diverse clustering-based dependency parsers in MSTParser framework.
- Worked on left-to-right target generation for hierarchical phrase-based translation.

Sep. 2010 – Jun. 2012 **Teaching Assistant**, *Simon Fraser University*, Vancouver, Canada.
- Computer architecture course.

Mar. 2009 – Jan. 2010 **R&D Engineer**, *ASR Gooyesh Pardaz Ltd.*, Tehran, Iran.
- Designed a robust natural language understanding component for a flight information system.

Jul. 2008 – Sep. 2008 **R&D Intern**, *ASR Gooyesh Pardaz Ltd.*, Tehran, Iran.
- Designed a robust parser for human-machine dialogue for banking systems.

Selected Skills

Programming languages	Java, Python, R, C++, MATLAB	Software toolkits	QuickNet, HTK, CRF++, MSTParser, Moses
Operating Systems	Linux, Mac OS, Microsoft Windows	Typesetting	LaTeX

Languages

Persian	Native	English	Full professional proficiency
French	Elementary proficiency	Arabic	Elementary proficiency

Publications

Peer-reviewed Journals

[1] **Marzieh Razavi** and Mathew Magimai.-Doss. A Posterior-Based Multi-Stream Formulation for G2P Conversion. *IEEE Signal Processing Letters*, February 2017.

[2] **Marzieh Razavi**, Ramya Rasipuram and Mathew Magimai.-Doss. Acoustic Data-Driven Grapheme-to-Phoneme Conversion in the Probabilistic Lexical Modeling Framework. *Speech Communication*, April 2016.

Peer-reviewed Conferences

[1] **Marzieh Razavi** and Mathew Magimai.-Doss. Improving Under-Resourced Language ASR through Latent Subword Unit Space Discovery. In *INTERSPEECH*, San Francisco, USA, September 2016.

[2] **Marzieh Razavi**, Ramya Rasipuram and Mathew Magimai.-Doss. Pronunciation Lexicon Development for Under-Resourced Languages Using Automatically Derived Subword Units: A Case Study on Scottish Gaelic. Winner of a *Best Student Paper Award* in *Language and Technology Conference (LTC)*, Poznan, Poland, November 2015.

[3] **Marzieh Razavi** and Mathew Magimai.-Doss. An HMM-Based Formalism for Automatic Subword Unit Derivation and Pronunciation Generation. In *IEEE ICASSP*, Brisbane, Australia, April 2015.

[4] Ramya Rasipuram, **Marzieh Razavi** and Mathew Magimai.-Doss. Integrated Pronunciation Learning for Automatic Speech Recognition Using Probabilistic Lexical Modeling. In *IEEE ICASSP*, Brisbane, Asutralia, April 2015.

[5] **Marzieh Razavi** and Mathew Magimai.-Doss. On Recognition of Non-Native Speech Using Probabilistic Lexical Model. In *INTERSPEECH*, Singapore city, Singapore, September 2014.

[6] **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 *IEEE ICASSP*, Florence, Italy, April 2014.

[7] Ramya Rasipuram, **Marzieh Razavi** and Mathew Magimai.-Doss. Probabilistic Lexical Modeling and Unsupervised Training for Zero-Resourced ASR. In *IEEE ASRU*, Olomouc, Czech Republic, December 2013.

[8] Gholamreza Haffari, **Marzieh Razavi** and Anoop Sarkar. An Ensemble Model that Combines Syntactic and Semantic Clustering for Discriminative Dependency Parsing. In *ACL*, Portland, USA, June 2011.

Research Reports

[1] **Marzieh Razavi**, Ramya Rasipuram and Mathew Magimai.-Doss. Towards Multiple Pronunciation Generation in Acoustic G2P Conversion Framework. `http://publications.idiap.ch/downloads/reports/2015/Razavi_Idiap-RR-34-2015.pdf`, *Idiap Research Institute*, October 2015.

Manuscripts under Submission

[1] **Marzieh Razavi**, Ramya Rasipuram and Mathew Magimai.-Doss. Towards Weakly Supervised Acoustic Subword Unit Discovery and Lexicon Development Using Hidden Markov Models. `http://publications.idiap.ch/downloads/reports/2016/Razavi_Idiap-RR-15-2017.pdf`, *Submitted to Speech Communication*, March 2017.

Honors & Awards

Nov. 2015 **Received one of the Best Student Paper Awards**, *Language and Technology Conference*, Poznan, Poland.

Jun. 2014 **Received the ISCA Student Travel Grant to Attend INTERSPEECH 2014**, *International Speech communication Association (ISCA)*.

Sep. 2011 **Recipient of SFU Graduate Fellowship**, *Simon Fraser University*, Vancouver, Canada.

Summer 2005 **Ranked 258th**, *Among more than 300000 students in the nationwide university entrance exam*, Tehran, Iran.

Activities

Professional Activities: ISCA, ACL, IEEE and IEEE SPS student member.

Extracurricular Activities: Poetry, Literature and Calligraphy.