IDIAP RESEARCH REPORT

ACOUSTIC DATA-DRIVEN GRAPHEME-TO-PHONEME CONVERSION IN THE PROBABILISTIC LEXICAL MODELING FRAMEWORK

Marzieh Razavi Ramya Rasipuram
Mathew Magimai.-Doss

Idiap-RR-10-2015

MAY 2015

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Acoustic Data-Driven Grapheme-to-Phoneme Conversion in the Probabilistic Lexical Modeling Framework

Marzieh Razavi[a,b,*], Ramya Rasipuram[a], Mathew Magimai Doss[a]

[a] *Idiap Research Institute, CH-1920 Martigny, Switzerland*
[b] *Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*

Abstract

One of the primary steps in building automatic speech recognition (ASR) as well as text-to-speech systems is development of a phonemic lexicon that provides a mapping between each word and its pronunciation as a sequence of phonemes. Phoneme lexicons can be developed by humans through use of linguistic knowledge, however, this would be a costly and time-consuming task. To facilitate this process, grapheme-to-phoneme conversion (G2P) techniques are used in which given an initial phoneme lexicon, the relationship between graphemes and phonemes is learned through data-driven methods. This article presents a novel G2P formalism which learns the grapheme-to-phoneme relationship through acoustic data and potentially relaxes the need for an initial phonemic lexicon in the target language of interest. The formalism involves a training part followed by an inference part. In the training part, the grapheme-to-phoneme relationship is captured in a probabilistic lexical modeling framework. Within this framework, a hidden Markov model (HMM) is trained in which each HMM state representing a grapheme is parameterized by a categorical distribution of phonemes. Then in the inference part, given the orthographic transcription of the word and the learned HMM, the most probable sequence of phonemes is inferred. In this article, we show that the recently proposed acoustic G2P approach in the Kullback-Leibler divergence-based HMM (KL-HMM) framework is a particular case of this formalism. We then benchmark the approach against two popular G2P approaches, namely joint multigram approach and decision tree-based approach. Our experimental studies on English and French show that despite relatively poor performance at the pronunciation level, the proposed approach can achieve comparable performance to state-of-the-art G2P methods at the ASR level.

Keywords: grapheme-to-phoneme conversion, probabilistic lexical modeling framework, Kullback-Leibler divergence-based hidden Markov model,

*Corresponding author
Email addresses: marzieh.razavi@idiap.ch (Marzieh Razavi), ramya.rasipuram@idiap.ch (Ramya Rasipuram), mathew@idiap.ch (Mathew Magimai Doss)

1. Introduction

Speech technology systems such as automatic speech recognition (ASR) and text-to-speech (TTS) systems aim to link two modes of communications, namely the spoken form (speech) and the written form (text). In order to model the relation between the two forms, a shared unit is commonly used. The shared units can typically be the whole words or subword units. However, subword units are preferred to words especially in large vocabulary tasks for two main reasons: 1) they are more trainable compared to the whole words themselves as the frequency of words in a text follow the Zipf's law[1](Powers, 1998), and 2) they are generalizable for the unseen words.

The most widely used subword units in current speech processing systems are phonemes[2] which can be related to both spoken and written forms. More precisely, phonemes are related to the speech signal as the envelope of magnitude spectrum of short-term speech signals typically depicts the characteristics of phonemes. They can be also related to the written symbols (i.e., graphemes). The link between phonemes and graphemes originates from the purpose of alphabetic orthographies which aim to present the phonetic structure of the spoken words in the graphic form (Frost, 1989). The relationship between the graphemes and phonemes can be shallow or deep depending on the language.

Typically, the development of phoneme-based speech technology systems consists of two steps: development of a lexicon consisting of the mapping between each word and its phoneme-based pronunciation followed by system training. The focus of this article is mainly on the first step (i.e., phonemic lexicon development). A phonemic lexicon can be developed manually through use of linguistic knowledge. However, manual development of lexicon 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 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. Towards that goal, grapheme-to-phoneme conversion (G2P) methods are applied in which given an initial phonemic lexicon called a *seed lexicon* as the training data, typically data-driven techniques such as decision trees (Black et al., 1998) or conditional random fields (Wang and King, 2011) are used to learn the grapheme-to-phoneme relationship. Most of these approaches rely solely on

[1]According to Zipf's law, the frequency of a word is inversely proportional to its rank in the frequency table.

[2]From the linguistics point of view, phonemes and phones are two different terminologies. Phonemes are "the smallest contrastive linguistic units which may bring about a change of meaning" (Chomsky and Halle, 1968) in a specific language while phones are units of the speech sounds which can be designed to cover the set of sounds in all languages. For the sake of clarity, throughout this article we use the term phoneme as in the literature the grapheme-to-phoneme terminology is more dominantly used.

the seed lexicon for learning the grapheme-to-phoneme relationship while no acoustic information is incorporated within the G2P process.

This article presents a novel G2P formalism in which the grapheme-to-phoneme relationship is learned through acoustic data. The formalism consists of two phases: a training phase and an inference phase. In the training phase, as the first step the relationship between acoustic feature observations and phonemes is learned through an acoustic model, such as an artificial neural network (ANN). The acoustic model can be trained on target language or language-independent data if the phonemic lexicon in the target language is not available. Then as the second step the relationship between the graphemes and phonemes is learned in a hidden Markov model (HMM) framework in which the outputs of the acoustic model are used as feature observations. In this HMM framework, each state represents a grapheme and is parameterized by a categorical distribution of phonemes. In the inference phase, given the orthographic transcription of the word, the grapheme-based HMM acts as a generative model and emits a sequence of phoneme posterior probabilities. The sequence of phoneme posterior probabilities is then decoded using an HMM in which each state represents a phoneme to infer the most probable pronunciation for each word.

In this article, we show that the recently proposed acoustic data-driven G2P approach in the framework of Kullback-Leibler divergence-based HMM (KL-HMM) (Rasipuram and Magimai-Doss, 2012a) is a particular case of this G2P formalism. We then build upon the previous studies on the acoustic G2P approach and study possible ways to refine the method by incorporating recent trends in ANN including using ANNs with more layers and output units. Furthermore, we benchmark the approach against two popular conventional G2P approaches, namely the joint multigram and the decision tree-based methods. We evaluate the proposed G2P approach at both pronunciation and application (ASR) levels. For the evaluation at the ASR level, we study different facets including combining the proposed G2P approach with conventional G2P approaches.

Our experimental studies on English and Swiss French show that the proposed approach can achieve comparable performance to the state-of-the-art G2P approaches at the ASR level. In addition, through combining the acoustic G2P approach with conventional G2P approaches improvements in the ASR performance can be achieved, in particular when limited amount of training data is available.

This article is organized as follows. Section 2 provides a background about the existing approaches for pronunciation generation in the literature. Section 3 proposes the novel G2P formalism for learning the grapheme-to-phoneme relationship through acoustic data and compares the acoustic G2P approach with the existing G2P methods. Section 4 describes the databases together with the experimental setups used in this study. Section 5 presents the pronunciation level results and analysis. Section 6 provides the experimental results at the ASR level. Finally Section 7 brings the conclusion.

2. Relevant literature

As explained briefly in Section 1, the first step towards building phoneme-based speech technology systems is development of a phonemic lexicon. Phonemic pronunciations can be typically obtained by humans through exploiting the linguistic knowledge. During the preparation of the pronunciation lexicon by linguists, care is taken to minimize word level confusions and consistency is ensured across the lexicon. The hand crafted phoneme pronunciation lexicon could possibly provide the optimum performance for ASR or TTS. However, design of the phonemic pronunciation lexicon of significant size by linguistic experts is a tedious and costly task. Furthermore, a finite lexicon will always have limited coverage for ASR and TTS synthesis systems. For this reason, ASR and TTS systems use G2P methods when hand crafted pronunciations fail to cover the vocabulary of a particular domain. In this section, we first elucidate two classes of G2P methods, namely knowledge-based and data-driven approaches, which have been explored in the literature. We then explain some of the limitations of the automatic grapheme-to-phoneme converters which have been addressed in the literature through pruning or weighing the pronunciation variants using acoustic data.

2.1. Knowledge-based approaches

Knowledge-based G2P approaches exploit rules derived by humans or from linguistic studies to convert the sequence of graphemes in a word to a sequence of phonemes. Rule-based G2P approaches are typically formulated in the framework of finite state automata (Kaplan and Kay, 1994). The primary advantage of rule-based approaches is that they provide complete coverage. However, as natural languages exhibit irregularities, it is necessary to cross-check if the rules are applicable to all the entries. Often rule-based G2P systems also need an exception list. Furthermore, design of rules requires specific linguistic skills that may not be always available. In order to reduce the amount of human effort and linguistic knowledge, data-driven approaches are usually employed.

2.2. Data-driven G2P approaches

Data-driven approaches for G2P predict the pronunciation of an unseen word based on the examples in the training data (i.e., the seed lexicon). Typically the G2P 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 phonemes (Damper et al., 2005; Jiampojamarn et al., 2007). In the second step, a learning method is employed to capture the grapheme-to-phoneme 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 approaches[3]. Therefore, what distinguishes different G2P approaches from each other is the learning and inference methods utilized. Among various G2P approaches proposed based on different techniques, local classification-based (Sejnowski and Rosenberg, 1987; Black et al., 1998; Pagel et al., 1998) and probabilistic sequence modeling-based approaches (Taylor, 2005; Bisani and Ney, 2008; Wang and King, 2011) have gained wide attention:

- *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 grapheme-to-phoneme relationship from the training data. For the inference part, the sequence of input graphemes are processed sequentially in which for each grapheme, the corresponding phoneme (or phoneme sequence) is locally generated. Therefore, we refer to these methods as local classification-based techniques.

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

$$F^* = \arg\max_F P(F|G) \tag{1}$$

$$= \arg\max_F P(F, G) \tag{2}$$

  where given a sequence of graphemes G, the goal is to search for a sequence of phonemes F^* that maximizes the posterior probability $P(F|G)$. Equation 1 can also be expressed as finding a sequence of phonemes F^* maximizing the joint probability $P(F, G)$ using Bayes rule (Equation (2)). Based on these expressions, different approaches have been explored:

  1. *HMM-based approach*: In (Taylor, 2005), the G2P problem is formulated in the standard HMM way by applying independent and identically distributed (i.i.d.) and first order Markov model assumptions as:

$$S^* = \arg\max_S P(G|S)P(S) \tag{3}$$

$$= \arg\max_S \prod_n P(g_n|s_n)P(s_n|s_{n-1}) \tag{4}$$

     where $S = [s_1, \ldots, s_n, \ldots, s_N]$ represents the hidden sequence of phonemes and $G = [g_1, \ldots, g_n, \ldots, g_N]$ denotes the sequence of grapheme observations. In this framework, each HMM represents a phoneme which emits (up to four) grapheme symbols. As opposed to local classification approaches in which the alignments are obtained as a pre-processing step,

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

in this framework the alignments can be derived during the Baum-Welch training. For the inference, the most probable sequence of phonemes that could have generated the input grapheme sequence is inferred using the Viterbi algorithm.

2. *Joint multigram approach*: In joint multigram or joint n-gram approaches, the joint probability of pairs of grapheme sequences and phoneme sequences is obtained based on the concept of graphones (Bisani and Ney, 2008). A graphone is a pair of a sequence of graphemes and a sequence of phonemes. Figure 1 shows the sequence of graphones for the word *phone* along with its pronunciation.

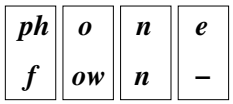


Figure 1: A possible sequence of graphones for the word *phone* and its associated pronunciation.

The joint probability $P(F, G)$ of a sequence of graphemes G and a sequence of phonemes F in Equation (2) is obtained by summing over matching alignments which are derived from sequences of graphones in the space of all possible sequence of graphones for the (F, G) pair.

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 phonemes can be derived by using the Viterbi algorithm. In (Novak et al., 2012), the inference of the best sequence of phonemes can be done in the weighted finite state transducer (WFST) framework.

3. *Conditional random field-based approach*: In conditional random field (CRF)-based approaches, the conditional probability $P(F|G)$ in Equation (1) is modeled using a log-linear representation (Wang and King, 2011; Lehnen et al., 2011). The CRF model is a discriminative model which can perform global inference. Therefore, it can exploit the advantages of both decision tree-based and joint multigram methods. 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 phoneme sequence is inferred using the Viterbi algorithm. In (Hahn et al., 2013), hidden conditional random fields (HCRFs) are used for the G2P task in which the alignment between the grapheme and phoneme sequence is modeled via a hidden variable.

2.3. Pronunciation extraction using acoustic data

The pronunciations derived from automatic grapheme-to-phoneme converters reflect the ambiguity and variation found in the lexical resources used to train the model. Therefore, the pronunciations or their variants may not reflect the natural phonological variation. For example, this can happen when a grapheme-to-phoneme converter trained on native pronunciations is used to extend the vocabulary of a non-native ASR system; or when the new vocabulary has unusual words.

To overcome this limitation, in the context of pronunciation variation modeling speech, spoken examples of words are used to obtain pronunciation variants (Strik and Cucchiarini, 1999; Cohen, 1989; Fosler-Lussier, 2000; Mokbel and Jouvet, 1999; Magimai-Doss and Bourlard, 2005). Most often, automatic phoneme transcriptions of spoken examples obtained from a phoneme 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 phoneme recognizer to obtain phoneme transcriptions of words. Possible alternate phoneme sequences for words are obtained by finding a best alignment between the output of the phoneme recognizer and pronunciations provided by the seed lexicon. The learned pronunciations of the words are collected in a dictionary.

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. The 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 (Riley, 1991). Figure 2 illustrates the typical pronunciation variant extraction process.

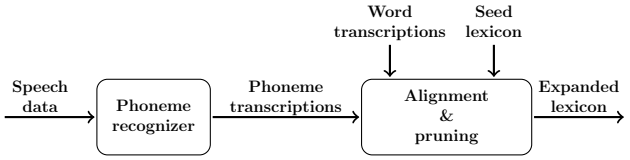


Figure 2: Pronunciation lexicon expansion with possible pronunciation variants for words obtained using speech samples.

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 interest to prune the pronunciation variants obtained through a grapheme-to-phoneme converter using the spoken word examples.

- In (McGraw et al., 2013), the pronunciation variants of words given by the graphone-based G2P approach (Bisani and Ney, 2008) are given pronunciation

weights using acoustic samples of words. The approach assumes that an expert provided pronunciation lexicon is available.

- In (Lu et al., 2013), an approach to enlarge the expert phonetic lexicon is proposed where the pronunciations of additional words are generated using their acoustic samples and a trained grapheme-to-phoneme converter. More precisely, first a grapheme-to-phoneme converter is trained using an expert lexicon. The grapheme-to-phoneme converter is used to generate pronunciation variants for new words. The weights for these multiple pronunciations are estimated based on acoustic evidence using the WFST-based EM algorithm. Finally, the acoustic model is updated using the augmented lexicon. The process is repeated until convergence.

As shown in Figure 3, the above two G2P approaches rely on a seed lexicon and a G2P converter. The acoustic samples are used only to weigh or select the alternate pronunciations given by a grapheme-to-phoneme converter.

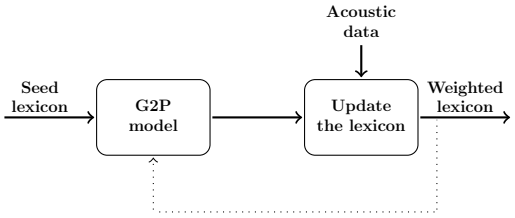


Figure 3: Acoustic data-driven G2P approaches proposed in the literature. The dotted line illustrates that some approaches iterate the G2P process.

In addition to the aforementioned approaches, other techniques have also been developed for adapting the G2P methods using spoken samples. In (Xiao et al., 2007), the parameters of the grapheme-to-phoneme converter are adapted using spoken examples for a name recognition task.

3. Acoustic G2P approach using probabilistic lexical modeling

In this section, we first present a novel G2P formalism which incorporates acoustic information to learn grapheme-to-phoneme relationship and demonstrate that the acoustic data-driven G2P approach in the KL-HMM framework is a particular case of this formalism. We then compare the acoustic G2P approach with other existing approaches in the literature.

3.1. Theoretical formulation

Given a sequence of graphemes $G = [g_1, \ldots, g_n, \ldots, g_N]$, the G2P problem in an HMM-based framework can be expressed as finding the most probable

phoneme sequence F^* that can be achieved by finding the most likely state sequence S^*:

$$S^* = \arg\max_{S \in \mathcal{S}} P(G, S|\Theta) \tag{5}$$

$$= \arg\max_{S \in \mathcal{S}} P(G|S, \Theta) P(S|\Theta) \tag{6}$$

where Θ denotes the parameters of the system, $\mathcal{S}$ denotes the set of possible HMM state sequences and $S = [s_1, \cdots, s_n, \cdots, s_N]$ denotes a sequence of HMM states which corresponds to a phoneme sequence hypothesis with $s_n \in \mathcal{F} = \{f_1, \ldots, f_k, \ldots, f_K\}$ where K is the number of phoneme units. By applying i.i.d. and first order Markov assumption, Equation 6 can be simplified as:

$$S^* = \arg\max_{S \in \mathcal{S}} \prod_{n=1}^{N} P(g_n|s_n = f_k, \Theta) P(s_n = f_k|s_{n-1} = f_{k'}, \Theta) \tag{7}$$

We can then write Equation 7 by applying the Bayes rule as follows:

$$S^* = \arg\max_{S \in \mathcal{S}} \prod_{n=1}^{N} \frac{P(s_n = f_k|g_n, \Theta) P(g_n|\Theta)}{P(s_n = f_k|\Theta)} P(s_n = f_k|s_{n-1} = f_{k'}, \Theta) \tag{8}$$

As $P(g_n|\Theta)$ does not affect the maximization, Equation 8 can be simplified as:

$$S^* = \arg\max_{S \in \mathcal{S}} \prod_{n=1}^{N} \underbrace{\frac{P(s_n = f_k|g_n, \Theta)}{P(s_n = f_k|\Theta)}}_{\text{local emission score}} \underbrace{P(s_n = f_k|s_{n-1} = f_{k'}, \Theta)}_{\text{transition probability}} \tag{9}$$

In Equation 9, assuming a uniform transition probability $P(s_n = f_k|s_{n-1} = f_{k'}, \Theta)$ and a uniform prior probability $P(s_n = f_k|\Theta)$, the estimation of the parameters would be restricted to learning the relationship between graphemes and phonemes, i.e., $P(s_n = f_k|g_n, \Theta)$. In this article, 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 in Section 3.4) or as presented in the following section, it can be estimated by exploiting acoustic data which can bring certain advantages.

3.2. Parameter estimation through acoustic data

Estimating the parameters $P(s_n = f_k|g_n)$ through acoustic data is not a trivial task. Recently within the ASR community approaches have been proposed which can model two types of units, namely graphemes and phonemes using acoustic data. These approaches can provide a means to learn the relationship between graphemes and phonemes (i.e., the parameters $P(s_n = f_k|g_n)$) through acoustic information. In this section, we first provide a background about these ASR approaches and then explain how they can be exploited for parameter estimation.

3.2.1. Relevant background

In a more recent work it was shown that in subword unit-based ASR approaches, the link between the lexical subword units and acoustic features can be factored through a latent variable referred to here as *acoustic units* into two models, namely the acoustic model and the lexical model (Rasipuram and Magimai.-Doss, 2015):

1. In the acoustic model the relationship between the acoustic features $\mathbf{x}_t$ and acoustic units $\{a^d\}_{d=1}^D$ is modeled. The acoustic units $\{a^d\}_{d=1}^D$ can be context-independent (CI) or clustered context-dependent (CD) subword units. The acoustic model can be either a Gaussian mixture model (GMM) or an artificial neural network (ANN). In likelihood-based ASR approaches, the acoustic model estimates likelihood vectors $\mathbf{v}_t = [v_t^1, \ldots, v_t^d, \ldots, v_t^D]^{\mathrm{T}}$ with $v_t^d = p(\mathbf{x}_t|a^d)$. In posterior-based ASR approaches, the acoustic model estimates posterior probability vectors $\mathbf{z}_t = [z_t^1, \ldots, z_t^d, \ldots, z_t^D]^{\mathrm{T}}$ with $z_t^d = P(a^d|\mathbf{x}_t)$.

2. In the lexical model the relationship between the acoustic units $\{a^d\}_{d=1}^D$ and lexical subword units $\{l^i\}_{i=1}^I$ is modeled as a set of categorical distributions $\{\mathbf{y}_i\}_{i=1}^I$, where $\mathbf{y}_i = [y_i^1, \ldots, y_i^d, \ldots, y_i^D]^{\mathrm{T}}$ and $y_i^d = P(a^d|l^i)$. The relationship between the acoustic and lexical units can be either a one-to-one deterministic map or a probabilistic map, leading to deterministic or probabilistic lexical modeling-based ASR approaches respectively. In deterministic lexical modeling-based ASR approaches (e.g. standard HMM/GMM or hybrid HMM/ANN), each lexical unit is deterministically mapped to an acoustic unit, i.e., $\mathbf{y}_i$ is a Kronecker delta distribution. The deterministic mapping is obtained either through knowledge (for CI lexical units) or learned during clustering and tying of states (for CD lexical units). In probabilistic lexical modeling-based ASR approaches, however, the relation between the acoustic and lexical units is probabilistically learned as explained briefly below.

As elucidated in (Rasipuram and Magimai.-Doss, 2015), there are different probabilistic lexical modeling-based ASR approaches such as probabilistic classification of HMM states (PCHMM) (Luo and Jelinek, 1999), tied posterior HMM (Rottland and Rigoll, 2000) and Kullback-Leibler divergence-based HMM (KL-HMM) (Aradilla et al., 2007). In these approaches, an HMM is trained in which each state represents a lexical unit l^i and is parameterized by a categorical distribution $\mathbf{y}_i$. The lexical model parameters $\{\mathbf{y}_i\}_{i=1}^I$ are estimated based on the acoustic unit evidence obtained from the acoustic model, i.e., $\mathbf{z}_t$ in the case of KL-HMM and $\mathbf{v}_t$ in the case of PCHMM and tied posterior HMM.

The probabilistic lexical modeling framework brings certain advantages over the deterministic lexical modeling-based ASR approaches:

Advantage-1: The acoustic and lexical units can be different. For example, the acoustic units can represent phonemes while the lexical units can represent graphemes (Rasipuram and Magimai.-Doss, 2015; Imseng et al., 2011).

Advantage-2: The acoustic and lexical units can represent subword units with different context lengths. 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).

Advantage-3: The acoustic model and the lexical model can be trained on different sets of data. For example, the acoustic model can be trained on language-independent data while the lexical model is trained on the target language data (Rasipuram and Magimai.-Doss, 2015).

3.2.2. Relevance to G2P

By exploiting the advantages of the probabilistic lexical, it is possible to learn the grapheme-to-phoneme relationship $P(s_n = f_k|g_n)$ through estimation of lexical model parameters $\mathbf{y}_i$ using acoustic information. More precisely:

1. The probabilistic lexical model enables modeling different types of units (*Advantage-1*). With graphemes as lexical units and phonemes as acoustic units in the probabilistic lexical modeling framework, the parameters of the lexical model $\mathbf{y}_i$ capture a grapheme-to-phoneme relationship. Therefore, the parameter estimation problem for the HMM explained in Section 3.1 amounts to learning the parameters $\{\mathbf{y}_i\}_{i=1}^{I}$ in the probabilistic lexical modeling framework in which the set of acoustic units is equal to $\mathcal{F} = \{f_1, \ldots, f_k, \ldots, f_K\}$ (in Section 3.1) and the set of lexical units $\mathcal{L}$ contains the possible graphemes in the target language (i.e., $\forall G_n = g_n : g_n \in \mathcal{L}$).

2. The grapheme-to-phoneme relationship can be regular or irregular, depending on the language. Languages with irregular grapheme-to-phoneme relationship require modeling of longer grapheme contexts to correctly capture the relationship between graphemes and phonemes. The probabilistic lexical model allows such a possibility as the length of the grapheme and phoneme contexts can be different (*Advantage-2*).

3. Conventional G2P approaches require a seed lexicon, which may not be available particularly for under-resourced languages. The probabilistic lexical model can relax the need for an initial phonemic lexicon by training the acoustic model on language-independent data (*Advantage-3*). More precisely, the grapheme-to-phoneme relationship can be learned by exploiting data from resource-rich languages to train the acoustic model and then training the grapheme-based lexical model on the limited amount of training data in the target language available. This is potentially interesting when there is lack of lexical resources in the target language.

The lexical model parameters $\{\mathbf{y_i}\}_{i=1}^{I}$ can be estimated using the Viterbi expectation-maximization algorithm given either acoustic unit posterior probability estimates $\mathbf{z}_t$ in posterior-based approaches (such as KL-HMM) or likelihood estimates $\mathbf{v}_t$ in likelihood-based approaches (such as PCHMM and tied posterior HMM). In the expectation (segmentation) step, an optimal grapheme

state sequence is obtained for each training utterance using the Viterbi algorithm. Then in the maximization step, given the optimal grapheme state sequences and the phonetic evidence, i.e., $\mathbf{z}_t$ or $\mathbf{v}_t$ belonging to each of these states, the new set of parameters $\{\mathbf{y_i}\}_{i=1}^{I}$ are estimated by either minimizing a cost function based on KL-divergence in the case of KL-HMM approach or maximizing a cost function based on likelihood in the case of PCHMM and tied posterior HMM approaches. Figure 4 illustrates the diagram of the EM step in the probabilistic lexical modeling framework in which the acoustic and lexical units represent phonemes and graphemes respectively.

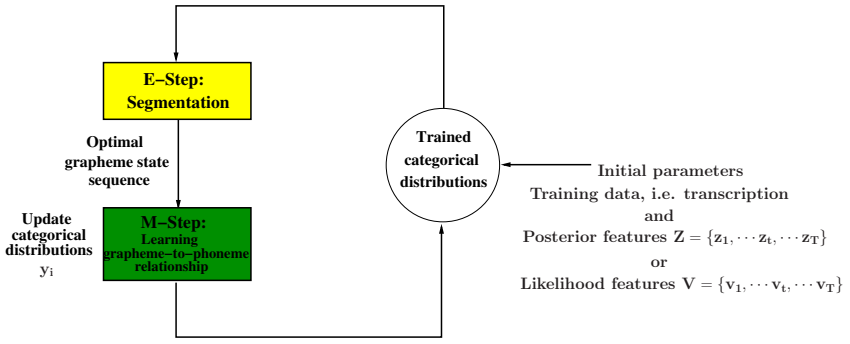


Figure 4: Illustration of parameter estimation in the probabilistic lexical modeling framework, where the acoustic units represent phonemes and lexical units represent graphemes.

3.3. Pronunciation Inference

Given the orthographic transcription of the word and the estimated parameters of the probabilistic lexical model, the lexical model can be used as a generative model where each state emits a single phoneme posterior probability vector. The most probable phoneme sequence is then inferred by decoding the sequence of phoneme posterior probabilities using the ergodic HMM presented in Section 3.1. Multiple pronunciations for a word could be extracted within this framework using n-best decoding.

Figure 5 provides a summary of the acoustic G2P approach using the probabilistic lexical modeling framework as a three-step process:

1. *Step 1*: An acoustic model (ANN or GMM) is trained to estimate phoneme posterior probabilities $\mathbf{z}_t$ or phoneme likelihoods $\mathbf{v}_t$.

2. *Step 2*: A grapheme-based probabilistic lexical model is trained to learn the relationship between graphemes and phonemes.

3. *Step 3*: Given the trained lexical model and the text, the most probable sequence of phonemes is inferred.

It can be seen that the recently proposed acoustic data-driven G2P approach (Rasipuram and Magimai-Doss, 2012a) in the KL-HMM framework is

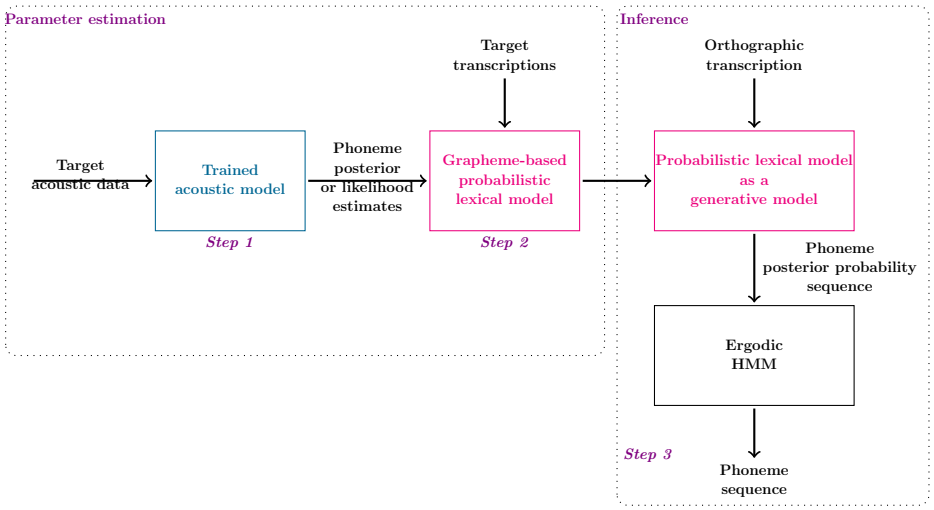


Figure 5: Block diagram of the acoustic G2P approach.

a particular case of this formalism where the acoustic model is estimating posterior probabilities $\mathbf{z}_t$ and the grapheme-to-phoneme relationship is captured through the parameters of the KL-HMM, i.e., a probabilistic lexical model.

Briefly, as illustrated in Figure 6, in this approach a grapheme-based ASR model is trained where the acoustic units $\{a^d\}_{d=1}^{D}$ are phonemes and the lexical units $\{l_i\}_{i=1}^{I}$ (modeled by HMM states) are based on graphemes. The acoustic model estimates phoneme posterior probabilities $\mathbf{z}_t$. The lexical model parameters $\{\mathbf{y}_i\}_{i=1}^{I}$ are trained using $\mathbf{z}_t$ as a feature observation with a cost function based on Kullback-Leibler divergence. More precisely, the local score at each HMM state is defined as the Kullback-Leibler divergence between the posterior feature $\mathbf{z}_t$ and categorical distribution $\mathbf{y}_i$ (Aradilla et al., 2007):

$$S_{\mathbf{KL}}(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} y_i^d \log(\frac{y_i^d}{z_t^d}) \tag{10}$$

As KL-divergence is not a symmetric measure, the local score can be estimated in other ways (Aradilla et al., 2008):

$$S_{\mathbf{RKL}}(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} z_t^d \log(\frac{z_t^d}{y_i^d}) \tag{11}$$

$$S_{\mathbf{SKL}}(\mathbf{y}_i, \mathbf{z}_t) = \frac{1}{2}(S_{\mathbf{KL}} + S_{\mathbf{RKL}}) \tag{12}$$

More information about the parameter estimation step in KL-HMM approach can be found in Appendix A.

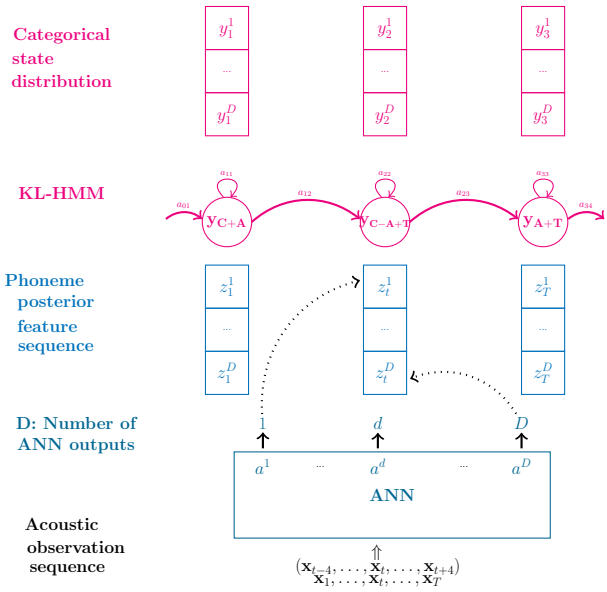


Figure 6: Illustration of KL-HMM approach in which graphemes are used as lexical units and the acoustic model is an ANN.

In this article, 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).

3.4. Comparison to existing approaches

The acoustic G2P approach can be considered to be similar to conventional data-driven G2P approaches in some aspects. By comparing the parameter estimation scheme presented in Section 3.2.2 and methods used in data-driven G2P approaches for capturing the grapheme-to-phoneme relationship, it can be observed that the Viterbi EM algorithm used for estimating the parameters $\{\mathbf{y_i}\}_{i=1}^{I}$ is similar to the alignment and learning step that is done in the G2P approaches. In other words, similar to the acoustic G2P approach, data-driven G2P approaches can be considered to consist of an *E-step* and an *M-step*:

1. *E-step:* This step acts as a common step among different data-driven G2P approaches which provides an alignment between the grapheme sequence and the phoneme sequence.

2. *M-step:* This step captures the relationship between graphemes and phonemes through a learning method. Depending on the G2P approach, the learning method can be different. For example, it can be through decision trees, neural networks, n-gram models or CRF.

In addition to the aforementioned aspects, there are other comparisons that can be drawn:

- *Comparison to the local classification-based approaches*: The local classification-based approaches can be seen as a particular case of the formalism in Section 3.1 where the transition and prior probabilities are uniform and phoneme 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 phoneme posterior probabilities would be zero or one. For the ANN-based approach, however, the output of the neural network directly provides phoneme posterior probability estimates.

  The main difference between the local classification-based approaches and the acoustic G2P approach is in the estimation of phoneme posterior probabilities. In the local classification-based approaches, the relationship between the graphemes and phonemes is learned from the seed lexicon while in the proposed approach the relationship is learned by the probabilistic lexical model through acoustics.

- *Comparison to the joint multigram approach*: The joint multigram and the proposed G2P approach are both capable of modeling the relationship between context-dependent graphemes and phonemes; and in both approaches the grapheme and phoneme context lengths can be different. The main difference between the two approaches is in the resources used and the learning methodology. In the joint multigram G2P approach, the grapheme-to-phoneme relationship is learned from the seed lexicon using the concept of graphone as a joint unit and the generative aspect is modeled through n-grams. On the other hand, in the proposed G2P approach conditional distribution of the phonemes given graphemes is modeled using acoustic information and the generative model is the HMM.

- *Comparison to the HMM-based approach*: Both the acoustic G2P approach and the HMM-based approach proposed by Taylor (2005) enable global inference in the generative framework of HMM. However, the emission scores and the HMM states are different in the two approaches. The HMM-based G2P approach estimates the likelihood of the graphemes given the phoneme states using seed lexicon as the training data. In the proposed approach, however, the posterior probability of phonemes given grapheme states is estimated using acoustic data.

- *Comparison to the pronunciation extraction using acoustic data:* The proposed approach and the methods for pronunciation extraction using phoneme recognizers (explained in Section 2.3) are similar in the sense that they both use an ergodic model to decode and obtain the phonetic transcription of the words (Mokbel and Jouvet, 1999). The difference appears at the level of obtaining the input to the ergodic model. In the proposed approach, the input to the ergodic model is the phoneme posterior probability sequence obtained from the grapheme-based probabilistic lexical model as shown in Figure 5

(Step 3). In the pronunciation extraction methods in the literature, however, the input is replaced by the phoneme posterior probability sequence or phoneme likelihood estimated on the acoustic data as shown in Figure 7. As a result, the pronunciation extraction approaches described in Section 2.3 are restricted to words for which acoustic samples are available, while this is not the case for the proposed G2P approach.

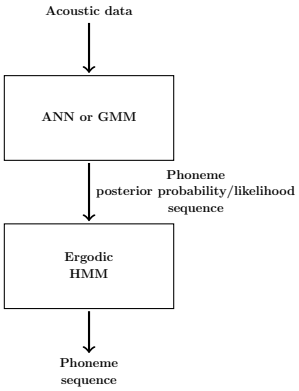


Figure 7: Block diagram of the pronunciation extraction methods in the literature using acoustic data.

Though the proposed approach does not rely on the availability of acoustic samples, it has the potential to exploit acoustic examples (if available) to select pronunciations borrowing the techniques from pronunciation extraction methods using phoneme recognizers.

- *Comparison to the acoustic G2P approaches in the literature*: The acoustic G2P approaches explained in Section 2.3 and the proposed approach both take advantage of acoustic information to obtain pronunciations. However, the acoustic G2P approaches in the literature use acoustic information to weight the pronunciations learned through G2P techniques, while the proposed G2P approach uses the acoustic information to learn the grapheme-to-phoneme relationship. As a consequence, in the acoustic G2P approaches in the literature acoustic samples are required while in the proposed approach this limitation is relaxed.

In this article, we benchmark the acoustic G2P approach against two conventional G2P approaches: 1) decision tree-based G2P which like the acoustic G2P approach is a particular case of the HMM-based formalism presented in Section 3.1, and 2) the state-of-the-art joint multigram G2P approach. Our studies focus on the first two advantages of the probabilistic lexical model and evaluate the G2P approaches on resource-rich languages in which a phonemic lexicon is available.

4. Experimental setup

In this section, we present the databases along with the experimental setups used for evaluating the G2P approaches.

4.1. Data sets

The performance of G2P approaches can depend on different factors:

- *Language*: The grapheme-to-phoneme relationship can be regular or irregular depending on the language. The G2P task for languages with irregular grapheme-to-phoneme relationship can be 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 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 G2P-generated pronunciations can have marginal or major effects on the performance of ASR systems.

In order to achieve efficient experimental design while considering the aforementioned factors thoroughly, we conducted our studies on two databases : 1) PhoneBook corpus, a small-vocabulary isolated word recognition English corpus, and 2) MediaParl corpus, a large-vocabulary continuous speech recognition (LVCSR) Swiss French corpus.

4.1.1. PhoneBook: isolated word recognition English corpus

PhoneBook is a speaker-independent task-independent isolated word recognition corpus (Pitrelli et al., 1995) for small size (75 words) and medium size (600 words) vocabularies. We use the medium size vocabulary task with 600 unique words (Dupont et al., 1997). The overview of the PhoneBook corpus in terms of number of utterances, hours of speech data, speakers and words present in train, cross-validation and test set is given in Table 1.

Table 1: Overview of the PhoneBook corpus in terms of number of utterances, hours of speech data, speakers and words present in the train, cross-validation and test sets.

Number of	Train	Cross-validation	Test
Utterances	19421	7290	6598
Hours	7.7	2.9	2.6
Speakers	243	106	96
Words	1580	603	600

The training set consists of 26,711 utterances (obtained by merging the small training set and cross-validation set as in (Dupont et al., 1997)), and test set consists of 6598 speech utterances. The test vocabulary consists of words and speakers which are unseen during training. PhoneBook pronunciation lexicon is transcribed using 42 phonemes (including silence).

In this article, we study the G2P on the PhoneBook English corpus as it is a challenging task for several reasons:

- The grapheme-to-phoneme relationship in English is highly irregular.
- The training and test vocabulary sets are totally different.
- The corpus contains uncommon English words and proper names (e.g. Witherington, Gargantuan, etc.).
- It can be seen as a resource-limited scenario as the amount of training data (in terms of number of training words and amount of speech data) is small.

4.1.2. MediaParl: LVCSR bilingual corpus

MediaParl is a bilingual corpus containing recordings of debates in Valais parliament in Switzerland in both Swiss German and Swiss French. Valais is a state in Switzerland including both French and German speakers with variety of accents. In this study, we used the French part of the corpus as French is a more challenging language for the G2P task due to its relatively more irregular grapheme-to-phoneme relationship compared to German. In our experiments, the database is partitioned into training, cross-validation and test set according to the structure provided in (Imseng et al., 2012a). Table 2 provides the number of utterances, hours of speech data, speakers and words present in the train, cross-validation and test set. 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.

Table 2: Overview of the MediaParl corpus in terms of number of utterances, hours of speech data, speakers and words present in the train, cross-validation and test set. For the test set, the amount of native and non-native data is shown as well.

Number of	Train	Cross-validation	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	1055	3376	4246

The French manual dictionary of the MediaParl corpus is provided in SAMPA format with a phoneme set of size 38 (including silence) and contains all the words in the train, cross-validation and test set. The dictionary includes the BDLex pronunciation lexicon[4]. For the words that are not found in the dictionary, a WFST-driven G2P system is used to generate the pronunciations[5] and afterwards the generated pronunciations are hand-corrected. The total vocabulary size is 12362. The dictionary size for the training set is 10800 (for 10555 words). The test set contains 4246 words of which 915 words are not seen during training.

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

Compared to the PhoneBook corpus, the MediaParl corpus can be interesting for the G2P task as:

- In French, the pronunciations could be more predictable given the orthography compared to English which may require having some prior knowledge about the word for "true" pronunciation prediction.
- The amount of training data is relatively more than PhoneBook corpus.
- The amount of unseen words in the test set is relatively small.
- MediaParl corpus contains debates as a form of spontaneous speech as well as non-native speech.

4.2. Grapheme-to-phoneme converters

The first step in evaluating G2P-based approaches is development of G2P-generated lexicons. For this purpose, we built grapheme-to-phoneme converters with the following setups:

- *Decision tree-based approach*: For building lexicons using the decision tree-based approach, we used the Festival toolkit (Taylor et al., 1998) . The width of grapheme context was set to 5 in both PhoneBook and MediaParl corpora.
- *Joint multigram approach*: For generating pronunciations based on the joint multigram approach, we used the Sequitur software developed at RWTH Aachen University[6]. The width of the graphone context was tuned on the cross-validation set. The optimal n-gram context size was 4 and 6 in the PhoneBook and MediaParl corpora respectively.
- *Acoustic G2P approach*: As the first step of the pronunciation generation using the acoustic G2P approach, ANNs more specifically multilayer perceptrons (MLPs) were trained. We used 39-dimensional PLP cepstral features with four preceding and four following frame context as 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, 2012b,a). However, recent advances in speech technology have shown that ANNs with deep architectures can help in improving the performance of the systems (Hinton et al., 2012). In order to investigate the effect of different MLP architectures on the performance of the acoustic G2P approach, we built MLPs with different number of layers and output units as follows:

  – *MLP-3-CI-M*: a three-layer MLP classifying M context-independent phonemes. For the PhoneBook corpus $M = 42$ and for the MediaParl corpus $M = 38$.

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

– *MLP-5-CI-M*: a five-layer MLP classifying CI phonemes.
– *MLP-5-CD-M*: a five-layer MLP modeling M clustered context-dependent (CD) phonemes as outputs. The output units were derived by clustering context-dependent phonemes in HMM/GMM framework using decision tree state tying. Different number of acoustic units were derived by adjusting the log-likelihood difference. 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 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 in the PhoneBook and MediaParl corpora respectively.

As the second step in pronunciation generation, a KL-HMM system modeling tri-graphemes (single preceding and following context[7]) was trained. The choice of local score to learn the KL-HMM parameters can be important as previously shown in (Rasipuram and Magimai.-Doss, 2013). By using the local score $S_{\mathbf{KL}}$, the system is more capable of capturing one-to-one grapheme-to-phoneme relationships. On the other hand, when using $S_{\mathbf{RKL}}$ as the local score, the system can better handle one-to-many relationships. For the case when using $S_{\mathbf{SKL}}$ as local score, the system is able to capture both one-to-one and one-to-many relations. In this article, the KL-HMM parameters were trained by minimizing the cost function based on the local score $S_{\mathbf{RKL}}$ as it is suitable for the scenarios where the grapheme-to-phoneme relationship is irregular. For tying KL-HMM states we applied KL-divergence based decision tree state tying method proposed in (Imseng et al., 2012b).

In the inference step, each MLP output unit was modeled with three left-to-right HMM states. For the case of PhoneBook database, silence was removed in the ergodic HMM as it could lead to deletion of some phonemes when generating pronunciations. However, for the case of MediaParl French corpus, 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 inference step is demonstrated through the example word "MAP" in Figure 8.

Note that the use of clustered CD phonemes as MLP output units could possibly help to better model the relationship between the phonemes and the graphemes (similar to the effect of graphones in the joint multigram approach). However, in the inference we are interested in inferring CI phoneme sequence. To resolve this issue, after training the KL-HMM, for each lexical unit l^i, the

[7]This is mainly due to the limitations of the HTK in tying longer contexts. In future work we aim to explore longer grapheme contexts.

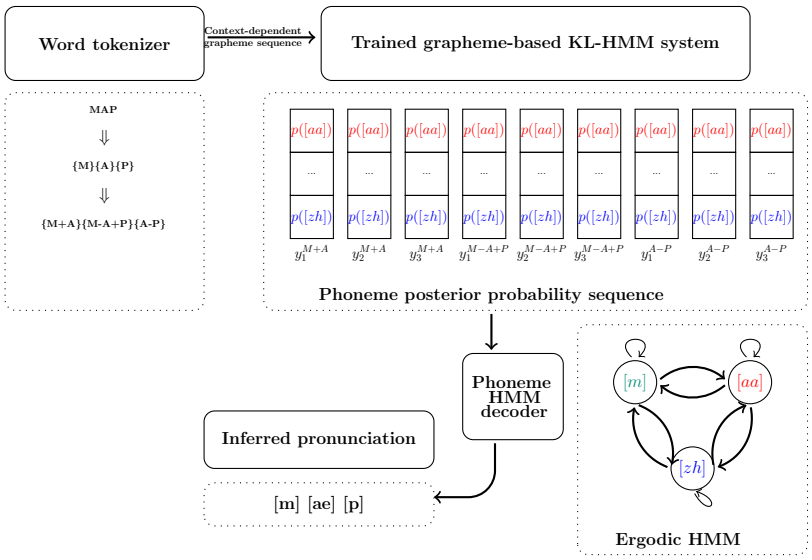


Figure 8: Block diagram of the inference phase in acoustic data-driven G2P task.

parameters $\{y_i^d = P(a^d|l^i)\}_{d=1}^{D}$ were marginalized, i.e., the posterior probabilities of the acoustic units $P(a^d|l^i)$ belonging to the same central phoneme were summed together.

4.3. Evaluation

We built G2P-based lexicons by using the grapheme-to-phoneme converters to generate pronunciations for the unseen words during training. We then evaluated G2P-generated lexicons at both pronunciation and ASR levels. For the evaluation at the pronunciation level, we computed phoneme and word accuracy on the G2P-based lexicons and analyzed the pronunciations using confusion matrix. For the evaluation at the ASR level, we considered different facets:

1. Evaluation using deterministic and probabilistic lexical modeling-based ASR systems.

2. Combination of acoustic G2P approach and conventional G2P approaches.

3. Comparison with the grapheme-based ASR system using KL-HMM.

The setup for each of these aspects is presented in this section.

4.3.1. Deterministic and probabilistic lexical modeling-based ASR systems

For evaluating the G2P-generated lexicons through deterministic lexical modeling-based ASR approaches, we used the HMM/GMM framework. We trained standard cross-word CD HMM/GMM systems with 39 dimensional PLP cepstral features $\mathbf{x}_t$ extracted using HTK toolkit (Young et al., 2006).

For the probabilistic lexical modeling-based ASR studies, we trained KL-HMM systems using phoneme posterior probabilities $\mathbf{z}_t$ obtained from *MLP-5-CI-M* as feature observations and modeled CD (tri) subword units. The KL-HMM parameters were trained by minimizing the cost function based on $S_{\mathbf{SKL}}$ as the local score. For tying KL-HMM (lexical) states we applied KL-divergence-based decision tree state tying method.

Currently, in standard HMM-based ASR system, it is common to use ANNs that classify clustered context-dependent phoneme states, referred to as CDNN, instead of GMMs as acoustic model to estimate emission likelihoods (Hinton et al., 2012). In more recent works, in the framework of KL-HMM, it has been shown that a performance comparable to the CDNN-based ASR system can be achieved with ANN trained to classify CI phonemes (Razavi et al., 2014; Razavi and Magimai.-Doss, 2014). In other words, probabilistic lexical modeling based ASR system using CI phonemes as acoustic units could be indicative of the performance of the CDNN based ASR system. Thus, we limited the deterministic lexical model studies to HMM/GMM systems.

4.3.2. Combination of G2P approaches

Conventional G2P approaches such as joint multigram and decision tree-based approaches learn the grapheme-to-phoneme relationship using a seed lexicon, while the acoustic G2P approach incorporates the acoustic information to learn this relationship. Therefore, it would be interesting to investigate the potential of acoustic G2P approach in providing complementary information to conventional G2P approaches. Combining the acoustic G2P approach and conventional G2P approaches can also be appealing in the sense that it can be related to the pronunciation variation methods incorporating acoustic information explained in Section 2.3. Towards these lines, we combined the acoustic G2P and conventional G2P approaches by building lexicons using pronunciations from either acoustic G2P and joint multigram approaches or acoustic G2P and decision tree-based approaches. We built HMM/GMM and KL-HMM systems in the same setup explained in Section 4.3.1.

4.3.3. Comparison with grapheme-based ASR using KL-HMM

The grapheme-based KL-HMM system was originally developed for ASR (Magimai.-Doss et al., 2011) and was later exploited for pronunciation generation. As grapheme-based approaches can avoid the need for a phonemic lexicon, it would be interesting to investigate whether doing lexicon development and ASR training in two separate stages as done in present phoneme-based ASR systems can bring any benefits over grapheme-based KL-HMM systems. For this purpose, we compared the grapheme-based KL-HMM system with the phoneme-based KL-HMM system using G2P-generated lexicons. The KL-HMM systems were built in the same setup explained in Section 4.3.1.

5. Pronunciation level results and analysis

In this section, we investigate the effect of different MLPs on the performance of the acoustic G2P approach and compare the acoustic G2P approach with joint

multigram and decision-tree-based approaches at the pronunciation level. In addition, we provide some pronunciation level analysis for the G2P approaches.

5.1. Results

Table 3 provides pronunciation level evaluation results in terms of phoneme and word accuracy for different G2P approaches. For the acoustic G2P approach, it can be observed that deeper MLP architectures generally perform better than three-layer MLP architectures. More precisely, it can be seen that for the PhoneBook corpus, through use of more layers and more outputs in the MLP architecture, the performance of the acoustic G2P approach at pronunciation level constantly improves. In the MediaParl corpus, using a five-layer MLP alone does not lead to improvements over three-layer MLPs. However, it can be observed that through using an MLP with more number of layers and marginalizing the posterior probabilities in the KL-HMM, significant improvements in terms of phoneme and word accuracy can be achieved.

Table 3: Pronunciation level evaluations in terms of phoneme accuracy (PA) and word accuracy (WA) using different G2P approaches.

MLP	PA on train	WA on train	PA on unseen	WA on unseen
MLP-3-CI-42	76.4	16.1	71.6	9.8
MLP-5-CI-42	77.2	17.9	72.4	10.8
MLP-5-CD-321	80.0	23.4	75.2	15.4
Joint multigram	98.8	93.9	89.2	50.5
Decision tree	87.7	38.9	81.5	31.0

(a) PhoneBook

MLP	PA on train	WA on train	PA on unseen	WA on unseen
MLP-3-CI-38	89.9	54.8	88.0	49.6
MLP-5-CI-38	89.9	54.5	87.8	49.5
MLP-5-CD-437	91.4	59.6	89.6	54.0
Joint multigram	99.8	99.3	97.4	89.0
Decision tree	98.1	90.9	96.0	82.2

(b) MediaParl

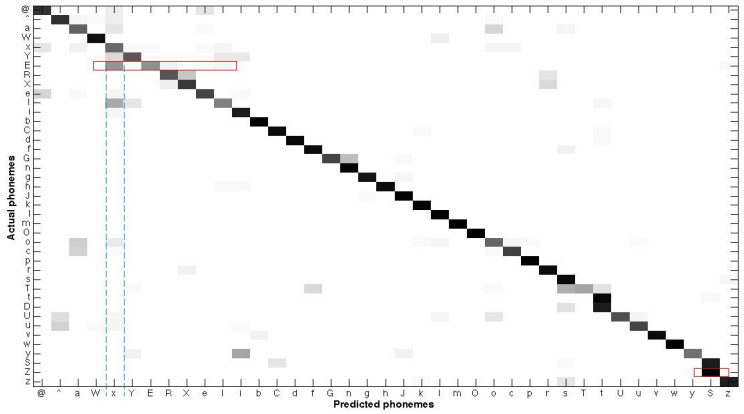
Additionally, it can be seen that for the PhoneBook corpus, the joint multigram approach is able to generate exact pronunciations for about 94 % of the words. This shows that the generated pronunciations with the joint multigram approach are consistent with the manually-generated pronunciations. On the other hand, for the acoustic G2P and decision tree-based approaches, the relatively poor word accuracy on the training words indicates that some discrepancies between the G2P-generated and manually-generated pronunciations exist. 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 approach.

The overall comparison of the results for different G2P approaches shows that conventional G2P approaches perform better than the acoustic G2P approach at the pronunciation level. This can be attributed to the fact that in conventional approaches, the grapheme-to-phoneme relationship is learned through direct use

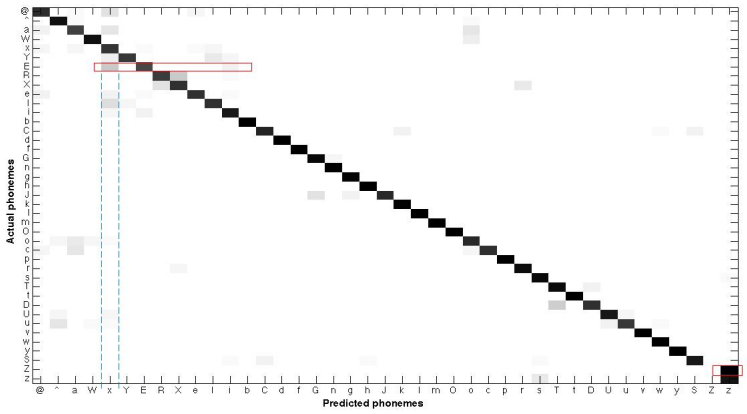
of the manually-generated train lexicon, while the acoustic G2P approach learns this relationship using acoustic information. As a result, conventional G2P approaches are able to generate closer pronunciations to the manual dictionary and therefore achieve better pronunciation level performance compared to the acoustic G2P approach.

5.2. Analysis

In this section, we provide the pronunciation level analysis for the joint multigram approach (as the state-of-the-art G2P approach) to be compared against the acoustic G2P method. Figures 9 and 10 visualize the confusion matrix in a gray scale color mapping for the pronunciation of the unseen words in the test set of the PhoneBook and MediaParl corpora respectively .



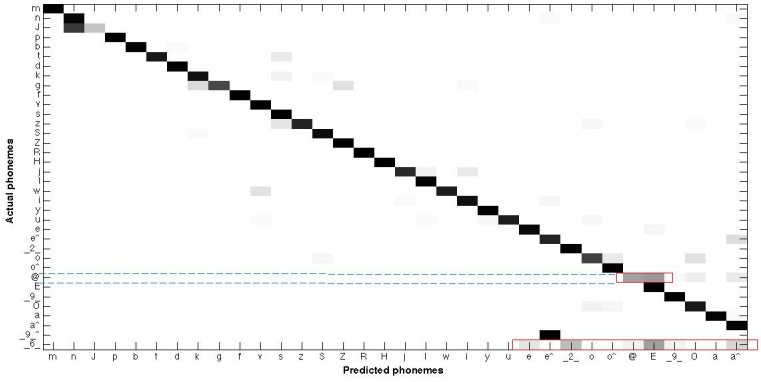
(a) Acoustic G2P Approach



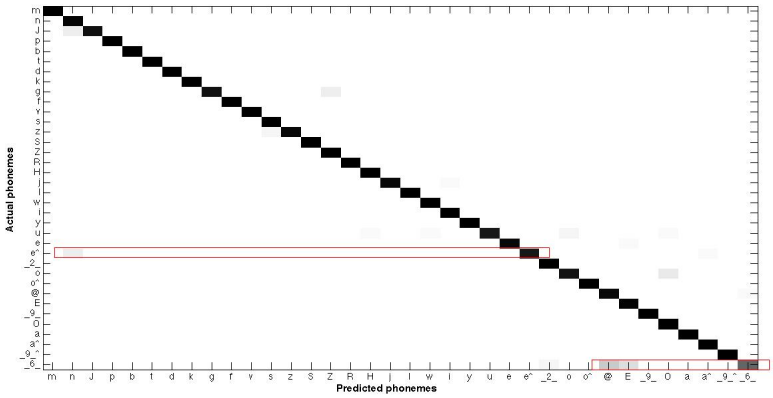
(b) Joint Multigram Approach

Figure 9: Relation between the actual and predicted phonemes in the pronunciations generated by the acoustic G2P and joint multigram approaches in the PhoneBook corpus.

It can be observed from Figure 9, that for the PhoneBook corpus, most of the confusions come from vowel phonemes such as /E/ (as found in the pronunciation of the word "aber": /a/ /b/ /E/ /R/) which are confused with similar phonemes such as /x/ (as appeared in the pronunciation of the word "abandonment": /x/ /b/ /@/ /n/ /d/ /x/ /n/ /m/ /x/ /n/ /t/) in both G2P approaches. Other confusions are related to consonant phonemes such as /Z/ which is confused with the phonemes /z/ and /S/ in joint multigram and acoustic G2P approaches. For the case of acoustic G2P approach, in fact the phoneme set size reduces as the phoneme /Z/ is replaced with the unvoiced phoneme /S/ which can be due to the confusion present at the output of MLP.



(a) Acoustic G2P Approach



(b) Joint Multigram Approach

Figure 10: Relation between the actual and predicted phonemes in the pronunciations generated by the acoustic G2P and joint multigram approaches in the MediaParl corpus.

Similarly for the case of MediaParl corpus, it can be seen that the confusions are mostly related to vowel phonemes such as /_6_/ (as found in the pronunciation of the word "arsenal": /a/ /R/ /s/ /_6_/ /n/ /a/ /l/) which is confused with the phoneme /E/ (as shown up in the pronunciation of the word "appel":

/a/ /p/ /E/ /l/). Similar to the PhoneBook corpus, in the acoustic G2P approach the phoneme set size is reduced since the phonemes /_6_/ and /_9_^/ are replaced with similar vowel phonemes.

To analyze the performance of the acoustic G2P and joint multigram approaches in terms of word accuracy at pronunciation level, we calculated the distribution of the unseen words in the test set based on Levenshtein distance between the generated pronunciation and the actual pronunciation. Table 4 provides the results when using pronunciations derived from the acoustic G2P and joint multigram approaches. For the acoustic G2P approach, it can be observed that 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.

Table 4: Distribution of the words in terms of Levenshtein distance for PhoneBook and MediaParl databases when using acoustic G2P and joint multigram approaches.

Levenshtein distance	Acoustic G2P	Joint multigram
0	93	304
1	0	0
2	3	1
3	13	5
4	25	19
5	57	31
6	101	50
7	92	50
8	86	45
9	51	36
10	39	26
11	19	20
12	15	10
13	7	4
14	0	0
15	1	1

(a) PhoneBook

Levenshtein distance	Acoustic G2P	Joint multigram
0	494	814
1	2	5
2	8	6
3	21	10
4	42	16
5	64	18
6	86	16
7	72	10
8	51	7
9	30	3
10	30	2
11	5	3
12	4	2
13	4	2
14	1	0
15	0	0
16	0	0
17	0	0
18	1	1

(b) MediaParl

To have a better sense about the quality of the pronunciations generated by acoustic G2P and joint multigram approaches, Tables 5 and 6 present some

of the generated pronunciations using both approaches in the PhoneBook and MediaParl corpora respectively. It can be observed from both tables that the joint multigram and acoustic G2P approaches show different kinds of capabilities in generating correct pronunciations. More precisely, in the English words "yowler", "uncharted" and "uninspired" the acoustic G2P approach is providing better pronunciations than the joint multigram approach. Similarly for the French words "les" and "tes" the acoustic G2P approach is 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 "aboutissait" and "acceptions" compared to the acoustic G2P approach. As the joint multigram and acoustic G2P approaches generate different types of errors, combination of the two approaches (without applying any pruning) can help in improving the ASR accuracy. We will see the effect of combination of G2P approaches on the ASR performance in Section 6.2.

Table 5: Sample words from the PhoneBook corpus along with their joint multigram-based, acoustic G2P-based and correct pronunciations.

Word	Joint multigram-based pronunciation	Acoustic G2P-based pronunciation	Correct pronunciation
yowler	/y/ /o/ /l/ /X/	/y/ /W/ /l/ /X/	/y/ /W/ /l/ /X/
uncharted	/ˆ/ /n/ /k/ /a/ /r/ /t/ /x/ /d/	/ˆ/ /n/ /C/ /a/ /r/ /t/ /x/ /d/	/ˆ/ /n/ /C/ /a/ /r/ /t/ /x/ /d/
uninspired	/ˆ/ /n/ /I/ /n/ /s/ /p/ /Y/ /r/ /d/	/ˆ/ /n/ /x/ /n/ /s/ /p/ /Y/ /X/ /d/	/ˆ/ /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 6: Sample words from the MediaParl corpus along with their joint multigram-based, acoustic G2P-based and correct pronunciations.

Word	Joint multigram-based pronunciation	Acoustic G2P-based pronunciation	Correct pronunciation
les	/l/	/l/ /E/	/l/ /E/
bourlard	/b/ /u/ /R/ /a/ /R/	/b/ /u/ /R/ /l/ /a/ /R/	/b/ /u/ /R/ /l/ /a/ /R/
tes	/t/	/t/ /E/	/t/ /E/
absorption	/a/ /b/ /s/ /O/ /R/ /p/ /s/ /j/ /oˆ/	/a/ /s/ /O/ /R/ /p/ /s/ /j/ /oˆ/	/a/ /b/ /s/ /O/ /R/ /p/ /s/ /j/ /oˆ/
aboutissait	/a/ /b/ /u/ /t/ /i/ /s/ /E/	/a/ /b/ /u/ /s/ /i/ /s/ /E/	/a/ /b/ /u/ /t/ /i/ /s/ /E/
acceptions	/a/ /k/ /s/ /E/ /p/ /t/ /j/ /oˆ/	/a/ /k/ /s/ /E/ /t/ /s/ /j/ /oˆ/	/a/ /k/ /s/ /E/ /p/ /s/ /j/ /oˆ/

6. ASR level results and analysis

This section presents the ASR evaluation results from different aspects explained in Section 4.3. For comparing the ASR performance of different systems, we applied the statistical significant test presented in (Bisani and Ney, 2004) with the confidence level of 95%.

6.1. Deterministic and probabilistic lexical modeling-based ASR systems

Table 7 presents the performance of HMM/GMM and KL-HMM systems in terms of word accuracy using the pronunciations from different G2P approaches for the unseen words. Similar to the pronunciation level results in Table 3, it can be observed that with improvements in the ANN architecture, the performance

of HMM/GMM systems also improves in most of the cases. However such improvements are not observed for the KL-HMM systems. This can be as a result of the capability of the probabilistic lexical model in better handling of possible errors in the pronunciations.

Table 7: Performance of HMM/GMM and KL-HMM systems in terms of word accuracy using different G2P approaches.

G2P Approach	Word accuracy	
	HMM/GMM	KL-HMM
Acoustic-G2P-MLP-3-CI-42	81.1	84.8
Acoustic-G2P-MLP-5-CI-42	82.1	85.1
Acoustic-G2P-MLP-5-CD-321	82.9	85.0
Joint multigram	88.8	89.4
Decision tree	83.6	84.9
Manual dictionary	98.2	98.2

(a) PhoneBook

G2P Approach	Word accuracy	
	HMM/GMM	KL-HMM
Acoustic-G2P-MLP-3-CI-38	72.0	73.3
Acoustic-G2P-MLP-5-CI-38	72.0	73.2
Acoustic-G2P-MLP-5-CD-437	72.2	73.3
Joint multigram	73.1	74.0
Decision tree	72.8	73.8
Manual dictionary	73.2	74.1

(b) MediaParl

The difference in the performance of the acoustic G2P and decision tree-based G2P approaches is not statistically significant despite using a shorter grapheme-context and the relatively poor pronunciation level performance in the acoustic G2P approach. Furthermore, the proposed approach performs comparable to the joint multigram G2P method in the MediaParl corpus. However, for the PhoneBook task, the joint multigram G2P approach performs significantly better than the acoustic G2P method. The difference in the behavior of the acoustic G2P approach in the two databases could be due to the following factors:

- Language: Since the grapheme-to-phoneme relationship in English is more irregular compared to French, it may require modeling of more than single preceding and following grapheme context.
- Discrepancy between the manually-generated and G2P-generated pronunciations: As it can be seen from Table 3, the word accuracy at the pronunciation level for the acoustic G2P approach is poor (in particular in the PhoneBook corpus). As a consequence, the phoneme contexts observed in the training lexicon can be different from the contexts seen in the generated lexicon. This effect could lead to discrepancies between the existing and G2P-generated pronunciations. For the MediaParl corpus, the unseen words are 20% of the overall words in the test set which do not appear frequently in the test set (the

most frequent unseen word has occurred only 7 times). As a result, the possible discrepancies between the existing and G2P-generated pronunciations for the unseen words may not be felt on the performance of the system. On the other hand in the PhoneBook corpus, the test set vocabulary is completely unseen and contains uncommon words. Therefore, the possible inconsistencies between the pronunciations could affect the ASR performance.

In order to ascertain that, we conducted experiments on the PhoneBook corpus by using the G2P-generated pronunciations in both train and test lexicons (no pronunciation from the manual dictionary was used). Table 8 presents the ASR performance in terms of word accuracy. It can be observed that in almost 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).

Through use of G2P-generated pronunciations in both train and test lexicons, the acoustic G2P approach achieves comparable performance to the joint multigram approach at ASR level, i.e., the difference between the ASR performance of the acoustic G2P and the joint multigram approach is not statistically significant.

Table 8: Performance of ASR systems in terms of word accuracy when using G2P-generated pronunciation at both train and test lexicons.

G2P Approach	Word accuracy	
	HMM/GMM	KL-HMM
Acoustic-G2P-MLP-5-CD-321	88.3	88.8
Joint multigram	89.1	89.4
Decision tree	88.6	87.5
Manual dictionary	98.2	98.2

6.2. Combination of G2P approaches

Table 9 reports the ASR performance of HMM/GMM and KL-HMM systems in terms of word accuracy using pronunciations from both acoustic G2P and conventional G2P approaches as explained in Section 4.3.2.

For the PhoneBook corpus, it can be observed that significant improvements in terms of word accuracy are achieved for both HMM/GMM and KL-HMM systems compared to the case using pronunciations based on a single G2P approach. This shows that the acoustic G2P approach can provide complimentary information for the conventional G2P approaches.

For the MediaParl corpus, it can be seen that the systems trained using the lexicon obtained from combination of G2P approaches yield the same performance as the systems trained using the manual dictionary. However, compared

Table 9: ASR performance in terms of word accuracy when combining pronunciations from different G2P approaches.

G2P Approach Combinations	Word accuracy	
	HMM/GMM	KL-HMM
Acoustic-G2P-MLP-5-CD-321 +Joint multigram	91.7	92.1
Acoustic-G2P-MLP-5-CD-321 +Decision tree	89.9	91.0
Manual dictionary	98.2	98.2

(a) PhoneBook

G2P Approach Combinations	Word accuracy	
	HMM/GMM	KL-HMM
Acoustic-G2P-MLP-5-CD-437 +Joint multigram	73.1	74.2
Acoustic-G2P-MLP-5-CD-437 +Decision tree	73.1	74.1
Manual dictionary	73.2	74.1

(b) MediaParl

to the PhoneBook corpus, the improvements in ASR accuracy through combination of G2P approaches are less noticeable. This can be due to availability of larger amount of training data and also smaller amount of unseen words in the MediaParl corpus which provides an ideal scenario for learning grapheme-to-phoneme relationship in conventional G2P approaches.

6.3. Comparison with grapheme-based ASR using KL-HMM

Table 10 provides the ASR word accuracies for the grapheme-based KL-HMM systems and phoneme-based KL-HMM systems using G2P-generated lexicons as explained in Section 4.3.3. The studies show that when the train and test set are in the same domain (as in the case of MediaParl database), building an ASR system as a two stage process helps. On the other hand, when the train and test set do not share information in terms of vocabulary (as in the case of PhoneBook database), the grapheme-based framework leads to a better performance.

Table 10: Comparison of the ASR results for the grapheme-based KL-HMM and the phoneme-based KL-HMM systems using the pronunciations derived from the combination of G2P approaches.

Database	Word accuracy	
	Grapheme-based KL-HMM	Phoneme-based KL-HMM
PhoneBook	93.6	92.1
MediaParl	71.7	74.2

7. Conclusions

In this article, we presented a novel G2P formalism in an HMM-based framework in which the grapheme-to-phoneme relationship is locally modeled as a

distribution of phoneme probabilities given a grapheme input. We showed that the existing local classification-based G2P approaches such as the decision tree-based G2P and the ANN-based G2P can be seen as a particular case of this formulation. Furthermore, we showed that the formalism together with recent developments in grapheme-based ASR using probabilistic lexical modeling naturally leads to a G2P approach where the grapheme-to-phoneme relationship is learned through acoustics.

We compared the proposed acoustic G2P approach against conventional G2P approaches on two different languages that have irregular grapheme-to-phoneme relationships. Our studies showed that the proposed acoustic G2P approach-based lexicon, despite poor performance at the pronunciation level, (a) yields ASR systems comparable to conventional G2P approach-based lexicons, and (b) yields a better ASR system when the lexicons are combined. These findings are particularly interesting as in the acoustic G2P approach, unlike conventional G2P approaches, only single left and single right grapheme contexts were modeled and a few of the phonemes were even filtered out during the pronunciation inference process. In the literature, evaluation of G2P approaches is often limited to pronunciation level evaluation. As a by-product, our studies also showed that the pronunciation level evaluation of G2P approaches is not fully indicative of the end system level performance, here in this case ASR system level performance. Such an observation has also been echoed in a recent study comparing the HCRF-based G2P approach with the joint multigram approach (Hahn et al., 2013).

A distinctive capability of the proposed acoustic G2P approach is that, unlike the conventional G2P approaches, it does not necessitate the availability of seed lexicon in the target language or domain to learn the grapheme-to-phoneme relationship (Rasipuram and Magimai-Doss, 2012a). More precisely, the acoustic to multilingual phoneme relationship can be learned using the acoustic and lexical resources of the auxiliary languages and domains, and the grapheme-to-phoneme relationships can be learned on the target language acoustic resources (Rasipuram and Magimai.-Doss, 2015; Rasipuram et al., 2013). In addition to that the proposed approach can be seamlessly integrated with automatic subword unit derivation through HMM-based spectral clustering and development of pronunciation lexicons (Razavi and Magimai.-Doss, 2015). These capabilities are potentially interesting for under-resourced languages. Our future work will focus along these directions to develop lexical resources for under-resourced languages.

Acknowledgment

This work was supported by Hasler foundation through the grant Flexible acoustic data driven grapheme to acoustic unit conversion (AddG2SU) and by the Swiss NSF through the grant Flexible Grapheme-Based Automatic Speech Recognition (FlexASR).

Appendix A. Parameter estimation in the KL-HMM approach

KL-HMM is fully parameterized by $\Theta_{kull} = \{\{\mathbf{y}_i\}_{i=1}^{I}, \{a_{ij}\}_{i,j=1}^{I}\}$ where I is the total number of states and state i is represented by categorical distribution $\mathbf{y}_i$, a_{ij} is the transition probability from state i to state j.

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 state 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 parameters Θ_{kull} are estimated by Viterbi expectation maximization algorithm which minimizes the cost function,

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

where $q_t \in \{1, \cdots, I\}$, $\mathcal{Q}$ denotes set of all possible HMM state sequences, $Q = \{q_1(n), \cdots, q_t(n), \cdots, q_T(n)\}$ denotes a sequence of HMM states and $\mathbf{z}_t(n) = [z_t^1(n), \cdots, z_t^d(n), \cdots, z_t^D(n)]^T$. More precisely, the training process involves iteration over the segmentation and the optimization steps until convergence. Given an estimate of Θ_{kull}, the segmentation step yields an optimal state sequence for each training utterance using Viterbi algorithm. The optimization step then estimates new set of model parameters given the optimal state sequences, i.e., alignment and $\mathbf{z}_t$ belonging to each of these states.

With $S_{\mathbf{RKL}}$ 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.,

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

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 $S_{\mathbf{KL}}$ 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.,

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

where y_i^{-d} 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 $S_{\mathbf{SKL}}$ 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).

References

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, 151–160, 1998.

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

R. Frost, Orthography and Phonology: The Psychological Reality of Orthographic Depth, Tech. Rep. SR-99/100, 162-171, Haskins Laboratories, 1989.

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

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

D. Wang, S. King, Letter-to-Sound Pronunciation Prediction Using Conditional Random Fields, IEEE Signal Processing Letters 18 (2) (2011) 122–125.

R. Rasipuram, M. Magimai-Doss, Acoustic Data-Driven Grapheme-to-Phoneme Conversion Using KL-HMM., in: Proceedings of ICASSP, 4841–4844, 2012a.

R. Kaplan, M. Kay, Regular Models of Phonological Rule Systems, Computational Linguistics 20 (1994) 331–378.

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

S. Jiampojamarn, G. Kondrak, T. Sherif, Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion, in: Proceedings of NAACL, 372–379, 2007.

T. J. Sejnowski, C. R. Rosenberg, Parallel Networks that Learn to Pronounce English Text, Complex Systems 1 (1987) 145–168.

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

P. Taylor, Hidden Markov Models for Grapheme to Phoneme Conversion., in: Proceedings of Interspeech, 1973–1976, 2005.

M. Bisani, H. Ney, Joint-Sequence Models for Grapheme-to-Phoneme Conversion, Speech Communication 50 (5) (2008) 434–451.

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

J. R. Novak, N. Minematsu, 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, 45–49, 2012.

P. Lehnen, S. Hahn, A. Guta, H. Ney, Incorporating Alignments Into Conditional Random Fields for Grapheme to Phoneme Conversion, in: Proceedings of ICASSP, 4916–4919, 2011.

S. Hahn, P. Lehnen, S. Wiesler, R. Schlter, H. Ney, Improving LVCSR with Hidden Conditional Random Fields for Grapheme-to-Phoneme Conversion., in: Proceedings of Interspeech, 495–499, 2013.

H. Strik, C. Cucchiarini, Modeling Pronunciation Variation for ASR: A Survey of the Literature., Speech Communication 29 (2-4) (1999) 225–246.

M. H. Cohen, Phonological Structures for Speech Recognition, Ph.D. thesis, University of California, Berkeley, 1989.

E. Fosler-Lussier, A Tutorial on Pronunciation Modeling for Large Vocabulary Speech Recognition., vol. 2705, Springer, 38–77, 2000.

H. Mokbel, D. Jouvet, Derivation of the Optimal Set of Phonetic Transcriptions for a Word from its Acoustic Realizations , Speech Communication 29 (1) (1999) 49 – 64.

M. Magimai-Doss, 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, 209–222, 2005.

M. Riley, A statistical model for generating pronunciation networks, in: Proceedings of ICASSP, vol. 2, 737–740, 1991.

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

L. Lu, A. Ghoshal, S. Renals, Acoustic Data-Driven Pronunciation Lexicon For Large Vocabulary Speech Recognition, in: Proceedings of ASRU, 374–379, 2013.

L. Xiao, A. Gunawardana, A. Acero, Adapting Grapheme-to-Phoneme Conversion for Name Recognition, in: Proceedings of ASRU, 130–135, 2007.

R. Rasipuram, M. Magimai.-Doss, Acoustic and Lexical Resource Constrained ASR using Language-Independent Acoustic Model and Language-Dependent Probabilistic Lexical Model, Speech Communication 68 (2015) 23–40.

X. Luo, F. Jelinek, Probabilistic Classification of HMM States for Large Vocabulary Continuous Speech Recognition, in: Proceedings of ICASSP, vol. 1, IEEE, 353–356, 1999.

J. Rottland, G. Rigoll, Tied posteriors: an approach for effective introduction of context dependency in hybrid NN/HMM LVCSR, in: Proceedings of ICASSP, 1241–1244, 2000.

G. Aradilla, J. Vepa, H. Bourlard, An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features, in: Proceedings of ICASSP, IV–657 – IV–660, 2007.

D. Imseng, R. Rasipuram, M. Magimai.-Doss, Fast and Flexible Kullback-Leibler Divergence based Acoustic Modeling for Non-native Speech Recognition, in: Proceedings of the ASRU, 2011.

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

G. Aradilla, H. Bourlard, M. M. Doss, Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task , in: Proceedings of Interspeech, 928–931, 2008.

J. Pitrelli, C. Fong, S. Wong, J. Spitz, H. Leung, PhoneBook: a Phonetically-Rich Isolated-Word Telephone-Speech Database, in: Proceedings of ICASSP, vol. 1, 101–104, 1995.

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

D. Imseng, et al., MediaParl: Bilingual Mixed Language Accented Speech Database, in: Proceedings of IEEE Workshop on SLT, 263–268, 2012a.

P. Taylor, A. Black, R. Caley, The Architecture of the Festival Speech Synthesis System, in: Proceedings of ESCA Workshop on Speech Synthesis, 1998.

D. Johnson, et al., ICSI Quicknet Software Package, http://www.icsi.berkeley.edu/Speech/qn.html, 2004.

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

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) (2012) 82–97.

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

D. Imseng, J. Dines, P. Motlicek, P. N. Garner, H. Bourlard, Comparing Different Acoustic Modeling Techniques for Multilingual Boosting, in: Proceedings of Interspeech, 2012b.

S. Young, et al., The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, UK, 2006.

M. Razavi, M. Magimai.-Doss, On Recognition of Non-Native Speech Using Probabilistic Lexical Model, in: Proceedings of Interspeech, 2014.

M. Magimai.-Doss, R. Rasipuram, G. Aradilla, H. Bourlard, Grapheme-based Automatic Speech Recognition using KL-HMM, in: Proceedings of Interspeech, 445–448, 2011.

M. Bisani, H. Ney, Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation, in: Proceedings of ICASSP, vol. 1, 409–412, 2004.

D. Jouvet, D. Fohr, I. Illina, Evaluating Grapheme-to-Phoneme Converters in Automatic Speech Recognition Context, in: Proceedings of ICASSP, 4821–4824, 2012.

R. Rasipuram, M. Razavi, M. Magimai-Doss, Probabilistic Lexical Modeling and Unsupervised Training for Zero-Resourced ASR, in: Proceedings of ASRU, 446–451, 2013.

M. Razavi, M. Magimai.-Doss, An HMM-Based Formalism for Automatic Subword Unit Derivation and Pronunciation Generation, in: Proceedings of ICASSP, 2015.

R. Veldhuis, The Centroid of the Symmetrical Kullback-Leibler Distance, IEEE Signal Processing Letters 9 (3) (2002) 96–99.