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) and text-to-speech systems is the 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. 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. In 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 performance of the proposed approach is not significantly different than the 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)

automatic speech recognition, lexicon development

1. Introduction

Speech technologies such as automatic speech recognition (ASR) and text-to-speech (TTS) systems aim to link two modes of communication, 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 easily trainable compared to the whole words as the frequency of words in a text follows Zipf's law[1](Powers, 1998), and 2) they are generalizable for unseen words.

The most widely used subword units in current speech processing systems are phones or phonemes[2]. Phonemes can be related to the spoken form (i.e., speech signal). More precisely, the envelope of magnitude spectrum of short-term speech signals typically depicts the characteristics of phonemes. They can also be related to the alphabetic written symbols (i.e., graphemes). The link between phonemes and graphemes originates from the alphabetic orthographies which aim to present the phonetic structure of the spoken words in a graphic form (Frost, 1989). The alphabetic orthographies can be deep or shallow depending on the language[3].

Typically, the development of phoneme-based speech technology systems consists of two steps: development of a lexicon consisting of a mapping between each word and its phoneme-based pronunciation followed by system training. The focus of this article is mainly on the phonemic lexicon development. A phonemic 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 evolution of languages and emergence of new words. Therefore, it is necessary to develop automatic pronunciation

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

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

[3]In shallow orthographies, the grapheme-to-phoneme correspondence is one-to-one (e.g., Finnish). In deep orthographies, however, the correspondence between the graphemes and phonemes is not direct. More precisely, the grapheme-to-phoneme relationship may be irregular (e.g., English) in which some prior knowledge about the word is required to accurately predict the relationship (i.e., different rules can be applied to various words). The grapheme-to-phoneme relationship may also be regular, i.e., predictable given a set of linguistic rules. However, accurate prediction of the grapheme-to-phoneme relationship in deep orthographies requires complex linguistic rules (e.g., French).

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*, typically data-driven and machine learning 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. The learned grapheme-to-phoneme relationship is then used to infer pronunciations for the unseen words. Most of the G2P approaches rely solely on 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 speech data along with word level transcriptions. 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). 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 ANNs 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 performance of the proposed approach is not significantly different than 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 a limited amount of data (for G2P model and acoustic model training) 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. Section 4 describes the databases along with the

evaluation setups in this study. Section 5 presents the pronunciation level setup, results and analysis. Section 6 provides the experimental setup and results at the ASR level. Finally Section 7 brings the conclusion.

2. Relevant literature

The first step towards building phoneme-based speech technology systems is the development of a phonemic lexicon. Phonemic pronunciations are typically hand-crafted by 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 an 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 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.

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[4]. 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 is processed sequentially in which for each grapheme, the corresponding phoneme (or phoneme 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 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 find a sequence of phonemes F^* that maximizes the posterior probability $P(F|G)$. Eqn. (1) can also be expressed as finding a sequence of phonemes F^* maximizing the joint probability $P(F, G)$ using the Bayes rule (Eqn. (2)). Various G2P approaches based on above expressions are described below:

  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:

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

$$= \arg\max_S P(G|S)P(S) \tag{4}$$

$$= \arg\max_S \prod_n P(g_n|s_n)P(s_n|s_{n-1}) \tag{5}$$

     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

[4]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.

(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 phonemes 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 phonemes F in Eqn. (2) is obtained based on the concept of graphones (Deligne et al., 1995). A graphone is a pair of a sequence of graphemes and a sequence of phonemes. Figure 1 shows a sequence of graphones for the word *phone* along with its pronunciation.

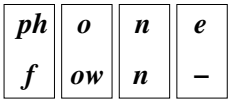


Figure 1: A possible sequence of graphones for the word *phone* and its associated pronunciation.

The joint probability $P(F, G)$ 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):

$$P(F, G) = \sum_{Q \in S(F,G)} p(Q) \tag{6}$$

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 best sequence of phonemes is obtained 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 Eqn. (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 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 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 sequence 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 in spontaneous speech when some of the sound units are dropped (Strik and Cucchiarini, 1999); or when a grapheme-to-phoneme converter trained on native pronunciations is used to extend the vocabulary of a non-native ASR system.

To overcome this limitation, in the context of pronunciation variation modeling, spoken examples of words are used to obtain pronunciation variants. 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. The phoneme recognizers can impose phonotactic constraints (Mokbel and Jouvet, 1999), exploit phone bigrams or trigrams (Fosler-Lussier, 2000), or be ergodic models (Magimai.-Doss and Bourlard, 2005). Possible alternate phoneme sequences for words are then obtained by finding the best alignment between the output of the phoneme recognizer and pronunciations provided by the seed lexicon.

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 illustrates the typical pronunciation variant extraction process.

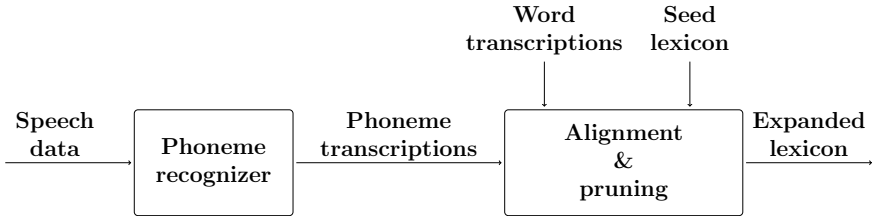


Figure 2: Pronunciation lexicon expansion with possible pronunciation variants for words obtained using speech samples.

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 an interest to prune the pronunciation variants obtained through a grapheme-to-phoneme converter using 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.

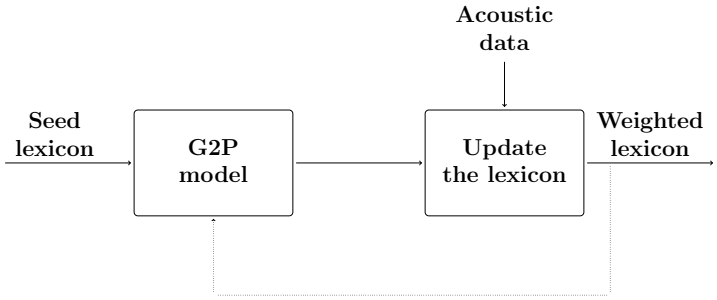


Figure 3: Acoustic data-driven G2P approaches proposed in the literature. The dotted line illustrates that some approaches iterate the G2P process.

In addition to the aforementioned approaches, 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 a 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^*:

$$S^* = \arg\max_{S \in \mathcal{S}} P(G, S|\Theta) \tag{7}$$

$$= \arg\max_{S \in \mathcal{S}} P(G|S, \Theta) P(S|\Theta) \tag{8}$$

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 assumptions, Eqn. (8) can be simplified as:

$$S^* = \arg\max_{S \in \mathcal{S}} \prod_{n=1}^{N} P(g_n|s_n = f_k, \Theta) P(s_n = f_k|s_{n-1} = f_{k'}, \Theta) \tag{9}$$

By applying the Bayes rule to Eqn. (9) we obtain:

$$S^* = \arg\max_{S \in \mathcal{S}} \prod_{n=1}^{N} \frac{P(s_n = f_k|g_n, \Theta) P(g_n|\Theta)}{P(s_n = f_k|\Theta)} P(s_n = f_k|s_{n-1} = f_{k'}, \Theta) \tag{10}$$

As $P(g_n|\Theta)$ does not affect the maximization, Eqn. (10) can be simplified as:

$$S^* = \arg\max_{S \in \mathcal{S}} \prod_{n=1}^{N} \underbrace{\frac{P(s_n = f_k|g_n, \Theta)}{P(s_n = f_k|\Theta)}}_{\text{local emission score}} \underbrace{P(s_n = f_k|s_{n-1} = f_{k'}, \Theta)}_{\text{transition probability}} \tag{11}$$

In Eqn. (11), 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 3.5) or as presented in the following section, it can be estimated by exploiting acoustic data which can bring certain advantages (discussed later in point 3 of Section 3.2.2).

3.2. Estimating $P(s_n = f_k|g_n)$ 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. Probabilistic lexical modeling

In a 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 either be 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 either be 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 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. More precisely, the parameter estimation is done using the Viterbi EM 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 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., $\mathbf{z}_t$ or $\mathbf{v}_t$ belonging to each of these states, the new set of parameters $\{\mathbf{y_i}\}_{i=1}^{I}$ is 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.

3.2.2. Relevance to G2P

The probabilistic lexical modeling framework brings certain advantages over the deterministic lexical modeling framework for learning the grapheme-to-phoneme relationship using acoustic information:

1. *The acoustic and lexical units can represent different types of subword units*: In the deterministic lexical modeling framework, as the acoustic and lexical units are deterministically related, they are constrained to be of the same type. For example, if the set of lexical units $\mathcal{L}$ is based on the phonemes (or graphemes), then the acoustic unit set $\mathcal{A}$ is also constrained to be based on phonemes (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 phonemes while the lexical units can represent graphemes (Rasipuram and Magimai.-Doss, 2015; Magimai.-Doss et al., 2011). In this case, the parameters of the lexical model $\mathbf{y}_i$ capture a probabilistic grapheme-to-phoneme 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 $\mathcal{L}$ is based on CI or CD subword units, then $\mathcal{A}$ 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 grapheme-to-phoneme correspondence which require modeling of longer grapheme contexts to correctly capture the relationship between graphemes and phonemes.

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 phonemes using acoustic and lexical resources from auxiliary languages, and then (b) learning a probabilistic relationship between graphemes of the target language and the multilingual phonemes using the target language acoustic data. Examples of similar work with the use of cross domain acoustic and lexical resources for grapheme-to-phoneme relationship learning can be found in (Magimai.-Doss et al., 2011; 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 grapheme-to-phoneme relationship. Thus, it can have potential implications for lexicon development for under-resourced languages and domains.

In this article, we exploit the advantages of the probabilistic lexical modeling framework to learn the grapheme-to-phoneme relationship through acoustic data. More precisely, we cast the parameter estimation problem for the HMM explained in Section 3.1 as learning the parameters $\{\mathbf{y}_i\}_{i=1}^{I}$ in the probabilistic lexical modeling framework in which the acoustic unit set $\mathcal{A}$ is equal to the set of phonemes $\mathcal{F} = \{f_1, \ldots, f_k, \ldots, f_K\}$ (in Section 3.1) and the lexical unit set $\mathcal{L}$ contains the possible graphemes in the target language (i.e., $\forall G_n = g_n : g_n \in \mathcal{L}$).

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 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 grapheme-to-phoneme relationships (Razavi et al., 2015a). However, selecting the best method for generating pronunciation variants is beyond the scope of this article.

3.4. Summary and implementation

Figure 4 provides a summary of the acoustic G2P approach using the probabilistic lexical modeling framework as a three-step process:

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

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

3. *Inference*: Given the trained lexical model and the orthographic transcription of the word, the most probable sequence of phonemes is inferred using the HMM framework in Section 3.1. The ergodic HMM in this article is implemented using the HTK toolkit (Young et al., 2006).

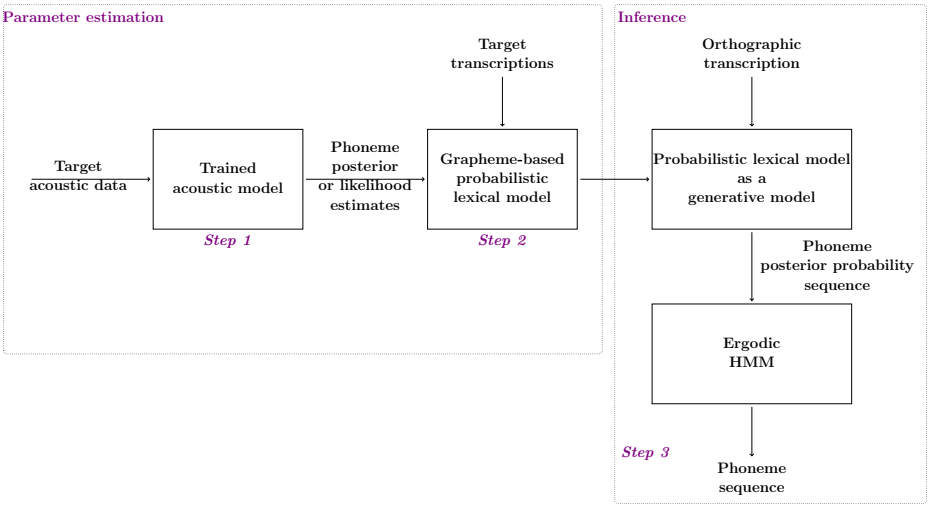


Figure 4: Block diagram of the acoustic G2P approach.

It can be seen that the recently proposed acoustic data-driven G2P 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 grapheme-to-phoneme relationship is captured through the parameters of the KL-HMM, i.e., a probabilistic lexical model.

As illustrated in Figure 5, 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), which can be estimated in different ways (Aradilla et al., 2008):

$$SC_{\mathbf{KL}}(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} y_i^d \log(\frac{y_i^d}{z_t^d}) \tag{12}$$

$$SC_{\mathbf{RKL}}(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} z_t^d \log(\frac{z_t^d}{y_i^d}) \tag{13}$$

$$SC_{\mathbf{SKL}}(\mathbf{y}_i, \mathbf{z}_t) = \frac{1}{2}(SC_{\mathbf{KL}} + SC_{\mathbf{RKL}}) \tag{14}$$

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

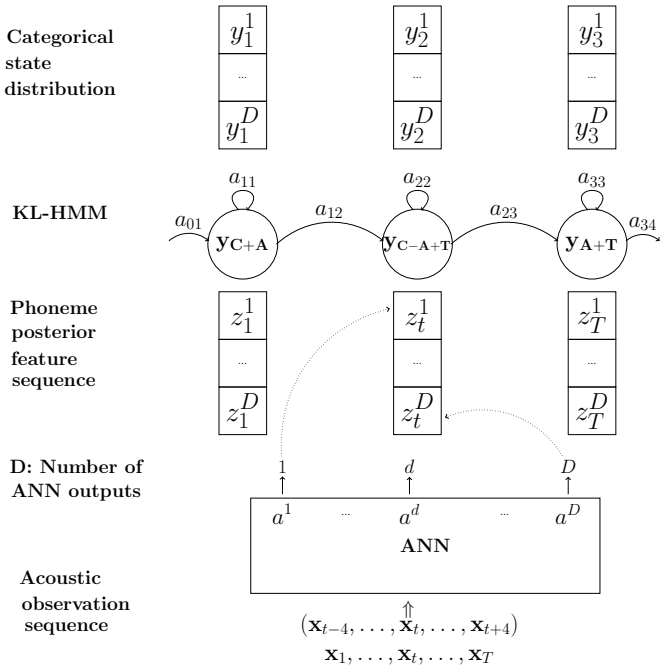


Figure 5: Illustration of KL-HMM approach in which graphemes are used as lexical units and the acoustic model is an ANN.

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).

3.5. Comparison to existing approaches

As explained in Section 3, in the acoustic G2P approach, the parameters of the probabilistic lexical model are estimated using the Viterbi EM algorithm as shown in Figure 6. Similar to the acoustic G2P approach, data-driven G2P 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 phoneme sequence is common to most of the G2P approaches.
- The *M-step* which captures the relationship between graphemes and phonemes is performed through different learning methods such as decision trees, neural networks, n-gram models or CRFs.

Table 1 further compares the acoustic G2P approach with the G2P approaches explained in Section 2 based on training criteria and required training data. The table also includes distinctive remarks in each approach.

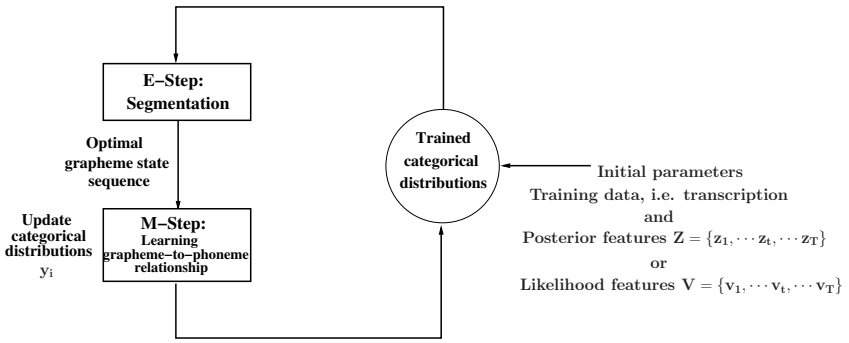


Figure 6: Illustration of parameter estimation in the probabilistic lexical modeling framework, where the acoustic units represent phonemes and lexical units represent graphemes.

The key distinctive factor in the acoustic G2P approach is exploiting acoustic data to learn the grapheme-to-phoneme relationship, in contrast to conventional data-driven G2P approaches which use only the seed lexicon. The proposed acoustic G2P approach is similar to the local classification based approaches, as they can be both seen as a particular case of the formalism in Section 3.1 where the transition and prior probabilities are uniform. In the local classification based approaches, the 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.

Table 1: Summary of different G2P approaches based on training criteria, required data and distinctive remarks.

Approach	Training criteria	Required data	Distinctive remarks
Local classification	Discriminative	Seed lexicon	Variation of the posterior-based approach in Eqn. (11) 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. (11) 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	Generative	Seed lexicon & Acoustic data	Exploits acoustic information to estimate $P(s_n = f_k\|g_n)$ in Eqn. (11).

In this article, we benchmark the acoustic G2P approach against two conventional G2P approaches: 1) decision tree-based G2P approach which like the acoustic G2P approach is a particular case of the HMM-based formalism

in Section 3.1, and 2) the state-of-the-art joint multigram G2P approach. We evaluate the G2P approaches on English and French as two languages with deep alphabetic orthographies.

4. Experimental setup

The performance of G2P approaches depends on various factors:

- *Language*: As explained in Section 1, alphabetic orthographies can be deep or shallow depending on the language. The G2P 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 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 this article, we considered the aforementioned factors thoroughly to design efficient experimental studies.

4.1. Datasets

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 (602 words) vocabularies. We use the medium size vocabulary task with 602 unique words (Dupont et al., 1997). The overview of the PhoneBook corpus is given in Table 2.

Table 2: 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	602

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 manually transcribed using 42 phonemes (including the phoneme *sil*). The manual lexicon contains only a single pronunciation per each word.

The G2P task on the PhoneBook corpus is challenging 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 there are only about 2000 training words and 10 hours of transcribed speech data available.

4.1.2. MediaParl: LVCSR bilingual corpus

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 study, we used the French part of the corpus as French is a challenging language for the G2P task due to its relatively complex grapheme-to-phoneme conversion rules 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 3 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.

Table 3: 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	10555	3376	4246

The preparation of the dictionary was started with the BDLex pronunciation lexicon (Imseng et al., 2012a)[5]. For the words that were not found in the

[5] `http://www.irit.fr/~Martine.deCalmes/IHMPT/ress_ling.v1/rbdlex_en.php`

BDLex dictionary, a WFST-driven G2P system was used to generate single-best pronunciations[6] and afterwards the generated pronunciations were hand-corrected. The manual dictionary of the French MediaParl corpus is in SAMPA format with a phoneme set of size 38 (including the phoneme *sil*) and contains all the words in the train, cross-validation and test set. The vocabulary size was 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 each 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. (2012a), liaison handling was not considered.

The G2P study on MediaParl corpus is different from the PhoneBook corpus for the following reasons:

- In French, the grapheme-to-phoneme relationship is regular (though the conversion rules can be complex), while in English the relationship is irregular.
- The amount of training data is bigger than for the PhoneBook corpus.
- The number of unseen words in the test set is relatively small (20% of the words in the test set).
- The MediaParl corpus contains not only spontaneous speech and debates but also non-native speech.

4.2. Evaluation

We used the G2P approaches to generate pronunciations for the words unseen during training. More precisely, the "G2P-based" lexicons in this article contain pronunciations from the manual dictionary for the words seen during training and the G2P-generated pronunciations for the unseen words. Towards 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 phoneme and word accuracy and analyzing the pronunciations using a confusion matrix. The pronunciation level studies are presented in Section 5. 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 6.

[6] http://code.google.com/p/phonetisaurus/

5. Pronunciation level studies

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

5.1. Pronunciation generation setup

We exploit the following G2P 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 approaches[7]. The hyper-parameters in each of the G2P approaches were tuned on the cross-validation set. The tuning on cross-validation set could possibly help in better generalization towards unseen contexts.

5.1.1. 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 phoneme accuracy 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 phonemes and hence smoothing the confidence scores (posterior probabilities) could be difficult (Wang and King, 2011). In this article, 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 per each unseen word in the PhoneBook and MediaParl corpora was 1.4 and 1.1 respectively.

5.1.2. Joint multigram approach

We used the Sequitur software developed at RWTH Aachen University[8]. 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.

[7]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 approaches could be suboptimal.

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

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 per each unseen word in the PhoneBook and MediaParl corpora was 4.9 and 2.7 respectively.

5.1.3. Acoustic G2P approach

The acoustic G2P approach includes three steps. In the first step, 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, 2012a,b). However, recent advances in speech technology have shown that ANNs with deep architectures can improve the performance of the ASR 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 the following MLPs with various number of layers and output units:

- *MLP-3-CI-M*: a three-layer MLP classifying M context-independent phonemes. For the PhoneBook corpus $M = 42$ and for the MediaParl corpus $M = 38$.
- *MLP-5-CI-M*: a five-layer MLP classifying CI phonemes.
- *MLP-5-CD-M*: a five-layer MLP modeling M clustered CD phonemes as outputs. The output units were derived by clustering CD phonemes in HMM/GMM framework using decision tree state tying. Various 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 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 context[9]) was trained.

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

The choice of local score to learn the KL-HMM parameters is important as previously shown in (Rasipuram and Magimai.-Doss, 2013). By using the local score $SC_{\mathbf{KL}}$, the system is better capable of capturing one-to-one grapheme-to-phoneme relationships. On the other hand, when using $SC_{\mathbf{RKL}}$ as the local score, the system can better handle one-to-many relationships. For the case when using $SC_{\mathbf{SKL}}$ 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 $SC_{\mathbf{RKL}}$ as it is suitable for the scenarios where the grapheme-to-phoneme relationship is irregular. For tying KL-HMM states we applied the KL-divergence based decision tree state tying method proposed by 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, silence was removed in the ergodic HMM as it could lead to deletion of some phonemes 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 inference step is demonstrated through the example word "MAP" in Figure 7.

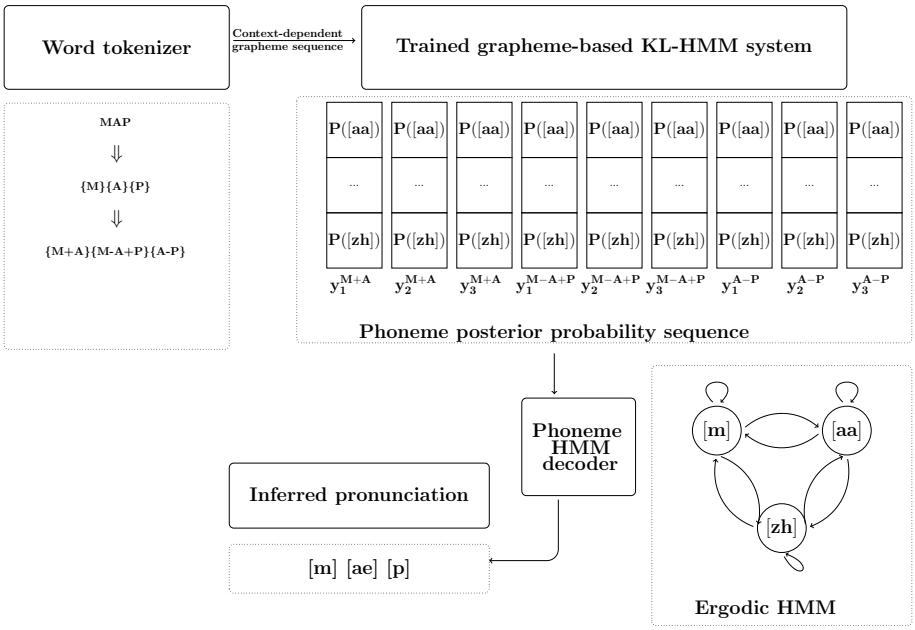
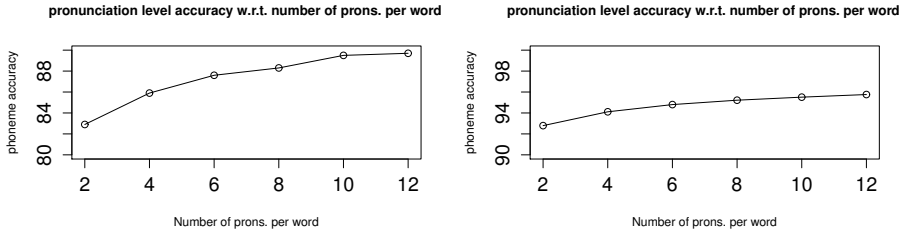


Figure 7: Block diagram of the inference phase in acoustic data-driven G2P task.

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 sequences.



(a) PhoneBook corpus

(b) MediaParl corpus

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

To resolve this issue, after training the KL-HMM, for each lexical unit l^i, the 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.

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 pronunciation level accuracy on the training words. Figure 8 shows the pronunciation level performance on the training words in terms of phoneme accuracy. For the PhoneBook corpus, it can be seen that when $N \geq 10$ the increase in the phoneme accuracy 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 phoneme and the spurious phonemes (consecutive appearance of the same phoneme) from the pronunciations. As a result of pruning, the number of unique pronunciations per each word was smaller than N. The average number of unique pronunciations per each unseen word in the PhoneBook and MediaParl corpora was 7.1 and 3.7 respectively.

5.2. Pronunciation level results

Table 4 provides pronunciation level evaluation results in terms of phoneme and word accuracy for different G2P approaches. To better analyze different G2P approaches, we have presented the results when generating pronunciations for the training words as well. For the acoustic G2P 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 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.

Table 4: Pronunciation level evaluations in terms of phoneme accuracy (PA) and word accuracy (WA) using different G2P approaches in the *single-best pronunciation* and *multiple pronunciation* scenarios. AG2P, JMM-G2P and DT-G2P represent acoustic G2P approach, joint multigram G2P approach and decision tree-based G2P approach respectively.

Approach	*Single-best pronunciation*		*Multiple pronunciation*	
	PA (WA) on train	PA (WA) on unseen	PA (WA) on train	PA (WA) 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)

(a) PhoneBook

Approach	*Single-best pronunciation*		*Multiple pronunciation*	
	PA (WA) on train	PA (WA) on unseen	PA (WA) on train	PA (WA) 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)

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

Finally, it is worth mentioning that the gap between the pronunciation level accuracy on the training and unseen words is significantly larger in the PhoneBook corpus compared to the MediaParl corpus. This is due to existence of uncommon words and availability of fewer amount of training data in the PhoneBook corpus, which makes generalizability of the G2P approaches towards unseen grapheme contexts more difficult.

5.3. Analysis

In this section, we provide the pronunciation level analysis for the joint multigram approach (as the state-of-the-art G2P approach) and the acoustic G2P approach using single-best pronunciations[10].

Table 5 shows examples of the phoneme confusions according to the confusion matrix of the generated pronunciations through acoustic G2P and joint multigram approaches for the PhoneBook corpus. It can be observed that most of the confusions come from vowel phonemes such as /E/ (as in the word "aber": /a/ /b/ /E/ /R/) which are confused with similar phonemes such as /x/ (as in the word "allow": /x/ /l/ /W/) in both G2P approaches. Confusions can also occur for consonant phonemes. For instance, the consonant phoneme /Z/ is confused with the phoneme /z/ and /S/ in the joint multigram and acoustic G2P approaches respectively. 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. It is interesting to note that the phoneme confusions in the two approaches can be different. For instance, in the acoustic G2P approach the phoneme /@/ 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 5: Examples of the phoneme confusions in the generated pronunciations through acoustic G2P (AG2P) and joint multigram (JMM-G2P) approaches for the PhoneBook corpus. The table presents phonemes together with their most confusable phonemes according to the confusion matrix.

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

Similarly for MediaParl, as shown in Table 6, it can be seen that the confusions are mostly related to vowel phonemes. For example, the phoneme /o/ (as in the word "ausse": /o/ /s/) is confused with the phoneme /O/ (as in the word "aussi": /O/ /s/ /i/) in both G2P approaches. Similar to the PhoneBook corpus, in the acoustic G2P approach the phoneme set size is reduced since the

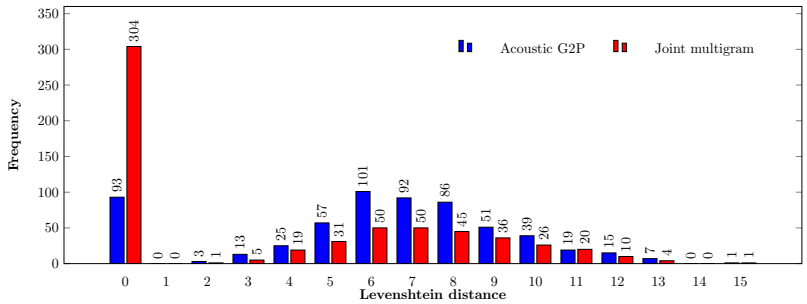
[10]The comparison is provided only for the single-best pronunciations, as the main goal in this section is to compare the potential of different G2P approaches, rather than investigating the effect of adding pronunciation variants.

phonemes /_6_/ and /_9_^/ are replaced with similar vowel phonemes. Furthermore, the phoneme confusions in the two approaches are different, similar to the observations in PhoneBook corpus. For instance, the phoneme /g/ is confused with the phonemes /k/ and /Z/ in the acoustic G2P approach and joint multigram approach respectively.

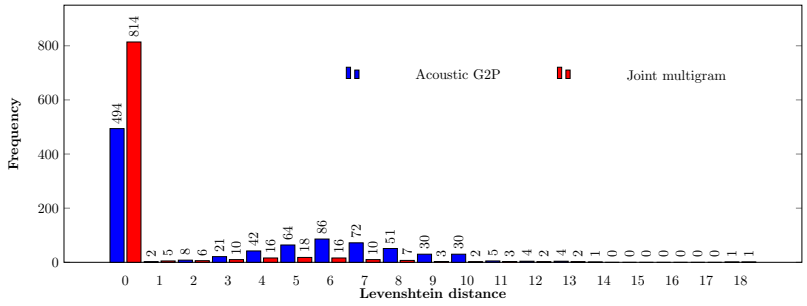
Table 6: Examples of the phoneme confusions in the generated pronunciations through acoustic G2P (AG2P) and joint multigram (JMM-G2P) approach for the MediaParl corpus. The table presents phonemes together with their most confusable phonemes according to the confusion matrix.

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

To analyze the performance of the acoustic G2P and joint multigram approaches in terms of word accuracy 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 9 depicts the results when using pronunciations derived from the acoustic G2P and joint multigram approaches.



(a) PhoneBook



(b) MediaParl

Figure 9: 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 and joint multigram approaches.

For the acoustic G2P 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 and joint multigram approaches, Tables 7 and 8 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 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 "anodin" 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 "examinerons" and "banale" compared to the acoustic G2P approach. As the joint multigram and acoustic G2P 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 approaches on the ASR performance in Section 6.2.

Table 7: Sample unseen words from the PhoneBook corpus along with their joint multigram-based (JMM-based), acoustic G2P-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	/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 8: Sample unseen words from the MediaParl corpus along with their joint multigram-based (JMM-based), acoustic G2P-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/ /ˆ/	/a/ /n/ /O/ /d/ /eˆ/
examinerons	/E/ /g/ /z/ /a/ /m/ /i/ /n/ /@/ /R/ /oˆ/	/E/ /z/ /a/ /m/ /i/ /n/ /E/ /R/ /oˆ/	/E/ /g/ /z/ /a/ /m/ /i/ /n/ /@/ /R/ /oˆ/
rèadaptation	/R/ /E/ /a/ /d/ /a/ /p/ /t/ /a/ /s/ /j/ /oˆ/	/R/ /E/ /a/ /d/ /a/ /t/ /a/ /s/ /j/ /oˆ/	/R/ /E/ /a/ /d/ /a/ /p/ /t/ /a/ /s/ /j/ /oˆ/
banale	/b/ /a/ /n/ /a/ /l/	/b/ /aˆ/ /n/ /a/ /l/	/b/ /a/ /n/ /a/ /l/

6. ASR level studies

We evaluated the G2P-based lexicons at the ASR level considering different facets:

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

2. Combination of different G2P approaches.

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

This section presents the ASR evaluation setup and results for each of these aspects. 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%.

Note that as explained in Section 4.2, the G2P-based lexicon contains pronunciations from the manual dictionary for the words seen during training, and G2P-generated pronunciations for the unseen words. 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 frequency words since, as explained in Section 4, 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.

6.1. Deterministic and probabilistic lexical modeling-based ASR studies

We used the HMM/GMM framework for the deterministic lexical modeling-based ASR studies. We trained standard cross-word CD HMM/GMM systems using the manual dictionary with 39 dimensional PLP cepstral features $\mathbf{x}_t$ extracted using HTK toolkit (Young et al., 2006). During decoding, the G2P-based lexicons were used.

For the probabilistic lexical modeling-based ASR studies, we trained KL-HMM systems using the manual dictionary with 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 $SC_{\mathbf{SKL}}$ as the local score. For tying KL-HMM (lexical) states we applied the KL-divergence-based decision tree state tying method.

It is worth mentioning that in deterministic lexical modeling based ASR framework, it is common to use ANNs that classify clustered CD phoneme states, referred to as CD neural networks (CDNNs), instead of GMMs as acoustic model to estimate emission likelihoods (Hinton et al., 2012). In recent works, it has been shown that the performance of a KL-HMM system using an ANN classifying CI phonemes is not significantly different than the HMM/CDNN ASR system (Razavi et al., 2014; Razavi and Magimai.-Doss, 2014). Therefore in this study, we limited the deterministic lexical model studies to only HMM/GMM systems, as the performance of a KL-HMM system using CI phonemes as acoustic units could be already an indicative of the performance of an HMM/CDNN ASR system.

Table 9 presents the performance of HMM/GMM and KL-HMM systems in terms of word accuracy (100 - ASR word error rate) using single-best and multiple pronunciations from different G2P 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.

Table 9: Performance of HMM/GMM and KL-HMM systems in terms of word accuracy using different G2P approaches. AG2P, JMM-G2P and DT-G2P represent acoustic G2P approach, joint multigram G2P approach and decision tree-based G2P approach respectively.

G2P approach	*Single-best pronunciation*		*Multiple pronunciation*	
	HMM/GMM	KL-HMM	HMM/GMM	KL-HMM
AG2P-*MLP-3-CI-42*	81.1	84.8	86.8	88.3
AG2P-*MLP-5-CI-42*	82.1	85.1	87.3	89.0
AG2P-*MLP-5-CD-321*	82.9	85.0	88.7	89.4
JMM-G2P	88.8	89.4	93.0	93.7
DT-G2P	85.2	86.9	86.9	88.5
Manual dictionary	98.2	98.2	98.2	98.2

(a) PhoneBook

G2P approach	*Single-best pronunciation*		*Multiple pronunciation*	
	HMM/GMM	KL-HMM	HMM/GMM	KL-HMM
AG2P-*MLP-3-CI-38*	72.0	73.3	72.5	73.6
AG2P-*MLP-5-CI-38*	72.0	73.2	72.5	73.7
AG2P-*MLP-5-CD-437*	72.2	73.3	72.6	73.6
JMM-G2P	73.1	74.0	73.2	74.0
DT-G2P	73.0	73.8	73.1	74.0
Manual dictionary	73.2	74.1	73.2	74.1

(b) MediaParl

6.1.1. ASR results using single-best pronunciations

For the acoustic G2P approach, it can be observed from Table 9 that similar to the pronunciation level results in Table 4, 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.

The performance of the acoustic G2P 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 approaches perform 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 irregular compared to French, it may require modeling of more than single preceding and single following grapheme context.

- Discrepancy between the manually-generated and G2P-generated pronunciations: As it can be seen from Table 4, the word accuracy at the pronunciation level for the acoustic G2P approach is poor (in particular in the PhoneBook corpus). This is partly due to replacement of vowel phonemes with similar vowels as observed in Tables 5 and 6. As a consequence, the phoneme contexts seen in the manual lexicon which are used for ASR system training are

different from the phoneme 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 10 presents the ASR performance in terms of word accuracy.

Table 10: Performance of ASR systems in terms of word accuracy 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 approach, joint multigram G2P approach and decision tree-based G2P approach respectively.

G2P Approach	Word accuracy	
	HMM/GMM	KL-HMM
AG2P-*MLP-5-CD-321*	88.3	88.8
JMM-G2P	89.1	89.4
DT-G2P	88.5	88.1
Manual dictionary	98.2	98.2

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). Furthermore, it can be seen that the difference between the ASR performance of the acoustic G2P and conventional G2P approaches is not statistically significant when using G2P-generated pronunciations in both train and test lexicons.

As it can be seen in Table 9, the KL-HMM systems generally perform better than HMM/GMM systems in both databases. This observation is consistent with our previous studies (Razavi et al., 2014; Razavi and Magimai.-Doss, 2014). In addition, the gap between the performance of different systems is relatively less in the KL-HMM framework compared to the HMM/GMM approach. The improvements in the performance and reduction in the gap can be attributed

to the probabilistic lexical modeling framework of KL-HMM, which enables capturing the possible inconsistencies and variations in pronunciations.

6.1.2. ASR results using multiple pronunciations

As it can be observed from Table 9, for the PhoneBook corpus, using multiple pronunciations leads to significant improvements over single-best pronunciations in the ASR word accuracy for all the G2P approaches. Furthermore, through use of multiple pronunciations, the gap between the acoustic G2P approach and conventional G2P 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 11 presents the ASR performance in terms of word accuracy. It can be seen that the G2P approaches can benefit from using G2P-generated pronunciations in both train and test lexicons.

Table 11: Performance of ASR systems in terms of word accuracy 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 approach, joint multigram G2P approach and decision tree-based G2P approach respectively.

G2P Approach	Word accuracy	
	HMM/GMM	KL-HMM
AG2P-*MLP-5-CD-321*	91.6	91.7
JMM-G2P	93.2	93.7
DT-G2P	89.8	89.2
Manual dictionary	98.2	98.2

6.2. Combination of G2P approaches

As discussed earlier in Section 3.5, different G2P approaches exploit different resources and techniques to learn the grapheme-to-phoneme relationship and infer pronunciations. It would be interesting to investigate whether combination of pronunciation lexicons obtained through various G2P approaches can bring any benefits for the ASR systems. Table 12 presents the average number of unique pronunciations per each unseen word for the PhoneBook and MediaParl corpora when combining G2P-based lexicons. The results show that combining the acoustic G2P approach with a conventional G2P approach leads to more diverse pronunciations than combination of conventional G2P approaches.

Table 13 reports the ASR performance of HMM/GMM and KL-HMM systems in terms of word accuracy when combining pronunciations from different G2P approaches. Similar to experimental studies in Section 6.1, we present the ASR results using a combination of single-best pronunciations and multiple pronunciations from each of the G2P approaches separately.

Table 12: Average number of pronunciations per unseen word obtained through combining different G2P 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 approaches. The second column shows the average number of pronunciations when combining pronunciation variants generated from each of the G2P approaches. AG2P, DT-G2P and JMM-G2P represent acoustic G2P approach, decision tree based G2P approach and joint multigram G2P approach respectively.

G2P Approach Combinations	PhoneBook		MediaParl	
	Comb. of 1-best G2P-based prons.	Comb. of multiple G2P-based prons.	Comb. of 1-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

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

For the PhoneBook corpus, significant improvements in terms of ASR word accuracy are achieved through combination of the G2P approaches compared to the case using single-best pronunciations from a G2P approach (Table 9).

For the MediaParl corpus, it can be seen that the systems using the lexicon obtained from combination of G2P approaches yield the same performance as the system using the manual dictionary. However, compared 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 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 9, 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 the same as the performance of the systems using multiple pronunciations through combination of single-best G2P-based pronunciations from various G2P approaches (with 2.4 and 1.6 pronunciations per unseen word in the PhoneBook and MediaParl respectively). 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 approach, but with a fewer number of pronunciation variants.

Table 13: ASR performance in terms of word accuracy when combining pronunciations from different G2P approaches. AG2P, JMM-G2P and DT-G2P represent acoustic G2P approach, joint multigram G2P approach and decision tree-based G2P approach respectively.

G2P approach	Combination of 1-best G2P-based pronunciations		Combination of multiple G2P-based pronunciations	
	HMM/GMM	KL-HMM	HMM/GMM	KL-HMM
AG2P-*MLP-5-CD-321* +JMM-G2P	91.7	92.1	94.9	95.0
AG2P-*MLP-5-CD-321* +DT-G2P	90.4	91.6	92.8	93.4
JMM-G2P +DT-G2P	92.6	93.0	94.5	95.0
AG2P-*MLP-5-CD-321* + JMM-G2P +DT-G2P	93.2	93.7	95.1	95.4
Manual dictionary	98.2	98.2	98.2	98.2

(a) PhoneBook

G2P approach	Combination of 1-best G2P-based pronunciations		Combination of multiple G2P-based pronunciations	
	HMM/GMM	KL-HMM	HMM/GMM	KL-HMM
AG2P-*MLP-5-CD-437* +JMM-G2P	73.1	74.2	73.2	74.1
AG2P-*MLP-5-CD-437* +DT-G2P	73.1	74.1	73.1	74.0
JMM-G2P +DT-G2P	73.2	74.1	73.3	74.1
AG2P-*MLP-5-CD-437* +JMM-G2P +DT-G2P	73.2	74.1	73.1	74.0
Manual dictionary	73.2	74.1	73.2	74.1

(b) MediaParl

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

It can be seen from Table 13 that for the PhoneBook corpus, a combination of pronunciation variants from each of the G2P 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 approach (Table 9). This can indicate that different G2P approaches bring complementary information to one another. For the MediaParl corpus, similar to the observations in Section 6.2.1, the combination of G2P 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.

6.3. 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 the lexicon obtained from combination of G2P approaches. The KL-HMM systems were built in the same setup explained in Section 6.1.

Table 14 presents the ASR results in terms of word accuracy. 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 approaches or from a single G2P approach.

Table 14: 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 (G2P)-based KL-HMM
PhoneBook	93.6	95.4
MediaParl	71.7	74.2

7. Conclusions

In this article, we presented a novel HMM-based G2P formalism in which the grapheme-to-phoneme relationship is locally modeled as a distribution of phoneme 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 approach where the grapheme-to-phoneme relationship is learned through acoustics. Furthermore, the existing local classification-based G2P approaches based on decision trees and ANNs can be seen as a particular case of this formulation.

We compared the proposed acoustic G2P approach against the conventional G2P approaches on two different languages with deep alphabetic orthographies and considered using both single-best pronunciations and multiple pronunciations per word. The studies showed that the acoustic G2P-based lexicon performs poor at the pronunciation level compared to conventional G2P 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 approaches reduces. Furthermore, despite the relatively poor performance at the pronunciation-level, the ASR system using the acoustic G2P-based lexicon performs statistically similar to the system using a lexicon

from conventional G2P approaches. Though the proposed acoustic G2P approach does not outperform the state-of-the-art G2P methods, it can bring two main advantages. We discuss them briefly below.

As observed through experimental studies, the acoustic G2P approach can bring complementary information compared to state-of-the-art G2P approaches. i.e., combination of lexicons from the acoustic G2P approach and conventional approaches can yield better ASR systems. In this article, we combined the proposed acoustic G2P approach and conventional G2P approaches at the lexicon level. However, the HMM-based G2P formulation can be effectively exploited to combine different streams of estimates of probability of phonemes given graphemes obtained from various G2P approaches during pronunciation inference (Razavi and Magimai.-Doss, 2015a).

Conventional G2P approaches necessitate the availability of a seed lexicon in the target language or domain. The proposed approach alleviates the need for a seed lexicon from the target language or domain given some amount of word level transcribed speech data. Specifically as discussed earlier in Section 3.2.2 (Point 3), the acoustic-to-phoneme relationship ($\mathbf{z}_t$ estimator) can be learned on language-independent data and the grapheme-to-phoneme relationship ($\{\mathbf{y}_i\}_{i=1}^{I}$) can be learned on target language or domain speech data. This could be potentially exploited to:

- Develop lexicons for new domains such as names, child speech and accented speech. For example, in the case of accented speech the proposed approach could be employed in the following manner. The acoustic-to-phoneme relationship (i.e., ANN) is learned on large amount of native speech while the grapheme-to-phoneme relationship (i.e., KL-HMM) is learned on available non-native speech. Given the learned grapheme-to-phoneme relationships the pronunciations are then inferred.

- Develop lexicons for under-resourced languages which lack phonetic lexicons. In such a case, we could employ the proposed acoustic G2P approach in the following manner. The acoustic-to-multilingual phoneme relationship is learned using auxiliary acoustic and lexical resources from resource-rich languages, and the grapheme-to-multilingual phoneme relationship is learned using target language speech data. The pronunciations are then finally inferred given the learned grapheme-to-multilingual phoneme relationships. In other words, the proposed approach provides a means to exploit lexical knowledge and acoustic resources from resource-rich languages towards development of lexicons for under-resourced languages.

Furthermore, the proposed acoustic G2P approach can be exploited for developing lexicons based on automatically derived subword units (Razavi and Magimai.-Doss, 2015b; Razavi et al., 2015b).

The proposed acoustic G2P approach has also room for further improvements. In this article, we investigated generating pronunciation variants through N-best decoding at the inference stage. However, multiple pronunciations can also be generated by employing different cost functions at the learning

stage (Razavi et al., 2015a). Our future work, in addition to the above discussed potential future directions, will also further investigate pronunciation variants generation and pronunciation pruning strategies for the proposed acoustic G2P approach.

Acknowledgment

This work was partly 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). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

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,

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

where $q_t \in \{1, \cdots, I\}$, $\mathcal{Q}$ denotes the 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 a new set of model parameters given the optimal state sequences, i.e., alignment and $\mathbf{z}_t$ belonging to each of these states.

With $SC_{\mathbf{RKL}}$ 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 $SC_{\mathbf{KL}}$ 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 $SC_{\mathbf{SKL}}$ 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 IEEE International Conference on Acoustics, Speech and Signal Processing (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 North American Chapter of the Association for Computational Linguistics (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 International Conference on 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. Deligne, F. Yvon, F. Bimbot, Variable-Length Sequence Matching for Phonetic Transcription Using Joint Multigrams, in: Proceedings of European Conference on Speech Communication and Technology, EUROSPEECH, 1995.

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. Schlüter, 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.

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.

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

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 Transactions 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 IEEE Automatic Speech Recognition and Understanding Workshop (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.

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. 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.

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

M. Razavi, R. Rasipuram, M. Magimai.-Doss, Towards Multiple Pronunciation Generation in Acoustic G2P Conversion Framework, Idiap-RR Idiap-RR-34-2015, Idiap, 2015a.

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

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 Spoken Language Technology, 263–268, 2012a.

K. Livescu, E. Fosler-Lussier, F. Metze, Subword Modeling for Automatic Speech Recognition: Past, Present, and Emerging Approaches, IEEE Signal Processing Magazine 29 (6) (2012) 44–57.

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

S. Hahn, P. Vozila, M. Bisani, Comparison of Grapheme-to-Phoneme Methods on Large Pronunciation Dictionaries and LVCSR Tasks, in: Proceedings of Interspeech, 2538–2541, 2012.

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 Idiap-RR-15-2013, 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.

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

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

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

M. Razavi, M. Magimai.-Doss, Posterior-Based Multi-Stream Formulation To Combine Multiple Grapheme-to-Phoneme Conversion Techniques, Idiap-RR Idiap-RR-33-2015, Idiap, 2015a.

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

M. Razavi, R. Rasipuram, 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.

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