IDIAP RESEARCH REPORT

TOWARDS MULTIPLE PRONUNCIATION GENERATION IN ACOUSTIC G2P CONVERSION FRAMEWORK

Marzieh Razavi Ramya Rasipuram
Mathew Magimai.-Doss

Idiap-RR-34-2015

OCTOBER 2015

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

TOWARDS MULTIPLE PRONUNCIATION GENERATION IN ACOUSTIC G2P CONVERSION FRAMEWORK

Marzieh Razavi[1,2], Ramya Rasipuram[1] and Mathew Magimai Doss [1]

[1] Idiap Research Institute, CH-1920 Martigny, Switzerland
[2] École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
{mrazavi, rramya, mathew}@idiap.ch

ABSTRACT

Recently an acoustic data-driven grapheme-to-phoneme (G2P) conversion approach has been proposed in which the G2P relationship is learned through acoustic data, and the learned relationship together with the orthographic transcription of the word are used to infer pronunciations. This paper extends the acoustic G2P conversion approach and proposes two methods for generating multiple pronunciations to better handle pronunciation variations. In the first method, multiple pronunciations are generated by using different cost functions at the learning stage to possibly capture different G2P relationships. The second method generates multiple pronunciations at the inference stage through N-best decoding. Our experimental studies on Phonebook task in English show that (a) the first method yields lower average number of pronunciations per word than the second method; and (b) both methods, without pronunciation selection or pruning, lead to improvements in the performance at the pronunciation level as well as the speech recognition level.

Index Terms— Kullback-Leibler divergence based HMM, Lexicon generation, multiple pronunciations, grapheme-to-phoneme conversion, pronunciation variability

1. INTRODUCTION

Grapheme-to-phoneme (G2P) conversion is the task of generating pronunciations for words given their orthographic transcriptions. It has applications in several human language technology systems such as automatic speech recognition (ASR), text-to-speech (TTS) and spelling correction. G2P conversion is a challenging problem as the pronunciation for a given word can vary depending on various factors such as the native language of the speaker, the origin of the word and the conversational context. One way to handle such variations is to include multiple alternative pronunciations of the same word in the pronunciation lexicon.

For generating pronunciations and their variants, several approaches have been proposed in the literature. They can usually be classified as either knowledge-based or data-driven approaches. In knowledge-based approaches, linguistic rules derived by humans (e.g., through linguistic studies) are exploited in order to convert the sequence of graphemes in a word to a sequence of phonemes and also to generate alternative pronunciations [1]. The drawback of such approaches is their dependence on human effort and linguistic information which may not be available for the target language (for example in the case of low-resourced languages). To resolve this issue, data-driven approaches have been proposed [2, 3, 4, 5, 6].

In data-driven approaches, an initial phoneme lexicon is used to learn the relationship between graphemes and phonemes through statistical methods such as artificial neural networks and joint n-grams [5, 4]. The data-driven G2P conversion approaches typically enable generating multiple pronunciations through n-best search [4]. However, not all the generated pronunciations are phonologically relevant. To address this issue, other approaches have been proposed which exploit acoustic information to select or weigh the pronunciations learned through G2P conversion techniques [7, 8]. These approaches, require acoustic samples for each word that they want to generate pronunciations.

Recently a novel data-driven G2P conversion approach has been proposed which exploits acoustic information to *learn* the relationship between graphemes and phonemes [9]. The acoustic G2P conversion approach involves a learning phase followed by the inference phase (*Section 2*). In the learning phase of the approach, the relationship between graphemes and phonemes is learned in the framework of Kullback-Leibler divergence based hidden Markov model (KL-HMM) through acoustic data. Then, in the inference phase, given the learned G2P relationship together with the orthographic transcription of the word, the most probable pronunciation is inferred. Previous studies have shown that the acoustic G2P conversion approach performs comparable to state-of-the-art G2P conversion approaches at the ASR level [10]. Moreover, it can potentially relax the need for an initial seed lexicon in the target domain [9] or language [10] which makes it suitable for scenarios where limited acoustic and lexical resources are available [11].

In the previous studies [9, 11, 10], the acoustic G2P conversion approach was assessed using single-best pronunciations. In this paper, we extend the acoustic G2P conversion approach by investigating its potential in generating multiple pronunciations. More specifically, we augment the acoustic G2P conversion approach by generating multiple pronunciations in two ways: 1) By using different cost functions at the learning phase in order to learn different G2P relationships, and 2) By inferring n-best pronunciations at the inference stage to capture alternative phonologically relevant pronunciations (*Section 3*).

We study the proposed methods to generate multiple pronunciations on a scenario where limited transcribed speech data is available along with its pronunciation lexicon and the goal is to augment the pronunciation lexicon with new unseen words (*Section 4*). Our studies show that the acoustic G2P conversion approach is indeed capable of generating multiple pronunciations which can improve the performance at both pronunciation and ASR levels. Furthermore, multiple pronunciations generated at the learning stage, with fewer number of average pronunciations per word and lower pronunciation level performance, lead to better ASR performance compared to the pronunciations generated through N-best decoding (*Section 5*).

2. BACKGROUND

In this section, we describe the learning and inference phases of the acoustic data-driven G2P conversion approach.

2.1. Learning Phase

In the learning phase, as the first step the relationship between acoustic feature observations $\mathbf{x}_t$ and acoustic units $\{a^d\}_{d=1}^D$ is learned through an acoustic model, such as an artificial neural network (ANN). The acoustic units can be either context-independent phonemes or clustered context-dependent phonemes. In this section, for the sake of clarity we assume the acoustic units $\{a^d\}_{d=1}^D$ are (context-independent) phonemes.

As the second step of the learning phase, the relationship between the graphemes and phonemes is learned in the KL-HMM framework in which [12, 13]:

1. The posterior probabilities of phonemes $\mathbf{z}_t$ estimated from the acoustic model are used as feature observations. More precisely, $\mathbf{z}_t = [z_t^1, \ldots, z_t^d, \ldots, z_t^D]^{\mathrm{T}}$ where $z_t^d = P(a^d|\mathbf{x}_t)$.
2. The KL-HMM states $\{l^i\}_{i=1}^I$ represent context-dependent grapheme states. Each HMM state is parameterized by a categorical distribution $\mathbf{y}_i = [y_i^1, \ldots, y_i^d, \ldots, y_i^D]^{\mathrm{T}}$ with $y_i^d = P(a^d|l^i)$ which models the relationship between the phonemes $\{a^d\}_{d=1}^D$ and the context-dependent grapheme state l^i. We refer to the set of categorical distributions $\{\mathbf{y}_i\}_{i=1}^I$ as the lexical model.
3. The KL-HMM parameters $(\{\mathbf{y_i}\}_{i=1}^I)$ capture the probabilistic relationship between graphemes and phonemes. To learn these parameters, a local score is defined at each state based on the KL-divergence between the categorical distribution $\mathbf{y}_i$ and the phoneme posterior feature $\mathbf{z}_t$. As KL-divergence is not a symmetric measure, the local score can be estimated in three ways:

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

$$S_{\mathbf{RKL}}(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} z_t^d \log(\frac{z_t^d}{y_i^d}) \tag{2}$$

$$S_{\mathbf{SKL}}(\mathbf{y}_i, \mathbf{z}_t) = \frac{1}{2}(S_{\mathbf{KL}} + S_{\mathbf{RKL}}) \tag{3}$$

4. The parameters are estimated through iterations of Viterbi segmentation and optimization steps until convergence. More precisely, in the segmentation step , the optimal state sequence for each training utterance is obtained by using the Viterbi algorithm which minimizes a cost function based on one of the KL-divergence based local scores:

$$\min_{Q \in \mathcal{Q}} \sum_{t=1}^{T} [S(\mathbf{y}_{q_t}, \mathbf{z}_t) - \log a_{q_{t-1}q_t}] \tag{4}$$

in which $q_t \in \{1, \cdots, I\}$, $\mathcal{Q}$ denotes set of all possible HMM state sequences, $Q = \{q_1, \cdots, q_t, \cdots, q_T\}$ denotes a sequence of HMM states and $a_{q_{t-1}q_t}$ denotes the transition probability from state q_{t-1} to state q_t. Depending on the local score S being $S_{\mathbf{KL}}$, $S_{\mathbf{RKL}}$, or $S_{\mathbf{SKL}}$ different cost functions can be obtained.

In the optimization step, given the alignment and $\mathbf{z}_t$ belonging to each state, the new set of parameters are estimated. For the cost function based on $S_{\mathbf{KL}}$, the optimal state distribution is the normalized geometric mean of the training acoustic state probability vectors assigned to the state, i.e.,

$$y_i^d = \frac{y_i^{-d}}{\sum_{d=1}^D y_i^{-d}} \quad where \quad y_i^{-d} = (\prod_{\mathbf{z}_t(n) \in Z(i)} z_t^d(n))^{\frac{1}{M(i)}} \quad \forall n,t \tag{5}$$

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 $Z(i)$.

For the cost function based on $S_{\mathbf{RKL}}$, the optimal state distribution is the arithmetic mean of the training acoustic state probability vectors assigned to the state, i.e.,

$$y_i^d = \frac{1}{M(i)} \sum_{\mathbf{z}_t(n) \in Z(i)} z_t^d(n) \quad \forall n, t \tag{6}$$

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

For the cost function based on $S_{\mathbf{SKL}}$, 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 [14].

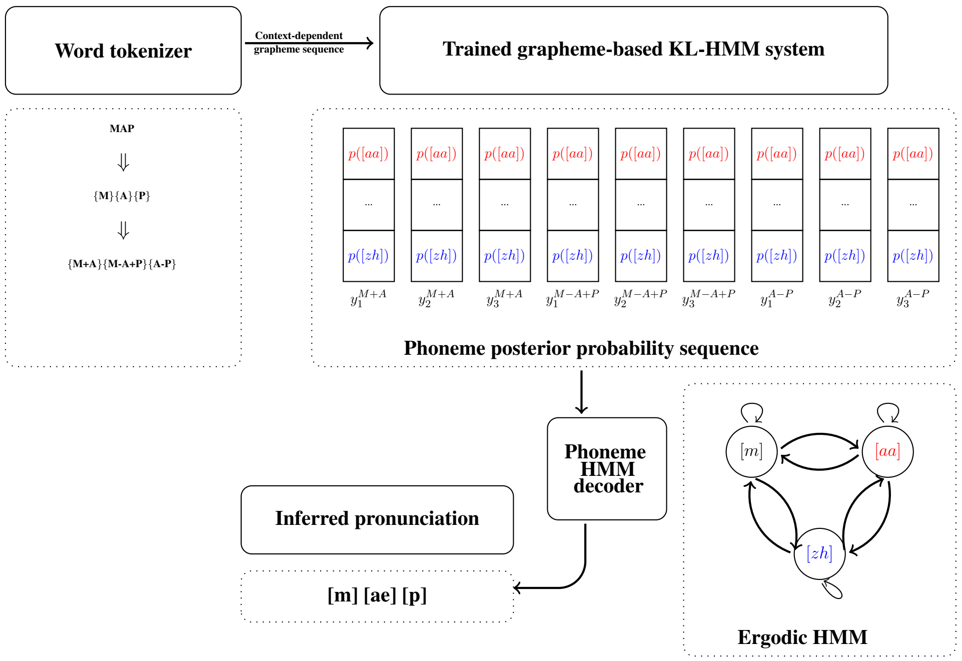
Word tokenizer
Context-dependent grapheme sequence
Trained grapheme-based KL-HMM system
MAP
⇓
{M}{A}{P}
⇓
{M+A}{M-A+P}{A-P}
$p([aa])$
...
$p([zh])$
y_1^{M+A} y_2^{M+A} y_3^{M+A} y_1^{M-A+P} y_2^{M-A+P} y_3^{M-A+P} y_1^{A-P} y_2^{A-P} y_3^{A-P}
Phoneme posterior probability sequence
Phoneme HMM decoder
Inferred pronunciation
[m] [ae] [p]
$[m]$
$[aa]$
$[zh]$
Ergodic HMM

Fig. 1. Block diagram of the inference phase in acoustic data-driven G2P conversion framework.

2.2. Inference Phase

In the inference phase, given the orthographic representation of the word together with the parameters of the KL-HMM, the most probable phoneme sequence is inferred. More precisely, as illustrated in Figure 1, the inference stage involves the following three steps:

Part (***A***): A given word is tokenized into its context-independent and then to its context-dependent graphemes.

Part (***B***): A sequence of phoneme posterior probability vectors is generated by concatenating the tokenized context-dependent graphemes and the trained KL-HMM parameters together.

Part (***C***): The most probable phoneme sequence is inferred by decoding the sequence of phoneme posterior probabilities using an ergodic HMM in which each state represents a phoneme.

The viability of the acoustic G2P conversion approach has been investigated in previous studies [9, 11, 10]. It was shown that despite relatively poor performance at the pronunciation level, the acoustic G2P conversion approach can perform comparable to the state-of-the-art G2P conversion approaches at the ASR level [10].

In the previous studies, the acoustic G2P conversion approach was limited to generate one pronunciation per word. However, it has been shown that ASR systems can possibly benefit from use of multiple pronunciations to better model the pronunciation variations [15]. In this paper we propose two approaches to generate multiple pronunciations in the acoustic G2P conversion framework.

3. MULTIPLE PRONUNCIATION GENERATION

In this section, we describe the proposed multiple pronunciation generation methods used at the learning stage and the inference stage of the acoustic G2P conversion framework.

3.1. Multiple Pronunciation Generation at Learning Phase

As explained in Section 2, the parameters of the KL-HMM system (capturing the probabilistic G2P relationship) are estimated by minimizing the KL-divergence between the state distributions and the posteriors belonging to the state. Depending on the local score (i.e., $S_{\mathbf{KL}}$, $S_{\mathbf{RKL}}$ or $S_{\mathbf{SKL}}$) used during parameter estimation, the state distribution is different:

1. According to Equation (5), the optimal state distribution for the local score $\mathbf{S_{KL}}$ is the geometric mean of probability vectors assigned to a state. It has been shown that using geometric mean for aggregation of probabilities leads to a less dispersive distribution [16, 17] (i.e., it is biased toward the dominant decision). As a result, it is expected that the cost function based on local score $\mathbf{S_{KL}}$ can better capture the one-to-one G2P relationships, as previously observed in [18] as well.
2. According to Equation (6), the optimal state distribution for the local score $\mathbf{S_{RKL}}$ is the arithmetic mean of probability vectors assigned to a state. It has been shown that using arithmetic mean for aggregation of probabilities leads to a dispersive distribution [16, 17] (i.e., it captures competing decisions). Therefore, it is expected that the cost function based on local score $\mathbf{S_{RKL}}$ can better capture the one-to-many G2P relationships, as also observed in [18] .

3. In the cost function based on the local score $\mathbf{S_{SKL}}$, both $S_{\mathbf{KL}}$ and $S_{\mathbf{RKL}}$ update steps are involved. It has been observed that the system based on the local score $\mathbf{S_{SKL}}$ can capture both one-to-one and one-to-many G2P relationships to a certain extent [18].

In other words, each cost function is capable of capturing a particular type (i.e., one-to-one or one-to-many) of G2P relationship. It is interesting to note that different types of G2P relationships can exist within a language. For instance, in English the grapheme [B] can be related to a single sound unit /b/ while [A] can be related to multiple sound units. Therefore, in order to capture different kinds of G2P relationships we could exploit different cost functions during KL-HMM training. The KL-HMM parameters learned through different cost functions can then be used to generate multiple pronunciations.

3.2. Multiple Pronunciation Generation at Inference Phase

In the acoustic G2P conversion approach, the pronunciations are inferred from the learned G2P relationships through acoustic data in the ergodic HMM framework. As the relationship between graphemes and phonemes is not always one-to-one, other alternative pronunciations can also be relevant to the data. In this paper, we propose to generate such alternative pronunciations through N-best Viterbi decoding which finds the N-best paths in the HMM. As the N-best pronunciations are generated from the learned G2P relationships using acoustic information, they can be expected to be phonologically relevant to the data.

Our hypothesis in this paper is that through generating multiple pronunciations in the acoustic G2P conversion approach it is possible to improve the performance at both pronunciation level and application level, in this case ASR. In the reminder of the paper, we validate our hypothesis by evaluating the two methods.

4. EXPERIMENTAL SETUP

In this paper, we consider a scenario in which limited transcribed speech data together with the pronunciation lexicon for the words in the training data is available, and the goal is to infer pronunciations for words which are not seen in the training data. For this purpose we have chosen the PhoneBook English corpus. This section describes the PhoneBook corpus together with the pronunciation generation and evaluation setups.

4.1. Database

PhoneBook is a speaker-independent task-independent isolated word recognition corpus [19] for small size and medium size vocabularies. In this paper, we use the medium size vocabulary task with 600 unique words [20]. The overview of the PhoneBook corpus in terms of number of utterances, hours of speech data, speakers and words present in train, cross-validation and test set is provided in Table 1.

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

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

The training set consists of 26,711 utterances (obtained by merging the small training set and cross-validation set as in [20]), and test set consists of 6598 speech utterances. PhoneBook pronunciation lexicon is transcribed using 42 phonemes (including silence). The test vocabulary consists of words and speakers which are unseen during training.

Our motivation behind using the Phonebook setup to study the proposed multiple pronunciation generation methods is as follows:

1. The G2P relationship in English is highly irregular.
2. The training, cross validation and test vocabulary sets are different.
3. The corpus contains uncommon English words and proper names (e.g. Witherington, Gargantuan, Laguardia, crawfordsville).
4. It can be seen as a resource-limited scenario as the amount of training data (in terms of number of training words and amount of speech data) is small.

4.2. Pronunciation Generation: Learning Phase

In the first step, the acoustic unit posterior probabilities $\mathbf{z}_t$ were estimated through an ANN. In a previous study [10], we trained different ANNs more specifically, multilayer perceptrons (MLPs) on the training data of PhoneBook corpus with different number of clustered context-dependent (CD) phonemes as MLP outputs. In this paper, we used the five-layer MLP modeling 321 clustered CD phonemes as output units as it was shown to lead to the highest frame accuracy on the cross-validation set. In this MLP, 39-dimensional PLP cepstral features with four preceding and four following frame context were used as the input. The MLP was trained with output non-linearity of softmax and minimum cross-entropy error criterion, using the Quicknet software [21].

In the second step, we trained KL-HMM systems modeling tri-graphemes using cost functions based on different local scores, i.e., $\mathbf{S_{KL}}$, $\mathbf{S_{RKL}}$ and $\mathbf{S_{SKL}}$. We refer to them as *kl-HMM*, *rkl-HMM* and *skl-HMM* respectively. Each grapheme unit was modeled with three HMM states. For tying KL-HMM states we applied KL-divergence based decision tree state tying method proposed in [22].

4.3. Pronunciation Generation: Inference Phase

As we are interested in inferring context-independent phoneme sequence, for each lexical unit l^i in the KL-HMM, we marginalized the parameters $\{y_i^d = P(a^d|l^i)\}_{d=1}^D$, i.e., the posterior probabilities of the acoustic units $P(a^d|l^i)$ belonging to the same central phoneme were summed.

Then given the KL-HMM parameters, we inferred pronunciations using the ergodic HMM in which each phoneme was modeled with three left-to-right HMM states. Silence was removed in the ergodic HMM as it could lead to deletion of some phonemes when generating pronunciations.

We refer to the lexicons obtained using the parameters of the *kl-HMM*, *rkl-HMM* and *skl-HMM* systems as *Lex-kl*, *Lex-rkl* and *Lex-skl* respectively.

4.4. Multiple Pronunciation Generation at Learning Phase

We generated different lexicons by combining the pronunciations obtained using different cost functions (i.e., *Lex-kl*, *Lex-rkl* and *Lex-skl*). Table 2 shows the properties of the generated lexicons and the average number of pronunciations per word in each lexicon. As expected, *Lex-kl-rkl-skl* has the largest number of pronunciations per word.

Lexicon	Description	average pron. per word
Lex-kl-rkl	*Lex-kl*+ *Lex-rkl*	1.54
Lex-kl-skl	*Lex-kl*+ *Lex-skl*	1.37
Lex-rkl-skl	*Lex-rkl*+ *Lex-skl*	1.4
Lex-kl-rkl-skl	*Lex-kl*+ *Lex-rkl*+ *Lex-skl*	1.77

Table 2. Description of the lexicons generated through combination of pronunciations obtained using different cost functions.

4.5. Multiple Pronunciation Generation at Inference Phase

We used N-best Viterbi algorithm to generate multiple pronunciations at the inference stage. In this paper, $N \in \{1, 2, 3\}$.

We investigated models based on the local scores $\mathbf{S_{KL}}$, $\mathbf{S_{RKL}}$ and $\mathbf{S_{SKL}}$. The trends were similar for all the local scores. However, due to space limitations, we only report the results using the local score $S_{\mathbf{SKL}}$ as they lead to the best performance at both pronunciation and ASR levels. Table 3 shows the descriptions of the lexicons generated through N-best decoding together with the average number of pronunciations per word in each lexicon.

It can be observed that the multiple pronunciation generation method at the inference stage leads to a larger number of pronunciations per word on average compared to the method used in the learning stage.

Lexicon	average prons. per word
Lex-skl-1best	1
Lex-skl-2best	1.56
Lex-skl-3best	2.38

Table 3. Description of the lexicons generated through N-best decoding.

4.6. Evaluation

We evaluated the generated lexicons at both pronunciation and ASR levels. For the evaluation at the pronunciation level, we computed phoneme and word accuracy on the test lexicons.

For the evaluation at the ASR level, we trained KL-HMM systems modeling context-dependent phonemes using phoneme posterior probabilities $\mathbf{z}_t$ obtained from a five-layer MLP classifying mono-phonemes as feature observations[1]. The KL-HMM parameters were trained by minimizing the cost function based on $S_{\mathbf{SKL}}$ as the local score. Throughout our studies we considered two scenarios:

1. *Case-1*: In this case, the KL-HMM system is trained using the manual dictionary and the generated G2P-based lexicons are used only at the decoding stage.

[1] In this paper, we used the MLP classifying mono-phonemes to be consistent with the previous studies [10]

2. *Case-2*: In previous studies, it has been found that discrepancies between the manual pronunciations used in the training lexicon and G2P-based pronunciations in the test lexicon can lead to inferior systems [10, 23]. Therefore we also studied the case where the G2P-based lexicons are used at both training and decoding stage.

5. EXPERIMENTAL ANALYSIS AND RESULTS

In the this section, we first analyze the multiple pronunciations generated and then present the evaluation results at both pronunciation and ASR levels.

5.1. Analysis

In Section 3.1, it was explained that the cost functions based on the local scores $S_{\mathbf{KL}}$ and $S_{\mathbf{RKL}}$ can better model one-to-one and one-to-many G2P relationships respectively, whereas the cost function using the local score $S_{\mathbf{SKL}}$ can model both one-to-one and one-to-many G2P relationships. In order to analyze the effect of the cost function and the local score, the G2P relationship captured by the parameters of different grapheme-based KL-HMM systems is presented in Table 4. More precisely, we exploit the intermediate context-independent KL-HMM systems developed during training the grapheme-based KL-HMM systems to obtain the grapheme-to-phoneme mapping as follows: for each state in the context-independent grapheme-based KL-HMM, we have chosen the phonemes with the probability higher than a threshold t ($t = 0.05$) according to the categorical distribution.

Graph.	$S_{\mathbf{KL}}$	$S_{\mathbf{RKL}}$	$S_{\mathbf{SKL}}$
B	b (0.9)	b (0.6)	b (0.8)
C	k (0.7), t (0.1)	k (0.3), s (0.3), C (0.1)	k (0.6), t (0.1)
F	f (0.8)	f (0.5), s (0.3)	f (0.7), s (0.2)
G	g (0.7), d (0.1)	g (0.2), jh (0.2), d (0.1)	g (0.4), d (0.2), k (0.1)
I	ay (0.5), aa (0.1), ae (0.1)	ay (0.2), aa (0.2), iy (0.1)	ay (0.5), aa (0.1)
N	n (0.8), m (0.1)	n (0.6), ng (0.2), m (0.1)	n (0.7), m (0.1)
U	uw (0.4), aa (0.2)	uw (0.2), y (0.1), aa (0.1)	uw (0.4), aa (0.2)
Y	iy (0.9)	iy (0.6), ay (0.1)	iy (0.8)

Table 4. G2P relationship captured by the parameters of the grapheme-based KL-HMM systems using $S_{\mathbf{KL}}$, $S_{\mathbf{RKL}}$ and $S_{\mathbf{SKL}}$ as the local scores.

It can be observed from Table 4 that the parameters of *kl-HMM* system can better capture one-to-one G2P relationships (e.g., see [B], [F]) whereas in *rkl-HMM* system, one-to-many G2P relationships are better captured (e.g., see [C], [N]). The parameters of the *skl-HMM* system tend to capture one-to-one G2P relationship similar to the case using the local score $S_{\mathbf{KL}}$. They can also capture one-to-many G2P relationships, but not to the same extent as local score $S_{\mathbf{RKL}}$ (e.g., see [G]).

As we will see later, the ability of the local score $S_{\mathbf{SKL}}$ to capture one-to-many G2P relationships can be beneficial during N-best decoding as it could lead to generating phonologically relevant pronunciations.

Table 5 provides example words together with their correct pronunciations from the manual dictionary and the generated pronunciations by using different cost functions or through N-best decoding. It can be observed that using each cost function has its own drawbacks and advantages. With the local score $S_{\mathbf{RKL}}$, the one-to-many relationships can be better captured compared to the local score $S_{\mathbf{KL}}$ (e.g., *diaries*). However, additional spurious relations can be possibly captured through use of the local score $S_{\mathbf{RKL}}$. This can be particularly seen in the word *shingler* where the spurious phoneme /n/ is inserted . In addition, the one-to many G2P relationships may lead to confusion between the phones. This can be seen in the word *chefs* where the correct phoneme is confused with the phoneme /x/ . Such confusions could be possibly better handled through the local scores $S_{\mathbf{KL}}$ (e.g., *chefs*) and $S_{\mathbf{SKL}}$ (e.g., *paxton*).

Through use of N-best pronunciations, phonological variations which can be more relevant to the data can be captured. This can be seen for the word *fuddle* in which the grapheme [U] is mapped to the sound /uw/ in the single-best pronunciation (as it had the highest probability according to Table 4) and in the second best pronunciation it is mapped to the sound /aa/ which is closer to the sound /ah/ provided in the manual dictionary.

5.2. Pronunciation Level Results

Figure 2 provides pronunciation level evaluation results in terms of phoneme accuracy (PA) and word accuracy (WA) when generating multiple pronunciations at the learning stage. The pronunciation level results when using *Lex-skl-1best* serves as the baseline in this section. It can be seen that through combining pronunciations obtained using different local scores, the performance of the acoustic G2P conversion approach at pronunciation level improves. The *Lex-kl-rkl-skl* leads to the best accuracy at the pronunciation level.

Figure 3 shows the pronunciation level performance when generating multiple pronunciations at the inference stage. It can be observed that through applying N-best decoding, improvements in terms of both phoneme and word accuracy over the baseline are achieved. The best pronunciation level accuracy is achieved when applying 3-best decoding.

It can be observed that the *Lex-skl-3best* lexicon obtained through N-best decoding performs better at the pronunciation level compared to the *Lex-kl-rkl-skl* lexicon obtained using different cost functions. This could be expected as the Lex-skl-3best on average has more pronunciations per word compared to the Lex-kl-rkl-skl lexicon (see Tables 2 and 3).

Word	Correct pron.	Generated Pronunciation	
		Learning stage	Inference stage
shingler	sh ih ng l axr	sh aa ng aa l axr ($S_{\mathbf{KL}}$) sh aa ng *n* l axr ($S_{\mathbf{RKL}}$) sh aa ng aa l axr ($S_{\mathbf{SKL}}$)	sh aa ng *aa* l axr (1-best) sh aa ng n l axr (2-best) -
chefs	sh eh f s	ch **eh** f s ($S_{\mathbf{KL}}$) ch *aa* f s ($S_{\mathbf{RKL}}$) ch *aa* f s ($S_{\mathbf{SKL}}$)	ch *aa* f s (1-best) ch *ih* f s (2-best) ch **eh** f s (3-best)
diaries	d ay axr iy z	d iy r iy z ($S_{\mathbf{KL}}$) d **ay** iy r iy z ($S_{\mathbf{RKL}}$) d **ay** iy r iy z ($S_{\mathbf{SKL}}$)	d ay iy r iy z (1-best) d ay iy r iy s (2-best) d ay ey r iy z (3-best)
fuddle	f ah d aa l	f *uw* d aa l ($S_{\mathbf{KL}}$) f **aa** d aa l ($S_{\mathbf{RKL}}$) f *uw* d aa l ($S_{\mathbf{SKL}}$)	f *uw* d aa l (1-best) f **aa** d aa l (2-best) f *ih* d aa l (3-best)
paxton	p ae k s t aa n	p *ey* k t aa n ($S_{\mathbf{KL}}$) p *ey* k t ow n ($S_{\mathbf{RKL}}$) p **ae** k t ow n ($S_{\mathbf{SKL}}$)	p ae k t *ow* n (1-best) p ae k t **aa** n (2-best) p ae t *ow* n (3-best)

Table 5. Example words together with their pronunciations from the manual dictionary as well as the generated pronunciations (at the learning stage and inference stage). We have highlighted the correctly generated phonemes with the bold font type in blue, and the spurious or confused phonemes with the italic font type in red.

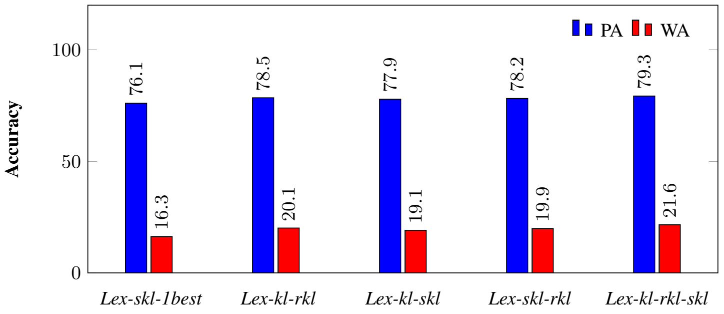

Fig. 2. Pronunciation level evaluations in terms of phoneme accuracy (PA) and word accuracy (WA) when using multiple pronunciations obtained using different cost functions.

5.3. ASR Level Results

Table 6 shows the performance of the KL-HMM systems in terms of word accuracy when exploiting the multiple pronunciation generated at learning stage. The KL-HMM system using *Lex-skl-1best* serves as our baseline systems for evaluating the pronunciation generation methods. It is worth mentioning that the ASR word accuracy when using the manual dictionary is about 98.2%. The acoustic G2P and standard G2P conversion approaches achieve a performance well below that [10]. This shows that G2P conversion in this corpus is indeed a difficult task.

It can be seen that through use of pronunciations derived using different learned G2P relationships, significant improvements over the baseline system are achieved in both *Case-1* and *Case-2* (with at least 99% confidence [24]). The KL-HMM system using *Lex-kl-rkl* performs comparable to the system using *Lex-kl-rkl-skl*. This could indicate that adding the pronunciations obtained from the KL-HMM system trained with $S_{\mathbf{SKL}}$ may not add further information.

Lexicon	*Lex-skl-1best*	*Lex-kl-rkl*	*Lex-kl-skl*	*Lex-skl-rkl*	*Lex-kl-rkl-skl*
WA-*Case-1*	85.0	86.9	86.6	86.2	**87.1**
WA-*Case-2*	88.2	89.9	89.3	89.6	**90.1**

Table 6. Performance of KL-HMM systems in terms of word accuracy when using multiple pronunciations obtained through different cost functions at KL-HMM training.

Table 7 presents the performance of the KL-HMM systems in terms of word accuracy using the N-best pronunciations. It can be observed that the KL-HMM systems using N-best pronunciations ($N > 2$), perform significantly better than the baseline system (with at least 99%

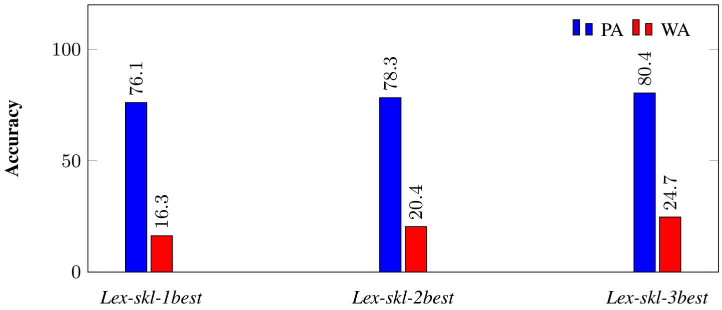

Fig. 3. Pronunciation level evaluations in terms of phoneme accuracy (PA) and word accuracy (WA) when applying N-best decoding.

confidence).

The improvements in accuracy in *Case-1* are more considerable compared to *Case-2* in both Tables 6 and 7. This can be attributed to reducing the inconsistencies between the train and test lexicons by providing alternative pronunciations at the test lexicon which are more consistent with the pronunciations at the train lexicon.

Lexicon	*Lex-skl-1best*	*Lex-skl-2best*	*Lex-skl-3best*
WA-*Case-1*	85.0	85.6	**87.0**
WA-*Case-2*	88.2	88.3	**89.2**

Table 7. Performance of KL-HMM systems in terms of word accuracy when using N-best pronunciations.

The best performing KL-HMM system in Table 6 performs slightly better than the best system in Table 7. This is in spite of the lower performance of the *Lex-kl-rkl-skl* at the pronunciation level and less number of pronunciations per word, which can indicate that the pronunciation level performance is not necessarily indicative of the performance at the ASR level.

6. DISCUSSION, CONCLUSIONS AND FUTURE WORK

In this paper, we investigated the potential of the acoustic G2P conversion approach in generating multiple pronunciations. We proposed and assessed two approaches to achieve this goal. The first approach generates multiple pronunciations through use of different cost functions at learning stage, and the second one generates pronunciations at the inference stage through N-best decoding. Our experimental studies on a resource-limited scenario on the PhoneBook corpus show that the ASR systems using multiple pronunciations perform significantly better than the systems using single-best pronunciations.

It is interesting to see that we can obtain performance improvements without using any pronunciation selection or pruning methods. However, as observed, the *Lex-skl-3best* lexicon with more average number of pronunciation per word and higher pronunciation accuracy yields an inferior ASR system when compared to *Lex-kl-rkl-skl*. Therefore, there may be benefits in pruning pronunciations. We aim to investigate this in our future studies.

The studies in this paper were conducted on a resource-constrained task in English. More recently, it was shown that the acoustic G2P conversion approach can be effectively employed in development of automatic suwbord unit based lexicons [25] and address the lexical resource constraint issues of real world under-resourced languages [26]. Our future work will investigate the potential of the proposed methods for multiple pronunciation generation in improving automatically derived subword unit based pronunciation lexicon.

In addition to under-resourced scenarios, the approach could be beneficial for tasks such as name recognition in which the pronunciation of a word can be affected by both the origin language of the word and the native language of the speaker [27]. In this case, it would be interesting to study multiple pronunciation generation by learning the relationship between graphemes and multilingual phones using acoustic resources of multiple languages. Our future work will also investigate this direction.

7. REFERENCES

[1] R.M. Kaplan and M. Kay, "Regular Models of Phonological Rule Systems," *Computational Linguistics*, vol. 20, pp. 331–378, 1994.

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

[3] D. Wang and S. King, "Letter-to-Sound Pronunciation Prediction Using Conditional Random Fields," *Signal Processing Letters, IEEE*, vol. 18, no. 2, pp. 122–125, 2011.

[4] M. Bisani and H. Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, May 2008.

[5] T. J. Sejnowski and C. R. Rosenberg, "Parallel Networks that Learn to Pronounce English Text," *Complex Systems*, vol. 1, pp. 145–168, 1987.

[6] A. Laurent, P. Delglise, and S. Meignier, "Grapheme to Phoneme Conversion Using an SMT System.," in *Proceedings of Interspeech*, 2009, pp. 708–711.

[7] I. McGraw, I. Badr, and J.R. Glass, "Learning Lexicons From Speech Using a Pronunciation Mixture Model," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 357–366, 2013.

[8] L. Lu, A. Ghoshal, and S. Renals, "Acoustic Data-Driven Pronunciation Lexicon For Large Vocabulary Speech Recognition," in *Proceedings of ASRU*, 2013, pp. 374–379.

[9] R. Rasipuram and M. Magimai-Doss, "Acoustic Data-driven Grapheme-to-Phoneme Conversion using KL-HMM," in *Proceedings of ICASSP*, Mar. 2012.

[10] M. Razavi, R. Rasipuram, and M. Magimai.-Doss, "Acoustic Data-Driven Grapheme-to-Phoneme Conversion in the Probabilistic Lexical Modeling Framework," Idiap-RR Idiap-RR-10-2015, Idiap, 5 2015.

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

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

[13] M. Magimai.-Doss, R. Rasipuram, G. Aradilla, and H. Bourlard, "Grapheme-based Automatic Speech Recognition using KL-HMM," in *Proceedings of Interspeech*, 2011, pp. 445–448.

[14] R. Veldhuis, "The Centroid of the Symmetrical Kullback-Leibler Distance," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 96–99, 2002.

[15] K. Livescu, E. Fosler-Lussier, and F. Metze, "Subword Modeling for Automatic Speech Recognition: Past, Present, and Emerging Approaches.," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 44–57, 2012.

[16] C. Genest and J. V. Zidek, "Combining Probability Distributions: A Critique and an Annotated Bibliography," *Statist. Sci.*, vol. 1, no. 1, pp. 114–135, 02 1986.

[17] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On Combining Classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[18] R. Rasipuram and M. Magimai.-Doss, "Probabilistic Lexical Modeling and Grapheme-based Automatic Speech Recognition," Idiap-RR Idiap-RR-15-2013, Idiap, 4 2013, Submitted to Speech Communication.

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

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

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

[22] D. Imseng, J. Dines, P. Motlicek, P. N. Garner, and H. Bourlard, "Comparing Different Acoustic Modeling Techniques for Multilingual Boosting," in *Proceedings of Interspeech*, Sept. 2012.

[23] D. Jouvet, D. Fohr, and I. Illina, "Evaluating Grapheme-to-Phoneme Converters in Automatic Speech Recognition Context," in *Proceedings of ICASSP*, 2012, pp. 4821–4824.

[24] M. Bisani and H. Ney, "Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation," in *Proceedings of ICASSP*, May 2004, vol. 1, pp. 409–412.

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

[26] M. Razavi, R. Rasipuram, and M. Magimai.-Doss, "On the Application of Automatic Subword Unit Derivation and Pronunciation Generation for Under-Resourced Language ASR: A Study on Scottish Gaelic," Idiap-RR Idiap-RR-13-2015, Idiap, 6 2015.

[27] B. Maison, S. E Chen, and P. S Cohen, "Pronunciation Modeling for Names of Foreign Origin," in *Proceedings of ASRU*. IEEE, 2003, pp. 429–434.