On Learning Grapheme-to-Phoneme Relationships through the Acoustic Speech Signal

Mathew Magimai.-Doss[1], Ramya Rasipuram[1,2]
[1]**Idiap Research Institute, Martigny, Switzerland,**
[2]**Ecole Polytechnique Fèdérale de Lausanne (EPFL), Lausanne, Switzerland**
e-mail: mathew@idiap.ch, ramya.rasipuram@idiap.ch

Abstract

Automatic speech recognition (ASR) systems, through use of the phoneme as an intermediary unit representation, split the problem of modeling the relationship between the written form, i.e., the text and the acoustic speech signal into two disjoint processes. The first process deals with modeling of the relationship between the written form and phonemes through development of a pronunciation dictionary using prior knowledge about grapheme-to-phoneme relationships. Given the pronunciation lexicon and the transcribed speech data, the second process then deals with modeling of the relationship between the phonemes and the acoustic speech signal using statistical sequence processing techniques, such as hidden Markov models. As a consequence of the two disjoint processes, development of an ASR system heavily relies on the availability of well-developed acoustic and lexical resources in the target language. This paper presents an approach where the relationship between graphemes and phonemes is learned through acoustic data, more precisely, through phoneme posterior probabilities estimated from the speech signal. In doing so, the approach tightly couples the above mentioned two processes and leads to a framework where, existing acoustic and lexical resources from different domains and languages can be effectively exploited to build ASR systems without development of a pronunciation lexicon and to develop lexical resources for resource scarce domains and languages. We demonstrate these capabilities of the proposed approach through cross domain studies in English, where the grapheme-to-phoneme relationship is deep.

Keywords: Automatic speech recognition, hidden Markov models, phonemes, graphemes, grapheme-to-phoneme conversion

1 Introduction

Speech technologies, such as automatic speech recognition (ASR) systems (Gold and Morgan, 1999), text-to-speech synthesis (TTS) systems (Taylor, 2009) interface or connect two different modes of human communication, namely, the spoken form (the acoustic speech signal) and the written form (the textual message). As a consequence, these systems need to model the relationship between the acoustic speech signal and units of written form, such as graphemes. However, modeling the relationship between the acoustic speech signal and graphemes directly is not trivial. The primary reason is that the realized acoustic speech signal is more related to the units of spoken form, i.e. phonemes, and the grapheme-to-phoneme relationships,

which depends upon whether the phoneme-grapheme relationships within the language are shallow or deep. For instance, languages, such as Spanish and Finnish, have shallow grapheme-to-phoneme relationship, while languages such as English and German have deep grapheme-to-phoneme relationship. In addition, languages tend to evolve over time, as a result the grapheme-to-phoneme relationship could undergo changes.

During the development of an ASR system, the problem of modeling the relationship between graphemes and acoustic signals is typically broken down into two parts through use of phonemes as the intermediary representation. In the first part, the relationship between the words (in written form) and the phonemes or phones in the language is modeled through a pronunciation lexicon (Gold and Morgan, 1999; Schultz and Kirchhof, 2006). In the second part, the relationship between phonemes and acoustic speech signals is usually modeled within the framework of hidden Markov models (HMM) using either Gaussian mixture models (GMM) (Rabiner, 1989) or artificial neural networks (ANN) (Morgan and Bourlard, 1995). A consequence of splitting the problem in two disjoint parts is that development of an ASR system heavily depends on prior acoustic and linguistic resources from the target language. For instance, development of a pronunciation lexicon requires knowledge of the grapheme-to-phoneme rules in a language, which are primarily derived from linguistic studies. Similarly, in the second part, a large amount of transcribed speech in the target language is needed, in addition to the availability of a pronunciation lexicon, in order to train better models.

This paper presents an approach that was originally developed at Idiap Research Institute to automatically learn grapheme-to-phoneme relationships through acoustic speech signals given prior resources such as, transcription of the speech signal and a seed lexicon. In this approach, the relationship between acoustic speech signal and phonemes is first modeled by an ANN. And, then a hidden Markov model (HMM) whose states represent graphemes and the state parameters characterize a probabilistic grapheme-to-phoneme relationship is trained. The parameters of the HMM are learned by using posterior probabilities of phonemes estimated by the ANN as feature observations (Magimai.-Doss et al., 2011; Rasipuram and Magimai.-Doss, 2015; Rasipuram, 2014). In this paper, we present a part of our research on English to demonstrate the viability of the approach and its ability to address lexical resource scarcity issues.

The remainder of the paper is organized as follows. Section 2 provides a brief background on the development of the pronunciation lexicon and HMM-based ASR systems. Section 3 presents the proposed approach. Sections 4-7 present the experimental studies. Finally, Section 8 summarizes and presents directions for future work.

2 Background

In this section, we provide a brief overview of the pronunciation lexicon development and the standard HMM-based ASR system.

2.1 Development of Pronunciation Lexicon

Pronunciation lexicon development can be seen as a process of converting a grapheme sequence $G = \{g_1, \cdots g_L\}$ obtained from the orthography of the word into a phoneme sequence $F = \{f_1, \cdots f_M\}$. Usually, the starting point for pronunciation lexicon development is grapheme-to-phoneme conversion rules derived from the linguistic studies of the language. Given these rules, two approaches can be adopted:

1. Human experts can be employed to predict the phoneme sequence. In this case, a human expert enters the sequence of phonemes in the orthography of the target word.
2. Employ computational phonological methods. For instance, formulation of the rules in terms of finite state automata and prediction of a phoneme sequence (Kaplan and Kay, 1994). This would still need supervision by humans, i.e., hand-correction.

These approaches can be employed if the vocabulary size is small. However, in the case of a large vocabulary, these approaches can be time consuming and tedious. Therefore in practice, a seed pronunciation lexicon consisting of a few words is developed first using human expertise. Automatic grapheme-to-phoneme conversion (G2P) techniques are then employed to learn the grapheme-to-phoneme relationships from the seed lexicon and to populate the pronunciation lexicon with new words. The challenge of how well the grapheme-to-phoneme relationship can be modeled automatically depends upon the language. As we will see shortly, the G2P techniques typically rely on modeling the contextual information in the grapheme sequence and, in some cases, the contextual information in the phoneme sequence as well. The underlying assumption here being that the relationship between context-dependent graphemes and phonemes is shallow. A number of machine learning techniques have been proposed for automatic G2P. These approaches can be broadly classified as,

1. Local classification-based approaches: In these approaches, the grapheme sequences (the orthography of a word) and the corresponding sequences in the seed pronunciation lexicon are first aligned. And, then a decision tree (Pagel et al., 1998) or an ANN (Sejnowski and Rosenberg, 1987) is trained to predict the corresponding phoneme for each grapheme in the orthography of the word given the context information (preceding and following graphemes). The pronunciations for new words are obtained by locally predicting the phonemes using the trained decision trees or ANN, and concatenating them.
2. Sequence classification-based approaches: The problem of G2P can be formulated as a sequence classification problem,

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

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

where, P (F |G) denotes the probability of the phoneme sequence F given the grapheme sequence G, P (F, G) denotes the joint probability of the phoneme sequence F and the grapheme sequence G, and F* is the inferred phoneme sequence.

The conditional random fields (CRF) based G2P technique (Wang and King, 2011) is based on Eqn. (1), while approaches such as, joint multigram or joint n-gram based technique (Bisani and Ney, 2008) and HMM-based technique (Taylor, 2005), are based on Eqn. (2). In addition to these approaches, there are other data-driven approaches, such as inductive learning of grapheme-to-phoneme rules (van Coile, 1990), Pronunciation by Analogy (Dedina and Nusbaum, 1991), Default&Refine (Davel and Barnard, 2008).

Currently, the joint n-gram approach is the state-of-the-art G2P technique. In this approach, the grapheme sequence and the phoneme sequence information are jointly modeled by units referred to as ”graphones”, which are created by pairing graphemes and phonemes after alignment and training of an n-gram model of the graphones using the seed lexicon (Bisani and Ney, 2008). Given G for an unseen word, F* is then inferred by Viterbi decoding (Forney, 1973).

2.2 HMM-based ASR

In HMM-based ASR system (Rabiner, 1989; Morgan and Bourlard, 1995; Gold and Morgan, 1999; Schultz and Kirchhof, 2006), the goal is to find the best matching word hypothesis *W** given the acoustic feature observation sequence $X = \{x_1, \cdots x_t, \cdots x_T\}$, where x_t is the acoustic feature observation, typically a parametric representation of short-term spectrum, at time frame *t* and *T* is the number of frames. Formally, it can be expressed as,

$$W^* = \arg\max_{W \in \mathcal{W}} P(W|X), \tag{3}$$

$$= \arg\max_{W \in \mathcal{W}} \frac{p(X|W) \cdot P(W)}{p(X)} \tag{4}$$

$$= \arg\max_{W \in \mathcal{W}} p(X|W) \cdot P(W) \tag{5}$$

where *W* denotes a word sequence hypothesis, $\mathcal{W}$ denotes the set of hypotheses, *P(W|X)* denotes the probability of the word sequence *W* given the acoustic feature sequence *X*, *p(X|W)* denotes the likelihood of the acoustic feature observation sequence *X* given the word sequence *W*, *P(W)* denotes the prior probability of the word sequence *W* and *p(X)* denotes the likelihood of the acoustic feature observation sequence *X*. Eqn. (5) results from the assumption that *p(X)* is independent of the word hypothesis *W*.

Typically, HMM-based ASR systems use phonemes as subword units. During training, the relation *p(X|W)* is modeled using the transcribed speech data and a well-developed pronunciation lexicon, while *P(W)* is modeled using the textual resources.

During recognition, given the pronunciation lexicon and the parameters of $p(X|W)$ and $P(W)$ estimators, W^* for a test utterance is obtained using a Viterbi decoder.

3 Learning Grapheme-to-Phoneme Relationship through Acoustics

In this section, we present a novel approach that was developed at Idiap under the FlexASR project[1] for learning grapheme-to-phoneme relationships through the acoustic speech signal and its subsequent use for automatic speech recognition. This is achieved through the recently proposed Kullback-Leibler divergence based HMM (KL-HMM) approach (Aradilla, 2008).

3.1 Kullback-Leibler divergence based HMM

Kullback-Leibler divergence based HMM is a new ASR approach where the posterior probability estimates of phonemes $z_t = [P(c_1|x_t), \cdots P(c_d|x_t), \cdots P(c_D|x_t)]^T$ are used as feature observations (Aradilla et al., 2008; Aradilla, 2008). Here, $\{c_1, \cdots, c_D\}$ denotes the set of D phoneme classes and x_t denotes the acoustic feature vector at time frame t. The phoneme posterior probabilities can be estimated by training an ANN (Morgan and Bourlard, 1995; Aradilla et al., 2008) or Gaussian mixture models (GMM) (Rabiner, 1989; Rasipuram and Magimai.-Doss, 2013). For the sake of clarity, we hereafter refer to z_t as the posterior feature. In the KL-HMM approach each HMM state i is parametrized by a categorical distribution $y_i = [y_1, \cdots y_k, \cdots, y_K]^T$, which is trained by minimizing a cost function based on Kullback-Leibler divergence between the state categorical distribution y_i and posterior feature observations (see Figure 1). More precisely, unlike the HMM/GMM system (Rabiner, 1989) where the local score is likelihood or the HMM/ANN system (Morgan and Bourlard, 1995) where the local score is scaled-likelihood, the local score $S(y_i, z_t)$ at each HMM state i in the case of the KL-HMM system is the Kullback-Leibler divergence between y_i and z_t, i.e.,

$$KL = S(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} y_i^d \log(\frac{y_i^d}{z_t^d}) \tag{6}$$

The above equation represents the case where $\boldsymbol{y_i}$ is the reference distribution and the local score is denoted as KL. However, given that KL-divergence is an asymmetric measure, there are two other possible ways to estimate KL-divergence, namely, the reverse KL-divergence (RKL, where the posterior feature z_t is the reference distribution) or the symmetric KL-divergence(SKL), as follows:

$$RKL \quad = S(\mathbf{y}_i, \mathbf{z}_t) \quad = \sum_{d=1}^{D} z_t^d \log(\frac{z_t^d}{y_i^d}) \tag{7}$$

[1] https://www.idiap.ch/scientific-research/projects/flexible-grapheme-based-automatic-speech-recognition

$$SKL \quad = S(\mathbf{y}_i, \mathbf{z}_t) \quad = \frac{1}{2}[KL + RKL]. \tag{8}$$

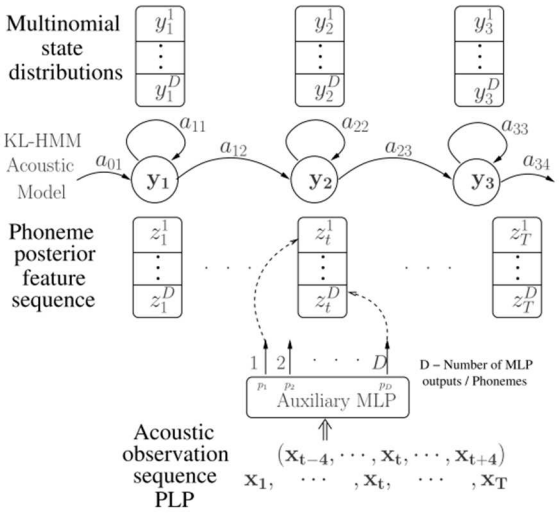

Figure 1: Illustration of KL-HMM approach with ANN as posterior feature estimator.

3.1.1 Training

The KL-HMM system is fully parameterized by $\Theta=\{\{y_i\}^I_{i=1}, \{a_{ij}\}^I_{i,j=1}\}$, where I is the total number of states, each state i is represented by a categorical distribution y_i and a_{ij} is the transition probability from state i to state j. Given a training set of N utterances, where each training utterance n is a sequence of phoneme posterior features $Z(n) = \{z_1(n), \cdots, z_{T(n)}(n)\}$, and $T(n)$ is the length of the training utterance n, the parameters Θ are estimated by the embedded Viterbi expectation maximization algorithm which minimizes the cost function,

$$\min_{\mathcal{Q}(n)} \sum_{n=1}^{N} \sum_{t=1}^{T(n)} [S(\mathbf{y}_{q_t}, \mathbf{z}_t(n)) - \log a_{q_{t-1}q_t}] \tag{9}$$

over all parameters Θ, where, $Q(n)$ denotes the set of possible state sequences allowed by utterance n and $qt \in \{1, \cdots I\}$. For more details about the training and update equations for each of the local scores, the reader is referred to Aradilla (2008).

3.1.2 Decoding

The decoding is performed using the standard Viterbi decoder. Given a sequence of phoneme posterior features $Z = \{z_1, \cdots, z_T\}$ and the trained parameters Θ, decoding involves recognition of the underlying hypothesis $\hat{m}$

$$\hat{m} = \arg\min_{\mathcal{Q}(m)} \sum_{t=1}^{T} [S(\mathbf{y}_{q_t}, \mathbf{z}_t) - \log a_{q_{t-1}q_t}] \tag{10}$$

where Q(m) denotes the set of possible state sequences allowed by the hypothesis m and $q_t \in \{1, \cdots I\}$. For further understanding about thesimilarities and dissimilarities between the KL-HMM approach and the standard HMM-based ASR approach, and the effect of different local scores on parameter estimation and decoding, the reader is referred to (Rasipuram, 2014; Rasipuram and Magimai.-Doss, 2015).

3.2 Proposed Approach

More recently, a grapheme-based ASR approach was proposed in the framework of KL-HMM (Magimai.-Doss et al., 2011; Rasipuram, 2014; Rasipuram and Magimai.-Doss, 2015) where,

- the relationship between acoustic features and phonemes is first modeled through a posterior feature estimator, e.g., ANN or GMM,
- then, a KL-HMM whose states represent graphemes is trained by using the phoneme posterior probabilities as feature observations. In doing so, the parameters of the KL-HMM tend to capture a probabilistic grapheme-to-phoneme relationship (see Section 5).

This approach has been found to yield significantly better performance than the standard HMM-based ASR approach, where the relationship between the acoustic feature and the graphemes is modeled directly (Kanthak and Ney, 2002; Killer et al., 2013). The remainder of the paper presents a part of our research that shows how the proposed approach can address lexical resource scarcity issues by using the KL-HMM as a recognition model and as a generative model.

4 Experimental Setup

This section presents the experimental setup for a case study on English to demonstrate the potential of the proposed approach. Our main reason for choosing English is that it has deep grapheme-to-phoneme relationships. Thus, modeling the grapheme-to-phoneme relationship effectively is not trivial.

4.1 Databases

In-domain corpus: We used the DARPA Resource Management (RM) corpus (Price et al., 1988) as the in-domain or target-domain corpus. The DARPA RM corpus consists of read queries on the status of Naval resources (Price et al., 1988). The task is artificial in many respects, including speech type, range of vocabulary and grammatical constraints. The speaker-independent ASR task training set consists of 3,990 utterances spoken by 109 speakers corresponding to approximately 3.8 hours of speech. The test set is a combination of four subsets provided by DARPA, namely, feb89, oct89, feb91, and sep92. Each of the subsets contain 300 utterances spoken by 10 speakers. Thus, the test set in total has 1,200 utterances, amounting to 1.1 hours of speech. The lexicon consists of 991 words. The phoneme-based lexicon was obtained

from UNISYN[2] lexicon. There are 42 context-independent phones[3], including silence. The test set is completely covered by a word pair grammar included in the task specification.

Out-of-domain corpus: We used the Wall Street Journal1 (WSJ1 – Paul and Baker, 1995), a read speech corpus, as the out-of-domain corpus. It consists of approximately 66 hours of speech recorded from 200 speakers. There are 10,000 unique words. The lexicon was obtained from UNISYN lexicon. There are 45 context-independent phones, including silence.

4.2 Modeling of the relationship between the acoustic signal and phonemes

Acoustic features: The acoustic feature vector is comprised of 13 dimensional PLP cepstral coefficients, their first order temporal derivatives and second order temporal derivatives estimated using a window of 30 ms with a 10 ms frame shift. The features were estimated using the HTK toolkit (Young et al., 2006). We used ANNs to model the relationship between the acoustic features and the phonemes to estimate phoneme posterior probabilities z_t. More precisely, we used two different ANNs, namely,

1. *In-domain ANN*: We used a three layer ANN (i.e., ANN with single hidden layer) that was trained on the DARPA RM corpus to classify 45 context-independent phones. This ANN was originally used in the study reported in (Dines and Magimai.-Doss, 2008).
2. *Out-of-domain ANN:* We used a three layer ANN that was trained on the WSJ1 corpus to classify 45 context-independent phones. This ANN was first used in the study reported in (Aradilla et al. 2008).

The input to the ANNs were 39 dimensional cepstral features with four frames preceding context and four frames following context, i.e., $(4 + 1 + 4) \times 39$ dimensional input. The ANNs were trained by minimizing a cost function based on cross entropy using the Quicknet software[4].

4.3 Studies

We present three different studies to demonstrate the potential of the proposed approach:

1. The first study presented in Section 5 demonstrates the capability of the proposed approach to learn a probabilistic grapheme-to-phoneme relationship. More specifically, in this study both the ANN and the KL-HMM are trained on the in-domain data and, the parameters of the KL-HMM are analyzed to show how the probabilistic grapheme-to-phoneme relationships are captured.
2. The second study presented in Section 6 focuses on the recognition model aspect of the KL-HMM. More precisely, we show that the ASR systems for

[2] http://www.cstr.ed.ac.uk/projects/unisyn/

[3] Phonemes /el/, /em/ and /en/ were merged with /l/, /m/ and /n/, respectively.

[4] http://www1.icsi.berkeley.edu/Speech/qn.html

a domain that lacks well-developed phonetic lexical resources can be effectively developed by

(a) training the ANN that models the relationship between the acoustic signal and phonemes on the out-of-domain data and,

(b) capturing the grapheme-to-phoneme relationships on the in-domain data.

3. The third study presented in Section 7 focuses on the generative model aspect of the KL-HMM. More specifically, we show that the learned grapheme-to-phoneme relationships can be exploited to perform grapheme-to-phoneme conversion by using the KL-HMM as a generative model. In that respect, building upon the second study, in this study we show how the out-of-domain acoustic resources and lexical resources can be exploited to build lexical resources for new domains.

All the KL-HMM systems reported in this paper are based on the local score SKL, Eqn. (8).

5 Analysis

This section presents an analysis of the KL-HMM parameters to show that indeed grapheme-to-phoneme relationships can be captured by the proposed approach.

5.1 Context-independent grapheme modeling

We trained a KL-HMM, where the feature observation was 45-dimensional phone posterior probabilities estimated by the in-domain ANN described earlier in Section 4.2 and the states represented 29 context-independent graphemes, including silence, hyphen, and apostrophe. Each grapheme was modeled by a single state. The parameters 29×45 were trained using the cost function based on SKL. The 45-dimensional parameter for each of the grapheme states was sorted in descending order and and the dimensions with probability value greater than or equal to 0.1 were selected. Table 1 shows the captured grapheme-to-phoneme relationships. It can be seen that the proposed approach is able to capture the dominant grapheme-to-phoneme relationships. In English, it is well known that the context-independent grapheme-to-phoneme relationship is variable, especially for vowels. This aspect was frequently observed. The context-independent grapheme H relates to aspirant sound /hh/. It can be seen that in addition to /hh/, the model captures the relation to stop consonants /dh/, /th/, /d/ and /t/, and silence. This indicates that the approach was able to implicitly capture the context in which grapheme H can occur, e.g. /dh/ reflects D followed by H. It can be observed that the parameters also capture acoustically confusable relationships that are potentially resulting from the assimilation process. For instance, see the relationships captured for graphemes D, G, M, S to name a few.

Table 1: Dominant grapheme-to-phoneme relationships (sorted according to the maximum probability value and with a probability value greater than or equal to 0.1) learned by KL-HMM states. For the sake of display the probability values were rounded off.

Grapheme	Captured phoneme relationship
A	/ae/ (0.5), /eh/ (0.2), /ey/ (0.1), /ax/ (0.1)
B	/b/ (0.9)
C	/k/ (0.6), /t/ (0.2), /ch/ (0.1), /s/ (0.1)
D	/d/ (0.7), /t/ (0.1)
E	/iy/ (0.3), /ax/ (0.2), /ih/ (0.2), /eh/ (0.1), /ey/ (0.1)
F	/f/ (0.9)
G	/g/ (0.7), /d/ (0.1), /k/ (0.1)
H	/dh/ (0.2) sil(0.2), /t/ (0.2), /th/ (0.1), /d/ (0.1), /hh/ (0.1)
I	/ih/ (0.5), /ax/ (0.2), /eh/ (0.1), /ay/ (0.1)
J	/jh/ (0.9)
K	/k/ (0.9)
L	/l/ (0.8)
M	/m/ (0.9), /n/ (0.1)
N	/n/ (0.8), /en/ (0.1)
O	/ao/ (0.2), /aa/ (0.2), /ow/ (0.2), /ah/ (0.1), /ax/ (0.1)
P	/p/ (0.9)
Q	/k/ (0.9)
R	/r/ (0.6), /axr/ (0.3), /er/ (0.1)
S	/s/ (0.8), /z/ (0.2)
T	/t/ (0.8)
U	/uw/ (0.3), /ax/ (0.3), /ih/ (0.1), /ah/ (0.1)
V	/v/ (0.9)
W	/w/ (0.9)
X	/k/ (0.5), /s/ (0.3), /t/ (0.1)
Y	/iy/ (0.5), /ey/ (0.3), /ih/ (0.1)
Z	/z/ (0.8), /s/ (0.1)
sil	sil (1.0)

5.2 Effect of context-dependent grapheme modeling

The underlying idea of grapheme-to-phoneme conversion approaches, discussed in Section 2.1, is that the relationship between graphemes and phonemes can become shallow when contextual information is modeled. The proposed approach provides similar capabilities. To illustrate it, we present an investigation, where

1. Single state grapheme models with three different types of contextual information: mono (context-independent), tri (word internal single preceding and single following graphemes), and quint (word internal two preceding and two following graphemes) were trained. Table 2 illustrates the different context models for word AREA as an example.

2. Entropy of the categorical distribution estimated for each of the grapheme models is computed.
   In the case of tri and quint, it is an average of the entropies of the grapheme models that share central graphemes. For example in Table 2, models *b-A+R* and *E-A+e* share the central grapheme A. Entropy of the categorical distribution is a good indicator of the one-to-one or shallow relationship and the one-to-many or deep relationship. More precisely, for the one-to-one relationship, the entropy is low while the entropy for one-to-many relationship is high.

Figure 2 plots the entropy for the different graphemes. It can be observed that

- vowel graphemes (A, E, I, O, U) and a few consonant graphemes (C, H, R, X) have high entropy for context mono which indicates that the parameters capture one-to-many G2P relationships. As the context is increased to tri and quint, the entropy decreases, which indicates that the context-dependent grapheme models are capturing a shallower grapheme-to-phoneme relationship, as compared to mono.
- a few consonant graphemes like B, K, P, and V have low entropy for context mono, which indicates that the context-independent grapheme itself models a one-to-one grapheme-to-phoneme relationship. However, the entropy slightly increases as the context is increased to tri and quint. A closer inspection of the parameters revealed that this was due to the phoneme context information captured by the grapheme KL-HMM models.

Table 2: Context expansion for the word AREA, where b denotes beginning of word and *e* denotes end of the word. The symbols '+' and '*' refer to the first and second following contexts, and '-' and '~' refer to the first and second preceding contexts, respectively.

Model	**Context expansion for word AREA**			
mono	A	R	E	A
tri	b-A+R	A-R+E	R-E+A	E-A+e
quint	b-A+R*E	b~A-R+E*A	A~R-E+A*e	R~E-A+e

6 Grapheme-based Automatic Speech Recognition

The development of a phoneme-based ASR system requires prior resources, such as acoustic resources (i.e., speech data with word level transcription) and a phonetic lexicon. Not all domains or languages may have well developed lexical resources. In the previous section, we presented analyses that showed how the proposed approach was able to capture grapheme-to-phoneme relationships. This suggests that the proposed approach has the capability to integrate lexicon learning as a phase in ASR system training.

Towards that end, in this section we present an ASR study to show that the proposed approach can effectively address the lack of lexical resource problem by exploiting out-of-domain acoustic and lexical resources. In particular, we present an ASR system where,

1. the relationship between the acoustic speech signal and phonemes is learned with out-of-domain acoustic and lexical resources and,
2. the grapheme-to-phoneme relationship is learned using in-domain acoustic resources. In doing so, the ASR system uses a lexicon based on graphemes, which is easy to obtain given the orthographic transcription of words.

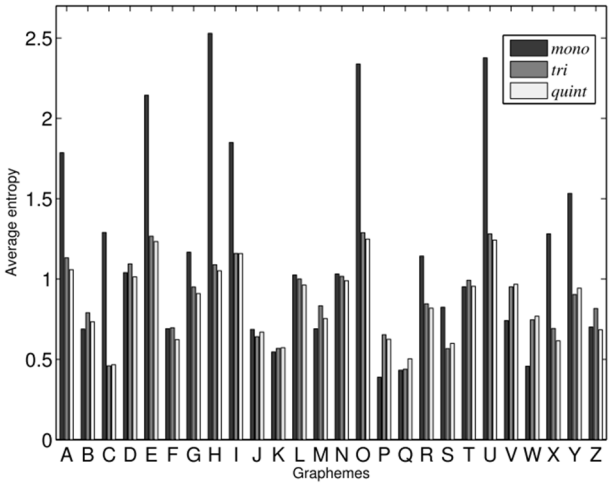

Figure 2: Entropy of grapheme models with increasing context. For contexts tri and quint, average entropy of all the grapheme models with same center grapheme is displayed.

In this study, the DARPA RM task serves as the in-domain task for which we treat as having no phoneme lexicon. The acoustic and lexical resources of WSJ1 corpus serve as the out-of-domain data. We used the out-of-domain ANN described earlier in Section 4.2 to estimate phoneme class conditional probabilities, i.e. z_t, and built two KL-HMM ASR systems using RM data:

1. A grapheme-based ASR system: the states of KL-HMM represent cross-word context-dependent graphemes. In this case, the KL-HMM models the grapheme-to-phoneme relationships. This system represents the case where no phoneme lexicon is available for the target domain, i.e. DARPA RM task.
2. A phoneme-based ASR system: the states of KL-HMM represent cross-word context-dependent phonemes. In this case, the KL-HMM models the phoneme-to-phoneme relationship. We used the well-developed phoneme lexicon of the RM corpus to build this system. So, this system represents the case where a well-developed phoneme lexicon is available for the target domain.

Table 3 presents the performances of the two systems. As illustrated, the grapheme-based ASR system was able to achieve performance comparable to the phoneme-based ASR system. Thus, indicating that the proposed approach can effectively address the lexical resource constraint problem. More details about this study,

including comparison with other approaches such as the standard HMM/GMM based ASR system can be found in Rasipuram's thesis (2014).

Table 3: Performance of grapheme-based and phoneme-based ASR systems expressed in terms of word error rate. A conventional context-dependent phoneme-based HMM/GMM ASR system achieves a performance of 4.1% word error rate (Hain and Woodland, 1999).

grapheme	phoneme
4.5%	4.1%

7 Acoustic Data-Driven Grapheme-to-Phoneme Conversion

In this section, we present a novel acoustic data-driven grapheme-to-phoneme conversion approach that exploits the learned grapheme-to-phoneme relationships. This approach was originally proposed by us in (Rasipuram and Magimai.-Doss 2012). As illustrated in Figure 3, in this approach,

1. Context-dependent grapheme KL-HMM models are trained first, as shown in the case of the grapheme-based ASR system presented in the previous section,
2. Given the orthographic transcription of the word, the grapheme KL-HMM models are then used to generate a sequence of phoneme posterior probabilities
3. Finally, the phoneme posterior probabilities are decoded using an ergodic HMM to obtain the phoneme sequence.

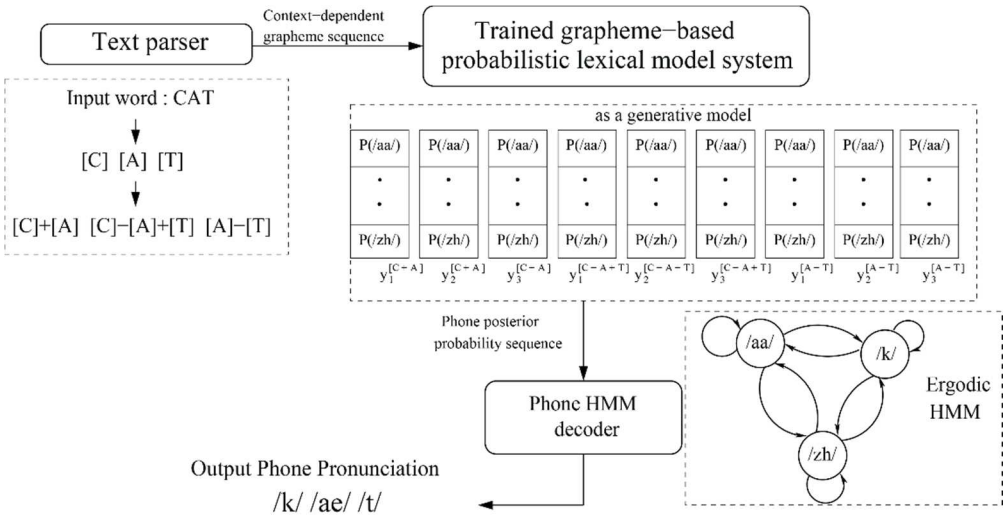

Figure 3: Acoustic data-driven G2P conversion.

One of the key advantages of the acoustic data-driven approach is that, as in the case of ASR, it can exploit the out-of-domain acoustic and lexical resources to build

lexical resources for a new domain or language. Towards that end, building on top of the ASR study presented in the previous section, we present a cross-domain study to demonstrate this capability. In this study, the DARPA RM task served as the target domain from which we were interested in building a phoneme lexicon using acoustic and lexical resources of WSJ1 corpus.

We used the trained context-dependent grapheme KL-HMM models of the ASR study presented in the previous section to develop the phoneme lexicon. We refer to this lexicon as *acoustic-g2p*. Table 4 presents pronunciation of a few words that were extracted using the acoustic data-driven grapheme-to-phoneme conversion approach, along with their respective pronunciations obtained from the RM lexicon.

Table 4: Pronunciation models of a few words generated using the acoustic data-driven G2P approach. By actual pronunciation, we refer to the pronunciation given in the RM lexicon.

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

In order to compare our approach to the state-of-the-art G2P approach, we trained a joint n-gram based G2P converter (Bisani and Ney, 2008), briefly presented in Section 2.1, on the WSJ1 lexicon using Sequitur tool[5] and developed a phoneme lexicon for the DARPA RM task. The graphone width was tuned by excluding 5% of the WSJ1 lexicon as the development set. We refer to this lexicon as *graphone-g2p*. We compare the *acoustic-g2p* lexicon and *graphone-g2p* lexicon by evaluating them at two different levels, namely,

1. At pronunciation level, by comparing the respective lexicons to the RM lexicon.
2. At ASR system level, by building a phoneme-based ASR system for each of the lexicons in the framework of KL-HMM.

Table 5 presents the evaluation at pronunciation level. It can be observed that the graphone-based G2P approach yields better pronunciations than the acoustic data-driven G2P approach.

Table 6 presents the evaluation at the ASR system level. It can be observed that, despite the wide differences in the performance at pronunciation level, the two lexicons yielded similar ASR systems. This suggests that in the acoustic G2P approach, errors at the pronunciation level could be due to substitution with an acoustically similar phone, which is reflected in the in-domain data.

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

Table 5: Evaluation of the extracted pronunciation models at the pronunciation level in terms of phone error rate (PER) and word error rate (WER).

Lexicon	PER	WER
acoustic-g2p	18.5%	65.4%
graphone-g2p	7.8%	27.6%

Table 6: Evaluation of the extracted pronunciation models at the ASR system level in terms of WER.

Lexicon	WER
acoustic-g2p	4.7%
graphone-g2p	4.4%

8 Summary and Discussion

This paper presented a novel approach for learning the relationships between graphemes and phonemes through the acoustic speech signal. In doing so, the approach jointly models the link between the written form and the spoken form as represented by the acoustic speech signal. We showed the potential of the approach in addressing lexical resource scarcity issues. More specifically, we illustrated the (a) development of an ASR system without explicit development of a pronunciation lexicon and (b) development of a pronunciation lexicon using the learned grapheme-to-phoneme relationship. In addition, we demonstrated that the parameters of the KL-HMM can be analyzed to understand the learned grapheme-to-phoneme relationships. The proposed approach opens potential directions for further research and development. In the remainder of this section, we briefly discuss a few of these directions.

It can be observed that the acoustic G2P-based systems (presented in Table 6) yield similar or slightly worse performance, as compared to the grapheme-based ASR system (presented in Table 3). This indicates that the proposed approach can potentially remove the necessity to explicitly build a lexicon, given auxiliary acoustic and lexical resources. Indeed, as shown recently (Imseng et al., 2011; Rasipuram, 2014; Rasipuram and Magimai.-Doss, 2015), ASR systems for new domains and languages can be rapidly developed by

1. training a language or domain independent ANN on multilingual data obtained from resource-rich languages to classify multilingual phones[6] and,
2. learning a probabilistic relationship between the target language graphemes and the multilingual phones on a relatively small amount of transcribed speech data.

[6] The central idea is that phonemes are sharable across languages. So, the relationship between phonemes and acoustic signal can be modeled in a language independent manner.

Furthermore, this approach allows the possibility to perform ASR in a new language without using any acoustic and pronunciation lexical resources of that language (Rasipuram et al., 2013a). In this case, the probabilistic grapheme-to-phoneme relationships are knowledge-based, which can be adapted in an unsupervised manner if untranscribed speech from the target language is available. In other words, the proposed approach can address both acoustic resource and lexical resource scarcity issues. This is particularly interesting for the development of ASR systems for minority languages that do not have well-developed resources, for instance see (Rasipuram et al., 2013b).

As noted in Section 2.1, the starting point for the development of a pronunciation lexicon is extraction of the grapheme-to-phoneme rules obtained from the linguistic studies of the target language and the viability of the G2P techniques (described in that section) rely on the availability of a seed lexicon in the target language. There are a number of languages in the world that do not have such well-developed linguistic resources (Besacier et al., 2014). As discussed above, the proposed approach enables development of ASR systems without explicit pronunciation lexicon development by borrowing resources from resource-rich languages and domains, and learning the relationship between graphemes and multilingual phones on target language speech data. This aspect can be exploited together with the acoustic data-driven G2P approach presented in Section 7 to build pronunciation lexicons for resource scarce languages. This is interesting not only for ASR, but also for TTS. In that regard, there is an on-going project AddG2SU at Idiap[7].

HMM-based ASR systems and statistical parametric speech synthesis systems (also referred to as HMM-based TTS systems [Zen et al., 2009]) have a few components in common, such as a pronunciation lexicon, modeling of the relationship between the acoustic speech signal and phonemes, which basically models the link between graphemes, phonemes and the acoustic speech signal in a similar manner. In other words, similar to HMM-based ASR system, the HMM-based TTS system must first model the relation between words (textual form) and phonemes through a pronunciation lexicon and then the relationship between phonemes and the acoustic signal is modeled via a generative model, such as GMMs. These two separate modeling steps in a TTS system could be linked through the acoustic G2P approach, presented in Section 7, to take advantage of the benefits provided by the proposed approach, in particular in addressing challenges related to resource scarcity. More precisely, this could be achieved by learning the probabilistic relationship between the graphemes and the clustered context-dependent phone HMM states that emit spectral-based acoustic feature vector. Furthermore, such an approach could possibly aid in bridging the gap between HMM-based ASR and TTS technologies (Dines et al., 2010).

[7] https://www.idiap.ch/scientific-research/projects/flexible-acoustic-data-driven-grapheme-to-subword-unit-conversion

Finally, the modeling of grapheme-to-phoneme relationships inherently assumes that the spoken language has a writing system. However, there are spoken languages that do not have a writing system (Besacier et al., 2014). As discussed above, the proposed approach enables borrowing of lexical or phonetic resources from other languages. Along similar lines, in conjunction with field linguistics, it could be possible to extend the proposed approach to borrow written scripts or graphemes from other languages to build a writing system for languages that do not have one. This is highly challenging, but interesting, from both spoken language research and spoken language preservation perspectives.

Acknowledgment

This work was supported by the Swiss NSF through the grants "Flexible Grapheme-Based Automatic Speech Recognition (FlexASR)" and the National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management" (www.im2.ch).

References

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

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

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

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

van Coile, B. 1990. Inductive learning of grapheme-to-phoneme rules. In *Proceedings of Int. Conf. Spoken Language Processing.* 765–768.

Davel, M. and E. Barnard 2008. Pronunciation prediction with Default&Refine. *Computer Speech and Language* 22. 374–393.

Dedina, M. and H. Nusbaum 1991. PRONOUNCE: a program for pronunciation by analogy. *Computer Speech and Language* 5. 55–64.

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

Dines, J. and M. Magimai.-Doss 2008. A Study of Phoneme and Grapheme based Context-Dependent ASR Systems. In Popescu-Bellis, A. and Renals, S. (eds.): *Machine Learning for Multimodal Interaction (MLMI). Lecture Notes in Computer Science No. 4892.* Edinburgh: Springer Verlag, 215–226.

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

Gold, B. and N. Morgan 1999. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music.* New York: John Wiley and Sons

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

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

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

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

Killer M., S. Stüker and T. Schultz 2003. Grapheme based Speech Recognition. In *Proceedings of EUROSPEECH*. 3141–3144.

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

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

Pagel, V., K. Lenzo and A. Black 1998. Letter to Sound Rules for Accented Lexicon Compression. In *Proceedings of Int. Conf. Spoken Language Processing*. 2015–2020.

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

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

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

Rasipuram, R. 2014. *Grapheme-based Automatic Speech Recognition using Probabilistic Lexical Modeling*. Ph.D. thesis, EPFL, Switzerland.

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

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

Rasipuram, R. and M. Magimai.-Doss 2015. Acoustic and Lexical Resource Constrained ASR using Language-Independent Acoustic Model and Language-Dependent Probabilistic Lexical Model. *Speech Communication* 68(4). 23–40.

Rasipuram, R., M. Razavi and M. Magimai.-Doss 2013a. Probabilistic Lexical Modeling and Unsupervised Training for Zero-Resourced ASR. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 446–451,

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

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

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

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

Taylor, P. 2009. *Text-to-Speech Synthesis.* Cambridge: Cambridge University Press.

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

Young, S., G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. C.Woodland 2006. *The HTK Book (for HTK Version 3.4).* Cambridge University Engineering Department, UK.

Zen, H., K. Tokhuda and A. Black 2009. Statistical Parametric Speech Synthesis. *Speech Communication* 51, 1039–1064.