Combining Acoustic Data Driven G2P and Letter-to-Sound Rules for Under Resource Lexicon Generation

Ramya Rasipuram[1,2] *and Mathew Magimai Doss*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland
ramya.rasipuram@idiap.ch, mathew@idiap.ch

Abstract

In a recent work, we proposed an acoustic data-driven grapheme-to-phoneme (G2P) conversion approach, where the probabilistic relationship between graphemes and phonemes learned through acoustic data is used along with the orthographic transcription of words to infer the phoneme sequence. In this paper, we extend our studies to under-resourced lexicon development problem. More precisely, given a small amount of transcribed speech data consisting of few words along with its pronunciation lexicon, the goal is to build a pronunciation lexicon for unseen words. In this framework, we compare our G2P approach with standard letter-to-sound (L2S) rule based conversion approach. We evaluated the generated lexicons on PhoneBook 600 words task in terms of pronunciation errors and ASR performance. The G2P approach yields a best ASR performance of 14.0% word error rate (WER), while L2S approach yields a best ASR performance of 13.7% WER. A combination of G2P approach and L2S approach yields a best ASR performance of 9.3% WER.

Index Terms: Kullback-Leibler divergence based HMM, Lexicon, grapheme, phoneme, grapheme-to-phoneme converter, letter-to-sound rules, multilayer perceptron.

1. Introduction

Automatic speech recognition (ASR) systems and text-to-speech synthesis (TTS) systems tend to model/represent each word in terms of subword units, typically phonemes. A pronunciation lexicon contains the mapping for each word to its phonetic transcription. During the development of ASR or TTS systems it is often presumed that a pronunciation lexicon is available. In other words, pronunciation lexicon is a prior resource. Many of the major languages, such as English, French, German have well-developed pronunciation dictionaries. However, there are languages which may not have such well-developed pronunciation dictionaries. Furthermore, development of lexicon for such languages usually may involve manual effort [1, 2]. The use of small amount of in-domain acoustic training data may help in avoiding the manual effort and improve pronunciation lexicon generation.

In a recent work, we proposed an acoustic data-driven grapheme-to-phoneme (G2P) conversion approach [3]. In this approach, with the aid of Kullback-Leibler divergence hidden Markov model (KL-HMM) [4], the probabilistic relationship between graphemes and phonemes is learned using acoustic data. The learned relationship along with the orthography of word is then used to infer the phoneme sequence or pronunciation model (briefly described in Section 2). In our previous work, we demonstrated the viability of the approach on a simulated scenario, where acoustic and lexical resources of one domain were used to create lexical resources for another domain.

In this paper, we extend our studies to under-resource lexicon development. More precisely, we study a case where small amount of transcribed data is available with a lexicon that contains the pronunciations for the few words that are present in the transcribed data. Given this, the goal is to develop a dictionary for completely unseen words.

On speaker-independent task-independent setup of PhoneBook corpus, we investigate our approach and compare it with standard letter-to-sound (L2S) rule conversion approach using festival toolkit [5] (Section 3). We evaluate the generated pronunciations in terms of pronunciation error (Section 4) and ASR performance (Section 5). Evaluation of the generated pronunciation lexicons in terms of ASR performance yielded a word error rate (WER) of 14.0% for the proposed G2P approach, and 13.7% for L2S approach. However, combining the pronunciation lexicon of both G2P and L2S based approaches yielded a performance of 9.3% WER, signifying the complementary information of the approaches.

2. Acoustic Data Driven G2P Conversion

In a recent work [3], we proposed a two phase G2P conversion approach as depicted in Figure 1. The training phase of proposed G2P approach involves training a grapheme based KL-HMM system which learns the probabilistic grapheme-to-phoneme relationship. Decoding phase infers the pronunciations of words, given the KL-HMM grapheme subword models and orthography of words.

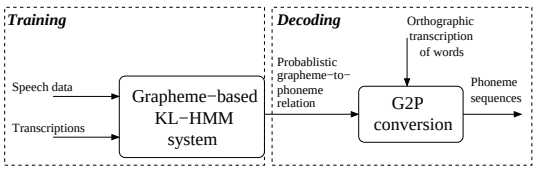

Figure 1: Block diagram of the proposed two phase G2P conversion approach.

2.1. Training Phase

In this phase, a grapheme-based KL-HMM system is trained using phoneme posteriors as features as detailed in [6] and briefly described here:

1. KL-HMM directly models posterior probabilities of phonemes estimated using a multilayer perceptron (MLP) as observation features in HMM.

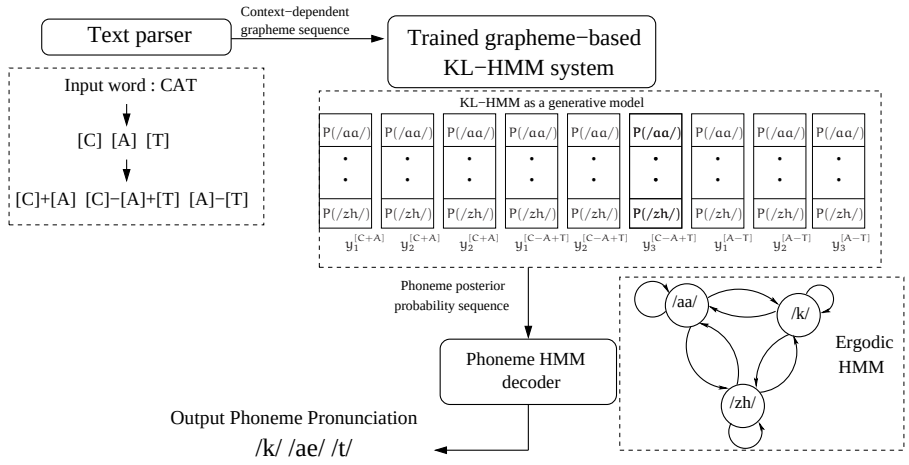

Figure 2: Acoustic data driven G2P conversion using KL-HMM grapheme subword models and orthographic transcription of words.

2. States of HMM are parameterized by multinomial state distributions. Multinomial state distributions are estimated using Viterbi expectation maximization algorithm which minimizes the cost function based on Kullback-Leibler (KL) divergence based local score. In this work, states represent context-dependent grapheme subword units.
3. KL-HMM system requires a posterior feature estimator, in this work we use a trained MLP
4. State multinomial distributions of grapheme subword models capture the probabilistic relation to phonemes

2.2. Decoding Phase

As shown in the block diagram of Figure 2, decoding phase involves inference of phoneme sequence given the KL-HMMs of grapheme subword units and the orthographic transcription of the word. More precisely, decoding phase involves the following steps:

1. The orthography of a given word is parsed to extract context-independent grapheme sequence and then context-dependent grapheme sequence. As shown in the figure, word CAT is parsed to extract its context-dependent grapheme sequence [C]+[A] [C]-[A]+[T] [A]-[T].
2. A word level HMM is created by concatenating the context-dependent grapheme sequence. A sequence of phoneme posterior probabilities is then obtained by stacking the multinomial distributions of the states as shown in Figure 2. In other words, the grapheme KL-HMM acts like a generative model where each state (in the left-to-right sequence) generates a single phoneme posterior probability vector.
3. Phoneme posterior probabilities in sequence are decoded by a phoneme HMM decoder, which is an ergodic HMM connecting all phonemes. More precisely, phoneme posterior probabilities in sequence are used as local scores similar to hybrid HMM/MLP system.

3. Experimental Setup

In this paper, we consider a scenario where limited transcribed speech data along with its pronunciation lexicon constituting pronunciations of words seen in the speech data is available. The goal is to infer pronunciation models for words which are not seen in the training data (For example, to augment the train pronunciation lexicon with new words).

We validate the proposed approach on PhoneBook speaker-independent task-independent 600 word isolated word recognition corpus [7]. PhoneBook corpus was chosen to evaluate the study because of two reasons, (1) test vocabulary consists of words and speakers which are unseen during training, i.e., training and test vocabulary/speakers are completely different (2) corpus consists of unusual/uncommon words, thus, extracting pronunciations for them is a difficult task. We use the small training setup defined in [8]. Table 1 gives the overview of the PhoneBook corpus in terms of number of utterances, speakers and words present in train, cross-validation and test sets. PhoneBook pronunciation lexicon is transcribed using 42 phonemes (including silence).

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

Number of	Train	Cross-validation	Test
Utterances	19421	7920	6598
Speakers	243	106	96
Words	1580	603	600

MLP was trained on limited training data of PhoneBook corpus to classify 42 context-independent phonemes. For MLP training, we followed the same setup as in [8], where 19421 utterances are used for training and 7920 utterances for cross validation. Thus, the data used to train the MLP did not contain any of the (test) words for which pronunciations are to be estimated.

Grapheme-based KL-HMM system is built using both training and cross validation utterances consisting of 27341 utterances covering 2183 words. Context-dependent (single preceding and single following) grapheme subword models are trained using phoneme posteriors estimated from the MLP as observation features. Unseen contexts back-off to the corresponding single preceding or single following or context-

independent graphemes. Cost function based on Reverse Kullback-Leibler divergence (RKL) [4, 6] was used to estimate the parameters of multinomial state distributions (of grapheme subword models) as it resulted in minimum KL-divergence on the training data compared to other KL-based local scores. Each grapheme subword is modeled as a three state left-to-right HMM.

We estimate the pronunciation models for 600 words in the test set using the proposed G2P approach. In the phoneme HMM decoder of the G2P approach, each phoneme was modeled by a 3-state HMM. We compare the proposed G2P approach with L2S approach, where the letter-to-sound rules are learned on train and cross-validation pronunciation lexicon (consisting of pronunciations for 2183 words) using festival tool kit [5]. Thus, grapheme-based KL-HMM system is trained on utterances covering 2183 words and L2S rules are learned on 2183 word pronunciation lexicon, where as MLP is trained on utterances covering 1580 words. Figure 3 gives an outline of the pronunciation model generation using G2P and L2S approaches.

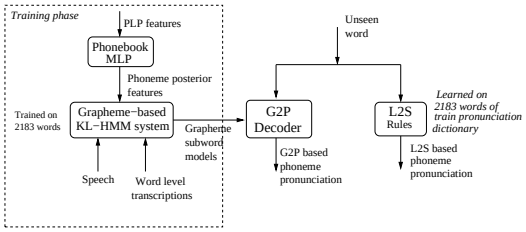

Figure 3: Illustration of pronunciation model generation using G2P and L2S approaches

4. Pronunciation Error Analysis

In this section, we compare the pronunciation models extracted using G2P and L2S approaches with the pronunciations given in the PhoneBook database. Table 2 presents the comparison in terms of phoneme error rate (PER) and word error rate (WER). It can be seen that pronunciation models extracted using L2S rules outperform the proposed G2P based approach. Table 3

Table 2: *Comparing the extracted pronunciation models with actual pronunciations in terms of phoneme error rate (PER) and word error rate (WER).*

System	PER	WER
G2P	27.7%	89.7%
L2S	18.5%	69.0%

shows the percentage of words covered in terms of Levenshtein distance (between the extracted pronunciation and actual pronunciation). About 69.8% of words in case of G2P approach and 81.1% of words in case of L2S approach lie within a Levenshtein distance of two. High error rates of the extracted pronunciations (when compared to the pronunciations given in the database) show the difficulty of the task and insufficiency of acoustic or linguistic knowledge to estimate pronunciations.

Table 3: *Percent of words covered with the given Levenshtein distance between the extracted pronunciation and actual pronunciation for G2P and L2S systems.*

Levenshtein distance	G2P	L2S
0	10.3	31.0
1	39.3	63.3
2	69.8	81.1
3	87.6	92.4
4	96.0	98.3
5	98.5	99.3
6	100	99.5
7	100	100

5. ASR Performance Analysis

The following context-independent phoneme subword KL-HMM-based ASR systems [4, 6] are built to evaluate different extracted pronunciation lexicons:

1. *BASE*: System using pronunciation lexicon given in the PhoneBook database.
2. *G2P*: System using pronunciation models extracted using the proposed G2P approach.
3. *L2S*: System using pronunciation models extracted using L2S rules.
4. *G2P+L2S*: System using pronunciation lexicon consisting of two pronunciations for each word, one from G2P and one from L2S. This lexicon was created to examine the complementary learning from acoustic data and linguistic data.

Posterior features for the KL-HMM system are estimated using PhoneBook MLP and the multinomial state parameters are estimated by optimizing the cost function based on symmetric Kullback Leibler divergence (SKL) local score [4, 6], since SKL resulted in better performance for ASR.

Table 4 gives the ASR performance in terms of word error rates (WER) for different systems on 600-word vocabulary PhoneBook task. *PB-train* refers to the case where only the test pronunciation lexicon is estimated using either G2P or L2S approaches where as the train pronunciation lexicon given in PhoneBook corpus is used to train KL-HMM monophone subword models. *PM-train* refers to the case when both train and test pronunciation dictionaries are extracted using either G2P or L2S approaches, i.e., the phoneme KL-HMM subword models are trained using extracted pronunciations.

Results in the Table 4 show that the performance of the system *G2P* is similar to the performance of the system *L2S*. Interestingly, the result shows that, despite the poor performance of pronunciation models extracted using G2P approach compared to L2S approach, ASR results show that the two approaches yield similar performance. Also, the system *G2P+L2S* yields significant performance improvement over the systems *G2P* or *L2S*. This shows that the pronunciation models learned from G2P and L2S approaches provide complementary information to ASR. Results also show that the ASR systems trained on the extracted pronunciations (column *PM-train* of Table 4) result in better performance compared to the ASR systems trained on the pronunciation lexicon given in database (column *PB-train* of Table 4). The reason for this could be the consistency between train and test pronunciation dictionaries. However, the

performance of the systems *G2P* and *L2S* are poor compared to the over optimistic system *BASE*, which has good quality pronunciations. Nevertheless, it is encouraging that the combined system approaches the performance of system *BASE*.

Table 4: *Word error rates (WER) of different ASR systems expressed in percentage.*

System	*PB-train*	*PM-train*
BASE	3.3%	–
G2P	15.8%	14.0%
L2S	16.0%	13.7%
G2P+L2S	10.3%	9.3%

6. Summary and Discussion

In this work, we compared our acoustic data-driven G2P approach with standard L2S rule based technique for under-resource lexicon development. We found that both techniques yield similar ASR performance. However, when combined yield significant performance gain, thus suggesting that our approach and traditional letter-to-sound conversion approach could be exploited together especially, for under-resource lexicon development. Also, by combining the two dictionaries, we combine knowledge driven approach (L2S) and data driven approach (G2P). In literature, there are similar efforts to use acoustic data and conventional grapheme-to-phoneme conversion approach, such as [9, 10], where they use multigram grapheme-to-phoneme conversion approach [11] and acoustic data together. One major distinction between the approach presented in [9, 10] and the approach investigated here is, we do not need acoustic data of the word when inferring the pronunciation model.

In this work, we observe large ASR performance difference between the baseline dictionary and the extracted dictionaries. Generally, L2S rules for pronunciation model generation yield better performance only when trained on very large pronunciation lexicon. In this work, L2S rules are trained on very limited training data. Compared to our previous work [3], the reason for the large ASR performance difference between baseline dictionary and the dictionary extracted using proposed approach could be,

1. In our previous work, MLP was trained on large amount of out-of-domain data (about 80 hours of speech). While, in this work the MLP is trained on only 5 hours of speech. Moreover, in the previous work speech data was from microphone, where as in the current work it is telephone speech.
2. In our previous work, grapheme KL-HMMs were trained on the words for which the pronunciations were to be extracted. So there was no issue of unseen context. While, in this work the words are neither seen in MLP training nor during KL-HMM training. So, many a times we observe that unseen context-dependent grapheme KL-HMM models back-off to context-independent grapheme. As observed in [6], the states of context-independent grapheme KL-HMM tend to capture information about different phonemes in different states. This may have led to more errors in the extracted pronunciations.

There are further refinements that are worth investigating to improve the acoustic data-drive G2P conversion approach in the context of under resource lexicon development such as,

- phonotactic constraints: in this work, the phoneme inference was obtained using ergodic HMM. It is possible to extract phoneme n-gram on the training dictionary and use it during phoneme decoding.
- n-best list: in this work we used 1-best output as the inferred pronunciation. It would be interesting to see the use of multiple pronunciations extracted with N-best list.
- MLP retraining: in this work the MLP was trained on the baseline dictionary. Similar to what we did with *PM-train*, we could retrain the MLP with alignments obtained with the extracted pronunciations for the training dictionary.

We will scrutinize these issues in our future work in addition to applying the approach on an under-resourced language, such as Scottish Gaelic.

7. Acknowledgements

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). The authors would like to thank their colleague John Dines for the help with festival toolkit and fruitful discussions.

8. References

[1] M. Davel and O. Martirosian, "Pronunciation Dictionary Development in Resource-Scarce Environments," in *Proc. of Interspeech*, 2009.

[2] S. Maskey, L. Tomokiyo, and A. Black, "Bootstrapping Phonetic Lexicons for New Languages," in *Proc. of ICSLP*, 2004.

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

[4] G. Aradilla, H. Bourlard, and M. Magimai-Doss, "Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task," in *Proc. of Interspeech*, 2008.

[5] P. Taylor, A. Black, and R. Caley, "The Architecture of the Festival Speech Synthesis System," in *The Third ESCA Workshop in Speech Synthesis*, 1998.

[6] M. Magimai.-Doss, R. Rasipuram, G. Aradilla, and H. Bourlard, "Grapheme-based Automatic Speech Recognition Using KL-HMM," in *Proc. of Interspeech*, 2011.

[7] J. Pitrelli, C. Fong, S. Wong, J. Spitz, and H. Leung, "Phonebook: A Phonetically Rich Isolated Word Telephone Speech Database," in *Proc. of ICASSP*, 1995.

[8] 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 *Proc. of ICASSP*, 1997.

[9] I. Badr, I. McGraw, and J. Glass, "Learning New Word Pronunciations from Spoken Examples," in *Proc. of Interspeech*, 2010.

[10] ——, "Pronunciation Learning from Continuous Speech ," in *Proc. of Interspeech*, 2011.

[11] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, May 2008.