View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Infoscience - École polytechnique fédérale de Lausanne

IDIAP RESEARCH REPORT

ON THE APPLICATION OF AUTOMATIC SUBWORD UNIT DERIVATION AND PRONUNCIATION GENERATION FOR UNDER-RESOURCED LANGUAGE ASR: A STUDY ON SCOTTISH GAELIC

Marzieh Razavi Ramya Rasipuram
Mathew Magimai.-Doss

Idiap-RR-13-2015

JUNE 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

On the Application of Automatic Subword Unit Derivation and Pronunciation Generation for Under-Resourced Language ASR: A Study on Scottish Gaelic

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

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

Abstract

Automatic speech recognition (ASR) systems typically require acoustic and lexical resources (such as a phonetic lexicon) which may not be available, in particular for under-resourced languages. A typical approach to address the issue of lexical resources in the literature is to use graphemes as the subword units. However, the success of grapheme-based ASR systems depends on the grapheme-to-phoneme relationship in the language. In this paper we investigate the potential of using automatically derived subword units (ASWUs) for under-resourced languages that lack lexical resources and have limited acoustic resources. More precisely, we exploit a recently proposed hidden Markov model (HMM) formalism in which the subword units and the associated pronunciations are derived using only target language transcribed speech data. Our experimental studies on Scottish Gaelic, a minority and under-resourced language, show that ASWUs can lead to significantly better ASR systems compared to grapheme subword units. Furthermore, the ASWU-based ASR systems yield comparable performance to the systems using multilingual acoustic and lexical resources.

Index Terms: automatic subword unit derivation, pronunciation generation, hidden Markov model, Kullback-Leibler divergence based hidden Markov model, under-resourced language, automatic speech recognition

1. Introduction

One of the primary steps in building automatic speech recognition (ASR) systems is to obtain a subword unit set together with a lexicon. The most widely used subword units in ASR systems are the linguistically motivated phone units. The phonetic lexicons can be obtained through use of linguistic knowledge by humans. To reduce the amount of human effort, grapheme-to-phone (G2P) conversion approaches are typically used given an initial lexicon [1]. However obtaining an initial lexicon requires linguistic knowledge as well as human effort which may not be available for all languages. As a result, resource-rich languages such as English and French have well developed lexical resources while under-resourced languages such as Scottish Gaelic and Heiti lack proper lexical resources.

The focus of the present paper is on under-resourced languages with no phonetic lexicon and limited transcribed speech data available. To avoid the need for phonetic lexical resources, graphemes can be alternatively used as subword units [2, 3, 4]. Since in under-resourced languages the acoustic resources can also be limited, multilingual and cross-lingual grapheme-based ASR approaches have also been proposed to exploit acoustic resources from other languages [5, 6]. Due to the possibility of mismatch between the grapheme sets of different languages, data-driven methods have been explored to map the graphemes [5]. In another work [6], porting multilingual grapheme models to a new language through poly-phone decision tree based tying was also investigated.

Despite advances in grapheme-based ASR, most of the proposed approaches have two main limitations: 1) as the acoustic observations are more related to phones, the performance of grapheme-based ASR systems depends on the G2P relationship of the language, and 2) as the G2P relationship can differ across languages, sharing grapheme subword models in multilingual grapheme-based ASR systems is not trivial. As a result, the idea of using multilingual grapheme models has not been generally successful. More recently an approach has been proposed which tries to overcome these limitations by training a multilingual phone classifier using auxiliary resources and builds an ASR system by learning the G2P relationship using acoustic data in the target language [7].

Yet another approach to circumvent the need for lexical resources could be to automatically derive subword units and generate pronunciations using acoustic information. There has been a sustained interest in the ASR community to automatically derive subword units from acoustic data [8, 9, 10, 11, 12, 13, 14]. The automatically derived subword units (ASWUs) have gained attention over the linguistically motivated units for three main reasons: 1) they tend to be more data-dependent as they are typically obtained through optimisation of an objective function using training speech data [11], 2) they can possibly help in better handling of pronunciation variations [15], and 3) they can avoid the need for linguistic knowledge which can be beneficial for under-resourced languages.

In this paper, we study the potential of using ASWUs for under-resourced language ASR. For that purpose, we exploit a hidden Markov model (HMM) formalism which could not only derive "phone-like" subword units using only transcribed speech data, but also can infer associated pronunciations for both seen and unseen words [16] (Section 2). Given that, we investigate two aspects: 1) how the ASWU-based system compares against grapheme-based ASR systems, and 2) how the ASWU-based system compares to systems using multilingual acoustic and lexical resources (Section 3). Our experimental studies on Scottish Gaelic, an under-resourced and minority language, show that ASWU-based ASR systems can achieve significantly better performance than grapheme-based systems and can yield comparable performance to the systems using auxiliary language resources (Section 4).

This work was supported by Hasler foundation through the grant AddG2SU. Thanks to University of Edinburgh for providing the data.

2. Background

This section describes some of the proposed approaches for obtaining ASWUs and generating pronunciations and explains the recently proposed HMM formalism used in this paper.

2.1. Relevant Literature

Various approaches have been proposed in the literature for automatic derivation of subword units and pronunciation generation. In the context of unsupervised learning of the subword units, approaches based on segmentation and clustering [17, 18] and spectral based clustering [12] have been proposed using unlabeled speech data. However, typically these approaches are either applied for other tasks such as keyword spotting or they do not provide a complete ASR system.

Other methods have also been proposed which approach this problem by assuming the availability of transcribed speech data. In [11], joint determination of subword units and pronunciations was investigated using a maximum likelihood criterion. However, the pronunciation generation was limited to the seen words during training. In [13] a hierarchical Bayesian model approach was proposed to jointly learn the subword units and pronunciations. In [14] a spectral based clustering approach was used to derive subword units from a context-dependent grapheme-based system. The pronunciations were then transformed using a statistical machine translation (SMT) approach. In a more recent work [16], a novel HMM formalism was proposed in which the ASWUs were obtained through HMM-based clustering and the pronunciations were generated through acoustic data without the necessity for an SMT approach. It was found that the ASWUs obtained through the HMM formalism can lead to better ASR systems compared to ASWUs derived using the approach in [14]. Therefore in this paper, the ASWUs are obtained using the recently proposed HMM formalism.

2.2. HMM Formalism

In this section, we briefly explain the formalism for automatic subword unit derivation and pronunciation generation.

2.2.1. Automatic Subword Unit Derivation

In this formalism, the subword units are derived from the clustered context-dependent units in a grapheme based system using maximum-likelihood criterion. More precisely, the ASWUs are the tied states of a grapheme based HMM/Gaussian mixture model (GMM) system obtained through decision tree clustering. It was demonstrated in the previous study on English that the ASWUs tend to be "phone-like" [16].

2.2.2. Pronunciation Generation

The ASWU-based pronunciations are generated in an acoustic data-driven manner using a recently proposed G2P conversion approach [19] in which the phones are replaced by ASWUs. The approach consists of a training phase and an inference phase. In the training phase, the relationship between graphemes and ASWUs is learned through acoustic data. More precisely, first the relationship between acoustic feature observations and ASWUs is learned through an acoustic model (e.g. an artificial neural network (ANN)). Then grapheme-to-ASWU relationship is learned in the framework of Kullback-Leibler divergence based HMM (KL-HMM) in which [20, 21]:

- The posterior probabilities of the ASWUs estimated from the trained acoustic model are used as feature observations.
- Each state represents a context-dependent grapheme state and is parameterized by a categorical distribution of ASWUs.
- The local score defined at each state is based on the KL-divergence between ASWU posterior feature and categorical distribution.
- The parameters (categorical distributions) are estimated through Viterbi Expectation-Maximization by minimizing a cost function based on KL-divergence local score.

In the inference phase, the learned grapheme-to-ASWU relationship is used to infer the pronunciation for each word. More precisely, given the orthographic transcription of the word, the grapheme-based KL-HMM acts as a generative model and emits a sequence of ASWU posterior probabilities. The sequence of ASWU posterior probabilities is then decoded using an ergodic HMM in which each state represents an ASWU to infer the most probable pronunciation for each word.

The block diagram of the grapheme-to-ASWU (G2ASWU) conversion approach is illustrated in Figure 1. More details about the original approach are found in [19].

3. Experimental Setup

This section explains the under-resourced language used in this study, the database and the experimental setups for subword unit derivation, pronunciation generation and evaluation.

3.1. Scottish Gaelic

Scottish Gaelic belongs to the class of Celtic languages. It is considered as an endangered language spoken by only 60,000 people. There are about 51 phonemes in the language [22]. However, the number of phonemes can change depending on the dialect. The language lacks a proper phonetic lexicon and the available transcribed speech data are limited.

Scottish Gaelic alphabet has 18 letters, consisting of five vowels and thirteen consonants. The long vowels are represented with grave accents (À, È, Ì, Ò, Ù). There are twelve basic consonant types in Scottish Gaelic (B, C, D, F, G , I , L, M, N, P, R, S, T):

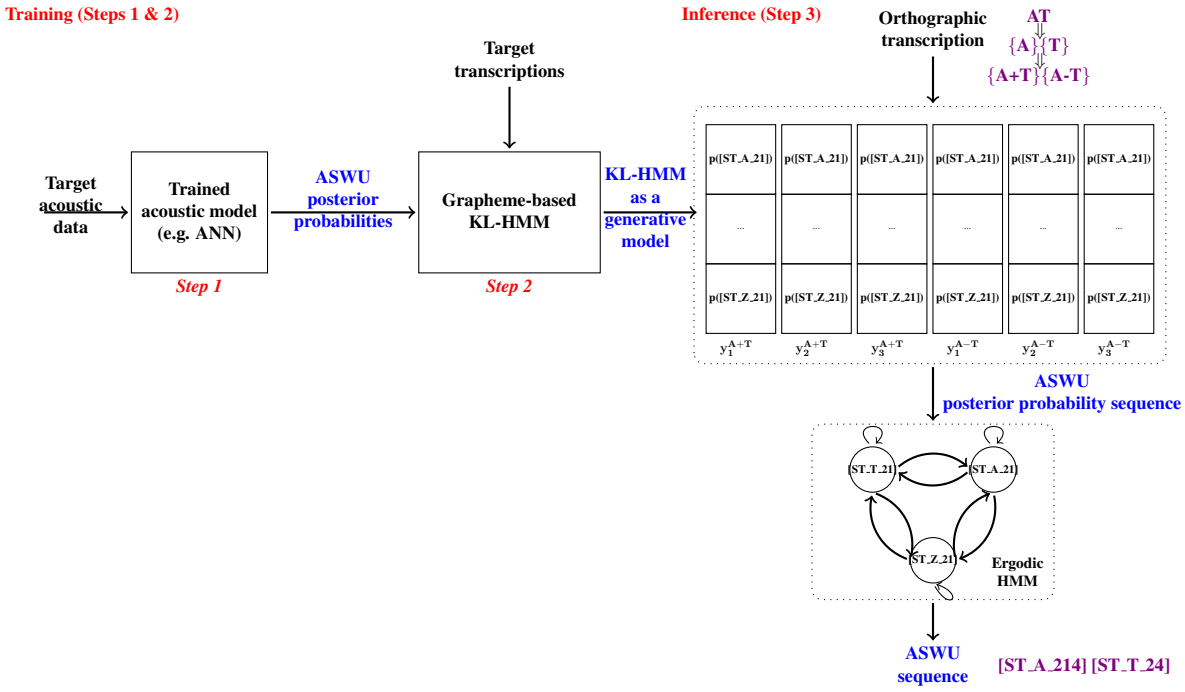

Figure 1: Block diagram of G2ASWU conversion approach. The subword units are represented in the form of HTK clustered states as $[ST\_G\_N]$, with G denoting a grapheme and N denoting a number.

- Each consonant is either fortis or lenis (i.e. they are produced with greater or lesser energy). The lenited consonants are presented in the orthography with a grapheme [H] next to them.
- Each consonant is either broad (velarized) or slender (palatalized). Broad consonants are surrounded by broad vowels (A, O or U), while slender consonants are surrounded by slender vowels (E or I).

Due to the effect of lenited and broad/slender letters on the pronunciation, typically the number of graphemes in a word is relatively larger than the number of phonemes. The G2P relationship in Scottish Gaelic is therefore many-to-one in most cases.

3.2. Database

The Scottish Gaelic corpus was collected by the University of Edinburgh in 2010 and contains recordings from broadcast news and discussion programs[1]. In this paper, the database is partitioned into training, development and test sets according to the structure provided in [23]. The training set contains 2389 utterances with 3 hours of speech and 22 speakers. The development set has 1112 utterances with 1 hour of speech and 12 speakers. The test set consists of 1317 utterances from 12 speakers amounting to 1 hour of speech. There are a total of 2246 unique words in the test set of which 772 are not seen during training.

The database does not provide any phonetic lexicon. The graphemic lexicon can be simply obtained from the orthography of the words. Moreover, the prior knowledge about broad and slender consonants can be applied to the word orthographies. Along these lines in this study, we obtain ASWUs and associated pronunciations in two scenarios:

- *orthography-based*: In this scenario, the graphemic lexicon is obtained directly from the orthography of the words without incorporating any knowledge about the language. As the corpus also contains borrowed English words, the graphemes J, K, Q, V, W, X, Y and Z are also present in the lexicon. Therefore the lexicon consists of 32 graphemes including silence. We refer to this lexicon as *Lex-Gr-Ortho-32*.
- *knowledge-based*: In this case, the graphemic lexicon is obtained by considering broad, slender and lenited consonants as separate graphemes. The lexicon contains 83 graphemes including silence and is referred to as *Lex-Gr-Knowl-83*.

3.3. Automatic Subword Unit Derivation

In order to automatically derive subword units, cross-word context-dependent grapheme-based HMM/GMM systems were trained using HTK toolkit [24]. Each grapheme was modeled with a single HMM state. The decision tree based clustering was done with singleton questions using maximum likelihood criterion to derive the subword units. Through adjusting the log-likelihood increase during decision-tree based state tying, different number of ASWUs were obtained. The number of ASWUs were 85, 91 and 95 in the *orthography-based* scenario and 132, 143 and 163 in the *knowledge-based* case.

[1] http://forum.idea.ed.ac.uk/tag/scots-gaelic

3.4. Pronunciation Generation

To generate ASWU-based pronunciations, first a five-layer multilayer Perceptron (MLP) was trained to classify the ASWUs. We used 39-dimensional PLP cepstral features with four preceding and four following frame context as MLP input. The optimal number of hidden units were obtained based on the frame accuracy on the development set. In most cases, each hidden layer had 1000 hidden units. The MLP was trained with output non-linearity of softmax and minimum cross-entropy error criterion, using Quicknet software [25].

In the next step, given the posterior probabilities of ASWUs as feature observations in the grapheme-based KL-HMM system, context-dependent (single preceding and single following) grapheme subword models were trained. The parameters of the KL-HMM were estimated by minimizing a cost function based on reverse KL-divergence local score [20]. For tying KL-HMM states we applied KL-divergence based decision tree state tying method proposed in [26]. Each grapheme subword unit was modeled with three HMM states. For the pronunciation inference, each ASWU in the ergodic HMM was modeled with three left-to-right HMM states.

Table 1 shows the ASWU-based lexicons together with the MLPs used. The MLPs in the *orthography-based* and *knowledge-based* scenarios are denoted as *MLP-Ortho-N* and *MLP-Knowl-N* respectively where N denotes the number of ASWUs. Similarly the lexicons are represented as *Lex-ASWU-Ortho-M* and *Lex-ASWU-Knowl-M* depending on the scenario with M denoting the actual number of subword units used. It can be seen that some of ASWUs are eliminated in the lexicons. For instance in the *Lex-ASWU-Ortho-76*, from the 85 ASWUs obtained through clustering, only 76 ASWUs are used. In other words, the G2ASWU conversion approach prunes out ASWUs that have less probable G2ASWU relationships.

Orthography-based			*Knowledge-based*		
Lexicon	# of units	MLP	Lexicon	# of units	MLP
Lex-ASWU-Ortho-76	*76*	*MLP-Ortho-85*	*Lex-ASWU-Knowl-103*	103	*MLP-Knowl-132*
Lex-ASWU-Ortho-82	*82*	*MLP-Ortho-91*	*Lex-ASWU-Knowl-111*	111	*MLP-Knowl-143*
Lex-ASWU-Ortho-84	*84*	*MLP-Ortho-95*	*Lex-ASWU-Knowl-128*	128	*MLP-Knowl-163*

Table 1: Summary of the ASWU-based lexicons and the MLPs used.

3.5. Evaluation

We built the following HMM/GMM and KL-HMM ASR systems [20] in *orthography-based* and *knowledge-based* scenarios to evaluate using different types of subword units.

HMM/GMM systems: We compared HMM/GMM systems trained using the ASWUs with the grapheme-based HMM/GMM system. In both cases, we trained cross-word context-dependent HMM/GMM systems with 39 dimensional PLP cepstral features extracted using HTK toolkit [24]. Each subword unit was modeled with three HMM states. For tying the HMM states, singleton questions were used. Each HMM state was modeled by a mixture of 8 Gaussians.

KL-HMM systems: We compared the ASWU-based KL-HMM systems with the grapheme-based KL-HMM systems with the following setups as shown in Figure 2:

- *KL-ASWU*: Depending on the scenario, we used either *MLP-Ortho-N* or *MLP-Knowl-N* as the acoustic model to estimate the posterior probabilities of ASWUs. We then trained context-dependent ASWU-based KL-HMM systems using the posterior probabilities of ASWUs as feature observations.
- *KL-GRAPH*: Instead of using the trained grapheme-based KL-HMM systems explained in Section 3.4 for pronunciation generation, we used them for speech recognition. More precisely, given the test posterior features estimated using either *MLP-Ortho-N* or *MLP-Knowl-N*, we used the trained grapheme-based KL-HMM systems directly for decoding.
- *KL-MULTI*: It was shown in previous studies that it is possible to improve grapheme-based ASR in under-resourced scenarios by using a multilingual phone classifier trained on resource-rich languages and then learning the G2P relationship in the grapheme-based KL-HMM framework using target language data [7, 23]. We compared the ASWU-based KL-HMM system with the grapheme-based KL-HMM system developed in [7] on the exactly same dataset using an MLP trained on 63 hours of speech from five languages to classify multilingual phones (of size 117). The MLP is referred to here as *MLP-MULTI-117*.

4. Experimental Results

Table 2 presents the HMM/GMM performance in terms of word accuracy (WA) in the *orthography-based* and *knowledge-based* scenarios explained in Section 3.2.

Orthography-based		*Knowledge-based*	
Lexicon	WA	Lexicon	WA
Lex-ASWU-Ortho-76	66.3	*Lex-ASWU-Knwol-103*	68.2
Lex-ASWU-Ortho-82	**66.4**	*Lex-ASWU-Knowl-111*	**68.4**
Lex-ASWU-Ortho-84	66.1	*Lex-ASWU-Knowl-128*	68.2
Lex-Gr-Ortho-32	64.6	*Lex-Gr-Knowl-83*	68.2

Table 2: Performance of HMM/GMM systems in terms of word accuracy in the *orthography-based* and *knowledge-based* scenarios.

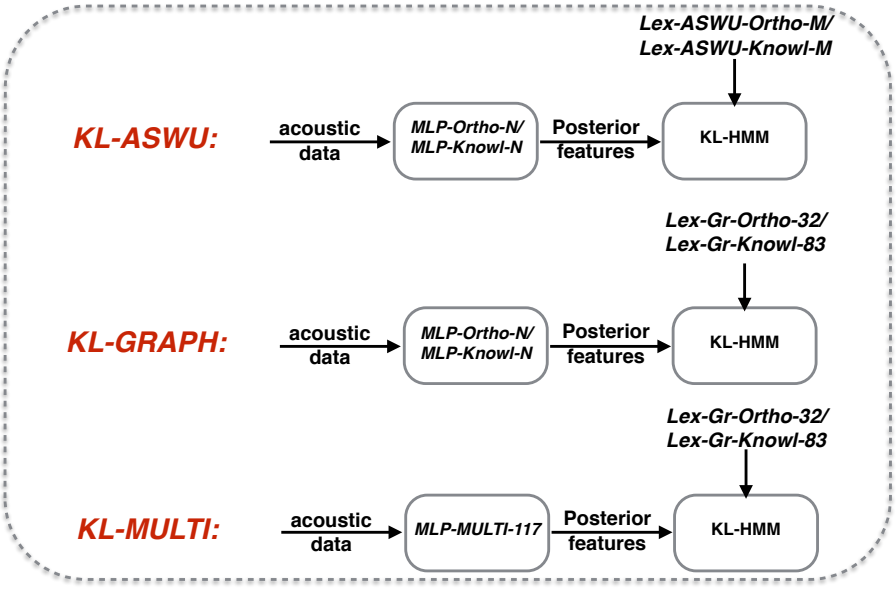

Figure 2: Different KL-HMM systems used in this paper.

It can be observed that in the *orthography-based* scenario, the ASWU-based HMM/GMM systems (irrespective of the number of ASWUs used) perform significantly better than the grapheme-based HMM/GMM system (with at least 95% confidence). In the *knowledge-based* scenario, however, only slight improvements over the grapheme-based system are achieved through use of ASWUs. As it can be seen in Table 2, varying the number of ASWUs does not seem to have a significant effect on the ASR performance. Therefore in the rest of the section the ASR performance using only one set of the ASWUs is reported.

Table 3 shows the performance of KL-HMM systems in terms of word accuracy in the *orthography-based* and *knowledge-based* scenarios. In the *orthography-based* scenario, similar to the observations in the HMM/GMM framework, use of ASWUs (in the *KL-ASWU* system) leads to a significantly better ASR performance compared to use of graphemes (in the *KL-GRAPH* system). More interestingly, the *KL-ASWU* system is able to outperform the *KL-MULTI* system which incorporates auxiliary data from other languages.

System	*Orthography-based*			*Knowledge-based*		
	Acoustic model	Lexicon	WA	Acoustic model	Lexicon	WA
KL-ASWU	*MLP-Ortho-91*	*Lex-ASWU-Ortho-82*	**69.5**	*MLP-Knowl-143*	*Lex-ASWU-Knowl-111*	72.0
KL-GRAPH	*MLP-Ortho-91*	*Lex-Gr-Ortho-32*	66.8	*MLP-Knowl-143*	*Lex-Gr-Knowl-83*	71.2
KL-MULTI	*MLP-MULTI-117*	*Lex-Gr-Ortho-32*	67.9	*MLP-MULTI-117*	*Lex-Gr-Knowl-83*	**72.7**

Table 3: Performance of KL-HMM systems in terms of word accuracy in the *orthography-based* and *knowledge-based* scenarios.

In the *knowledge-based* scenario, the improvements through use of ASWUs in the KL-HMM framework are more considerable than the HMM/GMM framework. More precisely, the *KL-ASWU* system performs significantly better than the *KL-GRAPH* system (with at least 95% confidence). Among different KL-HMM systems, the *KL-MULTI* system achieves the best performance. However, the difference between the *KL-ASWU* and *KL-MULTI* systems is not statistically significant.

The ASR systems in the *knowledge-based* scenario perform better than the systems in *orthography-based* scenario. This shows that integrating prior knowledge in the graphemic lexicon can be helpful for both ASWU- and grapheme-based systems. However, the improvements in the ASR performance through use of prior knowledge are more evident in the grapheme-based systems compared to ASWU-based systems. This could be attributed to the potential of ASWU-based pronunciations in automatically capturing some of the linguistic rules which can be alternatively obtained through prior knowledge.

5. Analysis

To analyze the generated pronunciations using ASWUs, Table 4 presents some of the words together with their pronunciations in the *orthography-based* and *knowledge-based* scenarios. In addition, for each word we have provided its knowledge-based pronunciation as found in the *Lex-Gr-Knowl-83*. For the sake of clarity, we have represented each ASWU of the form $[ST\_G\_N]$ with its corresponding mono-grapheme $[G]$. The broad and slender consonants are represented with the preceding $b\_$ and $s\_$ in the pronunciations in the knowledge-based lexicons. It is also worth mentioning that in Scottish Gaelic, broad consonants *MH* and *PH* are pronounced as the English sounds $/v/$ and $/f/$ respectively. The broad consonant *FH* is not pronounced at all (it corresponds to silence). In addition, the letter I following a vowel does not change its pronunciations.

It can be seen from the ASWU-based pronunciations in the *orthography-based* scenario that the recently proposed HMM formalism is capable of capturing some of the linguistic rules related to pronunciations. For instance, in the word *EUPHORT* the broad consonant PH is mapped to $[F]$ which corresponds to the /f/ sound. Similarly, in the word *MHÀIL*, the broad consonant *MH* corresponds to $[B]$ which is close to the $/v/$ sound. In fact, we can view the generated pronunciations in the form of sequence of mono-graphemes as a way of generating a more phonetic orthography for a given word.

Word	*Lex-ASWU-Ortho-91*	*Lex-ASWU-Knowl-143*	*Lex-Gr-Knowl-83*
MHÀIL	[B] [À] [L]	[b_BH] [À] [s_L]	[b_MH] [À] [I] [s_L]
FHUARAS	[H] [U] [A] [R] [A] [S]	[U] [A] [b_R] [A] [b_S]	[b_FH] [U] [A] [b_R] [A] [b_S]
EUPHORT	[E] [U][F] [O][R][T]	[E] [I] [b_F] [O] [b_R] [b_T]	[E] [U] [b_PH] [O] [b_R] [b_T]
VOTE	[B] [O] [T] [E]	[b_BH] [O][b_T][E]	[V][O][T][E]
YOU	[I] [O]	[I] [U]	[Y][O][U]
WARD	[U] [A] [R] [T]	[U][À][b_R][D]	[W][A][R][D]

Table 4: Sample examples for the generated pronunciations in the *orthography-based* and *knowledge-based* scenarios from the *Lex-ASWU-Ortho-91* and *Lex-ASWU-Knowl-143* respectively.

The ASWU-based pronunciations in the *knowledge-based* scenario, in addition to following linguistic rules, bring some other advantages. In the generated pronunciations, the letters with the same or similar sounds are represented with the same subword unit. For example, as the broad consonants *MH* and *BH* are both pronounced as /v/ sound, in the pronunciation of the word *MHaIL*, *b_BH* is used instead. This can be particularly useful in better handling of sparsity issues in under-resourced scenarios where the amount of data is sparse. Furthermore, the broad consonant *FH* is correctly omitted in the ASWU-based pronunciation of the word *FHUARAS* while it is appeared in the graphemic lexicon. However, for some of the borrowed English words (e.g. *YOU* and *WARD*), the generated pronunciations from the proposed approach (in both scenarios) seem to be influenced dominantly by Gaelic pronunciations. This could be attributed to the limited amount of English words available. As the ASWU-based and graphemic pronunciations seem to provide complementary information to each other, combining the two pronunciations may help in improving the ASR accuracy.

6. Conclusion and Future Directions

In this paper, we studied the potential of using ASWUs for under-resourced language ASR. Towards that goal, we exploited the recently proposed HMM formalism for automatic subword unit derivation and pronunciation generation. Our studies on Scottish Gaelic show that ASWU-based ASR systems can outperform grapheme-based ASR systems and can perform comparable to the ASR systems incorporating acoustic and lexical resources from other languages.

Our focus in this paper was to generate pronunciations using only under-resourced language data. An alternative way for pronunciation generation could be to train a multilingual grapheme-based KL-HMM system as done in *KL-MULTI* system and infer pronunciations using the learned relationship between the graphemes and the multilingual phones. Our future work aims to compare the ASWU-based system to the system using multilingual phone-based pronunciations obtained by that approach.

7. References

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

[2] S. Kanthak and H. Ney, "Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition." in *Proceedings of ICASSP*, 2002, pp. 845–848.

[3] M. Killer, S. Stüker, and T. Schultz, "Grapheme based speech recognition," in *Proceedings of Eurospeech*, 2003, pp. 3141–3144.

[4] T. Ko and B. Mak, "Eigentrigraphemes for under-resourced languages," *Speech Communication*, vol. 56, pp. 132–141, 2014.

[5] S. Stüker, "Integrating thai grapheme based acoustic models into the ML-MIX framework - for language independent and cross-language ASR," in *workshop on SLTU*, 2008, pp. 27–32.

[6] ——, "Modified polyphone decision tree specialization for porting multilingual grapheme based ASR systems to new languages." in *Proceedings of ICASSP*, 2008, pp. 4249–4252.

[7] R. Rasipuram and M. Magimai.-Doss, "Acoustic and lexical resource constrained asr using language-independent acoustic model and language-dependent probabilistic lexical model," *Speech Communication*, vol. 68, pp. 23–40, Apr. 2015.

[8] C. Lee, F. K. Soong, and B. Juang, "A segment model based approach to speech recognition," in *Proceedings of ICASSP*, 1988.

[9] T. Svendsen, K. Paliwal, E. Harborg, and P. Husoy, "An improved sub-word based speech recognizer," in *Proceedings of ICASSP*, 1989, pp. 108–111.

[10] T. Holter and T. Svendsen, "Combined optimisation of baseforms and model parameters in speech recognition based on acoustic subword units," in *Proceedings of ASRU*, Dec 1997, pp. 199–206.

[11] M. Bacchiani and M. Ostendorf, "Joint lexicon, acoustic unit inventory and model design," *Speech Communication*, vol. 29, no. 2, pp. 99–114, 1999.

[12] A. Jansen and K. Church, "Towards unsupervised training of speaker independent acoustic models," in *Proceedings of Interspeech*, 2011, pp. 1693–1692.

[13] C. Lee, Y. Zhang, and J. R. Glass, "Joint learning of phonetic units and word pronunciations for ASR." in *EMNLP*. ACL, 2013, pp. 182–192.

[14] W. Hartmann, A. Roy, L. Lamel, and J. Gauvain, "Acoustic unit discovery and pronunciation generation from a grapheme-based lexicon," in *Proceedings of ASRU*, 2013, pp. 380–385.

[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] M. Razavi and M. Magimai-Doss, "An HMM-based formalism for automatic subword unit derivation and pronunciation generation," *Proceedings of ICASSP*, 2015.

[17] C. Lee and J. R. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proceedings of ACL*, 2012, pp. 40–49.

[18] A. Garcia and H. Gish, "Keyword spotting of arbitrary words using minimal speech resources," in *Proceedings of ICASSP*, 2006, pp. 949–952.

[19] R. Rasipuram and M. Magimai-Doss, "Acoustic data-driven grapheme-to-phoneme conversion using KL-HMM," in *Proceedings of ICASSP*, Mar. 2012.

[20] G. Aradilla, H. Bourlard, and M. Magimai-Doss, "Using KL-based acoustic models in a large vocabulary recognition task." in *Proceedings of Interspeech*, 2008, pp. 928–931.

[21] 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.

[22] M. Wolters, "A diphone-based text-to-speech system for scottish gaelic," Ph.D. dissertation, M.S. thesis, University of Bonn, 1997.

[23] R. Rasipuram, P. Bell, and M. Magimai.-Doss, "Grapheme and multilingual posterior features for under-resourced speech recognition: a study on Scottish Gaelic," in *Proceedings of ICASSP*, 2013.

[24] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.0*. Cambridge University Press, 2000.

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

[26] D. Imseng *et al.*, "Comparing different acoustic modeling techniques for multilingual boosting," in *Proceedings of Interspeech*, Sep. 2012.