Towards Weakly Supervised Acoustic Subword Unit Discovery and Lexicon Development Using Hidden Markov Models

Marzieh Razavi[a,b,*], Ramya Rasipuram[c], Mathew Magimai.-Doss[a]

[a] *Idiap Research Institute, CH-1920 Martigny, Switzerland*
[b] *Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*
[c] *Apple Inc., Cupertino, CA, USA*

Abstract

State-of-the-art automatic speech recognition and text-to-speech systems are based on subword units, typically phonemes. This necessitates a lexicon that maps each word to a sequence of subword units. Development of a phonetic lexicon for a language requires linguistic knowledge as well as human effort, which may not be always readily available, particularly for under-resourced languages. In such scenarios, an alternative approach is to use a lexicon based on units such as, graphemes or subword units automatically derived from the acoustic data. This article focuses on automatic subword unit based lexicon development using methods that are employed for development of grapheme-based systems. Specifically, we present a novel hidden Markov model (HMM) based formalism for automatic derivation of subword units and pronunciation generation using only transcribed speech data. In this approach, the subword units are derived from the clustered context-dependent units in a grapheme based system using the maximum-likelihood criterion. The subword unit based pronunciations are then generated by learning either a deterministic or a probabilistic relationship between the graphemes and the acoustic subword units (ASWUs). In this article, we first establish the proposed framework on a well-resourced language by comparing it against related approaches in the literature and investigating the transferability of the derived subword units to other domains. We then show the scalability of the proposed approach on real under-resourced scenarios by conducting studies on Scottish Gaelic, a genuinely under-resourced language,

*Corresponding author
Email addresses: marzieh.razavi@idiap.ch (Marzieh Razavi), ramya.murali@gmail.com (Ramya Rasipuram), mathew@idiap.ch (Mathew Magimai.-Doss)

and comparing the approach against state-of-the-art grapheme-based ASR approaches. Our experimental studies on English show that the derived subword units can not only lead to better ASR systems compared to graphemes, but can also be transferred across domains. The experimental studies on Scottish Gaelic show that the proposed ASWU-based lexicon development approach scales without any language specific considerations and leads to better ASR systems compared to a grapheme-based lexicon, including the case where ASR system performance is boosted through the use of acoustic models built with multilingual resources from resource-rich languages.

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

1. Introduction

Speech technologies such as automatic speech recognition (ASR) systems and text-to-speech (TTS) systems typically model subword units as they are 1) more trainable compared to words and, 2) more generalizable toward unseen contexts or words. Subword modeling entails development of a pronunciation lexicon that represents each word as a sequence of subword units. Typically in the literature, the subword units are the phonemes or phones. Phonetic lexicon development requires linguistic expert knowledge about the phone set of the language and the relationship between the written form, i.e., graphemes and phonemes. Therefore, it is a time consuming and tedious task. To reduce the amount of human effort, grapheme-to-phoneme (G2P) conversion approaches have been proposed (Pagel et al., 1998; Sejnowski and Rosenberg, 1987; Taylor, 2005; Bisani and Ney, 2008). The G2P conversion approaches still require an initial phonetic lexicon in the target language to learn the relation between graphemes and phonemes through data-driven approaches. While majority languages such as English and French have well-developed phonetic lexicons, there are many other languages such as Scottish Gaelic and Vietnamese that lack proper phonetic resources.

In the absence of a phonetic lexicon, alternatively grapheme subword units based on the writing system have been explored in the literature (Kanthak and Ney, 2002a; Killer et al., 2003; Dines and Magimai.-Doss, 2007; Magimai-Doss et al., 2011; Ko and Mak, 2014; Rasipuram and Magimai.-Doss, 2015; Gales

et al., 2015). The main advantage of using graphemes as subword units is that they make development of lexicons easy. However, the success of grapheme-based ASR systems depends on the G2P relationship of the language. For languages with a regular or shallow G2P relationship such as Spanish, the performance of grapheme-based and phoneme-based ASR systems is typically comparable, whereas for languages with an irregular or deep G2P relationship such as English, the performance of a grapheme-based ASR system is relatively poor when compared to a phoneme-based system (Kanthak and Ney, 2002a; Killer et al., 2003).

Yet another way to handle lack of phonetic lexicon is to derive subword units automatically from the speech signal and build a lexicon based on that. In the literature, interest in acoustic subword unit (ASWU) based lexicon development emerged from the pronunciation variation modeling perspective, specifically with the idea of overcoming the limitations of linguistically motivated subword units, i.e., phones (Lee et al., 1988; Svendsen et al., 1989; Paliwal, 1990; Lee et al., 1988; Bacchiani and Ostendorf, 1998; Holter and Svendsen, 1997). However, recently, there has been a renewed interest from the perspective of handling lexical resource constraints (Singh et al., 2000; Lee et al., 2013; Hartmann et al., 2013). A limitation of most of the existing methods for acoustic subword units based lexicon development is that they are not able to handle unseen words.

In this article, building upon the recent developments in grapheme-based ASR, we propose an approach to derive "phone-like" subword units and develop a pronunciation lexicon given limited amount of transcribed speech data. In this approach, first a set of ASWUs is derived by modeling the relationship between the graphemes and the acoustic speech signal in a hidden Markov model (HMM) framework based on two well-known aspects,

1. alphabetic writing systems carry information regarding the spoken system. Alternatively, a written text embeds information about how it should be spoken. Though this embedding can be deep or shallow depending on the language; and

2. the envelope of the short-term spectrum tends to carry information related to phones.

The ASWU-based pronunciation lexicon is then developed by learning the grapheme-to-ASWU (G2ASWU) relationship through the acoustic signal, and

inferring pronunciations using G2ASWU conversion (analogous to G2P conversion). The G2ASWU conversion process inherently brings in the capability to generate pronunciation for unseen words. The viability of the proposed approach has been demonstrated through preliminary studies on English (Razavi and Magimai-Doss, 2015) and Scottish Gaelic (Razavi et al., 2015), where a probabilistic G2ASWU relationship was learned and pronunciation lexicon was developed.

This article builds on the preliminary works to first extend the approach to the case where a deterministic G2ASWU relationship is learned. We then study and contrast the two G2ASWU relationship learning methods and investigate the following aspects:

1. *Domain-independency of the ASWUs*: Subword units such as phones and graphemes are by default domain-independent. This enables using a lexicon based on either of them across different domains. ASWUs are derived from a limited amount of acoustic speech signal from a domain. Furthermore, the limited data can have undesirable variabilities based on the hardware used and the conditions under which the data is collected. Therefore a question that arises is whether the derived ASWUs are domain independent. Through a cross-domain study on English, we show that our approach indeed yields ASWUs that are domain independent. Furthermore, the proposed approach inherently enables transfering ASWU based lexicon developed on one domain to another.

2. *Potential of ASWUs in improving mulitilingual ASR*: It has been shown that both acoustic resource and lexical resource constraints can be effectively addressed by learning a probabilistic relationship between graphemes of the target languages and a multilingual phone set obtained from lexical resources of auxiliary languages using acoustic data (Rasipuram and Magimai.-Doss, 2015). Success of such approaches lies on the fact that there exists a systematic relationship between linguistically motivated grapheme units and phonemes. Therefore a question that arises is: Does the ASWU-based lexicon based on the proposed approach hold the advantage over grapheme-based lexicon in such a case? Alternately, do the ASWUs exhibit similar systematic relationship to multilingual phones and can it be exploited to further improve the under-resourced language ASR? Through a study on Scottish Gaelic, a genuinely under-resourced language, we show that there exists a systematic relationship between the

ASWUs and multilingual phones, which can not only be exploited to yield systems better than grapheme-based lexicons, but also to gain insight into the derived units.

It is worth mentioning that, to the best of our knowledge, this is the first work that aims to establish these aspects in the context of ASWU-based lexicon development. Consequently, it paves the path for adopting ASWU-based lexicon development and its use for ASR technology development, especially for under-resourced languages.

The remainder of the article is organized as follows. Section 2 provides a background about the grapheme-based ASR and related approaches in the literature for subword unit derivation and pronunciation generation. Section 3 describes the proposed approach. Section 4 presents investigations on the well-resourced majority language English and Section 5 presents the investigations on the under-resourced minority language Scottish Gaelic. Section 6 provides a brief analysis of the derived ASWUs and the generated pronunciations. Finally, Section 7 concludes the article.

2. Background

This section provides the relevant background for understanding the proposed approach for ASWU based lexicon development. Sections 2.1 and 2.2 first present a background on HMM-based ASR and grapheme-based ASR approaches, which form the basis for our proposed approach for automatic subword unit derivation and pronunciation generation. Section 2.3 then presents a survey on the existing approaches for derivation of ASWUs and lexicon development.

2.1. HMM-based ASR

In statistical automatic speech recognition, given the acoustic observation sequence $X = [\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T]$ with T denoting the total number of frames, the goal is to find the most probable sequence of words W^*,

$$W^* = \arg\max_{W \in \mathcal{W}} P(W|X, \Theta), \tag{1}$$

$$= \arg\max_{W \in \mathcal{W}} p(W, X|\Theta), \tag{2}$$

where $\mathcal{W}$ denotes the set of hypotheses and Θ denotes the set of parameters. Eqn. (2) is obtained result of applying Bayes' rule and assuming $p(X)$ to be constant w.r.t all word hypotheses. Hereafter for simplicity, we drop Θ from the equations.

The HMM-based ASR approach achieves that goal by finding the most probable sequence of states Q^* representing W^* by incorporating lexical and syntactic knowledge:

$$Q^* = \arg\max_{Q \in \mathcal{Q}} p(Q, X), \tag{3}$$

$$= \arg\max_{Q \in \mathcal{Q}} \prod_{t=1}^{T} p(\mathbf{x}_t | q_t = l^i) \cdot P(q_t = l^i | q_{t-1} = l^j), \tag{4}$$

$$= \arg\max_{Q \in \mathcal{Q}} \sum_{t=1}^{T} \log(p(\mathbf{x}_t | q_t = l^i)) + \log(P(q_t = l^i | q_{t-1} = l^j)), \tag{5}$$

where $\mathcal{Q}$ denotes all possible state sequences, q_t denotes the HMM state at time frame t and $l^i \in \{l^1, \cdots l^I\}$ denotes a subword unit or lexical unit. Eqn. (4) is derived as a consequence of i.i.d and first order Markov model assumptions.

Estimation of $p(\mathbf{x}_t | q_t = l^i)$ is typically factored through latent variables or acoustic units $\{a^d\}_{d=1}^{D}$ as (Rasipuram and Magimai.-Doss, 2015):

$$p(\mathbf{x}_t | q_t = l^i) = \sum_{d=1}^{D} p(\mathbf{x}_t, a^d | q_t = l^i), \tag{6}$$

$$= \sum_{d=1}^{D} p(\mathbf{x}_t | a^d, q_t = l^i) \cdot P(a^d | q_t = l^i), \tag{7}$$

$$= \sum_{d=1}^{D} p(\mathbf{x}_t | a^d) \cdot P(a^d | q_t = l^i) (\text{assuming } \mathbf{x}_t \perp\!\!\!\perp q_t | a^d), \tag{8}$$

$$= \mathbf{v}_t^{\mathrm{T}} \mathbf{y}_i, \tag{9}$$

where $\mathbf{v}_t = [v_t^1, \cdots, v_t^d, \cdots, v_t^D]^{\mathrm{T}}$ with $v_t^d = p(\mathbf{x}_t | a^d)$ and $\mathbf{y}_i = [y_i^1, \cdots, y_i^d, \cdots, y_i^D]^{\mathrm{T}}$ and $y_i^d = P(a^d | q_t = l^i)$.

As presented above in Eqn. (9), estimation of $p(\mathbf{x}_t | q_t = l^i)$ can be seen as matching acoustic information $\mathbf{v}_t$ with lexical information $\mathbf{y}_i$. In recent years, it has been shown that the match can also be obtained by matching posterior distributions of a^d conditioned on acoustic features and lexical information. One such approach is Kullback-Leibler divergence based HMM (KL-HMM) (Aradilla et al., 2008), where the local score is estimated as the Kullback-Leibler divergence between $\mathbf{y}_i$ and $\mathbf{z}_t$:

$$S_{KL}(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} y_i^d \cdot \log(\frac{y_i^d}{z_t^d}), \tag{10}$$

where $\mathbf{z}_t = [z_t^1, \cdots, z_t^d, \cdots, z_t^D]^{\mathrm{T}} = [P(a^1 | \mathbf{x}_t), \cdots, P(a^d | \mathbf{x}_t), \cdots, P(a^D | \mathbf{x}_t)]^{\mathrm{T}}$.

As KL-divergence is not a symmetric measure, the local score can be esti-

mated in other ways such as,

$$S_{RKL}(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} z_t^d \log(\frac{z_t^d}{y_i^d}), \tag{11}$$

or

$$S_{SKL}(\mathbf{y}_i, \mathbf{z}_t) = \frac{1}{2}(S_{KL} + S_{RKL}). \tag{12}$$

More details about KL-HMM approach are provided in Appendix A.

The HMM-based ASR approach has been primarily built with the idea of having a phonetic lexicon that transcribes each word as a sequence of phones. In conventional HMM-based ASR systems, lexical units $\{l^i\}_{i=1}^I$ model context-dependent phones and acoustic units $\{a^d\}_{d=1}^D$ are clustered context-dependent phone units. $\mathbf{v}_t$ and $\mathbf{z}_t$ are typically estimated using either Gaussian mixture models (GMMs) or artificial neural networks (ANNs); and $\{\mathbf{y}_i\}_{i=1}^I$ is a set of Kronecker delta distributions based on the one-to-one deterministic map between lexical unit l^i and acoustic unit a^d modeled by the state tying decision tree. We refer to this case where l^i and a^d are one-to-one related as *deterministic lexical modeling* framework. In (Rasipuram and Magimai.-Doss, 2015), it has been elucidated that there are HMM-based ASR approaches where the relationship between l^i and a^d is probabilistic. KL-HMM approach, probabilistic classification of HMM states (PC-HMM) approach (Luo and Jelinek, 1999) and tied posterior approach (Rottland and Rigoll, 2000) are examples of *probabilistic lexical modeling* framework. In KL-HMM, $\mathbf{y}_i$ is estimated based on $\mathbf{z}_t$ whereas in PC-HMM and tied posterior $\mathbf{y}_i$ is estimated based on $\mathbf{v}_t$. For a detailed overview on deterministic and probabilistic lexical modeling, the reader is referred to (Rasipuram and Magimai.-Doss, 2015).

2.2. Grapheme-based ASR

In the literature, the issue of lack of a well-developed phonetic lexicon has been addressed by using graphemes as subword units. Most of the studies in this direction have been conducted in the framework of deterministic lexical modeling, where $\{l^i\}_{i=1}^I$ model context-dependent graphemes, $\{a^d\}_{d=1}^D$ are clustered context-dependent grapheme units and $\mathbf{y}_i$ is a decision tree learned while state tying based on either singleton question set or phonetic question set (Kanthak and Ney, 2002b; Killer et al., 2003).

In the framework of probabilistic lexical modeling, it has been shown that grapheme-based ASR systems can be built with $\{a^d\}_{d=1}^D$ based on phones of auxiliary languages or domains, and $\{l^i\}_{i=1}^I$ based on the target language

graphemes. More precisely, a phone class conditional probability $\mathbf{z}_t$ estimator is trained with acoustic and lexical resources from auxiliary languages or domains, and $\mathbf{y}_i$, which captures a probabilistic G2P relationship, is trained on target language or domain acoustic data (Magimai.-Doss et al., 2011; Rasipuram and Magimai.-Doss, 2015). It has been shown that this approach can effectively address both acoustic resource and lexical resource constraints (Rasipuram and Magimai.-Doss, 2015; Rasipuram et al., 2013a). As a natural extension of the approach, an acoustic data-driven G2P conversion approach has been proposed, where the G2P relationship learned in this manner through acoustics is used to infer pronunciations (Rasipuram and Magimai-Doss, 2012; Razavi et al., 2016). We dwell about the acoustic data-driven G2P conversion approach more in the article later, as it is an integral part of the proposed ASWU based lexicon development approach.

2.3. Literature survey on ASWU derivation and pronunciation generation

The idea of using lexicons based on ASWUs instead of linguistically motivated units has been appealing to the ASR community for three main reasons: (1) ASWUs tend to rather be data-dependent than linguistic knowledge-dependent, as they are typically obtained through optimization of an objective function using training speech data (Lee et al., 1988; Bacchiani and Ostendorf, 1998), (2) they could possibly help in handling pronunciation variations (Livescu et al., 2012), and (3) they can avoid the need for explicit phonetic knowledge (Lee et al., 2013).

Typically, the ASWU-based lexicon development process, in addition to the speech signal, requires the corresponding transcription in terms of words, i.e., it is a weakly supervised process similar to acoustic model development in an ASR system.[1] This process involves two key challenges: (a) derivation of ASWUs, which is commonly done through segmentation and clustering and (b) pronunciation generation based on the derived ASWUs. The approaches proposed in the literature can be grouped into two categories based on how these two challenges are addressed. More precisely, there are approaches that decouple these two challenges and address them separately (Lee et al., 1988; Svendsen et al., 1989; Paliwal, 1990; Hartmann et al., 2013), and there are approaches

[1] More recently, in the context of "zero-resourced" ASR system development, there are efforts toward developing methods that are fully unsupervised (Chung et al., 2013; Lee et al., 2015). Such methods are at very early stages and are out of the scope of this article.

that address these two challenges in an unified manner with a common objective function (Holter and Svendsen, 1997; Bacchiani and Ostendorf, 1999, 1998; Singh et al., 2000, 2002; Lee et al., 2013). Here we discuss the prior works that are more relevant to our present work.

In (Hartmann et al., 2013) an approach was proposed based on the assumption that the orthography of the words and their pronunciations are related. In this approach, the subword units are obtained by clustering context-dependent (CD) grapheme models. This is achieved through a spectral based clustering approach (Ng et al., 2001). The pronunciations for seen and unseen words are generated by employing a statistical machine translation (SMT) framework. On the Wall Street Journal task, it was found that the resulting ASWU-based lexicon yields a better ASR system than the grapheme-based lexicon.

In (Bacchiani and Ostendorf, 1999, 1998), a segmentation and clustering approach was exploited for jointly determining the ASWUs and the associated pronunciations, where (1) in the segmentation step, pronunciation related constraints are applied such that a given word has the same number of segments across the acoustic training data, and (2) a maximum-likelihood criteria that is consistent for both segmentation and clustering is utilized. On read speech DARPA Resource Management task, it was shown that the proposed approach leads to improvements over a phone-based ASR system.

In (Singh et al., 2000, 2002), a maximum likelihood strategy was presented which decomposed the ASWU-based ASR system development as the joint estimation of the pronunciation lexicon (including determination of ASWU set size) and acoustic model parameters. More precisely, with an initial pronunciation lexicon based on context-independent graphemes, the acoustic model parameters and the pronunciation lexicon are updated iteratively. The lexicon update step is an iterative process within itself consisting of word segmentation estimation given the acoustic model and update of the lexicon based on the segmentation. After each iteration of lexicon update and acoustic model update convergence is determined by evaluating the ASR system on cross-validation data. If not converged, the ASWU set size is increased and the process is repeated. A proof of concept was demonstrated on DARPA Resource Management corpus.

In (Lee et al., 2013) a hierarchical Bayesian model approach was proposed to jointly learn the subword units and pronunciations. This is done by modeling two latent structures: (1) the latent phone sequence, and (2) the latent letter-to-sound (L2S) mapping rules, using an HMM-based mixture model in which each component represents a phone unit and the weights over HMMs are indicative

of the L2S mappings. It was shown that the proposed approach together with the pronunciation mixture model retraining leads to improvements over the grapheme-based ASR system on a weather query task.

3. Proposed Approach

This section presents an HMM-based formulation to derive ASWUs and develop an associated pronunciation lexicon. Essentially, the formulation builds on grapheme-based ASR in a deterministic lexical modeling framework as well as a probabilistic lexical modeling framework. More specifically, we show that:

1. The problem of derivation of ASWUs can be cast as a problem of finding phone-like acoustic units $\{a^d\}_{d=1}^{D}$ given transcribed speech, i.e., the speech signal and its orthographic transcription, in the grapheme-based ASR framework. Section 3.1 dwells on this aspect.

2. Given the derived ASWUs $\{a^d\}_{d=1}^{D}$ and the transcribed speech, the pronunciation lexicon development problem can be cast as a problem akin to acoustic data-driven G2P conversion (Razavi et al., 2016). Section 3.2 deals with this aspect.

3.1. Automatic subword unit derivation

State clustering and tying methods in HMM-based ASR have emerged from the perspective of addressing the data sparsity issue and handling unseen contexts (Young, 1992; Ljolje, 1994). However, this methodology can be adopted, as it is, to derive acoustic subword units in the framework of grapheme-based ASR. More precisely, we hypothesize and show that the clustered context-dependent grapheme units $\{a^d\}_{d=1}^{D}$ obtained in a context-dependent grapheme based ASR system can serve as ASWUs.

The reasoning behind our hypothesis is that the set of acoustic units $\{a^d\}_{d=1}^{D}$ is obtained by maximizing the likelihood of the training data, which is essentially determined by estimation of $p(\mathbf{x}_t|q_t = l^i)$, as during training the sequence model for each utterance is fixed given the associated transcription and lexicon. As observed earlier in Eqn. (9), $p(\mathbf{x}_t|q_t = l^i)$ estimation involves the matching of acoustic information $\mathbf{v}_t$ with lexical information $\mathbf{y}_i$. We know that standard features, such as cepstral features have been designed to model the envelope of the short-term spectrum, which carry information related to phones. Similarly it is very well known that context-dependent graphemes capture information related

to phones. This is one of the central assumptions in most of G2P conversion approaches, i.e., the relationship between context-independent graphemes and phones can be irregular but the relationship can become regular when contextual graphemes are considered. Therefore, as illustrated in Figure 1, for the likelihood of the training data to be maximized, clustered context-dependent grapheme units $\{a^d\}_{d=1}^{D}$ should model an information space that is common to both the short-term spectrum based feature $\mathbf{x}_t$ space and the context-dependent grapheme based lexical unit l^i space, which we hypothesize to be a phone-like subword unit space.

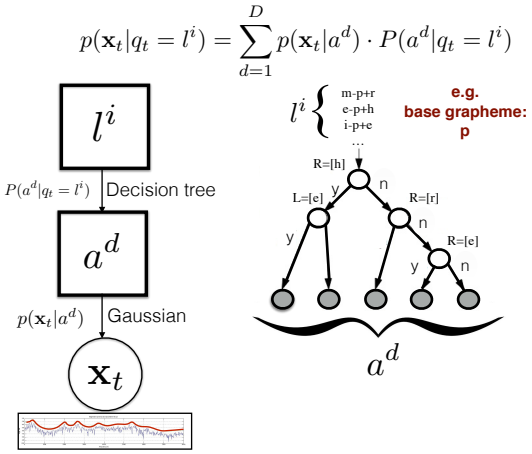


Figure 1: The clustered states a^d of a grapheme-based CD HMM/GMM system obtained through decision tree based clustering are exploited as ASWUs. As for the likelihood of the data to be maximized, $a^d s$ should be related to both CD graphemes l^i and cepstral features $\mathbf{x}_t$, they are expected to be phone-like.

Our argument is further supported by an ASR study that demonstrated the interchangeability of clustered context-dependent phoneme units space and clustered context-dependent grapheme units space in the framework of probabilistic lexical modeling (Rasipuram and Magimai-Doss, 2013) as well as by earlier works on grapheme-based ASR that have explored integration of phonetic information in clustering context-dependent grapheme units and state tying (Killer et al., 2003).

As shall be seen in the later sections, in the proposed approach the set of ASWUs $\{a^d\}_{d=1}^{D}$ is chosen in conjunction with grapheme-to-ASWU conversion via cross-validation.

3.2. Lexicon development through grapheme-to-ASWU conversion

In order to build speech technologies with the derived ASWUs, we need a mechanism to map the orthographic transcription of words to sequences of ASWUs for both seen and unseen words. For that purpose, an approach similar to automatic G2P conversion is desirable. However, conventional G2P conversion approaches are not directly applicable, as they necessitate a seed lexicon that maps a few word orthographies into sequence of phonemes (in our case ASWUs). In this section we present an approach that alleviates the necessity for a seed lexicon by exploiting acoustic information. This approach can be essentially considered as an extension of the grapheme-based ASR approach, where either a deterministic lexical model or a probabilistic lexical model $\{\mathbf{y}_i\}_{i=1}^{I}$ that captures G2ASWU relationship is learned and ASWU-based pronunciations are inferred. We present below these two frameworks.

3.2.1. Deterministic lexical modeling based G2ASWU conversion

This method of lexicon development is a straightforward extension of the ASWU derivation. More precisely, in the process of ASWU derivation a deterministic one-to-one map between context-dependent graphemes ($\{l^i\}_{i=1}^{I}$) and ASWUs ($\{a^d\}_{d=1}^{D}$) is learned. The pronunciations are inferred using this information similar to the decision tree based G2P conversion approach (Pagel et al., 1998), where given the grapheme context, a trained decision tree maps the central grapheme to a phoneme. In our case, the central grapheme is mapped to an ASWU.

3.2.2. Probabilistic lexical modeling based G2ASWU conversion

The other method for ASWU-based lexicon development is to exploit the acoustic data-driven G2P conversion approach using KL-HMM (Rasipuram and Magimai-Doss, 2012; Razavi et al., 2016), which can alleviate the necessity of a seed lexicon in the target domain or language. More precisely, the G2ASWU conversion involves,

1. getting an alignment in terms of the ASWUs $\{a^d\}_{d=1}^{D}$ using the trained grapheme-based HMM/GMM system followed by training of an ANN to estimate $\mathbf{z}_t$;[2] then

[2]If the estimation of $\mathbf{z}_t$ is based on Gaussians then it would amount to going from single Gaussian to GMMs (mixture increment step) of ASR system training.

2. training a context-dependent grapheme-based KL-HMM using $\mathbf{z}_t$ as feature observations; and finally

3. inferring the pronunciations given the KL-HMM parameters $\{\mathbf{y}_i\}_{i=1}^{I}$ and the orthographies of the words in the lexicon. More precisely, first a sequence of ASWU posterior probability vectors is obtained from the KL-HMM given the orthography of the target word. The sequence is then decoded by an ergodic HMM in which each state represents an ASWU to infer the pronunciation.

The main difference between this approach and the deterministic lexical modeling based G2ASWU conversion approach is that the G2ASWU mapping is probabilistic as opposed to being deterministic.

3.3. Summary of the proposed approach

Figure 2 summarizes our approach. As illustrated, the approach consists of three phases. *Phase I* involves derivation of ASWUs. *Phase II* involves learning G2ASWU relationship given the transcription and acoustic data. *Phase III* deals with lexicon development given the G2ASWU relationship and the word orthographies. *Phase II* is explicitly needed for learning the probabilistic G2ASWU relationship. In the case of deterministic G2ASWU conversion, it is implicit in *Phase I*. *Phase III* can be seen as decoding a sequence of ASWU posterior probability vectors $\mathbf{y}_i$. It is worth mentioning that the pronunciation inference step, i.e., *Phase III*, for both deterministic and probabilistic lexical modeling based approaches is the same. More precisely, in the case of deterministic lexical modeling based approach, the inference step is equivalent to decoding a sequence of Kronecker delta distributions resulting from the one-to-one mapping of CD graphemes (in the word orthography) to ASWUs using the decision tree (Razavi et al., 2016).

A central challenge in the proposed approach is how to determine the size of the ASWU set $\{a^d\}_{d=1}^{D}$. In the studies validating the proposed approach, presented in the remainder of the article, we show that this can be achieved via cross-validation. Specifically, a range of values for acoustic units set cardinality D can be considered based on the knowledge that the ratio of number of phonemes to number of graphemes is not an extremely high value, and can be selected via cross-validation at ASR level. For instance in English, if one considers the CMU dictionary, then the ratio is $\frac{38}{26}$ or $\frac{84}{26}$ (when lexical stress is considered). Alternately, the value of D can be chosen relative to the number of

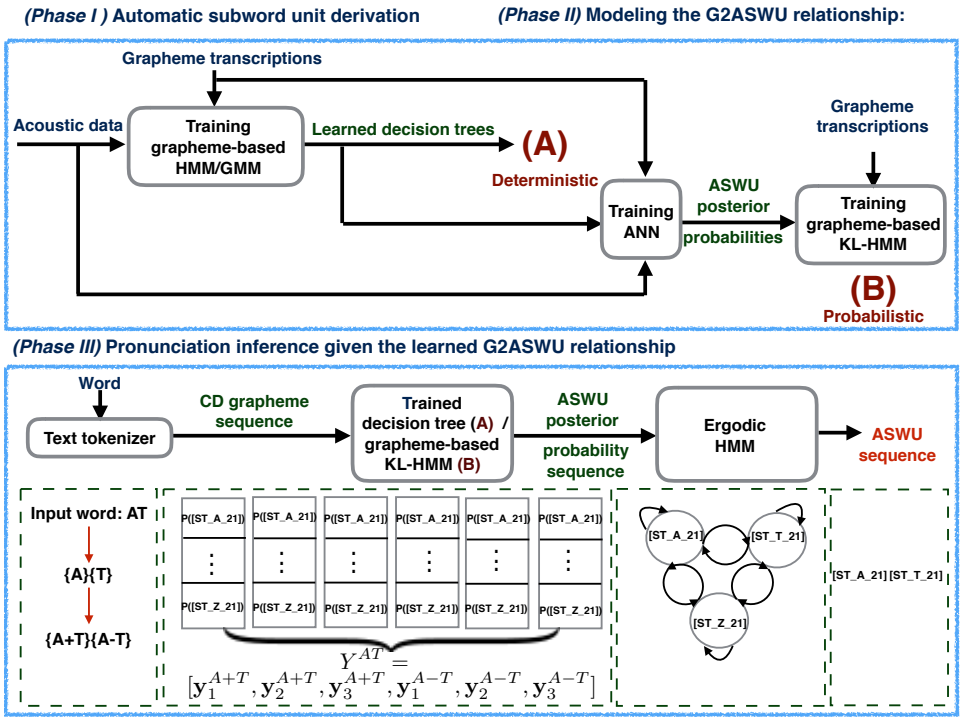


Figure 2: Block diagram of the HMM formalism for subword unit derivation and pronunciation generation. *Phase III* is shown for the case where the ASWU posterior probability vectors from KL-HMM are decoded. For the case where the ASWU posterior probability vectors are obtained from the decision trees (i.e., $\mathbf{y}_i$s are Kronecker delta distributions), only a single posterior probability vector per each context-dependent grapheme is generated, i.e., $Y^{AT} = [\mathbf{y}_1^{A+T}, \mathbf{y}_1^{A-T}]$

graphemes and is much smaller than the number of acoustic units considered for building context-dependent grapheme-based ASR systems, which is typically in the order of thousands.

4. In-Domain and Cross-Domain Studies on Resource-Rich Languages

In this section, we establish the proposed framework for subword unit derivation and lexicon development through experimental studies on a resource-rich language using only its word-level transcribed speech data. The rationale for studying on a well-resourced language is to enable analyzing the discovered subword units and relating them to phonetic identities. We selected English as the well-resourced language, as it is a challenging language for automatic pronunci-

ation generation due to its irregular G2P relationship, and has been the focus of many previous works on ASWU derivation and lexicon development. Our investigations are organized as follows:

1. *Evaluation of the proposed approach through in-domain studies*: We investigate the proposed approach for derivation of ASWUs and corresponding pronunciations on two English corpora, namely Wall Street Journal (WSJ) and Resource Management (RM). We evaluate the ASWU-based lexicons through in-domain ASR studies where the performance of the ASWU-based ASR systems is compared against grapheme-based and phoneme-based ASR systems (Section 4.2).

2. *Investigating the transferability of the ASWUs through cross-domain studies*: A central challenge in ASWU based lexicon development and its adoption for wider use is ascertaining whether the ASWUs derived from limited amount of acoustic resources generalize across domains, similar to linguistically motivated subword units phonemes and graphemes. To the best of our knowledge, none of the previous works have tried to ascertain that aspect. In that sense, we go a step further to conduct cross-domain studies where the ASWUs are derived from the WSJ0 corpus and the lexicon is developed for the RM corpus. We present three methods for development of lexicons in such a scenario, and investigate the transferability of the ASWUs by building and evaluating ASR systems using the developed lexicons (Section 4.3).

3. *Comparison to related approaches in the literature*: In Section 2.3, we discussed a few prominent approaches proposed in the literature for derivation of ASWUs and pronunciation generation. We compare the performance of the our approach with two of the related approaches in the literature studied on WSJ0 and RM corpora (Section 4.4). Indeed, one of the main reasons for selecting these two corpora is to enable comparison to these related works in the literature.

4.1. Databases

This section describes the setup on two corpora used in our experimental studies.

4.1.1. WSJ0 corpus

The WSJ corpus has been originally designed for large vocabulary speech recognition and natural language processing, and it contains a wide range of

vocabulary size (Paul and Baker, 1992). The WSJ corpus has two parts (Woodland et al., 1994) - WSJ0 (Garofolo et al., 1993) with 14 hours of speech and WSJ1 with 66 hours of speech. In this article, we use the WSJ0 corpus for training, which contains 7106 utterances (about 14 hours of speech) and 83 speakers. We report recognition studies on Nov92 test set, which contains 330 utterances from 8 speakers unseen during training. The training set contains 10k unique words. The recognition vocabulary size is 5k words. The language model consists of a bigram model. The grapheme-based lexicon was obtained from the orthography of the words and contained 27 subword units including silence. We refer to this lexicon as Lex-*WSJ*-Gr-27. The phoneme lexicon was based on UNISYN dictionary.

4.1.2. DARPA Resource Management corpus

The DARPA Resource Management (RM) task is a 1000 word continuous speech recognition task based on naval queries (Price et al., 1988). The training set consists of 3990 utterances spoken by 109 speakers amounting to approximately 3.8 hours speech data. The test set, formed by combining Feb89, Oct89, Feb91 and Sep92 test sets, contains 1200 utterances amounting to 1.1 hours of speech data. The word-pair grammer supplied with the RM corpus was used as the language model for decoding. The grapheme-based lexicon was obtained from the orthography of the words. In addition to the English characters, silence, symbol hyphen and symbol single quotation mark were considered as separate graphemes. Therefore, the lexicon contained 29 subword units. We refer to this lexicon as Lex-*RM*-Gr-29. The phoneme lexicon was based on UNISYN dictionary. As mentioned earlier, the RM corpus is mainly used to investigate transferability of the ASWUs across domains. So, it is worth pointing out that 507 out of the 990 words in the RM corpus do not appear in the WSJ0 training set vocabulary.

4.2. In-domain ASR studies

In this section we first explain the setup for derivation of ASWUs and development of ASWU-based lexicons. We then present the in-domain ASR studies for evaluation of the ASWU-based lexicons.

4.2.1. ASWU derivation and lexicon development setup

The setup for subword unit derivation and lexicon development through G2ASWU conversion is as follows:

Acoustic subword unit derivation: Toward automatic discovery of subword units, cross-word single preceding and single following CD grapheme-based HMM/GMM systems were trained with 39-dimensional PLP cepstral features ($c_0 - c_{12} + \Delta + \Delta\Delta$) extracted using HTK toolkit (Young et al., 2000). Each CD grapheme was modeled with a single HMM state. The subword units were derived through likelihood-based decision tree clustering using singleton questions. Different numbers of ASWUs were obtained by adjusting the log-likelihood increase during decision tree based state tying. The numbers of clustered units were obtained such that they are within the range of 2 to 4 times the number of graphemes, based on the general idea explained in Section 3.3. Therefore, for the WSJ0 corpus, ASWUs of size 60, 78 and 90 were investigated, and for the RM corpus, ASWUs of size 79, 92 and 109 were studied.

Deterministic lexical modeling based G2ASWU conversion: Given the learned decision trees for each ASWU set, the pronunciation for each word was inferred by mapping each grapheme in the word orthography to an ASWU by considering its neighboring (i.e., single preceding and single following) grapheme context. We denote the lexicons in the form of Lex-DB-Det-ASWU-M where DB and M correspond to the database and the number of ASWUs respectively. For example, the lexicon generated on WSJ0 corpus using 78 ASWUs is denoted as Lex-WSJ-Det-ASWU-78.

Probabilistic lexical modeling based G2ASWU conversion: In this case, given the obtained ASWUs:

1. A five-layer multilayer Perceptron (MLP) was trained to estimate the posterior probability of ASWUs. The input to the MLP was 39-dimensional PLP cepstral features with four preceding and four following frame context. The hyper parameters such as the number of hidden units per hidden layer were decided based on the frame accuracy on the development set. Each hidden layer had 2000 and 1000 hidden units in the WSJ0 and RM corpora respectively. The MLP was trained with output non-linearity of softmax and minimum cross-entropy error criterion using Quicknet software (Johnson et al., 2004).

2. Using the posterior probabilities of ASWUs as feature observations, a grapheme-based KL-HMM system modeling single preceding and single following grapheme context was then trained. Each CD grapheme was modeled

with three HMM states. The parameters of the KL-HMM were estimated by minimizing a cost function based on the reverse KL-divergence (S_{RKL}) local score (Aradilla et al., 2008), i.e., the MLP output distribution is the reference distribution, as previous studies had shown that training KL-HMM with S_{RKL} local score enables capturing one-to-many grapheme-to-phoneme relationships (Rasipuram and Magimai.-Doss, 2013). Unseen grapheme contexts were handled by applying the KL-divergence based decision tree state tying method proposed in (Imseng et al., 2012).

3. Given the orthography of the word and the KL-HMM parameters, the pronunciations were inferred by using an ergodic HMM in which each ASWU was modeled with three left-to-right HMM states.

During pronunciation inference, some of the ASWUs with less probable G2ASWU relationships were automatically pruned or filtered out. This can be observed from Table 1, which shows the properties of the ASWU-based lexicons together with the MLPs used for the WSJ0 and RM corpora respectively. The MLPs are denoted as MLP-DB-N, with DB and N denoting the database and the size of the ASWU set respectively. Similarly, the lexicons are shown as Lex-DB-Prob-ASWU-M, with M denoting the actual number of ASWUs used in the lexicon. As an example, it can be seen that in Lex-RM-Prob-ASWU-101, from the 109 original ASWU set, only 101 remained after G2ASWU conversion.

Table 1: Summary of the ASWU-based lexicons obtained through probabilistic lexical modeling based G2ASWU conversion for WSJ0 and RM corpora.

(a) WSJ0 corpus

Lexicon	MLP
Lex-WSJ-Prob-ASWU-58	MLP-WSJ-60
Lex-WSJ-Prob-ASWU-74	MLP-WSJ-78
Lex-WSJ-Prob-ASWU-88	MLP-WSJ-90

(b) RM corpus

Lexicon	MLP
Lex-RM-Prob-ASWU-77	MLP-RM-79
Lex-RM-Prob-ASWU-90	MLP-RM-92
Lex-RM-Prob-ASWU-101	MLP-RM-109

4.2.2. Selection of optimal ASWU-based lexicon

Given different lexicons obtained through deterministic and probabilistic G2ASWU conversion, the optimal lexicon was determined based on the ASR accuracy on the development set. More precisely, first HMM/GMM systems using different ASWU-based lexicons were trained with 39-dimensional PLP cepstral features. Then, the ASWU-based lexicon that led to the best performing HMM/GMM ASR system on the development set was selected.[3] In our experiments, in case of using the deterministic G2ASWU conversion for pronunciation generation, Lex-Det-*WSJ*-ASWU-90 and Lex-Det-*RM*-ASWU-92; and in case of using the probabilistic approach, Lex-Prob-*WSJ*-ASWU-88 and Lex-Prob-*RM*-ASWU-90 were selected as the optimal lexicons and are therefore used in the rest of the article.

4.2.3. Evaluation

To evaluate the generated ASWU-based lexicons, we compared the performance of ASWU-based ASR systems with the grapheme-based and phoneme based ASR systems. Toward that, we trained both context-independent and cross-word context-dependent subword unit-based HMM/GMM systems with 39-dimensional PLP cepstral features.[4] Each subword unit was modeled with three HMM states. For the CI grapheme-based systems, the number of Gaussian mixtures for each HMM state was decided based on the ASR word accuracy on the cross-validation set, resulting in 256 and 128 Gaussian mixtures for WSJ0 and RM corpora respectively. In case of using ASWUs, in order to have a comparable number of parameters to the grapheme based ASR system, each HMM state was modeled with 64 and 32 Gaussian mixtures in the WSJ0 and RM corpora respectively. Similarly, for phone subword units, the number of Gaussian mixtures for each HMM state was 128 and 64 in the WSJ0 and RM corpora. In the context-dependent case, for tying the HMM states, only singleton questions were used. Each tied state was modeled by a mixture of 16 and 8 Gaussians on

[3]It is worth mentioning that for WSJ0 and RM corpora there are no explicit development sets defined. To be more precise, in the case of RM the development set (1110 utterances) was merged with the training set (2880) to create training set of 3990 utterances in literature. So, we used the part of the data that was used for early stopping through cross validation in MLP training as the development data, and trained ASWU-based HMM/GMM systems on the remaining part of the training data. For instance, in the case of RM three HMM/GMM systems corresponding to the lexicons *Lex-RM-Prob-ASWU-77*, *Lex-RM-Prob-ASWU-90*, *Lex-RM-Prob-ASWU-101* were trained on 2880 utterances and the lexicon was selected using the 1110 utterances. We followed a similar procedure for WSJ0.

[4]The subword units are either graphemes or ASWUs or phonemes.

WSJ0 and RM corpora respectively. The number of tied states in all the systems trained on a corpus was roughly the same to ensure that possible improvements in the ASR accuracy are not due to the increase in complexity.[5]

Throughout this article, we report the ASR system performances in terms word recognition rate (`100 - word error rate`), denoted as WRR. Furthermore, for comparing the performance of different systems, we applied the statistical significant test presented in (Bisani and Ney, 2004) with the confidence level of 95%.

Table 2 presents the performance of ASR systems based on different lexicons. In the case of using CI units, the ASWU-based ASR systems perform significantly better than the grapheme-based ASR systems in both WSJ0 and RM corpora. In the case of CD units, it can be seen that for the WSJ0 corpus, the HMM/GMM system using ASWUs performs significantly better than the baseline grapheme-based ASR system. For the case of RM corpus, however, the improvements are not statistically significant. This could be due to the fact that in RM task almost all the words are seen during both training and evaluation.[6] In all cases, the ASWU based lexicon yields a system performance that lies between the performance of phoneme-based ASR system and grapheme-based ASR system.

When using CI subword units, it can be seen that the performance of the system using probabilistic lexical modeling based G2ASWU conversion is comparable or even better than the system using deterministic lexical modeling G2ASWU conversion, whereas when using CD subword units, this is not the case. A plausible reasoning for such a trend is that CI subword unit based systems using deterministic lexical modeling based G2ASWU conversion may require more parameters. We tested that by building CI ASWU-based ASR systems using deterministic and probabilistic lexical modeling based pronunciations with varying number of Gaussian mixtures (from 8 to 256). We observed that the difference between the best performing CI ASR systems using deterministic and lexical modeling based G2ASWU conversion is not statistically signif-

[5] For the WSJ0 corpus, the number of tied states was roughly 2000, and for the RM corpus the number of tied states was roughly 3000.

[6] Only two words of the test set are not seen during training (IT+S and REMARK).

icant[7], thus indicating that the deterministic lexical modeling based G2ASWU conversion approach leads to a better ASR system compared to the probabilistic approach. A potential explanation for this difference could be that, unlike the probabilistic lexical modeling based G2ASWU conversion approach, deterministic lexical modeling based G2ASWU conversion approach avoids ASWU deletions and could therefore generate a more consistent pronunciation lexicon for English.

Table 2: HMM/GMM ASR system performances in terms of WRR using CI and CD subword units.

(a) WSJ0 corpus.

Lexicon	CI	CD
Lex-*WSJ*-Gr-26	68.9	85.8
Lex-*WSJ*-Det-ASWU-90	78.6	88.7
Lex-*WSJ*-Prob-ASWU-88	78.7	87.3
Lex-*WSJ*-Ph-45	88.6	93.5

(b) RM corpus.

Lexicon	CI	CD
Lex-*RM*-Gr-29	84.2	94.0
Lex-*RM*-Det-ASWU-92	89.1	94.5
Lex-*RM*-Prob-ASWU-90	90.7	94.2
Lex-*RM*-Ph-45	93.5	95.9

4.3. Cross-domain ASR studies

This section presents a study that investigates the transferability of the ASWUs to a condition or domain unobserved during derivation of ASWUs. As noted earlier, for ASWUs to be adopted for the mainstream speech technology, this characteristic is highly desirable. Toward that we present a cross-database study where the ASWU derivation is carried out on out-of-domain (OOD) WSJ0 corpus and the lexicon is developed for target domain RM corpus. Similar to G2P conversion as elucidated in (Razavi et al., 2016), G2ASWU conversion (presented earlier in Section 3.2) can be seen as a two step process: 1) learning the relationship between the graphemes and the derived ASWUs, and 2) inferring the ASWU sequence (pronunciation) given the word orthography and the learned G2ASWU relationship. We present three methods for cross-domain ASWU-based lexicon development based on that understanding.

[7]For the WSJ0 corpus, the best performing CI ASR systems yielded WRR of 80.1 % and 79.7% ASR when using *Lex-WSJ-Det-ASWU-90* and *Lex-WSJ-Prob-ASWU-88*, respectively. For the RM corpus, the best performing CI ASR systems yielded WRR of 90.2% and 90.7% ASR word when using *Lex-RM-Det-ASWU-92* and *Lex-RM-Prob-ASWU-90*, respectively.

Method-I: Applying standard G2P conversion approach on the seed lexicon obtained from the OOD corpus

One possible way to generate pronunciations for the in-domain RM corpus is to use the ASWU-based lexicon from the WSJ0 corpus as the seed lexicon and train a G2ASWU converter. For this purpose, we investigated one of the state-of-the-art G2P conversion approaches, namely, the joint multigram approach (Bisani and Ney, 2008) for G2ASWU conversion. This was done by using the Sequitur software developed at RWTH Aachen University.[8] In our experiment, the maximum width of the graphone used was one, and the n-gram context size was 6.[9] As shown in Figure 3, first the G2ASWU relationship is learned on the ASWU-based lexicon for the WSJ0 corpus by training the G2ASWU converter. Then given the words in the RM corpus and the learned G2ASWU relationship, the pronunciations are inferred.[10]

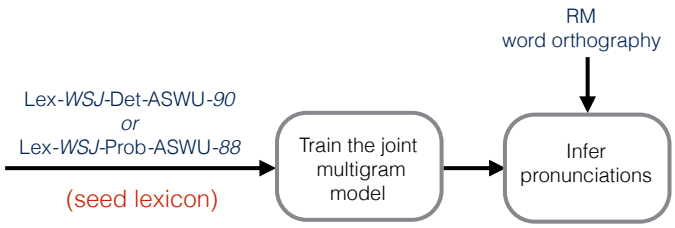


Figure 3: Diagram of joint multigram-based pronunciation generation for RM corpus using the seed lexicon trained on WSJ0 corpus (*Method-I*).

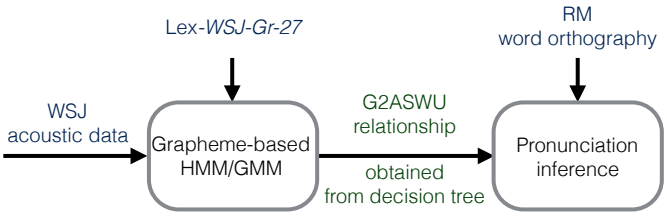
Method-II: Using the learned G2ASWU relationship on the OOD corpus for pronunciation inference on the in-domain corpus

Instead of using the ASWU-based lexicon from the WSJ0 corpus, only the learned G2ASWU relationships can be exploited for inferring pronunciations on the RM corpus. More precisely, we investigate use of the deterministic and probabilistic G2ASWU relationships obtained from (a) the decision trees learned on WSJ0, and (b) the KL-HMM trained on WSJ0, respectively to generate pronunciations for the RM corpus, as illustrated in Figure 4.

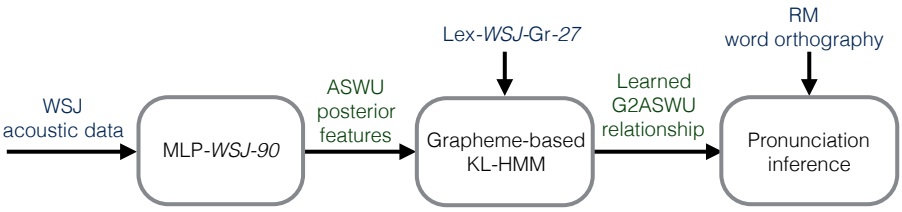
[8] http://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html

[9] As there are no canonical pronunciations in case of using ASWUs are available, we decided on the optimal n-gram context size based on the ASR accuracy on the cross validation set.

[10] The grapheme symbols such as single hyphen that appear in the RM word orthographies and have not been observed in the WSJ0 word orthographies were removed for the inference.



(a) Using a deterministic G2ASWU relationship learned on WSJ0 (*Method-II-a*). The grapheme-based HMM/GMM system is trained on WSJ0 corpus.



(b) Using a probabilistic G2ASWU relationship learned on WSJ0 (*Method-II-b*).

Figure 4: Illustration of pronunciation generation for RM corpus in *Method-II*.

Method-III: Learning the G2ASWU relationship on the in-domain corpus through acoustics

Instead of using the learned G2ASWU relationship on the WSJ0 corpus, we can use the trained MLP on WSJ0 corpus to estimate ASWU posterior probabilities for the RM speech data. Given the ASWU posterior probabilities as feature observations, a grapheme-based KL-HMM system can be trained on the RM corpus data. The pronunciation inference can then be done given the trained KL-HMM and the word orthographies, as shown in Figure 5.

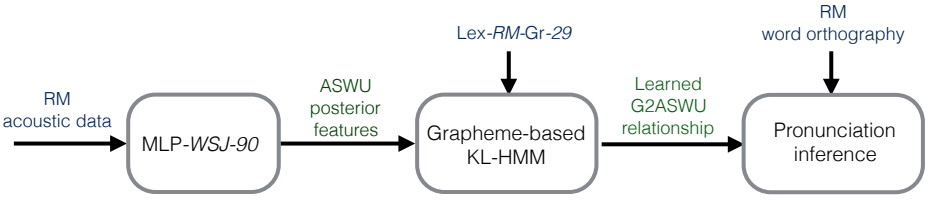


Figure 5: Illustration of pronunciation generation for RM corpus using Method III.

We generated ASWU-based lexicons for the RM corpus based on the three methods presented above. It is worth to reiterate that, in addition to acoustic differences between the two corpora, there are also differences at lexicon level, i.e., 507 out of the 990 words in the RM lexicon do not appear in the WSJ0

lexicon. For each of the lexicons developed, we trained context-independent and cross-word context-dependent ASWU-based HMM/GMM systems with 39-dimensional PLP cepstral features extracted using the HTK toolkit. Each sub-word unit was modeled with three HMM states. Each CI HMM state was modeled by 32 Gaussian mixtures similar to in-domain studies in Section 4.3. Each tied HMM state was modeled by a mixture of 8 Gaussians. The HMM states were tied using a singleton question set.

Table 3 presents the results in terms of WRR. It can be observed that the context-independent ASR systems, regardless of the method used for pronunciation generation, perform better than the grapheme-based CI ASR system (Table 2). The performance of the context-dependent ASWU-based ASR systems using the pronunciations generated through *Method-I* is inferior to the performance of the context-dependent grapheme-based ASR system (Table 2). The performance of the ASR systems using *Method-II* for pronunciation generation is comparable with the ASR systems obtained through in-domain studies (Table 2). Generating pronunciations using *Method-III* also leads to a comparable system to the in-domain ASWU-based ASR systems. Comparing the performance of the systems using *Method-I* for pronunciation generation with the systems using *Method-II* and *Method-III* shows that it is better to transfer the learned G2ASWU relationship or learn the G2ASWU relationship on target domain speech. A potential reason for that is that *Method-I* relies on availability of ground truths, like availability of seed lexicon obtained through linguistic expertise in G2P conversion, which in the present scenario is not available. Overall, *Method-II* leads to the best ASR performance. It may be possible to improve *Method-III* by acoustic model adaptation techniques to adapt the MLP trained on the out-of-domain data. This is open for further research. Together these studies show that, in the proposed approach, the derived ASWUs and the G2ASWU relationship learned from one domain are transferrable to another or target domain. Alternately, the proposed approach inherently enables such transfer.

4.4. Comparison to existing approaches

In this section, we compare the present work with two existing approaches in the literature that have reported studies on the WSJ0 and RM corpora with the same setup as that used in our studies. More precisely, on WSJ0 corpus, Section 4.4.1 compares our approach to the spectral clustering based approach proposed in (Hartmann et al., 2013). Section 4.4.2 studies the proposed ap-

Table 3: ASR system performances in terms of WRR on RM corpus using different cross-domain pronunciation generation methods.

Method	G2ASWU relationship	CI	CD
Method-I	Deterministic	87.5	92.3
	Probabilistic	85.2	91.3
Method-II	Deterministic	89.0	94.4
	Probabilistic	88.8	94.0
Method-III	Probabilistic	89.0	94.0

proach in comparison to the approach proposed by Bacchiani and Ostendorf in (Bacchiani and Ostendorf, 1999).

4.4.1. Comparison to Hartmann et al. (2013) approach

In essence, the proposed approach is similar to the spectral based clustering approach proposed in (Hartmann et al., 2013), as they both discover the ASWUs from the grapheme-based HMM/GMM system. However, there are two key differences between these approaches:

1. In our approach, the ASWUs are discovered through decision-tree based clustering of the HMM states, while in (Hartmann et al., 2013), the sub-word units are derived through spectral based clustering, which requires computation of similarity matrix between HMMs.

2. In our approach, the pronunciations are generated using the KL-HMM framework, while in (Hartmann et al., 2013), the pronunciations are transformed using a statistical machine translation approach.

As the experimental setup in this article on WSJ0 corpus and the work in (Hartmann et al., 2013) are the same, we provide a comparison between the baseline and the results in both works in Table 4. In (Hartmann et al., 2013) there are two grapheme baselines: one based on the standard orthography (denoted as grapheme-direct) and the other based on grapheme-to-grapheme (G2G) conversion (denoted as grapheme-transformed) employing an approach similar to machine translation. Similarly, in the ASWU based study they have two systems: one where the pronunciations are generated directly by mapping the graphemes to ASWUs based on the spectral clustering (denoted as ASWU-direct), and the other where ASWU-to-ASWU conversion is performed like G2G case mentioned above (denoted as ASWU-transformed). We ensured that our systems have comparable number of parameters in the case of both using CI

subword units and CD subword units based systems. It can be observed that the ASWU-based lexicon developed by our approach leads to a better ASR system. Furthermore, when comparing the best systems there is an absolute difference of 2.5% WRR, which indicates that the proposed approach in this article leads to a better ASR system.

Table 4: Comparison with the related work in (Hartmann et al., 2013).

Approach	Lexicon	CI	CD
Approach proposed in (Hartmann et al., 2013)	Grapheme-direct	60.1	84.2
	Grapheme-transformed	68.6	85.5
	ASWU-direct	70.7	85.6
	ASWU-transformed	76.7	86.2
Present work	Lex-*WSJ*-Gr-26	68.9	85.8
	Lex-*WSJ*-Det-ASWU-90	78.6	88.7
	Lex-*WSJ*-Prob-ASWU-88	78.7	87.3

4.4.2. Comparison to Bacchiani and Ostendorf (1999) approach

In a broad sense, the proposed approach and the joint subword unit derivation and pronunciation generation method proposed in (Bacchiani and Ostendorf, 1999) can be considered to be similar as,

1. both approaches consist of segmentation and clustering steps, except that in our approach the segmentation and clustering is guided through graphemes during the HMM/GMM training; and

2. both approaches apply the pronunciation length constraint which ensures uniformity in the number of segments for training tokens of a word. In our approach this is automatically achieved through use of a unique grapheme sequence representation for each word.

In our studies, we have used the RM corpus, which was also used in (Bacchiani and Ostendorf, 1999). However there are a few distinctions. In (Bacchiani and Ostendorf, 1999), the states of the HMMs were modeled by a single Gaussian as opposed to a mixture of Gaussians and the evaluation was carried out only on the *Feb89* test set. So we also trained a single Gaussian HMM/GMM system using the ASWU lexicon developed by our approach and evaluated on the *Feb89* test set. Table 5 presents the results in the case where the two approaches are similar in terms of number of ASWUs and clustered states. Table 6 provides a comparison between the best performance reported in (Bacchiani and

Ostendorf, 1999) and the performance achieved with the lexicon based on our approach on the *Feb89* test set with 2937 clustered states. These results indicate that the ASWU lexicon developed by the proposed approach can yield ASR systems comparable to the ASWU lexicon developed by Bacchiani and Ostendorf (1999) approach, which needs additional heuristics to constrain the ASWU derivation and pronunciation generation process and necessitates all the words to be observed. It seems that our approach requires a higher number of tied states to achieve its best performance, though.

Table 5: Comparison with the related work in (Bacchiani and Ostendorf, 1999) on *Feb89* test set using single Gaussian distributions.

	# of base units	# of clustered states	WRR
Approach proposed in (Bacchiani and Ostendorf, 1999)	124	1519	86.3
Present work	92	1559	86.9

Table 6: Comparison of the best result reported in (Bacchiani and Ostendorf, 1999) on *Feb89* test set with the result using the present work on the same test set using single Gaussian distributions.

	# of clustered states	WRR
Approach proposed in (Bacchiani and Ostendorf, 1999)	1499	91.2
Present work	2937	91.1

5. Application to an Under-Resourced Language

In the previous section, we demonstrated the potential of the proposed framework for subword unit derivation and pronunciation generation on the well-resourced language English. Most of the state-of-the-art speech recognition approaches have emerged through investigations on English. So it can be argued that our approach of deriving ASWUs using grapheme-based HMM/GMM system may be well suited just for English. Furthermore, the G2P relationship varies across languages. So a question that arises is whether the proposed approach is transferable to other languages or not.

In this section, our goal is two-fold: (1) to show the transferability of the approach to a new language, and (2) to show its utility to under-resourced

languages, specifically languages that do not have well-developed phonetic resources. In that direction, we present investigations on a genuinely under-resourced language, Scottish Gaelic. Unlike English, which belongs to the family of Germanic languages, Scottish Gaelic belongs to the family of Celtic languages. Our investigations are organized along two lines,

1. *Monolingual ASR studies*: We investigate the potential of the ASWU-based lexicons through monolingual ASR studies where we compare the performance of the ASWU-based ASR system with the alternative grapheme-based ASR system, as done in the studies on English.

2. *Multilingual ASR studies*: In (Rasipuram and Magimai.-Doss, 2015), it has been shown that performance of the under-resourced ASR system can be significantly improved by (a) training a multilingual acoustic model that estimates multilingual phone posterior probabilities using resources of resource rich languages, and then (b) learning a probabilistic lexical model that captures the grapheme-to-multilingual phone relationship on the target language speech. So we also investigate if the ASWU-based lexicons hold their benefit in such a multilingual ASR system scenario as well. As a product of the study, later in Section 6, we briefly explain how phonetic identities of the derived ASWUs could be discovered.

The remainder of the section is organized as follows. Section 5.1 presents the database and experimental setup used. Section 5.2 presents the details of the ASWU-based lexicon development. Finally, Section 5.3 and 5.4 presents the monolingual ASR and multilingual ASR studies, respectively.

5.1. Database

This section first describes the characteristics of the Scottish Gaelic language. It then explains the Scottish Gaelic corpus used in our studies.

5.1.1. Scottish Gaelic language

Scottish Gaelic belongs to the class of Celtic languages. There are six Celtic languages that are still spoken. These languages are divided into two groups of Goidelic languages and Brythonic languages. Scottish Gaelic belongs to Goidelic languages along with Irish and Manx. It can be considered as a truly endangered language as it is spoken by only about 60,000 people. There are about 51 phonemes in the language (Wolters, 1997). 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 also limited.

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

- Each consonant is either fortis or lenis (i.e., they are produced with greater or less 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).

Scottish Gaelic orthography is less complicated than English. The complications partly arise due to the reason that modern orthography is based on Classical Irish orthography and the letter-to-sound rule may depend on the dialect (Wolters, 1997). The number of graphemes in Gaelic words is typically greater than the number of phones in the word due to the effect of lenited and broad/slender graphemes on the pronunciation. The grapheme-to-phoneme relationship in Scottish Gaelic can therefore be many-to-one.[11] For example, the ratio of the number of graphemes to phonemes in the Gaelic word *SUIDHEACHADH* with pronunciation "sMj@x@G" (in the SAMPA format) is 1.7.

5.1.2. Scottish Gaelic corpus

The Scottish Gaelic corpus was collected by the University of Edinburgh in 2010 and contains recordings from broadcast news and discussion programs.In this article, the database is partitioned into training, development and test sets according to the structure provided in (Rasipuram et al., 2013b). The overview of the Scottish Gaelic corpus is given in Table 7.

The database does not provide any phonetic lexicon. The graphemic lexicon can be simply obtained from the orthography of the words. 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

[11] The many-to-one G2P relationship can actually be seen in other languages as well, e.g., English.

Table 7: Overview of the Scottish Gaelic corpus in terms of number of utterances, hours of speech data and speakers in the train, cross-validation and test sets.

Number of	Train	Cross-validation	Test
Utterances	2389	1112	1317
Hours	3	1	1
Speakers	22	12	12

including silence as shown in Table 8. We refer to this lexicon as Lex-*SG*-Gr-32. The lexicon contains 5083 unique words.

As the corpus does not provide a language model, we used a bigram language model trained on the sentences from the test set, as done in (Rasipuram et al., 2013b).[12]

Table 8: Graphemes used in the Scottish Gaelic corpus.

Vowels	A, E, I, O, U, À, È, Ì, Ò, Ù
Consonants	B, C, D, F, G , H, L, M, N, P, R, S, T
English Graphemes	J, K, Q, V, W, X, Y, Z

5.2. ASWU derivation and pronunciation generation setup

The setup for subword unit derivation and pronunciation generation for Scottish Gaelic is as follows:

Acoustic subword unit derivation: For automatic discovery of subword units, cross-word CD grapheme-based HMM/GMM systems were trained using 39-dimensional PLP cepstral features. Each CD grapheme was modeled with a single HMM state. Different numbers of ASWUs were obtained by adjusting the log-likelihood increase during decision tree clustering. The range for the number of ASWUs was decided to be similar to the range investigated in the studies on English, resulting in 85, 91 and 97 units.

Deterministic lexical modeling based G2ASWU conversion: For deterministic lexical modeling based G2ASWU conversion, the learned decision trees during ASWU derivation were exploited to map each grapheme in the word to an ASWU. We denote the lexicons generated using the deterministic lexi-

[12] This was mainly done as the corpus does not include a language model, and for Scottish Gaelic the resources are limited.

cal modeling based G2ASWU conversion as Lex-*SG*-Det-ASWU-*M* where *M* denotes the number of ASWUs.

Probabilistic lexical modeling based G2ASWU conversion: For probabilistic lexical modeling based G2ASWU conversion, first a five-layer MLP classifying ASWUs was trained in which each hidden layer had 1000 hidden units. The input to the MLP was 39-dimensional PLP cepstral features with four preceding and four following frame context. Then given the ASWU posterior probabilities from the ANN as feature observations, a CD grapheme-based KL-HMM was trained. Each CD grapheme in the KL-HMM was modeled with three left-to-right HMM states. For the pronunciation inference, the ASWU posterior probabilities were decoded through the ergodic HMM in which each ASWU was modeled with three left-to-right HMM states.

Table 9 shows the properties of the ASWU-based lexicons generated using a probabilistic lexical modeling based G2ASWU conversion. Similar to the studies on English, it can be observed that some of the ASWUs are pruned out during the pronunciation generation given the probabilistic G2ASWU mapping.

Table 9: Summary of the ASWU-based lexicons obtained through probabilistic lexical modeling based G2ASWU conversion for Scottish Gaelic corpus.

Lexicon	MLP
Lex-*SG*-Prob-ASWU-76	MLP-*SG*-85
Lex-*SG*-Prob-ASWU-82	MLP-*SG*-91
Lex-*SG*-Prob-ASWU-86	MLP-*SG*-97

We selected the optimal number of ASWUs and the corresponding lexicon based on the WRR on the development set. Lex-*SG*-Det-ASWU-85 and Lex-*SG*-Prob-ASWU-82 yielded the best ASR systems and are therefore used in the ASR studies presented below.

5.3. Monolingual ASR system studies

As mentioned earlier, there is no well-developed phonetic lexicons for Scottish Gaelic. So we evaluate the utility of the developed ASWU-based lexicon against a grapheme-based lexicon by conducting monolingual ASR studies. Specifically, we compare them across two frameworks, namely, HMM/GMM framework and KL-HMM framework.

HMM/GMM framework. We trained CI and cross-word CD HMM/GMM systems with 39-dimensional PLP cepstral features extracted using the HTK toolkit. Each subword unit was modeled with three HMM states. In the case of using CI subword units, the optimal number of Gaussian mixtures for the grapheme-based ASR system was 64 based on the best WRR obtained on the development set. For the ASWU-based ASR systems, the number of Gaussian mixtures was set to 16 so as to have a comparable number of parameters to the grapheme-based system. In the case of using CD subword units, for tying the HMM states singleton questions were used. Each HMM state was modeled by a mixture of 8 Gaussians. The number of tied states in all the systems were roughly the same.

KL-HMM framework. This is done by using the posterior based framework of KL-HMM directly for speech recognition. More precisely, instead of using the KL-HMM parameters for capturing a probabilistic G2ASWU relation for pronunciation inference, they are used in the KL-HMM ASR framework. In this case, we can visualize it as an approach that integrates pronunciation learning implicitly as a phase in ASR system training (Rasipuram et al., 2015). Our main motivation for performing this study was to ascertain whether doing lexicon development and ASR training as two separate stages can bring any advantage over doing direct speech recognition using grapheme-based KL-HMM system. For this purpose, we compared three KL-HMM systems, as illustrated in Figure 6, corresponding to lexicons Lex-*SG*-Gr-32, Lex-*SG*-Det-ASWU-85 and Lex-*SG*-Prob-ASWU-82, respectively. All the systems use the same MLP, which is *MLP-SG-91*, as the acoustic model to estimate posterior feature observations.

Table 10 presents the HMM/GMM systems and KL-HMM systems performance in terms of WRR. It can be observed that Lex-*SG*-Prob-ASWU-82 yields significantly better CI and CD systems than Lex-*SG*-Gr-32 in both the HMM/GMM framework and the KL-HMM framework. Lex-*SG*-Det-ASWU-85 yields a better system in the KL-HMM framework but a worse system in the HMM/GMM framework against Lex-*SG*-Gr-32. A possible reason for such a trend could be that, as discussed earler, in Scottish Gaelic the G2P relationship is many-to-one due to lenition and broad and slender consonants. So, when inferring pronunciations using the deterministic G2ASWU mappings, each grapheme in the word is invariably mapped into an ASWU. This can result in systematic erroneous pronunciations, which could lead to mismatch between

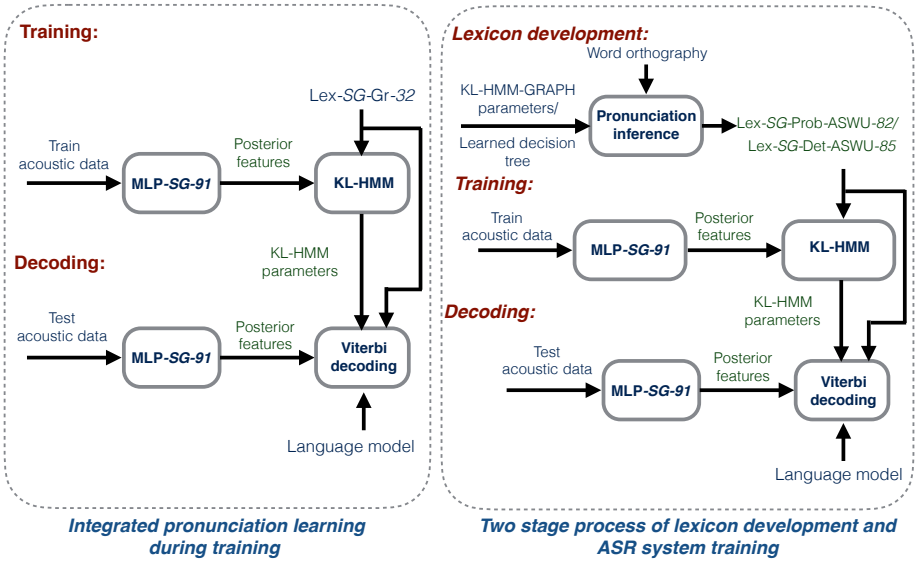


Figure 6: Illustration of KL-HMM based ASR system based on Lex-*SG*-Gr-32, Lex-*SG*-Det-ASWU-85 and Lex-*SG*-Prob-ASWU-82

acoustics and pronunciation model, as in the case of pronunciation variation. In the literature, it has been observed that KL-HMM approach is capable of handling pronunciation variation (Imseng et al., 2011; Razavi and Magimai.-Doss, 2014). As a consequence, unlike HMM/GMM framework, we observe that Lex-*SG*-Det-ASWU-85 yields a better system than *SG*-Gr-32 in KL-HMM framework.

Table 10: Performance of HMM/GMM and KL-HMM systems in terms of WRR using context-independent (CI) and context-dependent (CD) subword units. For the KL-HMM systems, MLP-*SG*-91 is used as the acoustic model.

Lexicon	HMM-GMM		KL-HMM	
	CI	CD	CI	CD
Lex-*SG*-Gr-32	46.0	64.6	35.6	66.8
Lex-*SG*-Det-ASWU-85	54.5	63.3	52.2	69.1
Lex-*SG*-Prob-ASWU-82	59.6	66.4	57.5	69.5

5.4. Multilingual ASR system studies

As mentioned earlier, the under-resourced ASR system performance can be improved by first using an acoustic model or ANN that classifies multilingual

phones and then learning a probabilistic relationship between the graphemes and multilingual phones using KL-HMM. We compared the grapheme-based lexicon and the ASWU-based lexicon in that framework by

1. first training a five-layer multilingual MLP on five auxiliary languages from SpeechDat(II) corpus namely British English, Swiss French, Swiss German, Italian and Spanish to estimate posterior probabilities of multilingual phones. The multilingual phoneset was formed by merging the phones that are shared across the aforementioned languages, leading to 117 phone units. We refer to this MLP as MLP-*MULTI*-117; and then

2. training a KL-HMM based ASR system corresponding to each of the lexicons Lex-*SG*-Gr-32, Lex-*SG*-Det-ASWU-85 and Lex-*SG*-Prob-ASWU-82, as illustrated in Figure 7.

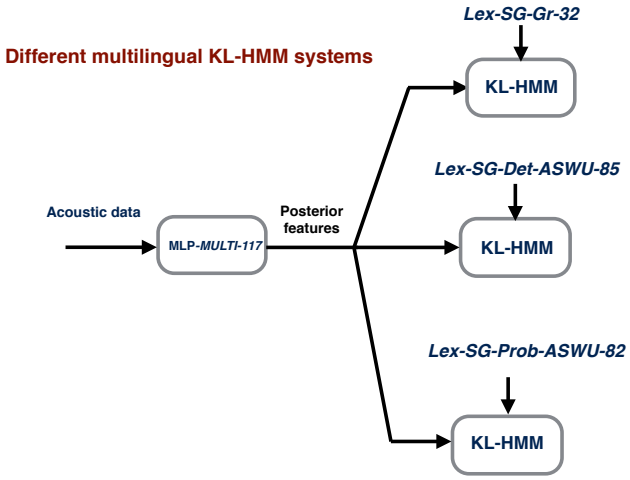


Figure 7: Illustration of KL-HMM based ASR system on Lex-*SG*-Gr-32, Lex-*SG*-Det-ASWU-85 and Lex-*SG*-Prob-ASWU-82 that exploits auxiliary multilingual resources.

Table 11 presents the performance of the different KL-HMM based systems in terms of WRR. It can be observed that the ASWU-based lexicon yields a significantly better system than the grapheme-based lexicon, thus showing that the proposed approach of ASWU-based lexicon development generalizes to multilingual resource sharing scenarios.

Table 11: Performance of KL-HMM based ASR systems exploiting auxiliary resources from resource-rich languages in terms of WRR. In these systems, MLP-*MULTI*-117 is used as the acoustic model.

Lexicon	CI	CD
Lex-*SG*-Gr-32	36.7	69.1
Lex-*SG*-Det-ASWU-85	52.1	70.7
Lex-*SG*-Prob-ASWU-82	57.7	**72.6**

6. Analysis

The ASR studies validated the proposed ASWU based lexicon from a speech technology perspective. As explained in Section 3.1, one of our hypotheses in this article is that the ASWUs obtained from the clustered CD grapheme units are "phone-like". This section focuses on that aspect through an analysis of the derived ASWUs (Section 6.1) and the generated pronunciations (Section 6.2). It is worth mentioning that a fully fledged quantitative analysis and concretely linking the derived ASWUs and the lexicons to existing linguistic knowledge would need a separate investigation, and is thus out of the scope of the article. In this section, our main goal is to provide a qualitative analysis and demonstrate how links to existing linguistic knowledge can be established to gain a better understanding. We notate phones as / / and graphemes as [ ]. Furthermore, we notate the derived ASWUs with the notation used by HTK to represent clustered CD units. For example, ASWU [ST_A_26] means a clustered CD unit with the center grapheme [A] (root node in the decision tree). For brevity, the analysis focuses only on the WSJ0 English corpus.

6.1. Relating the derived ASWUs to phonetic units

In order to analyze the relationship between the derived ASWUs and phonetic identities, we computed the KL-divergence between the Gaussian distribution modeling a mono-phone unit and the Gaussian distribution modeling an ASWU in the HMM/GMM setup on the WSJ0 corpus.[13] We computed the KL-divergence between single Gaussians, as this is the step at which the ASWUs are derived by clustering context-dependent graphemes. The KL-divergence between the Gaussian $\mathcal{N}_0(\mu_0, \Sigma_0)$ modeling a mono-phone unit as the reference distribution and the Gaussian $\mathcal{N}_1(\mu_1, \Sigma_1)$ modeling an ASWU as the measured

[13] In both cases, single-state models are used.

distribution is computed as (Duchi, 2007):

$$0.5\{\mathrm{Tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T\Sigma_1^{-1}(\mu_1 - \mu_0) - K - \ln\frac{|\Sigma_0|}{|\Sigma_1|}\},$$

where μ, Σ and K are the mean vector, the covariance matrix and dimension of the vector space respectively.

Table 12 provides a few ASWUs along with the five most related phones according to the KL-divergence matrix. For each grapheme, we have presented the ASWU that is most frequently used in the generated lexicon (they are marked in the table with a *). In addition to that, for each grapheme we have presented some of the other ASWUs that map to different sounds than the most frequently used ASWU.[14] Furthermore, the table also provides example English words that contain the ASWUs within their pronunciations. The example pronunciations were randomly selected from the lexicon. In each example, the grapheme that has been mapped to the ASWU in the pronunciation is highlighted.

It can be observed from the table that a consistent relationship between the ASWUs and phones exists. This relationship can be clearly observed in the case of consonant graphemes (such as [L], [M], [N] and [R]). For example, the ASWU belonging to grapheme [L] ([ST_L_24]) is more related to /l/ and /el/ sounds and the ASWU belonging to grapheme [R] ([ST_R_24]) is more related to /r/, /er/, and /axr/ sounds. These observations here are also consistent with the empirical observations made in an earlier grapheme-based ASR study on English (Rasipuram and Magimai.-Doss, 2013), where the grapheme-to-phoneme relationship is also learned through acoustics.

6.2. Generated pronunciations

This section provides a brief analysis on the generated pronunciations through deterministic and probabilistic G2ASWU modeling for the English WSJ0 corpus. Table 13 presents a few words selected from ASWU-based lexicons generated for the WSJ0 corpus. As one important aspect when generating pronunciations is generalization of the approach for the unseen contexts, we have provided examples from both the words that are seen during training, and the words that are not seen during training. We have highlighted the words that are unseen during training with underline. For each word, the first pronunciation is based on deterministic G2ASWU conversion and the second pronunciation

[14] Note that some of the ASWUs do not map to different sounds than the most frequently used ASWUs. They are only presented in the table as they are used in the generated pronunciations explained later in Section 6.2

Table 12: Relation between example automatically derived subword units on the WSJ0 corpus and phone units based on the KL-divergence matrix. The five most related phones are shown in the left to right order. The example pronunciations are obtained from *Lex-WSJ-Det-ASWU-90*. For each grapheme, the ASWU that is most frequently used in the generated lexicon is marked with a *.

ASWU	mapped phone	example word	ASWU	mapped phone	example word
[ST_A_28]*	/ae/,/ey/,/eh/,/ay/,/aw/	ATTACKED	[ST_N_24]*	/n/,/en/,/ng/,/m/,/em/	INTERMEDIATE
[ST_A_23]	/er/,/ey/,/r/,/ae/,/aw/	EARNED	[ST_N_23]	/n/,/en/,/ng/,/m/,/em/	BILLION
[ST_B_21]*	/b/,/d/,/v/,/dh/,/p/	BOOM	[ST_N_21]	/em/,/en/,/ng/,/n/,/m/	BLACKBURN
[ST_C_23]*	/k/,/t/,/p/,/d/,/th/	CREATING	[ST_O_27]*	/ah/,/ow/,/aa/,/l/,/ao/	DEPOSITS
[ST_C_21]	/s/,/z,/sh/,/f/,/zh/	CERTIFICATES	[ST_O_26]	/ax/,/uw/,/ah/,/uh/,/ih/	FOUNDATION
[ST_C _22]	/k/, /p/,/t/,/dh/,/th/	CONFRONTATION	[ST_O_29]	/aa/,/ah/,/aw/,/l/,/ow/	CONSEQUTIVE
[ST_D_23]*	/d/,/dx/,/b/,/g/,/dh/	LONGITUDINAL	[ST_P_21]*	/p/,/th/,/t/,/dh/,/k/	EXAMPLE
[ST_D_21]	/d/,/p/,/th/,/t/,/k/	BOND	[ST_Q_21]*	/k/,/p/,/th/,/dh/,/t/	CONSEQUENCES
[ST_E_29]*	/ih/,/ax/,/uh/,/uw/,/eh/	EXPENSIVE	[ST_R_24]*	/r/,/er/,/axr/,/uh/,/ay/	CONTRACTING
[ST_E_21]	/f/,/hh/,/th/,/em/,/p/	OTHERWISE	[ST_S_21]*	/s/,/f/,/z/,/th/,/hh/	DIRECTORS
[ST_E_26]	/axr/,/uw/,/uh/,/r/,/ih/	DRIVER	[ST_S_22]	/z/,/s/,/sh/,/zh/,/f/	PARTNERSHIPS
[ST_E_27]	/eh/,/ae/,/ih/,/ax/,/ay/	GENERATION	[ST_S_24]	/s/,/f/,/th/,/z/,/dh/	SKOLNIKS
[ST_F_22]*	/f/,/th/,/p/,/s/,/t/	FALLING	[ST_S_25]	/s/,/z/,/f/,/th/,/sh/	INCREASED
[ST_G_21]*	/g/,/dx/,/d/,/t/,/jh/	GOVERMENTS	[ST_T_25]*	/t/,/k/,/p/,/th/,/dh/	OMITTED
[ST_H_22]*	/sh/,/ch/,/zh/,/f/,/jh/	CHURN	[ST_T_24]	/p/,/th/,/f/,/dh/,/t/	BET
[ST_H_23]	/hh/,/th/,/f/,/p/,/en/	OUTRIGHT	[ST_U_24]*	/ax/,/uh/,/ah/,/ih/,/oy/	EQUALLY
[ST_I_27]*	/ih/,/eh/,/ax/,/uh/,/ah/	LOGIC	[ST_U_23]	/uw/,/ao/,/oy/,/axr/,/r/	NURSING
[ST_I_25]	/ih/,/uw/,/ax/,/iy/,/ey/	DISTILLERS	[ST_V_21]*	/v/,/b/,/d/,/dh/,/g/	COVERAGE
[ST_J_21]*	/jh/,/ch/,/t/,/dx/,/d/	JESSE	[ST_W_21]*	/w/,/l/,/oy/,/el/,/g/	DOWNGRADED
[ST_K_21]*	/k/,/t/,/p/,/d/,/dh/	BOOKS	[ST_X_21]*	/t/,/th/,/z/,/k/,/p/	EX
[ST_L_24]*	/l/,/el/,/ow/,/ao/,/aa/	EMPLOYS	[ST_Y_21]*	/iy/,/ng/,/y/,/ey/,/en/	COUNTRY
[ST_M_24]*	/m/,/n/,/em/,/ng/,/en/	GRUBMAN	[ST_Z_21]*	/z/,/s/,/th/,/f/,/jh/	FREEZES

is based on probabilistic G2ASWU conversion. With the information provided in Table 12, it can be observed that G2ASWU conversion approach is able to recognize different sounds of the same grapheme to provide a pronunciation similar to what is seen in a phone-based lexicon for both seen and unseen words during training. For example, in case of deterministic G2ASWU conversion, for the word $CENT$, the grapheme [C] is mapped to [ST_C_21], which in the earlier analysis was found to map to phone /s/, whilst for the word $CURB$ the grapheme [C] is mapped to [ST_C_23], which was found to be more related to /k/. The distinction between deterministic and probabilistic G2ASWU conversion can be very well observed through words $PHONE$ and $UPHELD$. In the case of the word $PHONE$, the deterministic G2ASWU conversion maps each grapheme to an ASWU unit while probabilistic G2ASWU conversion is able to map a group of graphemes to an ASWU, i.e., PH to /f/ and NE to /n/. In the case of the word $UPHELD$, it can be observed that probabilistic G2ASWU conversion leads to deletion of an unit while deterministic G2ASWU preserves the unit. We speculate that the inferior performance of probabilistic G2ASWU

conversion in the ASR studies on English is mainly due to such deletions.

Table 13: Few example words together with their generated pronunciations based on a deterministic or a probabilistic lexical modeling based G2ASWU conversion on the WSJ0 corpus.

Word	Lex-*WSJ*-Det-ASWU-90 Lex-*WSJ*-Prob-ASWU-88
PHONE	[ST_P_21] [ST_H_23] [ST_O_29] [ST_N_24] [ST_E_21] [ST_F_22] [ST_O_29] [ST_N_21]
UPHELD	[ST_U_24] [ST_P_21] [ST_H_23] [ST_E_29] [ST_L_24] [ST_D_21] [ST_O_27] [ST_P_21] [ST_H_23] [ST_L_24] [ST_D_21]
CENT	[ST_C_21] [ST_E_27] [ST_N_24] [ST_T_24] [ST_S_25] [ST_E_27] [ST_N_24] [ST_T_24]
CURB	[ST_C_23] [ST_U_23] [ST_R_24] [ST_B_21] [ST_C_22] [ST_U_23] [ST_R_24] [ST_B_21]
VERSIONS	[ST_V_21] [ST_E_26] [ST_R_25] [ST_S_22] [ST_I_25] [ST_O_26] [ST_N_23] [ST_S_21] [ST_V_21] [ST_E_26] [ST_R_25] [ST_S_22] [ST_T_22] [ST_O_26] [ST_N_23] [ST_S_21]
SLID	[ST_S_24] [ST_L_24] [ST_I_27] [ST_D_21] [ST_S_24] [ST_L_24] [ST_I_27] [ST_D_21]

It is worth mentioning that we have done the same kind of analysis for the RM corpus and we have observed similar trends. In the case of Scottish Gaelic, there is no well-developed phonetic lexicon available. Nevertheless, we have analyzed the ASWUs by building on the idea that speech sound units are shared across languages as the human speech production mechanism is common across languages. More precisely, by using the multilingual KL-HMM framework explained in Section 5.4 to capture the relationship between ASWUs and multilingual phones, we have tried to interpret the ASWUs in terms of meaningful linguistic units. The findings of this analysis and the analysis on the RM corpus can be found in (Razavi, 2017, Ch. 6).

7. Conclusions

This article presented a novel approach for subword unit derivation and pronunciation generation using only word level transcribed speech data. In this approach, the subword units are first derived by clustering context-dependent graphemes in an HMM-based ASR framework using maximum likelihood criteria; followed by modeling of the relationship between the graphemes and the derived units in a deterministic or probabilistic manner using acoustic data; and finally inferring pronunciations given the learned relationships and the word orthographies using an ergodic HMM. In comparison to existing approaches in the literature, a distinguishing aspect of the proposed approach is that it fits

within the well-known HMM framework for ASR and speech synthesis, and is therefore fairly straight-forward to implement given the available toolkits such as HTK (Young et al., 2000) and KALDI (Povey et al., 2011). The proposed approach assumes that a correspondence between the grapheme sequence in the written form of word and the phoneme sequence in the spoken form of the word exists. For logographic languages, where the graphemes represent morphemes or words, the approach could potentially be combined with transliteration.

Our experimental studies on two languages showed that the ASWU-based lexicon can be developed in a fully data-driven manner, i.e., the set of ASWUs and the corresponding lexicon can be selected through cross validation. The ASR studies on both the languages showed that the ASWU-based lexicons consistently yield significantly better ASR systems compared to the grapheme-based lexicons. For G2ASWU conversion, we investigated two approaches, namely, decision-tree based approach and KL-HMM based acoustic G2P approach. Our experimental studies also showed that both G2ASWU approaches are equally applicable, with the acoustic G2P approach holding advantage for languages with many-to-one G2P relationship. Also, in one of the first efforts, we showed that the discovered ASWUs and the learned G2ASWU relationship can be transferred across domains in a language and the G2ASWU conversion mechanism inherently enables such transfer. Furthermore, the analysis of the learned models and the generated pronunciations showed that the derived ASWUs to a good extent are systematically related to phonetic identities. In particular, studies on Scottish Gaelic showed that the multilingual resources not only help in building better ASWU-based ASR systems, but also enable discovery of the phonetic identities of the derived ASWUs (Razavi et al., 2015; Razavi, 2017, Ch. 6). This opens potential venues for further research and development to improve phonetic and lexical resources and technologies for under-resourced languages through transfer of linguistic knowledge and data across languages.

In the proposed approach the problem of ASWU derivation was as posed as a problem of finding a latent symbol space that can be related to acoustic data and associated transcriptions (or graphemes). In this work, we used standard cepstral features that tend to carry information related to phones to find the latent symbol space. However, there are alternative features or representations that carry phone related information and could be exploited to find phone-like latent symbol space. For instance using linguistically motivated articulatory features (AFs) (Jakobson et al., 1992; Ladefoged, 1993), which may be more robust representation when compared to spectral-based features and could help in

reducing the gap between ASWU-based approach and phoneme-based approach. This could be achieved without deviating from the HMM framework through the recently proposed AF-based ASR framework using KL-HMMs (Rasipuram and Magimai.-Doss, 2016), where it has been show that ASR systems can be developed by learning grapheme-to-AF relationship through acoustics. Alternately, we could cast the ASWU based lexicon development as a three step process, where first acoustic-to-AF relationship is learned on available multilingual resources; next grapheme-to-AF relationship is learned from the target language transcribed speech and clustered to derive ASWUs using KL-HMMs; and finally G2ASWU conversion is performed, as done in the present article. Our future work will focus toward this direction on both well-resourced and under-resourced languages along with development of methods to select multiple pronunciation variants.

Finally, it is worth mentioning that our focus in this article was on addressing the lack of phonetic resources in an under-resourced language through derivation of ASWUs and associated pronunciations in a data-driven fashion. Recently, end-to-end approaches have been proposed for ASR, which use a neural network to directly predict the characters given the utterance (Hannun et al., 2014; Graves and Jaitly, 2014; Hwang and Sung, 2016). These approaches do not require a phonetic lexicon for speech recognition, however, they are data-hungry and therefore may not suit well for under-resourced scenarios. On the other side, our multlingual studies on the Scottish Gaelic corpus have shown that by using the same acoustic model and by only modifying the lexical entities (ASWUs versus graphemes), the performance of ASR systems can be significantly improved. This implies that the ASWUs can provide better representations of words than graphemes, and introduces a new question: How would the approach used for end-to-end speech recognition perform when using ASWUs instead of graphemes? From all these perspectives, the utility of end-to-end ASR systems for under-resourced languages remains an open question and needs a separate investigation.

Appendix A. KL-HMM

This appendix explains the KL-HMM training and decoding procedure (Aradilla et al., 2008).

Appendix A.1. KL-HMM training

Given a training set of N utterances $\{Z(n), W(n)\}_{n=1}^{N}$, where for each training utterance n, $Z(n)$ represents a sequence of acoustic unit probability vectors $Z(n) = [\mathbf{z}_1(n), \cdots, \mathbf{z}_t(n), \cdots, \mathbf{z}_{T(n)}(n)]$ of length $T(n)$ and $W(n)$ represents the sequence of underlying words, the KL-HMM parameters are estimated by a Viterbi expectation-maximization procedure that minimizes the cost function,

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

where $Q = [q_1, \cdots, q_t, \cdots, q_{T(n)}]$ denotes a sequence of HMM states, $q_t \in \{1, \cdots, I\}$, $\mathcal{Q}$ denotes the set of all possible HMM state sequences, and $a_{q_{t-1}q_t}$ corresponds to transition probabilities.

In practice, the transition probabilities $a_{q_{t-1}q_t}$ are assumed to be constant (0.5), similar to the hybrid HMM/ANN approach. Therefore parameter estimation amounts to estimating $\{\mathbf{y}_i\}_{i=1}^{I}$. Given a uniformly initialized set of parameters $\{\mathbf{y}_i\}_{i=1}^{I}$ (i.e., $y_i^d = \frac{1}{D} \quad \forall i, D$) the segmentation step yields an optimal state sequence for each training utterance using Viterbi algorithm. Given the optimal state sequences, i.e., alignment and $\mathbf{z}_t$ belonging to each of these states, the optimization step then estimates a new set of model parameters by minimizing the cost function based on KL-divergence (Eqn. (A.1)) with the constraint that $\sum_{d=1}^{D} y_i^d = 1$. This process of segmentation and the optimization is iteratively done until convergence.

With S_{RKL} as the local score, the optimal state distribution is the arithmetic mean of the training acoustic state probability vectors assigned to the state, i.e.,

$$y_i^d = \frac{1}{M(i)} \sum_{\mathbf{z}_t(n) \in Z(i)} z_t^d(n) \quad \forall n, t \tag{A.2}$$

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

With S_{KL} as the local score, the optimal state distribution is the normalized geometric mean of the training acoustic state probability vectors assigned to the state, i.e.,

$$y_i^d = \frac{\hat{y}_i^d}{\sum_{d=1}^{D} \hat{y}_i^d} \quad where \quad \hat{y}_i^d = \Big(\prod_{\mathbf{z}_t(n) \in Z(i)} z_t^d(n) \Big)^{\frac{1}{M(i)}} \quad \forall n, t \tag{A.3}$$

where $\hat{y}_i^d$ represents the geometric mean of state i for dimension d, $Z(i)$ denotes the set of acoustic state probability vectors assigned to state i and $M(i)$ is the

cardinality of $Z(i)$.

With S_{SKL} as the local score, there is no closed form solution to find the optimal lexical state distribution. The optimal lexical state distribution can be computed iteratively using the arithmetic and the normalized geometric mean of the acoustic state probability vectors assigned to the state (Veldhuis, 2002).

Appendix A.2. KL-HMM decoding

Given the sequence of acoustic unit posterior probability vectors $Z = [\mathbf{z}_1, \cdots, \mathbf{z}_t, \cdots, \mathbf{z}_T]$ and the KL-HMM parameters, the best matching word sequence is obtained by minimizing the cost function,

$$W^* = \arg\min_{Q} \sum_{t=1}^{T} \left\{ S(\mathbf{y}_{q_t}, \mathbf{z}_t) - \log a_{q_{t-1}q_t} \right\} \tag{A.4}$$

where $Q = [q_1, \cdots, q_T]$ denotes a sequence of HMM states. It can be observed that Eqn. (A.4) is similar to Eqn. (5), except that maximizing the log-likelihood $p(\mathbf{x}_t|q_t = l^i)$ is replaced with minimizing a KL-divergence based score $S(\mathbf{y}_{q_t}, \mathbf{z}_t)$.

Acknowledgment

This work was supported by the Hasler foundation through the grant Flexible acoustic data driven grapheme to acoustic unit conversion (AddG2SU). All the research was conducted at the Idiap Research Institute. The authors would like to thank the reviewers for their valuable comments.

References

V. Pagel, K. Lenzo, A. Black, Letter to sound rules for accented lexicon compression, in: Proceedings of ICSLP, vol. 5, 2015–2020, 1998.

T. Sejnowski, C. Rosenberg, Parallel networks that learn to pronounce English text, Complex systems 1 (1) (1987) 145–168.

P. Taylor, Hidden Markov models for grapheme to phoneme conversion., in: Proceedings of Interspeech, 1973–1976, 2005.

M. Bisani, H. Ney, Joint-sequence Models for Grapheme-to-phoneme Conversion, Speech Communication 50 (5) (2008) 434–451.

S. Kanthak, H. Ney, Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition., in: Proceedings of ICASSP, 845–848, 2002a.

M. Killer, S. Stüker, T. Schultz, Grapheme based speech recognition, in: Proceedings of Eurospeech, 3141–3144, 2003.

J. Dines, M. Magimai.-Doss, A study of phoneme and grapheme based context-dependent ASR systems, in: Machine Learning for Multimodal Interaction, Springer, 215–226, 2007.

M. Magimai-Doss, R. Rasipuram, G. Aradilla, H. Bourlard, Grapheme-based Automatic Speech Recognition using KL-HMM, in: Proceedings of Interspeech, 2011.

T. Ko, B. Mak, Eigentrigraphemes for under-resourced languages, Speech Communication 56 (2014) 132–141.

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

M. Gales, K. Knill, A. Ragni, Unicode-based graphemic systems for limited resource languages, in: Proceedings of ICASSP, 5186–5190, 2015.

C.-H. Lee, F. K. Soong, B.-H. Juang, A segment model based approach to speech recognition, in: Proceedings of ICASSP, 1988.

T. Svendsen, K. Paliwal, E. Harborg, P. Husoy, An improved sub-word based speech recognizer, in: Proceedings of ICASSP, 108–111, 1989.

K. Paliwal, Lexicon-building methods for an acoustic sub-word based speech recognizer, in: Proceedings of ICASSP, 729–732, 1990.

M. Bacchiani, M. Ostendorf, Using automatically-derived acoustic sub-word units in large vocabulary speech recognition, in: International Conference on Spoken Language Processing, 1998.

T. Holter, T. Svendsen, Combined optimisation of baseforms and model parameters in speech recognition based on acoustic subword units, in: Proceedings of ASRU, 199–206, 1997.

R. Singh, B. Raj, R. Stern, Automatic generation of phone sets and lexical transcriptions, in: Proceedings of ICASSP, 1691–1694, 2000.

C. Lee, Y. Zhang, J. R. Glass, Joint Learning of Phonetic Units and Word Pronunciations for ASR., in: Proceedings of EMNLP, 182–192, 2013.

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

M. Razavi, M. Magimai-Doss, An HMM-based formalism for automatic subword unit derivation and pronunciation generation, in: Proceedings of ICASSP, 2015.

M. Razavi, R. Rasipuram, M. Magimai.-Doss, Pronunciation Lexicon Development for Under-Resourced Languages Using Automatically Derived Subword Units: A Case Study on Scottish Gaelic, in: 4th Biennial Workshop on Less-Resourced Languages, 2015.

G. Aradilla, H. Bourlard, M. Magimai-Doss, Using KL-based acoustic models in a large vocabulary recognition task., in: Proceedings of Interspeech, 928–931, 2008.

X. Luo, F. Jelinek, Probabilistic Classification of HMM States for Large Vocabulary Continuous Speech Recognition, in: Proceedings of ICASSP, 353–356, 1999.

J. Rottland, G. Rigoll, Tied posteriors: an approach for effective introduction of context dependency in hybrid NN/HMM LVCSR, in: Proceedings of ICASSP, 1241–1244, 2000.

S. Kanthak, H. Ney, Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition., in: Proceedings of ICASSP, 845–848, 2002b.

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

R. Rasipuram, M. Razavi, M. Magimai.-Doss, Probabilistic Lexical Modeling and Unsupervised Training for Zero-Resourced ASR, in: Proceedings of ASRU, 2013a.

R. Rasipuram, M. Magimai-Doss, Acoustic Data-driven Grapheme-to-Phoneme Conversion using KL-HMM, in: Proceedings of ICASSP, 2012.

M. Razavi, R. Rasipuram, M. Magimai.-Doss, Acoustic data-driven grapheme-to-phoneme conversion in the probabilistic lexical modeling framework, Speech Communication 80 (2016) 1–21.

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

C.-T. Chung, C.-A. Chan, L.-S. Lee, Unsupervised discovery of linguistic structure including two-level acoustic patterns using three cascaded stages of iterative optimization, in: Proceedings of ICASSP, 8081–8085, 2013.

C.-y. Lee, T. J. O'Donnell, J. Glass, Unsupervised lexicon discovery from acoustic input, Transactions of the Association for Computational Linguistics 3 (2015) 389–403.

M. Bacchiani, M. Ostendorf, Joint lexicon, acoustic unit inventory and model design, Speech Communication 29 (2) (1999) 99–114.

R. Singh, B. Raj, R. M. Stern, Automatic generation of subword units for speech recognition systems, IEEE Transactions on Speech and Audio Processing 10 (2) (2002) 89–99.

A. Y. Ng, M. I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: Advances in Neural Information Processing Systems, 849–856, 2001.

S. Young, The general use of tying in phoneme-based HMM speech recognisers, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 01, 569–572, 1992.

A. Ljolje, High accuracy phone recognition using context clustering and quasi-triphonic models, Computer Speech & Language 8 (2) (1994) 129–151.

R. Rasipuram, M. Magimai-Doss, Improving Grapheme-based ASR by Probabilistic Lexical Modeling Approach, in: Proceedings of Interspeech, 2013.

D. B. Paul, J. M. Baker, The Design for the Wall Street Journal-based CSR Corpus, in: Proceedings of the Workshop on Speech and Natural Language, 357–362, 1992.

P. C. Woodland, J. J. Odell, V. Valtchev, S. J. Young, Large Vocabulary Continuous Speech Recognition Using HTK, in: Proceedings ICASSP, 125–128, 1994.

J. Garofolo, D. Graff, D. Paul, D. Pallett, CSR-I (WSJ0) Complete LDC93S6A, Web Download. Philadelphia: Linguistic Data Consortium .

P. Price, W. M. Fisher, J. Bernstein, D. S. Pallett, The DARPA 1000-word Resource Management Database for Continuous Speech Recognition, in: Proceedings of ICASSP, IEEE, 651–654, 1988.

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

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

R. Rasipuram, M. Magimai.-Doss, Probabilistic Lexical Modeling and Grapheme-based Automatic Speech Recognition, Idiap-RR Idiap-RR-15-2013, 2013.

D. Imseng, et al., Comparing different acoustic modeling techniques for multilingual boosting, in: Proceedings of Interspeech, 2012.

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

M. Wolters, A Diphone-Based Text-to-Speech System for Scottish Gaelic, Master's thesis, University of Bonn, 1997.

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

R. Rasipuram, M. Razavi, M. Magimai.-Doss, Integrated Pronunciation Learning for Automatic Speech Recognition Using Probabilistic Lexical Modeling, in: Proceedings of ICASSP, 5176–5180, 2015.

D. Imseng, R. Rasipuram, M. Magimai.-Doss, Fast and flexible Kullback-Leibler divergence based acoustic modeling for non-native speech recognition, in: Proceedings of ASRU, 348–353, 2011.

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

J. Duchi, Derivations for Linear Algebra and Optimization, http://www.cs.berkeley.edu/∼jduchi/projects/general_notes.pdf, 2007.

M. Razavi, On Modeling the Synergy Between Acoustic and Lexical Information for Pronunciation Lexicon Development, Ph.D. thesis, École polytechnique fédérale de Lausanne (EPFL), 2017.

D. Povey, et al., The Kaldi Speech Recognition Toolkit, in: Proceedings of ASRU, 2011.

R. Jakobson, G. Fant, M. Halle, Preliminaries to Speech Analysis: the Distinctive Features and their Correlates, MIT Press, 1992.

P. Ladefoged, A Course in Phonetics, Harcourt Brace College Publishers, 1993.

R. Rasipuram, M. Magimai.-Doss, Articulatory Feature Based Continuous Speech Recognition using Probabilistic Lexical Modeling, Computer Speech and Language 36 (2016) 233–259.

A. Y. Hannun, A. L. Maas, D. Jurafsky, A. Y. Ng, First-pass large vocabulary continuous speech recognition using bi-directional recurrent DNNs, arXiv preprint arXiv:1408.2873 .

A. Graves, N. Jaitly, Towards End-To-End Speech Recognition with Recurrent Neural Networks., in: Proceedings of International Conference on Machine Learning (ICML), vol. 14, 1764–1772, 2014.

K. Hwang, W. Sung, Sequence to Sequence Training of CTC-RNNs with Partial Windowing, in: Proceedings of ICML, 2178–2187, 2016.

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