INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Improving Under-Resourced Language ASR Through Latent Subword Unit Space Discovery

Marzieh Razavi[1,2] *and Mathew Magimai.-Doss*[1]

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

Abstract

Development of state-of-the-art automatic speech recognition (ASR) systems requires acoustic resources (i.e., transcribed speech) as well as lexical resources (i.e., phonetic lexicons). It has been shown that acoustic and lexical resource constraints can be overcome by first training an acoustic model that captures acoustic-to-multilingual phone relationships on language-independent data; and then training a lexical model that captures grapheme-to-multilingual phone relationships on the target language data. In this paper, we show that such an approach can be employed to discover a latent space of subword units for under-resourced languages, and subsequently improve the performance of the ASR system through both acoustic and lexical model adaptation. Specifically, we present two approaches to discover the latent space: (1) inference of a subset of the multilingual phone set based on the learned grapheme-to-multilingual phone relationships, and (2) derivation of automatic subword unit space based on clustering of the grapheme-to-multilingual phone relationships. Experimental studies on Scottish Gaelic, a truly under-resourced language, show that both approaches lead to significant performance improvements, with the latter approach yielding the best system.

Index Terms: under-resourced ASR, acoustic model adaptation, subword unit set discovery, multilingual ANN

1. Introduction

A crucial step towards development of hidden Markov model (HMM)-based automatic speech recognition (ASR) systems is to model the relationship between the acoustic features $\{\mathbf{x}_t\}_{t=1}^{T}$ and the lexical subword units $\{l^i\}_{i=1}^{I}$. It has been elucidated in [1] that, when estimating emission likelihood score $p(\mathbf{x}_t|q_t = l^i)$ at each HMM state q_t, such a relationship can be factored through latent variables $\{a^d\}_{d=1}^{D}$ referred to here as *acoustic units* into two models, namely the acoustic model (which captures the relationship between acoustic features and acoustic units) and the lexical model (which captures the lexical unit-to-acoustic unit relationships):

$$p(\mathbf{x}_t|q_t = l^i) = \sum_{d=1}^{D} \underbrace{p(\mathbf{x}_t|a^d)}_{\text{acoustic model}} \underbrace{P(a^d|q_t = l^i)}_{\text{lexical model}} = \mathbf{v}_t^{\mathrm{T}} \cdot \mathbf{y}_i, \quad (1)$$

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

In standard HMM-based ASR approaches, the lexical units are typically context-dependent (CD) subword units. The acoustic units are the clustered CD subword unit states obtained through decision tree clustering. $\mathbf{v}_t$ is estimated by Gaussian mixture models (GMMs) [2] or artificial neural networks (ANNs) [3]. $\mathbf{y}_i$ is either deterministic, i.e., a Kronecker delta distribution, based on the decision tree learned during state tying [4] or probabilistic [5, 6, 1]. In the latter case, it is referred to as probabilistic lexical model.

Eqn. (1) can be regarded as a match between acoustic information and lexical information in the latent variable space, i.e., acoustic unit space. In recent years, a probabilistic lexical modeling approach in the framework of Kullback-Leibler divergence-based HMM (KL-HMM) [7, 8] has emerged. In this approach, the match between acoustic information and lexical information is obtained by matching posterior probability estimates of acoustic units $\mathbf{z}_t = [z_t^1, \ldots, z_t^d, \ldots, z_t^D]^{\mathrm{T}}$, $z_t^d = P(a^d|\mathbf{x}_t)$ (instead of $\mathbf{v}_t$) with $\mathbf{y}_i$ based on KL-divergence,

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

In this case, $\mathbf{y}_i$ is trained using a cost function based on KL-divergence, and $\mathbf{z}_t$ is estimated either using ANNs or GMMs [9].

As most of the ASR systems use phones as lexical subword units, in order to learn the acoustic model and the lexical model, two well developed resources are needed: acoustic resources (transcribed speech) and lexical resources (phonetic lexicons). Unfortunately, for many languages it can be difficult to obtain either acoustic or lexical resources or both. In the literature, the acoustic resource constraint has been approached through use of cross-lingual or multilingual resources [10, 11, 12, 13]. Typically, in these approaches a common phone set or acoustic unit space is defined based on cross-lingual or multilingual resources and a language-independent acoustic model is trained. The multilingual acoustic model is then adapted on target language (TL) data based on a deterministic lexical model learned on TL data. The adaptation process can also involve redefinition of acoustic unit space based on TL data [14, 15, 16]. In the absence of lexical resources in the literature, typically graphemes are used as subword units [17, 18, 19].

The acoustic and lexical resources constraint scenario, however, has been seldom addressed in the literature [20, 21]. The present paper focuses on this scenario, in particular on truly under-resourced languages which have limited acoustic resources and no phonetic lexicon. In that respect, probabilistic lexical modeling has emerged as a promising method. Specifically, the two constraints can be addressed by first training an acoustic model that learns the relationship between acoustic features and multilingual phones on language-independent data, and then training a lexical model that learns a probabilis-

This work was supported by Hasler Foundation through the grant AddG2SU: Flexible acoustic data driven grapheme to acoustic unit conversion.

http://dx.doi.org/10.21437/Interspeech.2016-1010

tic relationship between TL graphemes and multilingual phones on target language acoustic data [1]. A potential advantage of this approach is that the probabilistic relationship between TL graphemes and multilingual phones can be learned on a very limited acoustic resource to yield significantly better systems when compared to conventional grapheme-based ASR approaches. However, it has also been seen that such gains diminish as TL acoustic resources increase, the main reason being that the TL data is leveraged to only learn the lexical model.

This paper builds on the above described probabilistic lexical modeling framework (to address both acoustic and lexical resource constraints) to develop an approach where the TL data is leveraged for both acoustic modeling and lexical modeling. Specifically, in this approach, first a language-dependent acoustic unit space is defined based on the learned grapheme-to-multilingual phone relationships. The acoustic model is then adapted to classify the language-dependent acoustic units on TL data. Finally, a lexical model that captures the probabilistic relationship between the TL graphemes and language-dependent acoustic units is trained. Through an investigation on Scottish-Gaelic, a truly under-resourced minority language, we show that the proposed approach not only helps in significantly improving the performance of the ASR system (relative improvement of up to 16%) but also enables discovery of a language-specific phone set, which could be potentially exploited to develop lexical resources for the target under-resourced language.

The remainder of this paper is organized as follows. Section 2 presents the proposed approach along with a brief background on KL-HMM. Section 3 explains the experimental setup. Section 4 presents the experimental results and analysis. Finally, Section 5 concludes the paper.

2. Proposed approach

This section first provides a brief background on how acoustic and lexical resource constraints can be addressed using the KL-HMM framework by simply learning a lexical model that captures TL grapheme-to-multilingual phone relationships. It then presents the proposed approach to adapt both acoustic model and lexical model in this framework by discovering a language-dependent acoustic unit space.

2.1. Grapheme-based KL-HMM for low-resourced ASR

As briefly explained in Section 1 and illustrated in Figure 1, in under-resourced scenarios [1]:

1. A multilingual ANN is trained on acoustic and lexical data from resource-rich languages which is used to estimate multilingual phone posterior features $\{\mathbf{z}_t\}_{t=1}^{T}$.
2. Given the $\{\mathbf{z}_t\}_{t=1}^{T}$ estimator, a KL-HMM is trained on TL data in which each state representing a CD grapheme is parameterized by categorical distribution $\mathbf{y}_i$ of multilingual phones. The parameters $\{\mathbf{y}_i\}_{i=1}^{I}$ are estimated through Viterbi Expectation-Maximization by minimizing a cost function based on KL-divergence local score (Eqn. (2)). The unseen grapheme contexts are handled by tying KL-HMM states [22].

During recognition, the most probable sequence of words is inferred by using a standard Viterbi decoder where the state log-likelihood is replaced by KL-divergence.

2.2. Adaptation based on language-dependent acoustic unit discovery

We present two approaches to discover language-dependent acoustic units given the multilingual ANN and the grapheme-to-multilingual phone relationships learned by KL-HMM.

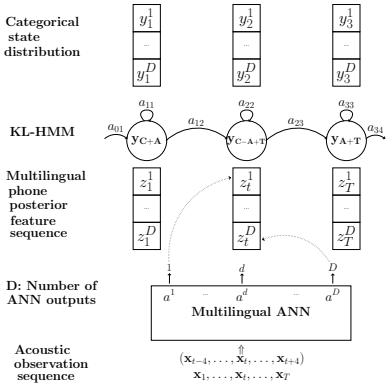

Figure 1: *Illustration of grapheme-based KL-HMM framework.*

2.2.1. Approach-I: Acoustic units as a subset of multilingual phone set

In this approach, as illustrated in Figure 2, given the multilingual phone posterior probabilities $\{\mathbf{z}_t\}_{t=1}^{T}$ estimated on TL data and the categorical distributions $\{\mathbf{y}_i\}_{i=1}^{I}$ from the trained CD grapheme KL-HMM, an optimal KL-HMM state sequence for the TL training utterance is obtained using the Viterbi algorithm (Figure 2-**Part I**). Then, for the aligned state l^i in time frame t, the multilingual phone a^d with the highest probability in $\mathbf{y}_i = [P(a^1|l^i), \ldots, P(a^D|l^i)]$ is selected as the phone label at time t, i.e., u_t (Figure 2-**Part II**). In doing so, we not only get labels for ANN adaptation, but also a subset of the multilingual phone set corresponding to the language.

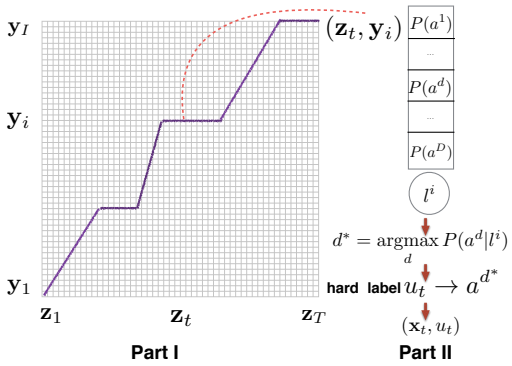

Figure 2: *Illustration of acoustic unit discovery from the multilingual phone set.*

2.2.2. Approach-II: Acoustic units as automatically derived subword units

Recently, it has been shown that phone-like automatic subword units can be derived from the clustered CD grapheme states in the HMM/GMM framework using standard cepstral features [23, 24]. In the present work, a similar approach is adopted where posterior features $\mathbf{z}_t$ instead of cepstral features are used. More precisely, the acoustic units are the clustered CD grapheme states of the KL-HMM explained in Section 2.1, which could again be expected to be phone-like. The labels for ANN adaption are obtained by aligning posterior features $\mathbf{z}_t$ with the KL-HMM after state tying similar to *Approach-I* (Figure 2-Part I) and using the decision tree to map the CD graphemes to clustered CD grapheme states.

Given the discovered language-dependent acoustic units and the corresponding alignment to the acoustic data, the ASR

system development finally involves: (a) adaptation of multilingual ANN by re-initializing the weights between last hidden layer and the output layer, which now models the discovered language-dependent acoustic units; and then (b) learning a new probabilistic lexical model that captures the relationship between TL graphemes and the discovered language-dependent acoustic units.

3. Experimental setup

We investigate the proposed approach on Scottish Gaelic, a genuinely under-resourced and minority language.

3.1. Scottish Gaelic

Scottish Gaelic is considered as an endangered language spoken by only 60,000 people which belongs to the class of Celtic languages. There are about 51 phonemes in the language [25]. However, the number of phonemes can vary depending on the dialect. The language lacks a proper phonetic lexicon and the available transcribed speech data are limited.

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

- Fortis (produced with greater energy) or lenis (produced with lesser energy): The lenited consonants appear in the orthography with a grapheme [H] next to them.
- 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).

3.2. Database

The Scottish Gaelic corpus was collected by the University of Edinburgh in 2010. It consists of recordings from broadcast news and discussion programs.[1] In this paper, the database is partitioned into training, cross-validation (CV) and test sets according to the structure provided in [26]. Table 1 provides an overview of the Scottish Gaelic corpus. The vocabulary size in the corpus was 5080. There are a total of 2246 unique words in the test set of which 772 are not seen during training.

Table 1: *Overview of the Scottish Gaelic corpus.*

Number of	Train	CV	Test
Utterances	2389	1112	1317
Hours	3	1	1
Speakers	22	12	12

The corpus does not provide any phonetic lexicon. The grapheme lexicon can be obtained from the orthography of the words. Additionally, prior knowledge about broad and slender consonants can be applied to the word orthographies. Along these lines we investigated using two different lexicons:

- *ortho*: In this scenario, the lexicon is obtained directly from the orthography of the words without incorporating any knowledge about the language. As the corpus also contains borrowed English words, the graphemes J, K, Q, V, W, X, Y and Z are also present in the lexicon. Therefore the lexicon consists of 32 graphemes including silence (sil).
- *ortho+know*: In this case, the lexicon is obtained by considering broad, slender and lenited consonants as separate graphemes. The lexicon contains 83 graphemes including sil.

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

As there is no language model provided in the corpus, we used an optimistic bigram language model trained on the sentences from the test set, similar to [26].

3.3. System setup

The setups for multilingual ANN training, KL-HMM training, acoustic and lexical model adaptation are as follows:

Multilingual ANN training: We trained a multilingual ANN, more precisely a 5-layer multilayer perceptron (MLP) using 63 hours of speech data from five languages in the SpeechDat(II) corpus, namely British English, Italian, Spanish, Swiss French and Swiss German. The input to the MLP was a 39 dimensional PLP cepstral feature with four frames preceding and four frames following context. Each hidden layer had 2000 units. The MLP output units were obtained by merging phones in the SAMPA format that share the same symbol across different languages to form a multi-lingual phone set of 117 units. The MLP was trained with output non-linearity of softmax and minimum cross-entropy error criterion using Quicknet software [27].

Approach-I: We trained single-preceding single-following context-dependent grapheme-based KL-HMMs on Scottish Gaelic data using the *ortho* lexicon. Each CD grapheme KL-HMM was modeled with three states. The acoustic unit space and the alignments were then obtained by aligning the trained KL-HMM with the posterior features as described in Section 2.2.1. This resulted in 71 phones, i.e., $\{a^d\}_{d=1}^{71}$.

Approach-II: We trained single-preceding single-following context-dependent grapheme-based KL-HMMs on Scottish Gaelic data using the *ortho* lexicon. Each CD grapheme KL-HMM was modeled with a single state and states were tied using a single question set following the previous work on automatic subword unit discovery [24]. In the KL-divergence based decision tree state tying method [22], similar to the log-likelihood based decision tree based approach in [4], stopping criterion based on minimum cluster occupancy and minimum decrease in the cost function threshold exists. We obtained different number of clustered CD grapheme states by adjusting the threshold based on decrease in the KL-divergence cost function during decision-tree based state tying and chose the one that yielded best system on the cross-validation data set. Specifically, this resulted in 306 clustered CD grapheme states, i.e., acoustic units $\{a^d\}_{d=1}^{306}$. The alignments for ANN adaptation were obtained as described in Section 2.2.2.

Acoustic model adaptation: For each of the approaches, given the acoustic units and the alignment to acoustic data, the multilingual ANN was adapted by re-initializing the weights between the last hidden layer and the output layer and the biases of the output layer. We investigated only last layer adaptation as well as adaptation of the whole ANN. For both approaches, we found adaptation of the whole ANN yields better frame level accuracy.

Lexical model re-training: For both approaches, given the posterior features from the adapted ANNs, single-preceding single-following CD grapheme-based KL-HMM systems were trained. Here we studied use of two different lexicons, namely *ortho* lexicon and *ortho+know* lexicon. Each CD grapheme was modeled by three states. To handle unseen grapheme contexts state tying was performed.

4. Results and analysis

This section first presents the ASR results. It then presents an analysis on the discovered units along with a comparative study w.r.t related approaches.

4.1. ASR Results

Table 2 presents the ASR performance in terms of word accuracy (WA). Our baseline represents the ASR performance of grapheme-based KL-HMM explained in Section 2.1 using either the *ortho* or *ortho+know* lexicons, i.e., no acoustic unit discovery or adaptation is done. We have also presented the performance of the grapheme-based HMM/GMM system on this corpus as reported in [1]. A comparison between the performance of the HMM/GMM system and our baseline shows the high quality of the baseline system. It can be seen that language-specific acoustic units discovery through either of approaches and subsequent adaptation of acoustic model and retraining of lexical model leads to statistically significant ASR improvements. The improvements can be consistently observed in both cases of using a *ortho* lexicon and *ortho+know* lexicon. Comparison of the results in *Approach-I* and *Approach-II* shows that the second approach yields significantly better ASR systems. This indicates that clustering the learned grapheme-to-multilingual phone relationships leads to a better language specific acoustic unit space representation compared to selecting a subset of multilingual phones based on the learned grapheme-to-multilingual phone relationships. Finally, *ortho+know* lexicon consistently yields better systems when compared to *ortho*. This is consistent with previous work in [1].

Table 2: *ASR WA of different systems using the ortho and ortho+know lexicons.*

	ortho	*ortho+know*
HMM/GMM	64.2 [1]	68.0 [1]
KL-HMM-no-adapt (Baseline)	69.2	74.0
KL-HMM-adapt-Approach-I	70.9	75.1
KL-HMM-adapt-Approach-II	**74.2**	**78.2**

4.2. Analysis

In this section we analyze the acoustic units discovered by the two approaches. Table 3 presents the multilingual phones discovered as acoustic units in *Approach-I*. It is worth mentioning that about 20 phones from Scottish Gaelic were not covered by the multilingual phone set used in our studies. These were mainly diphthongs. We have split the discovered phones into three categories: (a) phones existing in Scottish Gaelic according to [25] (b) phones existing in British English according to [28], and (c) phones from other languages. Among the phones existing in Scottish Gaelic and the multilingual phone set, only two phones are missing from the labels: {/ɤ/, /ʎ/}. However, it can be observed in the table that for each of the missing phones from Scottish Gaelic, an acoustically similar phone exists or is discovered. For example, /ʎ/ sounds acoustically similar to /j/. The appearance of phones from British English are due to existence of English words in the training data. Analysis of the phones from other languages shows that some of the phones are related to the diphthongs in the Scottish Gaelic, which were originally not covered by the multilingual phone set. A detailed phonetic analysis is part of our future work.

Unlike *Approach-I*, in the case of *Approach-II* the CD graphemes are not deterministically mapped to a multilingual phone. Rather they are probabilistically related to multilingual phones through the categorical distribution of the corresponding tied state or clustered CD grapheme state. We analyzed the categorical distributions of each tied state by selecting the phones with a probability higher than a threshold of 0.05. We observed that most of the phones were same as the ones presented in Table 3.[2] Furthermore, the phone /ʎ/ that was omitted in *Approach-I* was covered here.

Table 3: *Discovered acoustic units in Approach-I in IPA format.*

Phones in Scottish Gaelic	/m/, /n/, /ɲ/, /ŋ/, /p/, /b/, /t/, /d/, /k/, /g/, /f/, /v/, /s/, /ʃ/, /ç/,/ɤ/, /h/, /r/, /j/, /l/, /i/, /iː/, /u/, /uː/, /e/, /eː/, /o/, /oː/, /ə/, /ɛ/, /ɔ/, /ɔː/, /a/, /aː/, /aʊ/
Phones in British English	/aɪ/, /ɑ/ , /ɒ/, /ɑː/, /ʊ/, /ɪ/, /ɪə/, /tʃ/, /dʒ/, /w/, /z/, /θ/, /ð/ ,/eɪ/
Phones from other languages	/ɒʊ/, /ʒ/, /ʁ/, /jj/, /y/, /yː/, /õ/, /_œ_/, /_ɜ_ː/, /iːɐ/, /yːɐ/, /ʊɐ/, /bʁ/, /eːɐ/, /ɔɐ/ ,/ɔʏ /, /aɐ/, /ɑ̃/, /aːɐ/, /ɑs/, /ɛɒ/

4.3. Comparison to related approaches

As pointed earlier in Section 1, in phone-based ASR the multilingual ANN can be simply adapted by replacing the last layer with clustered CD phones of TL [14, 16]. A similar approach could be employed in our scenario in which the acoustic units are the clustered CD graphemes of TL. As noted earlier in Section 2.2.2 such an approach would be similar to the case where the acoustic units are automatic subword units derived using acoustic features [23, 24]. For sake of completeness, we investigated such a system by using the 1158 clustered grapheme states obtained through HMM/GMM using the *ortho* lexicon. Table 4 presents the results in terms of word accuracy. It can be seen that such a system performs better than *Approach-I*, however it does not beat the performance of *Approach-II*. This indicates that the clustered CD grapheme units obtained using posterior features could be better than clustered CD grapheme units obtained using acoustic features. This could have potential implications for development of automatic subword unit based lexicon development [29].

Table 4: *ASR results when obtaining the acoustic units from the clustered CD grapheme units of HMM/GMM system.*

	ortho	*ortho+know*
WA	73.8	77.0

5. Summary and future work

This paper proposed approaches to discover a language specific acoustic unit space based on grapheme-to-multilingual phone relationships, and using them to improve the grapheme-based ASR system of under-resourced languages through acoustic model and lexical model adaptation. It is worth mentioning that though the approaches have been developed in the KL-HMM based ASR framework, they can also be realized in the standard HMM-based ASR framework through corresponding probabilistic lexical modeling techniques [5, 6].

Our investigations on Scottish Gaelic showed that the proposed approaches, in addition to significantly improving the ASR system performance, help in discovering phone sets of the under-resourced language. Alternately, the proposed approaches provide means for automatic development of linguistic knowledge and resources for under-resourced languages by exploiting the existing multilingual acoustic and lexical resources. For our future studies, we aim to generate pronunciations based on the discovered acoustic units by exploiting the acoustic data-driven grapheme-to-phoneme conversion technique in the KL-HMM framework proposed in [30].

[2] The selected phones are not presented here due to space limitations.

6. References

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

[2] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, 1989.

[3] H. Bourlard and N. Morgan, *Connectionist speech recognition: A hybrid approach*. Norwell, MA, USA: Kluwer Academic Publishers, 1993.

[4] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," 1994.

[5] X. Luo and F. Jelinek, "Probabilistic classification of HMM states for large vocabulary continuous speech recognition," in *Proceedings of ICASSP*, vol. 1, 1999, pp. 353–356.

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

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

[8] M. Magimai.-Doss, R. Rasipuram, G. Aradilla, and H. Bourlard, "Grapheme-based automatic speech recognition using KL-HMM," in *Proceedings of Interspeech*, 2011, pp. 445–448.

[9] R. Rasipuram and M. Magimai-Doss, "Improving grapheme-based ASR by probabilistic lexical modeling approach," in *Proceedings of Interspeech*, 2013.

[10] J. Köhler, "Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks," in *Proceedings of ICASSP*, 1998, pp. 417–420.

[11] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition." *Speech Communication*, vol. 35, pp. 31–51, 2001.

[12] V.-B. Le and L. Besacier, "Automatic speech recognition for under-resourced languages: application to Vietnamese language," *Trans. Audio, Speech and Lang. Proc.*, vol. 17, no. 8, pp. 1471–1482, Nov. 2009.

[13] L. Burget *et al.*, "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in *Proceedings of ICASSP*, vol. 2010, no. 3, 2010, pp. 4334–4337.

[14] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proceedings of ICASSP*, 2013.

[15] G. Heigold *et al.*, "Multilingual acoustic models using distributed deep neural networks." in *Proceedings of ICASSP*, 2013, pp. 8619–8623.

[16] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Proceedings of ICASSP*, 2014.

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

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

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

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

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

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

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

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

[25] M. Wolters, "A diphone-based text-to-speech system for Scottish Gaelic," Master's thesis, University of Bonn, 1997.

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

[27] D. Johnson *et al.*, "ICSI quicknet software package," http://www.icsi.berkeley.edu/Speech/qn.html, 2004.

[28] P. Roach, "British English: Received pronunciation," *Journal of the International Phonetic Association*, vol. 34, no. 02, pp. 239–245, 2004.

[29] M. Razavi, R. Rasipuram, and 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.

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