ACOUSTIC AND LEXICAL RESOURCE CONSTRAINED ASR USING LANGUAGE-INDEPENDENT ACOUSTIC MODEL AND LANGUAGE-DEPENDENT PROBABILISTIC LEXICAL MODEL

Ramya Rasipuram Mathew Magimai.-Doss

Idiap-RR-02-2014

MARCH 2014

Acoustic and Lexical Resource Constrained ASR using Language-Independent Acoustic Model and Language-Dependent Probabilistic Lexical Model

Ramya Rasipuram[a,b], Mathew Magimai.-Doss[a]

{*ramya.rasipuram,mathew*}*@idiap.ch*

[a] *Idiap Research Institute, Martigny, Switzerland*
[b] *Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*

Abstract

One of the key challenge involved in building a statistical automatic speech recognition (ASR) system is modeling the relationship between lexical units (that are based on subword units in the pronunciation lexicon) and acoustic feature observations. To model this relationship two types of resources are needed, namely, acoustic resources (speech signals with word level transcriptions) and lexical resources (which transcribes each word in terms of subword units). Standard ASR systems typically use phonemes or phones as subword units. Not all languages have well developed acoustic resources and phonetic lexical resources. In this paper, we show that modeling of the relationship between lexical units and acoustic features can be factored into two parts through a latent variable, referred to as acoustic units, namely: (a) acoustic model that models the relationship between acoustic features and acoustic units and (b) lexical model that models the relationship between lexical units and acoustic units. Through this understanding, we elucidate that in standard hidden Markov model (HMM) based ASR system, the lexical model is deterministic (i.e., there exists an one-to-one relationship between lexical units and acoustic units), and it is the deterministic lexical model that imposes the need for well developed acoustic and lexical resources in the target language or domain when building ASR system. We then propose an approach that addresses both acoustic resource and lexical resource constraints. More specifically, in the proposed approach the acoustic model models the relationship between acoustic features and multilingual phones (acoustic units) on target language-independent data, and the lexical model models a probabilistic relationship between lexical units based on graphemes and multilingual phones on small amount of target language data. We show the potential and the efficacy of the proposed approach through experiments and comparisons with other approaches on three different ASR tasks, namely, non-native accented speech recognition, rapid development of ASR system for a new language and development of ASR system for a minority language.

Keywords:

Automatic speech recognition; Kullback-Leibler divergence based hidden Markov model; grapheme; phoneme; lexical modeling; lexicon.

1. Introduction

State-of-the-art automatic speech recognition (ASR) systems are based on hidden Markov models (HMMs). Development of HMM-based ASR system is often decomposed into two problems (Rabiner, 1989; Bourlard and Morgan, 1994). First, the relationship between "lexical units" (typically representing subword units) and acoustic feature observations, such as cepstral features is modeled. Second, the syntactic constraints of language are modeled.

The present paper focuses on the first problem where to model the relationship between lexical units and acoustic features, well developed acoustic resources (speech data with transcription) and lexical resources (phonetic dictionary) are required. While this is not an issue for resource rich languages, it is challenging for under-resourced languages and domains that may not have such resources (Besacier et al., 2014). In the literature lack of acoustic resources has been typically addressed by first modeling the relationship between lexical units and feature observations on out-of-domain or language-independent data and then adapting it on target language data through techniques such as bootstrapping, maximum a posteriori adaptation (MAP) technique, maximum likelihood linear regression (MLLR) (Kohler, 1998; Beyerlein et al., 2000; Schultz and Waibel, 2001; Le and Besacier, 2009; Burget et al., 2010). The lack of phonetic lexical resources has been addressed through the use of alternate subword units, such as graphemes (Schukat-Talamazzini et al., 1993; Kanthak and Ney, 2002; Killer et al., 2003; Dines and Magimai-Doss, 2007; Ko and Mak, 2014). However, the lack of both acoustic and lexical resources has been sporadically attempted (Stüker, 2008a,b).

In this paper, we first show that the modeling of relationship between lexical units and acoustic features can be factored into two parts or models through a latent variable, referred to as "acoustic units", namely,

1. *acoustic model* where the relationship between acoustic units and acoustic feature observation is modeled.
2. *lexical model* where the relationship between acoustic units and lexical units is modeled.

We then elucidate that in standard HMM-based ASR system the lexical model is *deterministic*. The deterministic lexical model imposes constraints such as, the acoustic units and the lexical units have to be of the same kind; the acoustic resources from target language or domain are required to train or adapt both acoustic model and lexical model.

In a recent work, we showed that there are probabilistic lexical modeling approaches such as, Kullback-Leibler divergence based hidden Markov model (KL-HMM) (Aradilla et al., 2008) where the relationship between lexical units and acoustic units is probabilistic (Rasipuram and Magimai.-Doss, 2013a). Probabilistic lexical modeling relaxes certain constraints imposed by deterministic lexical modeling and as a consequence, acoustic model and lexical model can be independently trained on different set of resources (Imseng et al., 2011, 2012; Rasipuram et al., 2013a); different kinds of subword units can be modeled in an ASR system (Magimai.-Doss et al., 2011; Imseng et al., 2011; Rasipuram et al., 2013a) and different types of contextual units can be modeled in an ASR system (Magimai.-Doss et al., 2011; Imseng et al., 2011, 2012; Rasipuram et al., 2013a). Motivated by these findings, this paper proposes an approach for rapid development of ASR systems in the framework of probabilistic lexical modeling with minimal acoustic and lexical resources from target language or domain. In the proposed approach,

- acoustic units are "multilingual phones" and lexical units are based on graphemes of the target language;
- an acoustic model is trained on language-independent acoustic and lexical resources;
- lexical model, which captures a probabilistic relationship between graphemes and multilingual phones, is trained on relatively small amount of target language-dependent acoustic data.

On three different ASR tasks we validate the proposed approach and compare it with standard approaches such as, acoustic model adaptation and use of Tandem features that exploit out-of-domain resources, and training acoustic model and lexical model on target language data alone.

The paper is organized as follows. Section 2 provides a background on standard HMM-based ASR system and elucidates the deterministic lexical model aspect in theory and practice. Section 3 presents implications of deterministic lexical modeling. Section 4 presents three different probabilistic lexical modeling approaches along with their potential implications. Sections 5 and 6 present the experimental setup and the results, respectively. Finally, in Section 7 we provide a discussion followed by conclusion.

2. Background

In statistical ASR approach, the goal is to find the best matching (most likely) word sequence W^* given the acoustic observation sequence $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T\}$, where t denotes the frame number and T the total number of frames. Formally,

$$\begin{aligned} W^* &= \arg\max_{W \in \mathcal{W}} P(W|X, \Theta) && (1) \\ &= \arg\max_{W \in \mathcal{W}} \frac{P(X|W, \Theta_A) \cdot P(W|\Theta_L)}{P(X|\Theta)} && (2) \\ &= \arg\max_{W \in \mathcal{W}} P(X|W, \Theta_A) \cdot P(W|\Theta_L) && (3) \end{aligned}$$

where $\mathcal{W}$ denotes the set of all possible word sequences, W denotes a word sequence, $\Theta = \{\Theta_A, \Theta_L\}$ denotes the set of acoustic and language model parameters. The acoustic model parameter set $\Theta_A = \{\theta_a, \theta_{pr}, \theta_l\}$ includes acoustic model (θ_a), lexicon (θ_{pr}) and lexical model (θ_l) parameters.

HMM-based ASR is a statistical ASR approach where given acoustic model, lexicon and language model, the most likely word sequence W^* is achieved by finding the most likely state sequence Q^*, i.e.,

$$\begin{aligned} Q^* &= \arg\max_{Q \in \mathcal{Q}} P(Q, X|\Theta) && (4) \\ &= \arg\max_{Q \in \mathcal{Q}} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t = l^i, \Theta_A) \cdot P(q_t = l^i|q_{t-1} = l^j, \Theta) && (5) \\ &= \arg\max_{Q \in \mathcal{Q}} \sum_{t=1}^{T} [\log p(\mathbf{x}_t|q_t = l^i, \Theta_A) + \log P(q_t = l^i|q_{t-1} = l^j, \Theta)] && (6) \end{aligned}$$

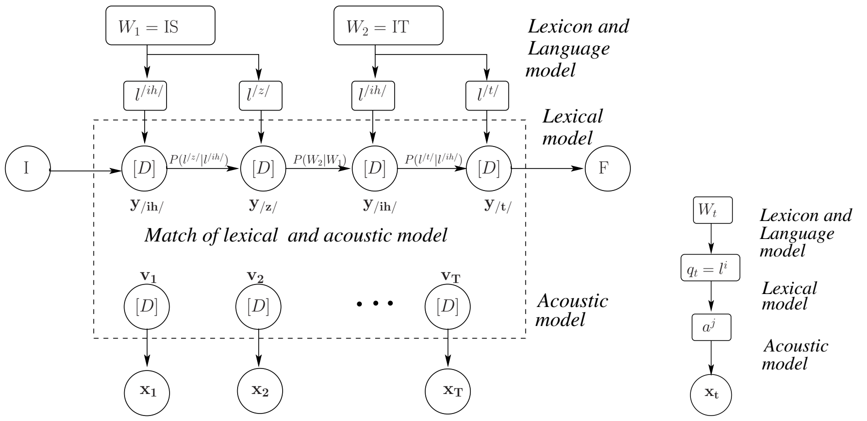
where $\mathcal{Q}$ denotes the set of possible HMM state sequences and each $Q = \{q_1, \ldots, q_t, \ldots, q_T\}$ denotes a sequence of lexical HMM states corresponding to a word sequence hypothesis[1], $q_t \in \mathcal{L} = \{l^1, \ldots l^i \ldots l^I\}$ and I is the number lexical units. In subword unit based ASR system, if phones are used as subword units then each lexical unit l^i represents a phone or a polyphone and if graphemes are used as subword units then each lexical unit l^i represents a grapheme or a polygrapheme. Eqn. (5) results after *i.i.d* and first order Markov assumptions. Eqn. (6) is as a result of log transformation to Eqn. (5). Usually, $\log p(\mathbf{x}_t|q_t = l^i, \Theta_A)$ is referred to as *local emission score* and $\log P(q_t = l^i|q_{t-1} = l^j, \Theta)$ is referred to as *transition score*. If l^j is the last lexical unit of a word and l^i is the first lexical unit of next word then $P(q_t = l^i|q_{t-1} = l^j, \Theta)$ is the language model probability otherwise it is the HMM state transition probability. The present paper deals only with the issues related to the estimation of local emission score.

Standard HMM-based ASR systems, for various reasons as elucidated shortly in the following subsections, implicitly model the dependency between acoustic feature observation $\mathbf{x}_t$ and lexical unit l^i through a *latent* variable a^d as

$$\begin{aligned} p(\mathbf{x}_t|q_t = l^i, \Theta_A) &= \sum_{d=1}^{D} p(\mathbf{x}_t, a^d|q_t = l^i, \Theta_A) \\ &= \sum_{d=1}^{D} p(\mathbf{x}_t|a^d, q_t = l^i, \theta_a, \theta_l) \cdot P(a^d|q_t = l^i, \theta_l) \\ &= \sum_{d=1}^{D} p(\mathbf{x}_t|a^d, \theta_a) \cdot P(a^d|q_t = l^i, \theta_l) && (7) \end{aligned}$$

[1] That is a sentence model consists of sequence of word models constrained by the language model, word models consist of sequence of subword models constrained by pronunciation lexicon and subword model consists of concatenation of one or more HMM states

Figure 1: Graphical model representation



(a) of system based on Eqns (6) and (7)

(b) of deterministic lexical model system at time frame t

We refer to latent variable a^d as acoustic unit. Furthermore, $\mathcal{A} = \{a^1, \dots a^d, \dots a^D\}$ is the set of acoustic units, D is the number of acoustic units, θ_a denotes the acoustic model parameter set and θ_l denotes the lexical model parameter set. The final relationship in Eqn. (7) is as a result of the assumption that given a^d, $\mathbf{x}_t$ is independent of l^i. In Eqn. (7), $p(\mathbf{x}_t|a^d, \theta_a)$ refers to the acoustic model likelihood, and $P(a^d|q_t = l^i, \theta_l)$ is the probability of acoustic unit given lexical unit given by the lexical model.

Figure 1(a) shows the graphical model representation of system based on Eqns (6) and (7) for the word sequence "IS IT". The figure shows that a sequence of words constrained by language model are represented by sequence of lexical units (l^{ih} l^z l^{ih} l^t) as given by pronunciation lexicon. For each lexical unit l^i, lexical model computes a D dimensional categorical variable $\mathbf{y}_i = [y_i^1, \dots, y_i^d, \dots y_i^D]^{\mathrm{T}}$, $y_i^d = P(a^d|l^i, \theta_l)$ that models a probabilistic relationship between the lexical unit l^i and D acoustic units. Given the acoustic feature observation sequence $\{\mathbf{x}_1, \dots, \mathbf{x}_T\}$, the acoustic model computes the sequence of acoustic likelihood vectors $\{\mathbf{v}_1, \dots, \mathbf{v}_T\}$, where $\mathbf{v}_t = [v_t^1 \dots, v_t^d, \dots, v_t^D]^{\mathrm{T}}$, and $v_t^d = p(\mathbf{x}_t|a^d, \theta_a)$. The local emission score at time frame t in Eqn (6) can be seen as a match between acoustic and lexical model scores as given in Eqn (7), which can be rewritten in terms of $\mathbf{y}_i$ and $\mathbf{v}_t$ as,

$$\log p(\mathbf{x}_t|q_t = l^i, \Theta_A) = \log \sum_{d=1}^{D} p(\mathbf{x}_t|a^d, \theta_a) \cdot P(a^d|q_t = l^i, \theta_l) = \log \mathbf{y}_i^{\mathrm{T}} \mathbf{v}_t \qquad (8)$$

In standard HMM-based ASR systems the lexical model is deterministic, i.e., each lexical unit l^i is deterministically mapped to an acoustic unit a^j ($l^i \mapsto a^j$), i.e.,

$$y_i^d = P(a^d | q_t = l^i, \theta_l) \; = \left\{ \begin{array}{ll} 1, & \text{if d = j ;} \\ 0, & \text{otherwise.} \end{array} \right. \tag{9}$$

The graphical model representation of deterministic lexical model based system at time frame t is illustrated in Figure 1(b). It is worth mentioning that in HMM-based ASR literature due to this deterministic relationship typically distinction is not made between acoustic and lexical units. Our main reason to refer lexical unit l^i and acoustic unit a^j distinctively here is to bring out the contributions of the present paper clearly.

Depending on the subword context modeled, there are two types of ASR systems, namely, context-independent subword unit based ASR system and context-dependent subword unit based ASR system. State-of-the-art ASR systems are typically based on context-dependent subword units.

2.1. Acoustic Modeling

In the literature, there are two main approaches for acoustic modeling, depending on the way acoustic units are modeled and the acoustic score $p(\mathbf{x}_t | a_t^d, \theta_a)$ is estimated, namely,

1. HMM/GMM approach (Rabiner, 1989) where acoustic score $p(\mathbf{x}_t | a^d, \theta_a)$ is estimated given a mixture of Gaussians that model an acoustic unit a^d. The acoustic model parameter set θ_a consists of set of acoustic units $\mathcal{A}$ and GMM parameters (means, variances and mixture weights) corresponding to acoustic units.
2. hybrid HMM/ANN approach (Morgan and Bourlard, 1995) where an artificial neural network (ANN) is first trained to estimate $P(a^d | \mathbf{x}_t, \theta_a)$ and then scaled-likelihood $p_{sl}(\mathbf{x}_t | a^d, \theta_a)$ is estimated as

$$p_{sl}(\mathbf{x}_t | a^d, \theta_a) = \frac{p(\mathbf{x}_t | a^d, \theta_a)}{P(\mathbf{x}_t)} = \frac{p(a^d | \mathbf{x_t}, \theta_a)}{P(a^d)} \tag{10}$$

   $P(a^d)$ is estimated on the training dataset through counting. The acoustic model parameter set θ_a consists of set of acoustic units $\mathcal{A}$, ANN parameters θ_a (weights and biases) and priors $\{P(a^d)\}_{d=1}^{D}$.

In the case of context-independent subword unit based ASR system, the acoustic unit set $\mathcal{A}$ is defined purely based on the pronunciation lexicon and the minimum duration constraint (knowledge driven). The number of acoustic units D is either $K \times M$, where K is the number of context-independent subword units in the lexicon and M is the minimum duration constraint (typically, $M = 3$) or simply the number of context-independent subword units K.

In the case of context-dependent subword unit based ASR system, $\mathcal{A}$ is usually obtained through decision tree based HMM state clustering and state tying technique that uses pronunciation lexicon, linguistic knowledge (question set for decision trees) and acoustic data. The number of acoustic units D varies depending on the hyper parameters such as, state occupancy count and log-likelihood threshold that are used during state tying and clustering (Young et al., 2006). However, D is well below the total number of context-dependent subword units possible for a given preceding context length c_l, following context length c_r and M, i.e., $M \cdot K^{c_r+c_l+1}$. In the literature, this is mainly done due to data sparsity and model complexity issues.

2.2. (Deterministic) Lexical Modeling

In context-independent subword unit based ASR system, the deterministic relationship between lexical and acoustic units is knowledge driven. Therefore, lexical model training is not involved. There are two types of context-independent subword unit based ASR systems one can encounter, where

1. both acoustic units and lexical units incorporate minimum duration constraint, i.e., $I = D = K \times M$. This is a system where the relationship between acoustic feature vectors ($\mathbf{x}_t$) and lexical units (l^i) is directly modeled.
2. acoustic units are context-independent subword units without any minimum duration constraint and lexical units are context-independent subword units with minimum duration constraint, i.e., $D = K$ and $I = K \times M$. In this case the deterministic relationship as in Eqn. (9) is represented by a look-up table with I rows.

Context-independent subword unit based HMM/GMM systems are of the first kind. In the past, context-independent subword unit based hybrid HMM/ANN systems were typically of the second kind (Morgan and Bourlard, 1995).

In context-dependent subword unit based ASR system, the deterministic relationship in Eqn. (9) is learned during acoustic model training, more precisely, at the stage of state clustering and tying. The total number of lexical units $I = M \cdot K^{c_r+c_l+1}$. The state tying process builds a look-up table with I rows that maps each lexical unit l^i to one of the D acoustic units. In toolkits such as HTK, this table is not explicitly seen. However, it can be obtained from the HMM definition file (or MMF file) and tied list after state clustering and tying (Young et al., 2006).

3. Implications of Deterministic Lexical Modeling

As described in the previous section, in standard HMM-based ASR systems the lexical model i.e., the relationship between lexical units $l^i \in \mathcal{L}$ and acoustic units $a^d \in \mathcal{A}$ is deterministic and the pronunciation lexicon (θ_{pr}) determines the lexical unit set $\mathcal{L}$ and the acoustic unit set $\mathcal{A}$. As a consequence,

- if $\mathcal{L}$ is based on phone subword units (phone-based ASR system) or grapheme subword units (grapheme-based ASR system) then $\mathcal{A}$ is also based on phones or graphemes, respectively.

- if $\mathcal{L}$ is based on context-independent subword units (context-independent subword unit based ASR system) or context-dependent subword units (context-dependent subword unit based ASR system) then $\mathcal{A}$ is also based on context-independent subword units or context-dependent subword units, respectively.

The first constraint deterministic lexical modeling imposes is the availability of sufficient and well developed acoustic data in the target language or domain to train effectively both acoustic model and lexical model. Unfortunately, many languages may not have such well developed acoustic resources. Most of the ASR systems use phones as lexical units. Therefore, the second constraint that arises as a result of deterministic lexical modeling is the availability of well developed phonetic lexicon. Again, many languages lack such well developed lexical resources. For a language, it can happen that there are different phonetic lexicons based on different phone sets. For instance, in English there are phonetic lexicons based on ARPABET, SRI phone set, UNISYN, SAMPA. The third constraint that deterministic lexical model introduces is that ASR system trained with one phone set can not be directly ported to or used for a new domain which has a lexicon based on different phone set.

3.1. Lack of acoustic resources

In the literature, lack of acoustic resources has been typically addressed using acoustic model adaptation techniques that exploit multilingual or crosslingual acoustic and lexical resources (Kohler, 1998; Beyerlein et al., 2000; Schultz and Waibel, 2001; Le and Besacier, 2009; Burget et al., 2010). Generally, the first step in these techniques is the definition of common or universal phone set across all out-of-domain languages and target language. This step ensures that the phone sets match across languages, thus, addressing the third constraint. The common or universal phone set can be defined either in knowledge-based manner or data-driven manner. Multilingual acoustic model (GMMs) and lexical model (state tying) are then trained on data from out-of-domain languages. The parameters of multilingual acoustic model are adapted on target language data using techniques such as, bootstrapping, maximum a posteriori adaptation (MAP) technique, maximum likelihood linear regression (MLLR) technique, sub-space Gaussian mixture model (SGMM) approach while the out-of-domain lexical model (state tying) is either retained (Kohler, 1998; Beyerlein et al., 2000; Le and Besacier, 2009) or redefined using target language data (Schultz and Waibel, 2001; Burget et al., 2010).

3.2. Lack of lexical resources

In practice, phone-based ASR system development can be seen as a two stage process. Development of pronunciation lexicon followed by ASR system training. Pronunciation lexicon development is a semi-automatic process, where usually given an existing manually developed (or verified) lexicon, a grapheme-to-phoneme converter (Bisani and Ney, 2008; Novak, 2011) is trained to extract pronunciations for new words or pronunciation variants. The augmented lexicon is then used to build ASR system. However, for some languages seed lexicon may not be available to train the grapheme-to-phoneme convertor. Therefore, alternate subword units like graphemes, which makes lexicon development easy, have been explored in the literature (Schukat-Talamazzini et al., 1993; Kanthak and Ney, 2002; Killer et al., 2003; Dines and Magimai-Doss, 2007; Ko and Mak, 2014).

The success of grapheme-based ASR system primarily depends on the grapheme-to-phoneme relationship of the language. The reason for this is as follows. It can be seen in Eqn. (7) that the acoustic model score $p(\mathbf{x}_t|a^d, \theta_a)$ models the dependency between acoustic feature observation $\mathbf{x}_t$ and acoustic unit a^d. As discussed earlier in this section, due to deterministic lexical modeling in standard HMM-based ASR systems, both acoustic unit a^d and lexical unit l^i are the same and represent graphemes. However, the acoustic feature observations or the cepstral features depict the envelop of short-term spectrum which is related to phones. As a result regular is the grapheme-to-phoneme relationship, better is the acoustic model. Indeed, the use of grapheme as subword units has mainly succeeded for languages such as, Spanish, Finnish where the grapheme-to-phoneme relationship is regular (Kanthak and Ney, 2002; Killer et al., 2003; Ko and Mak, 2014). For languages such as, English that have irregular grapheme-to-phoneme relationship, it has been found that grapheme-based system performs worse compared to phone-based system (Schukat-Talamazzini et al., 1993; Kanthak and Ney, 2002; Killer et al., 2003; Dines and Magimai-Doss, 2007; Ko and Mak, 2014).

3.3. Lack of acoustic and lexical resources

When the language lacks both acoustic and phone lexical resources, multilingual and crosslingual grapheme-based approaches that can leverage from resources available in other languages have been explored (Kanthak and Ney, 2003; Stüker, 2008a,b). Similar to multilingual phone subword modeling, multilingual grapheme subword modeling is based on *universal* or *multilingual* grapheme set formed by merging graphemes that are common across different languages. However, unlike multilingual phone sets, its not trivial to port multilingual grapheme sets to new languages mainly because of two reasons. Firstly, grapheme sets of all languages may not match or overlap. To overcome this issue, either transliteration or data driven mapping has been employed (Stüker, 2008b). Secondly, sharing of acoustic models of grapheme subword units across languages is not evident, since the relationship between graphemes and phones may differ considerably across languages. Investigations until now have shown that multilingual grapheme-based ASR systems generally perform worse compared to monolingual grapheme-based ASR systems (Kanthak and Ney, 2003; Stüker, 2008a,b). This is unlike phone subword units where it has been shown that multilingual acoustic models can outperform monolingual acoustic models (Schultz and Waibel, 2001).

4. Probabilistic Lexical Modeling

In Section 2 we explained standard HMM-based ASR where the lexical model is deterministic. In this section, we present three approaches which learn probabilistic lexical model by training a second HMM, whose states represent lexical units and each state l^i is parameterized by a categorical distribution $\mathbf{y}_i$. The categorical distribution tends to capture a probabilistic relationship between a lexical unit l^i and D acoustic units. We present these techniques from the perspective of hybrid HMM/ANN system. However, it should be noted that these approaches are equally applicable to HMM/GMM system (Rasipuram and Magimai.-Doss, 2013b). These approaches presume that acoustic unit set $\mathcal{A}$ is defined and trained acoustic model is available.

4.1. Kullback-Leibler Divergence based HMM

In the first approach, probabilistic lexical model is learned through estimates of $P(a^d|\mathbf{x}_t, \theta_a)$ in the framework of Kullback-Leibler divergence based HMM (KL-HMM) (Aradilla et al., 2008). The feature observations for the second HMM are $\mathbf{z}_t = [z_t^1 \ldots, z_t^d, \ldots, z_t^D)]^{\mathrm{T}}$ where $z_t^d = P(a^d|\mathbf{x}_t, \theta_a)$. It is worth mentioning that KL-HMM was originally developed from the perspective of acoustic modeling (Aradilla et al., 2008), as an alternative to Tandem approach (Hermansky et al., 2000). However, as shown recently and briefly explained in this section, KL-HMM is a probabilistic modeling approach (Rasipuram and Magimai.-Doss, 2013a,b).

In KL-HMM, as both feature observations and state distributions are probability vectors, local score at each HMM state is the Kullback-Leibler (KL) divergence between feature observations $\mathbf{z}_t^d$ and categorical distributions $\mathbf{y}_i$,

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

The above equation represents the case where $\mathbf{y}_i$ is the reference distribution and the local score is denoted as S_{KL}. However, KL-divergence is an asymmetric measure. Thus, there are other possible ways to estimate KL-divergence, namely,

1. Reverse KL-divergence (S_{RKL}): In this case the acoustic state probability vector $\mathbf{z}_t$ is the reference distribution

$$S_{RKL}(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} z_t^d \log\left(\frac{z_t^d}{y_i^d}\right) \tag{12}$$

2. Symmetric KL-divergence (S_{SKL}):

$$S_{SKL}(\mathbf{y}_i, \mathbf{z}_t) = \frac{1}{2} \cdot [S_{KL} + S_{RKL}] \tag{13}$$

The categorical distributions $\{\mathbf{y}_i\}_{i=1}^{I}$ are estimated by Viterbi expectation maximization algorithm which minimizes a cost function based on S_{KL} or S_{RKL} or S_{SKL}. Finally, the decoding is performed by replacing the log-likelihood based score in the standard Viterbi decoder with KL-divergence based local score.

4.2. Tied Posterior

In the second approach, probabilistic lexical model is learned through scaled-likelihood estimates $p_{sl}(\mathbf{x}_t|a^d, \theta_a)$ (see Eqn. (10)). The approach referred to as tied posterior approach (Rottland and Rigoll, 2000), was originally proposed in the framework of hybrid HMM/ANN to build context-dependent subword unit based ASR system using an ANN trained to classify context-independent subword unit (phones). The emission likelihood at each context-dependent state $q_t = l^i_{cd}$ is estimated as,

$$p(\mathbf{x}_t|q_t = l^i_{cd}) = \sum_{d=1}^{D} w^d_i \cdot p_{sl}(\mathbf{x}_t|a^d_{ci}) \tag{14}$$

where a^d_{ci} is context-independent phone, D is the number of context-independent phones, $p_{sl}(\mathbf{x}_t|a^d_{ci})$ is the scale-likelihood (see Eqn. 10), $0 \leq w^d_i \leq 1$ is the weight corresponding to context-dependent phone l^i_{cd} and $\sum_{d=1}^{D} w^d_i = 1$. The weights w^d_i are estimated by maximizing the cost function based on log-likelihood i.e., $\log p(\mathbf{x}_t|q_t = l^i_{cd})$. Comparison between (14) and (7) shows that l^i_{cd} corresponds to lexical unit l^i, a^d_{ci} corresponds to acoustic unit a^d and w^d_i corresponds to $y^d_i = P(a^d|l^i, \theta_l)$. In other words tied posterior approach is a HMM-based ASR approach that incorporates probabilistic lexical modeling.

Tied posterior approach can be interpreted along the lines similar to KL-HMM approach where the states of the secondary HMM are parameterized by $\mathbf{y}_i$. However, the feature observations used to train the HMM in tied posterior approach are vectors of scaled-likelihood $\mathbf{v}_t = [v^1_t \ldots, v^d_t, \ldots, v^D_t]^{\mathrm{T}}$ where $v^d_t = p_{sl}(\mathbf{x}_t|a^d, \theta_a)$, and the local score is

$$S_{tied}(\mathbf{y}_i, \mathbf{v}_t) = \log\big(\sum_{d=1}^{D} y^d_i . v^d_t\big) = \log\big(\mathbf{y}^{\mathrm{T}}_i \mathbf{v}_t\big) \tag{15}$$

Like KL-HMM, the parameters $\{\mathbf{y}_i\}^I_{i=1}$ can be estimated using Embedded Viterbi training algorithm, and the decoding can be performed by replacing the log-likelihood based score in standard Viterbi decoder with the local score $S_{tied}(\mathbf{y}_i, \mathbf{v}_t)$.

4.3. Scalar Product HMM

In KL-HMM system, local score is based on KL-divergence. However, two posterior probability distributions can be compared with different cost functions such as, scalar product, Bhattacharya distance (Soldo et al., 2011). It is possible to envisage a secondary HMM where local score is based on scalar product

$$S_{SP}(\mathbf{y}_i, \mathbf{z}_t) = \log\big(\mathbf{y}^{\mathrm{T}}_i \mathbf{z}_t\big) \tag{16}$$

We refer to this system as scalar product HMM (SP-HMM). Again, $\{\mathbf{y}_i\}^I_{i=1}$ can be estimated using embedded Viterbi training algorithm, and the decoding can be performed by replacing the log-likelihood based score in standard Viterbi decoder with $S_{SP}(\mathbf{y}_i, \mathbf{v}_t)$.

The SP-HMM is of particular interest here for the following reasons,

1. it can be seen as a particular case of tied posterior approach where the priors in the scaled-likelihood estimation are dropped or assumed to be equal.
2. as discussed in this section, SP-HMM and KL-HMM differ only in terms of the cost function used for parameter estimation and the local score used during decoding.

In the case of KL-HMM, Tied and SP-HMM approaches, the lexical model parameter set $\theta_l = \{\mathbf{y}_i\}_{i=1}^{I}$. The parameter estimation and decoding with KL-HMM, Tied and SP-HMM approaches is elaborated in Appendix A. For more details about the parameter estimation in KL-HMM, the reader is referred to (Aradilla, 2008). An issue that is common to all the probabilistic lexical modeling approaches discussed in this section is the robust estimation of $\{\mathbf{y}_i\}_{i=1}^{I}$, especially when the lexical units represent context-dependent subword units. This can be addressed by clustering and tying the states of KL-HMM or tied posterior or SP-HMM systems using the approach proposed in (Imseng et al., 2012).

4.4. Similarities and dissimilarities between KL-HMM, Tied and SP-HMM

In the three probabilistic lexical modeling approaches discussed, local score estimation at time frame t can be seen as a match between "bottom-up" acoustic information $\mathbf{z}_t$ or $\mathbf{v}_t$ and "top-down" lexical information $\mathbf{y}_i$ related to latent variable a^d, as shown in Figure 1(a). Yet another similarity between the three approaches is that they reduce to standard hybrid HMM/ANN system described earlier in Section 2 when the lexical model is deterministic, i.e., $\mathbf{y}_i$ is Kronecker delta distribution.

Despite these similarities, KL-HMM approach has additional advantages compared to Tied and SP-HMM approaches. We discuss them briefly in this section. From the communication theory perspective (Bahl et al., 1983), standard HMM-based ASR approach can be seen as a communication problem where noisy output of acoustic channel (i.e., sequence of acoustic likelihood vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_T\}$ or sequence of acoustic posterior vectors $\{\mathbf{z}_1, \ldots, \mathbf{z}_T\}$) is decoded by a linguistic decoder, which implies comparison to possible sequences of lexical model parameter vectors (for e.g. $\{\mathbf{y}_i, \ldots \mathbf{y}_g\}$ where $i, g \in \{1, \ldots, I\}$) with lexical transition constraints ($P(q_t = l^i | q_{t-1} = l^j)$). Thus, standard HMM-based ASR inherently gives more importance to lexical model and consequently relies on purity or correctness of the lexical knowledge imparted into the system. This aspect has particularly been observed in the case of pronunciation variation modeling of conversational speech where one of the best approach is to add pronunciation variants, i.e., improve the deterministic lexical model (Strik and Cucchiarini, 1999).

The KL-HMM approach using local score $S_{KL}(\mathbf{y}_i, \mathbf{z}_t)$ where $\mathbf{y}_i$ is the reference distribution reflects the HMM-based ASR. More specifically,

$$\begin{aligned} S_{KL}(\mathbf{y}_i, \mathbf{z}_t) &= \sum_{d=1}^{D} y_i^d \log\left(\frac{y_i^d}{z_t^d}\right) \\ &= \sum_{d=1}^{D} y_i^d \log y_i^d - \sum_{d=1}^{D} y_i^d \log z_t^d \end{aligned} \tag{17}$$

The first part of Eqn. (17) or the entropy of probability distribution $\mathbf{y}_i$ takes into account the uncertainty in the lexical model, and the second part or the cross entropy compares the acoustic model against the lexical model. It is trivial to see the point made above about the purity of lexical knowledge by turning $\mathbf{y}_i$ into deterministic lexical model i.e., Kronecker delta distribution. In such a case, the hybrid HMM/ANN approach (Bourlard and Morgan, 1994) where acoustic model estimates $P(q_t = a^d|\mathbf{x}_t, \theta_a)$ rather than $p_{sl}(\mathbf{x}_t|q_t = a^d, \theta_a)$ can be seen as a special case of KL-HMM approach.

KL-HMM approach, however, is capable of reversing the importance given to acoustic model and lexical model by changing the local score to $S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$, i.e.,

$$\begin{aligned} S_{RKL}(\mathbf{y}_i, \mathbf{z}_t) &= \sum_{d=1}^{D} z_t^d \log\left(\frac{z_t^d}{y_i^d}\right) \\ &= \sum_{d=1}^{D} z_t^d \log z_t^d - \sum_{d=1}^{D} z_t^d \log y_i^d \end{aligned} \tag{18}$$

It can be observed from Eqn. (18) that the first quantity or the entropy of probability distribution $\mathbf{z}_t$ is independent of lexical state and the matching only takes place with the second quantity, i.e., the cross entropy between distributions $\mathbf{z}_t$ and $\mathbf{y}_i$, with $\mathbf{z}_t$ as the reference. The local score $S_{SKL}(\mathbf{y}_i, \mathbf{z}_t)$ is the case where equal importance is given to both acoustic model and lexical model.

Yet another distinction between KL-HMM and Tied/SP-HMM approach is that, in KL-HMM the local score is discriminative (Blahut, 1974), i.e., the acoustic model and lexical model is matched discriminatively, irrespective of the type of local score used, i.e., $S_{KL}(\mathbf{y}_i, \mathbf{z}_t)$ or $S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$ or $S_{SKL}(\mathbf{y}_i, \mathbf{z}_t)$. We use these distinctions to better explain our findings in Section 6.

The above differences among different KL-divergence based local scores is from decoding perspective. The details on the role of different cost functions in estimating $\mathbf{y}_i$ i.e., from training perspective is presented in (Rasipuram and Magimai.-Doss, 2013a).

4.5. Potential of Probabilistic Lexical Modeling

In the case of probabilistic lexical modeling each lexical unit l^i is related to all acoustic units $\{a^d\}_{d=1}^{D}$ in probabilistic manner. As a consequence,

- the parameters of acoustic model θ_a and lexical model θ_l can be trained on independent set of resources. In the light of that, previous works on KL-HMM such as (Imseng et al., 2011, 2012; Rasipuram et al., 2013a) suggest that ASR systems can be rapidly developed using language-independent acoustic model and training only the lexical model on target language or domain data.

- $\mathcal{L}$ and $\mathcal{A}$ can model different contextual units. For instance, as in previous work on KL-HMM (Magimai.-Doss et al., 2011; Imseng et al., 2011, 2012; Rasipuram et al., 2013a), $\mathcal{L}$ can be based on context-dependent subword units while $\mathcal{A}$ can be based on context-independent subword units. These ASR systems have been found to yield performance comparable to or better than standard context-dependent subword unit based HMM/GMM system.

- it is not necessary that subword unit set used for defining acoustic units should be same as subword unit set used for defining lexical units. The lexical model can capture the relationship between the distinct subword unit sets through acoustics. This flexibility has been exploited to build ASR systems where the acoustic unit set $\mathcal{A}$ is based on phones and the acoustic model is trained on auxiliary acoustic and lexical resources, and the lexical unit set $\mathcal{L}$ is based on graphemes and the lexical model is learned on target language or domain data (Magimai.-Doss et al., 2011; Imseng et al., 2011; Rasipuram et al., 2013a; Rasipuram and Magimai.-Doss, 2013b). It has been observed that this grapheme-based ASR approach could result in better or comparable performance even for languages such as English (where grapheme-to-phoneme relationship is irregular) compared to two stage approach where G2P training is followed by ASR system development (Rasipuram and Magimai.-Doss, 2013a).

Given these findings we hypothesize that, compared to conventional approach of rapid development of ASR system through acoustic model adaptation of deterministic lexical model based ASR system, ASR systems can be rapidly and effectively built with probabilistic lexical modeling approach,

- by training a “shared” multilingual phone based language-independent acoustic model and

- learning a probabilistic relationship between graphemes of target language and multilingual phones using target language or domain acoustic data (speech data with word level transcriptions).

5. Experimental Setup

We validate our hypothesis by training a single language-independent multilingual acoustic model and conducting ASR studies on three different resource-constrained tasks where only lexical model is trained, namely,

System	Acoustic model			Lexical Model		
	Acoustic units	Approach	Train/ Adapt	Lexical units	Approach	Train/ Adapt
KL-HMM	CI	ANN	LI	CD	Prob	LD
SP-HMM	CI	ANN	LI	CD	Prob	LD
Tied-HMM	CI	ANN	LI	CD	Prob	LD
Tandem	(CI+)cCD	(ANN+)GMM	(LI+)LD	CD	Det	LD
MAP	cCD	GMM	LI+LD	CD	Det	LI
MLLR	cCD	GMM	LI+LD	CD	Det	LI
HMM/GMM	cCD	GMM	LD	CD	Det	LD

Table 1: Overview of different systems. CI denotes context-independent subword units, cCD denotes clustered states of context-dependent subword unit based HMM/GMM system and CD denotes context-dependent subword units. LI denotes language-independent data is used to train or adapt the model, LD denotes language-dependent data is used to train or adapt the model and LI+LD denotes both language-independent and language-dependent data is used to train the model. In Tandem, the ANN trained to classify context-independent acoustic units is used to extract features for HMM/GMM system. This is indicated through (CI+), (ANN+) and (LI+) notation. *Det* denotes lexical model is deterministic and *Prob* denotes lexical model is probabilistic.

- Non-native accented speech recognition task that lacks both acoustic resources and "well developed" phonetic lexical resources (typically, phone lexicon includes native speaker pronunciations). In the literature, non-native accented ASR research mainly has focused on acoustic model adaptation. We investigate it on English where grapheme-to-phoneme relationship is irregular.

- Rapid development of ASR system for a new language that is not present in language-independent data using minimal acoustic and lexical resources. We demonstrate this aspect on Greek ASR task.

- Development of ASR system for a minority language, particularly, Scottish Gaelic which has only 60,000 speakers, lacks sufficient acoustic resources and does not have any phonetic lexical resources. The grapheme-to-phoneme relationship of Gaelic is regular, and many-to-one as the number of graphemes in a word is significantly higher than number of phones (Rasipuram et al., 2013a).

We compare systems based on probabilistic lexical modeling approaches described in Section 4 with standard HMM-based systems with different capabilities. Table 1 provides an overview of the systems that are investigated. The non-native and minority language ASR studies build on top of our preliminary investigations that focussed on KL-HMM and the use of word-internal context-dependent subword units (Imseng et al., 2011; Rasipuram et al., 2013a).

5.1. Databases and Setup

In this section, we describe the different databases and the setup of the systems used.

5.1.1. Language-Independent Dataset

A part of SpeechDat(II) corpus which contains gender, dialect and age balanced native speech from multiple language speakers, more precisely, British English, Italian, Spanish, Swiss French and Swiss German, is used as language independent dataset. Each language has approximately 12 hours of speech data, totally amounting to 63 hours. All the SpeechDat(II) lexicons use SAMPA symbols. A multilingual phone set of 117 units obtained by merging phones that share the same symbols across the above mentioned five languages, serves as the acoustic (or the subword) unit set.

5.1.2. Non-native HIWIRE

HIWIRE corpus contains utterances spoken by natives of France (31 speakers), Greece (20 speakers), Italy (20 speakers) and Spain (10 speakers) (Segura et al., 2007). The utterances contain spoken pilot orders made of 133 words. The database provides grammar with a perplexity of 14.9. The HIWIRE task does not have training data. It only includes adaptation data (50 utterances per speaker, approx. 150 min) and test data (50 utterances per speaker, approx 150 min). To simulate limited resources the amount of adaptation data is continuously reduced from 150 min to 3 min (specifically, 150 min, 120 min, 90 min, 64 min, 32 min, 16 min, 10 min, 3 min) by picking a subset of utterances as in (Imseng et al., 2011). The grapheme-based lexicon was transcribed using 27 graphemes (26 English alphabets, and silence).

A noticeable difference between (Imseng et al., 2011) and this paper is that in the previous work lexicon based on ARPABET phone set supplied with HIWIRE corpus was used where as in this work we use lexicon based on SAMPA phone set. The lexicon based on SAMPA phone set was created by borrowing pronunciations of 102 words that are in common from the SpeechDat(II) English lexicon. For the remaining 31 words, we obtained pronunciations by mapping ARPABET phones to SAMPA phones. The main reason to use SAMPA phone set based lexicon in this work is to have a shared subword units set between out-of-domain lexicon and target domain lexicon. This allowed the evaluation of acoustic model adaptation based systems (MAP and MLLR) discussed later in Section 5.2.2. Also, native English is present in out-of-domain resources. Therefore, in the case of KL-HMM, SP-HMM and tied approaches, the lexical model parameters trained on SpeechDat(II) English are adapted using HIWIRE adaptation data. Additionally, we could also investigate the case where no lexical model or acoustic model adaptation is performed.

5.1.3. Greek SpeechDat(II)

The experimental setup is based on (Imseng et al., 2012). Training set, development set and test set contains 13.5 hours of speech (1500 speakers), 1.5 hours of speech (150 speakers) and 6.9 hours of speech (350 speakers), respectively. Two optimistic language models, one from the sentences in the development set and other from the sentences in the test set are built. The phone lexicon is transcribed in SAMPA phone set using 31 phones (including silence). To simulate limited resources, the amount of available data was continuously reduced from 13.5 hours to 5 minutes (specifically, 800 min, 300 min, 150 min, 75 min, 37 min, 18 min, 9 min, 5 min). All the systems were evaluated on the same test set. The test set contains 10k unique words. The performance of phone-based KL-HMM, MAP, MLLR and HMM/GMM systems presented in (Imseng, 2013, Figures 4.3 and 4.4) is taken as reference in this paper.

As this study focusses on grapheme-based ASR systems, grapheme lexicon was developed using 25 graphemes that includes 24 Greek alphabets and silence. The acoustic model adaptation systems impose the constraint that subword unit sets of language-independent data and target language data match. As a result, grapheme-based acoustic model adaptation systems were not directly applicable to Greek ASR task, as Greek graphemes are different from Roman graphemes. This necessitated transliteration of Greek alphabets in terms of English (Roman) alphabets, as given in (Rasipuram et al., 2013b, Table 1), for grapheme-based acoustic model adaptation systems described later in Section 5.2.2.

5.1.4. Scottish Gaelic

The Scottish Gaelic speech corpus was collected by CSTR, University of Edinburgh. The experimental setup is similar to (Rasipuram et al., 2013a). Corpus consists of speech from 46 speakers. The training set consists of 22 speakers, 2389 utterances amounting to 3 hours of speech, the development set consists of 12 speakers, 1112 utterances amounting to 1 hour of speech and the test set consists of 12 speakers, 1317 utterances amounting to 1 hour of speech. The speakers in training data, development data and test data are different. The vocabulary size is 5k unique words. The database does not contain phone pronunciation lexicon. The grapheme-based lexicon containing 83 graphemes (5 vowels, 5 long vowels, 23 broad consonants, 23 slender consonants, 26 consonants and silence) is obtained by considering broad and slender Gaelic consonants as separate graphemes. We refer to this lexicon as *knowledge-based* grapheme lexicon.

In this study, we also investigate a grapheme lexicon that does not use any knowledge, such as broad and slender consonants. We refer to it as *orthography-based* lexicon. This lexicon is transcribed in traditional way from the orthography of words and includes 32 Gaelic graphemes (25 alphabets, 5 accents and silence).

Table 2 summarizes the information about different corpora used.

5.2. Systems

In this section, we provide details about different systems given in Table 1 by grouping them into three categories.

Corpus (Description)	Language	# of Subword units		Train data	Test data
		Phones	Graphemes	(in min)	(in min)
SpeechDat(II) (Native speech sampled at 8K used to train the acoustic model)	English	45	27	744	n.a
	French	42	43	810	n.a
	German	59	42	846	n.a
	Italian	52	34	690	n.a
	Spanish	32	34	690	n.a
(data used to train *multilingual acoustic model*)		*117*	*47*	*3780*	n.a
HIWIRE (Non-native speech from natives of France, Spain, Italy and Greece)	English	42	27	0 to 150	150
SpeechDat(II) (Native Greek speech)	Greek	31	25	5 to 800	360
Scottish Gaelic (Broadcast news data)	Scottish Gaelic	n.a.	83 or 32	180	60

Table 2: Overview of the tasks and the respective corpora used in the study

5.2.1. Probabilistic Lexical Modeling based Systems

We use an off-the-shelf three layer multilingual multilayer perceptron (MLP) trained on language-independent dataset to classify 117 context-independent multilingual phones as acoustic model. More recently, MLPs with deep architectures classifying context-dependent clustered phone units have gained lot of attention (Hinton et al., 2012). In the present work, we use the off-the-shelf MLP for the following reasons,

- The exactly same off-the-shelf MLP has been used in the previous studies on the ASR tasks described earlier (Imseng et al., 2011, 2012). Therefore, the results from the present study are directly comparable to the previous studies.
- In recent work, it has been shown that KL-HMM retains its benefit over standard hybrid HMM/ANN system even when MLP that classifies clustered context-dependent phone units is used (Imseng et al., 2013; Razavi et al., 2014).

The use of deep MLP architectures and context-dependent acoustic units in probabilistic lexical modeling framework is open for further research. Lexical model is trained for each of the probabilistic lexical modeling systems, namely, KL-HMM, SP-HMM and Tied-HMM as described earlier in Section 4. We used S_{RKL} as the local score for the KL-HMM system based on more recent investigations (Rasipuram and Magimai.-Doss, 2013a; Imseng et al., 2012; Rasipuram et al., 2013a).

5.2.2. Acoustic model adaptation based systems

We present ASR systems based on standard MAP and MLLR adaptation techniques. For this purpose, multilingual context-dependent phone-based and grapheme-based HMM/GMM systems were trained on the language-independent data set. The phone-based HMM/GMM system used multilingual phones as subword units.

All the five considered European languages use Roman alphabet. Therefore, multilingual grapheme-based HMM/GMM system was developed by forming multilingual grapheme set of 47 units by merging graphemes that are common across the languages in language-independent data set. Accents and diacritics are treated as separate graphemes.

Each context-dependent subword unit was modeled using 3 HMM states and each HMM state is modeled using mixture of 16 Gaussians. Then, MAP adaptation or MLLR adaptation is performed using speech data from the target language or domain. As described earlier in Section 5.1.3, for Greek task transliterated grapheme based lexicon was used while performing MAP or MLLR adaptation.

5.2.3. Standard language-dependent acoustic model and lexical model based ASR systems

These are HMM/GMM ASR systems where both acoustic model and lexical model are trained on language-dependent data. We investigate two systems, the first system uses standard cepstral features as feature observations (HMM/GMM system) and the second system uses Tandem features as feature observations (Tandem system) (Hermansky et al., 2000). As indicated in Table 1, Tandem system exploits both language-dependent and language-independent resources similar to probabilistic lexical model based systems and acoustic model adaptation based systems.

The Tandem features were extracted by transforming 117 dimensional output of the same multilingual MLP described earlier in Section 5.2.1, with log transformation followed by principal component analysis. The dimensionality of the output features is either kept the same or reduced to 39 dimensions.

The HMM/GMM systems used 39 dimensional PLP cepstral features ($c_0 - c_{12} + \Delta + \Delta\Delta$) as acoustic features. All the phone subword based systems use phonetic question set and grapheme subword based systems use singleton question set for decision tree state tying procedure. The number of mixture components for each of the tasks and the training conditions were tuned on the development set. Additionally, for tandem systems, the dimensionality of the feature observations (either 117 dimensions or 39 dimensions) was also tuned on the development set. HTK toolkit was used to build all HMM/GMM systems (Young et al., 2006).

6. Results

The present section is organized as follows. First, we present results on rapid development of ASR with both acoustic and lexical resource constraints on HIWIRE and Greek ASR tasks. Later, we present results on minority language speech recognition using Scottish Gaelic task. The performance of all the systems is reported in terms of word accuracy.

System	3 min		10 min		120 min		150 min	
	Graph	Phone	Graph	Phone	Graph	Phone	Graph	Phone
KL-HMM	90.7	93.3	94.0	94.6	98.0	98.0	98.1	98.1
SP-HMM	91.4	93.3	92.1	94.2	95.0	95.6	95.0	95.6
Tied-HMM	86.4	92.5	88.6	93.2	94.3	95.3	94.4	95.4
MAP	86.7	91.6	88.9	92.6	96.7	97.9	96.9	98.0
MLLR	86.2	92.4	87.3	94.3	92.2	96.0	91.9	96.0
Tandem	39.5	55.3	68.9	85.4	95.4	96.2	95.9	96.5
HMM/GMM	26.7	48.3	64.8	82.6	95.8	96.6	96.4	96.8

Table 3: Performance in terms of word accuracy on HIWIRE test set for various systems using various amounts of HIWIRE adaptation data

System	5 min		37min		300 min		800 min	
	Graph	Phone	Graph	Phone	Graph	Phone	Graph	Phone
KL-HMM	78.0	80.3	81.4	83.0	83.8	84.4	84.5	84.8
SP-HMM	71.3	73.8	75.9	76.3	77.8	79.3	78.7	79.6
Tied-HMM	66.6	68.6	71.3	73.6	74.8	76.3	76.4	77.6
MAP	54.7	77.4	68.7	79.3	78.0	82.7	78.0	83.9
MLLR	50.0	77.3	52.6	78.7	52.8	79.1	52.8	78.7
Tandem	55.7	66.9	76.0	79.7	81.6	83.8	82.4	84.9
HMM/GMM	54.6	63.5	74.5	81.2	82.3	84.5	83.5	85.2

Table 4: Performance in terms of word accuracy on Greek test set for various systems using various amounts of Greek data

6.1. Rapid ASR development

Tables 3 and 4 summarize the performance in terms of word accuracy on HIWIRE and Greek tasks for various amounts of language-dependent training data for KL-HMM, SP-HMM, Tied-HMM, Tandem, MAP, MLLR and HMM/GMM systems. The results are analysed using Figures 2 and 3 along two aspects, namely, comparison of different probabilistic lexical model based systems (Section 6.1.1), comparison of probabilistic lexical model based systems against acoustic model adaptation based systems and standard HMM/GMM systems (Section 6.1.2).

6.1.1. Probabilistic Lexical Modeling based Systems

Figures 2(a) and 2(b) present the performance on HIWIRE and Greek tasks respectively, for phone- and grapheme-based KL-HMM, SP-HMM and Tied posterior systems with increasing amount of training data. The figures show that KL-HMM system consistently performs better compared to SP-HMM and

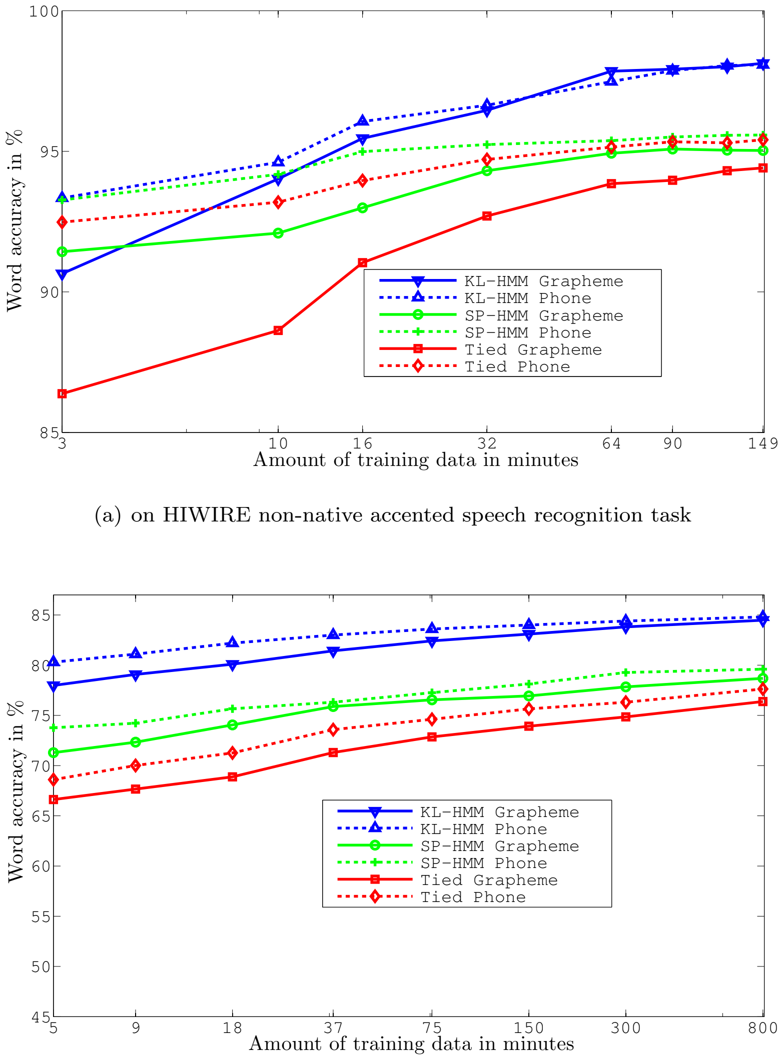
Tied-HMM systems for both phone and grapheme subword units. Furthermore, on HIWIRE task the difference is more pronounced when the systems use graphemes as subword units.

6.1.2. Comparison of probabilistic lexical modeling based system with other Systems

Figures 3(a) and 3(b) plot the performance on HIWIRE and Greek tasks respectively, with varying amount of training data for phone-based and grapheme-based KL-HMM, MAP, MLLR, Tandem and HMM/GMM systems. We can draw the following inferences from the figures,

1. KL-HMM based systems irrespective of the type of subword units used, phones or graphemes, tend to perform better than (when the training data is less) or comparable (when training data is increased) to phone-based or grapheme-based deterministic lexical model based systems. On both HIWIRE and Greek tasks, the difference in performance between phone and grapheme-based systems is minimal for KL-HMM approach compared to all other approaches.
2. On both HIWIRE (where grapheme-to-phoneme relationship is irregular) and Greek (where grapheme-to-phoneme relationship is regular) tasks it can be been observed that deterministic lexical model based systems are more suitable for phones than graphemes.
   On HIWIRE task where lexical units and acoustic units match or have shared unit set acoustic model adaptation based systems perform better than HMM/GMM or Tandem systems. However, the performance of acoustic model adaptation systems using graphemes is worse than with phones as subword units. On Greek task where transliterated grapheme based lexicon was used for acoustic model adaptation, grapheme-based systems perform significantly worse compared to phone-based acoustic model adaptation or HMM/GMM or Tandem systems. The results also show that in case of grapheme subword unit set mismatch, transliteration may not be the best possible alternative. In such cases, data driven mapping of grapheme subword units could potentially be investigated (Stüker, 2008b).
   When the available training data is larger, phone-based deterministic lexical model systems for both HIWIRE and Greek tasks perform comparable to phone-based KL-HMM system (though not the same technique, for e.g., in HIWIRE it is MAP and in Greek it is HMM/GMM and Tandem). However, in case of grapheme-based systems, this trend is not observed. The results, inline with the other multilingual grapheme-based ASR studies (Kanthak and Ney, 2003; Killer et al., 2003; Stüker, 2008a) show that the use of multilingual grapheme models across languages does not appear evident, since the relationship between graphemes and phones may differ considerably across languages.
3. Monolingual HMM/GMM systems and acoustic model adaptation based systems with shared unit set (i.e., on HIWIRE task) that exploit multilingual speech tend to converge with the increase in acoustic resources.

Figure 2: Comparison between probabilistic lexical modeling based systems with increasing amount of target domain or language training data



(a) on HIWIRE non-native accented speech recognition task

(b) on Greek ASR task

4. Compared to HMM/GMM approach, Tandem approach is beneficial mainly in low acoustic resource conditions.
5. Comparing MAP and MLLR approaches, it can be observed that MLLR is better than MAP mainly in very low acoustic resource conditions.

As mentioned in Section 5.1.2, it is possible to directly decode HIWIRE test set using language-independent acoustic and lexical models without any adaptation. Table 6.1.2 presents the performance on HIWIRE task for KL-HMM, SP-HMM, Tied-HMM and language-independent HMM/GMM systems. The lexical model for KL-HMM, SP-HMM and Tied-HMM systems is trained on SpeechDat(II) English data. It can be observed that for both phone and grapheme subword units KL-HMM system performs better than SP-HMM, Tied-HMM and LI HMM/GMM systems. Also, it is interesting to note that irrespective of subword units used, the performance of all the probabilistic lexical model based systems (that use context-independent phones as acoustic units) is better than LI HMM/GMM system (that use context-dependent phones as acoustic units).

System	Grapheme	Phone
KL-HMM	90.0	94.0
SP-HMM	87.3	93.2
Tied-HMM	86.0	91.6
LI HMM/GMM	84.2	91.3

Table 5: Performance in terms of word accuracy on HIWIRE test set using system trained on SpeechDat(II) data. LI HMM/GMM system refers to multilingual HMM/GMM system trained on language-independent (LI) data

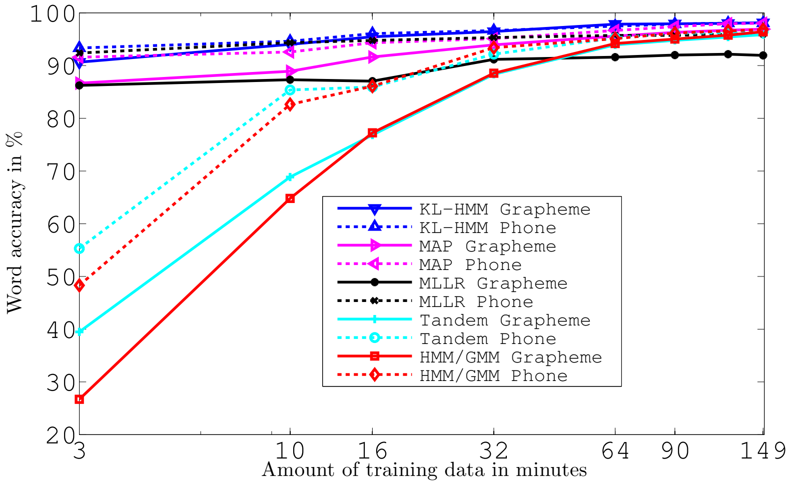
6.2. Scottish Gaelic ASR

Table 6 presents the performance on test set of Scottish Gaelic corpus for KL-HMM, SP-HMM, Tied-HMM, Tandem and HMM/GMM systems for *orthography-based* and *knowledge-based* grapheme lexicons. MAP system was not investigated for knowledge-based lexicon due to the mismatch between acoustic unit set and lexical unit set. It can be observed that the systems using *knowledge-based* grapheme lexicon perform better than systems using *orthography-based* grapheme lexicon. This shows that integrating orthographic knowledge specific to language in grapheme lexicon can help in improving the performance of grapheme-based ASR system. KL-HMM systems perform better than all other systems. Tandem system performs better than HMM/GMM system. Furthermore, MAP, SP-HMM and Tied-HMM systems perform worse compared to Tandem and HMM/GMM systems. Finally, in the case of orthography-based lexicon MAP system is not able to capitalize from the language-independent data.

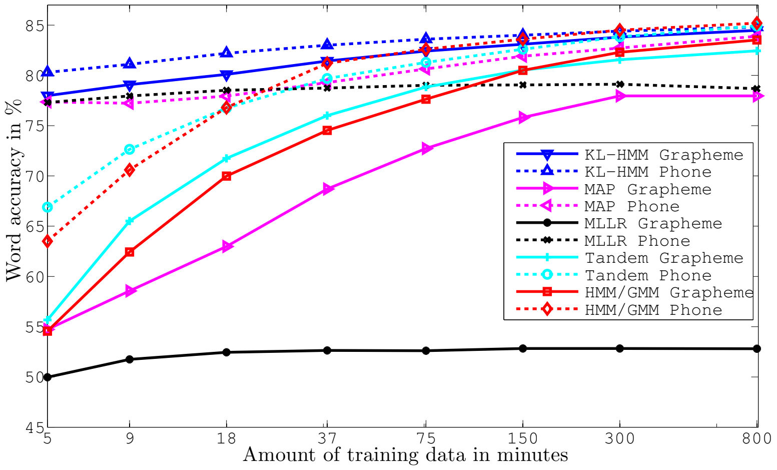
6.3. Analysis

From the experiments presented earlier in this section, it can be observed that despite using exactly same acoustic model, the performance trends of dif-

Figure 3: Comparison of phone-based and grapheme-based KL-HMM systems against acoustic model adaptation based systems and standard HMM/GMM system with increasing amount of target domain or language training data



(a) on HIWIRE non-native accented speech recognition task



(b) on Greek ASR task

System	*Orthography-based* lexicon	*Knowledge-based* lexicon
KL-HMM RKL	67.9	72.7
SP-HMM	52.0	56.7
Tied-HMM	54.5	59.7
MAP	55.1	–
Tandem	66.5	69.9
HMM/GMM	64.2	68.0

Table 6: Performance in terms of word accuracy on Gaelic test set for various systems

ferent probabilistic lexical modeling approaches KL-HMM, SP-HMM and Tied-HMM are different. KL-HMM system performs better than deterministic lexical model based systems in both under-resourced and well resourced conditions. While, SP-HMM and Tied-HMM systems show gains over the deterministic lexical model based systems mainly in under-resourced conditions (see Tables 3 and 4). We attribute the superiority of KL-HMM system to its abilities discussed in Section 4.4 such as, being able to give more importance to acoustic model than lexical model through the use of local score $S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$, discriminative local score etc.

In order to ascertain it, we conducted a study on HIWIRE task with 150 minute target data condition where the lexical model of KL-HMM system trained using local score S_{RKL} is used and decoding is performed with different local scores, namely, $S_{KL}(\mathbf{y}_i, \mathbf{z}_t)$, $S_{SKL}(\mathbf{y}_i, \mathbf{z}_t)$, $S_{tied}(\mathbf{y}_i, \mathbf{v}_t)$ and $S_{SP}(\mathbf{y}_i, \mathbf{z}_t)$. The study was conducted for both grapheme-based and phone-based systems. Table 7 presents the results of this study.

Local score for decoding	grapheme	phone
$S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$	98.1	98.1
$S_{KL}(\mathbf{y}_i, \mathbf{z}_t)$	97.8	97.6
$S_{SKL}(\mathbf{y}_i, \mathbf{z}_t)$	98.1	98.1
$S_{SP}(\mathbf{y}_i, \mathbf{z}_t)$	96.5	96.7
$S_{tied}(\mathbf{y}_i, \mathbf{z}_t)$	97.3	97.1

Table 7: Comparison across different local scores used during decoding. All the system used KL-HMM trained lexical model with local score $S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$.

It can be observed that decoding with KL-divergence based local scores $S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$, $S_{SKL}(\mathbf{y}_i, \mathbf{z}_t)$ and $S_{KL}(\mathbf{y}_i, \mathbf{z}_t)$ results in better performance compared to decoding with $S_{SP}(\mathbf{y}_i, \mathbf{z}_t)$ and $S_{tied}(\mathbf{y}_i, \mathbf{v}_t)$ local score, ascertaining the fact that KL-divergence is a discriminative local score. Furthermore, decoding with $S_{KL}(\mathbf{y}_i, \mathbf{z}_t)$, $S_{SP}(\mathbf{y}_i, \mathbf{z}_t)$ and $S_{tied}(\mathbf{y}_i, \mathbf{v}_t)$ yields lower performance than decoding with $S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$. However, decoding with $S_{SKL}(\mathbf{y}_i, \mathbf{z}_t)$ that gives equal importance to both acoustic and lexical model yields performance similar to $S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$. It can also be noted that decoding the KL-HMM lexical model with $S_{SP}(\mathbf{y}_i, \mathbf{z}_t)$ and $S_{tied}(\mathbf{y}_i, \mathbf{v}_t)$ results in better performance compared to SP-HMM trained and Tied-HMM trained lexical model, respectively (see Table 3). This indicates that KL-HMM approach with local score S_{RKL} is yielding a better lexical model compared to SP-HMM or Tied-HMM approaches. Deeper investigations on these aspects is out-of-the-scope of the present paper.

6.4. Comparisons to Literature

In the literature, there are studies that have been reported on HIWIRE task (Segura et al., 2007; Gemello et al., 2007). Despite using the same adaptation and test sets, the studies reported in this paper and the literature differ in terms of the sampling frequency of speech data, type and amount of the out-of-domain data used. First, we compare studies where any kind of adaptation was not performed,

- in (Segura et al., 2007), TIMIT trained monophone HMM/GMM system without adaptation was found to achieve performance of 91.4% word accuracy.
- in (Gemello et al., 2007), monophone hybrid HMM/ANN system using MLP trained on TIMIT, WSJ0, WSJ1 and Vehiclus-ch0 corpora was found to achieve performance of 90.5% word accuracy. Furthermore, monophone hybrid HMM/ANN system using MLP trained on LDC Macrophone and SpeechDat Mobile corpora and HIWIRE speech downsampled to 8kHz was found to achieve performance of 88.4% word accuracy.

As shown in Table 6.4, the phone-based KL-HMM system performs better compared to the studies reported in the literature and grapheme-based KL-HMM system performs comparable compared to studies reported in the literature. It can also be observed from Tables 6.4 and 6.1.2 that the phone-based LI HMM/GMM system performs similar to the above mentioned systems from the literature, where as the grapheme-based LI HMM/GMM system performs worse.

System	Out-of-domain data	Sampling frequency	Performance
HMM/GMM	TIMIT	16kHz	91.4
Hybrid HMM/ANN	TIMIT, WSJ0, WSJ1, Vehiclus-ch0	16kHz	90.5
Hybrid HMM/ANN	LDC Macrophone, SpeechDat Mobile	8kHz	88.4
KL-HMM Grapheme	SpeechDat(II)	8kHz	90.0
KL-HMM Phone	SpeechDat(II)	8kHz	94.0

Table 8: Comparison of the performance in terms of word accuracy on HIWIRE test set without any adaptation.

There are also studies on HIWIRE that report results with acoustic model adaptation where 150 min of HIWIRE adaptation data is used,

- in (Segura et al., 2007), it has been found that TIMIT trained HMM/GMM system with MLLR adaptation achieves performance of 97.25% word accuracy.
- in (Gemello et al., 2007), linear hidden network (LHN) based adaptation in hybrid HMM/ANN framework achieved performance of 98.2% on 16kHz sampled HIWIRE data. MLP trained on data from TIMIT, WSJ0, WSJ1 and Vehiclus-ch0 was adapted on HIWIRE data using LHN.

As shown in the Table 6.4, hybrid HMM/ANN system using LHN based adaptation performs similar to phone-based and grapheme-based KL-HMM systems. In (Imseng et al., 2011), on HIWIRE task the performance of grapheme-based KL-HMM system using low amounts of HIWIRE adaptation data (3min, 10min) was significantly poor compared to phone-based KL-HMM system. However, in this work the gap has significantly reduced as the lexical model parameters trained on SpeechDat(II) English are adapted using HIWIRE adaptation data.

System	Out-of-domain data	Sampling frequency	Performance
MLLR	TIMIT	16kHz	97.25
LHN	TIMIT, WSJ0, WSJ1, Vehiclus-ch0	16kHz	98.2
KL-HMM Grapheme	SpeechDat(II)	8kHz	98.1
KL-HMM Phone	SpeechDat(II)	8kHz	98.1

Table 9: Comparison of the performance in terms of word accuracy on HIWIRE test set with adaptation

In the case of Greek task, as previously mentioned phone-based KL-HMM, MLLR, MAP, and HMM/GMM systems reported in (Imseng et al., 2012) and (Imseng, 2013, Figure 4.3 in Page 59 and Figure 4.4 in Page 60) have been used as reference. However, the phone-based Tandem systems reported in (Imseng, 2013) and this work differ. Unlike (Imseng, 2013), in our studies the dimensionality of Tandem features was either 117 dimensions (all the dimensions) or 39 dimensions (same as the dimension of standard cepstral feature vector). The dimension of features was tuned on the development set for each of the training condition. We found dimensionality reduction to be beneficial, especially in the low acoustic resource conditions. For example, on 5 min acoustic resource case, preformance of phone-based Tandem system reported in (Imseng, 2013) was 30.2% word accuracy where as in this work with reduced feature dimensionality we achieved 66.9% word accuracy.

In the previous study on Scottish Gaelic ASR (Rasipuram et al., 2013a), knowledge-based grapheme lexicon that tagged word beginning and end graphemes was used and word-internal context-dependent graphemes were modeled. The KL-HMM and HMM/GMM systems achieved word accuracy of 72.8% and 64.8%, respectively. In this work, the same knowledge-based grapheme lexicon was used but without any word begin and end tags. As a result, the total number of grapheme subword units is less. Furthermore, in this paper we modeled cross-word context-dependent subword based systems. As it can be seen from Table 6, the knowledge-based HMM/GMM system yields 3.2% absolute improvement in WER compared to previous work and grapheme KL-HMM system achieves performance comparable to the previous study.

7. Discussion and Conclusion

In this work, we showed that ASR systems can be rapidly built using language-independent acoustic model and training only the lexical model on small amount of target language data. In a recent work (Rasipuram et al., 2013b), we have shown that the lexical model can be completely knowledge driven and ASR systems could be developed for new languages without using any acoustic and lexical resources from the language, i.e., (near) zero resource ASR system.

In this work, we compared probabilistic lexical model systems where only lexical model is trained on target language data with deterministic lexical model based systems where either acoustic model is adapted on target language data or both acoustic model and lexical model are trained on target language data. In our studies we observed that with increase in target language acoustic data, the gap between KL-HMM system and acoustic model adaptation based systems reduces. This suggests that there may be benefits in combining acoustic model adaptation and probabilistic lexical modeling.

- When using ANN-based acoustic model, this can be achieved by training a hierarchical neural network (Pinto et al., 2011) or adapting a neural network with target language data (Swietojanski et al., 2012). A recent study on Scottish Gaelic in the framework of KL-HMM has shown the potential of acoustic model adaptation using hierarchical neural network approach (Rasipuram et al., 2013a).

- KL-HMM approach is not restricted to ANN-based acoustic modeling alone (Rasipuram and Magimai.-Doss, 2013b). Therefore, using GMMs as acoustic model this can be achieved by adapting the GMMs through MAP technique (as done in the present paper) followed by KL-HMM training.

As mentioned earlier in Section 3, in deterministic lexical modeling framework, acoustic model adaptation and lexical model adaptation can be combined in different ways. For instance, (a) by jointly learning acoustic model parameters (GMMs) and probabilistic lexical model parameters in the framework of HMM/GMM systems as in (Luo and Jelinek, 1999), or (b) by combining acoustic model adaptation with polyphone decision tree state tying (PDTS) (Schultz and Waibel, 2001), or (c) using SGMM approach (Burget et al., 2010). Comparing probabilistic lexical modeling and deterministic lexical modeling along these lines with graphemes as subword units is part of our future work.

In conclusion, our studies showed that with probabilistic lexical modeling especially using KL-HMM approach, ASR systems can be rapidly developed for new languages and domains by training language or domain independent acoustic model and learning the grapheme-to-phone relationship on small amount of target language or domain data. In doing so, we not only address the lack of acoustic resource (speech data with transcription) problem but also the lack of lexical resource (phonetic pronunciation dictionary) problem.

Our studies, in addition to showing the efficacy of the proposed approach, also explicated that it is the constraints imposed by the deterministic lexical model that demand the availability of well-developed acoustic resources and phonetic lexical resources from the target language. Furthermore, our investigations also showed that deterministic lexical model based ASR approaches are more suitable for phone-based ASR than grapheme-based ASR, while probabilistic lexical model based ASR approach is suitable for both.

Appendix A. Parameter Estimation of Probabilistic Lexical Model Approaches

Given a trained ANN and training set of N utterances $\{X(n), W(n)\}_{n=1}^{N}$ where for each training utterance n, $X(n)$ represents sequence of cepstral features of length $T(n)$ and $W(n)$ represents the sequence of underlying words, the set of acoustic state probability vectors $\{Z(n), W(n)\}_{n=1}^{N}$ or the set of likelihood vectors $\{V(n), W(n)\}_{n=1}^{N}$ are estimated where $Z(n)$ represents a sequence of acoustic state probability vectors of length $T(n)$, $V(n)$ represents a sequence of acoustic likelihood probability vectors of length $T(n)$.

The KL-HMM system is parameterized by $\Theta_{kull} = \{\{\mathbf{y}_i\}_{i=1}^{I}, \{a_{ij}\}_{i,j=1}^{I}\}$. The training data $\{Z(n), W(n)\}_{n=1}^{N}$ and the current parameter set Θ_{kull}, are used to estimate the new set of parameters $\hat{\Theta}_{kull}$ by Viterbi expectation maximization algorithm which minimizes the cost function,

$$\hat{\Theta}_{kull} = \underset{\Theta_{kull}}{\arg\min} \Big[\sum_{n=1}^{N} \min_{Q \in \mathcal{Q}} \sum_{t=1}^{T(n)} \big[S_{RKL}(\mathbf{y}_{q_t}, \mathbf{z}_t(n)) - \log a_{q_{t-1}q_t}\big]\Big] \tag{A.1}$$

The parameters of the tied approach $\Theta_{tied} = \{\{\mathbf{y}_i\}_{i=1}^{I}, \{a_{ij}\}_{i,j=1}^{I}\}$ are estimated by Viterbi expectation maximization algorithm that maximizes the cost function,

$$\hat{\Theta}_{tied} = \underset{\Theta_{tied}}{\arg\max} \Big[\sum_{n=1}^{N} \max_{Q \in \mathcal{Q}} \sum_{t=1}^{T} \big[S_{tied}(\mathbf{y}_{q_t}, \mathbf{v}_t(n)) + \log(a_{q_{t-1}q_t})\big]\Big] \tag{A.2}$$

where $Q = \{q_1, \cdots q_t, \cdots, q_{T(n)}\}$, $q_t \in \{1, \cdots, I\}$ and $\mathcal{Q}$ denotes set of all possible HMM state sequences.

The training process involves iteration over the segmentation and the optimization steps until convergence. Given current set of parameters, the segmentation step yields an optimal state sequence for each training utterance using Viterbi algorithm. Given optimal state sequences and acoustic state posterior vectors belonging to each of these states, the optimization step then estimates new set of model parameters by minimizing Eqn. (A.1) or maximizing (A.2) subject to the constraint that $\sum_{d=1}^{D} y_i^d = 1$.

For local score S_{RKL} (Equation (12)), the optimal lexical state distribution is the arithmetic mean of the training acoustic state probability vectors assigned to the state, i.e.,

$$y_i^d = \frac{1}{M(i)} \sum_{\mathbf{z}_t(n) \in Z(i)} z_t^d(n) \quad \forall d \tag{A.3}$$

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

The optimal state distribution for tied approach is,

$$y_i^d = \frac{1}{M(i)} \sum_{\mathbf{v}_t(n) \in V(i)} \frac{y_i^d . v_t^d(n)}{\sum_{d=1}^{D} y_i^d . v_t^d(n)} \quad \forall d \tag{A.4}$$

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

SP-HMM is a special case of tied approach with optimal state distribution as,

$$y_i^d = \frac{1}{M(i)} \sum_{\mathbf{z}_t(n) \in Z(i)} \frac{y_i^d . z_t^d(n)}{\sum_{d=1}^{D} y_i^d . z_t^d(n)} \quad \forall d \tag{A.5}$$

Acknowledgment

This work was supported by the Swiss NSF through the grants "Flexible Grapheme-Based Automatic Speech Recognition (FlexASR) " and the National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management" (www.im2.ch). The authors would like to thank their colleagues at Idiap, especially, Dr. David Imseng for his help with Greek task and Marzieh Razavi. The paper has benefitted from the discussions authors had with Prof. Tanja Schultz and her group.

References

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

H. Bourlard, N. Morgan, Connectionist Speech Recognition - A Hybrid Approach, Kluwer Academic Publishers, 1994.

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

J. Kohler, Language Adaptation of Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks, in: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, 417–420 vol.1, 1998.

P. Beyerlein, W. Byrne, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, W. Wang, Towards Language Independent Acoustic Modeling, in: Proc. of ICASSP, 1029–1032, 2000.

T. Schultz, A. Waibel, Language-independent and language-adaptive acoustic modeling for speech recognition, Speech Communication 35 (2001) 31–51.

V.-B. Le, L. Besacier, Automatic Speech Recognition for Under-Resourced Languages: Application to Vietnamese Language, IEEE Trans. on Audio, Speech, and Language Processing 17 (2009) 1471–1482.

L. Burget, et al., Multilingual Acoustic Modeling for Speech Recognition based on Subspace Gaussian Mixture Models, in: Proc. of ICASSP, 4334–4337, 2010.

E. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, S. Rieck, Automatic Speech Recognition Without Phonemes, in: Proc. of European Conf. on Speech Communication and Technology (EUROSPEECH), 1993.

S. Kanthak, H. Ney, Context-Dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition, in: Proc. of ICASSP, 845–848, 2002.

M. Killer, S. Stüker, T. Schultz, Grapheme based Speech Recognition, in: Proc. of EUROSPEECH, 2003.

J. Dines, M. Magimai-Doss, A Study of Phoneme and Grapheme based Context-Dependent ASR Systems, in: Proc. of Machine Learning for Multimodal Interaction (MLMI), 215–226, 2007.

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

S. Stüker, Modified Polyphone Decision Tree Specialization for Porting Multilingual Grapheme Based ASR Systems to New Languages, in: Proc. of ICASSP, 4249–4252, 2008a.

S. Stüker, Integrating Thai Grapheme Based Acoustic Models into the ML-MIX Framework - For Language Independent and Cross-Language ASR, in: Proc. of SLTU, 2008b.

G. Aradilla, H. Bourlard, M. M. Doss, Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task , in: Proc. of Interspeech, 928–931, 2008.

R. Rasipuram, M. Magimai.-Doss, Probabilistic Lexical Modeling and Grapheme-based Automatic Speech Recognition, `http://publications.idiap.ch/downloads/reports/2013/Rasipuram_Idiap-RR-15-2013.pdf`, Idiap Research Report, 2013a.

D. Imseng, R. Rasipuram, M. Magimai.-Doss, Fast and Flexible Kullback-Leibler Divergence based Acoustic Modeling for Non-native Speech Recognition, in: Proc. of Automatic Speech Recognition and Understanding (ASRU), 348–353, 2011.

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

R. Rasipuram, P. Bell, M. Magimai.-Doss, Grapheme and Multilingual Posterior Features for Under-Resourced Speech Recognition: A Study on Scottish Gaelic, in: Proc. of ICASSP, 2013a.

M. Magimai.-Doss, R. Rasipuram, G. Aradilla, H. Bourlard, Grapheme-based Automatic Speech Recognition using KL-HMM, in: Proc. of Interspeech, 2693–2696, 2011.

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

S. Young, et al., The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, UK, 2006.

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

J. Novak, Phonetisaurus: A WFST-driven Phoneticizer, http://code.google.com/p/phonetisaurus/, 2011.

S. Kanthak, H. Ney, Multilingual Acoustic Modeling using Graphemes, in: Proc. of EUROSPEECH, 1145–1148, 2003.

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

H. Hermansky, D. Ellis, S. Sharma, Tandem Connectionist Feature Extraction for Conventional HMM Systems, in: Proc. of ICASSP, vol. 3, 1635–1638, 2000.

J. Rottland, G. Rigoll, Tied Posteriors: An Approach for Effective Introduction of Context Dependency in Hybrid NN/HMM LVCSR, in: Proc. of ICASSP, 1241–1244, 2000.

S. Soldo, M. Magimai.-Doss, J. P. Pinto, H. Bourlard, Posterior Features for Template-based ASR, in: Proc. of ICASSP, 4864–4867, 2011.

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

L. R. Bahl, F. Jelinek, R. Mercer, A Maximum Likelihood Approach to Continuous Speech Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-5 (2) (1983) 179–190.

H. Strik, C. Cucchiarini, Modeling pronunciation variation for ASR: A survey of the literature, Speech Communication 29 (1999) 225–246.

R. E. Blahut, Hypothesis Testing and Information Theory, IEEE Trans. on Information Theory IT-20 (4).

J. Segura, T. Ehrette, A. Potamianos, D. Fohr, I. Illina, P.-A. Breton, V. Clot, R. Gemello, M. Matassoni, P. Maragos, The HIWIRE Database, a Noisy and Non-native English Speech Corpus for Cockpit Communication, http://cvsp.cs.ntua.gr/projects/pub/HIWIRE/WebHome/HIWIRE_db_description_paper.pdf, 2007.

D. Imseng, Multilingual speech recognition A posterior based approach, http://publications.idiap.ch/downloads/papers/2013/Imseng_THESIS_2013.pdf, Ph.D. theisis, École Polytechnique Fédérale de Lausanne (EPFL), 2013.

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

G. Hinton, et al., Deep Neural Networks for Acoustic Modeling in Speech Recognition, IEEE Signal Processing Magazine 29 (6) (2012) 82–97.

D. Imseng, P. Motlicek, P. N. Garner, H. Bourlard, Impact of deep MLP architecture on different acoustic modeling techniques for under-resourced speech recognition, in: Proc. of ASRU, 2013.

M. Razavi, R. Rasipuram, M. Magimai-Doss, On Modeling Context-dependent Clustered States: Comparing HMM/GMM, Hybrid HMM/ANN and KL-HMM Approaches, http://publications.idiap.ch/downloads/reports/2013/Razavi_Idiap-RR-43-2013.pdf, To appear in Proc. of ICASSP, 2014.

R. Gemello, F. Mana, S. Scanzio, Experiments on Hiwire Database using Denoising and Adaptation with a Hybrid HMM-ANN Model, in: Proc. of Interspeech, 2429–2432, 2007.

J. P. Pinto, G. S. V. S. Sivaram, M. Magimai.-Doss, H. Hermansky, H. Bourlard, Analysis of MLP Based Hierarchical Phoneme Posterior Probability Estimator, IEEE Trans. on Audio, Speech, and Language Processing 19 (2011) 225–241.

P. Swietojanski, A. Ghoshal, S. Renals, Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR, in: Proc. of ICASSP, 246–251, 2012.

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