Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 68 (2015) 23–40

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

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]

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

Received 20 February 2014; received in revised form 25 November 2014; accepted 19 December 2014
Available online 29 December 2014

Abstract

One of the key challenges involved in building statistical automatic speech recognition (ASR) systems is modeling the relationship between subword units or "lexical units" and acoustic feature observations. To model this relationship two types of resources are needed, namely, acoustic resources i.e., speech data with word level transcriptions and lexical resources where each word is transcribed in terms of subword units. Standard ASR systems typically use phonemes or phones as subword units. However, not all languages have well developed acoustic and phonetic lexical resources. In this paper, we show that the relationship between lexical units and acoustic features can be factored into two parts through a latent variable, namely, an acoustic model and a lexical model. In the acoustic model the relationship between latent variables and acoustic features is modeled, while in the lexical model a probabilistic relationship between latent variables and lexical units is modeled. We elucidate that in standard hidden Markov model based ASR systems, the relationship between lexical units and latent variables is one-to-one and the lexical model is deterministic. Through a literature survey we show that this deterministic lexical modeling imposes the need for well developed acoustic and lexical resources from the target language or domain to build an ASR system. We then propose an approach that addresses both acoustic and phonetic lexical resource constraints in ASR system development. In the proposed approach, latent variables are multilingual phones and lexical units are graphemes of the target language or domain. We show that the acoustic model can be trained on domain-independent or language-independent resources and the lexical model that models a probabilistic relationship between graphemes and multilingual phones can be trained on a relatively small amount of transcribed speech data from the target domain or language. The potential and the efficacy of the proposed approach is demonstrated through experiments and comparisons with other approaches on three different ASR tasks: non-native and accented speech recognition, rapid development of an ASR system for a new language, and development of an ASR system for a minority language.

Keywords: Automatic speech recognition; Kullback–Leibler divergence based hidden Markov model; Grapheme subword units; Phoneme subword units; Lexical modeling; Pronunciation lexicon

1. Introduction

State-of-the-art automatic speech recognition (ASR) systems are based on hidden Markov models (HMMs). The development of an HMM-based ASR system is often decomposed into two problems (Rabiner, 1989; Bourlard and Morgan, 1994). First, the relationship between

[*] Corresponding author at: Idiap Research Institute, Martigny, Switzerland. Tel.: +41277217 711.
E-mail addresses: ramya.rasipuram@idiap.ch (R. Rasipuram), mathew@idiap.ch (M. Magimai-Doss).

http://dx.doi.org/10.1016/j.specom.2014.12.006
0167-6393/© 2014 Elsevier B.V. All rights reserved.

subword units or “lexical units” and acoustic feature observations is modeled. Second, the syntactic constraints of the language are modeled.

The present paper focuses on the first problem. To model the relationship between lexical units and acoustic features, transcribed speech data and a phonetic lexicon 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, the lack of transcribed speech data has been typically addressed through multilingual and crosslingual approaches (Kohler, 1998; Schultz and Waibel, 2001; Burget et al., 2010; Swietojanski et al., 2012; Huang et al., 2013). In these approaches, first the relationship between lexical units and acoustic feature observations is learned on domain- or language-independent data and later adapted on target language or domain data. If the phonetic lexicon in the target language is not available, then the use of alternate subword units such as graphemes has been explored (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 rarely been studied in the past (Stüker, 2008b; Stüker, 2008a). The focus of this paper is on building ASR systems for languages and domains that lack both a phonetic lexicon and transcribed speech data.

In this paper, we first show that the modeling of the relationship between lexical units and acoustic feature observations can be factored into two parts or models, namely, the acoustic model and the lexical model through a latent variable.

1. In the acoustic model, the relationship between latent variables and acoustic features is modeled.
2. In the lexical model, a probabilistic relationship between latent variables and lexical units is modeled.

We then elucidate that in standard HMM-based ASR systems the lexical model is *deterministic*. The deterministic lexical model imposes constraints such as: the latent variables 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 the acoustic model and the lexical model.

In recent work, we showed that there are approaches such as the Kullback–Leibler divergence-based hidden Markov model (Aradilla et al., 2008), where the relationship between lexical units and latent variables is probabilistic (Rasipuram and Magimai-Doss, 2013b). Probabilistic lexical modeling relaxes certain constraints imposed by deterministic lexical modeling. As a consequence, the acoustic and lexical models can be independently trained on different sets of resources. Further, different kinds of subword units can be modeled in an ASR system; and different types of contextual units can be modeled in an ASR system (Magimai-Doss et al., 2011; Imseng et al., 2011; Imseng et al., 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 the target language or domain. In the proposed approach:

- Latent variables 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.
- The lexical model that captures a probabilistic relationship between graphemes and multilingual phones, is trained on a relatively small amount of target language-dependent acoustic data.

The potential and efficacy of the proposed approach is demonstrated through experiments and comparisons with other standard approaches on three ASR tasks. The standard ASR approaches considered for comparison are the acoustic model adaptation and Tandem approaches that exploit language-independent resources, and the HMM/Gaussian mixture model (GMM) approach that uses only the target language data.

The paper is organized as follows: Section 2 provides a background on standard HMM-based ASR systems 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, their potential implications and the proposed approach. Sections 5 and 6 present the experimental setup and the results, respectively. Finally, in Section 7 we provide a discussion followed by a conclusion.

2. Background

In a statistical ASR approach, the goal is to find the best matching or the 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,

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

$$= \arg\max_{W \in \mathcal{W}} \frac{P(X|W, \Theta_A) \cdot P(W|\Theta_L)}{P(X|\Theta)} \tag{2}$$

$$= \arg\max_{W \in \mathcal{W}} P(X|W, \Theta_A) \cdot P(W|\Theta_L) \tag{3}$$

where $\mathcal{W}$ denotes the set of all possible word sequences. The first term on the right hand side of Eq. (3) is the likelihood of the acoustic observation sequence X given a word sequence W and is referred to as the acoustic likelihood. The second term on the right hand side of Eq. (3) is the prior probability of a word sequence W or the language model probability. The parameter set $\Theta = \{\Theta_A, \Theta_L\}$ includes the parameters of the acoustic likelihood estimator (Θ_A) and the parameters of the language model (Θ_L).

2.1. Standard HMM-based ASR

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

$$Q^* = \arg\max_{Q \in \mathcal{Q}} P(Q, X|\Theta) \tag{4}$$

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

$$= \arg\max_{Q \in \mathcal{Q}} \sum_{t=1}^{T} [\log p(\mathbf{x}_t|q_t = l^i, \Theta_A) + \log P(q_t = l^i|q_{t-1} = l^j, \Theta)] \tag{6}$$

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, $q_t \in \mathcal{L} = \{l^1, \ldots l^i \ldots l^I\}$ and I is the number of lexical units. In a subword unit based ASR system, if phones are used as subword units then each lexical unit l^i represents a phone or a polyphone. If graphemes are used as subword units then each lexical unit l^i represents a grapheme or a polygrapheme.

Eq. (5) arises from the HMM and language model assumptions. The two HMM assumptions are: (1) the output observation at time t is dependent only on the current state and (2) the first order Markov assumption which states that the current state is dependent only on the previous state. If l^j is the last lexical unit of a word and l^i is the first lexical unit of the 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. Eq. (6) is the result of log transformation of Eq. (5). Usually, $p(\mathbf{x}_t|q_t = l^i, \Theta_A)$ is referred to as the *local emission score* and $P(q_t = l^i|q_{t-1} = l^j, \Theta_A)$ is referred to as the *transition score*. The present paper deals only with the issues related to the estimation of the local emission score.

2.2. Framework of probabilistic lexical modeling

The local emission score $p(\mathbf{x}_t|q_t = l^i, \Theta_A)$ or the relationship between the acoustic feature observation $\mathbf{x}_t$ and the lexical unit l^i can be factored through a *latent* variable a^d as following:

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

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

$$= \sum_{d=1}^{D} \underbrace{p(\mathbf{x}_t|a^d, \theta_a)}_{\textbf{acoustic model}} \cdot \underbrace{P(a^d|q_t = l^i, \theta_l)}_{\textbf{lexical model}} \tag{9}$$

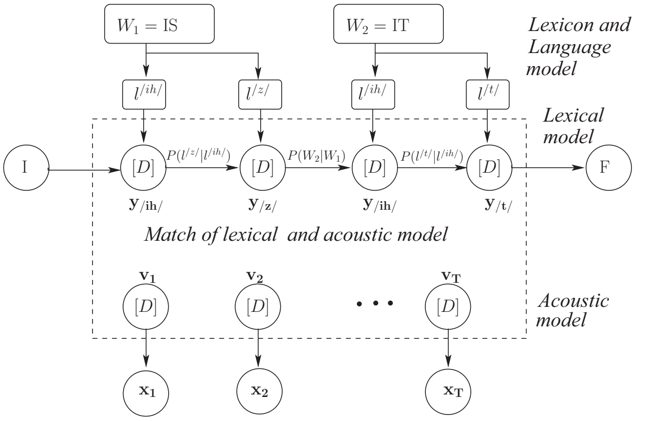

Fig. 1. The graphical model representation of a system incorporating probabilistic lexical modeling.

We refer to the latent variable a^d as the acoustic unit and the set of acoustic units $\mathcal{A} = \{a^1, \ldots a^d, \ldots a^D\}$ where D is the total number of acoustic units. The relationship in Eq. (9) is a result of the assumption that given a^d, $p(\mathbf{x}_t|a^d, q_t = l^i, \theta_a, \theta_l)$ is independent of l^i. In Eq. (9), $p(\mathbf{x}_t|a^d, \theta_a)$ is the acoustic unit likelihood, and $P(a^d|l^i, \theta_l)$ is the probability of the acoustic unit given the lexical unit and is given by the lexical model. In this paper, we refer to $p(\mathbf{x}_t|a^d, \theta_a)$ as the acoustic model evidence and $P(a^d|l^i, \theta_l)$ as the lexical model evidence. The parameters of the acoustic likelihood estimator Θ_A now encompass the *acoustic model* (θ_a), the *pronunciation lexicon* (θ_{pr}) and the *lexical model* (θ_l) parameters, therefore, $\Theta_A = \{\theta_a, \theta_{pr}, \theta_l\}$.

The graphical model representation of a system based on Eqs. (6) and (10) for the word sequence "IS IT" is illustrated in Fig. 1. In the figure, I and F refer to the non-emitting initial and final HMM states. The figure shows that the sequence of words constrained by the language model is represented by a sequence of lexical units (l^{ih} l^{z} l^{ih} l^{t}) as given by the pronunciation lexicon. For each lexical unit l^i, the lexical model computes a D dimensional categorical variable $\mathbf{y}_i = [y_i^1, \ldots, y_i^d, \ldots, y_i^D]^{\mathrm{T}}$, $y_i^d = P(a^d|l^i, \theta_l)$ that models a probabilistic relationship between a lexical unit l^i and D acoustic units. Given the acoustic feature observation $\mathbf{x}_t$ at time t, the acoustic model computes an acoustic unit likelihood vector $\mathbf{v}_t = [v_t^1, \ldots, v_t^d, \ldots, v_t^D]^{\mathrm{T}}$ where $v_t^d = p(\mathbf{x}_t|a^d, \theta_a)$. Having defined $\mathbf{y}_i$ and $\mathbf{v}_t$, Eq. (9) can be written as the following:

$$p(\mathbf{x}_t|q_t = l^i, \Theta_A) = \mathbf{y}_i^{\mathrm{T}}\mathbf{v}_t \tag{10}$$

Eq. (10) can be seen as a match between the acoustic and lexical model evidence, which in this case turns out to be the scalar product of $\mathbf{y}_i$ and $\mathbf{v}_t$.

2.3. Deterministic lexical model based ASR

Standard HMM-based ASR systems, for various reasons as elucidated shortly in the following subsections, implicitly model the dependency between acoustic feature

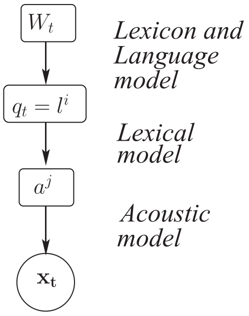

Fig. 2. The graphical model representation of a deterministic lexical model based ASR system.

observation $\mathbf{x}_t$ and a lexical unit l^i through the *latent* variable or the acoustic unit a^d. However, in standard HMM-based ASR systems each lexical unit l^i is deterministically mapped to an acoustic unit $a^j (l^i \mapsto a^j)$, i.e., the lexical model is deterministic,

$$y_i^d = P(a^d | q_t = l^i, \theta_l) = \begin{cases} 1, & \text{if } d = j; \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

The graphical model representation of an ASR system at time frame t in which the lexical model is deterministic is illustrated in Fig. 2. A lexical unit is given deterministically by the current word and its subword units. The lexical unit is mapped to an acoustic unit and the acoustic feature observation is conditioned on the acoustic unit.

It is worth mentioning that in HMM-based ASR literature, due to this deterministic relationship, typically no distinction is made between the acoustic and lexical units, or the acoustic and lexical models. Our main reason to refer to the lexical and acoustic units, or the acoustic and lexical models distinctly here is to bring out the contributions of the present paper clearly.

2.3.1. Lexical and acoustic units

Depending on the subword context modeled, there are two types of ASR systems: (1) context-independent subword unit based ASR systems, where lexical units are context-independent subword units, and (2) context-dependent subword unit based ASR systems, where the lexical units are context-dependent subword units.

In the case of context-independent subword unit based ASR systems, the acoustic unit set $\mathcal{A}$ is knowledge driven and defined based on the pronunciation lexicon. The number of acoustic units $D = K \times M$, where K is the number of context-independent subword units in the lexicon and M is the number of HMM states for each context-independent subword unit, typically, $M = 3$.

In the case of context-dependent subword unit based ASR systems, the number of lexical units $I = M \cdot K^{c_r+c_l+1}$ where c_l is the preceding context length, c_r is the following context length. Generally, not all context-dependent subword units will appear sufficiently often in the training data. Hence a sharing approach is used to enable multiple lexical units to share an acoustic model. This is done using a decision-tree based state clustering and tying technique that uses a pronunciation lexicon, linguistic knowledge to prepare a phonetic question set and acoustic data (Young et al., 1994). The number of acoustic units D varies depending on hyper parameters such as the state occupancy count and the log-likelihood threshold that are used during decision-tree based state clustering. However, the number of acoustic units D is well below the number of lexical units I.

2.3.2. Acoustic modeling

The two main approaches used in the literature to model the acoustic units are Gaussian mixture models (GMMs) and artificial neural networks (ANNs). The resulting ASR systems are usually referred to as HMM/GMM (Rabiner, 1989) and hybrid HMM/ANN (Morgan and Bourlard, 1995) systems, respectively.

1. In the HMM/GMM approach, the 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 the set of acoustic units $\mathcal{A}$ and the GMM parameters of the acoustic units.
2. In the hybrid HMM/ANN approach, an artificial neural network is first trained to estimate $P(a^d | \mathbf{x}_t, \theta_a)$ and then the scaled-likelihood $p_{sl}(\mathbf{x}_t | a^d, \theta_a)$ is estimated as

$$p_{sl}(\mathbf{x}_t | a^d, \theta_a) = \frac{p(\mathbf{x}_t | a^d, \theta_a)}{p(\mathbf{x}_t)} = \frac{P(a^d | \mathbf{x_t}, \theta_a)}{P(a^d)} \tag{12}$$

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

2.3.3. Deterministic lexical modeling

In context-independent subword unit based ASR systems, the deterministic relationship between lexical and acoustic units is knowledge driven. Therefore, lexical model training is not involved, and the deterministic map between lexical and acoustic units is the lexical model. The GMMs in the case of the HMM/GMM approach or the ANN in the case of the hybrid HMM/ANN approach is the acoustic model.

In context-dependent subword unit based ASR systems, lexical units are context-dependent subword units whereas acoustic units are clustered context-dependent subword units. As mentioned in Section 2.3.1, the decision trees and the phonetic question set are used to deterministically relate a lexical unit to an acoustic unit. Therefore, in context-dependent subword unit based HMM/GMM systems, the decision trees are the lexical model and the GMMs are the acoustic model. Similarly, in the case of hybrid HMM/ANN systems, decision trees are the lexical model and the ANN is the acoustic model (Dahl et al., 2012; Hinton et al., 2012).

3. Implications of deterministic lexical modeling

As described in the previous section, in standard HMM-based ASR systems the lexical model 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 or grapheme subword units then $\mathcal{A}$ is also based on phones or graphemes, respectively.
- If $\mathcal{L}$ is based on context-independent subword units or context-dependent subword units then $\mathcal{A}$ is also based on context-independent subword units or context-dependent subword units, respectively.

The performance of deterministic lexical model based ASR systems is dependent on the accuracy of the deterministic mapping which is in turn determined by the availability of well-developed resources. More specifically, deterministic lexical modeling imposes the following three constraints:

1. The availability of sufficient and well developed acoustic data in the target language or domain to effectively train both an acoustic model and a lexical model.
2. The availability of a well developed phonetic lexicon, as most of the ASR systems use phones as lexical units.
3. The ASR system trained with one phone set cannot be directly ported to or used as it is for a new domain which has a lexicon based on a different phone set. 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, CMUBET, SAMPA, etc.

Unfortunately, many languages do not have well-developed acoustic and lexical resources (Besacier et al., 2014). In the following subsections, we provide a literature survey on how the resource constraints have been addressed in the framework of deterministic lexical modeling.

3.1. Lack of acoustic resources

In the literature, the lack of acoustic resources has been typically addressed through approaches 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). The first step in most of these approaches is the definition of a common or universal phone set across all out-of-domain languages and the target language. This step ensures that the phone sets match across languages, thus addressing the third constraint mentioned above. The common or universal phone set can be defined either in a knowledge-based manner (Kohler, 1998; Beyerlein et al., 2000; Schultz and Waibel, 2001; Le and Besacier, 2009) or in a data-driven manner (Sim and Li, 2008; Sim, 2009). Multilingual acoustic models are first trained on the language-independent data and then adapted on the target language data.

In the framework of HMM/GMM systems, multilingual acoustic models or the GMMs serve as the seed models to be adapted on the target language data using techniques such as maximum a posteriori adaptation (MAP), maximum likelihood linear regression (MLLR) and subspace Gaussian mixture models (SGMM). The out-of-domain lexical model or the decision trees are 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). In the framework of hybrid HMM/ANN systems, the multilingual ANN can be used for the target language local emission score estimation after phone set mapping (Sim and Li, 2008; Sim, 2009). Other possibilities are training a hierarchical neural network (Pinto et al., 2011), adapting the multilingual ANN or the last layer of the multilingual ANN on the target language data (Swietojanski et al., 2012; Ghoshal et al., 2013; Huang et al., 2013), etc.

Alternatively, in the case of tandem approaches, the multilingual ANN is used to generate data-driven bottleneck or tandem features for the target language. These data-driven features are used to train an HMM/GMM system for the target language (Stolcke et al., 2006; Thomas and Hermansky, 2010; Thomas and Ganapathy, 2012). To fit the target language better, the multilingual ANN is sometimes adapted on the target language data with (Thomas and Hermansky, 2010) or without (Thomas and Ganapathy, 2012; Swietojanski et al., 2012) phone set mapping. However, in the tandem approach, as the acoustic and lexical models are trained on the target language data, minimal resources from the target language are necessary to robustly estimate the parameters.

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. Usually, given an existing manually developed or verified lexicon, a grapheme-to-phoneme (G2P) converter is trained to extract pronunciations for new words or to add pronunciation variants (Bisani and Ney, 2008; Novak, 2011). The augmented lexicon is then used to build an ASR system. However, for some languages, a seed lexicon may not be available to train a G2P convertor. Therefore, alternate subword units like graphemes, which make 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 systems primarily depends on the G2P relationship of the language. The reason for this is as follows: It can be seen in Eq. (9) that the acoustic model score $p(\mathbf{x}_t|a^d, \theta_a)$ models the dependency between the acoustic feature observation $\mathbf{x}_t$ and the acoustic unit a^d. As discussed in this section, due to the deterministic lexical modeling in standard HMM-based ASR systems, both the acoustic and lexical units are based on graphemes. However, the acoustic feature observations, or the cepstral features, depict the envelope of the short-term spectrum. The envelope of the short-term spectrum is related to phones. As a result, the more regular the G2P relationship is, the better is the acoustic model. Therefore, the use of graphemes as subword units has mainly succeeded for languages such as Spanish and Finnish where the G2P relationship is regular (Kanthak and Ney, 2002; Killer et al., 2003; Ko and Mak, 2014). For languages such as English which have an irregular G2P relationship, it has been found that grapheme-based ASR systems perform worse compared to phone-based systems (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 acoustic resources available in other languages have been explored (Kanthak and Ney, 2003; Stüker, 2008b,a). Similar to multilingual phone subword modeling, multilingual grapheme subword modeling is based on the *universal* or *multilingual* grapheme set formed by merging graphemes that are common across different languages. However, unlike multilingual phone sets, it is not trivial to port multilingual grapheme sets to new languages, mainly for two reasons: Firstly, grapheme sets of languages may not match or overlap. To overcome this issue, either transliteration or data driven mapping has been employed (Stüker, 2008a). 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. This is unlike phone subword units where it has been shown that multilingual acoustic models can outperform monolingual acoustic models.

4. Probabilistic lexical modeling

The two conditions, namely, $0 < P(a^d|l^i, \theta_l) < 1$ and $\sum_{d=1}^{D} P(a^d|l^i, \theta_l) = 1$, in Eq. (9) characterize an ASR approach where each lexical unit is probabilistically related to all acoustic units. We refer to them as probabilistic lexical model based ASR systems.

The probabilistic lexical modeling approaches presented in this paper presume that an acoustic unit set $\mathcal{A}$ is defined and a trained acoustic model is available. Therefore, in the first step, a standard HMM-based ASR system i.e., either an HMM/GMM system or a hybrid HMM/ANN system is trained. The acoustic model is the GMMs in the case of HMM/GMM or the ANN in the case of hybrid HMM/ANN. In the second step, the acoustic model from the first step is used with the pronunciation lexicon and acoustic training data to train the parameters of the probabilistic lexical model. More specifically, the parameters of the probabilistic lexical model are learned by training an HMM, whose states represent lexical units and each state l^i is parameterized by a categorical distribution $\mathbf{y}_i$. In this case, the lexical model parameter set consists of $\theta_l = \{\mathbf{y}_i\}_{i=1}^{I}$. We present these techniques from the perspective of the hybrid HMM/ANN. That is, in this paper we use an ANN as the acoustic model.

4.1. Kullback–Leibler divergence based HMM

In the first approach, lexical model parameters are learned through acoustic unit posterior probability estimates $P(a^d|\mathbf{x}_t, \theta_a)$ in the framework of Kullback–Leibler divergence based HMM (KL-HMM) (Aradilla et al., 2008). The feature observations used to train the HMM are the acoustic unit probability vectors $\mathbf{z}_t = \left[z_t^1 \ldots, z_t^d, \ldots, z_t^D\right]^{\mathrm{T}}$ where $z_t^d = P(a^d|\mathbf{x}_t, \theta_a)$. It is worth mentioning that KL-HMM was originally developed as an alternative acoustic modeling technique (Aradilla et al., 2008) to the 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, 2013b,a). In this paper, we explain and interpret all the literature on KL-HMM in terms of probabilistic lexical modeling.

In a KL-HMM, as both the feature observations and the state distributions are probability vectors, the local score or the match between acoustic and lexical model evidence at each HMM state can be the Kullback–Leibler (KL) divergence between the feature observation $\mathbf{z}_t^d$ and the categorical distribution $\mathbf{y}_i$,

$$S_{KL}(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} y_i^d \log\left(\frac{y_i^d}{z_t^d}\right) \tag{13}$$

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

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

$$S_{RKL}(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} z_t^d \log\left(\frac{z_t^d}{y_i^d}\right) \tag{14}$$

2. Symmetric KL-divergence (S_{SKL}): The local score S_{SKL} is the average of the local scores S_{KL} and S_{RKL}.

$$S_{SKL}(\mathbf{y}_i, \mathbf{z}_t) = \frac{1}{2} \cdot [S_{KL} + S_{RKL}] \tag{15}$$

The categorical distributions $\{\mathbf{y}_i\}_{i=1}^{I}$ are estimated by the Viterbi expectation maximization (EM) algorithm which minimizes a cost function based on the local score 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 a KL-divergence based local score.

4.2. Tied posterior

In the second approach, lexical model parameters are learned through scaled-likelihood estimates $p_{sl}(\mathbf{x}_t|a^d, \theta_a)$ (see Eq. (12)). The approach, referred to as the tied-posterior approach, was originally proposed in the framework of hybrid HMM/ANN to build context-dependent subword unit based ASR systems using an ANN trained to classify context-independent subword units (Rottland and Rigoll, 2000).

In the tied-posterior based HMM (tied-HMM) approach, the emission likelihood at each context-dependent HMM state $q_t = l^i_{cd}$ is estimated as,

$$p(\mathbf{x}_t|q_t = l^i_{cd}) = \sum_{d=1}^{D} w^d_i \cdot p_{sl}(\mathbf{x}_t|a^d_{ci}) \tag{16}$$

where a^d_{ci} is a context-independent phone, D is the number of context-independent phones, $p_{sl}(\mathbf{x}_t|a^d_{ci})$ is the scale-likelihood, $0 \leqslant w^d_i \leqslant 1$ is the weight corresponding to the 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 log-likelihood using the EM algorithm. Comparison between Eqs. (16) and (9) shows that l^i_{cd} corresponds to the lexical unit l^i, a^d_{ci} corresponds to the acoustic unit a^d and w^d_i corresponds to $y^d_i = P(a^d|l^i, \theta_l)$. In other words, the tied-HMM approach is an HMM-based ASR approach that incorporates probabilistic lexical modeling.

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

$$S_{tied}(\mathbf{y}_i, \mathbf{v}_t) = \log\left(\sum_{d=1}^{D} y^d_i . v^d_t\right) = \log\left(\mathbf{y}_i^{\mathrm{T}}\mathbf{v}_t\right) \tag{17}$$

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

4.3. Scalar product HMM

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

$$S_{SP}(\mathbf{y}_i, \mathbf{z}_t) = \log\left(\mathbf{y}_i^{\mathrm{T}}\mathbf{z}_t\right) \tag{18}$$

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

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

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

Parameter estimation for the KL-HMM, tied-HMM and SP-HMM approaches is elaborated in Appendix A. More details about the parameter estimation for the KL-HMM approach can be found in the thesis by Aradilla (2008). An issue that is common to all 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 HMM states of the KL-HMM, tied-HMM or SP-HMM systems using the approach proposed by Imseng et al., 2012.

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

In the three probabilistic lexical modeling approaches discussed, the 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 $\mathbf{y}_i$, as shown in Fig. 1. Yet another similarity between the three approaches is that they reduce to the standard hybrid HMM/ANN system described in Section 2 when the lexical model is deterministic, i.e., $\mathbf{y}_i$ is a Kronecker delta function. Despite these similarities, the KL-HMM approach has additional advantages compared to the tied-HMM and SP-HMM approaches. We discuss them briefly in this section.

From the communication theory perspective, the standard HMM-based ASR approach can be seen as a communication problem where the noisy output of the acoustic channel is decoded by a linguistic decoder (Bahl et al., 1983). That is, a sequence of acoustic unit likelihood vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_T\}$ or a sequence of acoustic unit posterior vectors $\{\mathbf{z}_1, \ldots, \mathbf{z}_T\}$ is compared with possible sequences of lexical model parameter vectors (for example, $\{\mathbf{y}_i, \ldots \mathbf{y}_g\}$ 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 the lexical model and consequently relies on the 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 approaches is to add pronunciation variants, i.e., improve the deterministic lexical model (Strik and Cucchiarini, 1999).

The KL-HMM approach using the 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,

$$\begin{aligned} S_{KL}(\mathbf{y}_i, \mathbf{z}_t) &= \sum_{d=1}^{D} y_i^d \log\left(\frac{y_i^d}{z_t^d}\right) \\ &= \sum_{d=1}^{D} y_i^d \log y_i^d - \sum_{d=1}^{D} y_i^d \log z_t^d \end{aligned} \tag{19}$$

The first part of Eq. (19) which is the entropy of the probability distribution $\mathbf{y}_i$ takes into account the uncertainty in the lexical model. 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 a Kronecker delta distribution, i.e., a deterministic lexical model. In such a case, the hybrid HMM/ANN approach (Bourlard and Morgan, 1994) can be seen as a special case of the KL-HMM approach when the acoustic unit probability estimate $P(q_t = a^d | \mathbf{x}_t, \theta_a)$ rather than the acoustic unit likelihood estimate $p_{sl}(\mathbf{x}_t | q_t = a^d, \theta_a)$ is used as the local emission score.

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

$$\begin{aligned} S_{RKL}(\mathbf{y}_i, \mathbf{z}_t) &= \sum_{d=1}^{D} z_t^d \log\left(\frac{z_t^d}{y_i^d}\right) \\ &= \sum_{d=1}^{D} z_t^d \log z_t^d - \sum_{d=1}^{D} z_t^d \log y_i^d \end{aligned} \tag{20}$$

It can be observed from Eq. (20) that the first quantity, the entropy of probability distribution $\mathbf{z}_t$, is independent of the lexical unit. The matching only takes place with the second quantity which is the cross entropy between distributions $\mathbf{z}_t$ and $\mathbf{y}_i$, with $\mathbf{z}_t$ as the reference. The local score $S_{SKL}(\mathbf{y}_i, \mathbf{z}_t)$ gives equal importance to the acoustic and lexical models.

Another difference between the KL-HMM and tied-HMM/SP-HMM approaches is that the KL-divergence based local scores can be linked to hypothesis testing (Blahut, 1974). The acoustic model evidence and lexical model evidence is matched discriminatively irrespective of the local score used. We use these distinctions to better explain our findings in Section 6.

The above differences among different KL-divergence based local scores are from the decoding perspective. The details on the role of different cost functions from the training perspective were presented by Rasipuram and Magimai-Doss (2013b).

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 a probabilistic manner. As a consequence probabilistic lexical model based ASR systems have the following advantages:

1. The parameters of the acoustic model θ_a and the lexical model θ_l can be trained on an independent set of resources. In this light, previous work on KL-HMM suggests that ASR systems can be rapidly developed using a domain-independent or language-independent acoustic model and by training only the lexical model on the target language or domain data (Imseng et al., 2011; Imseng et al., 2012; Rasipuram et al., 2013a).
2. $\mathcal{L}$ and $\mathcal{A}$ can model different contextual units. For instance, as in the previous work, $\mathcal{L}$ can be based on context-dependent subword units while $\mathcal{A}$ can be based on context-independent subword units (Rottland and Rigoll, 2000; Magimai-Doss et al., 2011; Imseng et al., 2011; Imseng et al., 2012; Rasipuram et al., 2013a). These ASR systems have been found to yield performance comparable to or better than standard context-dependent subword unit based HMM/GMM systems.
3. It is not necessary that the subword unit set used for defining the acoustic units should be the same as the subword unit set used for defining the 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 is based on phones and the lexical unit set is based on graphemes (Magimai-Doss et al., 2011; Imseng et al., 2011; Rasipuram et al., 2013a; Rasipuram and Magimai-Doss, 2013a).

4.6. Proposed grapheme-based ASR approach

In this paper, we propose a grapheme-based ASR approach where,

- First, an acoustic model that models multilingual phones is trained on language-independent acoustic and lexical resources.
- Then, the lexical model which captures a probabilistic relationship between target language graphemes and multilingual phones is trained on a relatively small amount of target language-dependent acoustic data.

The proposed approach is motivated from the following observations:

1. Multilingual phone-based acoustic models are sharable across languages. As discussed in Section 3.1, many acoustic model adaptation approaches addressing acoustic resource constraints in ASR system development exploit this aspect.
2. As mentioned in Section 4.5, when the acoustic units are based on phones and the lexical units are based on graphemes, probabilistic lexical modeling techniques such as KL-HMM are capable of learning a probabilistic G2P relationship. In a cross-domain English ASR study, it was observed that this aspect can be exploited to build grapheme-based ASR systems (Magimai-Doss et al., 2011; Rasipuram and Magimai-Doss, 2013b; Rasipuram, 2014). These grapheme-based ASR systems performed similarly to phone-based ASR systems, where the target domain phone lexicon is built by training a G2P converter on a cross-domain phone lexicon. This suggests that probabilistic lexical modeling approaches with lexical units based on graphemes and acoustic units based on phones could address lexical resource constraints by integrating lexicon learning as a phase in training the ASR system.
3. The probabilistic G2P relationship could be learned on a relatively small amount of target-domain transcribed speech (Imseng et al., 2011). Further, such a grapheme-based ASR system performed better than conventional phone-based acoustic model adaptation systems.

Given these observations, we hypothesize that the proposed grapheme-based ASR approach can address both acoustic and lexical resource constraints better than acoustic model adaptation based approaches developed in the framework of deterministic lexical modeling.

5. Experimental setup

The hypothesis is validated by training a single language-independent multilingual acoustic model and conducting ASR studies on the following three different resource-constrained tasks where only a lexical model is trained:

- Non-native accented speech recognition task that lacks both acoustic and “well developed” phonetic lexical resources. Typically, the phone lexicon consists native speaker pronunciations. In the literature, non-native accented ASR research has mainly focused on acoustic model adaptation. We investigate it on English where the G2P relationship is irregular.
- Rapid development of an 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 a Greek ASR task.
- Development of an ASR system for a minority and under-resourced language, Scottish Gaelic, which has only 60,000 speakers. The endangered status of Scottish Gaelic makes low-cost speech technology important for language conservation efforts. Scottish Gaelic also lacks sufficient acoustic resources and does not have any phonetic lexical resources. The G2P relationship of Scottish Gaelic is regular, and many-to-one as the number of graphemes in a word is significantly higher than the number of phones (Rasipuram et al., 2013a).

We compare the probabilistic lexical modeling based ASR 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 the KL-HMM approach 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 the SpeechDat(II) corpus, specifically, British English, Italian, Spanish, Swiss French and Swiss German, is used as the language-independent dataset. Each language has approximately 12 h of speech data, in total amounting to 63 h. All the SpeechDat(II) lexica 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

The HIWIRE corpus contains English 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 a grammar with a perplexity of 14.9. The HIWIRE task does not have training data. It only contains adaptation data of 50 utterances per speaker, approximately 150 min and test data of 50 utterances per speaker, approximately 150 min. To simulate limited resources the amount of adaptation data is reduced from 150 to 3 min (specifically, 150, 120, 90, 64, 32, 16, 10 and 3 min respectively) by picking various subsets of utterances (Imseng et al., 2011). The grapheme-based lexicon was transcribed using 27 graphemes comprising 26 English graphemes and silence.

A noticeable difference between the work of Imseng et al. (2011) and this paper is the following: In the previous work a phone-lexicon based on the ARPABET phone set supplied with the HIWIRE corpus was used, whereas in this work we use a phone-lexicon based on the SAMPA phone set. The phone-lexicon based on the SAMPA phone set was created by borrowing pronunciations of 102 words that are in common from the SpeechDat(II) English lexi-

Table 1
Overview of different systems. CI denotes context-independent subword units, cCD denotes clustered context-dependent subword states 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.

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

con. For the remaining 31 words, we obtained pronunciations by mapping ARPABET phones to SAMPA phones. The main reason to use the SAMPA phone set based lexicon in this work is to have a shared subword unit set between the out-of-domain lexicon and the target-domain lexicon. This allowed the evaluation of acoustic model adaptation based systems (MAP and MLLR) discussed in Section 5.2.2. Also, native English is present in out-of-domain resources. Therefore, in the case of the KL-HMM, SP-HMM and tied-HMM approaches, the lexical model parameters trained on SpeechDat(II) English are adapted using the HIWIRE adaptation data. Additionally, the use of phone-lexicon based on the SAMPA phone set allowed us to 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 that of Imseng et al. (2012). The training set contains 13.5 h of speech from 1500 speakers; the development set contains 1.5 h of speech from 150 speakers; and the test set contains 6.9 h of speech from 350 speakers. 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 the SAMPA phone set. To simulate limited resources, the amount of available data was reduced from 13.5 h down to 5 min (specifically, 800, 300, 150, 75, 37, 18, 9 and 5 min respectively). All the systems were evaluated on the same test set. The test set contains 10,000 unique words. The performance of the phone-based KL-HMM, MAP, MLLR and HMM/GMM systems presented by Imseng (2013) [Figs. 4.3 and 4.4] is taken as the reference in this paper.

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

5.1.4. Scottish Gaelic

The Scottish Gaelic speech corpus[1] was collected by CSTR,[2] University of Edinburgh. The experimental setup is similar to that of Rasipuram et al. (2013a). The Gaelic corpus consists of speech from 46 speakers. The training set consists of 22 speakers and 2389 utterances amounting to 3 h of speech; the development set consists of 12 speakers and 1112 utterances amounting to 1 h of speech; and the test set consists of 12 speakers and 1317 utterances amounting to 1 h of speech. The speakers in the training, development and test sets are different. The vocabulary size is 5000 unique words. The database does not contain a phone pronunciation lexicon. The grapheme-based lexicon contains 83 graphemes comprising 5 vowels, 5 long vowels, 23 broad consonants, 23 slender consonants, 26 consonants and silence. This grapheme lexicon is obtained by considering broad and slender Gaelic consonants as separate graphemes. We refer to this lexicon as the *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 the *orthography-based* lexicon. This lexicon is prepared in the traditional way from the orthography of words. The *orthography-based* lexicon consists of 32 Gaelic graphemes comprising 25 graphemes, 5 accents and silence.

Table 2 summarizes the information about the different corpora used.

5.2. Systems

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

[1] http://forum.idea.ed.ac.uk/idea/gaelic-speech-recognition-and-scots-gaelic-sound-archive.

[2] The Centre for Speech Technology Research (CSTR).

Table 2
Overview of the tasks and the respective corpora used in the study. The details of the data used to train the multilingual acoustic model are given in italic font.

Corpus (Description)	Language	# of Subword units		Train data	Test data
		Phones	Graphemes	(in min)	(in min)
SpeechDat(II) (Native speech sampled at 8 K)	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 the multilingual acoustic model)		*117*	*47*	*3780*	*n.a*
HIWIRE (Non-native speech from natives of France, Spain, Italy and Greece)	English	42	27	0–150	150
SpeechDat(II) (Native Greek speech)	Greek	31	25	5–800	360
Scottish Gaelic (Broadcast news data)	Scottish Gaelic	n.a.	83 or 32	180	60

5.2.1. Probabilistic lexical modeling based systems

As an acoustic model, we use a standard three-layer multilingual multilayer perceptron (MLP) trained on the language-independent dataset to classify 117 context-independent multilingual phones. 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 three-layer MLP for the following reasons:

- The same MLP has been used in the previous ASR studies on the HIWIRE and Greek tasks (Imseng et al., 2011; Imseng et al., 2012). Therefore, the results from the present study are directly comparable to the previous studies.
- In recent work, it has been shown that the KL-HMM retains its benefit over standard hybrid HMM/ANN systems even when an 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 a probabilistic lexical modeling framework is open for further research. A lexical model is trained for each of the probabilistic lexical modeling systems, namely, KL-HMM, SP-HMM and tied-HMM as described in Section 4. We used S_{RKL} as the local score for the KL-HMM system based on recent investigations (Rasipuram and Magimai-Doss, 2013b; 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 the Roman alphabet. Therefore, a multilingual grapheme set of 47 units was formed by merging graphemes that are common across all languages in the language-independent data set. Accents and diacritics are treated as separate graphemes. The grapheme-based HMM/GMM system used multilingual graphemes as subword units.

Each context-dependent subword unit was modeled using three-HMM states and each HMM state was modeled using a mixture of 16 Gaussians. Then, MAP or MLLR adaptation[3] was performed using speech data from the target language or domain. For MLLR adaptation, we used a regression class tree to group the Gaussians in the model set into regression classes and we used up to 32 regression classes.

As described in Section 5.1.3, for the Greek task a transliterated grapheme-based lexicon was used while performing MAP or MLLR adaptation.

5.2.3. HMM/GMM and tandem ASR systems

These are HMM/GMM ASR systems where both the acoustic model and the lexical model are trained on the language-dependent data. We investigate two systems: the HMM/GMM system that uses standard cepstral features as feature observations, and the tandem system that uses tandem features as feature observations (Hermansky

[3] We also applied MLLR and MAP in sequence (Oh and Kim, 2009). On the Greek task, it was observed that the performance of phone-based MLLR + MAP systems was better than that of the phone-based MAP or MLLR systems when at least 37 min of Greek acoustic data is available. However, such gains were not observed for grapheme-based MLLR + MAP systems. Further, the performance of the KL-HMM systems was always better than that of the MLLR + MAP systems. Therefore, the results of MLLR and MAP adaptation in sequence are not reported in the paper.

et al., 2000). As indicated in Table 1, the tandem system exploits both language-dependent and language-independent resources similarly to probabilistic lexical model based systems and acoustic model adaptation based systems.

The tandem features were extracted by transforming the 117-dimensional outputs of the multilingual MLP described 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.

The HMM/GMM systems used 39-dimensional PLP cepstral feature vectors as acoustic features. All the phone subword based systems use a phonetic question set and grapheme subword based systems use a singleton question set for the 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 or 39 dimensions) was tuned on the development set. The HTK toolkit was used to build all the HMM/GMM systems (Young et al., 2006).

6. Results

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

6.1. Rapid ASR development

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

Fig. 3(a) plots the performances of the phone- and grapheme-based KL-HMM, SP-HMM and tied-HMM systems with increasing amounts of training data on the HIWIRE task. Similarly, Fig. 3(b) plots the performances on the Greek ASR task with increasing amounts of training data. The figures show that the KL-HMM system consistently performs better compared to the SP-HMM and tied-HMM systems for both phone and grapheme subword units. Furthermore, on the HIWIRE task, the difference between the KL-HMM system and the SP-HMM or tied-HMM systems is more for grapheme-based systems than for phone-based systems.

6.1.2. Comparison of probabilistic lexical modeling based systems with other systems

Fig. 4(a) plots the performances of the phone- and grapheme-based KL-HMM, MAP, MLLR, tandem and HMM/GMM systems with increasing amounts of training data on the HIWIRE task. Similarly, Fig. 4(b) plots the performances on the Greek ASR task with increasing amounts of training data. We can draw the following inferences from the figures:

1. Irrespective of the subword units used, KL-HMM systems perform better than deterministic lexical model based systems when there is limited training data and comparable to deterministic lexical model based systems as the training data is increased.
2. On both tasks, the difference in performance between phone and grapheme-based systems is minimal for the KL-HMM approach compared to all other approaches.
3. On both the HIWIRE (where the G2P relationship is irregular) and Greek (where the G2P relationship is regular) tasks it can be observed that deterministic lexical model based systems are more suitable for phones than graphemes.
   (a) On the HIWIRE task, the acoustic model adaptation based systems perform better than the HMM/GMM or tandem systems. However, the perfor-

Table 3
Performance in terms of word accuracy on the HIWIRE test set for the various cross-word context-dependent ASR systems trained on varying amounts of HIWIRE adaptation data.

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 4
Performance in terms of word accuracy on the Greek test set for the various cross-word context-dependent ASR systems trained on varying amounts of Greek data.

System	5 min		37 min		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

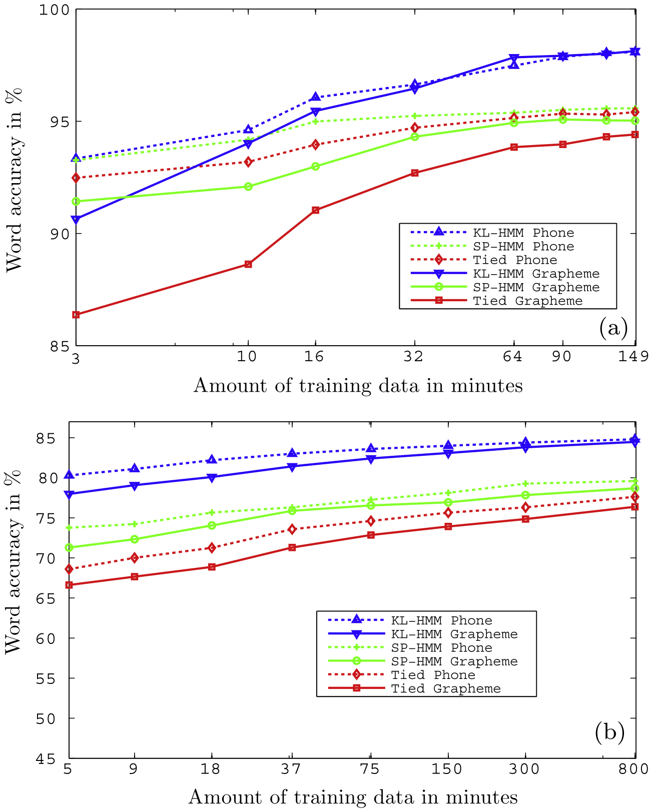

Fig. 3. Comparison between probabilistic lexical modeling based systems with increasing amounts of target domain or language training data. (a) on the HIWIRE non-native accented speech recognition task, (b) on the Greek ASR task.

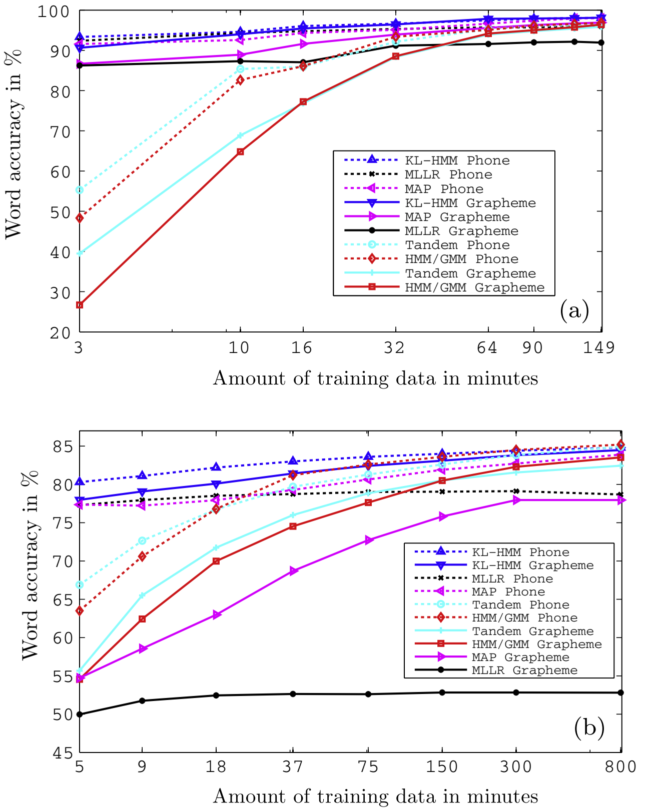

Fig. 4. Comparison of the phone-based and grapheme-based KL-HMM systems against the acoustic model adaptation based systems and the standard HMM/GMM system with increasing amounts of target domain training data. (a) on the HIWIRE non-native accented speech recognition task, (b) on the Greek ASR task.

mance of acoustic model adaptation systems using graphemes is worse than with phones as subword units. On the Greek task, where the transliterated grapheme-based lexicon was used, grapheme-based acoustic model adaptation systems perform significantly worse than phone-based acoustic model adaptation or HMM/GMM or tandem systems. The results also show that in the 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, 2008a).

(b) When the available training data is larger, on both tasks, phone-based deterministic lexical model systems perform comparably to the phone-based KL-HMM system. For example, with larger adaptation/training data sizes, on the HIWIRE task, MAP and KL-HMM systems perform similarly and on the Greek task, KL-HMM, HMM/

GMM and tandem systems perform similarly. However, in the case of grapheme-based systems this trend is not observed. The results, inline with the other multilingual grapheme-based ASR studies show that the use of multilingual grapheme models across languages does not appear evident (Kanthak and Ney, 2003; Killer et al., 2003; Stüker, 2008b).

4. Monolingual HMM/GMM systems and acoustic model adaptation based systems with the shared unit set (i.e., on the HIWIRE task) that exploit multilingual speech converge with the increase in acoustic resources.
5. Compared to the HMM/GMM approach, the tandem approach is beneficial mainly in low acoustic resource conditions.
6. Comparing MAP and MLLR approaches, 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 the HIWIRE test set using language-independent acoustic and lexical models without any adaptation. The performance on the HIWIRE task for the KL-HMM, SP-HMM, tied-HMM and the language-independent HMM/GMM systems is given in Table 5. The lexical model for the KL-HMM, SP-HMM and tied-HMM systems is trained on the SpeechDat(II) English data. It can be observed that, for both phone and grapheme subword units the KL-HMM system performs better than the SP-HMM, tied-HMM and LI HMM/GMM systems. Also, it is interesting to note that irrespective of the subword units used, the performances of all the probabilistic lexical model based systems (that use context-independent phones as acoustic units) are better than that of the LI HMM/GMM system (that uses context-dependent phones as acoustic units).

6.2. Scottish Gaelic ASR

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

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

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 6
Performance in terms of word accuracy on the Gaelic test set for the various cross-word context-dependent ASR systems.

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

6.3. Analysis

From the experiments presented earlier in this section, it can be observed that despite using exactly the same acoustic model, the performance trends of the various probabilistic lexical modeling approaches are different. The KL-HMM system performs better than the deterministic lexical model based systems in under-resourced conditions and performs similar to the deterministic lexical model based system in well-resourced conditions. While, the 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 the KL-HMM system to its abilities discussed in Section 4.4.

In order to ascertain the reason for difference in performance trends among the various probabilistic lexical modeling approaches, we conducted the following study. On the HIWIRE task, with the 150 min target data condition, the lexical model trained using the KL-HMM *RKL* approach is decoded with a Viterbi decoder using various 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. The results of this study are given in Table 7.

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 local scores $S_{SP}(\mathbf{y}_i, \mathbf{z}_t)$ and $S_{tied}(\mathbf{y}_i, \mathbf{v}_t)$. This result indicates that KL-divergence based local scores are better than scalar product based local scores. Furthermore, decoding with

Table 7
Comparison across different local scores used during decoding. The system trained with the KL-HMM *RKL* approach is decoded with all the other local scores.

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

$S_{KL}(\mathbf{y}_i, \mathbf{z}_t)$, $S_{SP}(\mathbf{y}_i, \mathbf{z}_t)$ 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 the acoustic and lexical models yields performance similar to $S_{RKL}(\mathbf{y}_i, \mathbf{z}_t)$. It can also be noted that the lexical model trained using the KL-HMM approach and decoded with $S_{SP}(\mathbf{y}_i, \mathbf{z}_t)$ and $S_{tied}(\mathbf{y}_i, \mathbf{v}_t)$ local scores results in better performance compared to the lexical model trained using the SP-HMM and tied-HMM approaches and decoded with $S_{SP}(\mathbf{y}_i, \mathbf{z}_t)$ and $S_{tied}(\mathbf{y}_i, \mathbf{v}_t)$ local scores. This indicates that the KL-HMM approach with the local score S_{RKL} yields a better lexical model compared to the SP-HMM or tied-HMM approaches. Deeper investigations on these aspects are out of the scope of the present paper.

6.4. Comparisons to the literature

In the literature, there are studies that have been reported on the 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 with studies in which no kind of adaptation was performed.

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

As shown in Table 8, the phone-based KL-HMM system performs better than the approaches proposed in the literature. The grapheme-based KL-HMM system performs comparable to the phone-based systems reported in the literature. It can also be observed from Tables 8 and 5 that the phone-based LI HMM/GMM system performs similarly to the systems from the literature, whereas the grapheme-based LI HMM/GMM system performs worse.

Table 8
Comparison of word accuracies (WA) on the HIWIRE test set without any adaptation.

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

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

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

As shown in Table 9, the hybrid HMM/ANN system using LHN based adaptation performs similarly to the phone-based and grapheme-based KL-HMM systems. According to Imseng et al. (2011), on the HIWIRE task, the performance of grapheme-based KL-HMM systems using low amounts of HIWIRE adaptation data (like 3–10 min) was significantly worse than that of phone-based KL-HMM systems. The reason for this could be that the lexical model parameters were directly trained on the limited HIWIRE adaptation data. In this work, this gap in performance has significantly reduced as the lexical model parameters trained on SpeechDat(II) English are adapted using HIWIRE adaptation data.

In the case of the Greek task, as previously mentioned phone-based KL-HMM, MLLR, MAP and HMM/GMM systems reported by Imseng (2013) [Fig. 4.3 in Page 59 and 4.4 in Page 60] have been used as reference. However, the phone-based tandem systems reported by Imseng (2013) and this paper differ mainly in terms of the dimensionality of the tandem features used. Imseng (2013) always used 117-dimensional tandem features. In this work, the dimension of features i.e., either 117 or 39 was tuned on the development set for each of the training conditions. We found dimensionality reduction to be beneficial, especially in the low acoustic resource conditions. For example, on the 5 min acoustic resource case, performance of the phone-based tandem system reported by

Table 9
Comparison of word accuracies (WA) on the HIWIRE test set with adaptation.

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

Imseng (2013) was 30.2% word accuracy with 117-dimensional tandem features. In this paper, with 39-dimensional tandem features we achieved a performance of 66.9% word accuracy.

In previous study on Scottish Gaelic ASR, a knowledge-based grapheme lexicon that tagged word beginning and end graphemes was used and word-internal context-dependent graphemes were modeled (Rasipuram et al., 2013a). The KL-HMM and HMM/GMM systems achieved a 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 smaller. Furthermore, in this paper we modeled cross-word context-dependent subword units. As can be seen from Table 6, the knowledge-based HMM/GMM system yields an absolute improvement of 3.2% word accuracy compared to the previous work and the grapheme KL-HMM system achieves performance comparable to that of the previous study.

7. Discussion and conclusion

In this work, we showed that ASR systems can be rapidly built using a language-independent acoustic model and training only the lexical model on a small amount of target language data. In recent work, it has been 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 systems (Rasipuram et al., 2013b). Further, it was also shown that if untranscribed speech data from the target language is available then the lexical model parameters can be adapted in an unsupervised manner to improve the performance of the ASR system.

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

- When using an ANN-based acoustic model, this can be achieved by training a hierarchical neural network (Pinto et al., 2011) or adapting the neural network with target language data (Swietojanski et al., 2012; Ghoshal et al., 2013; Huang et al., 2013). A recent study on Scottish Gaelic in the framework of KL-HMM has shown the potential of acoustic model adaptation using the hierarchical neural network approach (Rasipuram et al., 2013a).
- The KL-HMM approach is not restricted to ANN-based acoustic modeling alone (Rasipuram and Magimai-Doss, 2013a). Therefore, using GMMs as the acoustic model this can be achieved by adapting the GMMs through the MAP technique followed by KL-HMM training; or the parameters of GMMs and probabilistic lexical model can be jointly estimated using the approach proposed by Luo and Jelinek (1999).

As mentioned in Section 3, in the deterministic lexical modeling framework, acoustic model adaptation and lexical model adaptation can be combined in different ways. For instance, (a) by combining acoustic model adaptation with polyphone decision tree state tying (Schultz and Waibel, 2001) or (b) using the 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.

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 the deterministic lexical model based ASR approach is more suitable for phone-based ASR than grapheme-based ASR, while the probabilistic lexical model based ASR approach is suitable for both.

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

Acknowledgments

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 and Marzieh Razavi for their help with the Greek task and Dr. Phil Garner for proof reading the article. The paper has benefitted from the discussions the authors had with Prof. Tanja Schultz and her group. The authors would like to thank the anonymous reviewers for their constructive comments, insightful inputs and suggestions.

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 the sequence of cepstral features of length $T(n)$ and $W(n)$ represents the sequence of underlying words, the set of acoustic unit probability vectors $\{Z(n), W(n)\}_{n=1}^{N}$ or the set of likelihood vectors $\{V(n), W(n)\}_{n=1}^{N}$ are estimated. $Z(n)$ represents a sequence of acoustic unit 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 $\widehat{\Theta}_{kull}$ by the Viterbi expectation maximization algorithm which minimizes,

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

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

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

where $Q = \{q_1, \ldots q_t, \ldots, q_{T(n)}\}, q_t \in \{1, \ldots, I\}$ 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 the Viterbi algorithm. Given optimal state sequences and acoustic unit posterior vectors belonging to each of these states, the optimization step then estimates new set of model parameters by minimizing Eq. (A.1) or maximizing (A.2) subject to the constraint that $\sum_{d=1}^{D} y_i^d = 1$.

The optimal state distribution for the local score S_{RKL}, is the arithmetic mean of the training acoustic unit 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 unit probability vectors assigned to state l^i and $M(i)$ is the cardinality of $Z(i)$.

The optimal state distribution for the tied-HMM 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 unit probability vectors assigned to state l^i and $M(i)$ is the cardinality of $V(i)$.

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

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

References

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

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

Bahl, L.R., Jelinek, F., Mercer, R., 1983. A maximum likelihood approach to continuous speech recognition. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-5, 179–190.

Besacier, L., Barnard, E., Karpov, A., Schultz, T., 2014. Automatic speech recognition for under-resourced languages: a survey. Speech Commun. 56, 85–100.

Beyerlein, P., Byrne, W., Huerta, J.M., Khudanpur, S., Marthi, B., Morgan, J., Peterek, N., Picone, J., Wang, W., 2000. Towards language independent acoustic modeling. In: Proc. of ICASSP, pp. 1029–1032.

Bisani, M., Ney, H., 2008. Joint-sequence models for grapheme-to-phoneme conversion. Speech Commun. 50, 434–451.

Blahut, R.E., 1974. Hypothesis testing and information theory. IEEE Trans. Inform. Theory IT-20.

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

Burget, L., et al., 2010. Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models. In: Proc. of ICASSP, pp. 4334–4337.

Dahl, G.E., Yu, D., Deng, L., Acero, A., 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Language Process. 20, 30–42.

Dines, J., Magimai-Doss, M., 2007. A study of phoneme and grapheme based context-dependent ASR systems. In: Proc. of Machine Learning for Multimodal Interaction (MLMI), pp. 215–226.

Gemello, R., Mana, F., Scanzio, S., 2007. Experiments on hiwire database using denoising and adaptation with a hybrid HMM-ANN model. In: Proc. of Interspeech, pp. 2429–2432.

Ghoshal, A., Swietojanski, P., Renals, S., 2013. Multilingual training of deep neural networks. In: Proc. of ICASSP, pp. 7319–7323.

Hermansky, H., Ellis, D., Sharma, S., 2000. Tandem connectionist feature extraction for conventional HMM systems. In: Proc. of ICASSP, pp. 1635–1638.

Hinton, G. et al., 2012. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process. Mag. 29, 82–97.

Huang, J.T., Li, J., Yu, D., Deng, L., Gong, Y., 2013. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In: Proc. of ICASSP, pp. 7304–7308.

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

Imseng, D., Rasipuram, R., Magimai-Doss, M., 2011. Fast and flexible kullback-leibler divergence based acoustic modeling for non-native speech recognition. In: Proc. of ASRU, pp. 348–353.

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

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

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

Kanthak, S., Ney, H., 2003. Multilingual acoustic modeling using graphemes. In: Proc. of EUROSPEECH, pp. 1145–1148.

Killer, M., Stüker, S., Schultz, T., 2003. Grapheme based speech recognition. In: Proc. of EUROSPEECH.

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

Kohler, J., 1998. Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks. In: Proc. of ICASSP, vol. 1, pp. 417–420.

Le, V.B., Besacier, L., 2009. Automatic speech recognition for under-resourced languages: application to vietnamese language. IEEE Trans. Audio Speech Language Process. 17, 1471–1482.

Luo, X., Jelinek, F., 1999. Probabilistic classification of HMM states for large vocabulary continuous speech recognition. In: Proc. of ICASSP, pp. 353–356.

Magimai-Doss, M., Rasipuram, R., Aradilla, G., Bourlard, H., 2011. Grapheme-based automatic speech recognition using KL-HMM. In: Proc. of Interspeech, pp. 2693–2696.

Morgan, N., Bourlard, H., 1995. Continuous speech recognition: an introduction to the hybrid HMM/connectionist approach. IEEE Signal Process. Mag., 25–42

Novak, J., 2011. Phonetisaurus: a WFST-driven phoneticizer. <http://code.google.com/p/phonetisaurus/>.

Oh, Y., Kim, H.K., 2009. MLLR/MAP adaptation using pronunciation variation for non-native speech recognition. In: Proc. of ASRU, pp. 216–221.

Pinto, J.P., Sivaram, G.S.V.S., Magimai-Doss, M., Hermansky, H., Bourlard, H., 2011. Analysis of MLP based hierarchical phoneme posterior probability estimator. I EEE Trans. Audio Speech Language Process. 19, 225–241.

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

Rasipuram, R., 2014. Grapheme-based automatic speech recognition using probabilistic lexical modeling. Ph.D. Theisis, École Polytechnique Fédérale de Lausanne (EPFL). <http://www.idiap.ch/rramya/Rasipuram_thesis.pdf>.

Rasipuram, R., Magimai-Doss, M., 2013a. Improving grapheme-based ASR by probabilistic lexical modeling approach. In: Proc. of Interspeech.

Rasipuram, R., Magimai-Doss, M., 2013b. Probabilistic lexical modeling and grapheme-based automatic speech recognition. Idiap Research Report. <http://publications.idiap.ch/downloads/reports/2013/Rasipuram_Idiap-RR-15-2013.pdf>.

Rasipuram, R., Bell, P., Magimai-Doss, M., 2013a. Grapheme and multilingual posterior features for under-resourced speech recognition: a study on scottish gaelic. In: Proc. of ICASSP.

Rasipuram, R., Razavi, M., Magimai-Doss, M., 2013b. Probabilistic lexical modeling and unsupervised training for zero-resourced ASR. In: Proc. of ASRU.

Razavi, M., Rasipuram, R., Magimai-Doss, M., 2014. On modeling context-dependent clustered states: comparing HMM/GMM, hybrid HMM/ANN and KL-HMM approaches. In: Proc. of ICASSP.

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

Schukat-Talamazzini, E., Niemann, H., Eckert, W., Kuhn, T., Rieck, S., 1993. Automatic speech recognition without phonemes. In: Proc. of EUROSPEECH.

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

Segura, J., Ehrette, T., Potamianos, A., Fohr, D., Illina, I., Breton, P.A., Clot, V., Gemello, R., Matassoni, M., Maragos, P., 2007. 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>.

Sim, K., 2009. Discriminative product-of-expert acoustic mapping for cross-lingual phone recognition. In: Proc. of ASRU, pp. 546–551.

Sim, K., Li, H., 2008. Robust phone set mapping using decision tree clustering for cross-lingual phone recognition. In: Proc. of ICASSP, pp. 4309–4312.

Soldo, S., Magimai-Doss, M., Pinto, J.P., Bourlard, H., 2011. Posterior features for template-based ASR. In: Proc. of ICASSP, pp. 4864–4867.

Stolcke, A., Hwang, M., Lei, X., Morgan, N., Vergyri, D., 2006. Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons. In: Proc. of ICASSP, pp. 321–324.

Strik, H., Cucchiarini, C., 1999. Modeling pronunciation variation for ASR: a survey of the literature. Speech Commun. 29, 225–246.

Stüker, S., 2008a. Integrating thai grapheme based acoustic models into the ML-MIX framework – for language independent and cross-language ASR. In: Proc. of SLTU.

Stüker, S., 2008b. Modified polyphone decision tree specialization for porting multilingual grapheme based ASR systems to new languages. In: Proc. of ICASSP, pp. 4249–4252.

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

Thomas, S., Ganapathy, S., Hermansky, H., 2012. Multilingual MLP features for low-resource LVCSR systems. In: Proc. of ICASSP, pp. 4269–4272.

Thomas, S., Hermansky, H., 2010. Cross-lingual and multistream posterior features for low resource lvcsr systems. In: Proc. of Interspeech, pp. 877–880.

Young, S.J., Odell, J.J., Woodland, P.C., 1994. Tree-based state tying for high accuracy acoustic modelling. In: Proc. of HLT, pp. 307–312.

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