IDIAP RESEARCH REPORT

PROBABILISTIC LEXICAL MODELING AND GRAPHEME-BASED AUTOMATIC SPEECH RECOGNITION

Ramya Rasipuram Mathew Magimai.-Doss

Idiap-RR-15-2013

APRIL 2013

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Probabilistic Lexical Modeling and Grapheme-based Automatic Speech Recognition

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

ramya.rasipuram@idiap.ch, mathew@idiap.ch

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

Abstract

Standard hidden Markov model (HMM) based automatic speech recognition (ASR) systems use phonemes as subword units. Thus, development of ASR system for a new language or domain depends upon the availability of a phoneme lexicon in the target language. In this paper, we introduce the notion of probabilistic lexical modeling and present an ASR approach where a) first, the relationship between acoustics and phonemes is learned on available acoustic and lexical resources (not necessarily from the target language or domain), and then b) probabilistic grapheme-to-phoneme relationship is learned using the acoustic data of targeted language or domain. The resulting system is a grapheme-based ASR system. This brings in two potential advantages. First, development of lexicon for target language or domain becomes easy i.e., creation of a grapheme lexicon where each word is transcribed by its orthography. Second, the ASR system can exploit both acoustic and lexical resources of multiple languages and domains. We evaluate and show the potential of the proposed approach through a) an in-domain study, where acoustic and lexical resources of target language or domain are used to build an ASR system, b) a monolingual cross-domain study, where acoustic and lexical resources of another domain are used to build an ASR system for a new domain, and c) a multilingual cross-domain study, where acoustic and lexical resources of multiple languages are used to build multi-accent non-native speech recognition system.

Keywords:
Automatic speech recognition, Kullback-Leibler divergence based hidden Markov model, grapheme, phoneme, lexical modeling, lexicon, non-native speech recognition.

1. Introduction

Standard hidden Markov Model (HMM) based automatic speech recognition (ASR) systems generally model words in terms of subword units. ASR systems largely use linguistically motivated subword units, typically *phonemes/phones*[1]. A lexicon is used to map the orthographic transcription of a word to sequence of phonemes. Thus, lexicon is one of the prior resources that is needed to build an ASR system. Generally, lexicon is developed by applying grapheme-to-phoneme conversion rules obtained from the linguistic studies and later manually verfied. Therefore, development of lexicon using phonemes as subword units not only requires linguistic resources but also minimum phonetic expertise. In other words, phoneme lexicon development is a semi-automatic process.

An alternative to phoneme subword units is graphemes[2], which makes lexicon development easy [2, 3, 4], [5, Chapter 4], [6, 7, 8, 9, 10, 11, 12, 13, 14]. However, modeling the relationship between graphemes and standard spectral-based feature observations, such as PLP cepstral coefficients which capture phoneme related information (from envelop of short-term spectrum), is not always trivial. The reason being grapheme-to-phoneme (G2P) relationship depends upon the language. For language such as Spanish the relationship is regular, while for language such as English the relationship is irregular. To overcome the problem of irregular relationship, in literature modeling of context-dependent graphemes has been explored [2, 3, 4], [5, Chapter 4], [6, 7]. The implicit assumption here being that relationship between context-independent graphemes and context-independent phonemes can be irregular, but relationship between context-dependent graphemes and context-independent phonemes could be regular. However, in case of languages like English, context-dependent grapheme-based ASR systems have been still found to yield considerably lower performance compared to context-dependent phoneme-based ASR systems [2, 3, 4, 6, 7].

In practice, to build ASR systems for new languages and domains, typically, a two stage approach is taken. More precisely, first, a G2P converter [15, 16, 17, 18, 19] is used to build a phoneme lexicon for the words in the new domain, and then an ASR system is trained. The G2P convertor is typically trained independent of the ASR system on "existing"

[1] A phoneme is the smallest contrastive unit in the phonology of a language, while phone is the acoustic realization of phoneme [1]. For sake of clarity, we will use the term phoneme.

[2] Graphemes are alphabets of a language.

lexical resources. Thus, this process assumes a minimal availability of lexical resources for the language of interest. Furthermore, the output of G2P converter may have to be manually verified to avoid ASR performance degradation due to pronunciation errors [20].

This paper introduces the notion of *probabilistic lexical modeling* and presents an ASR approach, where unlike the two stage approach mentioned above, first, acoustic-to-phoneme relationship is modeled with available acoustic and lexical resources (not necessarily from the language or domain of interest), and then probabilistic G2P relationship is learned given the acoustic information. In doing so, the proposed approach results in a grapheme-based ASR system that a) provides the advantages of grapheme lexicon, b) has the capability to handle pronunciation errors and pronunciation variation, and c) can exploit both acoustic and lexical resources of multiple languages and domains. Finally, the proposed approach could potentially remove the necessity for training an explicit G2P converter.

In Section 2, we provide a brief background on HMM-based ASR to elucidate that a) in HMM-based there are two kind of states, namely, acoustic states and lexical states, and b) in standard HMM-based ASR systems the relationship between the acoustic states and the lexical states is *deterministic*. In Section 3, we reformulate HMM-based ASR in terms of acoustic states and lexical states for both *likelihood-based* approach and *posterior-based* approach, and introduce the framework for probabilistic lexical modeling. In Section 4, we show that a) standard HMM-based ASR system, b) recently proposed KL-HMM system which uses phoneme class conditional probabilities as feature observation [21], c) tied posterior approach [22], d) probabilistic classification of HMM states (PCHMM) approach [23] to name a few are special cases of the probabilistic lexical modeling framework.

Section 5 motivates and presents an overview of the experimental studies. In Section 6, probabilistic lexical modeling approaches are evaluated on in-domain ASR task where acoustic and lexical resources of the target domain are used to build an ASR system. ASR studies indeed show that the proposed grapheme-based ASR system can learn the probabilistic grapheme-to-phoneme relationship and yields a competitive system. In Section 7, cross domain acoustic and lexical resources are used to build a grapheme-based ASR system for a new domain. In this case, the proposed grapheme-based system is compared with conventional approach where first, grapheme-to-phoneme conversion is performed and then a phoneme-based system is trained. Finally, in Section 8, we investigate the potential of the proposed ASR approach to exploit cross-domain multilingual acoustic and lexical resources for a multi-accent non-native speech recognition task. Sec-

tion 9 presents an analysis to show how probabilistic lexical modeling captures the relationship between grapheme lexical states and phoneme acoustic states. Finally, in Section 10 we present discussion and conclude.

2. HMM-based ASR

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

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

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

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

where $\mathcal{W}$ denotes the set of all possible word sequences, W denotes a word sequence, $\Theta = \{\Theta_A, \Theta_L\}$ denotes the set of parameters, more specifically, acoustic model and lexical parameters set Θ_A and language model parameter set Θ_L.

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

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

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

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

where $\mathcal{Q}$ denotes a set of possible HMM state sequences and $Q = \{q_1, \cdots, q_t, \cdots, q_T\}$ denotes a sequence of HMM states. Eqn. (5) results after *i.i.d* and first order Markov assumptions. Usually, $\log p(\mathbf{x}_t|q_t, \Theta_A)$ is referred to as *local emission score* and $\log P(q_t|q_{t-1}, \Theta)$ is referred to as *transition score*.

The HMM-based ASR literature is mainly dominated by the approach of modeling state emission likelihood. However, in theory, HMMs can be also

trained and decoded using a posteriori probabilities $P(q_t|\mathbf{x}_t, \Theta_A)$ as emission probabilities [24, 25]. We differentiate between these two approaches by referring them as *likelihood-based approach* and *posterior-based approach*, respectively.

Irrespective of the approach taken, in practice, in HMM-based ASR system there are two kinds of states, namely *acoustic states* denoted as q_t^{aco} corresponding to acoustic model and *lexical states* denoted as q_t^{lex} corresponding to lexical model. For instance,

- in tied state context-dependent HMM/GMM system or HMM/context-dependent neural network system [26], the acoustic states are the clustered states (also known as physical states) and the lexical states are the states of context-dependent subword models (also known as logical states), e.g. /k/-/ae/+/t/, /b/-/ae/+/t/.
- in context-independent subword unit based hybrid HMM/ANN system, typically during the training phase the ANN is trained to classify D context-independent phonemes, and during the decoding phase a minimum duration constraint is applied for each phoneme [25]. In this case, there are D acoustic states and $n \cdot D$ lexical states.

The parameter set $\Theta_A = \{\theta_a, \theta_{pr}, \theta_l\}$ can be partitioned as acoustic model parameters θ_a, prior lexical knowledge parameters θ_{pr}, and lexical model parameters θ_l. The acoustic model parameters θ_a consists of set of acoustic states $\{1, \cdots, D\}$ and its corresponding parameters. In the case of HMM/Gaussian mixture model (HMM/GMM) system, it includes the means, the variances and the mixture weights of Gaussians for each acoustic state, and in the case of hybrid HMM/artificial neural network (HMM/ANN) system it includes ANN weights and biases [25]. The parameter set θ_{pr} consists of set of subword units, pronunciation models of the words (i.e., lexicon) and set of lexical states $\{1, \cdots, I\}$ where I denotes the number of lexical states. The lexical model parameter set θ_l consists of a set parameters $\{\mathbf{y}_i\}_{i=1}^{I}$ that capture the relationship between acoustic states and lexical states.

Having defined the parameter set Θ_A, the search for most likely sequence of states

1. in the case of likelihood-based approach can be rewritten as

$$Q_{lex}^* = \underset{Q_{lex} \in \mathcal{Q}_{\updownarrow}}{\arg\max} \sum_{t=1}^{T} [\log p(\mathbf{x}_t|q_t^{lex} = i, \Theta_A) + \log P(q_t^{lex} = i|q_{t-1}^{lex} = h, \Theta)] \tag{7}$$

2. in the case of posterior-based approach can be rewritten as

$$Q^*_{lex} = \arg\max_{Q_{lex} \in \mathcal{Q}_\updownarrow} \sum_{t=1}^{T} [\log P(q_t^{lex} = i|\mathbf{x}_t, \Theta_A) + \log P(q_t^{lex} = i|q_{t-1}^{lex} = h, \Theta)] \tag{8}$$

where Q_{lex} denotes a sequence of lexical states $\{q_1^{lex}, \cdots, q_t^{lex}, \cdots, q_T^{lex}\}$, $i, h \in \{1, \cdots, I\}$, $\mathcal{Q}_\updownarrow$ is a set of all possible lexical state sequences.

In HMM-based ASR system, in practice, the relationship between lexical state q_t^{lex} and acoustic observation $\mathbf{x}_t$ is not always modeled directly. As indicated earlier, the relationship is typically modeled through intermediate acoustic states q_t^{aco} and using lexical model parameters θ_l. The acoustic states can be more abstract and data-driven such as, clustered states of context-dependent subword unit based ASR system, or they can be same as lexical states i.e., lexical knowledge-driven, like in the case of context-independent subword unit based ASR system. In standard HMM-based ASR systems the lexical model is typically deterministic, i.e., θ_l is a table that maps a lexical state onto an acoustic state. So, if lexical state i is mapped to acoustic state j, according to the table, then $p(\mathbf{x}_t|q_t^{lex} = i, \Theta_A) = p(\mathbf{x}_t|q_t^{aco} = j, \theta_a)$ and $P(q_t^{lex} = i|\mathbf{x}_t, \Theta_A) = P(q_t^{aco} = j|\mathbf{x}_t, \theta_a)$ in Eqn. (7) and Eqn. (8), respectively, where $j \in \{1, \cdots, D\}$. This aspect is implicit in many ASR systems.

In the following section, we present a reformulation of HMM-based ASR in terms of lexical states and acoustic states and introduce the notion of probabilistic lexical modeling where the relationship between the lexical states and the acoustic states is probabilistic.

3. Reformulation of HMM-based ASR in terms of Acoustic States and Lexical States

In Section 3.1, we formulate estimation of lexical state emission likelihood (in the case of likelihood-based approach) or emission probability (in the case of posterior-based approach) using acoustic and lexical information. In Section 3.2, we elucidate that in the log-likelihood space, HMM-based ASR approach can be seen as a template matching approach, where a sequence corresponding to lexical evidence is compared with a sequence corresponding to acoustic evidence. Central to both these formulations is an underlying idea that lexical model parameter $\mathbf{y}_i$ of each lexical state i is a conditional probability vector $[y_i^1, \cdots, y_i^d, \cdots, y_i^D]^{\mathrm{T}}$, where $y_i^d = P(q_t^{aco} = d|q_t^{lex} = i, \theta_a, \theta_{pr})$, $0 \leq y_i^d \leq 1$ and $\sum_d y_i^d = 1$.

3.1. Estimation of lexical state likelihood or a posteriori probability

In likelihood-based approach, likelihood of a lexical state i can be estimated as,

$$\begin{aligned} p(\mathbf{x}_t|q_t^{lex}=i,\Theta_A) &= \sum_{d=1}^{D} p(\mathbf{x}_t, q_t^{aco}=d|q_t^{lex}=i,\Theta_A) \\ &= \sum_{d=1}^{D} P(q_t^{aco}=d|q_t^{lex}=i,\theta_a,\theta_{pr}) \cdot p(\mathbf{x}_t|q_t^{aco}=d,q_t^{lex}=i,\Theta_A) \\ &\approx \sum_{d=1}^{D} P(q_t^{aco}=d|q_t^{lex}=i,\theta_a,\theta_{pr}) \cdot p(\mathbf{x}_t|q_t^{aco}=d,\theta_a) \end{aligned} \tag{9}$$

where, $p(\mathbf{x}_t|q_t^{aco}=d)$ denotes likelihood of acoustic state d, $P(q_t^{aco}=d|q_t^{lex}=i,\theta_a,\theta_{pr})$ denotes the probability of acoustic state d given lexical state i, D is the number of acoustic states and I is the number of lexical states. Eqn. (9) assumes that $\mathbf{x}_t$ is independent of lexical state q_t^{lex} given the acoustic state q_t^{aco}.

Along similar lines, in the posterior-based approach, a posteriori probability of a lexical state i could be estimated as,

$$P(q_t^{lex}=i|\mathbf{x}_t,\Theta_A) = \sum_{d=1}^{D} P(q_t^{aco}=d|q_t^{lex}=i,\theta_a,\theta_{pr}) \ \cdot P(q_t^{aco}=d|\mathbf{x}_t,\theta_a) \tag{10}$$

where $P(q_t^{aco}=d|\mathbf{x}_t)$ is the posterior probability of acoustic state d given the acoustic observation $\mathbf{x}_t$.

In case of deterministic lexical modeling $\mathbf{y}_i$ is a Kronecker delta distribution, i.e., if lexical state i is mapped onto acoustic state j then,

$$y_i^d = P(q_t^{aco}=d|q_t^{lex}=i,\theta_a,\theta_{pr}) = \begin{cases} 1 & \text{if } d=j \\ 0 & \text{if } d\neq j \end{cases} \tag{11}$$

Applying Eqn. (11) in Eqn. (9) and Eqn. (10), we can see that lexical state likelihood is $p(\mathbf{x}_t|q_t^{aco}=j,\theta_a)$ and lexical state posterior probability is $P(q_t^{aco}=j|\mathbf{x}_t,\theta_a)$, respectively.

3.2. HMM-based ASR: a template matching approach

The HMM-based ASR theory though has been developed in likelihood space, in practice, ASR systems operate in log-likelihood space. One of

the main reason for this is to handle numerical problems. In this section, we will see that, with the reformulation presented in the previous section, log-likelihood space leads to better understanding of HMM-based ASR approach. More specifically, we will see that HMM-based ASR can be seen as a template matching approach, where the reference template or sequence is based on lexical information and the test template or sequence is based on acoustic information.

Briefly, in template matching, given a reference template $A = \{a_1, \cdots, a_m, \cdots, a_M\}$ and a test template $B = \{b_1, \cdots, b_n, \cdots, b_N\}$, a global score $G(A, B)$ is obtained by matching the templates using dynamic programming (with local constraints and boundary constraints) in the following manner:

1. initial condition: $C(1, 1) = S(a_1, b_1)$
2. for each pair of matching point (m, n) in the reference template and the test template, applying the recurrence equation

$$C(m,n) = S(a_m, b_n) + \max \begin{cases} tc(m-1,n) + C(m-1,n) \\ tc(m-1,n-1) + C(m-1,n-1) \\ tc(m,n-1) + C(m,n-1) \end{cases} \tag{12}$$

3. final condition: $G(A, B) = C(M, N)$

where $C(m, n)$ is the cumulative score and $S(a_m, b_n)$ is the local score at matching point (m, n), and $tc(m-1, n-1)$ is the cost to transit from matching point $(m-1, n-1)$ to matching point (m, n). It is to be noted that the second part in the above equation is based on the local constraints applied. In HMM-based ASR, the local constraints are defined by the topology of HMM. Furthermore, whether to maximize or minimize the cumulative score depends upon how the local score or transition costs are computed.

Finding the optimal state sequence Q^*[3], see Eqn. (7) or Eqn. (8), involves dynamic programming. Comparing Eqn. (12) with Eqn. (7) and Eqn. (8), it can be seen that the local score $S(,)$ in likelihood-based approach is $\log p(\mathbf{x}_t|q_t^{lex} = i, \Theta_A)$ and in posterior-based approach is $\log P(q_t^{lex} = i|\mathbf{x}_t, \Theta_A)$, respectively and the transition score $tc(,)$ is $\log P(q_t^{lex} = i|q_{t-1}^{lex} = h, \Theta)$.

[3]The problem of finding the optimal state sequence is also used during training of HMMs, for instance, in Viterbi EM training [27]

Based on the formulation presented in the previous section, we can see that the estimation of local score for both likelihood-based approach and posterior-based approach is matching of lexical evidence and acoustic evidence. More precisely, in

- likelihood-based approach

$$\log p(\mathbf{x}_t|q_t^{lex} = i, \Theta_A) = \log(\mathbf{y}_i^{\mathrm{T}}\mathbf{v}_t) = S(\mathbf{y}_i, \mathbf{v}_t) \tag{13}$$

  where $\mathbf{v}_t = [p(\mathbf{x_t}|q_t^{aco} = 1, \theta_a), \cdots, p(\mathbf{x_t}|q_t^{aco} = d, \theta_a), \cdots, p(\mathbf{x_t}|q_t^{aco} = D, \theta_a)]^{\mathrm{T}}$ or simply denoted as $[v_t^1, \cdots, v_t^d, \cdots, v_t^D]^{\mathrm{T}}$.

- posterior-based approach

$$\log P(q_t^{lex} = i|\mathbf{x}_t, \Theta_A) = \log(\mathbf{y}_i^{\mathrm{T}}\mathbf{z}_t) = S(\mathbf{y}_i, \mathbf{z}_t) \tag{14}$$

  where $\mathbf{z}_t = [P(q_t^{aco} = 1|\mathbf{x}_t, \theta_a), \cdots, P(q_t^{aco} = d|\mathbf{x}_t, \theta_a), \cdots, P(q_t^{aco} = D|\mathbf{x}_t, \theta_a)]^{\mathrm{T}}$ or simply denoted as $[z_t^1, \cdots, z_t^d, \cdots, z_t^D]^{\mathrm{T}}$.

In other words, HMM-based ASR approach can be seen as a template matching approach where,

1. the reference template is a sequence of lexical state parameter vectors $U = \{\mathbf{u}_1, \cdots, \mathbf{u}_m, \cdots, \mathbf{u}_M\}$ corresponding to a word sequence obtained by using lexical information. M denotes the number of states in the sequence and $\mathbf{u}_m \in \{\mathbf{y}_1, \cdots, \mathbf{y}_I\}$.
2. the test template is sequence $V = \{\mathbf{v}_1, \cdots, \mathbf{v}_t, \cdots, \mathbf{v}_T\}$ (in the case of likelihood-based approach) or $Z = \{\mathbf{z}_1, \cdots, \mathbf{z}_t, \cdots, \mathbf{z}_T\}$ (in the case of posterior-based approach).
3. the cost function $S(,)$ to estimate local score is scalar product.

Viewing HMM-based ASR as template matching approach provides venues for exploring new ways to match lexical evidence and acoustic evidence. In the likelihood-based approach, it can be observed that we are matching a conditional probability vector $\mathbf{y}_i$ with a likelihood vector $\mathbf{v}_t$, i.e., two statistical quantities that have different properties. For instance, $\sum_{d=1}^{D} y_i^d = 1$ but $\sum_{d=1}^{D} v_t^D \gtreqless 1$. Thus, it is difficult to visualize a cost function other than scalar product.

Unlike likelihood-based approach, in the case of posterior-based approach it can be observed that both reference vector $\mathbf{y}_i$ and the test vector $\mathbf{z}_t$ are conditional probability vectors, i.e., the reference vector and test vector lie in the same acoustic state posterior probability space, but conditioned on different information. In the reference vector, the acoustic state posterior

probability is conditioned on prior lexical information, while in the test vector, acoustic state posterior probability is conditioned on the feature observation $\mathbf{x}_t$. As a consequence, in posterior-based approach two sequences of conditional probability vectors are matched similar to the recently proposed template-based ASR approach using "posterior features" [28, 29].

In the template-based ASR approach using posterior features, it has been observed that there are different cost functions that can be used to locally match lexical evidence and acoustic evidence. For instance, Kullback-Leibler (KL) divergence, Bhattacharya distance, cosine distance, scalar product, cross entropy [29, 30]. Furthermore, it has been observed that local score based on KL-divergence or Bhattacharya distance tend to yield better system than scalar product. This indicates that posterior-based approach could also benefit from the use of alternate cost functions. Indeed, we will see this aspect in this paper.

As per the framework presented in both Sections 3.1 and 3.2, to perform ASR we need to estimate $\mathbf{y}_i$, and $\mathbf{v}_t$ or $\mathbf{z}_t$ depending upon the approach. The acoustic model i.e., $\mathbf{v}_t$ estimator or $\mathbf{z}_t$ estimator can be based on GMMs or ANNs, irrespective of what is estimated. ANNs can directly estimate a posteriori probabilities [25]. In other words, the output of ANNs is a posteriori probability of acoustic states. So, in the likelihood-based approach, when using ANNs, the likelihood of acoustic states is replaced by scaled-likelihood

$$\begin{aligned} p_{sl}(\mathbf{x}_t|q_t^{aco}=d,\theta_a) &= \frac{p(\mathbf{x}_t|q_t^{aco}=d,\theta_a)}{p(\mathbf{x}_t|\theta_a)} \\ &= \frac{P(q_t^{aco}=d|\mathbf{x}_t,\theta_a)}{P(q_t^{aco}=d|\theta_a)} \end{aligned} \tag{15}$$

Similarly GMMs estimate likelihood of acoustic states. So, in the case of posterior-based approach, when using GMMs, the a posteriori probability of acoustic states can be estimated using Bayes rule. In short, likelihood-based approach and posterior-based approach is equally applicable to HMM/GMM system and hybrid HMM/ANN system. The only difference is that the acoustic model in one case is generative and in another case is discriminative, respectively. In the following section, we present approaches to estimate the lexical model parameter set $\theta_l = \{\mathbf{y}_i\}_{i=1}^{I}$.

4. Probabilistic Lexical Modeling: KL-HMM and Related Approaches

In standard HMM-based ASR systems, most often the relationship between lexical states and acoustic states is one-to-one, i.e., $\mathbf{y}_i$ is a Kronecker

delta distribution. This can happen a) in context-dependent subword unit based ASR system while decision tree state tying, b) in context-independent subword unit based ASR system when prior knowledge such as, minimum duration constraint is applied, and c) even when the relationship between the lexical states and acoustic observations is directly modeled. In this case, lexical states and acoustic states are same entities and $D = I$. The output of ANNs or GMMs can be seen as acoustic evidence matched with deterministic lexical evidence. In this section, we present ASR approaches where $\mathbf{y}_i$ is no more deterministic, rather, $\mathbf{y}_i$ is learned from the acoustic data given the acoustic model, i.e., estimator of $\mathbf{v}_t$ or $\mathbf{z}_t$. We refer to such approaches as *probabilistic lexical modeling* approaches. In that regard, we first present and show that KL-HMM approach is a posterior-based probabilistic lexical modeling approach. We then present a few other related probabilistic modeling approaches.

4.1. Kullback-Leibler divergence based HMM

Kullback-Leibler divergence based HMM (KL-HMM) is a recently proposed approach where $\mathbf{z}_t = [z_t^1, \cdots, z_t^d, \cdots, z_t^D]^{\mathrm{T}}$, the phoneme class conditional probability vector or the acoustic state probability vector (originally it was referred to as *posterior feature*) estimated by an ANN is used directly as feature observation for a second HMM whose states are parameterized by categorical distributions $\mathbf{y}_i = [y_i^1, \cdots y_i^d, \cdots, y_i^D]^{\mathrm{T}}$ [21, 31]. Figure 1 illustrates the KL-HMM approach.

The local score $S(\mathbf{y}_i, \mathbf{z}_t)$ at each HMM state i is

$$S(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} y_i^d \log(\frac{y_i^d}{z_t^d}) \tag{16}$$

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

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

$$S(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} z_t^d \log(\frac{z_t^d}{y_i^d}) \tag{17}$$

2. Symmetric KL-divergence (SKL):

$$S(\mathbf{y}_i, \mathbf{z}_t) = \frac{1}{2} \cdot [KL + RKL] \tag{18}$$

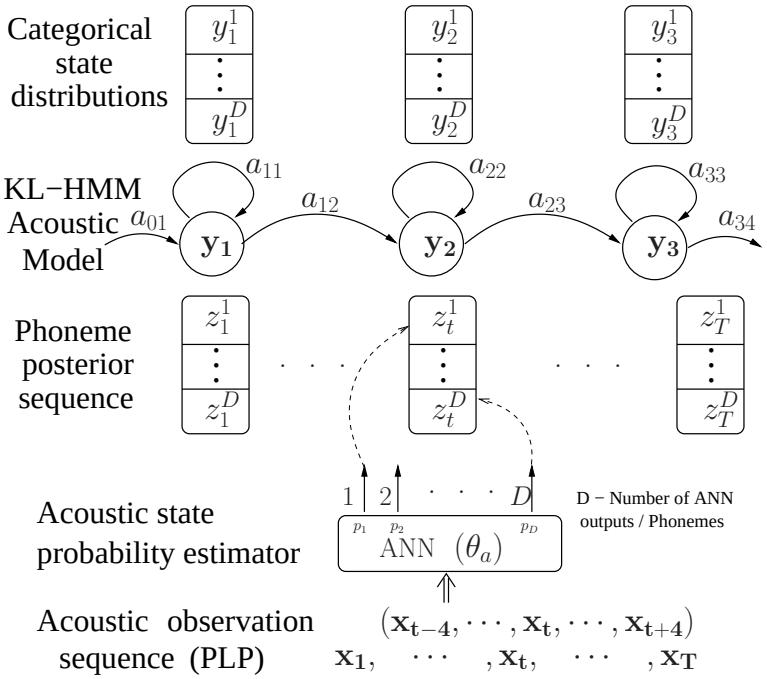

Figure 1: Illustration of KL-HMM acoustic model

It is clear from the formulation presented in Section 3, especially Section 3.2, that KL-HMM is a posterior-based ASR system, where the lexical model is probabilistic and instead of scalar product, a cost function based on KL-divergence is used. In theory, KL-divergence can be linked to hypothesis testing [32, 33]. More precisely, KL-divergence between two distributions is an estimate of expected log-likelihood ratio. As a consequence, irrespective of how $\mathbf{z}_t$ is estimated i.e., using a generative model (GMM) or discriminative model (ANN), locally (in time) KL-HMM is a discriminative model. More precisely, lexical evidence and acoustic evidence are discriminatively matched locally. However, globally KL-HMM is still a generative model in the acoustic state probability vector space, i.e., $\mathbf{z}_t$.

KL-HMM approach can also be interpreted in information theoretic sense. For instance, local score KL is the expected number of extra bits needed if $\mathbf{y}_i$ is the true distribution and the acoustic state symbols are coded by distribution $\mathbf{z}_t$. Similarly, local score RKL is the expected number of extra bits needed if $\mathbf{z}_t$ is the true distribution and the acoustic state symbols are coded by distribution $\mathbf{y}_i$. From this interpretation, it can be observed

that a) local score KL gives emphasis to the lexical model. This is good when lexical information is reliable; b) local score RKL gives emphasis to the acoustic model. This is good when lexical information or model is less reliable. Indeed, later in Section 6 we will see that for context-independent graphemes local score RKL is better; and c) local score SKL gives equal emphasis to the lexical model and the acoustic model. Dealing these aspects in more detail is out-of-the-scope of the present paper.

4.1.1. Training

KL-HMM is fully parameterized by $\Theta_{kull} = \{\{\mathbf{y}_i\}_{i=1}^{I}, \{a_{ij}\}_{i,j=1}^{I}\} = \{\theta_l, \{a_{ij}\}_{i,j=1}^{I}\}$ where I is the total number of states and state i is represented by categorical distribution $\mathbf{y}_i$, a_{ij} is the transition probability from state i to state j. Given a training set of N utterances, where each training utterance n is a sequence of acoustic state probability vectors $Z(n) = \{\mathbf{z}_1(n), \cdots, \mathbf{z}_t(n), \cdots, \mathbf{z}_{T(n)}(n)\}$ of length $T(n)$, the parameters Θ_{kull} are estimated by Viterbi expectation maximization algorithm which minimizes the cost function,

$$\min_{Q \in \mathcal{Q}} \sum_{n=1}^{N} \sum_{t=1}^{T(n)} [S(\mathbf{y}_{q_t}, \mathbf{z}_t(n)) - \log a_{q_{t-1}q_t}] \tag{19}$$

where $q_t \in \{1, \cdots, I\}$, $\mathcal{Q}$ denotes set of all possible HMM state sequences, $Q = \{q_1, \cdots q_t, \cdots, q_T(n)\}$ denotes a sequence of HMM states and $\mathbf{z}_t(n) = [z_t^1(n), \cdots, z_t^d(n), \cdots, z_t^D(n)]^T$. More precisely, the training process involves iteration over the segmentation and the optimization steps until convergence. Given an estimate of Θ_{kull}, the segmentation step yields an optimal state sequence for each training utterance using Viterbi algorithm. The optimization step then estimates new set of model parameters given the optimal state sequences, i.e., alignment and $\mathbf{z}_t$ belonging to each of these states. As mentioned earlier, there are different possible local scores based on KL-divergence. Each of these local scores lead to a different optimal state categorical distribution [34],

1. For local score KL (Equation (16)), the optimal lexical state distribution is the normalized geometric mean of the training acoustic state probability vectors assigned to the state. More precisely,

$$y_i^d = \frac{\bar{y}_i^d}{\sum_{d=1}^{D} \bar{y}_i^d} \text{ where } \bar{y}_i^d = \Big(\prod_{\mathbf{z}_t(n) \in Z(i)} z_t^d(n) \Big)^{\frac{1}{M(i)}} \quad \forall n, t \tag{20}$$

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

2. For local score RKL (Equation (17)), the optimal lexical state distribution is the arithmetic mean of the training acoustic state probability vectors assigned to the state, i.e.,

$$y_i^d = \frac{1}{M(i)} \sum_{\mathbf{z}_t(n) \in Z(i)} z_t^d(n) \quad \forall n, t \tag{21}$$

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

3. For local score SKL (Equation (18)), there is no closed form solution to find the optimal lexical state distribution. The optimal lexical state distribution can be computed iteratively using the arithmetic and the normalized geometric mean of the acoustic state probability vectors assigned to the state [35].

4.1.2. Decoding

The decoding is performed using standard Viterbi decoder. Given a sequence of acoustic state probability vectors $Z = \{\mathbf{z}_1, \cdots, \mathbf{z}_t, \cdots, \mathbf{z}_T\}$ and the trained parameters $\Theta_{kull} = \{\{\mathbf{y}_i\}_{i=1}^I, \{a_{ij}\}_{i,j=1}^I\}$, decoding involves recognition of the underlying hypothesis $\hat{m}$:

$$\hat{m} = \arg\min_{Q \in \mathcal{Q}} \sum_{t=1}^{T} [S(\mathbf{y}_{q_t}, \mathbf{z}_t) - \log a_{q_{t-1} q_t}] \tag{22}$$

where $\mathcal{Q}$ denotes the set of possible state sequences allowed by the hypothesis m.

4.2. Related Approaches

Probabilistic classification of HMM states [23] is a probabilistic lexical modeling approach in framework of HMM/GMM system. In this approach, first, decision tree clustering of context-dependent phonemes is performed and the resulting leaf nodes are chosen as HMM states. Later, state-to-class probabilities are estimated using EM algorithm along with GMM parameters. The approach can be seen as likelihood-based probabilistic lexical modeling in the framework of HMM/GMM system, where lexical states are context-dependent phonemes, acoustic states are decision tree clustered HMM states, and the emission likelihood is computed using Eqn. (9). In [36],

this approach was applied to model pronunciation variability in spontaneous speech.

Hidden sequence modeling [37] is an approach where each phoneme is represented by a mixture of HMM state sequences corresponding to different variants. State level probabilistic modeling and hidden model sequences compensate for substitution, insertion or deletion errors resulting because of pronunciation variation.

Tied posterior [22] is an approach that was proposed in the framework of hybrid HMM/ANN to build context-dependent subword unit based ASR system using an ANN trained to classify context-independent phoneme. In this approach, the emission likelihood at each context-dependent state $q_t^{cd} = i$ is estimated as

$$p(\mathbf{x}_t|q_t^{cd} = i) = \sum_{d=1}^{D} w_i^d \cdot p_{sl}(\mathbf{x}_t|q_t^{ci} = d) \tag{23}$$

where $p_{sl}(\mathbf{x}_t|q_t^{ci} = d)$ is the scale-likelihood (see Eqn. 15) given context-independent phoneme d, D is the number of context-independent phonemes, $0 \leq w_i^d \leq 1$ is the weight given to scaled-likelihood of context-independent phoneme d at state $i \in \{1, \cdots I\}$ and $\sum_d w_i^d = 1$. The weights are estimated by maximizing log-likelihood. Comparing Eqn. (23) with Eqn. (9), it can be seen that tied posterior approach is a likelihood based probabilistic lexical modeling approach, where q_t^{cd} is the lexical state, q_t^{ci} is the acoustic state and w_i^d is the lexical model parameter y_i^d.

In KL-HMM system, the local score is based on KL-divergence. However, as observed in Section 3.2, different cost functions can be used in the case of posterior based approach. So, it is possible to train a posterior-based system along similar lines of KL-HMM by replacing the local score based on KL-divergence by the local score based on scalar product (see Eqn. (14)). The resulting ASR system is consistent with the posterior-based ASR approach presented in Section 3.1. We refer to this system as scalar product HMM (SP-HMM). SP-HMM could be also seen as a special case of tied posterior approach where the priors of the acoustic states are dropped or assumed to be equal.

5. Overview of Grapheme-based ASR Studies

The present paper investigates the potential of probabilistic lexical modeling to build grapheme-based ASR system. In that regard, we study an approach where the lexical states are graphemes and the acoustic states

are phonemes. We hypothesize that with probabilistic lexical modeling it is possible to build ASR system which uses grapheme lexicon and achieves performance as good as ASR systems, where first grapheme-to-phoneme conversion is performed to build a lexicon and then an ASR system (with probabilistic lexical modeling capabilities) is trained. Towards that goal, we present different studies which compare ASR system using grapheme lexicon with ASR system using phoneme lexicon, namely,

1. in-domain ASR study (Section 6), where both the acoustic model parameters θ_a and the lexical model parameters θ_l are trained using acoustic and lexical resources of in-domain data.
2. cross-domain ASR study (Section 7), where the goal is to build an ASR system for a new domain using cross-domain acoustic and lexical resources. In this study, the acoustic model parameters θ_a are trained using acoustic and lexical resources (phoneme lexicon) of cross-domain, and the lexical model parameters θ_l are trained on in-domain data. In this case, the phoneme-based ASR system is built using phoneme lexicon developed by training a grapheme-to-phoneme converter on the cross-domain lexicon.
3. multi-accent non-native ASR study (Section 8), where there is limited availability of both acoustic and lexical resources. Towards that we show the potential of probabilistic lexical modeling to exploit acoustic and lexical resources of multiple languages. We investigate two cases,
   (a) in the first case, both acoustic model parameters θ_a and lexical model parameters θ_l are trained on out-of-domain data.
   (b) in the second case, the acoustic model parameters θ_a are trained using out-of-domain acoustic and lexical resources, and the lexical model parameters θ_l are trained with in-domain resources. Again, like the cross-domain study, here the phoneme-based ASR system is built by using a lexicon developed by training a grapheme-to-phoneme converter on the cross domain lexicon.

Finally, in Section 9 we analyze the lexical model parameters of grapheme-based system of in-domain study to elucidate the influence of cost function and subword context on the captured grapheme-to-phoneme relationship.

As discussed earlier in Section 3, irrespective of the approach used, i.e. likelihood-based or posterior-based, the acoustic observations can be modeled by either GMMs or ANNs. In this work, we use ANNs to model the acoustic observations. The main reasons being [25] a) ANNs can directly estimate a posteriori probabilities, b) ANNs can exploit acoustic context, and c) ANN is discriminatively trained and thus, has the capability to handle

undesirable variabilities such as, speaker, environment. Furthermore, in our studies ANN is always trained to classify context-independent phonemes.

We build and compare following systems a) hybrid HMM/ANN system (denoted as Hybrid); b) tied-posterior based system (denoted as Tied); c) KL-HMM system (denoted as KL-HMM); and d) Scalar product HMM (denoted as SP-HMM). It could be observed that all these systems use context-independent subword unit based acoustic model. We also build standard cross word context-dependent subword unit based HMM/GMM system with decision tree based state tying using HTK [38] and use it as a reference point for studies reported in Sections 6 and 7. Input to all the ANNs used in this paper is 39 dimensional perceptual linear prediction (PLP) cepstral coefficients ($c_0 - c_{12} + \Delta + \Delta\Delta$) with four frame preceding and four frame following context. The 39 dimensional PLP features used to train the ANN are also used to train the HMM/GMM systems. Table 1 summarizes the different systems and their capabilities.

Table 1: Overview of different systems. *post* denotes posterior-based ASR approach and *like* denotes likelihood-based ASR approach. CI denotes context-independent subword units and CD denotes context-dependent subword units. P and G denote phoneme lexicon and grapheme lexicon, respectively. *Det* denotes lexical model is deterministic and *Prob* denotes lexical model is probabilistic.

System	Appr -oach	θ_a Acoustic states	θ_{pr}		θ_l
			Lexicon	Lexical states	Lexical model
Hybrid	*like*	CI	P	CI	*Det*
KL-HMM	*post*	CI	P or G	CI or CD	*Prob*
SP-HMM	*post*	CI	P or G	CI or CD	*Prob*
Tied	*like*	CI	P or G	CI or CD	*Prob*
HMM/GMM	*like*	CD	P or G	CD	*Det*

We present ASR studies on English. In English, grapheme-to-phoneme relationship in English is highly irregular. Thus, presents a difficult case. As a result, learning the relationship between grapheme and phoneme is relatively difficult.

6. In-Domain ASR Study

We present in-domain ASR studies on medium vocabulary continuous speech recognition using DARPA Resource Management task. A part of this study is presented in [12]. The setup is exactly same as reported in [6]. The training set and development set consists of 2,880 utterances and 1,100

utterances, respectively, spoken by 109 speakers corresponding to approximately 3.8 hours of speech data. The test set contains 1,200 utterances amounting to 1.1 hours in total and is formed by combining Feb'89, Oct'89, Feb'91 and Sep'92 test sets. The test set is completely covered by a word pair grammar (perplexity 60) included in the task specification. In this work, the training set of 2,880 utterances is used to train the ASR systems as opposed to 3,990 utterances (formed by combining training and development set) usually reported in literature [37]. The development set 1110 utterances are used only to tune the word insertion penalty.

The lexicon consists of 991 words. The phoneme lexicon was obtained from UNISYN dictionary [39]. There are 42 context-independent phonemes including silence. The grapheme lexicon was transcribed using 29 context-independent graphemes (which includes silence, symbol hyphen and symbol single quotation mark).

We use an *off-the-shelf* ANN, more precisely multilayer perceptron, (reported in previous study [6]) trained on RM corpus to classify 45 context-independent phonemes as acoustic model. System Hybrid does not include any trainable lexical model parameters. The training phase of System KL-HMM, System SP-HMM and System Tied involves estimation of lexical parameters θ_l. Three different KL-HMM systems are built where the parameters $\mathbf{y}_i$ are estimated by minimizing cost function based on local scores KL, RKL and SKL, respectively.

In probabilistic lexical modeling framework, we model subword contexts *mono* (context-independent subword units), *tri* (context-dependent subword units with single preceding and following context) and *quint* (context-dependent subword units with two preceding and following contexts). For context-dependent studies, we train word internal context models. Each subword unit is modeled by a 3 state left-to-right HMM.

Phoneme-based HMM/GMM system uses phonetic question set for state tying, where as grapheme-based HMM/GMM system uses singleton question set. Decision tree based state tying resulted in 1220 clustered states for phoneme system and 1109 clustered states for grapheme system.

Table 2 compares the total number of parameters in different systems (Hybrid, KL-HMM and HMM/GMM) for phoneme and grapheme subword units. Lexical model parameter θ_l in the case of System HMM/GMM is deterministic, i.e. a table which maps each possible triphone or trigraph to a clustered state. Similarly, for System Hybrid the lexical model is deterministic, i.e., it is a table of 45x3 entries, where 3 is the minimum duration constraint. System SP-HMM and System Tied have same number of parameters as System KL-HMM . It can be noted that probabilistic lexical

modeling does not increase the acoustic model complexity across systems modeling lexical contexts *mono*, *tri* and *quint*. It only changes the lexical model complexity. Furthermore, it can be observed that System KL-HMM has fewer acoustic model parameters and more lexical model parameters compared to System HMM/GMM.

Table 2: Number of parameters for each system. θ_a denotes acoustic model parameters, θ_l denotes lexical model parameters. P, and G denote phoneme and grapheme subword unit based systems, respectively.

System (Context)	θ_a	θ_l (P)	Total (P)	θ_l (G)	Total (G)
Hybrid (*mono*)	0.5M	135	0.5M	–	–
KL-HMM (*mono*)	0.5M	6K	≈0.5M	4K	≈0.5K
KL-HMM (*tri*)	0.5M	0.3M	0.8M	0.2M	0.7M
KL-HMM (*quint*)	0.5M	0.5M	1.0M	0.5M	1.0M
HMM/GMM (*tri*)	0.8M	0.2M	1.0M	0.07M	0.87M

Table 3 presents the word error rate (WER) on the test set of RM corpus for various systems. ASR studies using System KL-HMM were first reported in [12]. It is to be noted that the hybrid HMM/ANN approach requires a one-to-one map between acoustic and lexical states, therefore, in order to model contexts *tri* and *quint*, context-dependent ANN needs to be trained. In this work, as we limit our studies to context-independent phoneme ANN, WER of hybrid HMM/ANN phoneme lexicon systems with context *tri* and *quint*, and WER of hybrid HMM/ANN grapheme lexicon systems for all contexts is not applicable.

The key observations are as follows,

- For phoneme lexicon with context *mono*, System Hybrid yields slightly poor performance compared to System KL-HMM SKL, System KL-HMM KL, System SP-HMM and System Tied which have probabilistic lexical modeling capabilities. The performance of phoneme-based KL-HMM systems modeling *tri* context increases compared to *mono* context for all local scores. However, modeling *quint* context does not always improve over *tri*.

- The grapheme-based systems in the framework of probabilistic lexical modeling yield significantly poor performance compared to their respective phoneme-based systems for context *mono*. However, as the

Table 3: Word error rate expressed in % on the test set of RM corpus for various systems with phonemes and graphemes as subword units. Boldface indicates the best performance obtained for each of the subword units.

System	Phoneme			Grapheme		
	Subword context			Subword context		
	mono	*tri*	*quint*	*mono*	*tri*	*quint*
Hybrid	7.4	–	–	–	–	–
KL-HMM KL	7.1	5.5	5.4	42.1	7.7	6.1
KL-HMM RKL	8.0	5.9	5.8	25.8	6.5	5.7
KL-HMM SKL	7.1	**5.1**	5.2	32.9	5.9	**5.2**
SP-HMM	6.9	6.0	6.0	22.9	7.1	6.3
Tied	6.8	5.8	5.8	21.2	6.5	5.5
HMM/GMM	10.5	**4.9**	–	36.0	7.1	–

context of grapheme units is increased, performance of the systems improve. Grapheme-based System KL-HMM SKL modeling *quint* context yields the best peformance, which is comparable to the performance of phoneme-based System KL-HMM SKL modeling *tri* context. This indicates that in the framework of probabilistic lexical modeling grapheme-based system could yield performance comparable to phoneme-based system.

- KL-HMM system based on local score SKL, except for context *mono*, performs better than System SP-HMM and System Tied. It can be observed that for context *mono* System Tied yields better performance compared to System SP-HMM and KL-HMM systems. This could be attributed to the scaling of ANN output by priors.

- It is interesting to note that System KL-HMM SKL, which uses context-independent phoneme acoustic states and probabilistic lexical model, performs close to System HMM/GMM, which uses context-dependent phoneme acoustic states and deterministic lexical model.

- Probabilistic lexical modeling approaches using grapheme lexicon outperform System HMM/GMM using grapheme lexicon.

7. Cross Domain ASR Study

In this section, we present cross-domain ASR study where cross-domain or task-independent acoustic and lexical resources are used to build ASR

system for a new task. In that regard, we present an experimental study where, Wall Street Journal 1 (WSJ1) corpus, which contains approximately 80 hours of speech data and 10K word lexicon (total 13K lexical entries), is used to build ASR system for RM task (presented in Section 6).

We use an *off-the-shelf* ANN [31] trained on WSJ1 (to classify 45 context-independent phonemes) as the acoustic model to estimate acoustic state probabilities of RM data. The phoneme lexicon was built by training a joint n-gram based G2P convertor on the WSJ lexicon. We used sequitur G2P toolkit [17] for this purpose. We compare systems based on four different lexicons,

1. *GRAPH* - same grapheme lexicon that was used in in-domain study presented in Section 6.
2. *WSJ-G2P* - phoneme lexicon obtained by G2P conversion
3. *Mixed-WSJ-G2P* - RM corpus and WSJ corpus share 568 common words. So, we created another phoneme lexicon where the pronunciation for common words is obtained from WSJ lexicon and the rest using G2P converter.
4. *BASE-RM* - same well developed phoneme lexicon that was used in in-domain study presented in Section 6. G2P converter can introduce pronunciation errors, so this lexicon serves as the optimistic case.

The systems investigated are System Hybrid, System KL-HMM SKL, System SP-HMM, System Tied and System HMM/GMM. For the sake of clarity, we present KL-HMM studies using only local score SKL. The reader may refer to [12] for results on the other local scores. Training of HMM/GMM system using lexicon *WSJ-G2P* resulted in 1250 clustered states and using lexicon *Mixed-WSJ-G2P lexicon* resulted in 1261 clustered states. For lexicon *BASE-RM*, we used System HMM/GMM from the in-domain study reported in Section 6. Adaptation of HMM/GMM system trained on WSJ1 was not considered as RM has sufficient speech data (3.8 hours) for full model re-estimation [40].

When compared to the number of parameters presented in Table 2, it is to be noted that by using an ANN trained on cross-domain corpus with more speech data, only the acoustic model complexity changes. The number of lexical model parameters remain the same. The ANN trained on WSJ corpus had 1.5M parameters. Complexity of System HMM/GMM does not change much.

Table 4 reports the performances on the test set of RM corpus for contexts *mono*, *tri* and *quint* using the above four lexicons and different systems.

ASR studies using System KL-HMM (for grapheme and phoneme lexicon) were first reported in [12].

Table 4: Word error rate expressed in % on the test set of RM corpus. Boldface indicates best system for each lexicon.

Lexicon	System	Subword context		
		mono	*tri*	*quint*
GRAPH	Hybrid	–	–	–
	KL-HMM SKL	32.4	6.0	**4.7**
	$SP-HMM$	23.9	6.8	5.8
	Tied	23.1	5.9	5.4
	HMM/GMM	36.0	7.1	–
WSJ-G2P	Hybrid	22.0	–	–
	KL-HMM SKL	15.6	**5.4**	5.7
	$SP-HMM$	11.7	6.3	6.0
	Tied	11.1	6.2	6.1
	HMM/GMM	17.3	7.1	–
Mixed-WSJ-G2P	Hybrid	12.5	–	–
	KL-HMM SKL	9.5	5.1	**5.0**
	$SP-HMM$	8.8	5.9	5.2
	Tied	8.9	5.8	5.7
	HMM/GMM	11.9	5.2	–
BASE-RM	Hybrid	8.6	–	–
	KL-HMM SKL	6.9	**4.7**	**4.7**
	$SP-HMM$	6.9	5.5	5.7
	Tied	6.8	5.2	5.5
	HMM/GMM	10.5	4.9	–

The main observations are as follows,

- Our original hypothesis was that the proposed grapheme-based system can yield performance as good as phoneme-based system which uses lexicon developed by G2P conversion. Comparing the best performances across lexicons, it can be observed that System KL-HMM SKL with lexicon *GRAPH* yields (a) better performance than phoneme-based systems using lexicon *WSJ-G2P* or lexicon *Mixed-WSJ-G2P*, (b) performance comparable to the optimistic case where the phoneme-based system uses well developed phoneme lexicon, and (c) better performance than System HMM/GMM across all lexicons.
- System Hybrid using WSJ-G2P or Mixed-WSJ-G2P lexicons yields

poor performance compared to system using BASE-RM lexicon. This is mainly due to the pronunciation errors of G2P converter. However, it can be observed that with probabilistic lexical modeling (comparison across *mono*) the degradation in performance is comparatively low.

The impact of pronunciation errors can also be observed on the performance of System HMM/GMM. Especially, in the case of lexicon *WSJ-G2P* it can be observed that the performance is significantly lower than the optimistic case, i.e., when lexicon *BASE-RM* is used. Furthermore, on the same lexicon i.e., *WSJ-G2P*, System KL-HMM SKL, System SP-HMM and System Tied with context-independent acoustic states and same context-dependent lexical states (context *tri*) outperform System HMM/GMM. This difference in performance is result of mainly two factors, first, System KL-HMM SKL, System SP-HMM and System Tied use an acoustic model that is trained on more data and on well developed cross domain lexicon. Second, as observed earlier probabilistic lexical modeling tends to compensate for pronunciation errors.

- among the systems using probabilistic lexical modeling, similar to in-domain study, System KL-HMM SKL yields the best system.

- Comparing with Table 3, it is interesting to note that performance of System KL-HMM SKL, System SP-HMM and System Tied with context modeling is slightly better despite the ANN being trained on cross-domain data. This indicates that these systems could benefit from larger acoustic training data.

8. Multi-Accent Non-Native ASR Study

In this section, we investigate the probabilistic lexical modeling approaches for multi-accent non-native speech recognition. Two main challenges faced while building ASR systems for non-native tasks are pronunciation variability and lack of proper lexical and acoustic resources. Spoken words in non-native speech are pronounced differently from native pronunciations. Furthermore, when multiple accents are involved, pronunciations may also differ across different accents. The lack of proper acoustic resources makes it difficult to train conventional ASR systems for individual accents. The probabilistic lexical modeling approach can provide two potential advantages this scenario: (1) with graphemes as subword it can eliminate the

need to build exclusive phoneme lexical resources for the task, (2) can exploit the existing acoustic-phonetic resources available in other domains and languages.

The work presented here is an extension of our previous work on HIWIRE multi-accent non-native speech recognition task [13] where using out-of-domain data from SpeechDat(II) corpus we showed that a) the proposed grapheme-based system can yield same performance as phoneme-based system (that has access to well developed lexicon), b) using out-of-domain acoustic-phonetic resources from multiple languages is more advantageous compared to using resources from a language, and c) the system can be rapidly developed using less amount of acoustic data.

In this section, we compare probabilistic lexical modeling using KL-HMM and tied posterior approaches for phoneme, grapheme, and G2P lexicon based ASR systems. Furthermore, we investigate two cases where (1) both acoustic state posterior estimator and lexical model parameters are trained on cross-domain data and (2) acoustic state posterior estimator is trained on cross-domain data and lexical model parameters trained on cross-domain data are adapted using in-domain data.

8.1. Setup

HIWIRE is a non-native English speech corpus that contains utterances spoken by natives of France (31 speakers), Greece (20 speakers), Italy (20 speakers) and Spain (10 speakers) [41]. The speech is sampled at 16kHz. The utterances contain spoken pilot orders made up of 133 words. The database provides a grammar with a perplexity of 14.9.

The HIWIRE task does not have training data. It only includes adaptation data (50 utterances per speaker) and test data (50 utterances per speaker). This experimental protocol was originally defined in [41] and was used in the previous study [13].

We use the two ANNs, more precisely MLPs, from the previous study [13], namely, monolingual MLP and multilingual MLP as monolingual (MLP-MONO) acoustic model and multilingual (MLP-MULTI) acoustic model, respectively. The monolingual MLP was trained on SpeechDat(II) British English to classify 45 phonemes. The multilingual MLP was trained by pooling acoustic and lexical resources from five different languages of SpeechDat(II) corpus, namely British English, Italian, Spanish, Swiss French and Swiss German to classify 117 phonemes. All the SpeechDat(II) lexicons use SAMPA symbols, therefore output of the multilingual MLP is formed by merging phonemes that share the same symbol across

languages to build a SAMPA multilingual phoneme set. Table 8.1 gives an overview of the acoustic models MLP-MONO and MLP-MULTI.

Table 5: Overview of the acoustic models. The the MLP output dimension (D) and the total amount of training data (in hours) and are given.

Acoustic Model	Phoneme set	D	Data
MLP-MONO	SAMPA English	45	12.4 hrs
MLP-MULTI	SAMPA multilingual	117	63.0 hrs

SpeechDat(II) is telephone speech corpus, hence, the HIWIRE speech was down sampled to 8kHZ before extracting PLP cepstral features, and then forward passed through MLP-MONO and MLP-MULTI. For more details the reader is referred to [13, 42].

We build ASR systems using KL-HMM with local score SKL (denoted as System KL-HMM) and tied posterior (System Tied) approaches modeling single preceding and single following subword context. Unseen context-dependent models were backed-off to a seen model, i.e., the context of unseen context-dependent subword unit encountered is decreased gradually until we encounter an observed subword model. In the worst case scenario, the back-off will lead to context-independent subword models. We compare the use of following four lexicons:

1. *GRAPH*: grapheme lexicon for both SpeechDat(II) English and HIWIRE contains 29 context-independent graphemes including silence and symbols (hyphen and single quotation mark).
2. *SD-EN-G2P*: phoneme lexicon where pronunciations of all the 133 words in the HIWIRE corpus are obtained using G2P convertor. The G2P system was trained on SpeechDat(II) British English lexicon using sequitur toolkit [17].
3. *Mixed-SD-EN-G2P*: phoneme lexicon where the pronunciations of 102 words of HIWIRE task that are in common with SpeechDat(II) are borrowed from SpeechDat(II) lexicon, and the pronunciations for remaining 31 words are obtained using G2P convertor.
4. *BASE-HIWIRE*: well developed phoneme lexicon based on SAMPA phone set. The original lexicon supplied with the HIWIRE corpus contains pronunciations based on ARPABET (US English), while SpeechDat(II) lexicon is based on SAMPA phone set. So, we created a HIWIRE lexicon based on SAMPA phone set by borrowing pronunciations of 102 words that are in common from the SpeechDat(II) lexicon. For the remaining 31 words, we obtained pronunciations by using the mapping between ARPABET and SAMPA phone sets.

We build systems using one of the above lexicons and acoustic models MLP-MONO or MLP-MULTI.

8.2. Experiments and Results

No adaptation: in this study, both acoustic model parameters θ_a and lexical model parameters θ_l are estimated using SpeechDat(II) English data.

Table 6 presents the performance for various systems using different lexicons and acoustic models (MLP-MONO and MLP-MULTI). Results show that KL-HMM approach consistently outperforms tied posterior approach. It can be observed that the system using *GRAPH* lexicon outperforms the systems using *SD-EN-G2P*, *Mixed-SD-EN-G2P* and *BASE-HIWIRE* lexicons. For all the lexicons, acoustic model MLP-MULTI performs better than acoustic model MLP-MONO. The system using *GRAPH* lexicon yields the best performance of 8.1% WER.

Table 6: Word error rates in percentage on the test set of HIWIRE corpus for KL-HMM and tied posterior systems using various lexicons. Models trained on SpeechDat(II) English data are used to decode HIWIRE utterances without any adaptation. Boldface indicates the best system for each acoustic model

Lexicon	MLP-MONO		MLP-MULTI	
	Tied	KL-HMM	Tied	KL-HMM
GRAPH	12.9	**12.1**	10.7	**8.1**
SD-EN-G2P	16.6	14.7	13.6	10.7
Mixed-SD-EN-G2P	16.1	14.3	12.9	10.5
BASE-HIWIRE	13.1	12.7	10.8	9.5

Lexical model adaptation: In this study, only the lexical model parameters θ_l are re-estimated on the adaptation data of HIWIRE task using the parameters estimated on SpeechDat(II) as initial parameters.

Table 7 presents the performance for various systems using different lexicons, acoustic models and lexical models. It can be observed that by just adapting the lexical model there is large improvement in performance for all the systems. The system using *GRAPH* lexicon achieves the best performance of 1.8% WER which is significantly[4] better than systems using lexicons *SD-EN-G2P*, *Mixed-SD-EN-G2P* and *BASE-HIWIRE*. Again, we observe that KL-HMM system outperforms tied-posterior system.

[4]Statistically significant with 99% confidence [43].

Table 7: Word error rates in percentage on the test set of HIWIRE corpus for systems KL-HMM and Tied using various lexicons. Models trained on SpeechDat(II) English data and adapted on HIWIRE adaptation data are used to decode HIWIRE utterances. Boldface indicates the best system for each acoustic model

Lexicon	MLP-MONO		MLP-MULTI	
	Tied	KL-HMM	Tied	KL-HMM
GRAPH	4.3	**2.6**	4.1	**1.8**
SD-EN-G2P	7.2	3.2	5.5	2.2
Mixed-SD-EN-G2P	7.3	3.2	5.3	2.2
BASE-HIWIRE	6.2	3.5	4.4	2.6

It can be noted that the performance of KL-HMM system with *BASE-HIWIRE* lexicon (2.6% WER) is different than the performance of KL-HMM system with baseline phoneme lexicon (1.9% WER) reported in [13]. The performance difference is because of the different phoneme lexicons used in the two studies. *BASE-HIWIRE* is based on SAMPA phone set and is obtained from British English lexicon of SpeechDat(II) corpus, while, the lexicon in [13] is based on ARPABET and is obtained from U.S English lexicon CMUDict. This difference in ASR performance precisely highlights the extra care that needs to be taken when developing phoneme lexicon. This is not an issue with grapheme lexicon in the present approach.

On the same HIWIRE task, in the literature it has been found that TIMIT trained HMM/GMM system without adaptation achieves a performance of 8.6% WER and with MLLR adaptation achieves the best performance of 2.75% WER [41]. In [44], linear hidden network (LHN) based speaker adaptation in hybrid HMM/ANN framework was investigated. ANN trained on data from TIMIT, WSJ0, WSJ1 and Vehiclus-ch0 was adapted to each speaker using LHN. The system achieved a WER of 1.8% on HIWIRE task. The systems presented in this section were trained on grapheme lexicon, multilingual data and 8kHz sampling where as in [44] LHN is trained on phoneme lexicon, English data, but more in quantity, and 16kHz sampling. Furthermore, we did not perform any speaker adaptation or acoustic model adaptation.

Experimental studies presented in this section show that the proposed grapheme-based ASR approach can handle pronunciation variability of non-native speech by exploiting existing acoustic-phonetic resources available in other languages and domains, and can yield performance comparable or

better than phoneme-based ASR system.

9. Analysis of Lexical Model

The experimental studies presented in the previous sections validated the proposed grapheme-based ASR approach. In this section, we elucidate that indeed in the proposed approach grapheme-to-phoneme relationship is captured by the lexical model. We show this while explaining the effect of the cost function and the grapheme context on the grapheme-to-phoneme relationship being learned by the lexical model. For the sake of clarity, we restrict the analysis to KL-HMM systems and in-domain models. Also, we will see that the update step in KL-HMM system is easy to interpret in terms of combination of probabilities.

9.1. Effect of Different Local Scores

In grapheme-based KL-HMM system where the lexical states represent graphemes and the acoustic states represent phonemes, the lexical model parameters $\mathbf{y}_i$ model the probabilistic relationship between graphemes and phonemes. In English, the relationship between graphemes and phonemes can be one-to-one or one-to-many. The accuracy of the captured probabilistic relationship is governed by the local score used during KL-HMM parameter estimation. It can be observed that,

- local score KL yields an update step (see Equation (20)) that estimates normalized geometric mean of the acoustic state probability vectors that belong to the state. In classifier fusion literature [45, 46, 47], it is shown that such combination of probabilities (referred to as product combination or logarithmic opinion pool) leads to a less dispersive distribution, i.e., it captures the dominant decision.

  In the case of graphemes that have more one-to-one relationship with phonemes, ANN decisions across frames belonging to a state can be more homogeneous, i.e., decisions belong to the same phoneme class. While, in the case of graphemes that have one-to-many relationship with phonemes, ANN decisions can be more heterogeneous, i.e., the decisions of ANN differ. Given this, it can be hypothesized that the system using local score KL has the capability to capture one-to-one G2P relationship better than one-to-many G2P relationship.

- local score RKL yields an update step (see Equation (21)) that estimates arithmetic mean of the acoustic state probability vectors that

belong to the state. In classifier fusion literature [45, 46, 47], it is shown that such combination of probabilities (referred to as sum combination or linear opinion pool) leads to a dispersive distribution compared to product combination. This is particularly beneficial when there are more heterogeneous decisions to combine, like in the case of one-to-many G2P relationship.

- local score SKL employs both KL and RKL update steps. Thus, it can be hypothesized that the system based on SKL will retain the strengths of the two local scores KL and RKL.

In order to analyze and visualize the effect of local scores better,

1. we trained grapheme-based KL-HMM systems using three local scores (KL, RKL and SKL), where lexical states are context-independent graphemes, acoustic states are context-independent phonemes, and each lexical state is modeled by a single state HMM, and
2. the categorical distribution of each grapheme model is sorted according to the probability value, and the dimensions with probability value greater than or equal to 0.1 are picked.

Table 8 presents the G2P relationship captured by different grapheme-based KL-HMM systems. From the table the effect of local scores on the captured G2P relationship can be observed as the following:

1. the system using local score KL captures one-to-one G2P relationship (e.g., see [B], [L], [M], [P]) better than one-to-many G2P relationship (e.g., see vowel graphemes, [C], [H], [X]).
2. local score RKL in addition to appropriate one-to-many G2P relationship (e.g., see vowel graphemes, [C], [G], [H]) also captures additional confusable and spurious relations. This can be particularly seen in the case of one-to-one G2P correspondence (e.g., see [B], [M]).
3. local score SKL tends to capture one-to-one G2P relationship similar to local score KL. It is able to capture one-to-many G2P relationship better than local score KL but not to the same extent as local score RKL (e.g., see [G], [H], [N]).

The analysis shows that local scores KL and RKL can model better one-to-one and one-to-many lexical-to-acoustic state relationships, respectively where as, the local score SKL can model both one-to-one and one-to-many lexical-to-acoustic state relationships.

Table 8: Grapheme-to-phoneme relationship (sorted according to the maximum probability value and with a probability value greater than or equal to 0.1) learned by KL-HMM states

Grap -pheme	Phonemes		
	KL	RKL	SKL
A	ae(0.7) eh(0.2) ey(0.1)	ae(0.3) ey(0.3) eh(0.1) ax(0.1)	ae(0.5) eh(0.2) ey(0.1) ax(0.1)
B	b(1.0)	b(0.5) ah(0.2)	b(0.9)
C	k(1.0)	k(0.5) ch(0.2) s(0.1) t(0.1)	k(0.6) t(0.2) ch(0.1) s(0.1)
D	d(0.9) t(0.1)	d(0.5) t(0.2) sil(0.1)	d(0.7) t(0.1)
E	ax(0.4) ih(0.3) eh(0.1) iy(0.1)	iy(0.3) eh(0.1) ax(0.1) ih(0.1) ey(0.1)	iy(0.3) ax(0.2) ih(0.2) eh(0.1) ey(0.1)
F	f(1.0)	f(0.7) v(0.1) sil(0.1)	f(0.9)
G	g(0.9)	g(0.4) jh(0.2) sil(0.1) k(0.1) d(0.1)	g(0.7) d(0.1) k(0.1)
H	t(0.7) d(0.1) sil(0.1)	sh(0.3) dh(0.2) hh(0.1) th(0.1)	dh(0.2) sil(0.2) t(0.2) th(0.1) d(0.1) hh(0.1)
I	ih(0.8) ax(0.1)	ih(0.4) ay(0.2) ax(0.1) iy(0.1)	ih(0.5) ax(0.2) eh(0.1) ay(0.1)
J	jh(1.0)	jh(0.7) ch(0.1) d(0.1) t(0.1)	jh(0.9)
K	k(1.0)	k(0.7) sil(0.1) t(0.1)	k(0.9)
L	l(1.0)	l(0.5) el(0.1) ao(0.1) ow(0.1)	l(0.8)
M	m(1.0)	m(0.7) n(0.1)	m(0.9) n(0.1)
N	n(0.9)	n(0.5) en(0.1) ng(0.1)	n(0.8) en(0.1)
O	ao(0.4) aa(0.3) ow(0.1) ah(0.1)	ao(0.2) aa(0.2) ow(0.2) sh(0.1) ah(0.1) ax(0.1)	ao(0.2) aa(0.2) ow(0.2) ah(0.1) ax(0.1)
P	p(1.0)	p(0.8)	p(0.9)
Q	k(1.0)	k(0.5) w(0.2) uw(0.1) y(0.1)	k(0.9)
R	r(0.8) axr(0.2)	r(0.4) axr(0.3) aa(0.1) er(0.1)	r(0.6) axr(0.3) er(0.1)
S	s(0.9) z(0.1)	s(0.6) z(0.2)	s(0.8) z(0.2)
T	t(0.9)	t(0.5) sil(0.1) d(0.1) k(0.1)	t(0.8)
U	ax(0.4) uw(0.3) ih(0.1)	uw(0.3) y(0.2) ax(0.1) ah(0.1)	uw(0.3) ax(0.3) ih(0.1) ah(0.1)
V	ay(0.9)	v(0.5) ay(0.3)	v(0.9)
W	w(1.0)	w(0.6) aw(0.1) uw(0.1)	w(0.9)
X	k(0.9) t(0.1)	s(0.4) k(0.4)	k(0.5) s(0.3) t(0.1)
Y	iy(0.8) ey(0.1)	iy(0.4) ay(0.1) ey(0.1) oy(0.1)	iy(0.5) ey(0.3) ih(0.1)
Z	z(0.9)	ay(0.4) z(0.3) s(0.1)	z(0.8) s(0.1)
sil	sil(1.0)	sil(1.0)	sil(1.0)

9.2. Effect of Increasing Grapheme Subword Unit Context

In [12], we analyzed the models of consonant grapheme [C] and vowel grapheme [A] (which have varying degree of irregular mapping to phoneme) to gain insight into the effect of contextual modeling. It was observed that vowel grapheme [A] needed longer context to dominantly capture the relation to appropriate phoneme compared to consonant grapheme [C]. Also, as the context of grapheme subword units is increased, KL-HMM parameters start to capture phoneme contextual information. In order to get a global picture on the effect of context, we trained single state grapheme models for three contexts (*mono*, *tri* and *quint*) using local score SKL, and then computed the entropy of the grapheme models. In the case of *tri* and *quint*, average entropy of all the grapheme models with same center grapheme was computed. Figure 2 shows the entropy of the grapheme models with

increasing context. It can be observed that,

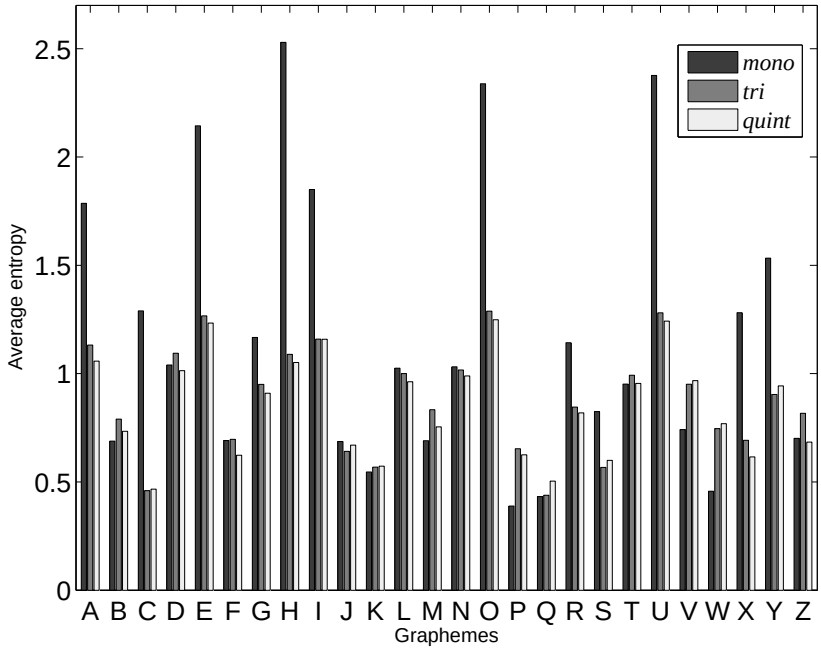

Figure 2: Entropy of grapheme models with increasing context. For contexts *tri* and *quint*, average entropy of all the grapheme models with same center grapheme is displayed.

- Vowel graphemes ([A], [E], [I], [O], [U]) and some consonant graphemes ([C], [H], [R], [X]) have high entropy for context *mono* signifying the fact that the models capture one-to-many G2P relationship. As the context increases, entropy decreases i.e., models tend to capture one-to-one G2P relationship.

- Few consonant graphemes like [B], [K], [P], [V] have low entropy for context *mono* which suggests that context-independent grapheme itself models one-to-one G2P relationship. However, the entropy slightly increases as the context increases. A closer look at the models revealed that this was due to the context information captured by the models.

In other words, similar to G2P conversion systems [15, 16, 17, 18, 19], one-to-many G2P relationship tends to become more regular or close to

one-to-one as longer grapheme context is modeled.

10. Discussion and Conclusion

The present paper presented a reformulation of HMM-based ASR in terms of acoustic states and lexical states. Under this reformulation, we elucidated that HMM-based ASR is a template matching approach, where a sequence corresponding to lexical evidence is matched with a sequence corresponding to acoustic evidence. Furthermore, in addition to explaining deterministic lexical modeling aspect implicit in standard ASR system, the reformulation unifies approaches developed with different perspectives into a single probabilistic lexical modeling framework. For instance, PCHMM is a soft state tying approach; tied posterior approach was developed more from the perspective of semi-continuous HMM [48]; and KL-HMM was mainly developed from the perspective of posterior features, as an alternate technique to Tandem features [49], where the ANN output is not post processed. In the context of posterior-based probabilistic lexical modeling approach, we also introduced SP-HMM. Finally, the paper also explained that both likelihood-based ASR approach and posterior-based ASR approach are independent of the approach taken to model the acoustics, i.e., to use GMMs or ANNs.

There are several potential advantages of modeling the probabilistic relationship between lexical states and acoustic states, namely,

1. flexibility to model graphemes as subword units and learning the G2P relationship using both acoustic and lexical data. We found this approach to be particularly beneficial in comparison to the approach, where ASR system is built with a phoneme lexicon that was developed by learning G2P relationship independent of the acoustic data. Furthermore, the approach was also found to be advantageous for non-native ASR, where developing a phoneme lexicon containing pronunciations that matches well with non-native pronunciations is a challenging task.
2. it could handle the short comings in the lexicon development. The short comings can be due to reasons such as, pronunciation errors (as in the case of G2P conversion), pronunciation mismatch (as in the case of non-native speech). In that respect, the cross-domain ASR study and the multi-accent non-native ASR study demonstrated the effectiveness of probabilistic lexical modeling.
3. it naturally leads to parameter sharing, irrespective of whether the acoustic model originally shares the parameters across all the acoustic states or not. For instance, in ANNs all the parameters (weights

and biases) are shared across all the acoustic states, while in the case GMMs the parameters are not always shared across all the acoustic states. The advantage of such natural parameter sharing is that the complexity could be effectively distributed across acoustic model and lexical model. To some extent, we observed that in our studies. More precisely, by using ANNs that classify context-independent phonemes (or acoustic states), we observed that grapheme-based and phoneme-based systems with probabilistic lexical modeling capabilities were able to achieve performance comparable to standard context-dependent subword unit HMM/GMM system by modeling the lexical state context.

4. as shown in this paper as well as in other studies using KL-HMM, such as [50], probabilistic lexical modeling framework can be effectively used to exploit acoustic and lexical resources of multiple languages and domains. This together with the advantage of using grapheme lexicon is interesting for development of ASR systems for under-resourced languages or domains [14].

In our studies, we found KL-HMM consistently yields a better system compared to tied posterior and SP-HMM. In Section 4.1, through the link between KL-divergence and hypothesis testing, we observed that KL-HMM leads to a system which is locally discriminative and globally generative in the acoustic state probability space. The generative aspect can be exploited for acoustic data-driven grapheme-to-phoneme conversion [51, 52].

In this work, we used ANNs that classified context-independent phonemes. However, it is possible to model context-dependent phonemes as acoustic states and yield similar gains in performance using probabilistic lexical modeling. This has been observed for both phoneme-based ASR [23, 36, 53] and grapheme-based ASR systems [54].

The grapheme-based approach presented in this paper needs a phoneme-based acoustic model. In other words, the approach depends upon the availability of phoneme lexicon. In a recent study, we have found that "phoneme-like" acoustic states could be derived directly from the acoustic data and used to build grapheme-based ASR systems that yield competing performance [54].

One salient aspect that stands out in the present work as well as in other related works on KL-HMM e.g., [50, 14] is that the acoustic model can be trained on data from different domains and languages. The lexical model parameters can be trained on domain specific or language specific data to yield a competitive system. Our future work will continue to explore this

aspect to build flexible grapheme-based ASR systems incorporating latest developments in acoustic modeling, such as use of context-dependent neural networks [26].

Acknowledgment

This work was supported by the Swiss NSF through the grants "Flexible Grapheme-Based Automatic Speech Recognition (FlexASR) " and the National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management" (www.im2.ch). The authors would like to thank their current and past colleagues at Idiap, especially Guillermo Aradilla, David Imseng, John Dines, Hervé Bourlard, and reviewers of the Interspeech 2011 paper [12] for their critical inputs and suggestions. The authors would also like to thank the anonymous reviewers for their insightful comments.

References

[1] D. O'Shaughnessy, Speech Communication Human and Machine, Addison-Wesley, 1987.

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

[3] M. Killer, S. Stüker, T. Schultz, Grapheme based Speech Recognition, in: Proc. of European Conference on Speech Communication and Technology (EUROSPEECH), 2003.

[4] B. Mimer, S. Stüker, T. Schultz, Flexible Decision Trees for Grapheme based Speech Recognition, in: Elektronische Sprachsignalverarbeitung, Cottbus, Germany, 2004.

[5] T. Schultz, K. Kirchhoff, Multilingual Acoustic Modeling, in: Multilingual Speech Processing, Academic Press, 2006.

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

[7] Y.-H. Sung, T. Hughes, F. Beaufays, B. Strope, Revisiting Graphemes with Increasing Amounts of Data, in: Proc. of ICASSP, 4449–4452, 2009.

[8] S. Stüker, T. Schultz, A Grapheme Based Speech Recognition System for Russian, in: Proc. of Speech and Computer (SPECOM), 2004.

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

[10] F. Biadsy, P. Moreno, M. Jansche, Google's Cross-Dialect Arabic Voice Search, in: Proc. of ICASSP, 4441–4444, 2012.

[11] T. Schlippe, E. G. K. Djomgang, N. T. Vu, S. Ochs, T. Schultz, Hausa Large Vocabulary Continuous Speech Recognition, in: Proc.of the Spoken Languages Technologies for Under-resourced Languages(SLTU), 2012.

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

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

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

[15] P. Taylor, A. Black, R. Caley, The Architecture of the Festival Speech Synthesis System, in: Proc. of ESCA Workshop on Speech Synthesis, 1998.

[16] S. F. Chen, Conditional and Joint Models for Grapheme-to-Phoneme Conversion, in: Proc. of EUROSPEECH, 933–936, 2003.

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

[18] S. Jiampojamarn, G. Kondrak, Online Discriminative Training for Grapheme-to-Phoneme Conversion, in: Proc. of Interspeech, 1303–1306, 2009.

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

[20] T. Schlippe, S. Ochs, T. Schultz, Grapheme-to-Phoneme Model Generation for Indo-European Languages , in: Proc. of ICASSP, 4801–4804, 2012.

[21] G. Aradilla, J. Vepa, H. Bourlard, An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features, in: Proc. of ICASSP, IV–657 – IV–660, 2007.

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

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

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

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

[26] G. E. Dahl, D. Yu, L. Deng, A. Acero, Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition , IEEE Trans. on Audio, Speech, and Language Processing 20 (2012) 30–42.

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

[28] G. Aradilla, H. Bourlard, Posterior-Based Features and Distances in Template Matching for Speech Recognition, in: Multimodal Interaction and Related Machine Learning Algorithms (MLMI), 204–214, 2007.

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

[30] S. Soldo, M. Magimai.-Doss, J. P. Pinto, H. Bourlard, On MLP-based Posterior Features for Template-based ASR, http://publications.idiap.ch/downloads/reports/2009/Soldo_Idiap-RR-37-2009.pdf, Idiap Research Report, 2009.

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

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

[33] S. Eguchi, J. Copas, Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma, Journal of Multivariate Analysis 97 (9).

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

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

[36] M. Saraclar, H. Nock, S. Khudanpur, Pronunciation Modeling By Sharing Gaussian Densities Across Phonetic Models, in: Computer Speech and Language, 137–160, 2000.

[37] T. Hain, P. Woodland, Dynamic HMM Selection for Continuous Speech Recognition, in: Proc. of EUROSPEECH, 1327–1330, 1999.

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

[39] S. Fitt, Documentation and User Guide to UNISYN Lexicon and Postlexical Rules, Tech. Rep., Center for Speech Technology Research, University of Edinburgh, 2000.

[40] J. Kohler, Language Adaptation of Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks, in: Proc. of ICASSP, vol. 1, 417–420 vol.1, 1998.

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

[42] D. Imseng, H. Bourlard, M. Magimai.-Doss, Towards Mixed Language Speech Recognition Systems, in: Proc. of Interspeech, 278–281, 2010.

[43] M. Bisani, H. Ney, Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation, in: Proc. of ICASSP, vol. 1, 409–412, 2004.

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

[45] C. Genest, J. V. Zidek, Combining Probability Distributions: A Critique and an Annotated Bibliography, Statistical Science 1 (1986) 114–148.

[46] J. Kittler, M. Hatef, R. P. Duin, J. Matas, On Combining Classifiers, IEEE Trans. on Pattern Analysis and Machine Intelligence 20 (3) (1998) 226–239.

[47] A. E. Abbas, A Kullback-Leibler View of Linear and Log-Linear Pools, Decision Analysis 6 (2009) 25–37.

[48] X. D. Huang, M. A. Jack, Semi-continuous hidden Markov models for speech signal, Computer Speech and Language 3 (3) (1989) 239–251.

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

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

[51] R. Rasipuram, M. Magimai.-Doss, Acoustic Data-driven Grapheme-to-Phoneme Conversion using KL-HMM, in: Proc. of ICASSP, 4841–4844, 2012.

[52] R. Rasipuram, M. Magimai.-Doss, Combining Acoustic Data Driven G2P and Letter-to-Sound Rules for Under Resource Lexicon Generation, in: Proc. of Interspeech, 2012.

[53] R. Rasipuram, M. Magimai.-Doss, KL-HMM and Probabilistic Lexical Modeling, `http://publications.idiap.ch/downloads/reports/2013/Rasipuram_Idiap-RR-04-2013.pdf`, Idiap Research Report, 2013.

[54] R. Rasipuram, M. Magimai.-Doss, Improving Grapheme-based ASR by Probabilistic Lexical Modeling Approach, `https://publidiap.idiap.ch/downloads//reports/2013/Rasipuram_Idiap-RR-14-2013.pdf`, Idiap Research Report, 2013.