PROBABILISTIC LEXICAL MODELING AND UNSUPERVISED TRAINING FOR ZERO-RESOURCED ASR

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

[1] Idiap Research Institute, CH-1920 Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

ABSTRACT

Standard automatic speech recognition (ASR) systems rely on transcribed speech, language models, and pronunciation dictionaries to achieve state-of-the-art performance. The unavailability of these resources constrains the ASR technology to be available for many languages. In this paper, we propose a novel zero-resourced ASR approach to train acoustic models that only uses list of probable words from the language of interest. The proposed approach is based on Kullback-Leibler divergence based hidden Markov model (KL-HMM), grapheme subword units, knowledge of grapheme-to-phoneme mapping, and graphemic constraints derived from the word list. The approach also exploits existing acoustic and lexical resources available in other resource rich languages. Furthermore, we propose unsupervised adaptation of KL-HMM acoustic model parameters if untranscribed speech data in the target language is available. We demonstrate the potential of the proposed approach through a simulated study on Greek language.

Index Terms— zero-resourced speech recognition, probabilistic lexical modeling, Kullback-Leibler divergence based hidden Markov model, graphemes, phonemes, unsupervised adaptation.

1. INTRODUCTION

Building a state-of-the-art ASR system for a new language, new domain, or a new accent requires large amounts of recorded and transcribed speech data, prior phoneme lexical resources and large amounts of text data. Obtaining these resources can be costly in terms of time and money. The unavailability of such resources in many languages has motivated approaches that exploit resources available in resource rich languages and domains to build better acoustic models for ASR systems in new languages [1, 2, 3, 4, 5, 6, 7]. The research in this domain has mainly focussed on multi-lingual and cross-lingual training in the framework of HMM/Gaussian mixture model (GMM) systems [1, 3], hybrid HMM/artificial neural networks (ANN) systems [5, 6], subspace GMM systems [2, 4] etc.

In the framework of hybrid HMM/ANN systems, Kullback-Leibler divergence based hidden Markov model (KL-HMM) is a recently proposed approach that is shown to be especially useful to build ASR systems for new and under-resourced languages [8, 9, 5, 6]. In KL-HMM approach, phoneme class conditional probabilities estimated by an ANN are directly used as feature observations [10]. In a more recent work, we showed that KL-HMM system is a HMM-based ASR system in which the relationship between physical (acoustic) states modeled by ANN and logical (lexical) units modeled by KL-HMM is probabilistic [11, 12]. More specifically, KL-HMM approach can be seen as probabilistic lexical modeling approach.

In previous KL-HMM studies, it has been shown that: (1) the ANN could be trained on resource rich languages [8, 9, 5, 6], (2) the phoneme lexical resources can be replaced with grapheme lexicon and grapheme subword units derived from the orthography of words [8, 6, 11] and (3) the KL-HMM parameters that model the probabilistic relationship between acoustic states and lexical units could be trained on small amount of speech data from the target language and its word level transcriptions [9, 5, 6].

In this paper, we propose a novel approach for zero-resourced ASR in the framework of KL-HMM that replaces KL-HMM parameters trained on transcribed speech data with a simple knowledge based KL-HMM parameter set. This approach can facilitate building ASR systems for languages without transcribed speech, and phoneme pronunciation dictionary. The only knowledge that is assumed to be available is the list of possible words in the language of interest. More specifically, in the proposed approach, (a) the ANN is trained on out-of-domain resources, (b) the phoneme lexical resources are replaced with grapheme lexicon and grapheme subword unit set derived from the given list of words in the target language, and (c) the relationship between acoustic states modeled by ANN and lexical units (graphemes) of the target language is defined using minimal phonetic knowledge of the target language. Furthermore, unsupervised adaptation of the initial knowledge based KL-HMM system is proposed if untranscribed speech data in the target language is available. The unsupervised adaptation is similar in spirit to [13], however, the proposed approach exploits resources available in other languages and domains.

We evaluate the proposed approach on SpeechDat(II) database considering Greek as the target zero-resourced language. Data from five other European languages of SpeechDat(II) corpus is used as out-of-domain resources. Our studies revealed that the proposed approach facilitates building ASR systems in zero-resourced setup with minimal knowledge or supervision. The KL-HMM system using just the word list from the target Greek language results in word error rate (WER) of 43.0% (67% relative increase in WER compared to optimized Greek system trained on 13.5 hours of transcribed speech that results in WER of 14.2%). Furthermore, by unsupervised adaptation of KL-HMM parameters and trigraph modeling WER of 27.7% is achieved (36.6% relative reduction in WER compared to the system using only the word list and 51% relative increase in WER compared to the optimized system).

The rest of the paper is organized as follows: In Section 2 we introduce the KL-HMM approach and the interpretation of KL-HMM

This work was supported by the Swiss NSF through the grants Flexible Grapheme-Based Automatic Speech Recognition (FlexASR), by Hasler foundation through the grants Flexible acoustic data driven grapheme to acoustic unit conversion (AddG2SU) 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 David Imseng and Alexandros Lazaridis for their help with the experimental setup.

as probabilistic lexical modeling approach. In Section 3, we explain the proposed approach in detail. Section 4 presents the experimental setup and Section 5 presents the experimental results on the Greek language. Finally, in Section 6 we conclude.

2. KL-HMM AND PROBABILISTIC LEXICAL MODELING

Kullback-Leibler Divergence based HMM (KL-HMM) is a posterior based ASR approach, where posteriori probabilities of phonemes (phoneme posterior features) estimated using an ANN are directly used as feature observations [10]. Let $\mathbf{z}_t$ denote the phoneme posterior feature vector estimated at time frame t,

$$\begin{aligned}\mathbf{z}_t &= [z_t^1, \cdots, z_t^d, \cdots, z_t^D]^T \\ &= [P(p_1|\mathbf{x}_t), \cdots, P(p_d|\mathbf{x}_t), \cdots, P(p_D|\mathbf{x}_t)]^T\end{aligned}$$

where $\mathbf{x}_t$ is the acoustic feature (e.g., cepstral feature) at time frame t, $\{p_1, \cdots p_d, \cdots p_D\}$ is the phoneme set, D is the number of phonemes, and $P(p_d|\mathbf{x}_t)$ denotes the a posteriori probability of phoneme p_d given $\mathbf{x}_t$.

Each HMM state i in the KL-HMM system is parameterized by a categorical distribution $\mathbf{y}_i = [y_i^1, \cdots, y_i^d, \cdots, y_i^D]^{\mathrm{T}}$. The local score $S(\mathbf{y}_i, \mathbf{z}_t)$ at each HMM state i is

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

The parameters $\{\mathbf{y}_i\}_{i=1}^I$ are trained by optimizing a cost function based on KL-divergence. Figure 1 illustrates the KL-HMM approach.

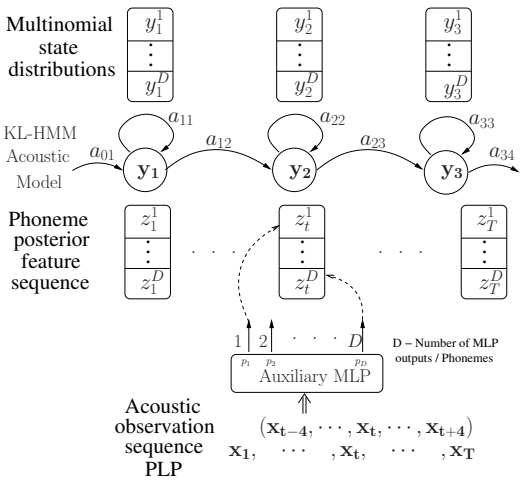

Fig. 1. Illustration of KL-HMM acoustic model

2.1. Training

KL-HMM is fully parameterized by $\Theta_{kull} = \{\{\mathbf{y}_i\}_{i=1}^I, \{a_{ij}\}_{i,j=1}^I\}$ 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 $\{Z(n), W(n)\}_{n=1}^N$, where for each training utterance n, $Z(n)$ represents 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 $T(n)$ and $W(n)$ represents the sequence of underlying words, the parameters Θ_{kull} are estimated by Viterbi expectation maximization algorithm which minimizes the cost function,

$$\min_{Q \in \mathcal{Q}} \sum_{n=1}^{N} \sum_{t=1}^{T(n)} [S(\mathbf{y}_{q_t}, \mathbf{z}_t(n)) - \log a_{q_{t-1}q_t}] \tag{2}$$

where $q_t \in \{1, \cdots, I\}$, $\mathcal{Q}$ denotes set of all possible HMM state sequences, $Q = \{q_1(n), \cdots, q_t(n), \cdots, q_T(n)\}$ 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. The optimal state distribution is the arithmetic mean of the training acoustic state probability vectors assigned to the state, i.e.,

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

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

2.2. Decoding

The decoding is performed using standard Viterbi decoder and the log-likelihood based score in the standard Viterbi decoding is replaced with KL-divergence based local score $-S(\mathbf{y}_i, \mathbf{z}_t)$. More precisely, 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 $\hat{m}$:

$$\hat{m} = \arg\min_{Q \in \mathcal{Q}} \sum_{t=1}^{T} [S(\mathbf{y}_{q_t}, \mathbf{z}_t) - \log a_{q_{t-1}q_t}] \tag{4}$$

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

2.3. Interpretation of KL-HMM as Probabilistic Lexical Modeling Approach

KL-HMM until now has been investigated as an approach where posterior probabilities of phonemes can be directly used as feature observations in HMM system [10, 8, 9, 5, 6]. Recently, we showed that the KL-HMM can be seen as probabilistic lexical modeling approach that is applicable to both HMM/GMM and hybrid HMM/ANN based ASR systems [11, 12]. More precisely, KL-HMM approach can be viewed as posterior based ASR approach which replaces the deterministic mapping between lexical units and acoustic states in standard HMM-based ASR system with probabilistic map [11, 12]. This is achieved in two steps:

1. first, an acoustic state posterior probability estimator is trained which estimates $\mathbf{z}_t = [P(q_t^{aco} = 1|\mathbf{x}_t) \cdots P(q_t^{aco} = D|\mathbf{x}_t)]^{\mathrm{T}}$. In this work, it is an ANN and the acoustic states are outputs of ANN.

2. second, a KL-HMM system is trained using $\mathbf{z}_t$ as feature observations. The states of second HMM represent lexical units (i.e., context-dependent subword units) that are parameterized by $\{\mathbf{y}_i\}_{i=1}^{I}$ and model the probabilistic relation between lexical units and acoustic states, i e., $\mathbf{y}_i = [P(q_t^{aco} = 1|q_t^{lex} = i) \cdots P(q_t^{aco} = D|q_t^{lex} = i)]^{\mathrm{T}}$

If $\mathbf{z}_t$ is estimated using ANN then KL-HMM can be seen as probabilistic lexical modeling applied to hybrid HMM/ANN system [12], whereas if $\mathbf{z}_t$ is estimated using GMMs of clustered HMM states then it can be seen as probabilistic lexical modeling applied to HMM/GMM system [11].

2.4. Previous Studies on KL-HMM

KL-HMM approach has been investigated in variety of real world ASR scenarios like:

- Grapheme subword unit based ASR systems [8, 6]: The goal was to build ASR systems for languages where the phoneme pronunciation lexicon is not available. Therefore, graphemes are used as subword units and the pronunciation lexicon is simply derived using the orthography of words. ANN is trained on languages where acoustic and pronunciation lexical resources are available and the KL-HMM parameter set Θ_{kull} is trained on data from the task of interest. The KL-HMM parameters in this case learn the probablistic grapheme-to-phoneme relationship.
- Non-native speech recognition [9]: Here, the goal was to build ASR system for non-native speech with limited acoustic resources. In this case, the phoneme pronunciation lexicon and the ANN are based on native language speakers. The KL-HMM parameter set Θ_{kull} is trained on small amount of non-native speech. It was shown that the KL-HMM parameters can account for phonetic variation inherent in non-native speech. Furthermore, KL-HMM system using grapheme lexicon could achieve perform similar to or better than KL-HMM system using phoneme lexicon. It was also observed that phoneme-based KL-HMM system was sensitive to the lexicon used (US English and British English).
- Rapid development of ASR systems for new and under-resourced languages [9, 5, 6]: The goal here was to build ASR systems for languages with limited training data. In this case, acoustic and lexical resources available in other resource rich languages are exploited to train ANN. The KL-HMM parameter set Θ_{kull} is trained on the data from task of interest. It has been shown that KL-HMM approach is especially useful when there is very limited amount of in-domain training data.

Nevertheless, all the previous studies on KL-HMM approach relied on speech data and its corresponding word level transcriptions to learn the KL-HMM parameter set Θ_{kull}.

3. KL-HMM APPROACH FOR ZERO-RESOURCED ASR

In this paper, we consider a zero-resourced ASR scenario in which it is assumed that we have no transcribed training data, no language models, and no pronunciation lexicon. However, we assume that we have knowledge of the possible words in the language and therefore its character or grapheme set is also known. We propose a novel approach to build a zero-resourced ASR system in the KL-HMM framework, in which various resources required to build KL-HMM based system are derived in the following way:

- Acoustic posterior probability estimator: in this work, it is an ANN trained on out-of-domain data from multiple languages. The acoustic units or the outputs of ANN represent multilingual phonemes. The test utterance of the target language is forward passed through this ANN to obtain posterior feature sequence $\mathbf{z}_t$.
- Pronunciation lexical resources: grapheme lexicon and grapheme subword unit set are obtained from the list of possible words in the target language. Therefore, the lexical units are context-independent graphemes of the target language.
- Grapheme language model: we train a bigram grapheme language model based on the word list in the target language.
- KL-HMM parameter set: an initial knowledge-based KL-HMM parameter set $\Theta_l^{kn} = \{\{\mathbf{y}_i\}_{i=1}^{I}\}$ is defined in the following way: (1) first associate each grapheme lexical unit to one or more phoneme outputs of ANN, (2) if a lexical unit i is mapped to an acoustic unit d then $P(q_t^{aco} = d|q_t^{lex} = i) = s$ whereas if lexical unit j is not mapped to acoustic unit d then $P(q_t^{aco} = d|q_t^{lex} = j) = \frac{1-s}{I-1}$, I being the total number of acoustic units and s is chosen such that $s \geq 0.5$. In case of one-to-many map between lexical unit and acoustic states, the value of s is divided accordingly, (3) each context-independent grapheme is modeled as a 3 state HMM. Figure 2 illustrates the knowledge-based KL-HMM parameter training.

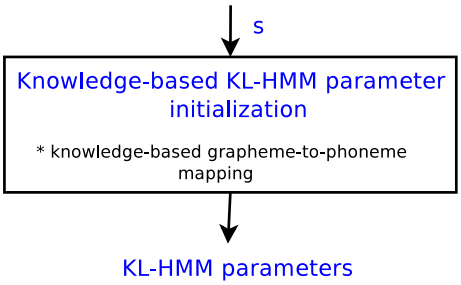

Fig. 2. Illustration of the knowledge-based KL-HMM parameter training (Case 2 in Section 4)

With knowledge-based KL-HMM parameter initialization, we investigate a case where no speech training data from the target language is used to train an ASR system.

In addition to the word list, if we assume that speech data from the target language i.e., $\{X(n)\}_{n=1}^{N}$ is also available, but without its word level transcriptions, then the knowledge-based KL-HMM parameter set can be updated in an unsupervised training scenario as follows:

- the speech data of the target language is forward passed through ANN to derive posterior probability estimates $\{Z(n)\}_{n=1}^{N}$ corresponding to $\{X(n)\}_{n=1}^{N}$.
- the knowledge-based grapheme sequence decoder (which is an ergodic KL-HMM using knowledge-based KL-HMM parameters and bigram grapheme language model) is used to generate the grapheme level transcriptions for the speech data.
- the decoded grapheme transcriptions and their posterior probability estimates $\{Z(n)\}_{n=1}^{N}$ are used to update the initial knowledge-based KL-HMM parameter set.

The unsupervised training process is illustrated in Figure 3 and consists of three steps, initial knowledge-based grapheme sequence decoder, KL-HMM training and updated grapheme sequence decoder. The KL-HMM parameter set can be updated iteratively, using newly generated grapheme transcriptions along with posterior probability estimates $\{Z(n)\}_{n=1}^{N}$.

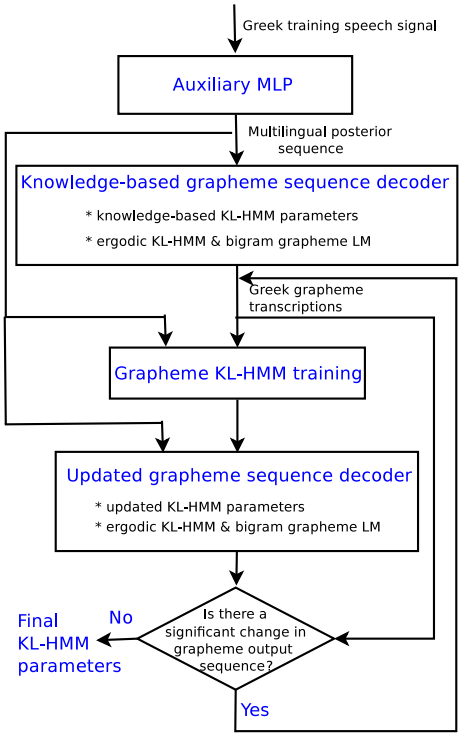

Fig. 3. Illustration of the proposed zero resourced ASR training using KL-HMM (Case 3 in Section 4)

Furthermore, with the unsupervised training process, context-dependent grapheme subword units can also be modeled in the following way: instead of aligned context-independent grapheme sequences use decoded context-independent grapheme sequences to obtain context-dependent grapheme sequences for the all the utterances of the training data, build traditional context-dependent grapheme subword based KL-HMM system using posterior probability estimates $\{Z(n)\}_{n=1}^{N}$ and their decoded grapheme sequences.

4. EXPERIMENTAL SETUP

To evaluate the proposed approach, we use SpeechDat(II) database and consider Greek as the target language. Five other European languages from SpeechDat(II) namely British English (EN), Swiss French (SF), Swiss German (SZ), Italian (IT) and Spanish (ES) are considered as auxiliary languages for which acoustic and lexical resources are available. All the ASR systems are based on KL-HMM approach.

We use an off-the-shelf multi-lingual ANN trained by pooling acoustic and lexical resources from the aforementioned five languages of SpeechDat(II) corpus as posterior feature estimator in all the experiments. All the SpeechDat(II) lexicons use SAMPA[1] symbols, therefore output of the multi-lingual ANN is formed by merging phonemes that share the same symbol across different languages to form SAMPA multi-lingual phoneme set of 117 units. Approximately, 12 hours of speech data from each language (totally amounting to 63 hours) is used to train the ANN. The input to the ANN is 39 dimensional PLP cepstral feature with 4 frames preceding context and 4 frames following context.

All the systems use grapheme subword units and grapheme lexicon, unless specified. The grapheme subword unit set contains 25 graphemes (including sil) as shown in Table 1. The grapheme lexicon which includes 35146 entries is obtained from the orthography of the words in the SpeechDat(II) Greek corpus. We study three different scenarios or cases that differ in terms of available resources for the ASR system training. More precisely, the three following cases differ in terms of the training data used in KL-HMM parameter set Θ_{kull} estimation:

- *Case 1 or Full resourced ASR study*: The Greek SpeechDat(II) corpus contains approximately 13.5 hours of training (1500 speakers), 1.5 hours of development (150 speakers) and 6.9 hours of testing (350 speakers) data. The parameters of the KL-HMM system are trained using the training part of the Greek corpus.
- *Case 2 or Zero resourced ASR with list of words available*: Only the list of possible Greek words is assumed to be available. We use a list of 35146 words from the SpeechDat(II) Greek corpus. The knowledge-based KL-HMM parameter set is defined following the procedure given in Section 3. Table 1 provides the grapheme-to-multilingual phoneme map used in defining the knowledge-based KL-HMM parameter set. Empirically it was observed on the development data that when the value of s was above 0.7, there was not much change in grapheme sequence decoded using an ergodic KL-HMM. So, we only present ASR results for the case $s = 0.8$. When a grapheme is mapped to more than one phoneme, then the value of 0.8 is equally split between dimensions of the categorical distribution accordingly. This system is illustrated in Figure 2.
- *Case 3 or Zero resourced ASR with untranscribed speech and word list*: Untranscribed Greek speech data is also assumed to be available along with the list of words. The speech data corresponding to the training set of Greek SpeechDat(II) corpus is used. The Greek training speech data is forward passed through the multi-lingual ANN to obtain multi-lingual posterior probabilities. The knowledge based KL-HMM parameter set is updated in unsupervised way using multi-lingual posterior features and decoded grapheme transcriptions as illustrated in Figure 3.

The bigram grapheme language model is built from the orthography of words (35146 words) and has perplexity of 9. The transition probabilities of ergodic grapheme KL-HMM decoder (given in Section 3) are derived from this bigram language model.

KL-HMM systems model either context-independent (mono) or context-dependent (tri) subword units. KL-divergence based decision tree state tying method proposed in [5] is used to tie context-dependent subword unit based systems. Since we do not have an appropriate language model for Greek, we build two optimistic language models as in [5], one from the sentences in the development (dev) set and other from the sentences in the test set.

For evaluating the systems we report grapheme error rate (GER) on the train set and word error rate (WER) on the dev and the test sets. In a real world zero-resourced ASR scenario, it may not be

[1] http://www.phon.ucl.ac.uk/home/sampa/

Grapheme	Trans.	Phoneme	Grapheme	Trans.	Phoneme
α	a	a, a:	ν	n	n
β	b	b, v	ξ	x	x
γ	g	g, G, j	o	o	o
δ	d	d, D	π	p	p
ϵ	e	e	ρ	r	r
ζ	z	dz, z	σ	s	s
η	h	E:,i	τ	t	t
θ	th	T	υ	y	i, y, y:
ι	i	i,i:	ϕ	f	f
κ	k	k	χ	ch	c,x
λ	l	l	ψ	ps	s
μ	m	m	ω	w	o, O:

Table 1. Greek graphemes and their transliterated format (Trans.) together with the corresponding multi-lingual phonemes

possible to compute the GER on train set as reference transcriptions are not available. However, the GER is reported in this paper as it is a simulated study (i.e., reference transcriptions are available) and can provide better understanding of training procedure. To clarify, we did not tune the insertion penalty and the language scale factor while reporting the GER on the train set.

5. RESULTS AND ANALYSIS

Table 2 presents the GER and the WER on the train and the test sets respectively for *Case 1* i.e., full resourced KL-HMM based ASR system. For the sake of comparison, we also provide WER of KL-HMM systems modeling context-independent and context-dependent phoneme subword units from [5]. Grapheme error rate for phoneme subword based systems is not available. In spite of training the KL-HMM parameters on supervised training data, the GER on the training set is 47.5% for context-independent grapheme system and 42.5% for context-dependent grapheme system. Phoneme subwords achieve better performance when context-independent units are modeled, whereas grapheme subword units achieve better performance when context-dependent subword units are modeled. This could be due to the one-to-one and regular grapheme-to-phoneme relationship of the Greek language.

Subword units	Mono		Tri	
	GER/train	WER/test	GER/train	WER/test
Graphemes	47.5	23.4	42.5	14.2
Phonemes [5]	n.a.	20.5	n.a.	15.2

Table 2. *Case 1* or full resourced KL-HMM systems trained on 800 minutes of data

Table 3 reports the GER and the WER for *Case 2* and *Case 3* KL-HMM systems modeling context-independent grapheme subword units. Column 1 (or iteration 0) of the table indicates *Case 2*. The results indicate that the proposed approach can build ASR systems in zero-resourced setup with minimal knowledge. However, compared to the full resourced monograph KL-HMM system GER increases by 7% relative and WER increases by 47% relative, which is expected because of minimal supervision and training resources.

The results also show that unsupervised adaptation of the knowledge based KL-HMM parameters i.e., *Case 3* (or iteration 1) significantly improves the performance of KL-HMM system (12% absolute improvement in WER compared to *Case 2*). However, the unsupervised adaptation seems to converge in just one iteration as GER and WER on train and test sets do not change significantly. This can be due to relatively small number of KL-HMM parameters ($3 * 25 * 117$).

	Iteration	GER/train	WER/dev	WER/test
Case 2	0	54.9	39.1	43.0
Case 3	1	51.4	29.3	31.2
Case 3	2	51.4	29.9	31.6

Table 3. The GER on train and the WER on the test and the dev sets for *Case 2* and *Case 3* context-independent grapheme KL-HMM systems

Table 4 reports the GER and the WER on the train and the test sets respectively for *Case 3* KL-HMM systems modeling context-dependent subword units. The mono grapheme transcriptions from iteration 1 of Table 3 serve as the reference transcriptions to train context-dependent unsupervised KL-HMM system. Results show that the unsupervised context-dependent subword model training of KL-HMM parameters improves the performance of system (4% absolute decrease in WER compared to unsupervised context-independent subword model training of KL-HMM parameters). Again, the grapheme transcriptions derived from this system are used to iteratively update the unsupervised context-dependent subword KL-HMM parameters. Even though there is a slight improvement in GER, ASR results on the test and dev sets show that further update of the context-dependent models does not yield any improvement in performance.

	Iteration	GER/train	WER/dev	WER/test
Case 3	1	50.9	25.9	27.7
Case 3	2	50.1	26.1	27.7
Case 3	3	50.1	25.9	27.7

Table 4. The GER on train and the WER on test and dev sets for the *Case 3* or context-dependent grapheme KL-HMM systems trained using unsupervised training procedure

To summarize, our results indicate that the knowledge based monograph KL-HMM system using just the word list from the target Greek language achieves WER of 43.0%. By unsupervised adaptation of the monograph KL-HMM parameters using speech data from target domain, the system achieves WER of 31.2% (27% relative reduction in WER compared to the system using only the word list). Furthermore, by trigraph modeling and unsupervised adaptation of the KL-HMM parameters the system achieves WER of 27.7% (36.6% relative reduction in WER compared to the system using only the word list). However, results indicate that the KL-HMM system trained using unsupervised procedure increases the WER by 51% relative compared to the optimized system trained on 13.5 hours of supervised data.

In order to better understand the trends observed in Tables 2, 3 and 4, we analyzed the GER obtained for different systems. Table 5 presents a more detailed analysis of the performance of various systems. It can be observed from the table that the high grapheme error rate is because of very high deletion rates, followed by substitutions rates. Further analysis based on the confusion matrix revealed that more than 50% of the deletions are because of 5 vowels. The high number of grapheme vowel deletions could be because of the vowel digraphs in Greek.

Scenario	Context	Corr	Sub	Del	Ins	Err
Case 1	mono	54.6	17.4	28.1	2.0	47.5
Case 2	mono	48.9	19.6	31.4	3.8	54.9
Case 3	mono	50.9	19.3	29.7	2.3	51.4
Case 1	tri	59.8	14.9	25.2	2.3	42.5
Case 3	tri	50.9	18.1	30.9	1.9	50.9

Table 5. The GER on train set split into percent correct (Corr), percent substituted (Sub), percent deleted (Del), percent inserted (Ins) and overall error (Err) for various scenarios

The GER of the system based on *Case 2* was also analyzed in terms of number of utterances in various GER ranges. Figure 4 plots the histogram of the GER against the number of utterances. The figure shows that about one third of utterances have more than 50% GER. From the histogram we can infer that careful selection of utterances may lead to better systems in case of unsupervised adaptation.

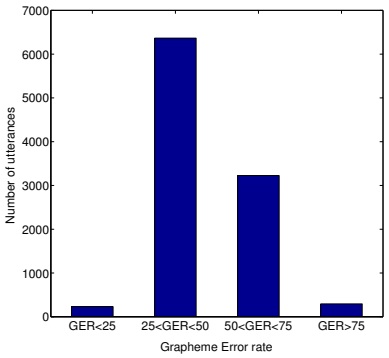

Fig. 4. Histogram of the GER and the number of utterances for *Case 2* system

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a novel approach to build ASR systems in zero-resourced scenario where a) the acoustic model or an ANN is trained using acoustic and lexical resources from auxiliary languages and b) a probabilistic lexical model (or KL-HMM parameter set) for the target language is estimated by exploiting the knowledge of grapheme-to-phoneme mapping. Furthermore, if untranscribed speech data from the target language is available then the knowledge-based KL-HMM parameter set could be adapted/re-estimated in unsupervised manner using graphemic constraints learned from available word list. Our investigations on a simulated Greek ASR task showed that the proposed approach can facilitate building ASR systems in zero-resourced setting with minimal knowledge or supervision.

The proposed approach could be improved further by,

- incorporating more knowledge about the Greek language, like digraphs, while building grapheme lexicon and bigram grapheme language model.
- replacing the bigram grapheme network with trigram grapheme network to better incorporate the graphemic constraints.
- adopting confidence measures during unsupervised adaptation to prune utterances with high grapheme error rate.
- adopting semi-supervised training if small amount of transcribed speech data from the target language is available.

We intend to investigate these aspects in our future work and extend the studies to other languages where the grapheme-to-phoneme relationship may not be as regular as Greek.

7. REFERENCES

[1] L. Viet-Bac and L. Besacier, "Automatic Speech Recognition for Under-Resourced Languages: Application to Vietnamese Language," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, pp. 1471–1482, Nov 2009.

[2] L. Lu, A. Ghoshal, and S. Renals, "Regularized Subspace Gaussian Mixture Models for Cross-lingual Speech Recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2011*, 2011, pp. 365–370.

[3] N. Thang Vu, F. Kraus, and T. Schultz, "Rapid building of an ASR system for Under-Resourced Languages based on Multilingual Unsupervised Training," in *Proc. of Interspeech*, 2011, pp. 3145–3148.

[4] Y. Qian, D. Povey, and J. Liu, "State-Level Data Borrowing for Low-Resource Speech Recognition based on Subspace GMMs," in *Proc. of Interspeech*, 2011.

[5] D. Imseng, J. Dines, P. Motlicek, P. N. Garner, and H. Bourlard, "Comparing different acoustic modeling techniques for multilingual boosting," in *Proc. of Interspeech*, 2012.

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

[7] S. Thomas, M. Seltzer, K. Church, and H. Hermansky, "Deep Neural Network Features and Semi-supervised training for Low Resource Speech Recognition," in *Proc. of ICASSP*, 2013.

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

[9] D. Imseng, R. Rasipuram, and 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)*, 2011, pp. 348–353.

[10] G. Aradilla, H. Bourlard, and M. Magimai Doss, "Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task ," in *Proc. of Interspeech*, 2008, pp. 928–931.

[11] R. Rasipuram and M. Magimai.-Doss, "Improving Grapheme-based ASR by Probabilistic Lexical Modeling Approach," in *Proc. of Interspeech*, 2013.

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

[13] F. Wessel and H. Ney, "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 23–31, 2005.