Open access • Proceedings Article • DOI:10.1109/ASRU.2011.6163956

Fast and flexible Kullback-Leibler divergence based acoustic modeling for non-native speech recognition — Source link

David Imseng, Ramya Rasipuram, Mathew Magimai.-Doss

Institutions: Idiap Research Institute

Published on: 01 Dec 2011 - IEEE Automatic Speech Recognition and Understanding Workshop

Topics: Word error rate, Hidden Markov model, Multilayer perceptron, Feature (machine learning) and Kullback–Leibler divergence

Fast and flexible Kullback-Leibler divergence based acoustic modeling for non-native speech recognition

David Imseng [#†1], Ramya Rasipuram [#†2], Mathew Magimai.-Doss [#3]

[#] *Idiap Research Institute, P.O. Box 592, CH-1920 Martigny, Switzerland*
[†] *Ecole Polytechnique Fédérale Lausanne (EPFL), CH-1015 Lausanne, Switzerland*
[1] dimseng@idiap.ch, [2] ramya.rasipuram@idiap.ch, [3] mathew@idiap.ch

***Abstract*—One of the main challenge in non-native speech recognition is how to handle acoustic variability present in multi-accented non-native speech with limited amount of training data. In this paper, we investigate an approach that addresses this challenge by using Kullback-Leibler divergence based hidden Markov models (KL-HMM). More precisely, the acoustic variability in the multi-accented speech is handled by using multilingual phoneme posterior probabilities, estimated by a multilayer perceptron trained on auxiliary data, as input feature for the KL-HMM system. With limited training data, we then build better acoustic models by exploiting the advantage that the KL-HMM system has fewer number of parameters. On HIWIRE corpus, the proposed approach yields a performance of 1.9% word error rate (WER) with 149 minutes of training data and a performance of 5.5% WER with 2 minutes of training data.**

***Index Terms*—Non-native speech recognition, hidden Markov model, posterior features, Kullback-Leibler divergence, multilayer perceptron**

I. Introduction

Non-native speech recognition is a challenging task for reasons such as, a) there are large number of non-native accents, and b) usually only a small amount of non-native speech data is available for training. In literature, several methods based on acoustic model adaptation, pronunciation model adaptation, or both have been proposed to improve automatic speech recognition (ASR) system performance on non-native speech. In acoustic model adaptation based approaches, the native language acoustic models are pooled and adapted to the non-native speaker/accent using small amount of non-native speech. In the framework of hidden Markov models/Gaussian mixture models (HMM/GMM) system, traditional adaptation methods such as, maximum likelihood linear regression (MLLR), maximum a posteriori (MAP), and model interpolation have been used [1], [2]. While, in the framework of hybrid hidden Markov modesl/multilayer perceptron (HMM/MLP) system, linear hidden network (LHN) based adaptation has been used to improve the performance [3]. In the area of pronunciation modeling, attempts have been made to detect and correct the non-native pronunciations using small amount of non-native speech data [4].

Kullback-Leibler divergence based HMM (KL-HMM) is a recently proposed acoustic modeling approach where the acoustic class conditional probabilities are *directly* used as feature observation [5], [6]. We refer to these features as *posterior features*. As described in detail in Section II, in this approach the emission distribution of each HMM state is modeled by a multinomial distribution, and the cost function used to optimize the multinomial distribution is based on Kullback-Leibler divergence.

The KL-HMM system provides flexibilities such as transfer learning, fewer number of parameters, choice of acoustic classes (i.e., posterior features), and use of alternate subword units such as, graphemes. These flexibilities can be exploited to address the challenges involved in non-native speech recognition. In that regard, this paper investigates

1) an approach where the acoustic variation in multi-accented speech is modeled through the use of *universal phoneme class conditional probabilities* as posterior features in the KL-HMM system. These features are estimated by training an MLP on (auxiliary) multilingual speech data. We compare it against the approach where the MLP is trained on monolingual speech data, in our case English data.
2) the use of graphemes as subword units and compare it with the standard approach of using phonemes as subword units. The use of graphemes eases pronunciation lexicon generation. In addition, it could avoid the necessity to generate multiple pronunciation variants.
3) fast acoustic model training/adaptation using small amounts of non-native speech data by exploiting the flexibility that KL-HMM system has fewer number of parameters, especially when the posterior extractor is trained on auxiliary data.

Experimental studies conducted on the HIWIRE corpus [1] shows that a) universal posterior features yield better performance than monolingual posterior features, b) systems based on phoneme subword units and grapheme subword units perform equally well, and c) a "reasonable" ASR performance could be achieved with training data as low as two minutes speech.

The remainder of the paper is organized as follows. Section II describes the KL-HMM system. Section III then motivates the use of KL-HMM for non-native speech recognition. Section IV describes the different systems that are investigated and experimental results are presented in Section V. Finally, Section VI summarizes and concludes the paper.

II. KULLBACK-LEIBLER DIVERGENCE BASED ACOUSTIC MODELING

As mentioned briefly earlier in Section I, KL-HMM directly use the acoustic class conditional probabilities as features, i.e. posterior features [5], [6]. Posterior features can be seen as a data driven feature. More precisely, posterior feature extraction involves the transformation of standard acoustic feature vector x_t of dimension F,

$$x_t = \begin{pmatrix} x_t(1) \\ \vdots \\ x_t(F) \end{pmatrix}$$

at time frame t into a class conditional probability vector z_t of dimension K,

$$z_t = \begin{pmatrix} z_t(1) \\ \vdots \\ z_t(K) \end{pmatrix} = \begin{pmatrix} P(c^1|x_t,\theta) \\ \vdots \\ P(c^K|x_t,\theta) \end{pmatrix}$$

where, $\{c^1, \cdots, c^K\}$ denotes the acoustic classes and θ the parameters of the model/classifier that is used to estimate the probabilities. The model/classifier can be a well trained discriminative classifier such as, an MLP or a generative classifier such as, GMM[1].

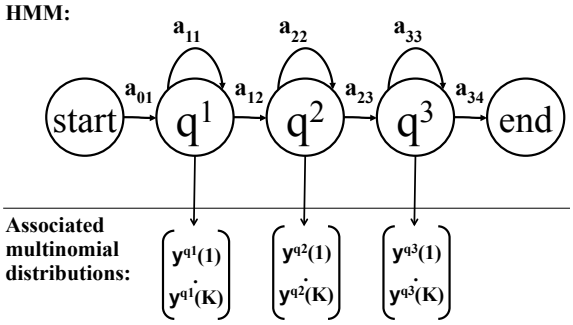

Fig. 1. Each state is parametrized by a multinomial distribution of dimensionality K. The transition probabilities are also parameters of the HMM.

Formally, in the KL-HMM system, z_t is the feature observation and each state is parametrized by a multinomial distribution. Figure 1 illustrates a KL-HMM consisting of three emitting states q^1, q^2, q^3 and two non-emitting start and end states. Each emitting state $q^d : d \in \{1, \ldots, D\}$, where D is the total number of states, is parametrized by a multinomial distribution

$$\begin{pmatrix} y^{q^d}(1) \\ \vdots \\ y^{q^d}(K) \end{pmatrix}$$

with K being the dimensionality of the posterior feature. The KL-HMM acoustic model is completely parameterized by $\Theta_{KL} = \{Y, A\}$, where $Y = \{y^{q^1}, \cdots, y^{q^D}\}$ is the set containing state multinomial distributions and A is the state transition probability matrix.

[1]Note that GMM can be trained discriminatively, but then they may not be truly generative models

1) Training: Let us assume that we have access to a set of T training frames along with labels, i.e:

- a sequence of posterior probability feature vectors $Z = \{z_1, \cdots, z_T\}$.
- transcription in terms of subword units, e.g. phonemes.

Then, the training phase involves the estimation of Θ_{KL} by optimizing a cost function based on the Kullback-Leibler (KL) divergence. More precisely, this is done by Viterbi expectation maximization training algorithm which minimizes

$$\arg\min_{\mathcal{Q}} \sum_{t=1}^{T} [f(y^{q_t}, z_t) - \log(a_{q_{t-1},q_t})] \tag{1}$$

where, $q_t \in \{q^1, \cdots, q^D\}$ and $\mathcal{Q}$ is the set of all possible state sequences $\{q_1, \cdots, q_T\}$ allowed by the HMM corresponding to the training frames. The local score $f(y^{q_t}, z_t)$ is the KL-divergence between the multinomial state distribution y^{q_t} and the observation posterior feature vector z_t.

2) Decoding: Given a sequence of posterior feature vectors $Z^b = \{z_1, \cdots z_{T^b}\}$ corresponding to a test utterance b and the set of trained parameters $\Theta_{KL} = \{Y, A\}$, the decoding phase involves recognition of hypothesis $\hat{m}_b$ as follows:

$$\hat{m}_b = \arg\min_{\mathcal{Q}_m} \sum_{t=1}^{T^b} [f(y^{q_t}, z_t) - \log(a_{q_{t-1},q_t})]$$

where, $\mathcal{Q}_m$ represents the set of all possible state sequences allowed by hypothesis m.

KL-divergence being an asymmetric measure, the local score $f(y^{q^d}, z_t)$ can be estimated as:

$$f_{KL}(y^{q^d}, z_t) = \sum_{k=1}^{K} y^{q^d}(k) \log \frac{y^{q^d}(k)}{z_t(k)} \tag{2}$$

$$f_{RKL}(y^{q^d}, z_t) = \sum_{k=1}^{K} z_t(k) \log \frac{z_t(k)}{y^{q^d}(k)} \tag{3}$$

$$f_{SKL}(y^{q^d}, z_t) = \frac{1}{2} f_{KL}(y^{q^d}, z_t) + \frac{1}{2} f_{RKL}(y^{q^d}, z_t) \tag{4}$$

Through the use of different local scores, KL-HMM establishes a framework that unifies different types of acoustic models, such as discrete HMM and HMM/MLP [6, Chapter 6]. For instance, when using MLP for posterior feature extraction the system using the local score $f_{KL}(y^{q^d}, z_t)$ can be linked to HMM/MLP systems. While, the system using the local score $f_{RKL}(y^{q^d}, z_t)$ can be linked to discrete HMM systems, where the MLP acts as a vector quantizer [6]. ASR studies until now have shown that $f_{SKL}(y^{q^d}, z_t)$ yields the best system [5], [6], [7].

In this work, as done in the original work [6], an MLP that is trained to classify context-independent phonemes is used. The choice of MLP for posterior feature extraction can be motivated by reasons such as, a) a well trained MLP can directly estimate a posteriori probabilities of output classes [8], b) discriminative training can provide invariance towards undesirable variabilities such as, speaker and environment, c) posterior feature estimation could be improved by

combining multiple feature streams at MLP output level [9] or using hierarchical approaches, and d) MLPs can be effectively employed for transfer learning [10], i.e. they can be trained on auxiliary data and used for different tasks.

III. Motivation to use KL-HMM for Non-native Speech Recognition

KL-HMM provide certain advantages which can be effectively exploited for non-native speech recognition such as,

1) Transfer learning: the posterior feature estimator could be trained on auxiliary data. In the context of non-native speech recognition, this can be effectively used to pool resources from multiple databases and languages.
2) Choice of posterior feature space: the posterior feature space can be phonemes that are specific to a language, universal phonemes, or articulatory features. Thus, KL-HMM systems provide a framework to introduce multilingual knowledge. This may be essential for improving the performance of ASR system on multi-accented non-native speech. Along this direction, in this work we propose to use universal phonemes as acoustic classes.
3) Choice of subword units: as the phonetic information is captured via posterior feature space, KL-HMM allow the possibility to use alternate subword units, such as graphemes. One of the main advantage of using graphemes as subword units is that it avoids the need to build a lexicon. In KL-HMM systems, when graphemes are used as subword units, the parameters of the system, i.e. K dimensional multinomial distribution per state, can capture the relation between written and spoken form of the language [7]. This could be useful in the context of non-native speech recognition. For instance, in [4] graphemic constraints were introduced in the phonetic confusion because the writing of uttered words may influence the pronunciations produced by non-native speakers. Thus, in addition to phonemes, in this paper we also investigate the use of graphemes as subword units for non-native speech recognition.
4) Fewer number of parameters: As described earlier in Section II, each emitting state is modeled by a K dimensional multinomial distribution. Thus, the KL-HMM system has fewer number of parameters that needs to be estimated during training. This suggests that KL-HMM systems may require less training data to adapt to multiple accents. We also explore this direction in this paper.

IV. Experimental setup

In this section, we first describe the datasets we used followed by the details about posterior feature extraction and the investigated systems.

A. Dataset

We use HIWIRE [1] for our experimental studies. HIWIRE is a non-native English speech corpus that contains English utterances pronounced by natives of France (31 speakers), Greece (20 speakers), Italy (20 speakers) and Spain (10 speakers). The utterances contain spoken pilot orders made up of 133 words and the database also provides a grammar with a perplexity of 14.9. The phoneme dictionary is in CMU format and makes use of 38 ARPABET[2] phonemes. The grapheme dictionary was transcribed using 29 context-independent graphemes including silence. The abbreviation words present in the dictionary were transcribed using a look up table specifying the way individual letters are pronounced as shown in Table I.

TABLE I
Example of lookup table entries used for transcribing letters in the abbreviations, for grapheme dictionary

Letter	Grapheme pronunciation
D	[D] [E] [E]
F	[E] [F]
I	[E] [Y] [E]
S	[E] [S]
T	[T] [E] [E]

The phoneme dictionary consists of pronunciation variants, i.e. multiple pronunciations for some words, whereas grapheme dictionary consists of single pronunciation for each word. HIWIRE consists of 100 recordings per speaker, of which the first 50 utterances are commonly defined to serve as adaptation data and the rest of the 50 utterances as testing data.

For training posterior feature extractors i.e. MLPs, we used SpeechDat(II) [3] data. More specifically, we used data from the British English, Italian, Spanish, Swiss French and Swiss German SpeechDat(II) databases. All SpeechDat(II) databases contain native speech. Furthermore, the data is also gender-balanced, dialect-balanced according to the dialect distribution in a language region, and age-balanced. The databases have been recorded over the telephone at 8 kHz and are subdivided into different corpora. We only used *Corpus S*, that contains ten read sentences from each of the 2000 speakers per language. To split the databases into training (1500 speakers), development (150 speakers) and testing (350 speakers) sets, we used the standard procedure that maintains the gender-, dialect- and age-distributions of the database, as described in [11]. Only the training portion was used for this study.

B. Posterior features

As discussed in Section II, KL-HMM use posterior probabilities of acoustic classes, i.e. posterior features as feature observation. To estimate these features, we train MLPs to classify context-independent phonemes using Quicknet[4] software. The feature input to the MLPs is 39 Mel-Frequency Perceptual Linear Prediction (MF-PLP) cepstral features ($C_0 - C_{12} + \Delta + \Delta\Delta$) with a temporal context of four preceding frames and four following frames. In this paper, we investigate posterior

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[3] http://www.speechdat.org/SpeechDat.html
[4] http://www.icsi.berkeley.edu/Speech/qn.html

features estimated using two different source phoneme sets in SAMPA[5] format.

- English phoneme set: we use only the British English data to train a MLP to estimate English SAMPA phoneme posteriors. We denote this MLP as MLP-EN.
- Universal phoneme set: since all the SpeechDat(II) dictionaries use SAMPA symbols, we merged phonemes that share the same symbol across languages to build a universal phoneme set. We then train an MLP to estimate universal phoneme posteriors using the data from five different European languages (British English, Italian, Spanish, Swiss French and Swiss German). We denote this MLP as MLP-UNI.

The number of parameters in each MLP was set to 10% of the number of available training frames and the MLPs were trained with cross entropy error criteria. For more details about the MLP training the reader is referred to [12]. Table II summarizes the posterior feature extraction systems.

TABLE II
OVERVIEW OVER ALL THE PHONEME POSTERIOR ESTIMATORS. THE TOTAL AMOUNT OF TRAINING DATA AS WELL AS THE PHONEME SET INCLUDING THE NUMBER OF PHONEMES K ARE GIVEN.

	Phoneme set	K	Data (in hours)
MLP-EN	SAMPA English	45	12.4
MLP-UNI	SAMPA universal	117	63.0

After training, MLP-EN and MLP-UNI were used to extract English posterior features and universal posterior features, respectively on the HIWIRE corpus, and used as feature observations for the KL-HMM system. It is to be noted that since HIWIRE was recorded at 16 kHz and SpeechDat(II) was recorded at 8 kHz, the HIWIRE recordings were down-sampled to 8 kHz before extracting the MF-PLP features.

C. Systems

We study non-native speech recognition in the framework of KL-HMM using two types of subword units, namely, phonemes and graphemes. The KL-HMM are trained with either English phoneme posterior features estimated by MLP-EN or universal phoneme posterior features estimated by MLP-UNI. Each subword unit is represented by a three state left-to-right HMM. The multinomial state distributions are estimated by optimizing an objective function based on a symmetric variant of the Kullback-Leibler divergence (as discussed in Section II) over the adaptation set of the HIWIRE corpus. The insertion penalty and the language scaling factor were tuned on the adaptation set of the HIWIRE corpus. We study both context-independent subword unit (mono-phoneme/mono-grapheme) and word internal context-dependent subword unit (tri-phoneme/tri-grapheme) based systems. In the case of context-dependent subword unit based system, for the unseen contexts the corresponding context-independent subword unit models were used. Table III gives an overview of the investigated systems.

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

TABLE III
OVERVIEW OF THE DIFFERENT SYSTEMS INVESTIGATED. WE COMBINE EACH FEATURE TYPE WITH EACH WORD UNIT TYPE.

System	Subword unit	Features
PHONE-EN	Phonemes	English posteriors
GRAPH-EN	Graphemes	English posteriors
PHONE-UNI	Phonemes	Universal posteriors
GRAPH-UNI	Graphemes	Universal posteriors

V. RESULTS

We first present results for different features and subword units and then for low amounts of adaptation data.

A. Varying posterior features and subword units

Table IV shows the word error rates (WERs) on the test set of the HIWIRE database for context-independent subword unit systems. Results reveal that phoneme subword units yield better performance than grapheme subword units. This is not surprising as the correspondence between phoneme and grapheme is weak in English language. As a result, the HMM of context-independent grapheme subword unit captures only gross phonetic information in their multinomial state distributions, and are thus ambiguous [7]. Universal phoneme posterior features yield significantly better performance than English phoneme posterior features for both phoneme and grapheme subword units.

TABLE IV
WORD ERROR RATES ON VARIOUS LANGUAGES OF HIWIRE DATABASE USING CONTEXT INDEPENDENT SUBWORD UNITS (MONO-PHONEMES AND MONO-GRAPHEMES). FR DENOTES FRENCH ACCENT, GR DENOTES GREEK ACCENT, IT DENOTES ITALIAN ACCENT, AND SP DENOTES SPANISH ACCENT.

System	FR	GR	IT	SP	Total
PHONE-EN	4.8	3.3	6.0	5.5	4.8
GRAPH-EN	13.0	10.7	14.0	13.9	12.8
PHONE-UNI	2.6	1.8	3.8	3.5	2.8
GRAPH-UNI	10.2	6.1	8.6	9.2	8.6

Table V shows the WERs on the test set of the HIWIRE database for context-dependent subword unit systems. Interestingly, grapheme-based systems yield same performance as phoneme-based systems. For the English language, it has been observed that grapheme-based systems may require longer subword contexts to be modeled to effectively disambiguate between phonemes, and achieve performance as good as phoneme based system [7]. However, in this case, single preceding and single following context is sufficient. The main reasons for this trend could be that a) longer context may be more important for native speech than non-native speech which could contain more variation at phonetic level. In such a case, a smaller grapheme context could be sufficient. This point needs further investigation and is part of our future work. b) HIWIRE task is relatively constrained task when compared to large vocabulary tasks. As already observed with context-independent subword unit systems, universal phoneme posterior features outperform English phoneme posterior features. Thus, signifying the importance of multilingual features for non-native speech recognition.

TABLE V
WORD ERROR RATES ON NON-NATIVE ACCENTS OF HIWIRE DATABASE USING SINGLE PRECEDING AND SINGLE FOLLOWING CONTEXT-DEPENDENT SUBWORD UNIT MODELS. FR DENOTES FRENCH ACCENT, GR DENOTES GREEK ACCENT, IT DENOTES ITALIAN ACCENT, AND SP DENOTES SPANISH ACCENT.

System	FR	GR	IT	SP	Total
PHONE-EN	2.2	1.8	4.2	3.0	2.7
GRAPH-EN	2.3	2.2	3.6	2.9	2.7
PHONE-UNI	1.8	1.2	2.5	2.1	1.9
GRAPH-UNI	1.7	1.4	2.5	2.0	1.9

B. Fast training/adaptation

To investigate the behavior of the KL-HMM system when there is only little amount of training data, we decreased the amount of adaptation data progressively. The standard adaptation set of HIWIRE consists of 50 sentences per speaker. We decreased the amount of adaptation data by only taking 40, 30, 20, ten, five and three sentences per speaker. Three sentences per speaker is about ten minutes of adaptation data. To ensure full coverage in terms of context-independent phonemes and graphemes, we picked different sentences for the three, five and ten sentences scenario for the phoneme- and grapheme-based systems, respectively. We went further down and randomly picked utterances until full coverage of context-independent phonemes and graphemes was achieved. This resulted in three minutes of data. Instead of randomly picking sentences, we also explored manual utterance selection. This allowed us to decrease the amount of data to two minutes.

For this study, we trained context-dependent phoneme- and grapheme-based systems using universal phoneme posterior features. As mentioned earlier in Section IV-C, unseen context-dependent subword unit models were backed-off to context-independent subword unit models.

Table VI compares the phoneme- and grapheme-based KL-HMM systems. It can be observed that with 60 minutes or more, the two systems yield same performance. The phoneme-based system is clearly superior to the grapheme-based system when very little adaptation data is used. As discussed earlier, in the case of graphemes, contextual modeling is important, especially for languages like English. Therefore, the grapheme-based system requires more adaptation data than the phoneme-based system.

TABLE VI
UNIVERSAL POSTERIORS TRI-GRAPHEME AND TRI-PHONEME SUBWORDS SYSTEM PERFORMANCE FOR DIFFERENT AMOUNTS OF DATA.

Minutes	Sentences per speaker	Graphemes	Phonemes
2	-	21.6	5.5
3	-	13.8	4.4
10	3	5.2	3.9
16	5	5.1	3.2
32	10	3.1	2.5
64	20	2.1	2.0
90	30	1.9	2.0
122	40	1.9	1.9
149	50	1.9	1.9

Figure 2 compares the performances of phoneme- and grapheme-based systems to the results reported in the literature on the same setup. It can be observed that KL-HMM systems outperforms the MLLR-based speaker adaptation for all amounts of adaptation data that have been investigated. In [3], two different linear hidden network (LHN) based adaptation approaches have been investigated, namely, LHN based speaker adaptation and LHN based data adaptation. For the LHN based speaker adaptation, an extra hidden layer was trained for each speaker separately. While, in the case of LHN based data adaptation an extra hidden layer was added and trained on the whole adaptation data. It can be seen that the KL-HMM system clearly outperforms the system based on speaker adaptation and achieves performance similar to the system based on data adaptation.

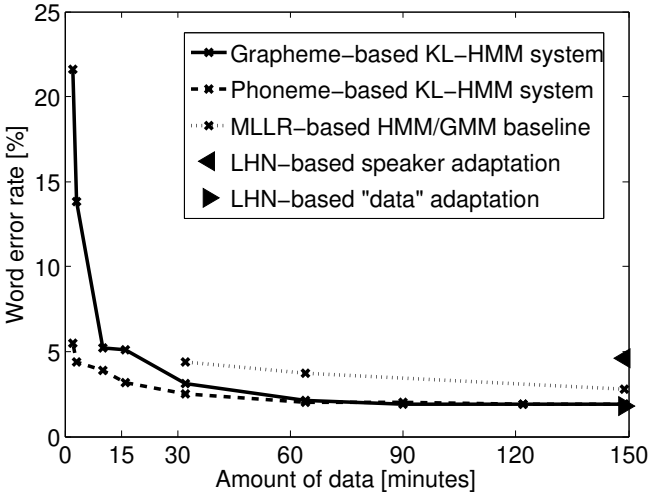

Fig. 2. We compare the KL-HMM systems with different amounts of adaptation data to previous studies reported in the literature. The MLLR studies are published in [1] and the LHN approaches in [3].

Finally, in the literature the best result reported on HIWIRE task is 1.4% WER [2]. However, this performance has been obtained by a modified setup where, after excluding the data of the test speaker, the rest of the corpus is used for training/adaptation. In that sense, the performances achieved with the KL-HMM system can be considered as one of the best.

VI. SUMMARY AND CONCLUSION

In this paper, we investigated how the flexibilities provided by KL-HMM can be exploited for non-native speech recognition. The main findings from our investigations are summarized as follows,

- KL-HMM framework is able to exploit multilingual information in the form of universal phoneme posterior probabilities to improve performance on non-native speech.
- Graphemes can be used as an alternative subword unit to phonemes. This could possibly help in reducing dictionary building efforts.
- KL-HMM could be trained rapidly with small amount of non-native speech data.

- KL-HMM outperform previously reported MLLR-based and LHN-based speaker adaptation techniques on the HIWIRE dataset and yields the same performance as LHN-based data adaptation technique.

In our future work, we intend to investigate a) the use of articulatory features for non-native speech recognition, b) the effect of unseen non-native accents, c) longer subword unit context modeling, and d) approaches to tie context-dependent subword unit models to handle unseen contexts.

ACKNOWLEDGMENT

This research was supported by the Swiss NSF through the project Interactive Cognitive Systems (ICS) under contract number 200021_132619/1, through the project FlexASR, and through the National Center of Competence in Research on "Interactive Multimodal Information Management" (www.im2.ch). The authors would like to thank Prof. Hervé Bourlard for providing critical inputs and guidance during the course of this work.

REFERENCES

[1] J. C. Segura *et al.*, "The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication," 2007. [Online]. Available: http://cvsp.cs.ntua.gr/projects/pub/HIWIRE/WebHome/HIWIRE_db_description_paper.pdf

[2] G. Bouselmi, D. Fohr, and I. Illina, "Multi-accent and accent-independent non-native speech recognition," in *Proc. of Interspeech*, 2008.

[3] R. Gemello, F. Mana, and S. Scanzio, "Experiments on HIWIRE database using denoising and adaptation with a hybrid HMM-ANN model," in *Proc. of Interspeech*, 2007, pp. 2429–2432.

[4] G. Bouselmi, D. Fohr, I. Illina, and J.-P. Haton, "Multilingual non-native speech recognition using phonetic confusion-based acoustic model modification and graphemic constraints," in *Proc. of Interspeech*, 2006, pp. 109–112.

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

[6] G. Aradilla, "Acoustic models for posterior features in speech recognition," Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne, 2008.

[7] M. Magimai.-Doss, R. Rasipuram, G. Aradilla, and H. Bourlard, "Grapheme-based automatic speech recognition using KL-HMM," in *Proc. of Interspeech*, 2011.

[8] H. Bourlard and N. Morgan, *Connectionist Speech Recognition – A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[9] F. Valente, "A novel criterion for classifiers combination in multistream speech recognition," *IEEE Signal Processing Letters*, vol. 16, no. 7, pp. 561–564, July 2009.

[10] L. Toth, J. Frankel, G. Gosztolya, and S. King, "Cross-lingual portability of MLP-based tandem features - a case study for English and Hungarian," in *Proc. of Interspeech*, 2008.

[11] D. Imseng, H. Bourlard, and M. Magimai.-Doss, "Towards mixed language speech recognition systems," in *Proc. of Interspeech*, 2010, pp. 278–281.

[12] D. Imseng, H. Bourlard, M. Magimai.-Doss, and J. Dines, "Language dependent universal phoneme posterior estimation for mixed language speech recognition," in *Proc. of ICASSP*, 2011, pp. 5012–5015.