IDIAP RESEARCH REPORT

ON RECOGNITION OF NON-NATIVE SPEECH USING PROBABILISTIC LEXICAL MODEL

Marzieh Razavi Mathew Magimai-Doss

Idiap-RR-11-2014

JUNE 2014

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

On Recognition of Non-Native Speech Using Probabilistic Lexical Model

Marzieh Razavi[1,2] *and Mathew Magimai Doss*[1]

[1]Idiap Research Institute, CH-1920 Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
marzieh.razavi@idiap.ch, mathew@idiap.ch

Abstract

Despite various advances in automatic speech recognition (ASR) technology, recognition of speech uttered by non-native speakers is still a challenging problem. In this paper, we investigate the role of different factors such as type of lexical model and choice of acoustic units in recognition of speech uttered by non-native speakers. More precisely, we investigate the influence of the probabilistic lexical model in the framework of Kullback-Leibler divergence based hidden Markov model (KL-HMM) approach in handling pronunciation variabilities by comparing it against hybrid HMM/artificial neural network (ANN) approach where the lexical model is deterministic. Moreover, we study the effect of acoustic units (being context-independent or clustered context-dependent phones) on ASR performance in both KL-HMM and hybrid HMM/ANN frameworks. Our experimental studies on French part of MediaParl as a bilingual corpus indicate that the probabilistic lexical modeling approach in the KL-HMM framework can capture the pronunciation variations present in non-native speech effectively. More precisely, the experimental results show that the KL-HMM system using context-dependent acoustic units and trained solely on native speech data can lead to better ASR performance than adaptation techniques such as maximum likelihood linear regression.

Index Terms: Non-native speech recognition, Kullback-Leibler divergence based hidden Markov model, Probabilistic lexical modeling

1. Introduction

There is growing interest in the speech community to improve the speech recognition for non-native speech as notable number of people in today's world using speech technology applications are non-native speakers. Non-native speech recognition can be a challenging problem due to existence of various accents [1] while only small amount of non-native speech data is available.

Several adaptation methods have been proposed to improve the automatic speech recognition (ASR) on non-native speech data. A wide range of such techniques exploit acoustic model adaptation. For example, in the framework of hidden Markov model/Gaussian mixture model (HMM/GMM), Gaussian parameters are adapted using maximum likelihood linear regression (MLLR) or maximum a posteriori (MAP) estimation [2, 3]. On the other hand, in the framework of hybrid HMM/artificial neural network (ANN), linear hidden network based adaptation techniques have been applied [4]. Other approaches have also been proposed in which a multilingual acoustic model is exploited for the task of non-native speech recognition [5]. Another existing class of adaptation techniques is applied at the pronunciation level. In [6], pronunciation model adapatation using small amount of non-native speech data was explored. Furthermore, approaches to adapt context-dependent state clustering methods such as polyphone decision tree specialization (PDTS) method for non-native speech recognition have been proposed [7, 8].

All the adaptation methods discussed above have been proposed within the framework of standard HMM-based ASR systems. In such systems, as explained in Section 2, the relationship between the acoustic/physical states and lexical/logical states is deterministic (deterministic lexical model). An alternative approach which has been shown to be successful in improving non-native speech recognition is a posterior-based ASR approach called Kullback-Leibler divergence based HMM (KL-HMM) [9, 10]. KL-HMM can be viewed as an ASR approach where the relation between the acoustic units (modeled through ANN) and lexical units (modeled through KL-HMM) is probabilistic [11] (probabilistic lexical model). It has been shown that exploiting resources from multiple auxiliary languages and training the KL-HMM on small amount of adaptation data leads to improvements in non-native speech recognition [12]. In a more recent work, a speaker adaptation technique has been applied to the posterior features to improve the ASR performance in the KL-HMM framework [13].

The existing approaches in the KL-HMM framework have focused on improving the performance of the system through use of small amount of non-native speech data. However, the potential of KL-HMM as a probabilistic lexical modeling approach in handling pronunciation variations without using any adaptation data has not been investigated yet. Such study can be appealing as with growing applications in speech technology, assuming the presence of adaptation data may not be feasible. For example, call routing or tourism information systems need to deal with a variety of pronunciations while there is no adaptation data available.

In this paper, we study the role of probabilistic lexical model in the KL-HMM framework in handling pronunciation variations by comparing it against the deterministic lexical model in the framework of hybrid HMM/ANN. Our experimental studies on French part of MediParl corpus show that the probabilistic lexical model in the KL-HMM framework trained only on native speech data can result in significant improvements compared to the deterministic lexical model in hybrid HMM/ANN approach. Furthermore, our studies also show that the KL-HMM system using context-dependent acoustic units can yield better performance than systems based on speaker adaptation.

The rest of this paper is structured as follows. Section 2

This work was supported by Hasler foundation through the grant AddG2SU and the National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management. The authors would like to thank Ramya Rasipuram for her valuable comments.

provides some background on different HMM-based ASR systems and explains our hypothesis in the present study. Section 3 describes the MediaParl corpus used for the experimental studies and the experimental setup. Sections 4 and 5 provide experimental results and comparison to previous work. Finally, Section 6 brings the conclusion.

2. Background

In a recent study we elucidated that ASR can be viewed as a matching process between acoustic information and lexical information via a latent symbol set [14] as illustrated in Figure 1.

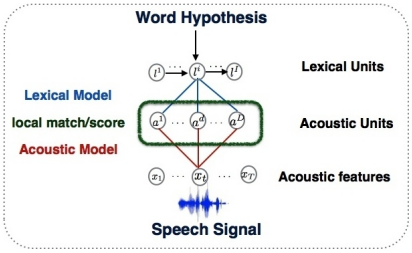

Figure 1: Schematic view of HMM-based ASR approach

In that sense, four fundamental questions can arise:

1. What should the type of latent symbols (acoustic units) be?
2. How to model the relation between acoustic signal and acoustic units (acoustic model)?
3. How to model the relation between the acoustic units and lexical subword units (lexical model)?
4. What should the cost function to locally match the acoustic evidence and lexical evidence be?

Based on the answers to these questions, different systems can be developed. Language modeling and efficient search of output word hypothesis using dynamic programming are common aspects for all these systems. In this paper we are interested in three systems, namely, HMM/GMM, hybrid HMM/ANN and KL-HMM.

For the case of HMM/GMM systems, the aforementioned questions are answered as follows [15]:

1. The acoustic units $\{a^d\}_{d=1}^D$ can be context-independent (CI) or clustered context-dependent (cCD) phones.
2. The relation between the acoustic observation $\mathbf{x}_t$ (e.g. cepstral features) and acoustic units $\{a^d\}_{d=1}^D$ is modeled though GMMs which estimate a likelihood probability vector $\mathbf{v}_t = [v_t^1, \ldots, v_t^d, \ldots, v_t^D]^{\mathrm{T}}$ with $v_t^d = p(\mathbf{x}_t|a^d)$.
3. The relation between the acoustic units $\{a^d\}_{d=1}^D$ and lexical unit l^i, $i \in \{1, \cdots I\}$ is one-to-one deterministic map. i.e, if the lexical unit l^i is deterministically mapped to the acoustic unit a^k, then the relation is modeled through a Kronecker delta distribution $\mathbf{y}_i = [y_i^1, \ldots, y_i^d, \ldots, y_i^D]^{\mathrm{T}} = \delta_{d=k}$ with $y_i^d = p(a^d|l^i)$. The deterministic mapping is obtained either through knowledge (for CI lexical units) or learned during clustering and tying of states (for CD lexical units).
4. The cost function C is then the log of dot product between acoustic model likelihood vector $\mathbf{v}_t$ and lexical model posterior probability vector $\mathbf{y}_i$, i.e. $C = \log \mathbf{y}_i^{\mathrm{T}} \mathbf{v}_t$.

In the hybrid HMM/ANN systems, on the other hand, some of the questions are answered differently. More precisely, the type of acoustic units and lexical modeling are similar to HMM/GMM systems. However, an ANN is used as the acoustic model to estimate posterior probabilities $\{p(a^d|\mathbf{x}_t)\}_{d=1}^D$ and then scale-likelihood vector $\mathbf{v}_t$ with $v_t^d = \frac{p(a^d|\mathbf{x}_t)}{p(a^d)}$ is estimated. The cost function C is then $C = \log \mathbf{y}_i^{\mathrm{T}} \mathbf{v}_t$.

In the case of KL-HMM, the main advantage results from the different approach taken for lexical modeling. More precisely, in this framework the acoustic units and acoustic model can be similar to the aforementioned approaches. However, the relation between the acoustic units $\{a^d\}_{d=1}^D$ and lexical unit l^i is modeled through a categorical distribution $\mathbf{y}_i$ with $y_i^d = p(a^d|l^i)$ as KL-HMM parameters. To learn the KL-HMM parameters, a cost function C is defined based on the KL-divergence between the acoustic unit posterior probability vector $\mathbf{z}_t = [z_t^1, \ldots, z_t^d, \ldots, z_t^D]^{\mathrm{T}}$ with $z_t^d = p(a^d|\mathbf{x}_t)$ (estimated using an ANN or GMM) as the feature observation and the categorical distribution $\mathbf{y}_i$, i.e.

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

As KL-divergence is not a symmetric measure, the local score can be estimated in other ways such as

$$C = S_{RKL}(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} z_t^d \log(\frac{z_t^d}{y_i^d}) \tag{2}$$

or

$$C = S_{SKL}(\mathbf{y}_i, \mathbf{z}_t) = \frac{1}{2}(S_{KL} + S_{RKL}) \tag{3}$$

The parameters $\{\mathbf{y_i}\}_{i=1}^I$ are then estimated using the Viterbi expectation-maximization algorithm which minimizes a cost function based on KL-divergence scores.

These properties of different approaches are summarized in Table 1. The deterministic lexical model in HMM/GMM and HMM/ANN systems imposes certain constraints. For example, the acoustic and lexical units should be of the same type. i.e., if the lexical units are context-independent or context-dependent, then the acoustic units are also constrained to be context-independent or context-dependent respectively. However the probabilistic lexical model in the KL-HMM framework removes such constraints.

Systems	Acoustic unit	Lexical unit	Acoustic Model	Lexical Model	Cost function
HMM/GMM	CI cCD	CI CD	Generative	Deterministic	$\log \mathbf{y}_i^{\mathrm{T}} \mathbf{v}_t$
HMM/ANN	CI cCD	CI CD	Discriminative	Deterministic	$\log \mathbf{y}_i^{\mathrm{T}} \mathbf{v}_t$
KL-HMM	CI/cCD	CI/CD	Discriminative	Probabilistic	$S_{KL}(\mathbf{y}_i, \mathbf{z}_t)$

Table 1: Comparison of properties of different approaches.

Our hypothesis in this paper is that the soft mapping between acoustic and lexical units provided by the probabilistic lexical model, even though learned on native speech, can help in modeling pronunciation variabilities present in non-native speech. In the following sections, we validate our hypothesis by comparing hybrid HMM/ANN as an instance of deterministic lexical modeling approach with KL-HMM as an example of probabilistic lexical modeling approach in non-native speech recognition task.

3. Experimental Setup

In this section, we first describe the MediaParl corpus used in the experiments and then explain the setup of HMM/GMM, hy-

brid HMM/ANN and KL-HMM systems used for the experimental studies.

3.1. Dataset

The experimental studies in this paper are conducted on MediaParl corpus [16]. MediaParl is a bilingual corpus containing recordings of debates in Valais parliament in Switzerland in both Swiss German and Swiss French. Valais is a state in Switzerland including both French and German speakers with variety of accents specially among German speakers. Therefore, MediaParl provides a suitable framework for speech related studies in particular for non-native speech recognition.

In our experiments, the database is partitioned into training, development and test sets according to the structure provided in [16]. Table 2 provides the number of train, dev and test utterances for both French and German along with information about different speakers in the test set. All the speakers in the training and development set are native speakers. In the test set, four speakers are German native speakers and for three speakers, French is the native language. Speakers 109 and 191 are German native speakers who are also fluent in French.

Language	Train Utter.	Dev Utter.	Test Utter.						
			059 (DE-N)	079 (DE-N)	109 (DE-N)	191 (DE-N)	094 (FR-N)	096 (FR-N)	102 (FR-N)
French (FR)	5471	646	31	22	233	165	313	89	72
German (DE)	5955	879	195	698	402	310	72	8	7

Table 2: Data Partitioning in MediaParl Corpus. DE-N and FR-N represent German and French native speakers respectively.

As it can be observed from Table 2, the number of non-native utterances in German part of MediaParl is relatively small (only 87 utterances). Therefore, in this study only French is considered as the target language of interest.

The French dictionary of the MediaParl corpus is provided in SAMPA format with a phone set of size 38 (including sil) and contains all the words in the train, development and test set. The dictionary includes the BDLex pronunciation lexicon[1] and the vocabulary size is 12,362.

For the language model, a bigram model is trained on transcriptions of the training set as well as EuroParl corpus (which consists of about 50 million words for each language).

3.2. Systems

In our experiments, HMM/GMM, hybrid HMM/ANN and KL-HMM systems were studied with the following setups:

HMM/GMM systems: We trained standard cross-word context-dependent HMM/GMM systems with 39 dimensional PLP cepstral features extracted using HTK toolkit [17]. The number of Gaussians and number of clustered states were tuned on the development set. The best performing system had 3928 clustered states with 16 Gaussians per clustered state which was served as baseline for the studies in this paper.

Multilayer perceptrons (MLPs): For the hybrid HMM/ANN and KL-HMM systems we studied two ANNs, more precisely, MLPs to investigate the effect of acoustic units (being context-independent or clustered context-dependent phones) on the ASR performance. As the input to the MLP, PLP cepstral features (of dimension 39) with four frames preceding context and four frames following context were used. The MLPs were trained using Quicknet software [18] with output non-linearity of softmax and minimum cross-entropy error criterion.

We exploited the following MLPs:

- *MLP-CI-38*: a 5-layer MLP classifying context-independent phones as output units (with about 8.8M parameters).
- *MLP-CD-N*: a 5-layer MLP modeling $N = 437$ context-dependent clustered phones as outputs. The acoustic units were derived by clustering context-dependent phones in the HMM/GMM framework using decision tree state tying. The MLP had roughly the same number of parameters as *MLP-CI-38* ($\approx$ 8.8M).

Hybrid HMM/ANN systems: As explained in section 2, the scaled likelihoods $\mathbf{v}_t$ in hybrid HMM/ANN system were estimated by dividing the posterior probabilities $P(a^d|\mathbf{x}_t)$ derived from MLP by the priori probability of acoustic unit $P(a^d)$ estimated from relative frequencies in the training data.

KL-HMM systems: The KL-HMM systems used acoustic units posterior probabilities as feature observations and modeled either context-independent or context-dependent (tri) phones as lexical units. The KL-HMM parameters were trained by minimizing the cost functions based on local scores KL, SKL and RKL (as described in Section 2) and the local score with minimum KL-divergence on training data was used. The two KL-HMM systems using context-independent and clustered context-dependent phones used SKL and RKL as the local score respectively. In order to tie the KL-HMM (lexical) states KL-divergence based decision tree state tying method proposed in [19] was applied.

4. Results and Analysis

Table 3 presents the results in terms of word error rate (WER) in hybrid HMM/ANN and KL-HMM systems using context-independent acoustic units. It can be observed that the KL-HMM system outperforms hybrid HMM/ANN for both native and non-native speech. For the case of native speech, the probabilistic lexical model can help in capturing the pronunciation variations present in debates as type of spontaneous speech. In addition, pronunciation variabilities can occur as a result of differences between the Swiss French and French accent [2]. The soft mapping between acoustic and lexical units in the KL-HMM framework can help in handling these pronunciation variations. For the case of non-native speech, as hypothesized, the pronunciation variation information captured by the probabilistic lexical model from the native speakers' speech is helpful for non-native speech recognition.

	Native	*Non-native*	*Overall*	*Lexical units*
Hyb-MLP-CI-38	*26.7*	*40.0*	*33.1*	*CI*
KL-HMM-MLP–CI-38	*23.3*	*37.3*	*30.0*	*CI*

Table 3: Experimental results in hybrid HMM/ANN and KL-HMM using CI acoustic units

Table 4 presents the results using clustered context-dependent acoustic units (of size 437)[3]. As in the KL-HMM

[1] http://www.irit.fr/ Martine.deCalmes/IHMPT/ress_ling.v1/rbdlex_en.php

[2] For instance, the closed vowel /o/ in Valaisan accent is pronounced as open vowel /O/ in words like "eau" [20]

[3] HMM/GMM results are presented here as the baseline to which other approaches are compared.

framework, the acoustic and lexical units do not require to be of the same type (CI or CD), we have also presented the results using context-independent acoustic units and context-dependent lexical units. Similar to the previous scenario using context-independent acoustic units, it can be observed that the KL-HMM approach outperforms the hybrid HMM/ANN system which is indicative of the role of the probabilistic lexical model in capturing pronunciation variabilities. Similar to our previous work [14], it is interesting to note that *KL-HMM-MLP-CI-38* system using context-dependent lexical units achieves comparable results to the *Hyb-MLP-CD-437* system.

System	*Native*	*Non-native*	*Overall*	*Lexical units*
HMM/GMM-CD-3928	*19.8*	*34.4*	*26.8*	*CD*
Hyb-MLP-CD-437	*19.4*	*32.0*	*25.5*	*CD*
KL-HMM-MLP-CI-38	*20.0*	*32.3*	*25.9*	*CD*
KL-HMM-MLP-CD-437	*16.0*	*29.1*	*22.3*	*CD*

Table 4: Experimental results in hybrid HMM/ANN and KL-HMM using CD acoustic and lexical units

To further analyze the performance of hybrid HMM/ANN and KL-HMM systems presented in Table 4, we have depicted the results in terms of WER per speaker. It can be observed that for almost all the speakers, the KL-HMM approach outperforms the hybrid HMM/ANN approach. The results are consistent for both native and non-native speakers.

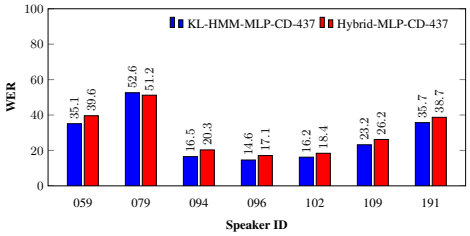

Figure 2: Comparison of hybrid HMM/ANN and KL-HMM ASR performance per speaker

In our previous study [14], it was observed that the hybrid HMM/ANN system requires slightly more number of acoustic units than the KL-HMM system. So we conducted experiments by increasing the number of acoustic units. More precisely, we used *MLP-CD-N* with $N \in \{817, 1084\}$ with roughly same number of parameters as before ($\approx$ 8.8M) to classify the acoustic units. The different number of acoustic units were derived by adjusting the log-likelihood difference during the decision tree state tying in the HMM/GMM framework. Figure 3 presents the results in terms of WER with different number of acoustic units for both native and non-native speech. It can be observed from Figure 3 that while increasing the number of clustered context-dependent acoustic units (from 437 to 817) leads to slight improvement in native speech recognition, it slightly hurts the non-native speech recognition. This indicates that increasing the number of acoustic units does not necessarily lead to improvement in the overall results.

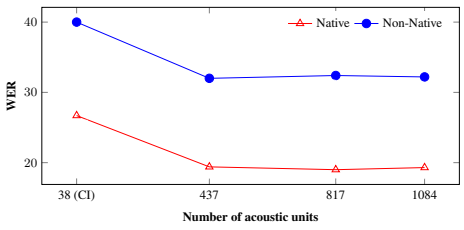

Figure 3: Effect of number of acoustic units on native and non-native speech recognition in hybrid HMM/ANN framework

5. Comparison to Previous Work

In this study, we analyzed the potential of the KL-HMM approach in improving non-native speech recognition without using any adaptation data. In this section, we compare our results with a previous study on the MediaParl corpus using speaker adaptation techniques [13]. Figure 4 shows the results in terms of WER, when using MLLR, speaker adaptive KL-HMM and *KL-HMM-MLP-CD-437*[4]. It can be observed that *KL-HMM-MLP-CD-437* system can perform better than both MLLR and speaker adaptive KL-HMM approaches for both native and non-native speech. While the speaker adaptive KL-HMM technique leads to improvement in non-native speech recognition compared to MLLR, it does not help in improving the recognition performance for native speech. However, *KL-HMM-MLP-CD-437* is useful for both native and non-native speech without using any adaptation data. We should note that in [13], a 3-layer MLP classifying context-independent phones was used as the acoustic model. It would be interesting to see whether speaker adaptive KL-HMM system yields any further improvement for context-dependent acoustic units with the same amount of adaptation data.

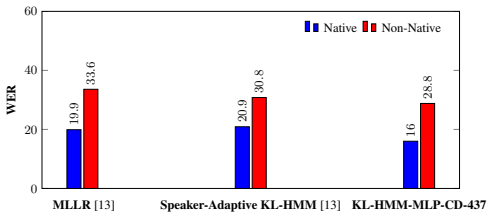

Figure 4: Comparison of different approaches

6. Discussion and Conclusion

In this paper, we studied the role of the probabilistic lexical model in the KL-HMM framework in improving non-native speech recognition without using any adaptation data. Our experimental studies showed that by moving from the deterministic lexical modeling approach of hybrid HMM/ANN to the probabilistic lexical modeling approach of KL-HMM, notable improvements in non-native speech recognition can be achieved. The observations in this study are inline with the studies such as probabilistic classification of HMM states (PC-HMM) [21] which has shown to be successful in handling pronunciation variations in spontaneous speech [22]. The PC-HMM approach has been argued to be similar to the KL-HMM approach in the sense that both can be viewed as probabilistic lexical modeling approaches [23]. In the future work, we aim to investigate the effect of cross-lingual knowledge transfer (for example, by incorporation of acoustic information from the native language of the speaker) on the non-native speech recognition within the KL-HMM framework.

[4] In [13], speaker 059 was omitted due to lack of sufficient adaptation data. Therefore, for the sake of comparability we also report the results without speaker 059.

7. References

[1] D. Van Compernolle, "Recognizing speech of goats, wolves, sheep and ... non-natives," *Speech Communication*, vol. 35, no. 1, pp. 71–79, 2001.

[2] J. Segura, T. Ehrette, A. Potamianos *et al.*, "The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication," *Online. http://www. hiwire. org*, 2007.

[3] G. Bouselmi, D. Fohr, I. Illina *et al.*, "Multi-accent and accent-independent non-native speech recognition." in *Proceedings of Interspeech*, 2008, pp. 2703–2706.

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

[5] V. Fischer, E. Janke, and S. Kunzmann, "Likelihood combination and recognition output voting for the decoding of non-native speech with multilingual HMMs." in *Proceedings of Interspeech*, 2002.

[6] G. Bouselmi, D. Fohr, I. Illina, J.-P. Haton *et al.*, "Multilingual non-native speech recognition using phonetic confusion-based acoustic model modification and graphemic constraints," in *Proceedings of ICSLP*, 2006.

[7] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *in Proceedings of ICASSP*, vol. 1. IEEE, 2003, pp. I–540.

[8] U. Nallasamy, F. Metze, and T. Schultz, "Enhanced polyphone decision tree adaptation for accented speech recognition." ISCA, 2012. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2012.html#NallasamyMS12

[9] G. Aradilla, J. Vepa, and H. Bourlard, "An acoustic model based on Kullback-Leibler divergence for posterior features," in *Proceedings of ICASSP*, 2007, pp. IV–657 – IV–660.

[10] G. Aradilla, H. Bourlard, and M. M. Doss, "Using KL-based acoustic models in a large vocabulary recognition task," in *Proceedings of Interspeech*, 2008, pp. 928–931.

[11] R. Rasipuram and M. Magimai-Doss, "Improving grapheme-based ASR by probabilistic lexical modeling approach," in *Proceedings of Interspeech*, 2013.

[12] D. Imseng, R. Rasipuram, and M. Magimai-Doss, "Fast and flexible kullback-leibler divergence based acoustic modeling for non-native speech recognition," in *Proceedings of ASRU*, Dec. 2011, pp. 348–353.

[13] D. Imseng and H. Bourlard, "Speaker adaptive kullback-leibler divergence based hidden markov models," in *Proceedings of ICASSP*, 2013.

[14] M. Razavi, R. Rasipuram, and M. Magimai-Doss, "On modeling context-dependent clustered states: Comparing HMM/GMM, Hybrid HMM/ANN and KL-HMM Approaches," *To appear in Proceedings of ICASSP*, 2014.

[15] R. Rasipuram and M. Magimai.-Doss, "Acoustic and lexical resource constrained ASR using language-independent acoustic model and language-dependent probabilistic lexical model," Idiap, Idiap-RR Idiap-RR-02-2014, 3 2014.

[16] D. Imseng *et al.*, "Mediaparl: Bilingual mixed language accented speech database," in *Proceedings of IEEE Workshop on SLT*, Dec. 2012, pp. 263–268.

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

[18] D. Johnson *et al.*, "ICSI Quicknet Software Package," http://www.icsi.berkeley.edu/Speech/qn.html, 2004.

[19] D. Imseng *et al.*, "Comparing different acoustic modeling techniques for multilingual boosting," in *Proceedings of Interspeech*, Sep. 2012.

[20] H. Caesar, "Integrating language identification to improve multilingual speech recognition," Idiap, Idiap-RR Idiap-RR-24-2012, 7 2012.

[21] X. Luo and F. Jelinek, "Probabilistic classification of hmm states for large vocabulary continuous speech recognition," in *Proceedings of ICASSP*, vol. 1. IEEE, 1999, pp. 353–356.

[22] M. Saraclar, H. Nock, and S. Khudanpur, "Pronunciation modeling by sharing gaussian densities across phonetic models," *Computer Speech & Language*, vol. 14, no. 2, pp. 137–160, 2000.

[23] R. Rasipuram and M. Magimai-Doss, "Probabilistic lexical modeling and grapheme-based automatic speech recognition," Idiap, Idiap-RR Idiap-RR-15-2013, 4 2013.