IDIAP RESEARCH REPORT

KL-HMM AND PROBABILISTIC LEXICAL MODELING

Ramya Rasipuram Mathew Magimai.-Doss

Idiap-RR-04-2013

FEBRUARY 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

KL-HMM and Probabilistic Lexical Modeling

Ramya Rasipuram and Mathew Magimai.-Doss

Abstract

Kullback-Leibler divergence based hidden Markov model (KL-HMM) is an approach where a posteriori probabilities of phonemes estimated by artificial neural networks (ANN) are modeled directly as feature observation. In this paper, we show the relation between standard HMM-based automatic speech recognition (ASR) approach and KL-HMM approach. More specifically, we show that KL-HMM is a probabilistic lexical modeling approach which is applicable to both HMM/GMM ASR system and hybrid HMM/ANN ASR system. Through experimental studies on DARPA Resource Management task, we show that KL-HMM approach can improve over state-of-the-art ASR system.

Index Terms

Automatic speech recognition, hidden Markov model, Lexical modeling, Posterior features, Kullback-Leibler divergence based HMM

I. Introduction

In standard hidden Markov model (HMM) based automatic speech recognition (ASR) systems, the feature observations are typically short-term spectral based features such as, mel frequency cepstral coefficients (MFCCs), perceptual linear prediction (PLP) cepstral coefficients and the emission distribution is modeled by either Gaussian mixture models (GMMs) or artificial neural networks (ANNs) [1], [2]. The system using GMMs is referred to as HMM/GMM system and the system using ANNs is referred to as hybrid HMM/ANN system.

In more recent works, different approaches have been proposed for modeling the output of the ANN i.e. a posteriori probabilities of acoustic classes (e.g., phonemes) as feature observation such as, Tandem approach [3], Kullback-Leibler divergence based HMM approach [4], [5], Dirichlet mixture model approach [6]. In Tandem approach, the a posteriori probabilities are transformed, more precisely whitened and decorrelated, and used as feature input for HMM/GMM system. While, in KL-HMM approach and Dirichlet mixture model approach the a posteriori probabilities of phone classes are directly used as feature observation and modeled by HMM.

The focus of this paper is on KL-HMM approach which until now has largely been investigated from *posterior feature* modeling perspective (Section II). In this paper, we first elucidate that standard HMM-based ASR system

R. Rasipuram is with Idiap Research Institute, Martigny, Switzerland and Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. M. Magimai.-Doss is with Idiap Research Institute, Martigny, Switzerland. Both authors have equally contributed to this work. 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)

uses *deterministic* lexical model (Section III). We then show that KL-HMM is a probabilistic lexical modeling approach, where the local emission score is estimated by matching lexical evidence and acoustic evidence (Section IV). While doing so, we also introduce a new approach, referred to as scalar product HMM (SP-HMM), and show its link to tied posterior approach [7]. Finally, we present experimental studies in the framework of HMM/GMM system which shows that KL-HMM approach and SP-HMM approach can yield improvements over state-of-the-art ASR system (Section V).

II. Kullback-Leibler Divergence Based HMM

In KL-HMM approach [5], posterior probabilities of phonemes, also referred to as *posterior feature*, estimated by ANN is used as feature observation. Let $\mathbf{z}_t = [z_t^1, \cdots, z_t^D]^{\mathrm{T}} = P(p_1|\mathbf{x}_t), \cdots, P(p_D|\mathbf{x}_t)]^{\mathrm{T}}$ denote the posterior feature vector estimate at time frame t, 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 $i \in \{1, \cdots I\}$ in the KL-HMM system is parameterized by a categorical distribution $\mathbf{y}_i = [y_i^1, \cdots, y_i^D]^{\mathrm{T}}$. The local score at each HMM state is estimated as Kullback-Leibler (KL) divergence between $\mathbf{y}_i$ and $\mathbf{z}_t$, i.e.,

$$KL = \sum_{d=1}^{D} y_i^d \log(\frac{y_i^d}{z_t^d}) \tag{1}$$

In this case, $\mathbf{y}_i$ serves as the reference distribution and $\mathbf{z}_t$ serves as the test distribution. KL-divergence being an asymmetric measure, there are also other ways to estimate the local score,

1) Reverse KL-divergence (RKL):

$$RKL = \sum_{d=1}^{D} z_t^d \log(\frac{z_t^d}{y_i^d}) \tag{2}$$

2) Symmetric KL-divergence (SKL):

$$SKL = \frac{1}{2} \cdot [KL + RKL] \tag{3}$$

The HMM state parameters $\{\mathbf{y}_i\}_{i=1}^{I}$ are estimated by using Viterbi expectation maximization algorithm which minimizes a cost function based on one of the above local scores. During testing, decoding is performed using standard Viterbi decoder. For more details the reader is referred to [5], [4].

III. Standard HMM based ASR

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

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

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

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

where $\mathcal{Q}$ denotes set of all possible HMM state sequences, $Q = \{q_1, \cdots q_t, \cdots q_T\}$ denotes a sequence of HMM states, T denotes number of frames, and $\Theta = \{\Theta_A, \Theta_L\}$ denotes the set of parameters, more specifically acoustic model and lexical model parameters set Θ_A and language model parameters Θ_L. 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*.

In HMM/GMM system, the emission likelihood $p(\mathbf{x}_t|q_t, \Theta_A)$ is estimated using GMMs. In hybrid HMM/ANN system, the emission likelihood is estimated using ANN. More precisely, the ANN estimates a posteriori probability of state $P(q_t|\mathbf{x}_t, \Theta_A)$ which is then converted into scaled-likelihood $p_{sl}(\mathbf{x}_t|q_t, \Theta_A)$,

$$p_{sl}(\mathbf{x}_t|q_t, \Theta_A) = \frac{p(\mathbf{x}_t|q_t, \Theta_A)}{p(\mathbf{x}_t|\Theta_A)} = \frac{P(q_t|\mathbf{x}_t, \Theta_A)}{P(q_t|\Theta_A)} \tag{7}$$

and used as local emission score. Though the literature is dominated by the approach of using likelihood as local emission score, in theory, HMMs can be also trained and decoded using $P(q_t|\mathbf{x}_t, \Theta_A)$ as emission probabilities [2]. We differentiate between these two approaches by referring to as *likelihood based approach* and *posterior based approach*, respectively.

In practice, in HMM-based ASR system there are two kinds of HMM 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 context-dependent subword unit based ASR system, the clustered states are the acoustic states and the lexical states are the states of context-dependent subword model, e.g. /k/-/ae/+/t/.
- in hybrid HMM/ANN system, typically during the training phase the ANN is trained to classify K context-independent phonemes, and during the decoding phase a minimum duration constraint is applied for each phoneme [2]. In this case, there are K acoustic states and $n \cdot K$ lexical states, where n is the minimum duration.

Let $\Theta_A = \{\theta_a, \theta_l\}$, where θ_a denotes the parameters of acoustic model and θ_l denotes the parameters of lexical model. The acoustic model parameters in the case of GMMs are the Gaussian means, variance and weights of each acoustic state. In the case of ANNs, the acoustic model parameters are the weights and biases. In standard HMM-based ASR systems, the relationship between lexical states and acoustic states is one-to-one, i.e. *deterministic*. Thus, θ_l consists of the set of subword units, pronunciation models of words and a table that maps lexical states (corresponding to the subword units) onto acoustic states.

During both training phase and decoding phase, the emission likelihood is estimated by matching the acoustic state evidence with the lexical model. This is trivial as the relationship between the acoustic states and the lexical states is one-to-one. More precisely, given the one-to-one relationship, $p(\mathbf{x}_t|q_t^{lex} = i, \Theta_A) = p(\mathbf{x}_t|q_t^{aco} = d, \theta_a)$ in the case of likelihood based approach and $P(q_t^{lex} = i|\mathbf{x}_t, \Theta_A) = P(q_t^{aco} = d|\mathbf{x}_t, \theta_a)$ in the case of posterior based approach, where $i \in \{1, \cdots I\}$ here denotes a lexical state, $d \in \{1, \cdots D\}$ here denotes an acoustic state, I here denotes the number of lexical states and D here denotes the number of acoustic states. Here after, for simplicity we will drop the notations for parameters.

IV. Relation between KL-HMM and Standard HMM-based ASR

A strict one-to-one relationship between lexical states and acoustic states makes the ASR system overly rely on prior knowledge resources in the lexical model, namely subword units and pronunciation models. This can lead to mismatch between lexical model and acoustic model (e.g., pronunciation variation), which in turn can affect ASR performance. One way to handle this issue is to model the soft/probabilistic relationship between lexical states and acoustic states.

A. Probabilistic Lexical Modeling and KL-HMM

The probabilistic relationship between lexical states and acoustic states can be modeled as $P(q_t^{aco} = d|q_t^{lex} = i), \forall i \in \{1, \cdots I\}, d \in \{1, \cdots D\}$. Let $\mathbf{y}_i = [P(q_t^{aco} = 1|q_t^{lex} = i) \cdots P(q_t^{aco} = D|q_t^{lex} = i)]^{\mathrm{T}}$ be the vector representing the relationship between lexical state i and the D acoustic states. Having said that, there are two main questions, namely

1) How to estimate lexical evidence $P(q_t^{aco} = d|q_t^{lex} = i)$ or simply, $\mathbf{y}_i$?
2) How to integrate/match lexical evidence with acoustic evidence, which in the case of likelihood based approach is $p(\mathbf{x}_t|q_t^{aco} = d)$ and in the case of posterior based approach is $P(q_t^{aco} = d|\mathbf{x}_t)$?

KL-HMM is a posterior based probabilistic lexical modeling approach, where

1) first, an acoustic state posterior probability estimator is trained with deterministic lexical model as done in standard HMM-based ASR system.
2) then, a second HMM is trained by using acoustic state posterior probability estimates $\mathbf{z}_t = [P(q_t^{aco} = 1|\mathbf{x}_t) \cdots P(q_t^{aco} = D|\mathbf{x}_t)]^{\mathrm{T}}$ as feature observations. The states of the second HMM represent the lexical states, which are parametrized by $\{\mathbf{y}_i\}_{i=1}^I$. The parameters $\{\mathbf{y}_i\}_{i=1}^I$ are trained by optimizing a cost function based on KL-divergence as mentioned earlier in Section II.

In theory, KL-divergence can be linked to hypothesis testing [8], [9]. So, KL-HMM can be seen as a probabilistic lexical modeling approach, where the local emission score is estimated by discriminatively matching the lexical evidence and the acoustic evidence.

There are also other ways to achieve probabilistic lexical modeling. For instance,

- in likelihood based approach, this can be achieved by modeling $p(\mathbf{x}_t|q_t^{lex} = i)$ as

$$= \sum_{d=1}^{D} p(\mathbf{x}_t, q_t^{aco} = d|q_t^{lex} = i), \quad \forall i \in \{1, \cdots I\} \tag{8}$$

$$= \sum_{d=1}^{D} p(\mathbf{x}_t|q_t^{aco} = d, q_t^{lex} = i) \cdot P(q_t^{aco} = d|q_t^{lex} = i) \tag{9}$$

$$\approx \sum_{d=1}^{D} p(\mathbf{x}_t|q_t^{aco} = d) \cdot P(q_t^{aco} = d|q_t^{lex} = i) \tag{10}$$

Eqn. (10) assumes that $\mathbf{x}_t \perp q_t^{lex} \mid q_t^{aco}$. Given a trained acoustic state likelihood estimator, $\mathbf{y}_i$ can be estimated using a cost function based on Eqn. (10). In the case where the acoustic states are modeled by ANN $p(\mathbf{x}_t|q_t^{aco} =$

$d)$ is replaced by $p_{sl}(\mathbf{x}_t|q_t^{aco} = d)$. It is interesting to note that, then, the likelihood based approach is exactly same as the tied posterior approach proposed in [7].

- in posterior based approach, yet another way is to model $P(q_t^{lex} = i|\mathbf{x}_t)$ as

$$\sum_{d=1}^{D} P(q_t^{aco} = d|q_t^{lex} = i) \cdot P(q_t^{aco} = d|\mathbf{x}_t) = \mathbf{y}_i^{\mathrm{T}}\mathbf{z}_t \tag{11}$$

  Given a trained acoustic model, $\mathbf{y}_i$ can be estimated by training a second HMM similar to KL-HMM, where $\mathbf{z}_t$ is used as feature observation, the states of the HMM are parametrized by $\mathbf{y}_i$, and a cost function based on Eqn. (11), i.e. dot/scalar product of posterior probability vectors is used. We refer to it as scalar product HMM (SP-HMM). It can be noticed that tied posterior approach [7] reduces to SP-HMM approach, when equal prior for acoustic states is assumed.

In a recent work, a template based ASR approach using posterior features (estimated by ANN or GMM) has been proposed [4], [10]. This approach can be linked to posterior based probabilistic lexical modeling approach. In this template based ASR system, first an ANN or GMM needs to be trained which can be seen as acoustic state posterior probability estimator. Then, reference templates (sequence of posterior features) are obtained and stored. Each time frame in a reference template can be interpreted as an abstract lexical state, and the posterior feature vector at each time frame in the reference template (though estimated using acoustics) can be seen as probabilistic relationship between abstract lexical state and acoustic state. During testing, the test template is matched with the reference templates. In the template based system, in addition to KL-divergence and scalar product other local matching functions such as, Bhattacharya distance, cosine distance have been investigated, and have been found to yield competitive systems [10]. This suggests that in posterior based probabilistic lexical modeling approach there are other local matching functions that could also be investigated.

It is worth mentioning that the approach of modeling probabilistic relationship between lexical states and acoustic states is ideologically similar to the hidden model sequence HMM (HMS-HMM) approach proposed in [11]. However, HMS-HMM approach is implementation wise very different. Also, it was particularly developed for context-dependent subword unit (phone) modeling. The approaches described in this section does not put any such limitation.

Finally, when compared to standard approach of using deterministic lexical model, it is important to note that probabilistic lexical modeling does not changes the acoustic model complexity. It only changes the lexical model complexity, where θ_l now consists of subword unit set, pronunciation model of words and $\{\mathbf{y}_i\}_{i=1}^{I}$.

B. Interpretation of Previous Work on KL-HMM

The above described relation to probabilistic lexical modeling helps us to better understand the potentials of KL-HMM approach and elucidate previous work. KL-HMM has been investigated for

1) development of context-dependent subword unit based ASR system without explicitly modeling the relationship between context-dependent subword unit and acoustic observations [5], [12], [13], [14]. Here, the

acoustic states are the context-independent phonemes and the lexical states are context-dependent subword units.

2) use of graphemes as subwords [12], [13]. In this case, the acoustic states represent context-independent phonemes and the lexical states represent context-independent or -dependent graphemes. Here, $\mathbf{y}_i$ captures the probabilistic relationship between graphemes and phonemes.
3) non-native speech recognition and rapid development of ASR system for new language using multilingual phonemes and auxiliary/out-of-domain data [12], [14]. In these works, the acoustic states are context-independent multilingual phonemes and the lexical states are context-dependent monolingual phonemes or graphemes. The acoustic states probability $\mathbf{z}_t$ estimator is trained on auxiliary data and $\mathbf{y}_i$ is trained on in-domain data.

V. Experiments and Results

In the past, KL-HMM approach has been investigated in the hybrid HMM/ANN framework, where the acoustic states modeled by ANN are context-independent phonemes [5], [12], [14]. In these studies, it has been often observed that KL-HMM approach performs better than state-of-the-art HMM/GMM system only when very little data is available, e.g. see [14]. In this section, we present ASR studies which show that KL-HMM or SP-HMM approach is equally applicable to state-of-the-art HMM/GMM framework, and can improve over standard HMM/GMM system.

We present ASR studies on DARPA Resource Management task [15]. We use the setup described in [11]. The only difference is that we use UNISYN dictionary [16] and except for 35 words rest of the words have single pronunciation. We compare the standard HMM/GMM approach, where the relationship between lexical and acoustic states is deterministic, with probabilistic lexical modeling approach. More precisely,

- Deterministic lexical model based system: we train and test a crossword triphone based HMM/GMM system with state tying using HTK, where each triphone is modeled by 3 states. The acoustic feature $\mathbf{x}_t$ is 39 dimensional PLP cepstral feature. The number of clustered/acoustic states $D = 1611$.
- Probabilistic lexical model based system: Given the clustered/acoustic state models of the deterministic lexical model system, the training phase involves estimation of
  1) acoustic state posterior feature $\mathbf{z}_t = [z_t^1 \cdots z_t^d \cdots z_t^D]^{\mathrm{T}}$ assuming equal priors for the acoustic states,
  $$z_t^d = P(q_t^{aco} = d|\mathbf{x}_t) = \frac{p(\mathbf{x}_t|q_t^{aco} = d)}{\sum_{j=1}^{D} p(\mathbf{x}_t|q_t^{aco} = j)} \tag{12}$$
  where $p(\mathbf{x}_t|q_t^{aco} = d)$ is the likelihood of acoustic state d.
  2) and then, $\mathbf{y}_i$ by SP-HMM approach or KL-HMM approach.

  We train and test word internal triphone system (without state tying) and cross word triphone system (with state tying), where, in both systems similar to HMM/GMM system each triphone is modeled by 3 states. The state tying is performed using the approach proposed in [14] with state occupancy count of one. In order to compare across different estimates of $\mathbf{y}_i$ and to limit the number of experiments, all the systems are decoded with local emission score based on Eqn. (11).

Table I presents the ASR performance of systems based on deterministic lexical model and probabilistic lexical model in terms of word error rate (WER). The performance of deterministic lexical model based system is comparable to 4.1% WER reported in [11]. It can be observed that by just modeling word internal triphones, the KL-HMM approach (with local score RKL) and the SP-HMM approach of estimating $\mathbf{y}_i$ yields improvement over deterministic lexical model based system. With crossword modeling, the KL-HMM approach with local score SKL also improves over the deterministic lexical model based system, while the KL-HMM approach with local score RKL performs significantly better than the deterministic lexical model based system. In the case of SP-HMM, cross word system is not reported as we could not apply the state tying approach.

TABLE I

WER FOR DIFFERENT SYSTEMS. WI DENOTES WORD INTERNAL TRIPHONE MODELING, XWRD DENOTES CROSS WORD TRIPHONE MODELING, AND N.A DENOTES NOT APPLICABLE. THE LOCAL SCORE FOR KL-HMM APPROACH IS MENTIONED BETWEEN PARENTHESIS.

Lexical model	$\mathbf{y}_i$ estimation	WI	XWRD
Deterministic	-	n.a.	4.2
Probabilistic	KL-HMM (KL)	7.1	6.6
Probabilistic	KL-HMM (RKL)	3.8	**2.9**
Probabilistic	KL-HMM (SKL)	4.6	3.7
Probabilistic	SP-HMM	3.9	-

In our recent work on grapheme-based ASR using KL-HMM approach [17], we have observed that local score RKL models well one-to-many relation between lexical states and acoustic states followed by local score SKL, while local score KL models well one-to-one relation between lexical states and acoustic states. The general idea of probabilistic lexical modeling approach is that the relation between lexical states and acoustic states may not be one-to-one but one-to-many. This aspect can be observed by comparing the performance across different local scores in the KL-HMM approach. KL-HMM approach with local score KL yields significantly poor performance compared to local scores RKL and SKL, as it may not be able to capture the one-to-many relationship. The best performance of 2.9% WER compares favourably to 3.1% WER obtained by HMS-HMM approach [11], which is ideologically similar. Furthermore, to the best of our knowledge, without any acoustic model adaptation, 2.9% is the lowest WER to be reported on RM task.

VI. DISCUSSION AND CONCLUSION

In standard HMM-based ASR system, the relation between lexical states and acoustic states is deterministic. In this paper, we showed that approaches such as, KL-HMM, SP-HMM, tied posterior are probabilistic lexical modeling approaches, where the probabilistic relation between lexical states and acoustic states is learned by training a second HMM which uses a posteriori probability or likelihood of acoustic states as feature observation. Furthermore, we showed how KL-HMM approach and SP-HMM approach can be applied to state-of-the-art HMM/GMM system to improve ASR performance.

Probabilistic lexical modeling approach, at the cost of increasing lexical model complexity, can help in handling pronunciation variation [13], modeling longer subword unit context without explicitly modeling the acoustic relationship [7], [5], modeling alternate subword units, such as graphemess [12], [17], and effective use of auxiliary resources (both acoustic and linguistic) [13], [14].

In this paper, we investigated the application of KL-HMM approach and SP-HMM approach to HMM/GMM system with context-dependent clustered acoustic states. When comparing this study to previous studies on KL-HMM, it can be observed that we increased the complexity of the acoustic model (going from context-independent phoneme states to clustered context-dependent phoneme states) and the complexity of lexical model (increasing the posterior feature dimension). However, it may be possible to build competitive ASR systems by keeping the acoustic model complexity low as done in previous studies, and only increase the lexical model complexity. It could be, for instance, done in an hierarchical framework where,

1) ANN or GMM is trained for estimating context-independent phoneme acoustic state posterior features
2) similar to previous studies [13], [14], [12], [17], model the context-independent phoneme acoustic state posterior features by a tied state context-dependent phoneme based ASR system using KL-HMM approach or SP-HMM approach
3) estimate a new set of posterior features corresponding to the clustered states
4) finally, train a second context-dependent phoneme based ASR system using KL-HMM approach or SP-HMM approach, where the feature observations now are the clustered state posterior features

Such a framework is not only interesting from general ASR perspective, but also for low acoustic resource scenarios, and multilingual ASR scenarios (where multilingual data is used to train an acoustic model that is shared across different languages), as context-independent phonemes could be considered more language independent than context-dependent phonemes.

In our future work, we will investigate the above described approach while extending our current investigations to conversational speech and grapheme-based ASR.

ACKNOWLEDGMENT

The authors would like to thank their current and past colleagues, especially Guillermo Aradilla, David Imseng, Dr. John Dines, and Prof. Hervé Bourlard, for their critical inputs and suggestions.

REFERENCES

[1] L. R. Rabiner and H. W. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs New Jersey: Prentice Hall, 1993.

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

[3] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. of ICASSP*, 2000.

[4] G. Aradilla, "Acoustic models for posterior features in speech recognition," Ph.D. dissertation, EPFL, Switzerland, 2008.

[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] B. Vardarajan, G. S. V. S. Sivaram, and S. Khudanpur, "Dirichlet Mixture Models of neural net posteriors for HMM-based speech recognition," in *Proc. of ICASSP*, 2011.

[7] J. Rottland and G. Rigoll, "Tied posteriors: An approach for effective introduction of context dependency in hybrid NN/HMM LVCSR," in *Proc. of ICASSP*, 2000.

[8] R. E. Blahut, "Hypothesis testing and information theory," *IEEE Trans. on Information Theory*, vol. IT-20, no. 4, 1974.

[9] S. Eguchi and J. Copas, "Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma," *Journal of Multivariate Analysis*, vol. 97, no. 9, 2006.

[10] S. Soldo *et al.*, "Posterior features for template-based ASR," in *Proc. of ICASSP*, 2011.

[11] T. Hain, "Hidden model sequence models for automatic speech recognition," PhD Dissertation, University of Cambridge, 2001.

[12] M. Magimai.-Doss *et al.*, "Grapheme-based automatic speech recognition using KL-HMM," in *Proc. of Interspeech*, 2011.

[13] 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 ASRU*, 2011, pp. 348–353.

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

[15] P. Price *et al.*, "A database of continuous speech recognition in a 1000 word domain," in *Proc. of ICASSP*, 1988.

[16] S. Fitt, "Documentation and User Guide to UNISYN Lexicon and Postlexical Rules," CSTR, University of Edinburgh, Tech. Rep., 2000.

[17] R. Rasipuram and M. Magimai.-Doss, "Modeling graphemes in the framework of Kullback-Leibler divergence based acoustic modeling," *Submitted to Speech Communication*.