ON MODELING CONTEXT-DEPENDENT CLUSTERED STATES: COMPARING HMM/GMM, HYBRID HMM/ANN AND KL-HMM APPROACHES

Marzieh Razavi[1,2], *Ramya Rasipuram*[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
{mrazavi, rramya, mathew}@idiap.ch

ABSTRACT

Deep architectures have recently been explored in hybrid hidden Markov model/artificial neural network (HMM/ANN) framework where the ANN outputs are usually the clustered states of context-dependent phones derived from the best performing HMM/Gaussian mixture model (GMM) system. We can view a hybrid HMM/ANN system as a special case of recently proposed Kullback-Leibler divergence based hidden Markov model (KL-HMM) approach. In KL-HMM approach a probabilistic relationship between the ANN outputs and the context-dependent HMM states is modeled. In this paper, we show that in KL-HMM framework we may not require as many clustered states as the best HMM/GMM system in the ANN output layer. Our experimental results on German part of MediaParl database show that KL-HMM system achieves better performance compared to hybrid HMM/ANN and HMM/GMM systems with much fewer number of clustered states than is required for HMM/GMM system. The reduction in number of clustered states has broader implications on model complexity and data sparsity issues.

Index Terms— HMM/GMM, hybrid HMM/ANN, Kullback-Leibler divergence based HMM, context-dependent subword units, non-native speech recognition

1. INTRODUCTION

In conventional HMM-based automatic speech recognition (ASR) systems, the emission distribution is modeled either using Gaussian mixture models (GMM) which leads to so-called HMM/GMM systems or artificial neural network (ANN) which leads to hybrid HMM/ANN systems [1]. Traditionally, in hybrid HMM/ANN systems, the outputs of ANN represented context-independent phones and each ANN output was related to a unique HMM state. This limited the HMM topology to context-independent subword units.

In recent years, improvements in computer hardware and machine learning techniques have enabled using efficient methods for training ANNs with more hidden layers and output units. The large number of ANN output units allows to train hybrid HMM/ANN systems that can deal with a possibly large number of HMM states needed for modeling context-dependent phones. More specifically, the ANN outputs are defined as the clustered context-dependent phones derived from the HMM/GMM system [2, 3].

Several studies have investigated the effect of number of hidden units and hidden layers of ANN on the ASR performance in hybrid HMM/ANN framework [2, 4]. In addition to number of layers and size of layers, the choice of ANN output units and the way output units are defined is also crucial. Most often, the output units of ANN are clustered states of context-dependent phones derived from the best performing HMM/GMM system [2, 3]. In [5], the number of ANN outputs was tuned on the development set for hybrid and Tandem systems.

In this paper, we first elucidate that hybrid HMM/ANN approach is a special case of recently proposed Kullback-Leibler divergence based hidden Markov model (KL-HMM) approach (Section 2). More specifically, in hybrid HMM/ANN approach the relationship between ANN outputs and HMM states is deterministic (deterministic lexical model), whereas in KL-HMM approach the relationship between ANN outputs and HMM states is probabilistic (probabilistic lexical model). Then, we study the effect of the number of ANN outputs on ASR performance for KL-HMM approach and compare it with hybrid HMM/ANN and HMM/GMM approaches. We hypothesize that with probabilistic lexical modeling, KL-HMM approach can result in better systems with fewer number of ANN outputs compared to number of clustered states in best-performing HMM/GMM system.

We evaluate the hypothesis with ASR studies on German part of MediaParl database that includes real speech data from Valais parliament of Switzerland (Sections 3 and 4). Finally we conclude in Section 5.

2. BACKGROUND AND MOTIVATION

In this section, we first provide a brief overview of Kullback-Leibler divergence based HMM (KL-HMM) approach and later relate it with standard HMM/ANN approach. Finally, we present motivation for the present study.

2.1. Kullback-Leibler divergence based HMM

Kullback-Leibler divergence based HMM (KL-HMM) is a posterior-based ASR approach, where posterior probabilities of acoustic units (for example, context-independent phones) estimated using an ANN are directly used as feature observations [6, 7]. Let $\mathbf{z}_t$ denote the acoustic unit posterior feature vector estimated at time frame t,

$$\begin{aligned}\mathbf{z}_t &= [z_t^1, \ldots, z_t^d, \ldots, z_t^D]^{\mathrm{T}} \\ &= [P(a_1|\mathbf{x}_t), \ldots, P(a_d|\mathbf{x}_t), \ldots, P(a_D|\mathbf{x}_t)]^{\mathrm{T}} \end{aligned} \quad (1)$$

where $\mathbf{x}_t$ is the acoustic feature (e.g., cepstral feature) at time frame t, $\{a_1, \ldots, a_d, \ldots, a_D\}$ is the set of acoustic units, D is the number

This work was supported by Hasler foundation through the grants Flexible acoustic data driven grapheme to acoustic unit conversion (AddG2SU), 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).

of acoustic units, and $P(a_d|\mathbf{x}_t)$ denotes the posterior probability of acoustic unit a_d given $\mathbf{x}_t$. We refer to $\mathbf{z}_t$ as acoustic unit posterior feature.

Let $\mathcal{L} = \{l_1, \dots, l_i, \dots, l_I\}$ be the set of lexical units (for example, context-dependent subword units). For the sake of clarity here we assume each lexical unit represents one HMM state. Each lexical unit l_i in the KL-HMM system is parameterized by a categorical distribution $\mathbf{y}_i = [y_i^1, \dots, y_i^d, \dots, y_i^D]^{\mathrm{T}}$ where $0 \leq y_i^d \leq 1$, $\sum_{d=1}^{D} y_i^d = 1$, and $y_i^d = p(a_d|l_i)$. Therefore, KL-HMM can be seen as probabilistic lexical modeling approach in which the relationship between acoustic units modeled by ANN and lexical units modeled by KL-HMM is probabilistic [8, 9, 10].

The local score at each HMM state is the Kullback-Leibler (KL) divergence between the acoustic unit posterior feature and the state distribution,

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

In this case $\mathbf{y}_i$ acts as the reference distribution. KL-divergence being an asymmetric measure, there are also other ways to estimate the local score. More specifically, reverse KL-divergence (RKL) where acoustic unit posterior feature $\mathbf{z}_t$ is the reference distribution and symmetric KL-divergence (SKL) which is the average of KL and RKL local scores.

The KL-HMM parameters $\{\mathbf{y}_i\}_{i=1}^{I}$ are estimated using Viterbi expectation maximization algorithm which minimizes a cost function that is based on KL-divergence. During testing, 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 $-S(\mathbf{y}_i, \mathbf{z}_t)$.

2.2. Relationship with Hybrid HMM/ANN approach

In standard hybrid HMM/ANN approach, the relationship between acoustic units modeled by ANN and lexical units modeled by HMM is deterministic [9]. The posterior probability of acoustic unit given by the ANN is converted to scaled-likelihood of HMM state and is used as local emission score [1],

$$p_{sl}(\mathbf{x}_t|q_t = l_i) = \frac{p(\mathbf{x}_t|q_t = l_i)}{P(\mathbf{x}_t)} = \frac{P(a_k|\mathbf{x}_t)}{P(a_k)} \tag{3}$$

The lexical unit l_i is deterministically mapped to acoustic unit a_k modeled by ANN, where $a_k \in \{a_1, \dots, a_D\}$.

In common practice, the outputs of ANN are divided with corresponding priors, to avoid the mismatch between relative frequencies in acoustic data and relative frequencies given by pronunciation and language models. However, in theory, HMMs can be trained and decoded using posterior probabilities of acoustic units directly [1, 11]. In that sense, hybrid HMM/ANN approach can be seen as special case of KL-HMM approach when the relationship between acoustic and lexical units is deterministic and the local score is KL-divergence as given in Eqn (2). More precisely, if lexical unit l_i is deterministically mapped to acoustic unit a_k ($l_i \mapsto a_k$), then the state distribution $\mathbf{y}_i$ is a Kronecker delta distribution, where

$$y^d_{q_t=i} = \begin{cases} 1, & \text{if d = k ;} \\ 0, & \text{otherwise.} \end{cases}$$

and the local score $S(\mathbf{y}_i, \mathbf{z}_t)$ is -log $P(a_k|\mathbf{x}_t)$.

2.3. Motivation for the Present Study

Earlier, in hybrid HMM/ANN systems, the acoustic units were context-independent phones and were typically modeled using ANN with one hidden layer. The main limitation of hybrid HMM/ANN systems was the difficulty in modeling context-dependent subword units. Since, each ANN output is deterministically related to a HMM state, it is necessary that ANN models all the context-dependent subword units. However, training ANN that models all possible context-dependent subword units is impractical, owing to complexity and data sparseness issues.

On the other hand, HMM/GMM systems are able to efficiently model context-dependent subword units through state clustering and tying. The decision tree based state clustering is used to automatically cluster the context-dependent subword units into clustered acoustic units (commonly referred to as physical states). A context-dependent subword unit is modeled with three HMM states and each HMM state is often deterministically related to one of the acoustic units (deterministic lexical modeling).

Recently, hybrid HMM/ANN systems are extended along the lines of HMM/GMM systems [2, 3]. More specifically, ANNs with more than one hidden layer are used to classify clustered acoustic units. Each HMM state representing a possible context-dependent phone (lexical unit) is deterministically related to one of the acoustic units. Most often, the number of output units of ANN is fixed to the number of clustered context-dependent phones of best performing HMM/GMM system [2, 3].

It has been shown that the KL-HMM approach can perform similar to or better than hybrid HMM/ANN and HMM/GMM systems when trained using context-independent phones as acoustic units and context-dependent phones as lexical units [12, 9]. In this paper, we extend the investigations using KL-HMM approach along the lines of recent hybrid HMM/ANN approaches. More precisely, we use context-dependent clustered states as acoustic units and context-dependent phones as lexical units. Furthermore, we hypothesize that KL-HMM approach may not require as many acoustic units as best performing HMM/GMM system because the approach models probabilistic relationship between acoustic and lexical units, in addition to using a discriminative acoustic model, i.e., ANN.

3. EXPERIMENTAL SETUP

In this section, we provide details of the MediaParl database used in experiments and explain the setup of HMM/GMM, HMM/ANN and KL-HMM systems used for evaluation.

3.1. Database

We used the German part of MediaParl database for evaluation [13]. MediaParl is a bilingual corpus containing data (debates) in both Swiss German and Swiss French which were recorded at the Valais parliament in Switzerland. Valais is a state which has both French and German speakers with high variability in local accents specially among German speakers. Therefore, MediaParl provides a real-speech corpus that is suitable for ASR studies in particular for accented and non-native speech. In this paper, we study ASR performance for both native and non-native speakers.

In our experiments, audio recordings with 16 kHz sampling rate are used. The database is partitioned into training, development and test sets following the structure given in [13]: 90% of the native speakers (who speak only one language) form the training set (of 14 hours and 73 speakers) and the remaining 10% form the development set (of 2 hours and 8 speakers). The test set (of 4 hours and 7 speakers) contains the speakers who speak in both French and German. Four speakers are bilingual speakers with German as their native language and for three speakers, French is the native language. We refer to the utterances of speakers with German as their native language as *native* speech (1605 utterances) and German utterances from speakers with French as their first language as *non-native* speech (87 utterances).

The MediaParl corpus dictionary is provided in SAMPA format with a phone set of size 57 (including sil) and contains all the words in the train, development and test set. The vocabulary size is 16755 words. For the language model, we used a bigram model on transcriptions of the training set as well as EuroParl corpus (which consists of about 50 million words for each language).

3.2. Systems

HMM/GMM systems: We trained standard context-independent and cross-word context-dependent HMM/GMM systems with 39 dimensional PLP cepstral features extracted using HTK toolkit [14]. We tuned the number of Gaussians for both context-independent and context-dependent systems on the development set. For context-dependent HMM/GMM systems we also tuned the number of clustered states. The best performing context-independent system used 128 Gaussians and the best performing context-dependent system had 3000 clustered states with 16 Gaussians per clustered state.

Multilayer perceptrons (MLPs): For hybrid HMM/ANN and KL-HMM systems, we studied various ANNs, more precisely, MLPs that vary in terms of number of MLP layers or output units. We used 39-dimensional PLP cepstral features with four frames preceding context and four frames following context as MLP input. All the MLPs were trained with output non-linearity of softmax and minimum cross-entropy error criterion, using Quicknet software [15]. We investigated the following MLPs:

- *MLP-3L-CI-57*: a standard 3-layer MLP modeling context-independent phones (which are of size 57) as output units. The number of parameters was approximately 0.8M.
- *MLP-5L-CI-57*: a 5-layer MLP modeling context-independent phones as output units. This MLP had an architecture of 351 x 2000 x 2000 x 2000 x 57 with about 8.8M parameters.
- *MLP-5L-CD-N*: a 5-layer MLP modeling N context-dependent clustered phones as outputs where $N \in \{195, 385, 549, 759, 1101, 3000\}$. The output units were derived by clustering context-dependent phones in HMM/GMM framework using decision tree state tying. The different number of acoustic units were derived by adjusting the log-likelihood difference. All the 5-layer MLPs had roughly the same number of parameters ($\approx$ 8.8M). More precisely, each 5-layer MLP had an architecture of 351 x 2000 x HU x 2000 x N and the number of middle layer hidden units HU was adjusted so as to keep the number of parameters constant.

Hybrid HMM/ANN systems: We estimated the scaled likelihoods in hybrid HMM/ANN system by dividing the posterior probabilities $P(a_k|\mathbf{x}_t)$ derived from MLP with the priori probability of acoustic unit $P(a_k)$ estimated from relative frequencies in the training data. These scaled likelihoods were used as emission probabilities for HMM states.

KL-HMM systems: KL-HMM systems used acoustic units posterior probabilities as feature observations and modeled context-dependent (tri) phones. 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 that resulted in minimum KL-divergence on training data was selected. In most of the cases RKL resulted as the local score. For tying KL-HMM (lexical) states we applied KL-divergence based decision tree state tying method proposed in [16].

For all the systems, each lexical unit was modeled with three HMM states. In addition, the parameters of systems (such as language scaling factor and word insertion penalty) were tuned on the development set.

4. RESULTS AND ANALYSIS

In this section, we first present ASR studies on MediaParl corpus and then analyze the results for native and non-native speech.

4.1. Effect of Number of Acoustic Units on ASR Performance

Figure 1 presents the results in terms of word error rate (WER) for HMM/GMM, hybrid HMM/ANN and KL-HMM systems with varying number of acoustic units.

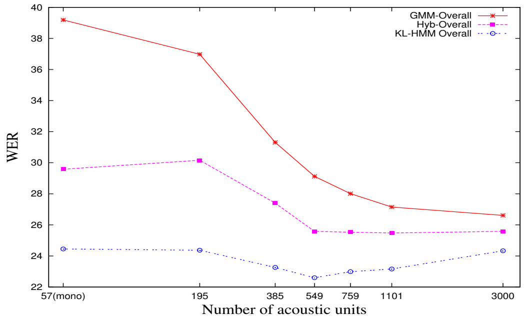

Fig. 1. Overall performance in terms of WER for different systems with varying number of acoustic units.

It can be observed from Figure 1 that:

- For HMM/GMM system, as the number of acoustic units is increased, the WER decreases.
- Similar trend exists for HMM/ANN system. However, when $N \geq 549$ the decrease in WER is not statistically significant.
- The WER of KL-HMM system is less sensitive to the number of acoustic units N. The system achieves optimal WER with fewer number of acoustic units ($N = 549$) compared to HMM/GMM framework ($N = 3000$).
- Irrespective of the number of acoustic units, KL-HMM system results in best performance.

Indeed, these results show the validity of hypothesis that KL-HMM system may not need as many acoustic units as best performing HMM/GMM system. The acoustic model complexity (in terms of number of parameters) of HMM/GMM system increases with increasing N whereas, in the case of hybrid HMM/ANN and KL-HMM systems, the acoustic model complexity is constant with increasing N ($\approx$ 8.8M). The hybrid HMM/ANN and KL-HMM systems could probably improve with increasing N if relatively large amount of training data was available.

Table 1 summarizes the best results for HMM/GMM, hybrid HMM/ANN and KL-HMM systems along with the total number of acoustic and lexical model parameters. The acoustic and lexical model parameters are calculated following the procedure given in [10].

Systems	θ_a	θ_l	WER
HMM/GMM	3.8M	185K	**26.6**
Hybrid-MLP-5L-CD-1101	8.8M	185K	**25.5**
KL-HMM-MLP-3L-CI-57	0.8M	412K	26.8
KL-HMM-MLP-5L-CI-57	8.8M	397K	24.4
KL-HMM-MLP-5L-CD-549	8.8M	5M	**22.6**

Table 1. Overall results in terms of WER for different systems when modeling context-dependent (tri) phones. θ_a , θ_l denote the number of parameters in acoustic and lexical model respectively.

KL-HMM and hybrid HMM/ANN systems are denoted along with the MLP used. The table also provides the WER of *KL-HMM-MLP-3L-CI-57* and *KL-HMM-MLP-5L-CI-57* systems to understand the effect of deeper MLP architecture on ASR performance in KL-HMM framework. From Table 1 it can be observed that:

- The *Hybrid-MLP-5L-CD-1101* system performs better than the best performing *HMM/GMM*[1] system with 4.3% relative improvement which is in line with recent studies on deep MLP architectures for acoustic modeling in ASR [3].
- The *KL-HMM-MLP-3L-CI-57* system performs similar to *HMM/GMM* system, despite using fewer number of parameters and modeling only context-independent phones as acoustic units.
- The *KL-HMM-MLP-5L-CI-57* system that uses more layers, performs better than *KL-HMM-MLP-3L-CI-57* system with 2.4% absolute improvement [17]. Also, the performance of *KL-HMM-MLP-5L-CI-57* system, in spite of using context-independent acoustic units, is better than *Hybrid-MLP-5L-CD-1101* system that uses context-dependent acoustic units.
- The *Hybrid-MLP-5L-CD-1101*, *KL-HMM-MLP-5L-CI-57* and *KL-HMM-MLP-5L-CD-549* systems have same complexity in terms of number of acoustic model parameters. The difference in the performance between the systems is due to the difference in the number of MLP outputs and the lexical model (deterministic or probabilistic).
- The *KL-HMM-MLP-5L-CD-549* system that uses context-dependent acoustic units performs significantly better than every other system in the table (with at least 99% confidence).

4.2. Analysis for Native and Non-Native Speech

Figure 2 shows the performance of the systems in terms of WER with varying number of acoustic units for both native and non-native speech on MediaParl test set. It can be observed that in general, as number of acoustic units is increased, the WER on native speech decreases. On the other hand, as number of acoustic units is increased, the WER on non-native speech eventually increases for all the three systems. This effect can be better seen for hybrid HMM/ANN and KL-HMM systems, where using context-independent phones as acoustic units resulted in lower WER for non-native speech. This suggests that, when the training data includes only the native speech, as done in this study, increasing the number of acoustic units may not be beneficial for non-native speech recognition. We intend to investigate this aspect further on French part of MediaParl database which contains more non-native utterances.

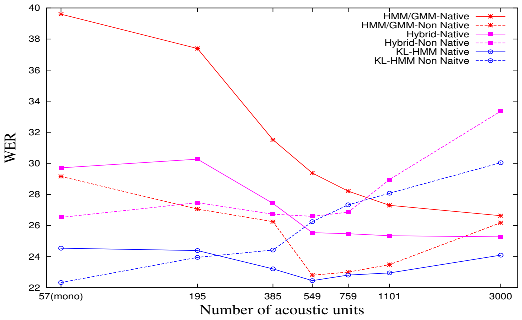

Fig. 2. Performance of the systems in terms of WER with varying number of acoustic units for both native and non-native speech.

5. DISCUSSION AND CONCLUSION

In this paper, we studied different systems, namely, standard HMM/GMM system, hybrid HMM/ANN system and KL-HMM system. On the outset, these systems seem to be very different. However, all these systems can be explained through one-and-same principle i.e., ASR by matching a sequence of "latent" symbols based on acoustic information with a set of reference sequences of latent symbols based on lexical and syntactic information. In other words, we are performing ASR by generating a sequence of latent symbols based on acoustic information and matching it with a sequence of latent symbols corresponding to each word hypothesis generated using lexical and syntactical information. In that sense, the systems investigated in this paper differ on four fundamental issues,

1. latent symbol set: context-independent phone or context-dependent clustered phone states.
2. modeling of relationship between latent symbols and acoustic signal (acoustic model): HMM/GMM system uses a generative model while hybrid HMM/ANN system and KL-HMM system use a discriminative model[2].
3. modeling of relationship between latent symbols and lexical (subword) units (lexical model) [8, 9]: in HMM/GMM and hybrid HMM/ANN systems the relationship is one-to-one deterministic map while in KL-HMM system the relationship is probabilistic.
4. cost function to match (locally) the symbol sequences [9]: in both HMM/GMM system and hybrid HMM/ANN system it is log of dot product between acoustic model likelihood vector and lexical model (Kronecker delta) posterior probability vector, while in the case of KL-HMM it is the KL-divergence between acoustic model posterior probability vector and lexical model posterior probability vector. The size of the likelihood/posterior probability vector depends upon the cardinality of the latent symbol set.

Language modeling and efficient search of output word hypothesis using dynamic programming are common aspects for all these systems. Our studies show that KL-HMM approach, which uses discriminative acoustic model, probabilistic lexical model and discriminative local score [18], can achieve better system than standard HMM/GMM system and hybrid HMM/ANN system with fewer latent symbols. Our future work will focus towards a) use of latent symbols obtained with more than single preceding and following context and b) use of latent symbols extracted without using phoneme information [8].

[1] The accuracy of HMM/GMM system reported in [13] was lower compared to our result, as speech data with sampling rate of 8 kHz was used.

[2] KL-HMM approach is also applicable to generative acoustic model [8]

6. REFERENCES

[1] N. Morgan and H. Bourlard, "Continuous Speech Recognition: An Introduction to the Hybrid HMM/Connectionist Approach," *IEEE Signal Processing Magazine*, pp. 25–42, May 1995.

[2] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition," in *IEEE Trans. on Audio, Speech, and Language Processing*, 2012.

[3] G. Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[4] O. Vinyals and N. Morgan, "Deep vs. Wide: Depth on a Budget for Robust Speech Recognition," in *Proc. of Interspeech*, 2013.

[5] Z. Tüske, R. Schlüter, H. Ney, and M. Sundermeyer, "Context-Dependent MLPs for LVCSR: TANDEM, Hybrid or Both?," in *Proc. of Interspeech*. 2012, ISCA.

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

[7] 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.

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

[9] Ramya Rasipuram and Mathew Magimai.-Doss, "Acoustic and Lexical Resource Constrained ASR using Language-Independent Acoustic Model and Language-Dependent Probabilistic Lexical Model," Idiap-RR Idiap-RR-02-2014, Idiap, March 2014.

[10] Ramya Rasipuram and Mathew Magimai-Doss, "Probabilistic lexical modeling and grapheme-based automatic speech recognition," Idiap-RR Idiap-RR-15-2013, Idiap, April 2013.

[11] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, Kluwer Academic Publishers, Norwell, MA, USA, 1993.

[12] D. Imseng, J. Dines, P. Motlicek, P. Garner, and H. Bourlard, "Comparing Different Acoustic Modeling Techniques for Multilingual Boosting," in *Proc. of Interspeech*, 2012.

[13] D. Imseng et al., "MediaParl: Bilingual Mixed Language Accented Speech Database," in *Proc. of the IEEE Workshop on SLT*, Dec. 2012, pp. 263–268.

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

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

[16] D. Imseng, J. Dines, P. Motlicek, P. Garner, and H. Bourlard, "Comparing Different Acoustic Modeling Techniques for Multilingual Boosting," in *Proc. of Interspeech*, Sept. 2012.

[17] D. Imseng, P. Motlicek, P. Garner, and H. Bourlard, "Impact of Deep MLP Architecture on Different Acoustic Modeling Techniques for Under-Resourced Speech Recognition," in *Proc. of ASRU*, Dec. 2013.

[18] R. Blahut, "Hypothesis Testing and Information Theory," *IEEE Trans. on Information Theory*, vol. IT-20, no. 4, 1974.