SPEAKER ADAPTIVE KULLBACK-LEIBLER DIVERGENCE BASED HIDDEN MARKOV MODELS

David Imseng[1,2] *and Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland
{dimseng,bourlard}@idiap.ch

ABSTRACT

Kullback-Leibler divergence based hidden Markov models (KL-HMM) have recently been introduced as an efficient and principled way to directly model sequences of posterior vectors to perform Automatic Speech Recognition (ASR). Through efficient feature level adaptation and parsimonious use of parameters, KL-HMM was successfully applied to accented and under-resourced speech recognition tasks. In this paper, inspired from Maximum A Posteriori (MAP) adaptation, we further boost KL-HMM performance by applying Bayesian speaker adaptation, directly applied to posterior features. This approach performs a simple, adaptive regression between phone posteriors estimated with a Multilayer Perceptron (MLP) on large amounts of speaker-independent training data, and speaker-specific phone posteriors generated by the speaker-independent MLP on very limited amount of speaker-specific adaptation data. Using Swiss French data (MediaParl), we show that such speaker adaptive KL-HMM can significantly outperform conventional adaptation techniques on non-native speech while yielding similar performance on native data.

Index Terms— Kullback-Leibler divergence, speaker adaptation, non-native speech, speech recognition

1. INTRODUCTION

Several speaker adaptation techniques have been proposed to improve Automatic Speech Recognition (ASR) performance. Speaker adaptation is also particularly relevant in the case of non-native ASR, given the high variability of accented speech and the usually small amount of non-native speech data available for training [1, 2, 3, 4]. In the context of HMM/GMM (Hidden Markov Models parametrized by Gaussian Mixture Models), traditional solutions include Maximum Likelihood Linear Regression (MLLR), Maximum a Posteriori (MAP), and model interpolation [1, 2, 3]. In the case of hybrid HMM/MLP systems (using a Multilayer Perceptron to estimate emission probabilities), a Linear Hidden Network (LHN) was typically used to adapt the MLP to a speaker [4].

In a recent study [5], we presented an approach to deal with the acoustic and pronunciation variability of non-native speech using a Kullback-Leibler divergence based hidden Markov model (KL-HMM) for acoustic modeling. KL-HMM is a particular form of HMM where each HMM state is parametrized by a trained posterior distribution (reference posterior) which models posterior features estimated by an MLP as KL-HMM acoustic features. The MLP can be trained on multilingual data to estimate universal phone class posterior probabilities (multilingual posterior features). The KL-HMM is then able to exploit the multilingual information on a feature level and learn the relation between non-native speech data and multilingual posterior features. In such a multilingual setup, KL-HMM was shown to outperform MLLR and LHN for non-native speaker adaptation [5]. In those experiments, it was also observed that KL-HMM was quickly yielding state-of-the-art performance with limited amount of training data.

In this paper, we propose and evaluate speaker adaptive KL-HMM. Similar to MAP that adapts the Gaussian Mixture Models (GMM) of an HMM/GMM based speech recognizer, the proposed speaker adaptive KL-HMM adapts the generic reference posteriors of the KL-HMM. Just using a couple of minutes of speech data, and using the same speaker-independent MLP to generate features, we train a speaker-specific KL-HMM. The generic KL-HMM reference posteriors are then adapted by performing a linear combination with the speaker-specific reference posteriors. For the sake of comparison (and in contrast to our earlier study [5]), we use a monolingual MLP to generate the features in this work. We expect improvement if a multilingual MLP is used, but leave the empirical verification for future work.

For this study, we used data from the bilingual *MediaParl*[1] database [6]. MediaParl is a Swiss accented bilingual database containing recordings in both accented French and German, as they are spoken at the Parliament in Valais, a state of Switzerland. The advantage of MediaParl is that it is a pretty large multilingual database and the test set consists of bilingual speakers, hence non-native and native speech recorded at same conditions.

In the sequel of this paper, we will first briefly review standard KL-HMM techniques (Section 2) and then introduce the speaker adaptive KL-HMM concept (Section 3). As described in Section 4, we use the French part of the MediaParl database to show that the performance of the speaker adaptive KL-HMM on native French speech is comparable to MLLR and significantly better than MLLR on non-native French speech. Initially, we also evaluated MAP. However, since we deal with extremely low amounts of data (up to five minutes), MLLR outperforms MAP. This result is consistent with earlier studies [1] and we therefore only report MLLR results.

2. KL-HMM

The notion of KL-HMM was initially introduced by Aradilla [7]. In this section, we briefly present the KL-HMM model and summarize the basic training and decoding techniques.

This research was supported by the Swiss NSF through the project Interactive Cognitive Systems (ICS) under contract number 200021_132619/1.

[1]Documented and publicly available at http://www.idiap.ch/dataset/mediaparl

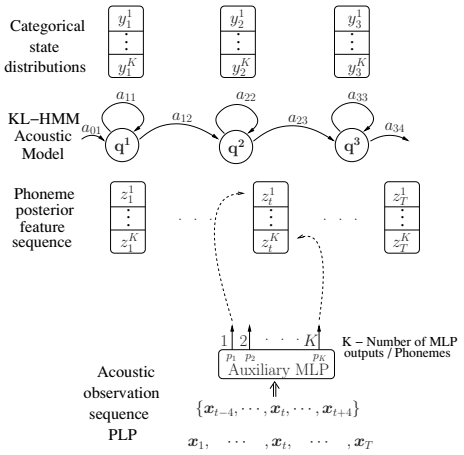

Fig. 1. KL-HMM - the emission probabilities are modeled with categorical distributions and the MLP output can directly be used.

2.1. KL-HMM modeling

As illustrated in Figure 1, a KL-HMM is a particular form of HMM in which the emission probability of a particular state q^d is parametrized by a *categorical distribution*[2] $\boldsymbol{y}_d = (y_d^1, \ldots, y_d^K)^\mathsf{T}$, where K is the dimensionality of the features (corresponding in our case to the number of MLP outputs) and T the transpose operation. The KL-HMM categorical distributions can directly be trained from phone class posterior probabilities $\boldsymbol{z}_t$.

In most of our initial KL-HMM work [5, 8], we often used the symmetric variant of the KL divergence. More recently, though, it was observed that the asymmetric KL divergence $KL(x||y)$, as defined below, was consistently more robust. This is also intuitively reasonable. Indeed, the underlying acoustic modeling problem is not symmetric since we observe the posterior features and train the categorical distributions. Therefore, we use the following Kullback-Leibler based distance as local score in this study:

$$d(\boldsymbol{z}_t, \boldsymbol{y}_d) = \sum_{k=1}^{K} z_t^k \log \frac{z_t^k}{y_d^k}. \tag{1}$$

A detailed description of training and decoding algorithms based on the symmetric variant of the KL divergence can be found in [8]. In this paper, we use the asymmetric KL divergence as given in (1). For clarity, we briefly review the training and decoding algorithms.

2.2. KL-HMM training

The categorical distributions $Y = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_D\}$ can be learned using an iterative Viterbi segmentation-optimization scheme. The cost function can be defined by integrating the local score, given in (1), over time t and states q^d, resulting in:

$$\mathcal{F}(Z, Y) = \sum_{t=1}^{T} \sum_{d=1}^{D} d(\boldsymbol{z}_t, \boldsymbol{y}_d) \delta_t^d, \tag{2}$$

[2]A *Categorical Distribution* is a multinomial distribution from which only one sample is drawn. In our case that one sample per HMM state is trained along an EM-like algorithm minimizing the accumulated KL divergence between the reference categorical distributions and the posterior sequences used as acoustic features.

where the Kronecker delta δ_t^d is defined as:

$$\delta_t^d = \begin{cases} 1, & \text{if } \boldsymbol{z}_t \text{ is associated with state } q^d \\ 0, & \text{otherwise.} \end{cases}$$

To associate each $\boldsymbol{z}_t$ with one of the states, the HMM aligns the phone class posterior probabilities Z with the states by minimizing $\mathcal{F}(Z, Y)$, given in (2).

Each $\boldsymbol{z}_t$ is then used to update a particular categorical distribution $\boldsymbol{y}_d$. To minimize $\mathcal{F}(Z, Y)$ subject to $\sum_{k=1}^{K} y_d^k = 1$, we take the partial derivative with respect to each variable y_d^k and set it to zero to find the minimum. Then, we introduce the Lagrange multipliers λ to enforce the sum to one constraint:

$$\frac{\partial}{\partial y_d^k} \mathcal{F}(Z, Y) + \lambda \left(\sum_{k=1}^{K} y_d^k - 1 \right) = 0. \tag{3}$$

Solving (3) yields:

$$y_d^k = \frac{1}{T_d} \sum_{\forall t^*} z_t^k. \tag{4}$$

where the sum extends over all t^* such that $\boldsymbol{z}_{t^*}$ is associated with state q_d and T_d stands for the number of frames associated with state q_d.

2.3. KL-HMM decoding

During decoding, we aim at finding the optimal KL-HMM state sequence Q minimizing[3]:

$$\mathcal{F}_{\mathcal{Q}}(Z, Y) = \min_{\mathcal{Q}} \sum_{t=1}^{T} d(\boldsymbol{z}_t, \boldsymbol{y}_{q_t}), \tag{5}$$

where $\mathcal{Q} = \{q_1, \ldots, q_T\}$ stands for all possible state sequences and $\boldsymbol{y}_{q_t}$ is the categorical distribution associated with q_t, the KL-HMM state at time t.

3. SPEAKER ADAPTIVE KL-HMM

Earlier studies have shown that KL-HMM performs extremely well when only a small amount of training data is available [9]. Even though it is not an adaptation technique, the categorical distributions are trained from scratch, it outperformed current state-of-the-art adaptation techniques such as MLLR. However, if the amount of data to train/adapt gets below a certain threshold, KL-HMM may overfit. Therefore, we introduce the concept of speaker adaptive KL-HMM in this section.

We assume to have a generic KL-HMM system with the categorical distributions $Y = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_D\}$ trained as described in Section 2.2. Furthermore, we suppose to have a small amount of speaker-specific adaptation data $X^s = \{\boldsymbol{x}_1^s, \ldots, \boldsymbol{x}_N^s\}$. Given the speaker-specific data X^s, we can generate posterior features $Z^s = \{\boldsymbol{z}_1^s, \ldots, \boldsymbol{z}_N^s\}$ by using the same speaker-independent MLP as in Section 2.1.

The posterior features Z^s can then be used to train a speaker specific KL-HMM with categorical distributions $Y^s = \{\boldsymbol{y}_1^s, \ldots, \boldsymbol{y}_D^s\}$ along the same procedure as described in Section 2.2. For the speaker-specific KL-HMM training, we use the generic categorical distributions Y as seed models (i.e. initialization: $Y^s = Y$). Due to

[3]For the sake of simplicity, the transition probabilities $a_{q_{t-1}q_t}$ are omitted in (5) because they are considered to be equal and fixed.

Language	Vocabulary size	Number of bigrams	Perplexities DEV	TST
French	12,035	1.5 M	147	152
German	16,727	1.9 M	295	360

Table 1. Statistics of the monolingual language models.

the small amount of adaptation data, we expect the speaker-specific KL-HMM parameters Y^s to overfit. To overcome that problem, we eventually combine the generic Y and the speaker specific Y^s on the state-level:

$$\boldsymbol{y}_d^{\text{adaptive}} = \alpha \boldsymbol{y}_d + (1-\alpha)\boldsymbol{y}_d^s \quad (6)$$

where $\boldsymbol{y}_d^{\text{adaptive}}$ stands for the categorical distribution of the speaker adaptive KL-HMM and $\alpha \in [0, 1]$ is a parameter of the combination.

4. EXPERIMENTS AND RESULTS

In this section, we evaluate the speaker adaptive KL-HMM and compare it to the standard KL-HMM, a conventional HMM/GMM system and MLLR.

4.1. Data

For our studies, we used the French part of the MediaParl database [6]. MediaParl is a Swiss accented bilingual database containing recordings in both French and German as they are spoken in Switzerland. The data were recorded at the Valais Parliament. Valais is a bilingual Swiss state with many local accents and dialects. Therefore, the database contains data with high variability and is suitable to study multilingual, accented and non-native speech recognition.

The MediaParl database contains a dictionary with all the words (no out of vocabulary words) and standardized training, development and test sets. The bigram language model (see Table 1) was trained on two sources: the transcriptions of the training set and texts from the corpus Europarl, a multilingual corpus of European Parliament proceedings [10]. Europarl is made up of about 50 million words for each language and is used to overcome data sparsity of the MediaParl texts. However, vocabularies were limited to the sole words from MediaParl.

The test set, shown in Table 2, contains all the seven speakers that speak in both languages. In this paper, we study fast speaker adaptation (minutes of data for each speaker) on the French part of the data. *Speaker 059* is discarded because a couple of French phonemes are not pronounced at all. For all the other speakers, we randomly select five minutes of adaptation data (and exclude that data from the test set). Only for *speaker 079* (2 minutes of French data in total) we use about half the data for adaptation and the other half for testing.

4.2. Systems

For our study, we compare four systems: conventional HMM/GMM, MLLR, KL-HMM and speaker adaptive KL-HMM.

4.2.1. HMM/GMM

For the standard HMM/GMM system, the adaptation data was not used. We used the training data from the French MediaParl to train a conventional crossword context-dependent speech recognizer from 39 Mel-Frequency Perceptual Linear Prediction (MF-PLP) features (C0-C12+Δ+$\Delta\Delta$), extracted with the HTS variant [11] of the HTK toolkit. Each triphone was modeled with three states and each state was modeled with 16 Gaussians. To tie rare states, we applied a conventional decision tree. The minimum description length criterion was used to determine the number of tied states [12]. For decoding, we used the bigram language model as described in Section 4.1.

Spkr ID	Sent. in French	Adapt data [min]	Test	Sent. in German	Mother tongue
059	31	-	-	195	German
079	22	1	1	698	German
094	313	5	60	72	French
096	89	5	15	8	French
102	72	5	7	7	French
109	233	5	46	402	German
191	165	5	28	310	German
Total	925	26	157	1692	

Table 2. MediaParl-TST: speakers using both languages form the test set. For each speaker the number of French and German sentences as well as the mother tongue is given.

4.2.2. Maximum likelihood linear regression

In an earlier study, we investigated MLLR as well as a constrained version of it (CMLLR) to evaluate whether a new language could be accommodated by linear transforms [8]. CMLLR has fewer parameters and we assumed that this could be advantageous if we only have access to a limited amount of data. However, even if we only used 5 minutes of adaptation data, MLLR outperformed CMLLR. Therefore, in this study, we only investigated standard MLLR . For this, we used the adaptation data described in Table 2 to perform speaker adaptation and employed a regression tree that allowed up to 16 regression classes.

4.2.3. KL-HMM

For the standard KL-HMM system, we first trained an MLP from the 39 MF-PLP features (see Section 4.2.1) in a nine frame temporal-context (four preceding and following frames). The number of parameters in the MLP was set to 10% of the number of available training frames, to avoid overfitting. We used Quicknet [13] software to train the MLP.

The MLP was trained on triphone targets. To obtain triphone targets, we used the HMM/GMM system presented in Section 4.2.1 with a modified decision tree. As described by [12], the MDL criterion has a hyper-parameter, c, which controls the weight of the term that penalizes models with large amounts of tied states. For the triphone target creation, we used $c = 16$ to obtain 659 tied states, used as MLP targets. Then we used all the French MediaParl training data to train a crossword context-dependent KL-HMM based speech recognizer. Similar to the HMM/GMM system, we did not use the adaptation data and we tied rare states by applying decision tree clustering reformulated as dictated by the KL criterion [9]. For decoding, we used the bigram language model as described in Section 4.1.

4.2.4. Speaker adaptive KL-HMM

The speaker adaptive KL-HMM was trained as described in Section 3. As seed models, we used the KL-HMM system presented in

Section 4.2.3. In Section 4.3.1, we discuss the choice of α.

4.3. Results

In this section, we first investigate the choice of the parameter α for the speaker adaptive KL-HMM and then we compare the four systems described in Section 4.2 against each other.

4.3.1. Tuning of the parameter α

Figure 2 shows the influence of the parameter α. If α is set to one, the speaker adaptive KL-HMM is equivalent to the standard KL-HMM. For each speaker, there is at least one α value for which the performance of the speaker adaptive KL-HMM is better than the performance of the standard KL-HMM. However, we also see that for some values of α, the performance decreases. It can clearly be seen that α-values close to zero perform bad in general, i.e. the adapted KL-HMM system overfits. The highest performance gains can be seen for two non-native speakers [4] (*079* and *191*). The French HMM/GMM baseline system reported in [6] performed particularly bad on theses two speakers, hence they seem to have a strong accent.

In the remainder of this paper, we will use the best performing α value that we found for each speaker (on the test set). In future, we will investigate how to automatically tune α.

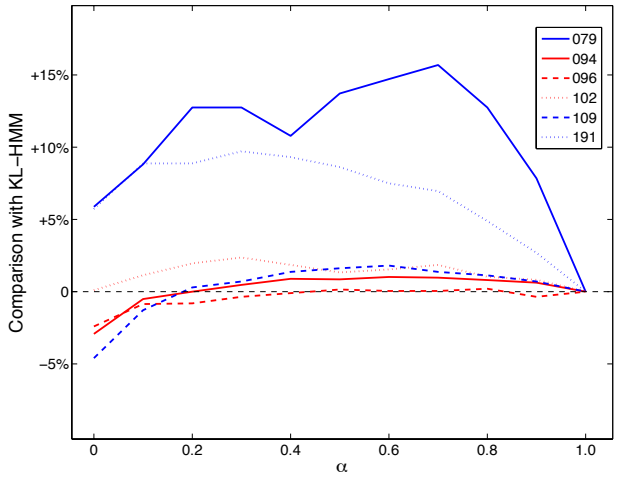

Fig. 2. Tuning of the parameter α: the performance of the speaker adaptive KL-HMM is compared with the standard KL-HMM performance (y-axis shows relative performance change). Each curve represents one speaker. Red curves represent native speakers and blue curves stand for non-native speakers.

4.3.2. System comparison

In Figure 3, the performance of a conventional HMM/GMM system, MLLR, KL-HMM and speaker adaptive KL-HMM are compared. On the left and on the right, the performance on native and non-native speech, respectively, is shown. Blue bars represent HMM based systems (Standard=HMM/GMM, Adapt=MLLR) and red bars represent KL-HMM based systems (Standard=KL-HMM, Adapt=speaker adaptive KL-HMM). At a first glance, we observe that for native speech, the HMM/GMM based systems perform better and for non-native speech, the KL-HMM based systems perform better. If we have a closer look, we can distinguish four different cases:

- Standard on native speech: the HMM/GMM system performs significantly better than the KL-HMM system
- Adapt on native speech: there is no significant difference between the MLLR and the speaker adaptive KL-HMM system
- Standard on non-native speech: there is no significant difference between the HMM/GMM and the KL-HMM system
- Adapt on non-native speech: the speaker adaptive KL-HMM performs significantly better than MLLR

For the significance test, we used the bootstrap estimation method [14] and a confidence interval of 99%. Overall, the speaker adaptive KL-HMM system performs best.

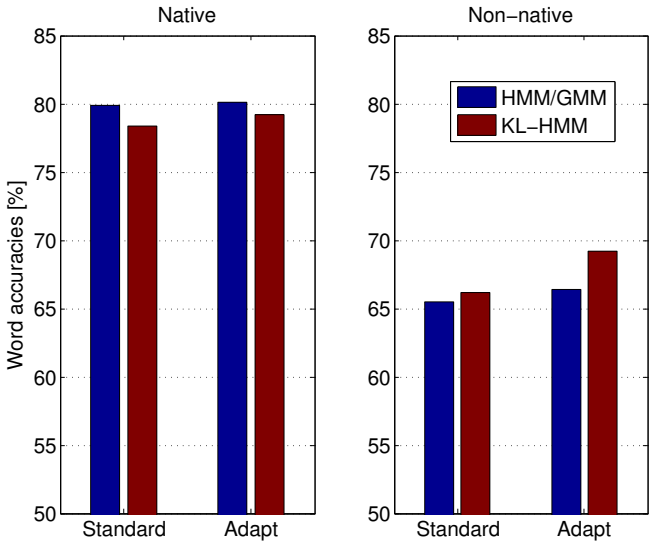

Fig. 3. Comparison of the four systems described in Section 4.2. The left and right plot shows word accuracies on native and non-native speech, respectively. *Standard* stands for HMM/GMM and KL-HMM (not used the adaptation data) and *adapt* stands for MLLR and speaker adaptive KL-HMM (used the adaptation data).

5. CONCLUSION

In this paper, we introduced a speaker adaptation approach for KL-HMM. Fast speaker adaptation is achieved by exploiting the parsimonious use of parameters of KL-HMM that efficiently uses very limited amounts of training data. Reference KL-HMM categorical distributions are then expressed as a linear function between phone posteriors estimated on large amounts of speaker-independent training data, and speaker-specific phone posteriors obtained on very limited amount of speaker-specific adaptation data. On non-native Swiss French data, the speaker adaptive KL-HMM has been shown to significantly outperform MLLR. On native speech, speaker adaptive KL-HMM still yields similar performance than MLLR.

In future, we plan to use large amounts of multilingual speaker-independent training data and expect further improvements. We will also investigate how the parameter of speaker adaptive KL-HMM, α, can automatically be tuned.

[4] We consider a speaker as a non-native French speaker if there were more German than French sentences recorded (see Table 2).

6. REFERENCES

[1] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *Proc. of ICASSP*, 2003, pp. 540–543.

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

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

[4] 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.

[5] 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.

[6] D. Imseng et al., "Mediaparl: Bilingual mixed language accented speech database," in *Proceedings of the 2012 IEEE Workshop on Spoken Language Technology*, 2012.

[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.

[8] D. Imseng, H. Bourlard, and P. N. Garner, "Using KL-divergence and multilingual information to improve ASR for under-resourced languages," in *Proc. of ICASSP*, 2012, pp. 4869–4872.

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

[10] Philipp Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proc. of the 10th Machine Translation Summit*, 2005, pp. 79–86.

[11] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, and K. Tokuda, "The HMM-based speech synthesis system version 2.0," in *Proceedings of the sixth ISCA Workshop on Speech Synthesis (ISCA SSW6)*, 2007, pp. 294–299.

[12] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL principle for speech recognition," in *Proc. of Eurospeech*, 1997, vol. I, pp. 99–102.

[13] D. Johnson et al., "ICSI quicknet software package," 2004.

[14] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. of ICASSP*, 2004, vol. 1, pp. I–409–412.