IDIAP RESEARCH REPORT

COMPARING DIFFERENT ACOUSTIC MODELING TECHNIQUES FOR MULTILINGUAL BOOSTING

David Imseng John Dines Petr Motlicek
Philip N. Garner Hervé Bourlard

Idiap-RR-01-2013

JANUARY 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

Comparing different acoustic modeling techniques for multilingual boosting

David Imseng, John Dines, Petr Motlicek, Philip N. Garner, Hervé Bourlard

June 15, 2012

Abstract

In this paper, we explore how different acoustic modeling techniques can benefit from data in languages other than the target language. We propose an algorithm to perform decision tree state clustering for the recently proposed Kullback-Leibler divergence based hidden Markov models (KL-HMM) and compare it to subspace Gaussian mixture modeling (SGMM). KL-HMM can exploit multilingual information in the form of universal phoneme posterior features and SGMM benefits from a universal background model that can be trained on multilingual data. Taking the Greek SpeechDat(II) data as an example, we show that KL-HMM performs best for small amounts of target language data.

Index Terms: Speech recognition, multilingual acoustic modeling, under-resourced languages

1 Introduction

Developing a state of the art speech recognizer from scratch for a given language is expensive. The main reason for this is the large amount of data that is needed to train current recognizers. Data collection involves large amounts of manual work, not only in time for the speakers to be recorded, but also for annotation of the subsequent recordings. Therefore, the need for training data is one of the main barriers in porting current systems to many languages. On the other hand, large databases already exist for many languages.

We have already shown that multilingual training data can boost the performance of a speech recognizer for a target language (the language that the system is supposed to recognize) for which there is very little available training data (Imseng et al., 2012, 2011). Further, we showed that Kullback-Leibler divergence based hidden Markov models (KL-HMMs) are very powerful when only small amounts of training data are available (Imseng et al., 2012, 2011). A KL-HMM is an HMM that uses a categorical distribution as its state emission distribution. The name comes from the fact that a Kullback-Leibler divergence based distance measure is employed. More specifically, each state of the HMM is modeled with a categorical distribution and phoneme posterior probabilities given the acoustics serve as features. The categorical distributions can be trained with a Viterbi segmentation optimization algorithm.

State-of-the-art Automatic Speech Recognition (ASR) systems typically employ context dependent modeling to better take into account the canonical-to-surface form variability of pronunciation inherent to acoustic modeling. Such context dependent modeling most commonly takes the form of the triphone whose representation comprises a phone along with its preceding and following phone context. In creating triphone (or higher order) context models we immediately run into the problem of sparsity of the training data, since many triphone contexts will occur infrequently or not at all. To overcome this, the decision tree clustering approach (Young et al., 1994) was introduced in which states of context dependent models are tied (thereby sharing data) according to shared properties (usually phonological) and by greedy optimization of a given criterion (usually maximum likelihood). An additional property of this approach is that it also permits the synthesis of contexts that were unseen in the training data.

However, no such decision tree clustering algorithms have been available to date for the KL-HMM framework. Therefore, in previous work, we used a back-off strategy during decoding where unseen triphones were modeled by the monophone model of the center phoneme (Imseng et al., 2012, 2011). In this paper, we present an algorithm that allows us to perform decision tree clustering for KL-HMM based ASR systems. Further we will also compare the KL-HMM system to the subspace Gaussian mixture modeling (SGMM) technique (Povey et al., 2010).

SGMMs have shown real potential for multilingual modeling (Burget et al., 2010). In case of conventional acoustic models (i.e., context-dependent triphones), the distribution of each HMM state is represented by relatively large number of parameters completely defining a Gaussian Mixture Model (GMM). The SGMM approach exploits GMMs as the underlying state distribution as well. However, for each specific HMM state, the high-dimensional super vector which is compounded from all the GMM parameters (i.e., only mean vectors and mixture component weights) is constrained to operate in a relatively low dimensional subspace.

Evaluation of this work was carried out using SpeechDat(II) data from five European languages as available multilingual information/data and the Greek SpeechDat(II) database as representative of an unseen language with little available data. Results reveal that the proposed decision tree algorithm allows KL-HMM to work best for small amounts of data and that the SGMMs are superior for larger amounts.

In Section 2 we introduce the decision tree clustering approach for KL-HMM, Section 3 presents the data that we used and the different systems are described in Section 4. The results follow in Section 5 before Section 6 concludes the paper.

2 Decision tree clustering for KL-HMM

We first briefly present the standard likelihood based decision tree clustering in Section 2.1. Then, we introduce the novel algorithm for KL-HMMs in Section 2.2.

2.1 Likelihood based decision criterion

Suppose that we have a set of states $\mathcal{S}$ that we wish to tie using the standard decision tree method (Young et al., 1994) such that at the parent node we have a set of questions $q \in Q$. Then each question can split $\mathcal{S}$ into two non-overlapping sub-sets $\mathcal{S}_y(q)$ and $\mathcal{S}_n(q)$, where subscripts y and n indicate the binary split that separates the set into *yes* and *no* responses to question q.

Given the following assumptions:

- The assignments of observations to states are not altered during the clustering procedure.
- The contribution of the transition probabilities to the total likelihood can be ignored.
- The total likelihood of the data can be approximated by a simple average of the log likelihoods weighted by the probability of state occupancy.

the splitting criterion can be approximated as (Young et al., 1994) :

$$L(\mathcal{S}) \simeq -\frac{1}{2}(\log[(2\pi)^K|\Sigma(\mathcal{S})|] + K) \sum_{s\in\mathcal{S}} \sum_{f\in F} \gamma_s(o_f) \tag{1}$$

where for training data pooled in set of states $s \in \mathcal{S}$; $L(\mathcal{S})$ is the log-likelihood, $\Sigma(\mathcal{S})$ is the variance of data in the set of states $\mathcal{S}$, F is the set of frames in the training data and $\gamma_s(o_f)$ is the posterior probability of state s for acoustic observation vector o_f. Assuming hard occupation decision for states, i.e. $\tilde{s} = \text{argmax}_s\ \gamma_s(o_f) :\ \gamma_{\tilde{s}} = 1, \gamma_{s\neq\tilde{s}\in\mathcal{S}} = 0$, we can further simplify (1):

$$L(\mathcal{S}) \simeq -\frac{1}{2}(\log[(2\pi)^K|\Sigma(\mathcal{S})|] + K) \sum_{s\in\mathcal{S}} N(s) \tag{2}$$

where $N(s)$ is the number of times that state s is observed in the training data.

Since questions split $\mathcal{S}$ into two non-overlapping sub-sets $\mathcal{S}_y(q)$ and $\mathcal{S}_n(q)$ at each node, we can choose the question q that maximizes the likelihood difference $\Delta L(q|\mathcal{S})$:

$$\Delta L(q|\mathcal{S}) = L(\mathcal{S}_y(q)) + L(\mathcal{S}_n(q)) - L(\mathcal{S})$$

To avoid over-fitting, the stopping criterion is usually based on a combination of minimum cluster occupancy and minimum increase in log-likelihood threshold. The latter can automatically be determined with the minimum description length (MDL) criterion (Shinoda and Watanabe, 1997).

It is evident from these equations that the likelihood does not depend on the training observations themselves but merely on the variance over training data corresponding to the states (which can be calculated from the state pdfs) and the state occupancy statistics. In the remainder of this section we show that a similar derivation exists for systems that use a Kullback-Leibler divergence based cost function to perform ASR.

2.2 Kullback-Leibler based decision criterion

Recent ASR studies have shown that Kullback-Leibler (KL) divergence based hidden Markov models (KL-HMMs) are very powerful when only small amounts of training data are available (Imseng et al., 2012). A KL-HMM is an HMM that uses a KL-divergence based cost function[1]. More specifically, each state s of the HMM is modeled with a categorical distribution y_s and phoneme posterior probabilities given the acoustics of time t, z_t serve as features. The categorical distributions can be trained with a Viterbi segmentation optimization algorithm, but it is not evident how to tie states with a decision tree. Therefore, we propose a modified version of the likelihood-based decision tree framework presented in Section 2.1.

Amongst different KL-divergence based cost-functions, usually the symmetric one performs best for recognition (Imseng et al., 2011). However, unfortunately, for the clustering algorithm that we propose, there is no closed form solution for the symmetric KL-divergence and we use the asymmetric KL-divergence between observed posterior vector, z_t, and state posterior vector, y_s, defined as:

$$D_{KL}(y_s \| z_t) = \sum_{k=1}^{K} y_s(k) \log \frac{y_s(k)}{z_t(k)} \tag{3}$$

where $k \in \{1 \dots K\}$ is the dimensionality index of the posterior distribution vector. The KL-divergence is always non-negative and zero if and only if the observed posterior vector and the state posterior vector are equal, i.e.:

$$D_{KL}(y_s \| z_t) \geqslant 0 \text{ and } D_{KL}(y_s \| z_t) = 0 \text{ iff } y_s = z_t$$

Hence, instead of maximizing the likelihood, we propose to minimize the KL-divergence:

$$D_{KL}(\mathcal{S}) = \sum_{s \in \mathcal{S}} \sum_{f \in F(s)} \sum_{k=1}^{K} y_{\mathcal{S}}(k) \log \frac{y_{\mathcal{S}}(k)}{z_f(k)} \tag{4}$$

where $\mathcal{S}$ is a set of states s and $F(s)$ the set of training vectors corresponding to state s. The state posterior vector associated with the set $\mathcal{S}$, $y_{\mathcal{S}}$, can be calculated as follows (Aradilla et al., 2007):

$$y_{\mathcal{S}}(k) = \frac{\tilde{y}_{\mathcal{S}}(k)}{Y_{\mathcal{S}}} = \frac{\left[\prod_{s \in \mathcal{S}} \prod_{f \in F(s)} z_f(k)\right]^{\frac{1}{N(\mathcal{S})}}}{\sum_{k=1}^{K} \tilde{y}_{\mathcal{S}}(k)} \tag{5}$$

with $N(\mathcal{S})$ being the number of frames associated to the set $\mathcal{S}$ and $Y_{\mathcal{S}}$ acting as a normalization factor.

The unnormalized state posterior associated with a single state s, $\tilde{y}_s(k)$, can be written as (Aradilla et al., 2007):

$$\tilde{y}_s(k) = \left[\prod_{f \in F(s)} z_f(k)\right]^{\frac{1}{N(s)}} \tag{6}$$

Combining (5) and (6) (Imseng and Dines, 2012):

$$\tilde{y}_{\mathcal{S}}(k) = \left[\prod_{s \in \mathcal{S}} (y_s(k) \cdot Y_s)^{N(s)}\right]^{\frac{1}{\sum_{s \in \mathcal{S}} N(s)}} \tag{7}$$

Hence we can express $y_{\mathcal{S}}(k)$ based on y_s, Y_s and $N(s)$, thus without having access to the individual observations z_f.

Further expanding (4) and simplifying leads to (Imseng and Dines, 2012):

$$D_{KL}(\mathcal{S}) = -\sum_{s \in \mathcal{S}} N(s) \log \sum_{k=1}^{K} \tilde{y}_{\mathcal{S}}(k) \tag{8}$$

[1] Kullback and Leibler originally introduced the *discrimination information* (Kullback and Leibler, 1951) that is nowadays often referred to as *Kullback-Leibler distance* or as a KL-divergence because it is not a metric.

Thus, the KL divergence of a set of states $\mathcal{S}$, $D_{KL}(\mathcal{S})$, can be calculated based on the statistics y_s, Y_s and $N(s)$ of the individual states.

For the splitting of a set of states $\mathcal{S}$, we propose to choose the question that maximizes the KL-divergence difference $\Delta D_{KL}(q|\mathcal{S})$:

$$\Delta D_{KL}(q|\mathcal{S}) = D_{KL}(\mathcal{S}) - (D_{KL}(\mathcal{S}_y(q)) + D_{KL}(\mathcal{S}_n(q)))$$

to minimize D_{KL}. Identically to the likelihood based decision tree, the stopping criterion can be based on a combination of minimum cluster occupancy and minimum decrease in the cost function threshold. But, in contrast to the likelihood based tree, it is not evident how to determine the latter automatically.

3 Database

For this study, we used data from the SpeechDat(II) databases. We used *corpus S*, which contains ten read sentences per speaker.

3.1 Source languages

We used the data of five European languages, namely British English, Italian, Spanish, Swiss French and Swiss German as source languages. As we will see in Section 4, we exploited the multilingual data in several different ways. For that purpose, a universal phoneme set was built by merging phonemes that share the same symbol across languages. The universal phoneme set consists of 116 SAMPA[2] phonemes and silence.

In total, there are 63 hours of SpeechDat(II) training data in this five languages, uttered by 7500 speakers (1500 per language).

3.2 Target language

In this study, Greek was the target language. The Greek SpeechDat(II) database contains a relatively large amount of data that was split into training (1500 speakers), development (150 speakers) and testing (350 speakers) sets as we already described in (Imseng et al., 2010).

To simulate limited resources, we continuously reduced the amount of available data by picking a subset of utterances for both the training and the development set. The amount of training data varied from 13.5 hours to 5 minutes. We did not change the test set and all the systems were evaluated on the same set. The test sentences use 10k different words.

Since we have no access to an appropriate language model, we simply built two different language models: one with all the sentences from the development set and one with all the sentences from the test set. These language models have perplexities of 43 and 44 respectively. The development language model was used during the parameter tuning (language scaling factor and word insertion penalty) on the development set and the test language model was used during the evaluation. In this sense, results should be considered as optimistic.

4 System description

In total we compared five systems. As baseline, we used a standard HMM/GMM system only trained on data from the target language. The remaining four systems were trained on multilingual data. Two systems were based on mixtures of Gaussian distributions and two on categorical distributions (KL-HMM).

4.1 Mixtures of Gaussians distributions

All the Gaussian systems used 39 Mel-Frequency Perceptual Linear Prediction (MF-PLP) features ($C_0 - C_{12} + \Delta + \Delta\Delta$), extracted with the HTS variant[3] of the HTK toolkit.

[2] http://www.phon.ucl.ac.uk/home/sampa/grk-uni.htm

[3] http://hts.sp.nitech.ac.jp/

4.1.1 Monolingual HMM/GMM system

The baseline, a conventional HMM/GMM system, was trained only on the available Greek data. The system based on context dependent phonemes (triphones) was trained from the MF-PLP features with the HTS toolkit. The triphone models were tied with the help of a decision tree that was based on the minimum description length criterion. The tied triphone models were then modeled with 2, 4, 8 and 16 Gaussian mixtures with diagonal covariance. Depending on the available amount of training data, the optimal choice for the number of Gaussians varied and was tuned on the development set.

4.1.2 Maximum likelihood linear regression

To evaluate whether the new language could be accommodated by linear transforms, we first trained a triphone HMM/GMM system on the multilingual data (using the universal phoneme set). Each triphone was modeled with 16 Gaussians. Then, we applied the standard maximum likelihood linear regression (MLLR) and used a regression tree that allowed up to 32 regression classes to adapt the universal models to the target language. Since not all the Greek phonemes were present in the universal phoneme set, we needed to map the palatal plosives c and ɟ to the velar plosives k and ɡ respectively.

4.1.3 Subspace Gaussian mixture models

Recently, a new acoustic modeling technique based on Subspace Gaussian Models (SGMMs) (Povey et al., 2010) has been proposed and applied in a multilingual framework (Burget et al., 2010). In our experimental work, we first trained a Universal Background Model (UBM) of GMMs using the data of the source languages. Then, the UBM was used to initialize the SGMM model. Finally, the rest of SGMM parameters (i.e., mean and weight projections, variances and state-specific parameters) was trained in a SGMM framework. As it was the case in monolingual HMM/GMM system, the choice of parameters (especially the size of state-specific vectors and total number of sub-states) varied depending on the available amount of training data. The number of parameters was fixed to be approximately similar to the HMM/GMM system. We used a little higher number (+10%) of SGMM sub-states than total number of Gaussians in the HMM/GMM system. The sub-space dimensions were set to be reasonably high according to the availability of the training data.

In our work with SGMMs, we also performed experiments to train all other than state-specific parameters with the data of the source languages, as proposed in (Burget et al., 2010). However, such SGMM configurations performed significantly worse.

4.2 KL-HMM

Both Kullback-Leibler divergence based HMM systems used universal phoneme posterior probabilities as features. The universal phoneme posterior probabilities were estimated with a multilingual Multilayer Perceptron (MLP) that was previously trained with the data of the source languages. For the training of the KL-HMM parameters, the Greek MF-PLP features were forward passed through the MLP to obtain universal phoneme posterior probabilities.

4.2.1 KL-HMM BO

The standard KL-HMM system was based on triphones. Since no decision tree was available, we limited ourselves to word-internal triphones only (as opposed to cross-word triphones for all the other systems). During decoding, we backed off (BO) to the context independent model of the center phoneme if a triphone was not seen during training. Each triphone was modeled with three states.

4.2.2 KL-HMM tree

The second KL-HMM system used the proposed decision tree approach and was therefore based on cross-word triphones. The total number of states was tuned on the development set and was usually higher than for the HMM/GMM system. However, the total number of parameters was still lower than for the HMM/GMM system because each state was modeled with one categorical distribution instead of a mixture of Gaussians.

5 Results

We evaluated all five systems presented in Section 4. We hypothesize, that the proposed KL-HMM system with decision tree outperforms all other approaches for very low amounts of data. Furthermore, we expect the SGMM system to perform best for larger amounts of data.

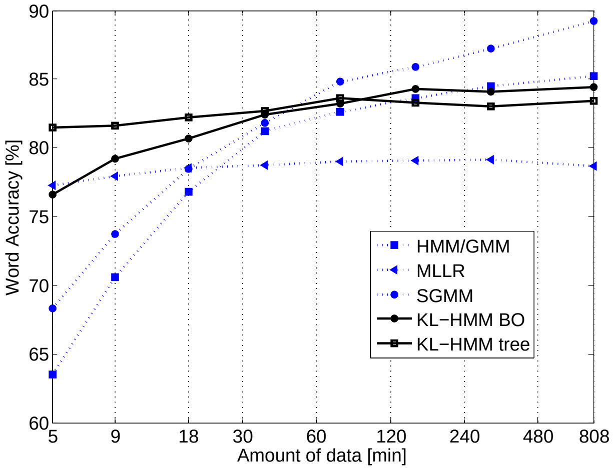

Figure 1: *Word accuracies for Greek ASR. The different systems are described in Section 4. Dashed curves represent GMM-based systems and solid ones KL-HMM-based systems.*

Figure 1 shows the results. The KL-HMM system with tree performs best for very low amounts of data. If there is less than 30 minutes of training data, the tree-based KL-HMM system significantly outperforms all other systems. It is remarkable that system *KL-HMM tree* reaches a performance of more than 81% word accuracy, if only five minutes of Greek training data are available. The overall behavior of the KL-HMM system with tree and the MLLR system are similar (almost flat), but the tree-based KL-HMM system performs about 4% absolute better.

If there is more than about an hour of data, the KL-HMM system without a tree performs slightly better than the KL-HMM system with tree. We believe that this is due to the mismatch of the cost functions during decoding and decision tree clustering. As already mentioned, we used the symmetric KL-divergence during decoding and the asymmetric version given in (3) for the decision tree clustering.

Furthermore, for more than one hour of data, the SGMM system reveals its potential. Whereas the standard HMM/GMM system performs only marginally better than the KL-HMM systems, the SGMM system reaches a word accuracy of 89% if all the Greek training data is used. Hence, both of our hypotheses were confirmed by these experiments.

6 Conclusions

In this paper, we proposed and evaluated an adapted version of decision tree state clustering for KL-MM systems. For the evaluation, we used multilingual data from five source languages to boost the performance of a Greek speech recognizer and simulated low-resource scenarios by restricting the amount of Greek training data.

The tree-based KL-HMM system successfully exploits multilingual information in the form of universal phoneme posterior features and outperforms all other systems for very low amounts of data (less than one hour). The SGMM system with a UBM trained on the source languages was shown to be superior for larger amounts of Greek training data.

References

G. Aradilla, J. Vepa, and H. Bourlard. An acoustic model based on Kullback-Leibler divergence for posterior features. In *Proc. of ICASSP*, pages IV–657–660, 2007.

L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát,

D. Povey, A. Rastrow, R. C. Rose, and S. Thomas. Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models. In *Proc. of ICASSP*, pages 4334–4337, 2010.

David Imseng and John Dines. Decision tree clustering for KL-HMM. Technical Report Idiap-Com-01-2012, Idiap Research Institute, 2012.

David Imseng, Hervé Bourlard, and Mathew Magimai.-Doss. Towards mixed language speech recognition systems. In *Proc. of Interspeech*, pages 278–281, 2010.

David Imseng, Ramya Rasipuram, and Mathew Magimai.-Doss. Fast and flexible Kullback-Leibler divergence based acoustic modeling for non-native speech recognition. In *Proc. of ASRU*, pages 348–353, 2011.

David Imseng, Hervé Bourlard, and Philip N. Garner. Using KL-divergence and multilingual information to improve ASR for under-resourced languages. In *Proc. of ICASSP*, pages 4869–4872, 2012.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22 (1):79–86, 1951.

D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A.Rastrow, R. C. Rose, P. Schwarz, and S. Thomas. Subspace gaussian mixture models for speech recognition. In *Proc. of ICASSP*, pages 4330–4333, 2010.

Koichi Shinoda and Takao Watanabe. Acoustic modeling based on the MDL principle for speech recognition. In *Proc. of Eurospeech*, pages I –99–102, 1997.

S. J. Young, J. J. Odell, and P. C. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the workshop on Human Language Technology*, pages 307–312, 1994.