IDIAP COMMUNICATION REPORT

DECISION TREE CLUSTERING FOR KL-HMM

David Imseng John Dines

Idiap-Com-01-2012

FEBRUARY 2012

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

Decision tree clustering for KL-HMM

David Imseng, John Dines

February 3, 2012

Abstract

Recent Automatic Speech Recognition (ASR) studies have shown that Kullback-Leibler diverge based hidden Markov models (KL-HMMs) are very powerful when only small amounts of training data are available. However, since the KL-HMMs use a cost function that is based on the Kullback-Leibler divergence (instead of maximum likelihood), standard ASR algorithms such as the commonly used decision tree clustering are not applicable in general. In this communication, we present an algorithm that allows us to perform decision tree clustering for KL-HMM based ASR systems.

1 Introduction

State-of-the-art Automatic Speech Recognition (ASR) systems typically employ context dependent modeling in order 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. In order to overcome this, the decision tree clustering approach (Odell, 1995) 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.

Recent ASR studies have shown 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., 2011). A KL-HMM is an HMM that uses a Kullback-Leibler divergence based cost function. 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.

The usage of a KL-HMM based ASR system has several advantages (Imseng et al., 2011):

- Transfer learning: the posterior feature estimator may be trained on auxiliary data.
- Choice of posterior feature space: the posterior feature space may be phonemes that are specific to a language, universal phonemes, or articulatory features, or any other posterior features that may be relevant to the classification task.
- Fewer number of parameters: this suggests that KL-HMM systems may require less training data than conventional models.

However, decision tree clustering algorithms for the KL-HMM framework to date have been available. Therefore, in previous work, we used a back-off strategy during decoding and modeled unseen triphones during decoding with the monophone model of the center phoneme (Imseng et al., 2011). In this communication, we present an algorithm that allows us to perform decision tree clustering for KL-HMM based ASR systems.

We first briefly present the standard likelihood based decision tree clustering in Section 2. Then, we introduce the novel algorithm for KL-HMMs in Section 3.

2 Likelihood based decision criteria

Suppose that we have a set of states $\mathcal{S}$ that we wish to tie using the decision tree method of Odell (Odell, 1995; 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.

Odell formulated a likelihood based decision criteria (Odell, 1995, chapter 3.7.1) and proposed a computationally efficient algorithm, based on 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.

Given these assumptions, the splitting criterion can be approximated as:

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

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):

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

where $F(s)$ is the set of training vectors corresponding to state s and $N(s)$ is the number of times that state s is observed in the training data.

As already mentioned, each question can split $\mathcal{S}$ into two non-overlapping sub-sets $\mathcal{S}_y(q)$ and $\mathcal{S}_n(q)$. Hence, at each node, we need to choose a question q that maximizes the likelihood difference $\Delta L(q|\mathcal{S})$:

$$\Delta L(q|\mathcal{S}) = (L(\mathcal{S}_y(q)) + L(\mathcal{S}_n(q))) - L(\mathcal{S}) \tag{3}$$

To avoid over-fitting, the stopping criteria 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 document we show that a similar derivation exists for systems that use a Kullback-Leibler divergence based cost function to perform ASR.

3 Kullback-Leibler based decision criteria

In this section, we develop a decision tree clustering algorithm that is based on the Kullback-Leibler (KL) divergence.

The KL-divergence between observed posterior vector, z_t, and state posterior vector, y_s, is defined as:

$$D_{KL}(y_s||z_t) = \sum_{k=1}^{K} y_s(k) \log \frac{y_s(k)}{z_t(k)} \tag{4}$$

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:

$$D_{KL}(\mathcal{S}) = \sum_{s \in \mathcal{S}} \sum_{f \in F(s)} \sum_{k=1}^{K} y_{\mathcal{S}}(k) \log \frac{y_{\mathcal{S}}(k)}{z_f(k)} \tag{5}$$

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:

$$y_{\mathcal{S}}(k) = \frac{\tilde{y}_{\mathcal{S}}(k)}{Y_{\mathcal{S}}} \tag{6}$$

where (Aradilla et al., 2007):

$$\tilde{y}_{\mathcal{S}}(k) = \left[\prod_{s \in \mathcal{S}} \prod_{f \in F(s)} z_f(k) \right]^{\frac{1}{N(\mathcal{S})}} \tag{7}$$

$$Y_{\mathcal{S}} = \sum_{k=1}^{K} \tilde{y}_{\mathcal{S}}(k) \tag{8}$$

with $N(\mathcal{S})$ being the number of frames associated to the set $\mathcal{S}$. We can further develop (7):

$$\tilde{y}_{\mathcal{S}}(k) = \left[\prod_{s \in \mathcal{S}} \tilde{y}_s(k)^{N(s)} \right]^{\frac{1}{N(\mathcal{S})}} \tag{9}$$

where $N(s)$ is the number of times that state s is observed in the training data. Combining (6) and (9), leads to

$$\tilde{y}_{\mathcal{S}}(k) = \left[\prod_{s \in \mathcal{S}} (y_s(k) \cdot Y_s)^{N(s)} \right]^{\frac{1}{\sum_{s \in \mathcal{S}} N(s)}} \tag{10}$$

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 (5) leads to:

$$\begin{aligned} D_{KL}(\mathcal{S}) &= \sum_{s \in \mathcal{S}} \sum_{f \in F(s)} \sum_{k=1}^{K} [y_{\mathcal{S}}(k) \log y_{\mathcal{S}}(k) - y_{\mathcal{S}}(k) \log z_f(k)] \\ &= \sum_{s \in \mathcal{S}} N(s) \sum_{k=1}^{K} y_{\mathcal{S}}(k) \log y_{\mathcal{S}}(k) - \sum_{s \in \mathcal{S}} \sum_{f \in F(s)} \sum_{k=1}^{K} y_{\mathcal{S}}(k) \mathrm{log} z_f(k) \\ &= N(\mathcal{S}) \sum_{k=1}^{K} y_{\mathcal{S}}(k) \log y_{\mathcal{S}}(k) - \sum_{s \in \mathcal{S}} \sum_{f \in F(s)} \sum_{k=1}^{K} y_{\mathcal{S}}(k) \mathrm{log} z_f(k) \end{aligned} \tag{11}$$

The second term of (11) can be simplified:

$$\begin{aligned}
\sum_{s\in\mathcal{S}}\sum_{f\in F(s)}\sum_{k=1}^{K} y_{\mathcal{S}}(k)\mathrm{log} z_f(k) &= \sum_{k=1}^{K} y_{\mathcal{S}}(k)\sum_{s\in\mathcal{S}}\sum_{f\in F(s)} \mathrm{log} z_f(k)\\
&= \sum_{k=1}^{K} y_{\mathcal{S}}(k)\log\prod_{s\in\mathcal{S}}\prod_{f\in F(s)} z_f(k)\\
&= \sum_{k=1}^{K} y_{\mathcal{S}}(k)N(\mathcal{S})\log\tilde{y}_{\mathcal{S}}(k)\\
&= N(\mathcal{S})\sum_{k=1}^{K} y_{\mathcal{S}}(k)\log\left[y_{\mathcal{S}}(k)\cdot Y_{\mathcal{S}}\right]\\
&= N(\mathcal{S})\sum_{k=1}^{K} y_{\mathcal{S}}(k)\left[\log y_{\mathcal{S}}(k)+\log Y_{\mathcal{S}}\right] \qquad (12)
\end{aligned}$$

(11) and (12) yield:

$$\begin{aligned}
D_{KL}(\mathcal{S}) &= N(\mathcal{S})\sum_{k=1}^{K} y_{\mathcal{S}}(k)\log y_{\mathcal{S}}(k) - N(\mathcal{S})\sum_{k=1}^{K} y_{\mathcal{S}}(k)\left[\log y_{\mathcal{S}}(k)+\log Y_{\mathcal{S}}\right]\\
&= -N(\mathcal{S})\sum_{k=1}^{K} y_{\mathcal{S}}(k)\log Y_{\mathcal{S}}\\
&= -N(\mathcal{S})\log Y_{\mathcal{S}}\sum_{k=1}^{K} y_{\mathcal{S}}(k) \qquad (13)
\end{aligned}$$

Since by definition: $\sum_{k=1}^{K} y_{\mathcal{S}}(k) = 1$, (13) further simplifies:

$$D_{KL}(\mathcal{S}) = -N(\mathcal{S})\log Y_{\mathcal{S}} \qquad (14)$$

Combining (8), (10) and (14) leads to:

$$D_{KL}(\mathcal{S}) = -\sum_{s\in\mathcal{S}} N(s)\log\sum_{k=1}^{K}\left[\prod_{s\in\mathcal{S}}(y_s(k)\cdot Y_s)^{N(s)}\right]^{\frac{1}{\sum_{s\in\mathcal{S}}N(s)}} \qquad (15)$$

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 $\Delta D_{KL}(q|\mathcal{S})$:

$$\Delta D_{KL}(q|\mathcal{S}) = D_{KL}(\mathcal{S}) - (D_{KL}(\mathcal{S}_y(q)) + D_{KL}(\mathcal{S}_n(q))) \qquad (16)$$

in order to minimize D_{KL}. Identically to the likelihood based decision tree, the stopping criteria 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.

In literature (Aradilla, 2008), KL-divergence related measures such as reverse-KL and symmetric-KL have also been proposed. Unfortunately, there is no closed form solution for D_{KL}, $Y_{\mathcal{S}}$ and $y_{\mathcal{S}}$ that is independent of the individual observations z_f for the reverse-KL and symmetric-KL cases.

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.

Guillermo Aradilla. *Acoustic Models for Posterior Features in Speech Recognition*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2008.

David Imseng, Ramya Rasipuram, and Mathew Magimai.-Doss. Fast and flexible kullback-leibler divergence based acoustic modeling for non-native speech recognition. In *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, 2011.

Julian James Odell. *The Use of Context in Large Vocabulary Speech Recognition*. PhD thesis, University of Cambridge, 1995.

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.