Probabilistic State Clustering Using Conditional Random Field For Context-Dependent Acoustic Modelling

Khe Chai SIM

School of Computing, National University of Singapore, Singapore

Abstract

Hidden Markov Models are widely used in speech recognition systems. Due to the co-articulation effects of continuous speech, context-dependent models have been found to yield performance improvements. One major issue with context-dependent acoustic modelling is the robust parameter estimation of unseen or rare models in the training data. Typically, decision tree state clustering is used to ensure that there are sufficient data for each physical state. Decision trees based on phonetic questions are used to cluster the states. In this paper, conditional random field (CRF) is used to perform probabilistic state clustering where phonetic questions are used as binary feature functions to predict the latent cluster weights. Experimental results on the Wall Street Journal reveals that CRF-based state clustering outperformed the conventional maximum likelihood decision tree state clustering with similar model complexities by about 10% relative.

Index Terms: state clustering, conditional random field, complexity control

1. Introduction

State-of-the-art automatic speech recognition typically represents phonemes using Hidden Markov Models. Co-articulation is a common phenomenon in continuous speech, where the acoustic characteristic of a phoneme is influenced by the preceding and succeeding phonemes. This phenomenon causes greater variability, which eventually results in poorer recognition performance. This motivated context-dependent modelling of the phonemes. Typically, triphone models are considered, where the identities of the phone immediately preceding and succeeding the current phone are used as contexts. Unfortunately, context dependent modelling leads to a prohibitively large number of distinct models, rendering it impractical both in terms of efficient computation and reliable parameter estimation. For example, a system with $N$ context independent phones yields $N^3$ triphone models. The issue becomes more apparent when wider contexts are used.

A typical solution to mitigate the problem of having an exorbitantly large number of context dependent models is to employ parameter tying schemes to achieve a compromise between accurate modelling and the amount of training data available. Decision tree state clustering technique is commonly used to tie sufficient data for each physical state. Decision trees based on phonetic questions are used to cluster the states. In this paper, conditional random field (CRF) is used to perform probabilistic state clustering where phonetic questions are used as binary feature functions to predict the latent cluster weights. Experimental results on the Wall Street Journal reveals that CRF-based state clustering outperformed the conventional maximum likelihood decision tree state clustering with similar model complexities by about 10% relative.

This paper proposes a novel approach of probabilistic state clustering, where soft assignments are used to map the context-dependent states probabilistically to a set of latent clusters. This allows modelling of a larger number of distinct states without a dramatic increase in model complexity, since the number of free model parameters is predominantly governed by the number of latent clusters, instead. The latent clusters are obtained via unsupervised clustering using Gaussian Mixture Models (GMMs). Probabilistic assignments of context dependent states to these latent clusters are estimated using Conditional Random Field (CRF) [3], where phonetic questions similar to those used in the conventional decision tree clustering are adopted as binary feature functions for CRF learning. The resulting CRF model is used to obtain the probabilistic assignment weights of all the states, including those which are not seen in training, based on the binary phonetic features derived from it. Since the CRF model may generate assignment weights for unseen states which are different from all the assignment weights of the seen states, CRF-based state clustering allows unseen models to be "synthesised" in a latent continuous space. Therefore, it is expected to be able to make better use of the model parameters, which leads to a compact model representation.

The remaining of this paper is organised as follows. Section 2 formulates the state clustering problem, which forms the fundamental of the subsequent discussions in this paper. Section 3 describes the decision tree clustering approach. Section 4 introduces the proposed CRF-based clustering method. Experimental results are presented in Section 5.

2. State Clustering

Hidden Markov Models (HMMs) are commonly used to model phonemes for speech recognition. HMMs are finite state transducers whose model parameters are given by the state transition probabilities and state observation probabilities. For continuous density HMMs, state observation probability distributions are typically modelled using a Gaussian Mixture Model (GMM):

$$p(o_t|s) = \sum_{m=1}^M p(o_t|s, m)P(m|s)$$

(1)

where $p(o_t|s, m)$ is a Gaussian distribution of component $m$ and state $s$ generating the observation vector, $o_t$ and $P(m|s)$ is the probability of component $m$ given state $s$. $M$ is the number of Gaussian components per state. For context-dependent modelling, the state observation probability distributions in Equa-
tion 1 may be rewritten as:

\[ p(\alpha_i | s_c) = \sum_{m=1}^{M} p(\alpha_i | s_c, m) P(m | s_c) \]  

(2)

where \( s_c \) denotes a context-dependent state. \( s \in S \) is the corresponding context-independent state and \( c \in C \) indicates the context. \( S \) and \( C \) are the sets of context-independent states and the possible contexts, respectively. The total number of unclustered states is given by \( |S| \times |C| \), where \(|\cdot|\) denotes the set cardinality.

Triphones, \( |C| = |S|^2 \). Typically, \(|S| \approx 40 \) for English, which leads to approximately \( 40^2 \) unclustered triphone states. This is prohibitively large for reliable model parameter estimation. State clustering provides a simple solution to reduce the complexity of context-dependent models. Let \( G \) defines a set of latent clusters. Then, Equation 2 can be expanded as:

\[ p(\alpha_i | s_c) = \sum_{g=1}^{p(g)} \sum_{m=1}^{M} p(\alpha_i | s_c, g, m) P(g, m | s_c) \]  

(3)

where the probability of a latent cluster \( g \) of a context-independent state \( s \) generating \( \alpha_i \) is given by a GMM:

\[ p(\alpha_i | s_c, g, m) = \sum_{g=1}^{p(g)} \sum_{m=1}^{M} p(\alpha_i | s_c, g, m) P(m | s_c, g) \]  

(4)

Equation 3 indicates a probabilistic clustering of \( s_c \), where \( P(g | s_c) \) denotes the non-negative assignment weights of state \( s_c \) to the latent cluster \( g \) such that \( \sum_{g \in G} P(g | s_c) = 1 \).

When hard assignments are used, the context-dependent states, \( s_c \), are clustered into disjoint sets such that \( \cup_{g \in G} g \) forms a set of all the context-dependent states and \( \cap_{g \in G} g = \emptyset \). In this case, the assignments are given by

\[ P(g | s_c) = \left\{ \begin{array}{ll} 1 & s_c \in g \\ 0 & s_c \notin g \end{array} \right. \]  

(5)

Note that state clustering based on hard assignment is employed by the conventional decision tree clustering scheme, which will be described in the next section.

### 3. Decision Tree Clustering

Decision tree state clustering \([1, 4]\) uses a decision tree to perform top-down clustering with hard assignment. Typically, one decision tree is constructed for each context-independent state, \( s \). Initially, all the context-dependent states corresponding to \( s \) are clustered together at the root node. The terminal nodes of the tree are split recursively while there are sufficient 'state occupancy counts' associated with them. A phonetic question is chosen from a set of questions to split the nodes such that the likelihood increase resulted from the split is maximised. Let \( S_0 \) denote the set of states prior to the split and \( S_1 = S_0 \cap \Sigma \) and \( S_2 = S_0 \setminus \Sigma \) denote the two sets of states after the split using question, \( q \). If each state \( s \) is represented by a single Gaussian distribution, \( N(\mu_s, \Sigma) \), the increase in likelihood due to a split using \( q \) is given by \([1]\):

\[ \Delta(q) = L(S_1) + L(S_2) - L(S_0) \]  

(6)

where

\[ L(S) = -\frac{1}{2} \log \left( (2\pi)^d |\Sigma(S)| \right) \sum_{s \in S} \beta_s \]  

(7)

\( \beta_s \) is the occupancy count for state \( s \) and \( \Sigma(S) \) is the pooled state variance of all the states in \( S \).

### 4. Conditional Random Field Clustering

Conditional Random Field (CRF) clustering uses CRF to predict the probabilistic clustering weights for context dependent states. A schematic diagram showing the CRF-based state clustering process is depicted in Figure 1. It consists of two main stages: 1) Unsupervised clustering of acoustic observations using GMM; 2) Prediction of clusters using CRF based on binary phonetic features. These two stages will be described further in the following sub-sections.

#### 4.1. Unsupervised Acoustic Feature Clustering

In Section 2, the formulation of probabilistic state clustering was presented. Closer examination reveals that the probabilistic clustering given in Equation 3 is equivalent to a semi-continuous HMM system (or a tied-mixture system) if \( p(\alpha_i | g, s) \) is modelled as a single Gaussian distribution. The set of clusters, \( \{g\} \), corresponds to the tied-mixture components and \( P(g | s_c) \) are the corresponding component weights.

Firstly, the latent clusters may be found using unsupervised clustering techniques. In this paper, Gaussian Mixture Models are used to cluster the acoustic observations. One GMM is trained for each context-independent state. This can be achieved by training an \( M \)-component GMM/HMM monophone system, as given by Equation 1. A tied-mixture system can be constructed by cloning the weights of the context-independent states to context-dependent states and retraining the mixture weights, \( P(g | s_c) \), using the well-known Baum-Welch algorithm \([5]\). However, it is not possible to update weights for states which do not appear in the training data. To overcome this problem, phonetic questions, similar to those used in the conventional decision tree state clustering, are used to predict
the clustering weights for the unseen states. Next, CRF models will be used to perform discriminative cluster prediction.

### 4.2. CRF-based Cluster Prediction

Conditional Random Field [3] is a discriminative log-linear model where the conditional probability of the class \(y\) given the input \(x\) is modelled as

\[
P(y|x) = \frac{1}{Z(x)} \exp \left( \sum_{k=1}^{K} \lambda_k f_k(y, x) \right)
\]

where \(\lambda_k\) are the model parameters which correspond to the \(k\)th feature as given by the feature function, \(f_k(y, x)\). \(K\) is the total number of features of the CRF model. Equation 8 can be used to predict the cluster weights of the context-dependent states, \(P(y|x_s)\) by using the context-dependent states as inputs \((x = s)\) and the latent clusters as the classes to be predicted \((y = g)\). The binary feature functions \(f_k(g, s)\) are \([0, 1]\) are given by the set of phonetic questions used for decision tree clustering. One CRF model is trained for each context-independent state.

The CRF models are trained using the training data pairs \((g = \hat{m}_t, s = \hat{q}_t) : \forall t\), where \(\hat{m}_t\) and \(\hat{q}_t\) are the target cluster and state of \(o_t\). These target labels can be obtained as follows:

\[
\hat{Q}_t^m = \arg\max \ p(\theta | \hat{Q}_t^m, x_t, s)
\]

\[
\hat{m}_t = \arg\max \ p(m_t | \hat{q}_t, s)
\]

where \(\hat{Q}_t^m = \{o_t : \forall t\}\) and \(\hat{Q}_t^q = \{q_t : q_t \in S, \forall t\}\) denote the observation and state sequences respectively. \(\hat{Q}_t^m\), the most likely state sequence given the GMM/HMM model, \(\theta\), can be obtained by performing a state-level Viterbi forced-alignment (c.f. Equation 9). Given the state alignment, \(\hat{Q}_t^m\), each observation, \(o_t\), can be clustered as \(\hat{m}_t\), one of the \(M\) components which yields the highest posterior probability (c.f. Equation 10). The parameters of the CRF models can be estimated by maximising the conditional log likelihood using iterative numerical optimisation [3]. Given the CRF models, the cluster weights can then be estimated using Equation 8 for all the context-dependent states (both seen and unseen).

### 4.3. Complexity Control

The CRF-based state clustered model has \(|S| \times |C| \times |G|\) free parameters associated with the cluster weights. Usually, there are still too many free parameters and further parameter tying is necessary. In this paper, context dependent states whose cluster weights shared the same rank ordering are clustered together. The resulting cluster weights are then given by

\[
\hat{P}(g|s) = \sum_{s_c \in \hat{S}} P(g|s_c)
\]

where \(\hat{S}\) denotes the set of states sharing the same weight ordering. To further control the model complexity, the tying can be performed based on the ranking of the top \(G\) clusters. Note that when \(G = 1\), CRF-based clustering reduces to a hard clustering similar to that using decision tree clustering. In this case, CRF-based clustering differs from decision tree clustering in the way phonetics questions are asked to perform clustering. As previously described in Section 3, decision tree clustering uses a hierarchy of questions to perform clustering where the questions are selected to maximise training data likelihood. On the other hand, CRF-based clustering learns a set of weights, \(\lambda_k\), for each phonetic question to determine its importance.

### 5. Experimental Results

In this section, experimental results comparing the conventional decision tree state clustering and the proposed CRF-based probabilistic clustering will be reported on the Wall Street Journal database (WSJCAM0) [6]. There are 18.3 hours of training data, comprising 9889 utterances spoken by 96 speakers. The 5k WSJ0 tasks are used for performance evaluation. The two testing data sets used are \(s_{1,dt5a}\) and \(s_{1,dt5b}\), which consist of 0.73 and 0.67 hours of speech data respectively.

### Table 1: Comparison of model complexity and WER performance of decision tree state clustered triphone systems on WSJ0 test sets.

<table>
<thead>
<tr>
<th>Number of Physical HMMs</th>
<th>Number of Distinct States</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3433</td>
<td>899</td>
<td>21.53</td>
</tr>
<tr>
<td>7804</td>
<td>1812</td>
<td>18.69</td>
</tr>
<tr>
<td>14379</td>
<td>3865</td>
<td>16.23</td>
</tr>
</tbody>
</table>

### Table 2: Comparison of model complexity of CRF-based state clustered triphone systems on trained on WSJ0

<table>
<thead>
<tr>
<th>No. of Latent Clusters</th>
<th>Top-G Ranking</th>
<th>No. of Phys. HMMs</th>
<th>No. of Distinct States</th>
</tr>
</thead>
<tbody>
<tr>
<td>936</td>
<td>1</td>
<td>3576</td>
<td>808</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>17365</td>
<td>3434</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>31235</td>
<td>9705</td>
</tr>
<tr>
<td>1872</td>
<td>1</td>
<td>8238</td>
<td>1498</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>28610</td>
<td>8514</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>42264</td>
<td>24980</td>
</tr>
<tr>
<td>3744</td>
<td>1</td>
<td>14304</td>
<td>2678</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>37126</td>
<td>16712</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>48214</td>
<td>44448</td>
</tr>
</tbody>
</table>

HMM acoustic models are trained using the Hidden Markov Model toolkit (HTK) [7]. 39 dimensional Mel Frequency Cepstral Coefficients (MFCC) consisting of 12 static coefficients, \(c_0\) energy and the first two derivative features are used as acoustic features. Word recognition is performed using a bigram full decoding followed by a trigram rescoring.

Firstly, the Word Error Rate (WER) performance results for context-dependent systems based on the conventional decision tree state clustered triphone are reported in Table 1. Three systems with varying number of distinct states after decision tree clustering are used as the baseline systems. The first two columns of Table 1 show the distinct number of physical HMMs and clustered states in the systems. There is a consistent performance improvement as the model complexity increases. The final system with about 3800 distinct states is roughly the optimum configuration where increase model complexity beyond which yields negligible improvement. These systems have a single Gaussian component per state. Hence, the total number of Gaussian components in those systems are the same as the number of distinct states.

Next, the performance of CRF-based state clustered triphone systems with different model complexities are summarised in Table 2. The CRF models are trained using the
In this paper, a novel probabilistic state clustering method was proposed where Conditional Random Field (CRF) was used to predict the probabilistic cluster weights. CRF-based clustering generalises the conventional decision tree state clustering technique by allowing each context-dependent state to be associated with multiple latent clusters in a probabilistic manner. Compared to the hard clustering approach in a decision tree clustering method, CRF-based clustering allows greater number of distinct context-dependent states to be modelled without dramatic increase in the model parameters. Furthermore, CRF is used to estimate the cluster weights of all the context-dependent states (including those that do not appear in the training data) based on the binary phonetic features derived from the phonetic questions similar to those used in the conventional decision tree clustering technique. Experimental results on the WSJ0 5k tasks showed approximately 10% relative improvements using CRF-based state clustering compared to decision tree state clustering.

### 7. References


---

Figure 2: Comparison of WER performance between decision tree and CRF clustering with different model complexities on si_dt5a (top) and si_dt5b (bottom).

---

http://crfpp.sourceforge.net/