Decision Tree-based Simultaneous Clustering of Phonetic Contexts, Dimensions, and State Positions for Acoustic Modeling

Heiga Zen, Keiichi Tokuda, Tadashi Kitamura

Department of Intelligence and Computer Science, Nagoya Institute of Technology, Nagoya, Japan
{zen,tokuda,kitamura}@ics.nitech.ac.jp

Abstract

In this paper, a new decision tree-based clustering technique called Phonetic, Dimensional and State Positional Decision Tree (PDS-DT) is proposed. In PDS-DT, phonetic contexts, dimensions and state positions are grouped simultaneously during decision tree construction. PDS-DT provides a complete distribution structure without any external control parameters. In speaker-independent continuous speech recognition experiments, PDS-DT achieved about 13%–15% error reduction over the phonetic decision tree-based state-tying technique.

1. Introduction

In large vocabulary continuous speech recognition systems, context-dependent model, typically triphone, and continuous density HMMs are often used. The use of triphones rather than monophones is known to provide higher recognition accuracy. While the large number of triphones can help to capture variations in speech data, it results in too many free-parameters in a system. Maintaining a good balance between the model complexity and model robustness is important in acoustic modeling. Therefore, various parameter clustering techniques have been proposed [1–5]. The use of Phonetic Decision Trees (P-DT) [5] is a good solution to this problem. It has two advantages over bottom-up based approaches [1–3]. First, by incorporating phonetic knowledge into questions, it can assign unseen triphones to the leaf nodes of decision trees. Second, the splitting procedure of the decision tree provides a way of keeping the balance of model complexity and robustness.

In decision tree-based acoustic modeling, a model-tying approach and state-tying approach were developed. The state-tying approach is widely used because it provides a more detailed level of sharing and outperforms the model-tying approach [5]. However, in state-tying approach, all dimensions of state output probability distribution have common sharing structure. In speech recognition, mel-cepstral coefficients (mel-cepstrum) and their time derivatives (Δmel-cepstrum) are widely used as acoustic features. Mel-cepstral coefficients in the lower quefrency range are generally considered to have more significant information than those in the higher quefrency range. Likewise, static coefficients have more significant information than their time derivatives. Therefore, assigning a greater number of distributions to the coefficients in the lower quefrency range than those in the higher quefrency range and to the static coefficients than their time derivatives may result in better recognition performance. The feature vector can be modeled more efficiently by a proper context-dependent sharing structure for each dimension of feature vector.

In [6], a clustering technique called Dimensional Split Phonetic Decision Trees was proposed. In this paper, it is referred to as Phonetic and Dimensional Decision Trees (PD-DT). It determines whether the distribution of each dimension should be split or not when applying a question. It gives a proper context-dependent sharing structure for grouped dimensions. The PD-DT is thought to adopt questions about contexts and dimensions.

Generally, a decision tree is built for each base phone in state-tying approach. However, it is possible to build a single tree by applying questions about state positions. The state-tying structure can be constructed across state positions. Recent research [7, 8] shows that the state-tying across state positions do not occur for the most of trees. Questions about state positions were applied at root node and its neighboring nodes of each decision tree, hence constructed sharing structures were almost the same as conventional decision trees for each state position of each base phone. However, each dimension could have different state-position and phonetic-context dependency.

This paper introduces questions about state positions in PD-DT, and proposes a technique for simultaneous clustering of phonetic contexts, dimensions and state positions. It is a unified technique for decision tree-based acoustic modeling based on MDL criterion [9]. Experimental results show the effectiveness of state-positional-split with dimensional-split.

In Section 2, implementation of PDS-DT is described. In Section 3, PDS-DT is evaluated in speaker-independent continuous speech recognition experiments. The last section presents conclusions and future topics.

2. Implementation of PDS-DT

This section introduces questions about state positions into PD-DT, and proposes PDS-DT.

2.1. PD-DT based clustering

In order to construct a proper context-dependent sharing structure for each dimension, Phonetic and Dimensional Decision Tree (PD-DT) is proposed [6]. In P-DT clustering based on Minimum Description Length (MDL) criterion [10], when splitting node $S$ with question $q$ into nodes $S_{q+}$ and $S_{q-}$, the change in model Description Length (DL), $\Delta_{DL}^{q}(S)$, is given by

$$\Delta_{DL}^{q}(S) = \frac{1}{2} \left[ \Gamma(S_{q+}) \log \left[ \frac{1}{2} \log \left( \frac{1}{2} \right) \right] + \Gamma(S_{q-}) \log \left[ \frac{1}{2} \log \left( \frac{1}{2} \right) \right] \right] - \Gamma(S) \log |S| + K \log \Gamma(S_{0}), \quad (1)$$

where $\Gamma(\cdot)$ is the accumulated state occupancy of each node, $\Sigma(\cdot)$ is the covariance matrix of each node, $K$ is the dimensionality of
the feature vector, and \( S_0 \) denotes the root node of the decision tree. In case of diagonal covariance, (1) can be rewritten as

\[
\Delta_{\text{DL}}^{(q)}(S, k) = \sum_{k=1}^{K} \Delta_{\text{DL}}^{(q)}(S, k),
\]

(2)

\[
\Delta_{\text{DL}}^{(q)}(S, k) = \frac{1}{2} \left[ \Gamma(S_q) \log \sigma^2(S_q, k) + \Gamma(S_{-q}) \log \sigma^2(S_{-q}, k) - \Gamma(S) \log \sigma^2(S, k) \right] + \log \Gamma(S_0),
\]

(3)

where \( \sigma^2(S_q, k) \), \( \sigma^2(S_{-q}, k) \), and \( \sigma^2(S, k) \) are the \( k \)-th elements of diagonal covariance matrices \( \Sigma(S) \), \( \Sigma(S_q) \), and \( \Sigma(S_{-q}) \), respectively. Thus, the change in DL for each dimension \( \Delta_{\text{DL}}^{(q)}(S) \) is given by the total change in DL for each dimension, \( \Delta_{\text{DL}}^{(q)}(S) \). In PD-DT, a question is applied only for the dimensions which satisfy \( \Delta_{\text{DL}}^{(q)}(S, k) < 0 \). The set of dimensions \( H^{(q)}(S) \) to be split using question \( q \), the change in DL \( \Delta_{\text{DL}}^{(q)}(S) \) for question \( q \), and the best question \( q_{\text{best}} \) for contextual split are given by

\[
H^{(q)}(S) = \left\{ k \left| k \in G(S), \Delta_{\text{DL}}^{(q)}(S, k) < 0 \right. \right\},
\]

(4)

\[
\Delta_{\text{DL}}^{(q)}(S, k) = \sum_{k \in G^{(q)}(S)} \Delta_{\text{DL}}^{(q)}(S, k),
\]

(5)

\[
q_{\text{best}} = \arg \min_{q \in Q} \Delta_{\text{DL}}^{(q)}(S),
\]

(6)

respectively, where \( G(S) \) is a set of dimensions existing in node \( S \), and \( Q \) is a set of questions.

The PD-DT based clustering shown in Fig.1 is outlined as follows:

Step 1: For all of nodes in decision tree, calculate \( \Delta_{\text{DL}}^{(\text{nodes})} \) and determine \( H^{(\text{nodes})} \).

Step 2: Choose node \( S \) which has the minimum \( \Delta_{\text{DL}}^{(\text{nodes})} \).

Step 3: Split node \( S \) dimensionally into two nodes \( S_1 \) and \( S_2 \) according to \( H^{(\text{nodes})}(S) \). \( S_1 \) is composed of dimensions which exist in \( H^{(\text{nodes})}(S) \), and \( S_2 \) is composed of dimensions which do not exist in \( H^{(\text{nodes})}(S) \) while existing in \( G(S) \).

Step 4: Split node \( S_1 \) contextually into \( S_{\text{yes}} \) and \( S_{\text{no}} \) by question \( q_{\text{best}} \).

Step 5: If \( \Delta_{\text{DL}}^{(\text{nodes})} \geq 0 \) for all of nodes, no splitting is conducted. Otherwise, return to step 1.

2.2. Questions about state positions

Generally, a decision tree is built for each state position of each base phone, with the goal of grouping context-dependent phone HMM states into several equivalent classes. However, model performance could be improved when building a single decision tree to group all states associated with a base phone, because it could make better use of the training data than clustering only across the same state position: the trees should maximize likelihood (or minimize model description length in MDL criterion) by adjusting the amount of tying across state positions as well as across contexts. This paper introduces the knowledge which the neighboring states can be tied into questions about state positions as follows:

- Is this the 1st state of the model?
- Is this the 1st or 2nd state of the model?

...
In PDS-DT, proper context-dependent and state-position-dependent sharing structure for grouped dimensions can be constructed. PDS-DT has no external control parameter because it is based on the MDL criterion. An overview of constructed sharing structure by the proposed technique is illustrated in Fig.2. In Fig.2, distributions whose index $\Delta_i$ is based on the MDL criterion. An overview of constructed sharing structure for grouped dimensions can be constructed when questions about state positions are applied at the root node and its neighboring nodes of the decision tree. In this paper, PS-DT in case 1 and case 2 are referred to as PS-DT(1) and PS-DT(2), respectively. Also, PDS-DT in case 1 and case 2 are referred to as PDS-DT(1) and PDS-DT(2), respectively.

2.4. Constructed sharing structure by PDS-DT

In PDS-DT, proper context-dependent and state-position-dependent sharing structure for grouped dimensions can be constructed. PDS-DT has no external control parameter because it is based on the MDL criterion. An overview of constructed sharing structure by the proposed technique is illustrated in Fig.2. In Fig.2, distributions whose index $\Delta_i$ is based on the MDL criterion. An overview of constructed sharing structure for grouped dimensions can be constructed when questions about state positions are applied at the root node and its neighboring nodes of the decision tree. In this paper, PS-DT in case 1 and case 2 are referred to as PS-DT(1) and PS-DT(2), respectively. Also, PDS-DT in case 1 and case 2 are referred to as PDS-DT(1) and PDS-DT(2), respectively.

To evaluate the proposed technique, a gender-dependent continuous speech recognition experiment was conducted. The ASJPB database (phonetically-balanced sentences) and ASJ-INAS database (Japanese newspaper article sentences speech corpus) were used. About 20,000 sentences spoken by about 130 male speakers were used for training. The IPA-98-TestSet that was not used for acoustic model training served as the test data. This test data consists of a total of 100 sentences spoken by 23 speakers.

The speech data was down-sampled from 20kHz to 16kHz, windowed at a 5-ms frame rate using a 25-ms Blackman window, and parameterized into 18 mel-cepstral coefficients with a mel-cepstral analysis technique [14]. Static coefficients excluding zero-th coefficients and their first and second derivatives including zero-th coefficients were used as feature parameters. Cepstral mean subtraction was applied to each utterance. Three, four, and five-state left-to-right HMMs were used to model 43 Japanese phonemes, and 118 questions about phonetic contexts were prepared in decision tree construction. In order to evaluate the sharing structure obtained by each clustering technique, single Gaussian with diagonal covariance matrix per leaf node was used in this experiment.

The one-pass Viterbi algorithm was used for decoding with the phonotactic constraints of phoneme sequences in Japanese. Recognition results were shown in phoneme error rates. The label insertion penalty was varied until the number of insertions and deletions in the recognizer output are approximately equal for each result.

3.2. Experimental results

Figure 3 shows the example of decision tree for phoneme $\text{N}$/ built by PDS-DT. As shown in Fig.3, PDS-DT constructed a different phonetic-context and state-position dependent sharing structure for grouped dimensions.

Table 1 shows the total numbers of distributions (leaf nodes) and free-parameters for each technique. It shows that the numbers of distributions of PD-DT and PDS-DT were much larger than that of P-DT and PS-DT. However, the numbers of free-parameters of PD-DT and PDS-DT were about 13%–18% smaller than that of P-DT and PS-DT. Table 1 may indicate that PD-DT and PDS-DT has more representation ability without increasing the number of free parameters in a system.
When each model has 3 or 4-state HMM topology, tying across state positions did not occur in PS-DT(1) and PS-DT(2). Therefore, constructed sharing structures of P-DT and PS-DT(2) were completely the same. However, in PDS-DT(1) and PDS-DT(2), distribution tying across state position was occurred. When each model has 5-state HMM topology, about 2% of the distributions was tied across state position.

Figure 4 shows the speech recognition performance of each technique. In the case of 3 or 4-state HMM topology, although PDS-DT achieved almost the same recognition performance compared with PD-DT, 13–15% error reduction over P-DT and PS-DT was achieved. However, in the case of 5-state HMM topology, PDS-DT achieved a 7% error reduction over PD-DT and a 16% error reduction over P-DT and PS-DT.

These results indicates new knowledge as follows:

- The dimensional-split in decision tree-based clustering improves recognition performance.
- Questions about state positions in decision tree-based state tying technique is not valid because tying across state positions does not occur.
- Questions about state positions with dimensional-split technique is effective when the number of states in HMM topology is large.

4. Conclusion and future works

This paper presents a new clustering technique called Phonetic, Dimensional and State Positional Decision Tree (PDS-DT). PDS-DT can construct a proper context-dependent and state-position-dependent sharing structure for grouped dimensions. In speaker-independent continuous phoneme recognition experiments, the proposed PDS-DT successfully reduced the phoneme error rates by 1–7% over the Phonetic and Dimensional Decision Tree (PDT) based clustering technique [6] resulting in 13–15% error reduction over the conventional Phonetic Decision Tree (P-DT) based state-tying technique [5, 10].

5. References