# DECISION TREE DISTRIBUTION TYING BASED ON A DIMENSIONAL SPLIT TECHNIQUE

*Heiga Zen , Keiichi Tokuda , and Tadashi Kitamura*

Department of Computer Science,
Nagoya Institute of Technology,
Nagoya, 466-8555, Japan
Email: {zen,tokuda,kitamura}@ics.nitech.ac.jp

## ABSTRACT

In this paper, a new clustering technique called Dimensional Split Phonetic Decision Tree (DS-PDT) is proposed. In DS-PDT, state distributions are split dimensionally when applying phonetic question. This technique is an extension of the decision tree based acoustic modeling. It gives a proper context-dependent sharing structure of each dimension automatically while maintaining the correlations among the dimensions. In speaker-independent continuous speech recognition experiments, DS-PDT achieved about 8% error reduction over the phonetic decision tree clustering.

## 1. INTRODUCTION

In large vocabulary continuous speech recognition systems, context-dependent model, typically triphone, and continuous density HMMs are often used. It is well known that the use of triphones rather than monophones provides 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. It is very important to maintain a good balance between the model complexity and model robustness. Therefore, various parameter clustering techniques has been proposed [1–4]. The use of phonetic decision trees (PDT) [4] is one of the good solutions of this problem. It has two advantages over the bottom-up based approaches. First, by incorporating the phonetic knowledge into the 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 the decision tree based acoustic modeling, a tree is constructed either for each phoneme or for each state of each phoneme. The state-based approach is widely used because it provides a more detailed level of sharing and outperforms the model-based approach [4]. In speech recognition, mel-cepstral coefficients (mel-cepstrum) and their time derivatives ($\Delta$mel-cepstrum, $\Delta^2$mel-cepstrum) are widely used as acoustic features. It is generally considered that the mel-cepstral coefficients in the lower quefrency range have more significant information than those in higher the quefrency range. Likewise, static coefficients have more significant information than their time derivatives. Therefore, assigning more 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. However, in decision tree based state tying technique, all dimensions have common sharing structures. It is considered that the feature vector can be modeled more efficiently by a proper context-dependent sharing structure for each dimension.

In this paper, a new clustering technique called Dimensional Split Phonetic Decision Tree (DS-PDT) is proposed. This technique determines whether the distribution of each dimension should be split or not when applying a phonological question. It gives a proper context-dependent sharing structure of each dimension automatically while maintaining the correlations among the dimensions.

In Section 2, implementation of DS-PDT is described. In Section 3, DS-PDT is evaluated in speaker-independent continuous speech recognition experiments. The last section gives conclusions and future works.

## 2. PROPER SHARING STRUCTURE FOR EACH DIMENSION OF DISTRIBUTIONS

In order to assign the proper number of distributions to each dimension and obtain the proper context-dependent sharing structure for each dimension, Feature-Dependent Successive State Splitting (FD-SSS) has been proposed [5]. FD-SSS is an extension of ML-SSS [6]. It determines the sharing structure of the distribution for each dimension separately. This idea can be applied to phonetic decision tree

The authors wish to thank Y. Nankaku, Dr. T. Yoshimura, and Dr. C. Miyajima for helpful discussions and assistance in this work.

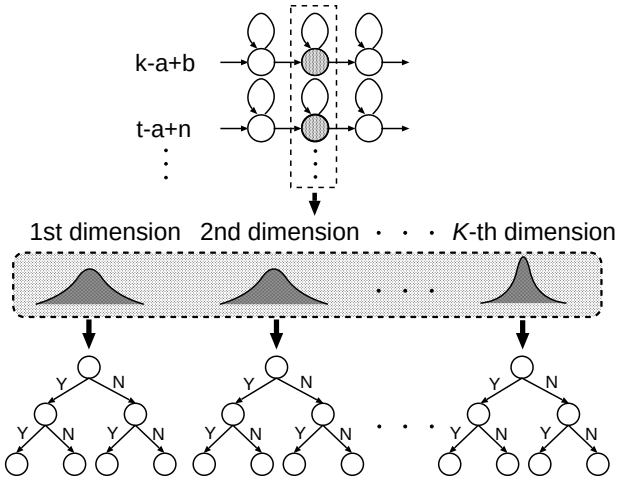


**Fig. 1**. FD-PDT based clustering.

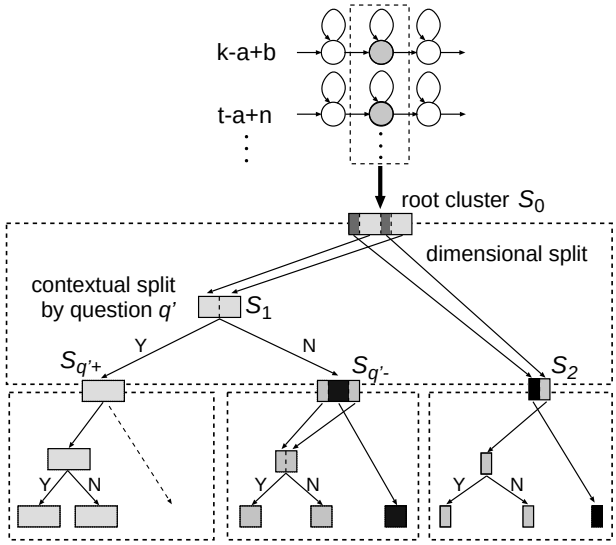


**Fig. 2**. DS-PDT based clustering.

clustering. We call it Feature-Dependent Phonetic Decision Tree (FD-PDT). The FD-PDT based clustering is outlined in Fig.1. In FD-PDT, decision trees are constructed for each dimension of each state of each phoneme.

However, in FD-PDT, each dimension of distributions is separated from the beginning of clustering. Although FD-PDT can provide some adequate context-dependent sharing structure for each dimension, generally there are correlations among different dimensions. They can be captured when distributions are defined across the dimensions and represented by full-covariance matrices or more than two Gaussian mixtures. Accordingly, it is necessary to model a context-dependent sharing structure for each dimension while maintaining the correlations among the dimensions.

### 2.1. Dimensional Split Phonetic Decision Tree

In order to obtain a proper context-dependent sharing structure for each dimension while maintaining the correlations among the dimensions, DS-PDT, which is an extension of PDT, is proposed. In PDT clustering based on MDL criterion [7], when splitting cluster $S$ with phonological question $q$ into clusters $S_{q+}$ and $S_{q-}$, the change of model Description Length (DL), $\Delta_q$, is given by

$$\Delta_q = \frac{1}{2}\Big\{ \Gamma(S_{q+}) \log \left|\Sigma_{S_{q+}}\right| + \Gamma(S_{q-}) \log \left|\Sigma_{S_{q-}}\right| - \Gamma(S) \log |\Sigma_S| \Big\} + K \log \Gamma(S_0) \ , \quad (1)$$

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

$$\Delta_q = \sum_{k=1}^{K} \Delta_q^{(k)} \ , \quad (2)$$

$$\Delta_q^{(k)} = \frac{1}{2}\Big\{ \Gamma(S_{q+}) \log \sigma^2_{S_{q+},(k)} + \Gamma(S_{q-}) \log \sigma^2_{S_{q-},(k)} - \Gamma(S) \log \sigma^2_{S,(k)} \Big\} + \log \Gamma(S_0) \ , \quad (3)$$

where $\sigma^2_{S,(k)}$ , $\sigma^2_{S_{q+},(k)}$, and $\sigma^2_{S_{q-},(k)}$ are the $k$-th elements of diagonal covariance matrices $\Sigma_S$, $\Sigma_{S_{q+}}$, and $\Sigma_{S_{q-}}$, respectively. Thus, the DL change of whole distribution $\Delta_q$ is given by the total of DL change of each dimension, $\Delta_q^{(k)}$. In DS-PDT, a phonological question is applied for the dimensions like $\Delta_q^{(k)} < 0$. The set of dimensions $H_q$ to be split using question $q$, the change of DL $\Delta_q$ for question $q$, and the best question $q'$ for contextual split are given by

$$H_q = \Big\{ k \ \Big| \ k \in G,\ \Delta_q^{(k)} < 0 \Big\} \ , \quad (4)$$

$$\Delta_q = \sum_{k \in H_q} \Delta_q^{(k)} \ , \quad (5)$$

$$q' = \arg\min_{q \in Q} \Delta_q, \quad (6)$$

respectively, where $G$ is a set of dimensions existing in cluster $S$, and $Q$ is a set of phonological questions.

The DS-PDT based clustering shown in Fig.2 is outlined as follows:

Step 1: For all of clusters in decision tree, calculate $\Delta_{q'}$ and determine $H_{q'}$.

Step 2: Choose cluster $S$ which has the minimum $\Delta_{q'}$.

Step 3: Split cluster $S$ dimensionally into two clusters $S_1$ and $S_2$ according to $H_{q'}$. $S_1$ is composed of dimensions which exist in $H_{q'}$, and $S_2$ is composed of dimensions which do not exist in $H_{q'}$ while existing in $G$.

Step 4: Split cluster $S_1$ contextually into $S_{q'+}$ and $S_{q'-}$ by phonological question $q'$.

Step 5: if $\Delta_{q'} \geq 0$ for all of clusters, no splitting conducted.

### 2.2. Relation among PDT, FD-PDT, and proposed DS-PDT

DS-PDT is equivalent to PDT if dimensional split $H_q$ is restricted by $H_q = G$. Also, DS-PDT is equivalent to FD-PDT if dimensional split $H_q$ is restricted by

$$H_q = \left\{ k \;\middle|\; \arg\min_{k \in G} \Delta_q^{(k)}, \Delta_q^{(k)} < 0 \right\} .$$

Thus, DS-PDT includes PDT and FD-PDT as special cases.

## 3. EXPERIMENTS

### 3.1. Experimental conditions

To evaluate the proposed technique, we conducted a speaker-independent speech recognition experiment. We used phonetically balanced 503 sentences uttered by 6 male speakers from the ATR speech database B-set. The 450 sentences were used for training, and the remaining 53 sentences were used for testing. We adopted a jack-knife approach, i.e., 5 speakers leaving out one of the speakers were used for training HMMs, and the excluded speaker was used for testing. All results were shown in the phoneme error rates averaged over 6 recognition tasks in the jack-knife approach.

Speech signal was sampled at 16kHz, windowed at a 5-ms frame rate using a 25.6-ms Blackman window, and parameterized into 19 mel-cepstral coefficients with a mel-cepstral analysis technique [8]. The 18 static coefficients excluding zero-th coefficient, their first and second derivatives including zero-th coefficients were used as feature parameters. We used 3-state left-to-right HMMs for modeling 37 Japanese phonemes, and 118 phonological questions were asked for splitting nodes in a decision tree.

The one-pass Viterbi algorithm was used for decoding with the phonotactic constraints of phoneme sequences in Japanese.

### 3.2. Constructed decision tree

Figures 3–5 show the examples of decision trees for the 2nd state of phoneme /k/. Figures 3, 4, and 5 correspond to PDT, FD-PDT, and proposed DS-PDT, respectively. As shown in Fig.4 and 5, FD-PDT and DS-PDT has different tying structures for different features.

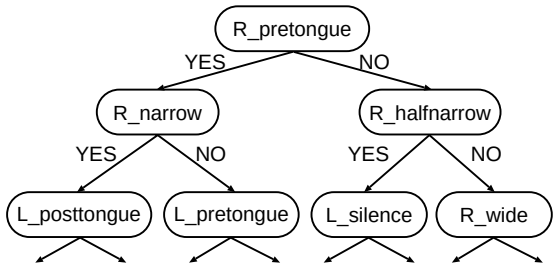


**Fig. 3**. Example of PDT for the 2nd state of phoneme /k/.

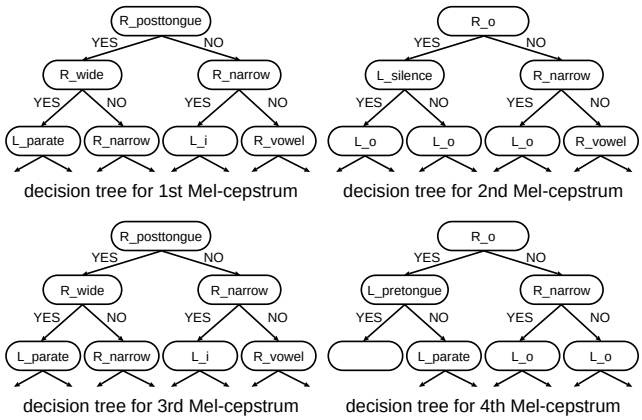


**Fig. 4**. Examples of FD-PDT for the 2nd state of phoneme /k/.

Table 1 shows the examples of number of applied questions. Table 2 shows the total numbers of distributions (leaf nodes) and free parameters. From Table 1 and 2, it can be seen that FD-PDT and DS-PDT have stronger context dependencies and a higher representation ability than PDT, without increasing the number of free parameters.

Figure 6 shows the total number of distributions assigned to each individual dimension in the PDT, FD-PDT and DS-PDT. It can be seen that the numbers of assigned distributions for FD-PDT and DS-PDT are different among the dimensions. This implies that the FD-PDT and DS-PDT suc-

**Table 1**. Examples of number of applied questions.

| method | 2nd state of “o” | 2nd state of “a” |
|---|---|---|
| PDT | 67 | 47 |
| FD-PDT | 104 | 106 |
| DS-PDT | 104 | 105 |

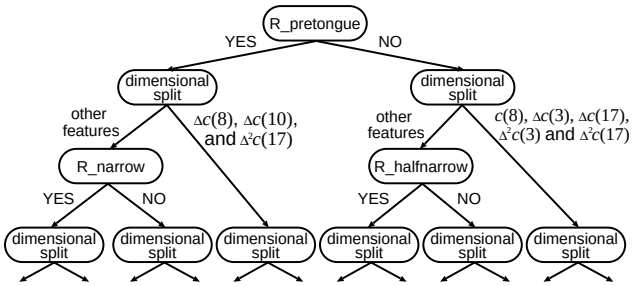


**Fig. 5**. Example of DS-PDT for the 2nd state of phoneme /k/.

**Table 2**. Numbers of distributions (leaf-nodes) and free parameters.

| method | #distributions | #parameters |
|---|---|---|
| PDT | 3,044 | 344,624 |
| FD-PDT | 136,592 | 273,182 |
| DS-PDT | 93,838 | 311,916 |

cessfully allocated an optimal number of free parameters to each dimension.

Figure 7 shows the experimental results. DS-PDT achieved about 8% error reduction over PDT in almost same number of free parameters. When the distribution of each leaf node represented by a single Gaussian, FD-PDT and DS-PDT gave almost the same recognition performance. However, by increasing the number of Gaussian mixtures, FD-PDT did not achieve a significant improvement, whereas DS-PDT recognition performance was improved. It is considered that DS-PDT can capture the correlations among the dimensions when each distribution has more than two Gaussian mixtures.

## 4. CONCLUSION

In this paper, we have presented a new clustering technique called Dimensional Split Phonetic Decision Tree (DS-PDT). DS-PDT can assign an optimal number of distributions to each dimension maintaining the correlation among the dimensions. In the speaker-independent continuous phoneme recognition experiments, proposed DS-PDT successfully reduced the phoneme error rates by about 8% over PDT.

Future works include application to large vocabulary continuous speech recognition.

## 5. REFERENCES

[1] K.-F. Lee, "Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.38, no.4, pp.599–609, 1990.

[2] J. Takami and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," *Proc. of ICASSP'92*, pp.573–576, 1992.

[3] M.-Y. Hwang, X. Huang, and F.Alleva, "Predicting Unseen Triphones with Senones," *Proc. of ICASSP'93*, pp.311–314, 1993.

[4] S. J. Young, J. J. Odell and P. C. Woodland, "Tree-Based State Tying for High Accuracy Acoustic Modeling," *ARPA Human Language Technology Workshop* , pp.286–291, 1994.

[5] S. Matsuda, M. Nakai, H. Shimodaira and S. Sagayama, "Feature-Dependent Allophone Clustering," *Proc. of ICSLP2000*, pp.413–416, 2000.

[6] M. Ostendorf and H. Singer, "HMM Topology Design Using Maximum Likelihood Successive State Splitting," *Computer Speech and Language*, vol.11, no.1, pp.17–41, 1997.

[7] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn.(E)*, vol.21, no.2, pp.79–86, 2000.

[8] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," *Proc. of ICASSP'92*, vol.1, pp.137–140, 1992.

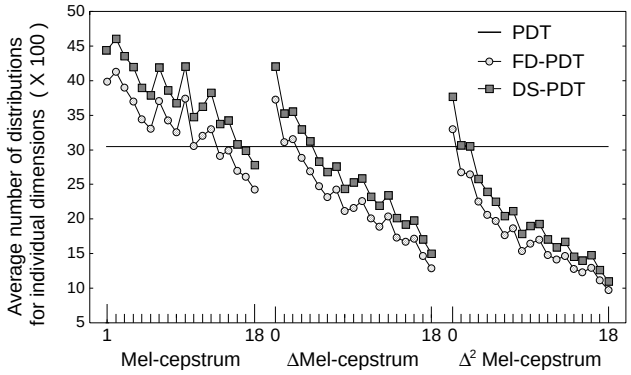


**Fig. 6**. Numbers of distributions of each dimension for PDT, FD-PDT, and DS-PDT.

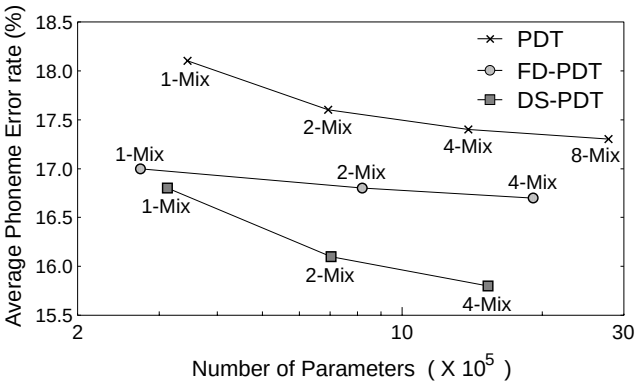


**Fig. 7**. Average phoneme error rates for PDT, FD-PDT, and DS-PDT.