DECISION TREE DISTRIBUTION TYING BASED ON A DIMENSIONAL SPLIT TECHNIQUE

Authors : Heiga Zen

Keiichi Tokuda

Tadashi Kitamura

Presented by : Heiga Zen

Department of Computer Science,

Nagoya Institute of Technology, Japan

ICSLP 2002

Introduction

Phonetic Decision Tree based State Tying Approach

- Reduce free parameters in a system
- Generate unseen models
- All dimensions have the same sharing structure

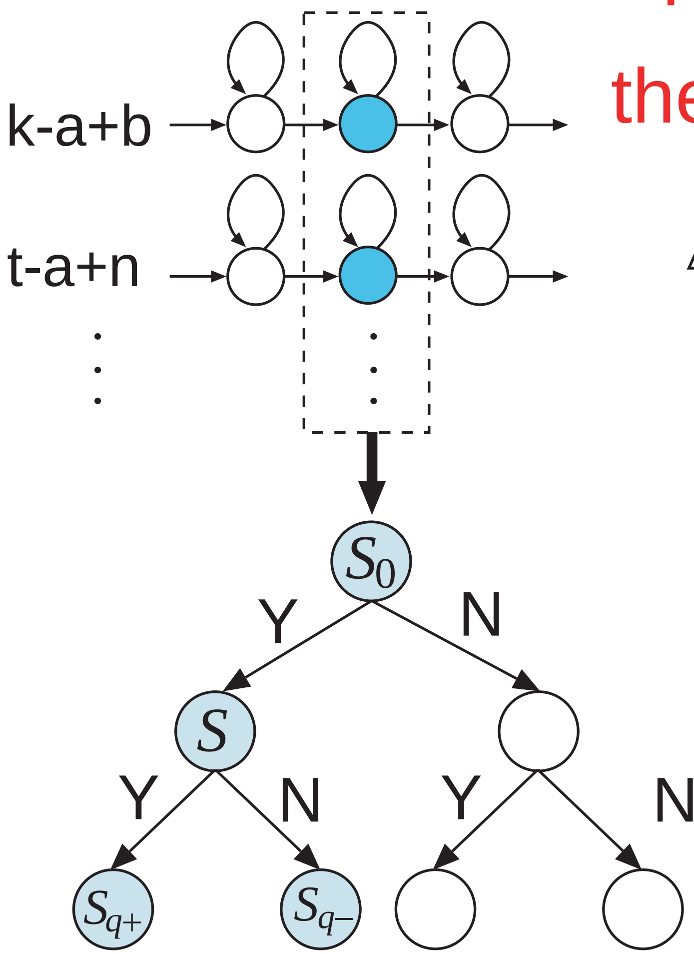
All dimensions should have the same sharing structure ?

Dimensional Split Phonetic Decision Tree

- Construct proper sharing structure of each dimension
- Capture correlations among dimensions

Phonetic Decision Tree Based on MDL Criterion

- PDT based on MDL criterion [Shinoda, et al. ; 1997]



Split S into S_{q+} and S_{q-} by question q ,

the difference of DL value, Δ_q is given by

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

K	: Dimensionality of Feature Vector
Σ	: Covariance Matrix of Each Cluster
$\Gamma(\cdot)$	: State Occupancy Count for Each Cluster

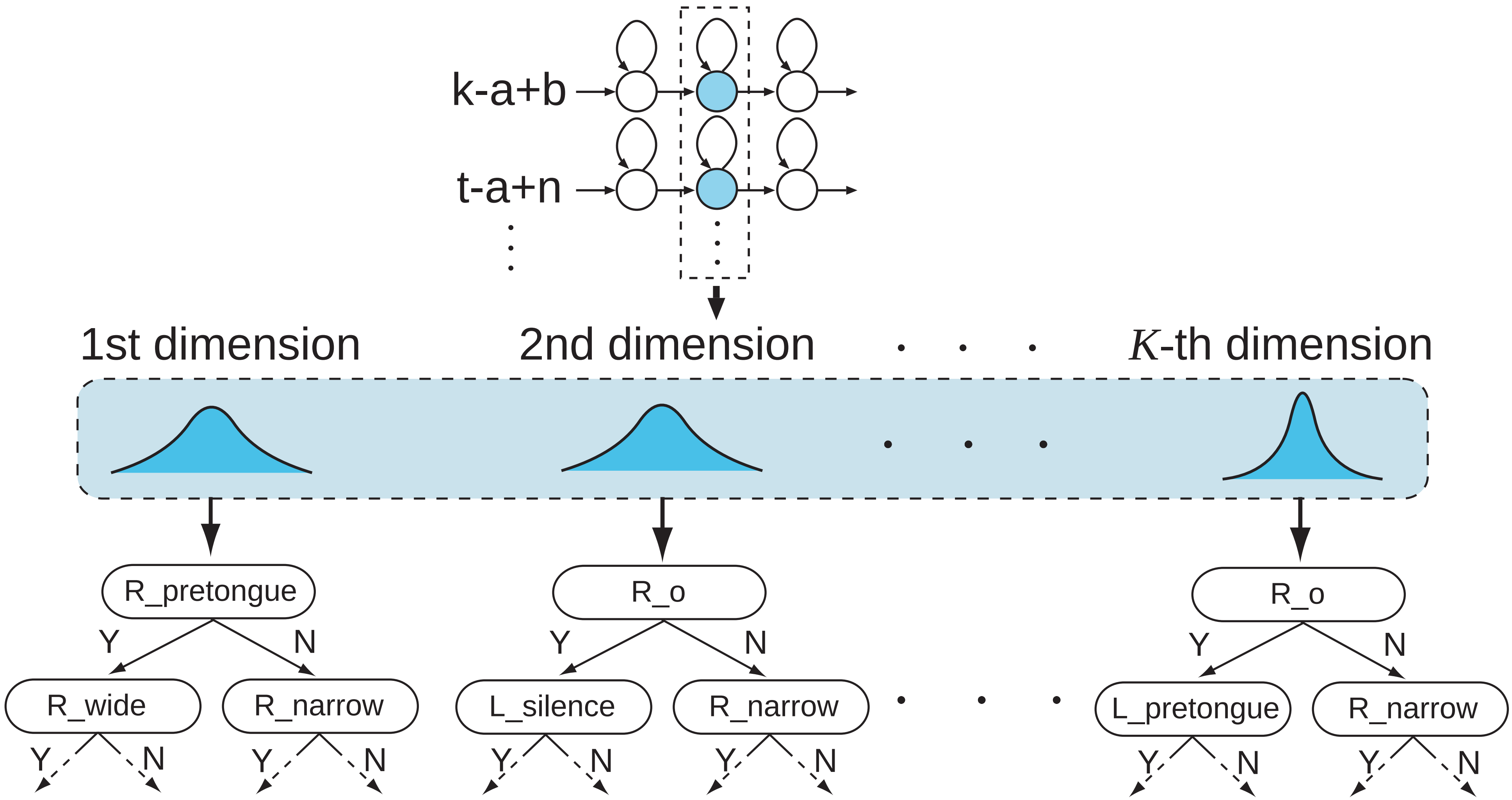
$$q_{\mathrm{best}} = \left\{ q \,\middle|\, \arg\min_{q \in Q} \Delta_q , \Delta_q < 0 \right\}$$

$q_{\mathrm{best}} = \phi \rightarrow$ stop

Feature Dependent Phonetic Decision Tree

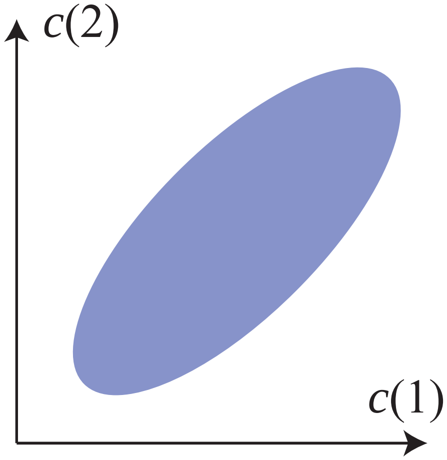
- Feature Dependent SSS [Matsuda, et al. ; 2000]

  → Feature Dependent Phonetic Decision Tree

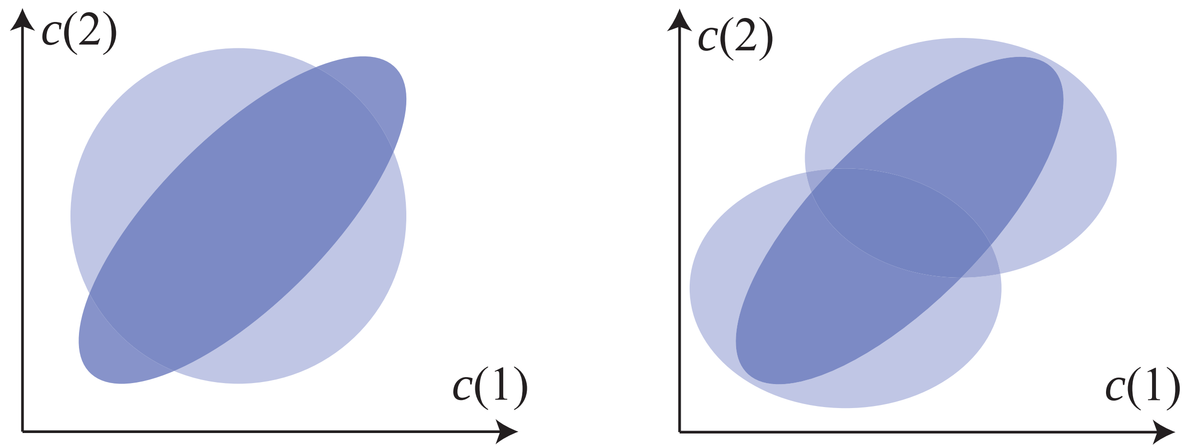


Problem in Feature Dependent Approach

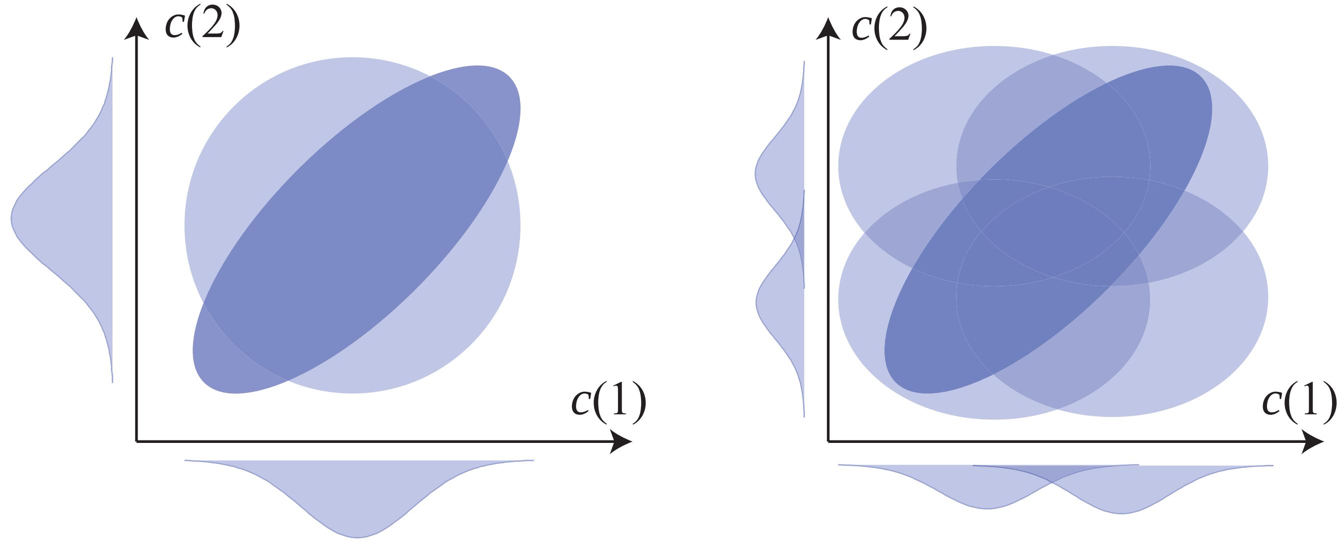
True Distribution



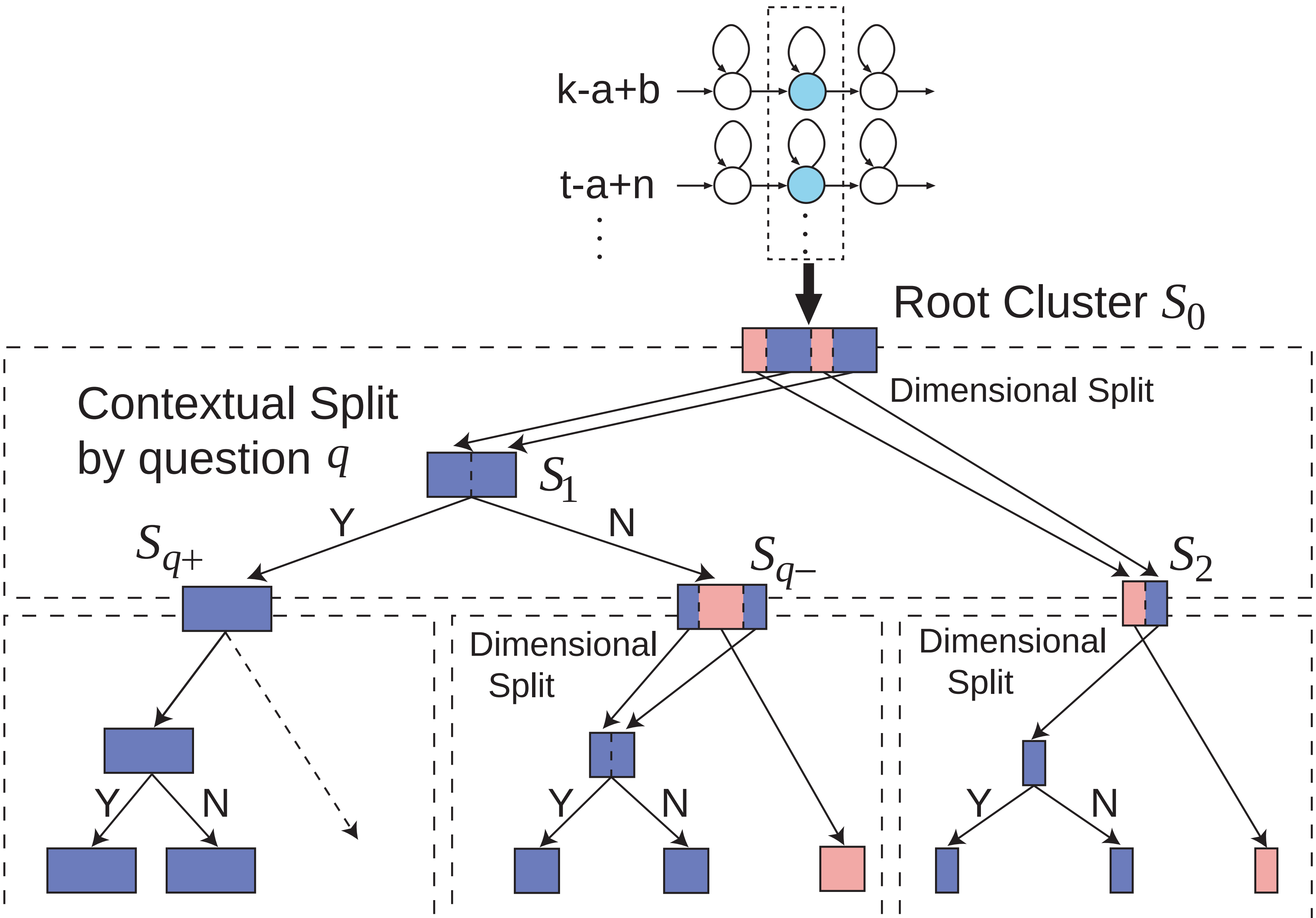
State Tying Approach



Feature Dependent Approach

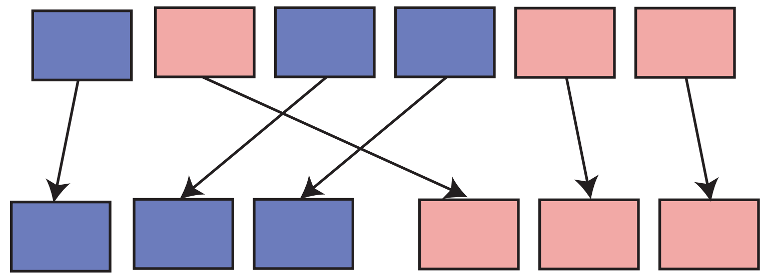


Dimensional Split Phonetic Decision Tree



Dimensional Split Technique

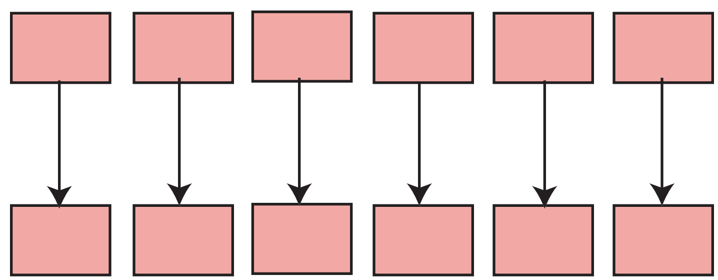
q_1



$H_{q_1} = \{1, 3, 4\}$

$\Delta_{q_1} = \Delta^1_{q_1} + \Delta^3_{q_1} + \Delta^4_{q_1}$
$= -0.8$

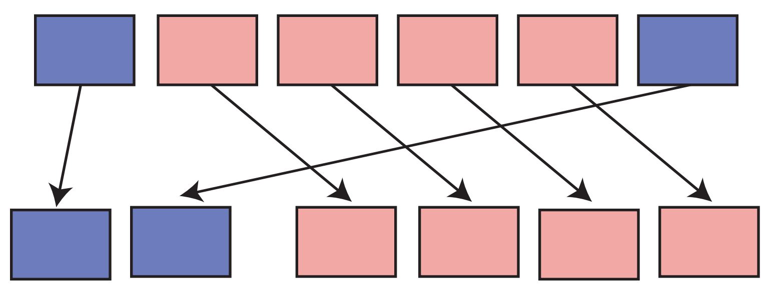
q_2



$H_{q_2} = \{\phi\}$

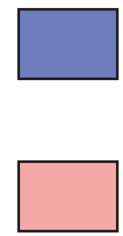
$\Delta_{q_2} = 0$

q_3



$H_{q_3} = \{1, 6\}$

$\Delta_{q_3} = \Delta^1_{q_3} + \Delta^6_{q_3}$
$= -1.5$



$\Delta^{(k)}_q < 0$

$\Delta^{(k)}_q > 0$

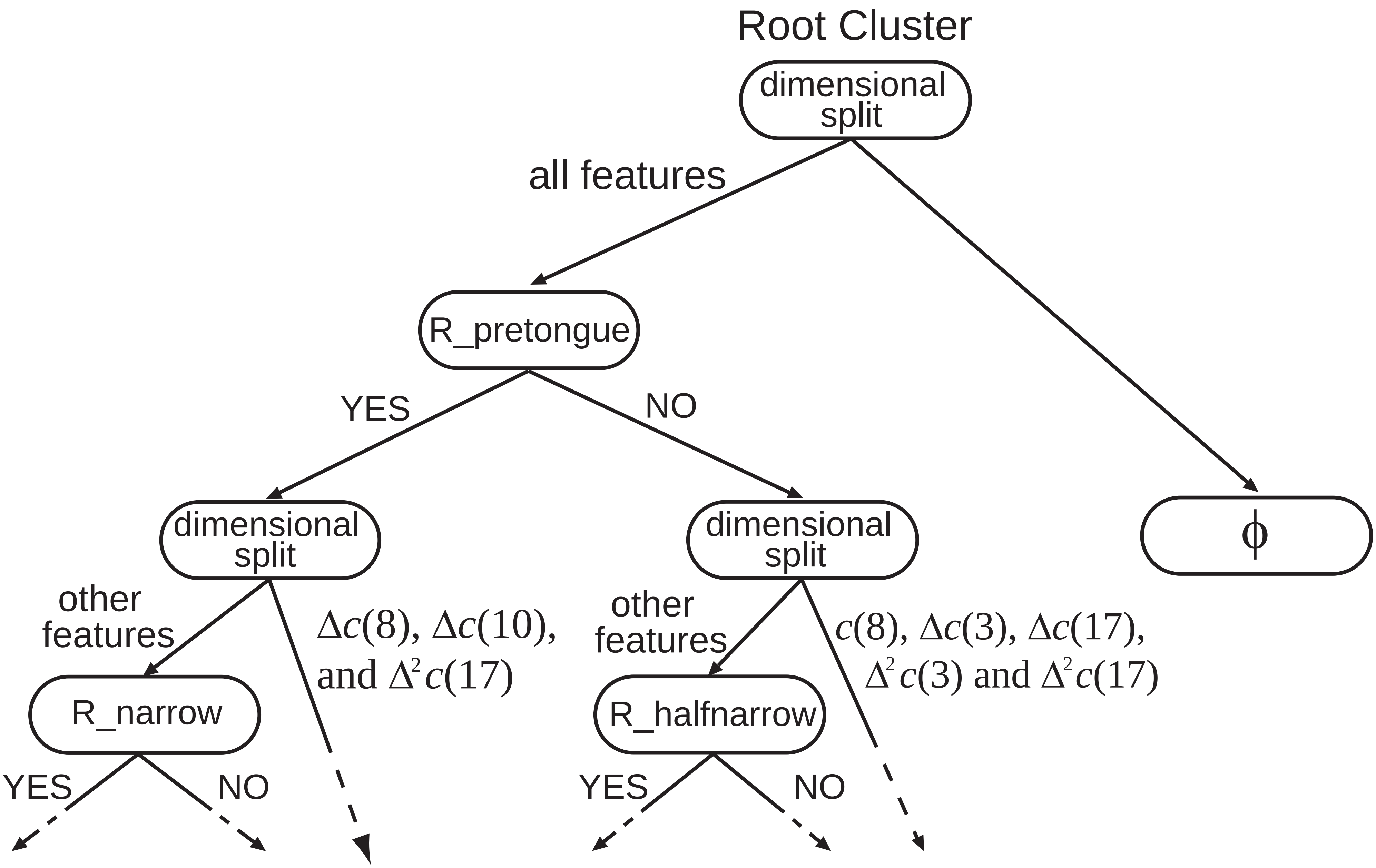
Best Question

$$q_{\text{best}} = \arg\min_{q \in Q} \Delta_q$$
$$= q_3$$

Dimensional Split

$$S = \{1, 2, 3, 4, 5, 6\}$$
$$S_1 = \{1, 6\}$$
$$S_2 = \{2, 3, 4, 5\}$$

Constructed Decision Tree by Dimensional Split Approach



Experimental Conditions

Database	ATR continuous speech database B-set, 6 male speakers
Training Data	5 male speakers, phonetically balanced 450 sentences
Test Data	53 sentences of remaining 1 male speaker
Test Method	Jack-knife approach
Recognition	Continuous phoneme recognition

Clustering Results

Number of leaf nodes and free parameters

	#leaf nodes	#parameters
PDT	3,044	344,624
FD-PDT	136,592	273,182
DS-PDT	93,838	311,916

PDT : Phonetic Decision Tree

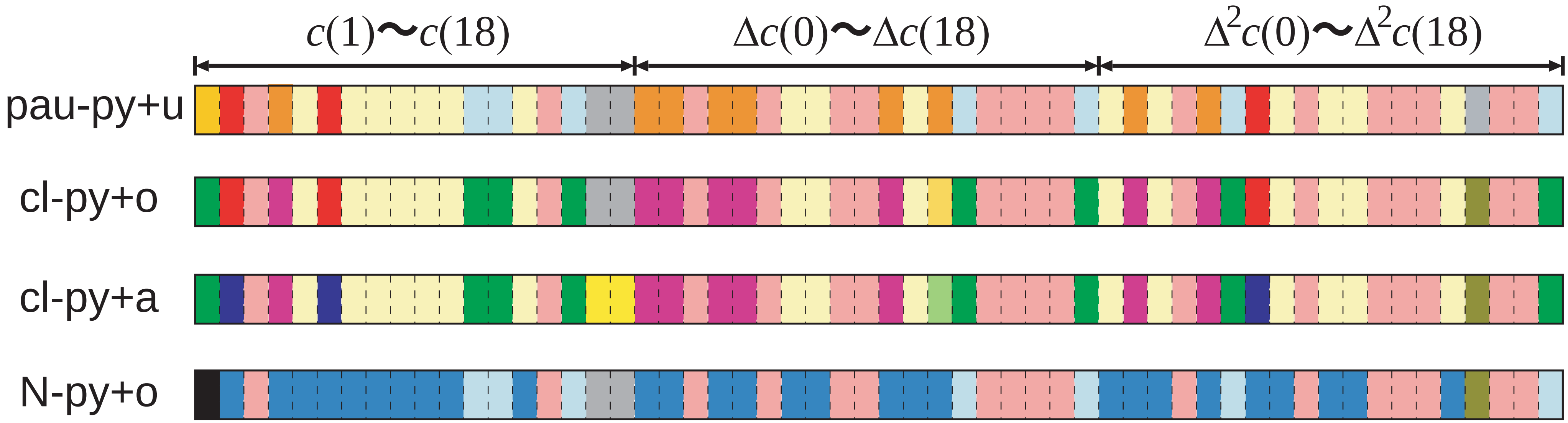
FD-PDT : Feature Dependent Phonetic Decision Tree

DS-PDT : Dimensional Split Phonetic Decision Tree

All of these methods were based on MDL criterion.

Example of Sharing Structure in DS-PDT

1st state of phoneme "py"



The same color regions correspond to the same leaf node.

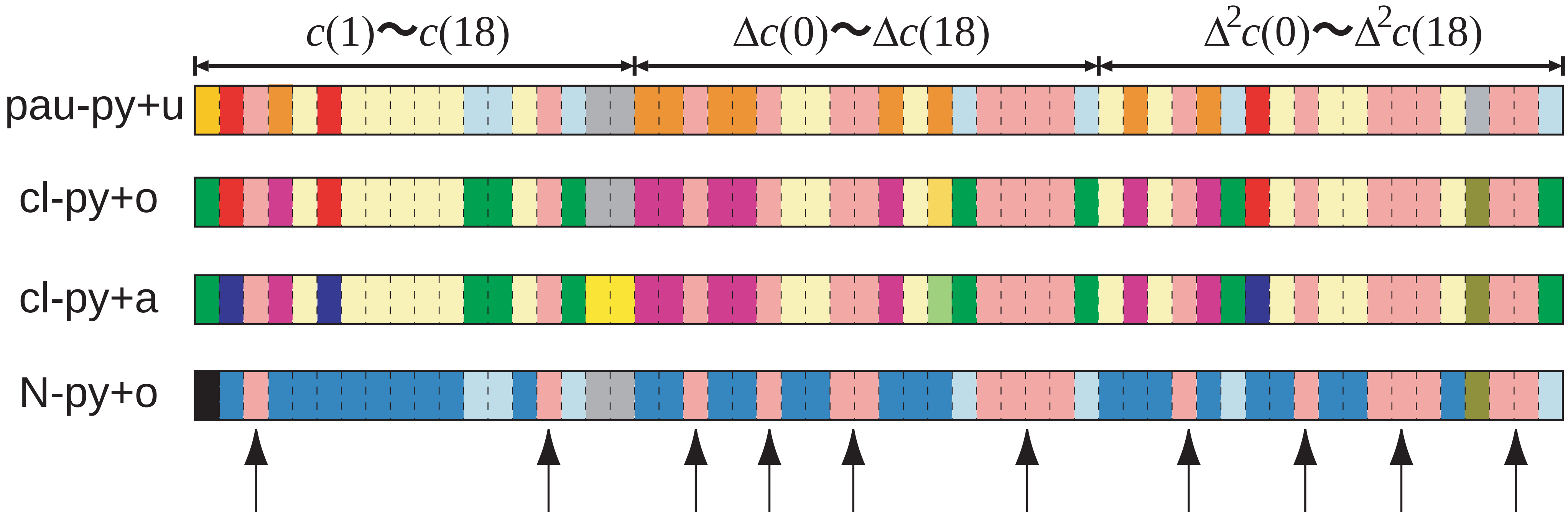
PDT clustered these 4 states into 2 states,

DS-PDT clustered these into 16 sub-distributions,

FD-PDT clustered these into 103 scalar distributions.

Example of Sharing Structure in DS-PDT

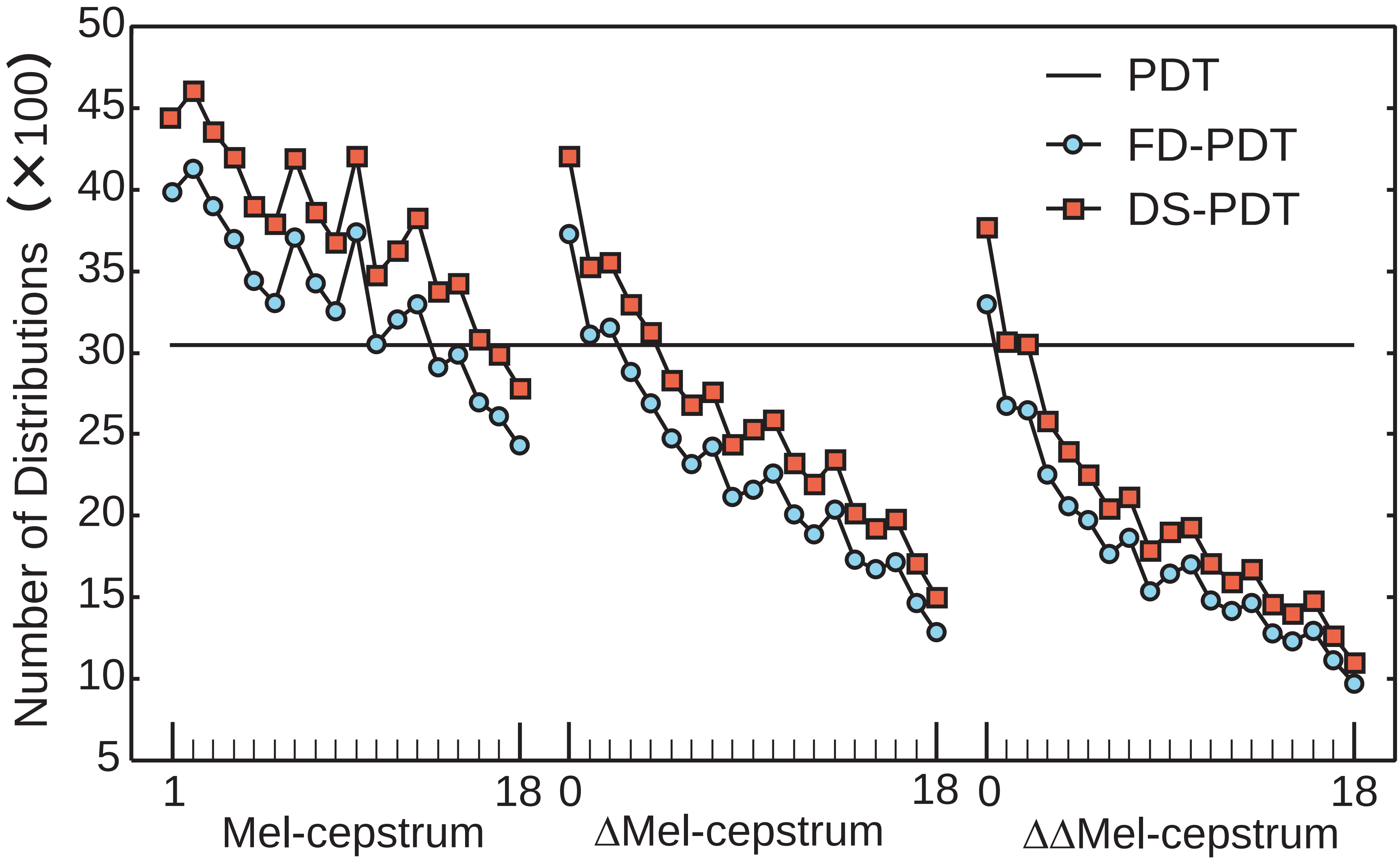
1st state of phoneme "py"



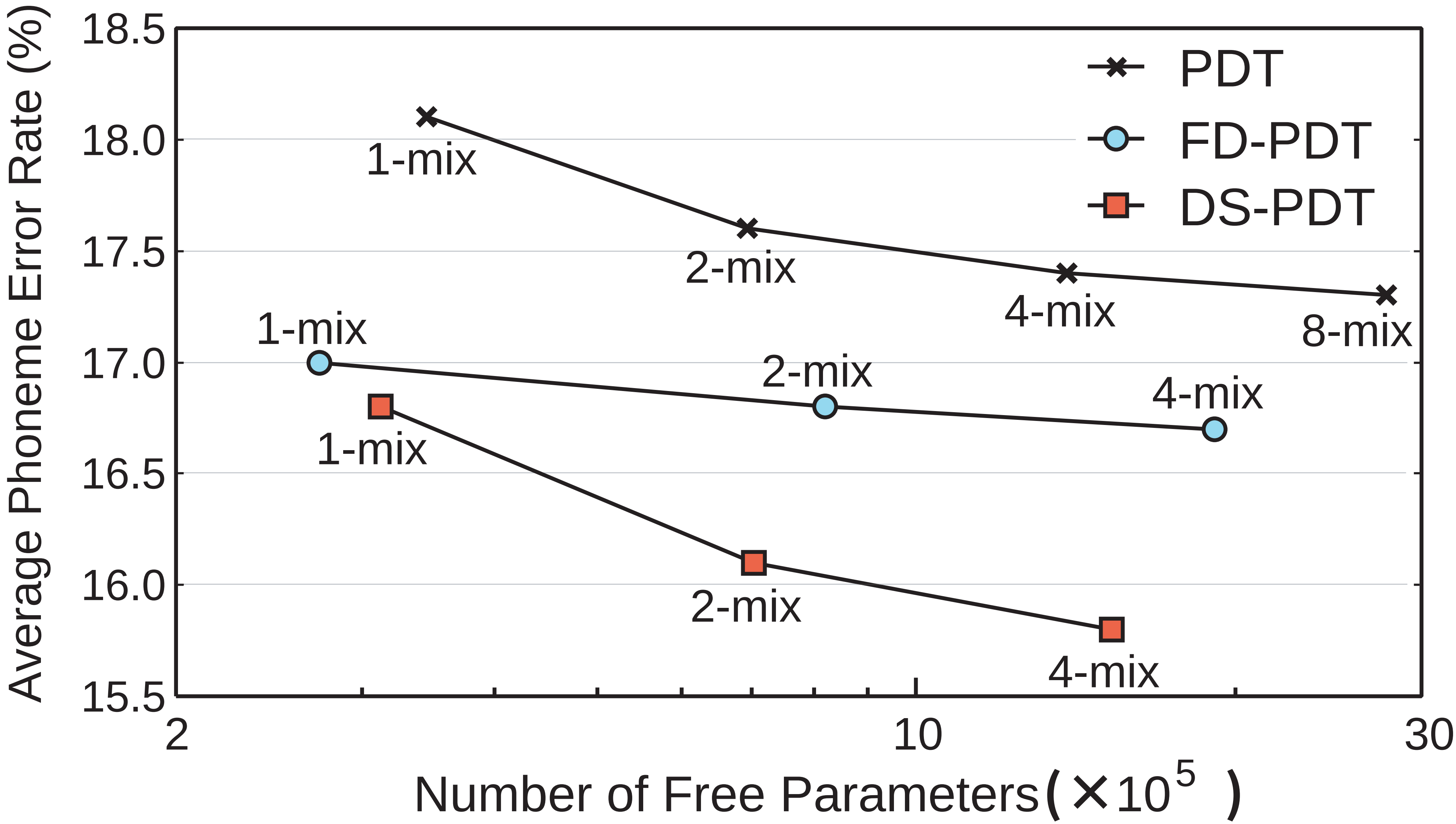
The same distribution is shared (17 / 56)

These features were independent from contextual environment in this task ?

Number of Distributions in Each Dimension



Experimental Results



Conclusion

- DS-PDT can construct proper sharing structure for each dimension.

- DS-PDT achieved about 8% error reduction over conventional PDT.

Future Work

- Experiment on large corpus.

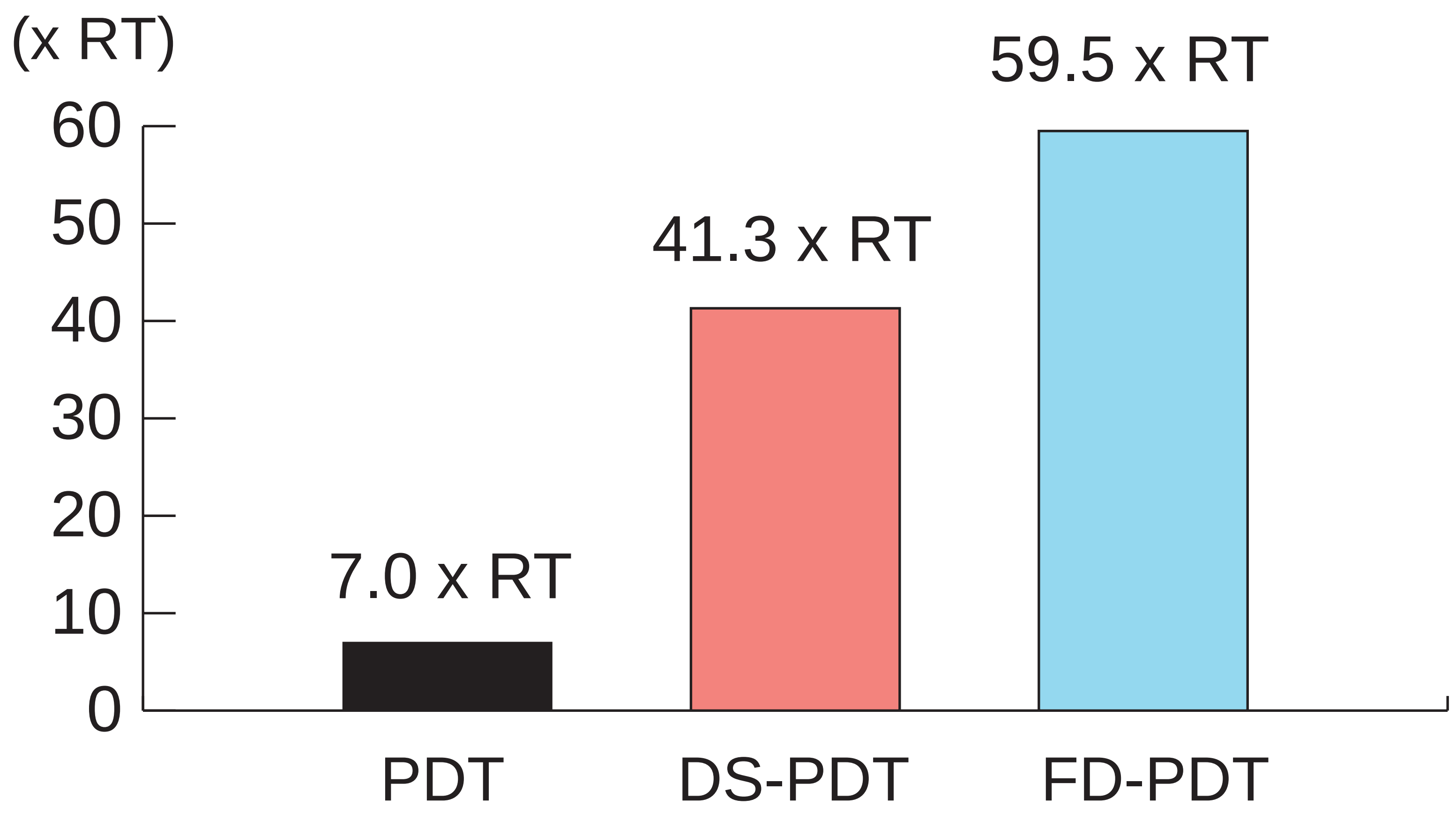
- Large vocaburaly continuous word recognition.

Computational Cost

Machine : Pentium 4 , 1.6 GHz

Decoder : HTK Viterbi Decoder

(modified for proposed acoustic models, not tuned)



Difference of DL in Each Dimension

The difference of Description Length in PDT based on MDL

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

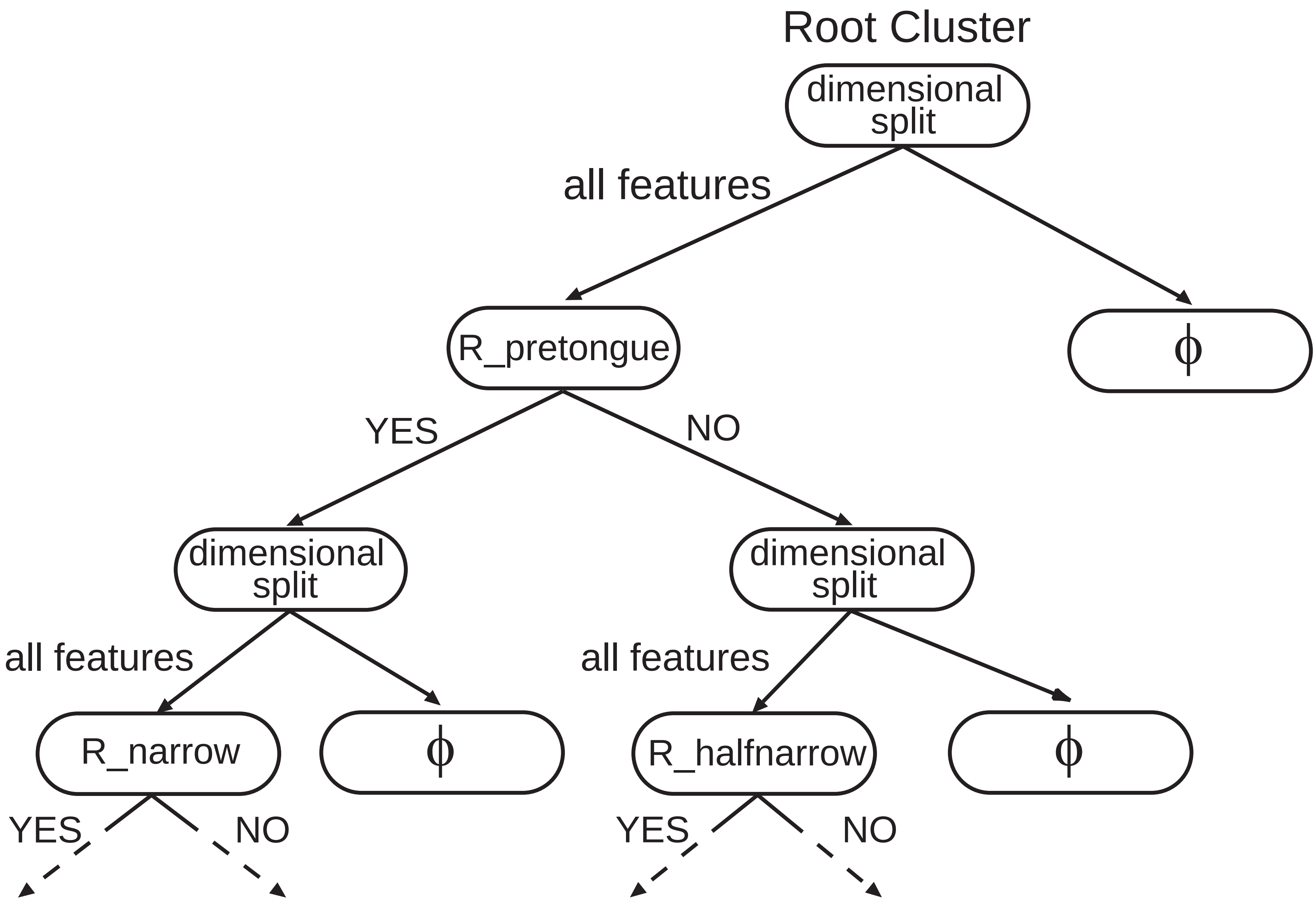
In case of diagonal covariance,

$$\Sigma_S = \mathrm{diag}\,[\sigma^2_{S,(1)}, \sigma^2_{S,(2)}, \ldots, \sigma^2_{S,(K)}]$$

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

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

Special Case of DS-PDT = State Tying Approach



Special Case of DS-PDT = Feature Dependent Approach