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

Heiga Zen, Keiichi Tokuda, Tadashi Kitamura (Nagoya Institute of Technology, Japan)

1. Introduction

Context-dependent HMMs (ex. triphone HMMs)

⇒ Too many free-parameters in system
Unseen model

Phonetic Decision Tree (P-DT) based state tying

- P-DT reduces free-parameters in a system
- P-DT can generate unseen models
- All dimensions have the same sharing structure
- Each state position has different decision tree

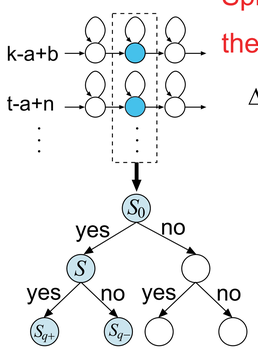
All dimensions have the same context-dependency ?
All dimensions have the same state-position-dependency ?

Phonetic, Dimensional & State Positional Decision Tree (PDS-DT) based clustering technique

- PDS-DT is an extension of P-DT
- Can construct different tying structure for each dimension
- Evaluate PDS-DT in speech recognition experiment

2. Phonetic Decision Tree-based State Tying Technique

P-DT based on ML criterion [S. J. Young et al. ; 1994]
P-DT based on MDL criterion [K. 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

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

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

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

$q_{\text{best}} = \phi$ → stop

3. Phonetic & State-Positional Decision Tree

Phonetic & State-positional Decision Tree (PS-DT)
[A. Lazarides, et al. ;1996, H. J. Nock ;1996]

- Introduce questions about state positions into P-DT
- Can construct state-tying structure across state positions

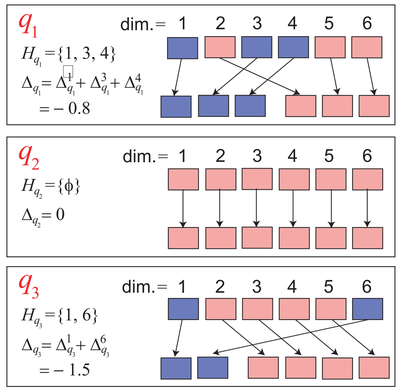
→ Almost the same as P-DT based state tying

4. Phonetic & Dimensional Decision Tree

MDL-based dimensional-split [H. Zen, et al., 2002]

→ Phonetic & Dimensional Decision Tree (PD-DT)

- Distritubtions are dimensionally split
- PD-DT construct proper context-dependent sharing structure for grouped dimensions



Best Question

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

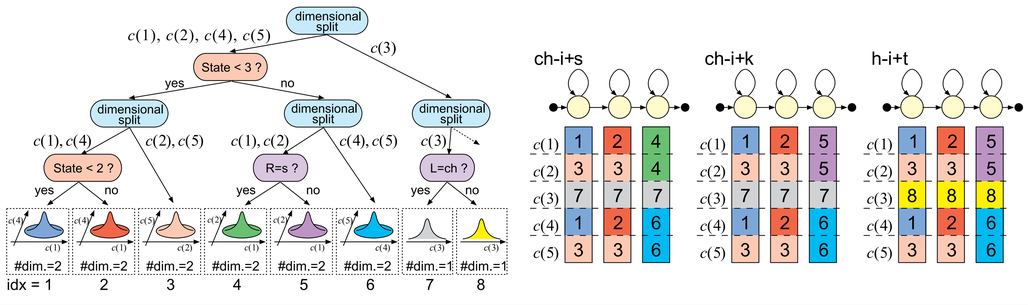
Dimensional Split

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

5. Phonetic, Dimensional & State-positional Decision Tree based Clustering

Phonetic, Dimensional & State-positional Decision Tree based Clustering

- Introduce dimensional-split technique into PS-DT
- Phonetic contexts, dimensions and state positions are clustered simultaneously
- Each distribution has different dimensionality
- Each distribution is composed of different dimensions
- Unified technique for decision-tree based acoustic modeling based on MDL criterion



6. Constructed Parameter Sharing Structure

Phonetic Decision Tree (P-DT)

- All dimensions have the same context-dependent sharing structure
- P-DT cannot construct state sharing structure across state positions

Phonetic & State Positional Decision Tree (PS-DT)

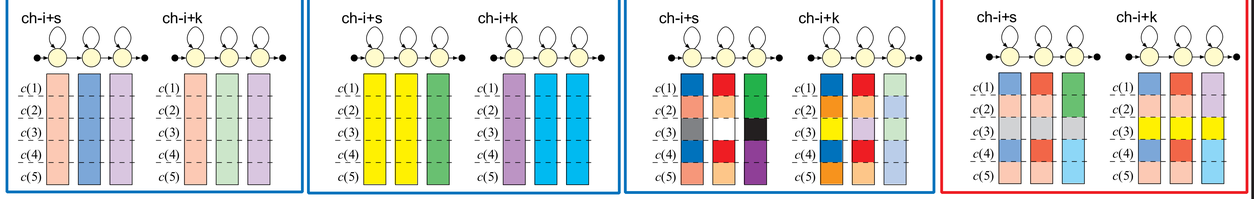
- All dimensions have the same context-dependent sharing structure
- PS-DT can construct state sharing structure across state positions

Phonetic & Dimensional Decision Tree (PD-DT)

- Each dimension has different context-dependent sharing structure
- PD-DT cannot construct distribution sharing structure across state positions

Phonetic, Dimensional & State Positional Decision Tree (PDS-DT)

- Each dimension has different context-dependent sharing structure
- PDS-DT can construct distribution sharing structure across state positions



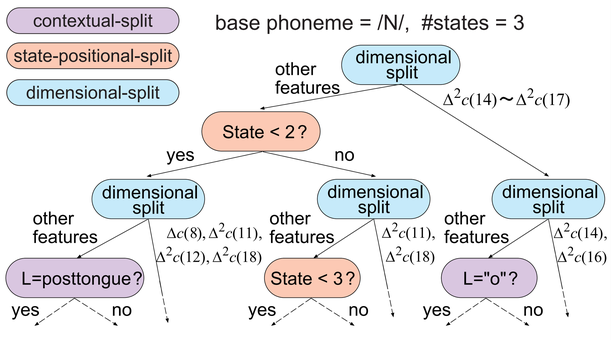
7. Experimental Conditions

Database	ASJ-PB database (phonetically-balanced) ASJ-JNAS database (newspaper article)
Training Data	about 25000 sentences (about 160 male speakers)
Test Data	100 sentences spoken by 23 male speakers

Sampling Frequency	16kHz
Window	Blackman window
Frame Length / Rate	25ms / 5ms
Speech Analysis	Mel-cepstral Analysis
Feature Vector	$c(1)\sim c(18)$, $\Delta c(0)\sim\Delta c(18)$, $\Delta^2 c(0)\sim\Delta^2 c(18)$
CMS	Each Utterance

HMM topology	3, 4, 5 states left-to-right HMM (no skip transition) with diagonal covariance matrices
#base phones	43 Japanese phonemes
#questions	118 phonetic questions about left & right phoneme

8. Example of constructed PDS-DT



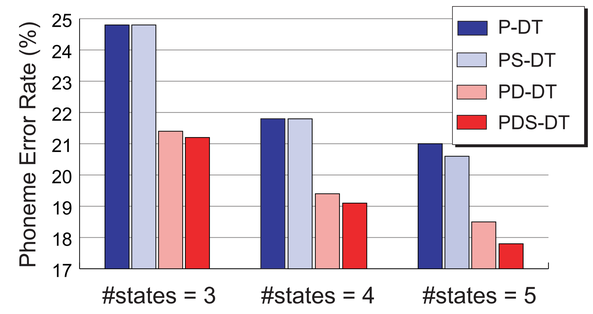
9. Constructed HMMs

Number of states, distributions and free-parameters

	#states	#dist.	#param.	%dist. tying across state positions
PS-DT	3	6,955	778,960	0 %
	4	8,255	924,560	0 %
	5	9,322	1,044,064	0.93 %
PDS-DT	3	185,743	644,004	0.23 %
	4	234,350	780,568	0.54 %
	5	275,220	893,574	2.11 %

Dimensional-Split + Question about state position
→ Tying across state position occurred

10. Recognition Experimental Results



PDS-DT achieved about 13-15% phoneme error reduction rate over PDT