

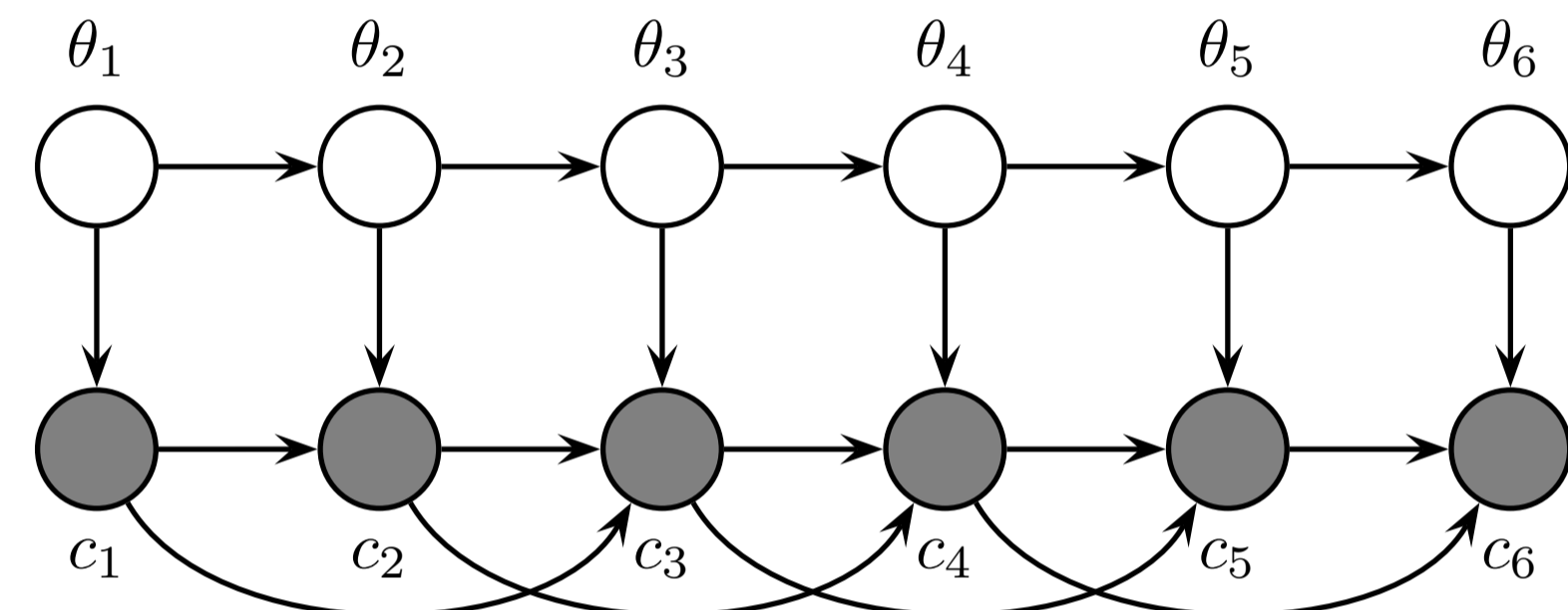
Outline

- ▶ **autoregressive HMM** is an alternative model for speech
 - ▶ simple, efficient parameter estimation
 - ▶ high-quality synthesis (Shannon and Byrne, 2009)
- ▶ but previous experiments re-used decision trees from a standard, non-autoregressive system
- ⇒ investigate **autoregressive decision tree clustering**

Autoregressive HMM

- ▶ alternative to the standard HMM synthesis framework
- ▶ modifies state output distributions

$$P(c, \theta) = \prod_t \underbrace{P(\theta_t | \theta_{t-1})}_{\text{transition prob}} \underbrace{P(c_t | c_{t-K:t-1}, \theta_t)}_{\text{state output dist}}$$



- ▶ hidden **state sequence** $\theta = \theta_{1:T}$
- ▶ observed acoustic **feature vector sequence** $c = c_{1:T}$
- ▶ turns problem of learning a model over sequences $P(c|\theta)$ from data

t	...	20	21	22	23	24	...
θ_t	...	k-aa+t	k-aa+t	k-aa+t	k-aa+t	aa-t+s	...
c_t	...	1.0	1.3	1.6	2.0	1.8	...

- ▶ into learning a function $(c_{t-K:t-1}, \theta_t) \mapsto c_t$ from data

$(c_{t-2}, c_{t-1}, \theta_t)$	$\mapsto c_t$
(0.6, 0.7, k-aa+t)	$\mapsto 1.0$
(0.7, 1.0, k-aa+t)	$\mapsto 1.3$
(1.0, 1.3, k-aa+t)	$\mapsto 1.6$
(1.3, 1.6, k-aa+t)	$\mapsto 2.0$
(1.6, 2.0, aa-t+s)	$\mapsto 1.8$

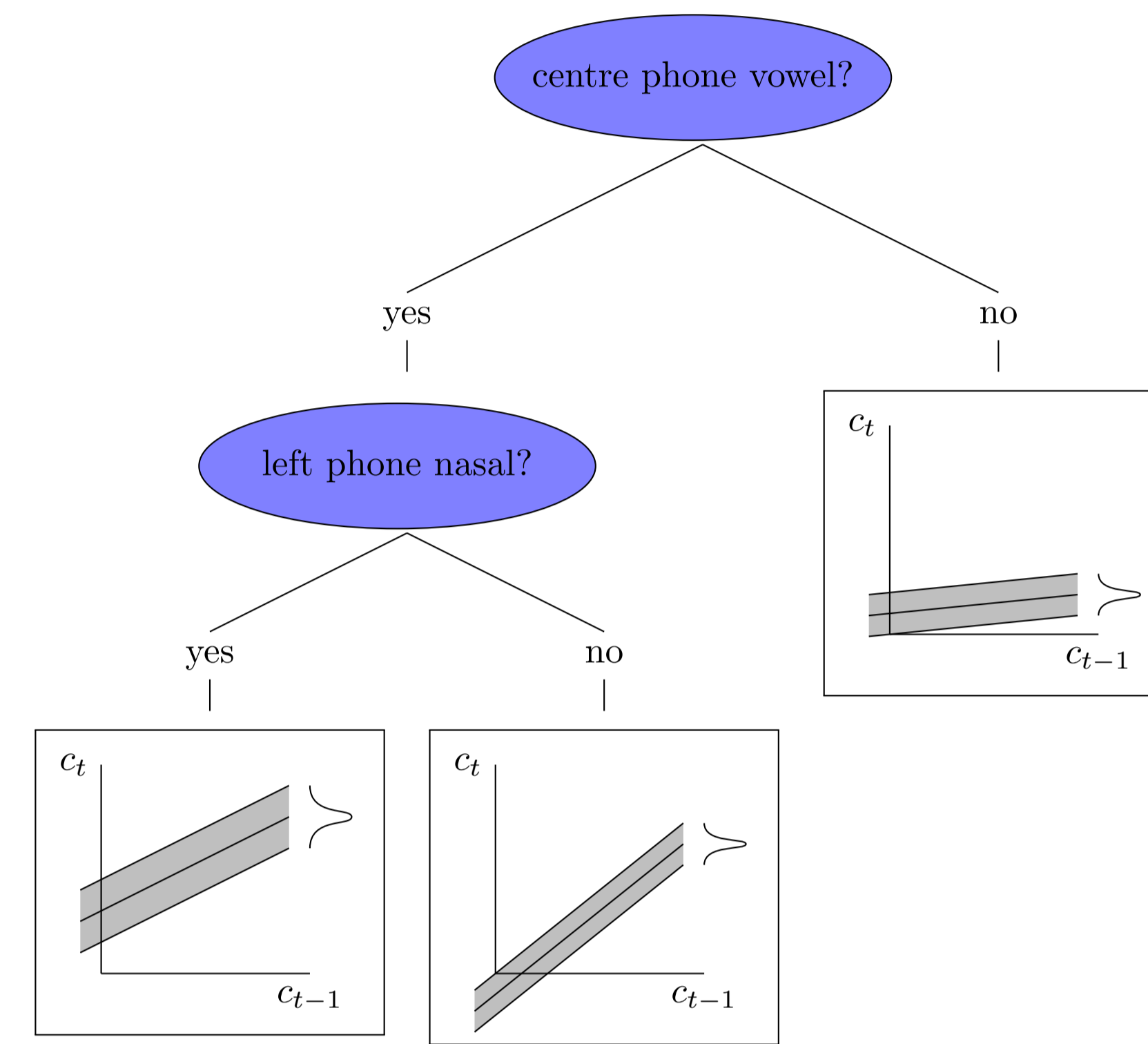
- ▶ a standard regression problem
- ⇒ can plug in **any** regression model

Autoregressive HMM advantages

- ▶ **consistent** modelling of dynamics of speech
 - ▶ standard HMM synthesis framework ignores static-dynamic constraints during training (Zen et al., 2004)
- ▶ **efficient** training using expectation-maximization
- ▶ **flexible** framework for further extensions
 - ▶ e.g. non-linear regression models

Decision tree clustering

- ▶ recursively subdivide state space using a predefined set of **questions**
- ▶ one state output distribution shared across all states in each leaf
- ▶ for the autoregressive HMM the state output distributions are models of $P(c_t | c_{t-K:t-1})$



Training:

- ▶ try to maximize **penalized likelihood**

$$\log P(c) - \xi \cdot \#\{\text{leaves}\}$$
- ▶ grow tree greedily starting with root node, by choosing question (or no question) that maximizes penalized likelihood
- ▶ **clustering threshold** ξ controls overall number of leaves
 - ▶ either set manually
 - ▶ or set according to **Minimum Description Length (MDL)** criterion (Shinoda and Watanabe, 2000)

$$\xi = \frac{1}{2} k \log N$$

$$k = \# \text{ free params per leaf, } N = \text{total } \# \text{ frames}$$

Experiments

- ▶ depth $K = 3$ (look at 3 previous frames)
- ▶ linear-Gaussian state output distributions (mean prediction is linear combination of previous frames)

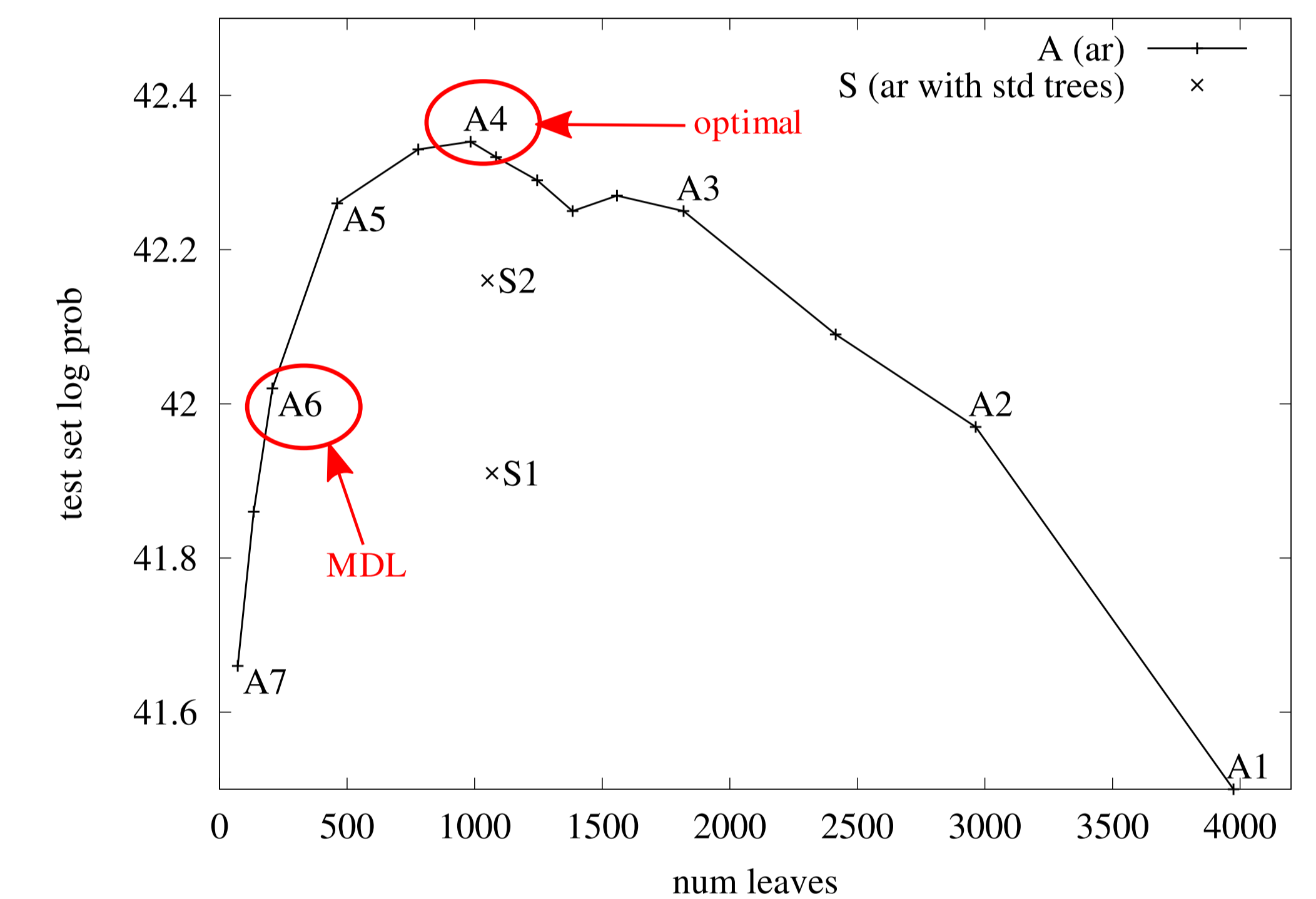
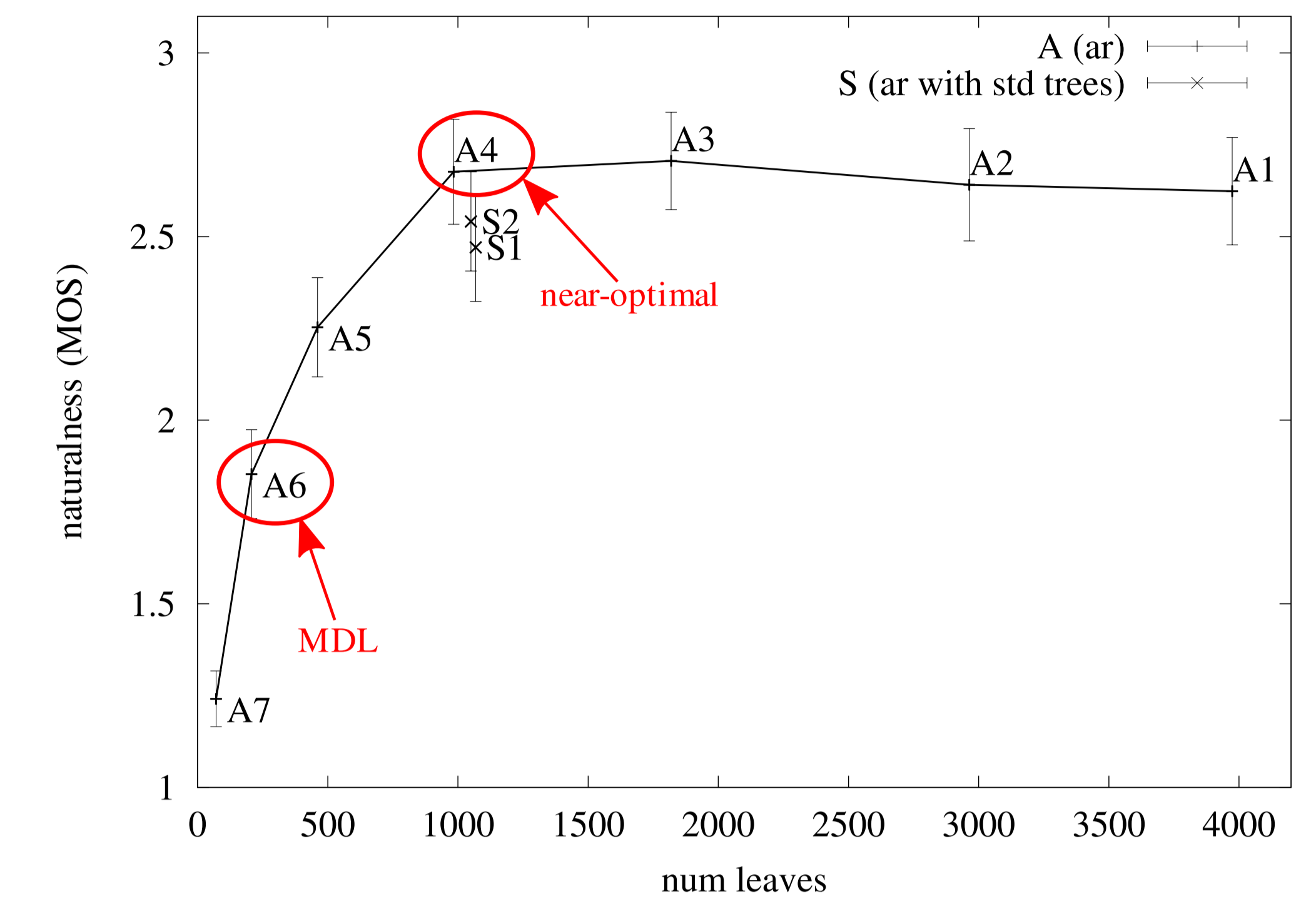
Compare:

- ▶ **baseline autoregressive systems (S1 and S2)** which re-use decision trees derived from a standard HMM system
- ▶ **autoregressive clustering systems (A1-7)** with various clustering thresholds ξ

Metrics:

- ▶ **naturalness** – opinion score in a subjective listening evaluation
 - ▶ 34 native English speakers, 50 utterances per listener
- ▶ **test set log probability** – log probability on a held-out test set

Results



- ▶ autoregressive clustering appears to give a small improvement in naturalness (opinion score median 2 to 3, mean 2.5 to 2.7)
- ▶ underfitting degrades naturalness
- ▶ overfitting well-tolerated
- ▶ MDL criterion not directly applicable to autoregressive HMM
- ▶ optimal model complexity gives near-optimal naturalness

Acknowledgements

- ▶ research funded by the European Community's Seventh Framework Programme (FP7/2007-2013), grant agreement 213845 (EMIME)
- ▶ many thanks to organizers of the Blizzard Challenge for providing scripts for our experimental evaluation

References

- M. Shannon and W. Byrne. Autoregressive HMMs for speech synthesis. In *Proc. Interspeech 2009*, 2009. <http://mi.eng.cam.ac.uk/~sms46/papers/shannon2009ahs.pdf>.
- K. Shinoda and T. Watanabe. MDL-based context-dependent subword modeling for speech recognition. *J. Acoust. Soc. Jpn. (E)*, 21(2):79–86, 2000.
- H. Zen, K. Tokuda, and T. Kitamura. An Introduction of Trajectory Model into HMM-Based Speech Synthesis. In *Proc. Fifth ISCA Workshop on Speech Synthesis*, 2004.