

# Autoregressive HMMs for speech synthesis

Matt Shannon   William Byrne



19 May 2010



- 1 Introduction
  - Introduction
- 2 Autoregressive HMM
  - Model
  - Advantages
- 3 Experiments
  - Autoregressive HMM
  - Autoregressive clustering
- 4 Summary

## Autoregressive HMM

- alternative to standard HMM synthesis framework
- modifies state output distributions
- provides a **consistent**, **efficient** and **flexible** framework for modelling speech

- 1 Introduction
  - Introduction
- 2 Autoregressive HMM
  - Model
  - Advantages
- 3 Experiments
  - Autoregressive HMM
  - Autoregressive clustering
- 4 Summary

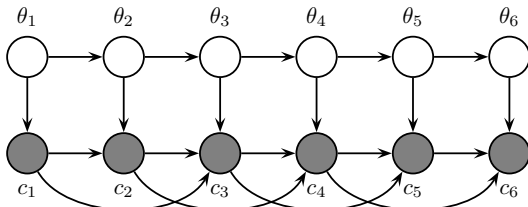
- hidden **state sequence**  $\theta = \theta_{1:T}$ 
  - e.g. states of full-context models (quinphones, POS, etc)
- observed acoustic **feature vector sequence**  $c = c_{1:T}$ 
  - e.g. 40-dim static mel-generalized cepstra

- hidden **state sequence**  $\theta = \theta_{1:T}$ 
  - e.g. states of full-context models (quinphones, POS, etc)
- observed acoustic **feature vector sequence**  $c = c_{1:T}$ 
  - e.g. 40-dim static mel-generalized cepstra

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

- hidden **state sequence**  $\theta = \theta_{1:T}$ 
  - e.g. states of full-context models (quinphones, POS, etc)
- observed acoustic **feature vector sequence**  $c = c_{1:T}$ 
  - e.g. 40-dim static mel-generalized cepstra

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



- turns problem of learning a model  $P(c|\theta)$
- into learning a function  $(c_{t-K:t-1}, \theta_t) \mapsto c_t$  from data:

$(c_{t-2}, c_{t-1}, \theta_t)$	$\mapsto$	$c_t$
(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
- $\Rightarrow$  can plug in **any** regression model

- consistent modelling of dynamics of speech
  - standard HMM synthesis framework ignores static-dynamic constraints during training
- efficient training using expectation-maximization
- synthesis using established excellent algorithms
  - e.g. synthesis considering global variance<sup>1</sup>
- flexible framework for further extensions
  - e.g. non-linear regression models

---

<sup>1</sup>T. Toda and K. Tokuda. Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis. In *Proc. Interspeech 2005*, 2005

- 1 Introduction
  - Introduction
- 2 Autoregressive HMM
  - Model
  - Advantages
- 3 Experiments
  - Autoregressive HMM
  - Autoregressive clustering
- 4 Summary

- depth  $K = 3$  (look at 3 previous frames)
- partition phonetic contexts ( $\theta_t$ ) using **decision tree**
- fit **linear regression model** in each region (each leaf node)
  - maps acoustic context  $c_{t-K:t-1}$  to acoustic output  $c_t$
- treat feature vector components as independent given state sequence (c.f. diagonal covariance matrices)
- generates speech of **comparable naturalness** to a standard HMM synthesis system (same MOS mean, median and box plot)<sup>2</sup>

---

<sup>2</sup>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>

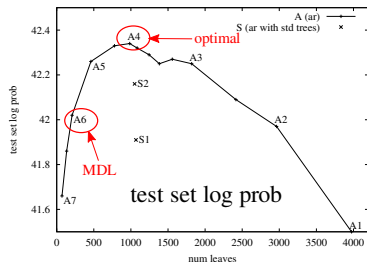
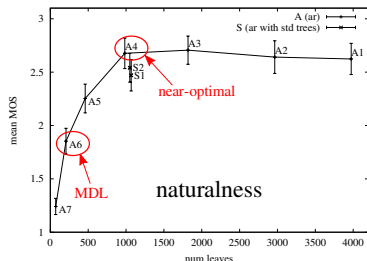
- decision tree clustering for the autoregressive HMM
  - previous experiments re-used trees from standard HMM system
- conceptually similar to standard case
- but need to pass accumulators to clustering algorithm
- improves naturalness slightly (MOS median 2 to 3, mean 2.5 to 2.7)<sup>3</sup>

---

<sup>3</sup>submitted to Interspeech 2010  
(<http://mi.eng.cam.ac.uk/~sms46/papers/shannon2010autoregressive-submitted.pdf>)

# Autoregressive clustering

- overfitting well-tolerated
- underfitting degrades naturalness
- minimum description length (MDL) criterion not directly applicable to autoregressive HMM
- optimal model complexity gives near-optimal naturalness
- (samples)



- 1 Introduction
  - Introduction
- 2 Autoregressive HMM
  - Model
  - Advantages
- 3 Experiments
  - Autoregressive HMM
  - Autoregressive clustering
- 4 Summary

- **consistent** treatment of static-dynamic constraints
- **efficient training** and **synthesis**
- **flexible** framework
- gives synthesized speech of **comparable naturalness** to standard HMM synthesis framework

- research funded by the European Community's Seventh Framework Programme (FP7/2007-2013), grant agreement 213845 (EMIME)
- we are very grateful to Matt Gibson for his substantial help in conducting the subjective listening evaluation, and to the organizers of the Blizzard Challenge for providing scripts to conduct this evaluation

- 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>.
- T. Toda and K. Tokuda. Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis. In *Proc. Interspeech 2005*, 2005.