

Autoregressive HMMs for speech synthesis

Matt Shannon William Byrne



Interspeech 2009



Outline

- 1 Introduction
 - Highlights
 - Background
- 2 Autoregressive HMM
 - Model
 - Training
 - Synthesis
- 3 Experiments
 - Experimental set-up
 - Results
- 4 Conclusion

Highlights

Autoregressive HMM for speech synthesis:

- synthesis with established **excellent synthesis algorithms**
- **consistent** – uses same model for training and synthesis
 - unlike standard HMM synthesis framework
- **easy and efficient training** using expectation maximization
 - unlike trajectory HMM
- **performance comparable** to standard HMM synthesis framework on Blizzard Challenge-style naturalness evaluation

Background

- HMM synthesis now rivals unit selection¹
- a key breakthrough – respecting **static-dynamic constraints** during synthesis²
- standard HMM synthesis framework
 - efficient EM training
 - but **inconsistent** – ignores static-dynamic constraints during training

¹A.W. Black, H. Zen, and K. Tokuda. Statistical parametric speech synthesis. In *Proc. ICASSP 2007*, pages 1229–1232, 2007

²K. Tokuda, T. Kobayashi, and S. Imai. Speech parameter generation from HMM using dynamic features. In *Proc. ICASSP 1995*, volume 1, 1995

Background (cont)

- trajectory HMM³ respects static-dynamic constraints during training
 - improved synthesis quality
 - consistent – uses same model for training and synthesis
 - but training more complicated
- remains a challenge to find a model that can **easily** and **consistently** be used for both training and synthesis
- we investigate the **autoregressive HMM**⁴⁵ for speech synthesis

³H. 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

⁴C. Wellekens. Explicit time correlation in hidden Markov models for speech recognition. In *Proc. ICASSP 1987*, volume 12, 1987

⁵P.C. Woodland. Hidden Markov models using vector linear prediction and discriminative output distributions. In *Proc. ICASSP 1992*, volume 1, pages 509–512, 1992

Outline

- 1 Introduction
 - Highlights
 - Background
- 2 Autoregressive HMM
 - Model
 - Training
 - Synthesis
- 3 Experiments
 - Experimental set-up
 - Results
- 4 Conclusion

The model

- 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

The model

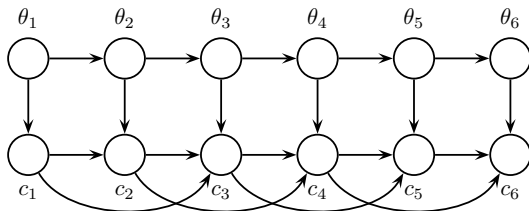
- hidden **state sequence** $\theta = \theta_{1:T}$
 - e.g. states of full-context models (quinphones, POS, etc)
- observed acoustic **feature vector sequence** $c = c_{1:T}$
 - e.g. 40-dim static Mel-generalized cepstra

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

The model

- 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

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



State output distributions

- state output distributions **conditional Gaussian**:

$$P(c_t | c_{1:t-1}, \theta_t) = \mathcal{N}(c_t | \mu_{\theta_t}(c_{1:t-1}), \Sigma_{\theta_t})$$

- mean functions** (μ_q) a linear map of a set of **summarizers** (f^d):

$$\mu_q(c_{1:t-1}) \triangleq \sum_{d=1}^D A_q^d f^d(c_{1:t-1}) + \mu_q^0$$

where each summarizer f^d gives a vector-valued **summary** of past output $c_{1:t-1}$

State output distributions (cont)

- we use summarizers (f^d) a linear combination of past output:

$$f^d(c_{1:t-1}) = \sum_{k=-K}^{-1} w_k^d c_{t+k}$$

call (w_k^d) **window coefficients**

- use diagonal matrices $A_{qij}^d = a_{qi}^d \delta_{ij}$ and $\Sigma_{qij} = \sigma_{qi}^2 \delta_{ij}$
- ⇒ feature vector seq components (c_i) independent given state seq θ

Example of state output distributions

For window set:

window	-3	-2	-1	0
w^1			1.0	
w^2		-1.0	1.0	
w^3	1.0	-2.0	1.0	

Example of state output distributions

For window set:

window	-3	-2	-1	0
w^1			1.0	
w^2		-1.0	1.0	
w^3	1.0	-2.0	1.0	

- state output distributions for feature vector index i :

$$P(c_{ti}|c_{1:t-1}, \theta_t) = \mathcal{N}(c_{ti}|\mu_{\theta_{ti}}(c_{1:t-1}), \sigma_{\theta_{ti}}^2)$$

- where mean functions:

$$\begin{aligned}\mu_{qi}(c_{1:t-1}) &= a_{qi}^1(c_{(t-1)i}) \\ &+ a_{qi}^2(c_{(t-1)i} - c_{(t-2)i}) \\ &+ a_{qi}^3(c_{(t-1)i} - 2c_{(t-2)i} + c_{(t-3)i}) \\ &+ \mu_{qi}^0\end{aligned}$$

Training

Expectation maximization:

- Forward-Backward algorithm for computing state occupancies $\gamma_q(t)$
- easy and efficient parameter re-estimation formulae (see paper)

Synthesis

For autoregressive HMM:

- $P(c|\theta)$ high-dimensional Gaussian over vector sequences
 - can efficiently compute mean and variance of this Gaussian
- ⇒ many **current synthesis algorithms directly applicable**:
- synthesis using dynamic features⁶
 - synthesis considering global variance⁷

⁶K. Tokuda, T. Kobayashi, and S. Imai. Speech parameter generation from HMM using dynamic features. In *Proc. ICASSP 1995*, volume 1, 1995

⁷T. Toda and K. Tokuda. Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis. In *Proc. Interspeech 2005*, 2005

Outline

- 1 Introduction
 - Highlights
 - Background
- 2 Autoregressive HMM
 - Model
 - Training
 - Synthesis
- 3 **Experiments**
 - Experimental set-up
 - Results
- 4 Conclusion

Experimental set-up

- Blizzard Challenge-style naturalness evaluation using MOS
- CMU ARCTIC database speaker slt (~ 1 hour)
- 4 systems:
 - natural speech
 - autoregressive HMM system (with synthesis considering GV)
 - baseline standard HMM synthesis framework system (with GV)
 - autoregressive HMM system (without GV)
- 50 utterances per listener
- 39 listeners completed (24 native, 15 non-native)

Systems in experiment

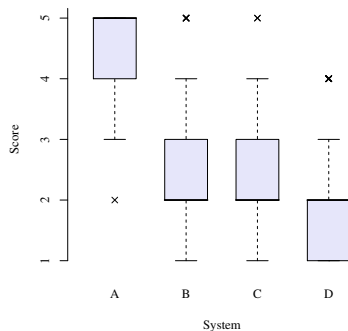
- even for autoregressive system, only the spectral features were modelled with the autoregressive HMM:

	AR system	standard system
spectral (MGC)	AR	standard
free params per state	5×40	6×40
log F0		standard
band aperiodicity		standard
clustering		standard(!)

- implemented in HTS (easy to adapt existing code!)

Results (native listeners)

system	native	
	mean	median
A (natural)	4.7	5
B (AR)	2.3	2
C (standard)	2.3	2
D (AR no GV)	2.0	2



Outline

- 1 Introduction
 - Highlights
 - Background
- 2 Autoregressive HMM
 - Model
 - Training
 - Synthesis
- 3 Experiments
 - Experimental set-up
 - Results
- 4 Conclusion

Conclusion

Autoregressive HMM for speech synthesis:

- **consistent** and **efficient** model for speech
 - ⇒ has advantages over standard HMM synthesis framework and trajectory HMM
- **comparable performance** to standard HMM synthesis framework on Blizzard Challenge-style naturalness evaluation
- easy to **adapt existing code** for autoregressive HMM

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 I

- A.W. Black, H. Zen, and K. Tokuda. Statistical parametric speech synthesis. In *Proc. ICASSP 2007*, pages 1229–1232, 2007.
- T. Toda and K. Tokuda. Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis. In *Proc. Interspeech 2005*, 2005.
- K. Tokuda, T. Kobayashi, and S. Imai. Speech parameter generation from HMM using dynamic features. In *Proc. ICASSP 1995*, volume 1, 1995.
- C. Wellekens. Explicit time correlation in hidden Markov models for speech recognition. In *Proc. ICASSP 1987*, volume 12, 1987.
- P.C. Woodland. Hidden Markov models using vector linear prediction and discriminative output distributions. In *Proc. ICASSP 1992*, volume 1, pages 509–512, 1992.
- 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.