Autoregressive HMMs for speech synthesis

Matt Shannon    William Byrne

19 May 2010
Outline

1. Introduction
   - Introduction

2. Autoregressive HMM
   - Model
   - Advantages

3. Experiments
   - Autoregressive HMM
   - Autoregressive clustering

4. Summary
Introduction

Autoregressive HMM

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

1. Introduction
   - Introduction

2. Autoregressive HMM
   - Model
   - Advantages

3. Experiments
   - Autoregressive HMM
   - Autoregressive clustering

4. Summary
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
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

\[
P(c, \theta) = \prod_t P(\theta_t | \theta_{t-1}) \cdot P(c_t | c_{t-K:t-1}, \theta_t)
\]

transition probs  state output dist
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

\[
P(c, \theta) = \prod_t P(\theta_t | \theta_{t-1}) P(c_t | c_{t-K:t-1}, \theta_t)
\]

![Diagram of a hidden Markov model with states and observed acoustic features.]

\( \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6 \)
\( c_1, c_2, c_3, c_4, c_5, c_6 \)
Model

- 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:

\[
\begin{array}{c|c}
(c_{t-2}, c_{t-1}, \theta_t) & c_t \\
\hline
(1.0, 1.3, k-aa+t) & 1.6 \\
(1.3, 1.6, k-aa+t) & 2.0 \\
(1.6, 2.0, aa-t+s) & 1.8 \\
\end{array}
\]

- a standard regression problem
- can plug in any regression model
Advantages

- 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\(^1\)
- flexible framework for further extensions
  - e.g. non-linear regression models

Outline

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)$^2$

---

Autoregressive clustering

- 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)\(^3\)

\(^3\)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)
Outline

1 Introduction
   - Introduction

2 Autoregressive HMM
   - Model
   - Advantages

3 Experiments
   - Autoregressive HMM
   - Autoregressive clustering

4 Summary
Summary

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

- 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