

The Effect of Using Normalized Models in Statistical Speech Synthesis

Matt Shannon¹ Heiga Zen² William Byrne¹

¹University of Cambridge



UNIVERSITY OF
CAMBRIDGE



²Toshiba Research Europe Ltd

TOSHIBA
Leading Innovation >>>

Interspeech 2011

Outline

Introduction

- Overview

- Predictive distribution

- Normalized models

Effect of normalization

- Plot sampled trajectories

- Test set log probs

All existing models are less than satisfactory

Improving the model

Summary

Overview

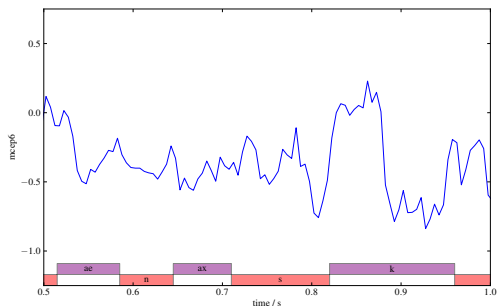
- ▶ standard approach to HMM-based speech synthesis is inconsistent
 - ▶ static and dynamic feature sequences are deterministically related (can compute one from the other)
 - ▶ this relationship taken into account during synthesis
 - ▶ but ignored during training
 - ▶ static and dynamic feature sequences treated as conditionally independent
 - ▶ so model assigns most of its probability mass to things that can never happen (sequences where the statics and dynamics don't match up)
- ▶ in fact, model used during training can be viewed as an **unnormalized** version of the model used during synthesis (see paper for details)
 - ▶ unnormalized means probabilities don't sum to 1
- ▶ this paper looks at **what effect this lack of normalization during training has on the predictive probability distribution used during synthesis**

Overview

- ▶ standard approach to HMM-based speech synthesis is inconsistent
 - ▶ static and dynamic feature sequences are deterministically related (can compute one from the other)
 - ▶ this relationship taken into account during synthesis
 - ▶ but ignored during training
 - ▶ static and dynamic feature sequences treated as conditionally independent
 - ▶ so model assigns most of its probability mass to things that can never happen (sequences where the statics and dynamics don't match up)
- ▶ in fact, model used during training can be viewed as an **unnormalized** version of the model used during synthesis (see paper for details)
 - ▶ unnormalized means probabilities don't sum to 1
- ▶ this paper looks at **what effect this lack of normalization during training has on the predictive probability distribution used during synthesis**
- ▶ working assumption is that we care about having an accurate predictive distribution

Predictive distribution

- ▶ audio represented as a sequence of feature vectors ($40 \times T$ matrix)
- ▶ for simplicity of visualization, we will focus on one component of this feature vector, say mcep6
- ▶ let \mathbf{c} be the **trajectory** of mcep6 values over time (T -dim vector)

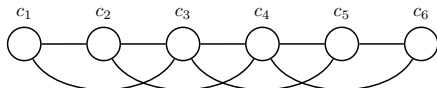


- ▶ let \mathbf{q} be the hidden state sequence (sequence of T states)
- ▶ **predictive distribution** $P(\mathbf{c}|\mathbf{q}, \lambda)$ is a distribution over trajectories

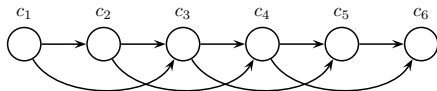
Normalized models

- ▶ there have been attempts to rectify the inconsistency present in the standard approach by using the same model for training and synthesis

- ▶ trajectory HMM¹ (globally normalized)



- ▶ autoregressive HMM² (locally normalized)



- ▶ in fact, the trajectory HMM can be viewed as precisely the model used during synthesis in the standard approach
- ▶ for both these models predictive distribution $P(\mathbf{c}|\mathbf{q}, \lambda)$ is Gaussian (T -dimensional distribution over trajectories)

¹H. Zen, K. Tokuda, and T. Kitamura. Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic features. *Computer Speech and Language*, 21(1):153–173s, 2007

²M. Shannon and W. Byrne. Autoregressive HMMs for speech synthesis. In *Proc. Interspeech 2009*, pages 400–403, 2009

Outline

Introduction

- Overview

- Predictive distribution

- Normalized models

Effect of normalization

- Plot sampled trajectories

- Test set log probs

All existing models are less than satisfactory

Improving the model

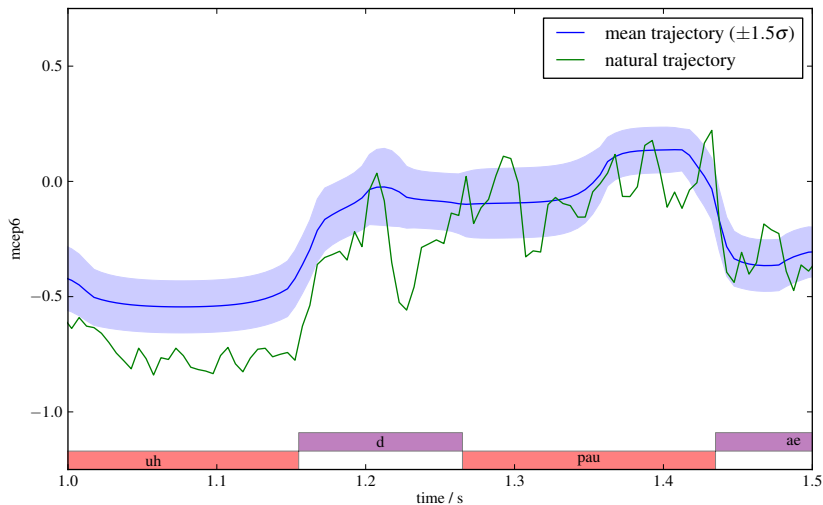
Summary

Effect of normalization

- ▶ to investigate the effect of normalization we compare
 - ▶ standard approach (unnormalized)
 - ▶ trajectory HMM (normalized)
 - ▶ autoregressive HMM (normalized)
- ▶ we compare their predictive distributions in a few ways
 - ▶ (subjective) visualize predictive distribution
 - ▶ plot mean trajectory with pointwise variance
 - ▶ plot sampled trajectories
 - ▶ (objective) compute test set log probs

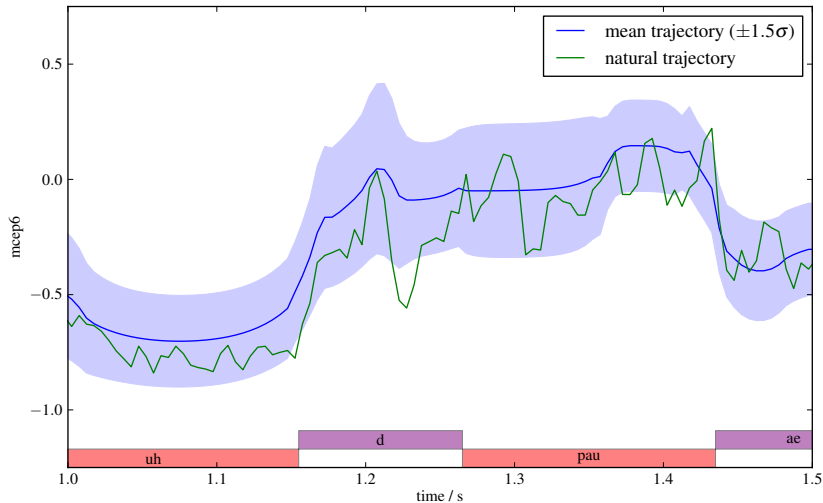
Mean trajectory with pointwise variance

Standard HTS training (unnormalized)



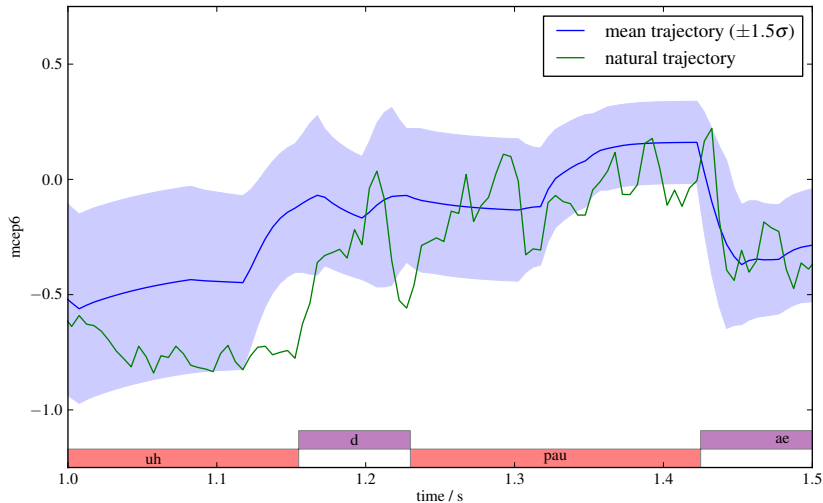
Mean trajectory with pointwise variance

Trajectory HMM (normalized)



Mean trajectory with pointwise variance

Autoregressive HMM (normalized)



Mean trajectory with pointwise variance

We can see

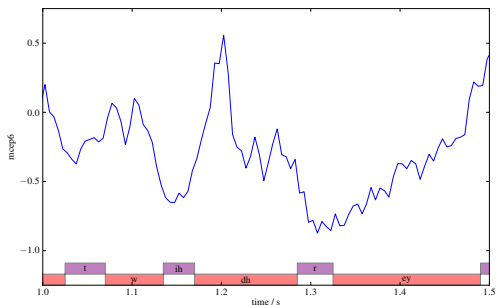
- ▶ variance for the unnormalized standard approach appears to be too small
- ▶ the variance for the normalized models is larger, and looks more reasonable
- ▶ normalization also changes the mean trajectory, though this depends more on the precise form of normalization used

Plot sampled trajectories

- ▶ another way to investigate predictive dist is to **sample** from it
 - ▶ maximum likelihood training implicitly assumes speaker generated the training corpus by sampling trajectory from $P(\mathbf{c}|\mathbf{q}, \lambda)$
- ⇒ a good way to assess accuracy of probabilistic model is to sample from our trained model $P(\mathbf{c}|\mathbf{q}, \lambda)$ and compare these samples to natural trajectories

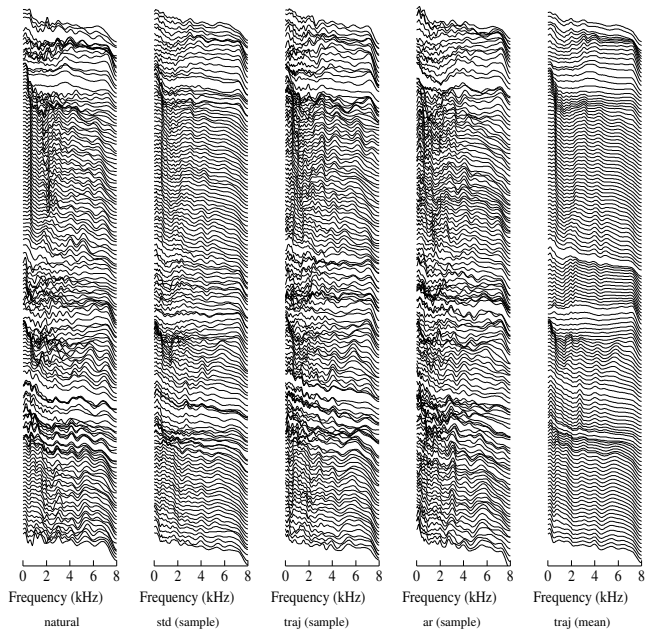
Plot sampled trajectories

- ▶ another way to investigate predictive dist is to **sample** from it
- ▶ maximum likelihood training implicitly assumes speaker generated the training corpus by sampling trajectory from $P(\mathbf{c}|\mathbf{q}, \lambda)$
- ⇒ a good way to assess accuracy of probabilistic model is to sample from our trained model $P(\mathbf{c}|\mathbf{q}, \lambda)$ and compare these samples to natural trajectories
- ▶ example of sampled trajectory



- ▶ in fact, we plot running spectra instead so we can get an idea of what's happening to all mcep components at once

Plot sampled trajectories



Plot sampled trajectories

We can see

- ▶ sampled trajectories from normalized models qualitatively similar to natural trajectories
 - ▶ same characteristic roughness
- ▶ sampled trajectories from unnormalized standard approach slightly too smooth
 - ▶ since standard approach underestimates predictive variance
- ▶ mean trajectory is much too smooth
 - ▶ expected, since maximum likelihood training assumes natural trajectories generated by sampling

Test set log probs

- ▶ compute log prob on a held-out test set of 50 utterances
- ▶ natural metric to evaluate **probabilistic** models

system	log prob
standard	29.3
trajectory HMM	47.6
autoregressive HMM	47.8

Test set log probs

- ▶ compute log prob on a held-out test set of 50 utterances
- ▶ natural metric to evaluate **probabilistic** models

system	log prob
standard	29.3
trajectory HMM	47.6
autoregressive HMM	47.8

- ▶ normalized models have **much** better test set log prob
- ⇒ suggests normalized models better probabilistic models of speech

Test set log probs

- ▶ compute log prob on a held-out test set of 50 utterances
- ▶ natural metric to evaluate **probabilistic** models

system	log prob
standard	29.3
trajectory HMM	47.6
autoregressive HMM	47.8

- ▶ normalized models have **much** better test set log prob
- ⇒ suggests normalized models better probabilistic models of speech
- ▶ why is score for standard approach score **so** low?
 - ▶ if we artificially boost predictive variance by a factor of 3, standard approach test set log prob goes to 46.9
- ⇒ standard approach systematically underestimates predictive variance

Outline

Introduction

- Overview

- Predictive distribution

- Normalized models

Effect of normalization

- Plot sampled trajectories

- Test set log probs

All existing models are less than satisfactory

Improving the model

Summary

All existing models are less than satisfactory

Sampled trajectories from the normalized models we have currently

- ▶ look more like natural speech than mean trajectories
- ▶ have some nice properties
 - ▶ e.g. sampled trajectories from these normalized models have almost completely natural global variance distributions, without using any additional global variance modelling
- ▶ sound terrible (!)
 - ▶ traj HMM with GV
 - ▶ traj HMM mean
 - ▶ traj HMM sampled

All existing models are less than satisfactory

Sampled trajectories from the normalized models we have currently

- ▶ look more like natural speech than mean trajectories
 - ▶ have some nice properties
 - ▶ e.g. sampled trajectories from these normalized models have almost completely natural global variance distributions, without using any additional global variance modelling
 - ▶ sound terrible (!)
 - ▶ traj HMM with GV
 - ▶ traj HMM mean
 - ▶ traj HMM sampled
- ⇒ existing models are not modelling something they should be modelling

Outline

Introduction

- Overview

- Predictive distribution

- Normalized models

Effect of normalization

- Plot sampled trajectories

- Test set log probs

All existing models are less than satisfactory

Improving the model

Summary

Improving the model

- ▶ we looked at one possible improvement to model
- ▶ trajectory HMM with full covariance matrices
 - ▶ explicitly model correlation between different feature vector components within one frame
 - ▶ these intra-frame correlations ignored by current normalized models
- ▶ subjective listening test results

system	trajectory	MOS
diag cov traj HMM	sampled	1.7
full cov traj HMM	sampled	2.0
full cov traj HMM	mean	3.4

Improving the model

- ▶ we looked at one possible improvement to model
- ▶ trajectory HMM with full covariance matrices
 - ▶ explicitly model correlation between different feature vector components within one frame
 - ▶ these intra-frame correlations ignored by current normalized models
- ▶ subjective listening test results

system	trajectory	MOS
diag cov traj HMM	sampled	1.7
full cov traj HMM	sampled	2.0
full cov traj HMM	mean	3.4

- ▶ using full covariance matrices does improve naturalness of sampled trajectories
 - ▶ but sampled trajectories still sound bad, and much worse than mean trajectory
- ⇒ full covariance trajectory HMM is a better probabilistic model of speech, but still not a good one!

Outline

Introduction

- Overview

- Predictive distribution

- Normalized models

Effect of normalization

- Plot sampled trajectories

- Test set log probs

All existing models are less than satisfactory

Improving the model

Summary

Summary

To summarize

- ▶ standard model used during training is **unnormalized**
 - ▶ **normalization** (trajectory HMM, autoregressive HMM) results in a better predictive distribution
 - ▶ subjectively better
 - ▶ the natural trajectory is massively outside the expected range less often with normalized models
 - ▶ sampled trajectories from normalized models look qualitatively similar to natural trajectories, whereas sampled trajectories from the standard approach are slightly too smooth
 - ▶ objectively better
 - ▶ greatly increases test set log probability
 - ▶ modelling intra-frame correlations further improves the predictive distribution
 - ▶ however for all models the predictive distribution is not good, and the models are far from satisfactory probabilistic models of speech
- ⇒ more work needed if we want good predictive distributions

References I

- M. Shannon and W. Byrne. Autoregressive HMMs for speech synthesis. In *Proc. Interspeech 2009*, pages 400–403, 2009.
- H. Zen, K. Tokuda, and T. Kitamura. Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic features. *Computer Speech and Language*, 21 (1):153–173s, 2007.