The effect of normalization – a case study in speech synthesis
Cambridge machine learning RCC

Matt Shannon

UNIVERSITY OF CAMBRIDGE

3 March 2011
Outline

Overview of statistical speech synthesis

Acoustic models

Interlude – guess the fake

Effect of normalization

Synthesis

Summary
Overview

Overall goal of speech synthesis

▶ convert text
  ▶ sequence of words
    [ his, blood, grew, hot, with, rage, at, the, thought ]
▶ to speech
  ▶ waveform
Overview

Overall goal of speech synthesis

- convert text
  - sequence of words
    - [his, blood, grew, hot, with, rage, at, the, thought]
- to speech
  - waveform

In statistical speech synthesis

- probabilistic model of speech given text
- training corpus of (word sequence, waveform) examples from a single speaker (e.g. 4000 sentences)
- use model to generate waveform for new, unseen word sequences
- ideal system is one that mimics original speaker exactly
Overview

Speech

► turns out it’s hard to model the waveform directly
⇒ convert waveform into a sequence of 40-dimensional feature vectors
► one vector every 5 milliseconds
► resynthesis algorithm to approximately reconstruct waveform from feature vector sequence
  ▶ resynthesized speech sounds extremely natural
  ▶ (can even do things properly and model waveform given feature vector sequence probabilistically\(^1\) though we won’t talk about this)
► notation
  ▶ \(c_t\) is feature vector at time step \(t\)
  ▶ \(c = c_{1:T}\) is the entire feature vector sequence

Overview

Text

- words are too big a unit (will never see all words)
- break words into sub-word units called phones
  - finite set (e.g. 41 phones)
  - e.g. “hot” has phones [hh, aa, t]
- assume precisely one phone is the ‘current phone’ at each time $t$
  - not much of a restriction
  - pau for silences

- notation
  - $l_j$ is $j^{th}$ phone
    - e.g. $l = [hh, aa, t]$
  - $j_t$ is index of current phone at time step $t$
    - e.g. $j = [0, \ldots, 0, 1, \ldots, 1, 2, \ldots, 2]$
      - 21 18 4
    - so $q_t \triangleq l_{j_t}$ is the ‘current phone’ at time $t$
To make this more concrete, here is component 6 of the feature vector over time

Labels at the bottom are phones
Overview

![Graph](image-url)
Overview

High-level generative model

- given word sequence
  
  \[ \text{[ his, blood, grew, hot, ...]} \]

- compute sequence of phones \( l = l_{1:j} \) (deterministic function)
  
  \[ l = [ \text{pau, hh, ih, z, b, l, ah, d, g, r, uw, hh, aa, t, ...} ] \]

- sample segmentation \( j | l, \nu \) (duration model)
  
  \[ j = [0, \ldots, 0, 1, \ldots, 1, 2, \ldots, 2, 3, \ldots, 3, \ldots] \]

- this determines phone-sequence-with-timings \( q_t \)
  
  \[ q = [ \text{pau, \ldots, pau, hh, \ldots, hh, ih, \ldots, ih, z, \ldots, z, \ldots} ] \]

- sample feature vector sequence \( c | q, \lambda \) (acoustic model)
Overview

High-level generative model

- given word sequence
  \[ \text{[ his, blood, grew, hot, ...]} \]
- compute sequence of phones \( l = l_1[:J] \) (deterministic function)
  \[ l = [\text{pau, hh, ih, z, b, 1, ah, d, g, r, uw, hh, aa, t, ...]} \]
- sample segmentation \( j|l, \nu \) (duration model)
  \( j = [0, \ldots, 0, 1, \ldots, 1, 2, \ldots, 2, 3, \ldots, 3, \ldots] \)
  \[ \begin{array}{cccc}
    & 31 & 18 & 5 & 14 \\
  \end{array} \]
- this determines phone-sequence-with-timings \( q_t \)
  \( q = [\text{pau, ...}, \text{pau, hh, ...}, \text{hh, ih, ...}, \text{ih, z, ...}, \text{z, ...}] \)
  \[ \begin{array}{cccc}
    & 31 & 18 & 5 & 14 \\
  \end{array} \]
- sample feature vector sequence \( c|q, \lambda \) (acoustic model)

So

- duration model \( P(j|l, \nu) \) models segmentation over time
- acoustic model \( P(c|q, \lambda) \) models acoustics given this segmentation
we won’t talk about duration model $P(j|l, \nu)$ today

in fact we will mostly assume the segmentation $j$ is known

but worth knowing that in practice we use a Markov chain as the segmentation process

constructed so that the duration of each phone has approximately correct distribution
Overview

Two provisos about text

▶ label $l_j$ for $j^{th}$ phone actually includes richer context, including current phone, previous phone, next phone and more
  ▶ provides better modelling of transitions between phones and of how acoustics change depending on phonetic context
  ▶ simple e.g. $l_j = (hh, ih, z)$ instead of $l_j = ih$

▶ actually break each phone down into 5 sub-phone units
  ▶ allows ‘start of $ih$’ to be modelled differently to ‘middle of $ih$’ for example
  ▶ simple e.g. $j_t = (2, 3)$ instead of $j_t = 2$
  ▶ finer-grained segmentation
  ▶ very important for good modelling

▶ $q_t$ takes into account these richer contexts
  ▶ simple e.g. $q_t = ((hh, ih, z), 2)$ instead of $q_t = ih$
  ▶ idea is that $q_t$ contains all phonetic context that might be relevant for the acoustics at or near time $t$
  ▶ $q$ is called state sequence (terrible name!)
Overview

Two provisos about speech

▶ feature vector $c_t$ actually includes 0/1-dimensional pitch and 5-dimensional aperiodicity information in addition to 40-dimensional spectral information

▶ typically assume different components of the feature vector sequence are independent given the state sequence $q$

$$P(c_{1:T}|q) = \prod_{i=1}^{40} P(c^i_{1:T}|q)$$

▶ not actually a bad assumption (special property of this feature vector representation)

For notational clarity from now on focus on a single component (e.g. component 6), so

▶ $c_t$ is a scalar value

▶ $c$ is a sequence of scalar values (a trajectory)
Overview

Summary

- represent speech as feature vector sequence $c$
- represent text as label sequence $l$ (e.g. each label is a phone)
- two-level generative model for $P(c|l)$
  - duration model $P(j|l, \nu)$ models segmentation across time
    - Markov chain
  - acoustic model $P(c|q, \lambda)$ models acoustics given this segmentation
    - haven't seen an acoustic model yet
    - idea is that dynamics near time $t$ depend on the current phonetic context $q_t$
    - simple example of phonetic context $q_t = ((hh, ih, z), 2)$
Overview

- "mcep6" graph with time axis in seconds.
- Key points: "pau", "hh", "z", "l", "d", "r", "hh", "pau", "ih", "b", "ah", "g", "uw", "aa".
- Time range: 0.0 to 1.0 seconds.
Outline

Overview of statistical speech synthesis

Acoustic models

Interlude – guess the fake

Effect of normalization

Synthesis

Summary
Acoustic models

More broadly, goal is

- come up with a good model of scalar sequence data $c = c_{1:T}$
- where dynamics of sequence near time $t$ depend on a discrete-valued quantity $q_t$

How would you model this?
Acoustic models

There are several acoustic models $P(c|q, \lambda)$ in common use:

1. directed graphical model (dependence on $q$ not shown)

\[
P(c|q, \lambda) = \prod_{t} P(c_t|q_t, c_{t-K:t-1}, \lambda)
\]

2. undirected graphical model (dependence on $q$ not shown)

\[
P(c|q, \lambda) = \frac{1}{Z(q, \lambda)} \prod_{t} \psi(c_{t-K:t+K}; q_t, \lambda)
\]

3. unnormalized version of above undirected graphical model

\[
"P"(c|q, \lambda) = \prod_{t} \psi(c_{t-K:t+K}; q_t, \lambda)
\]
Acoustic models

[Discussion of directed vs undirected]
Acoustic models

Directed vs undirected graphical models

- directed models explicitly model an imagined generative process (advantage)
- directed models explicitly model an imagined generative process (disadvantage)
- directed models allow easy sampling (ancestral)
- directed models easier to make Bayesian
- normalization constant in undirected models often makes inference hard or intractable
- undirected models more naturally cope with a set of soft constraints all of which should be simultaneously roughly satisfied
Acoustic models

[Discussion of normalized vs unnormalized]
Acoustic models

Normalized vs unnormalized graphical models

▶ unnormalized models often more tractable
▶ normalized models much better justified theoretically
▶ lose guarantee that you’re training method is doing anything sensible if training method is justified probabilistically but you’re using a non-probabilistic model

In practice unnormalized model is often just used during training, and a normalized distribution is used when making predictions. In this case

▶ link with product-of-experts
▶ might expect product-of-experts models where we train unnormalized to be over-confident, since experts modelling the same thing multiple times but don’t realize it
  ▶ simple example
Acoustic models

A bit more detail on existing models

- directed graphical model $P(c|q, \lambda) = \prod_t P(c_t|q_t, c_{t-K:t-1}, \lambda)$
  - locally normalized
  - $P(c|q, \lambda)$ factorizes over time with respect to $q$
  - factors $P(c_t|q_t, c_{t-K:t-1}, \lambda)$ typically linear Gaussian
  - called the autoregressive HMM\(^2\)

---


Acoustic models

A bit more detail on existing models

- directed graphical model $P(c|q, \lambda) = \prod_t P(c_t|q_t, c_{t-K:t-1}, \lambda)$
  - locally normalized
  - $P(c|q, \lambda)$ factorizes over time with respect to $q$
  - factors $P(c_t|q_t, c_{t-K:t-1}, \lambda)$ typically linear Gaussian
  - called the autoregressive HMM\(^2\)

- undirected graphical model $P(c|q, \lambda) = \frac{1}{Z(q, \lambda)} \prod_t \psi(c_{t-K:t+K}; q_t, \lambda)$
  - globally normalized at the level of $c$
  - normalization constant $Z$ depends on entire state sequence $q_{1:T}$ -- does not factorize over time with respect to $q$
  - factors $\psi(c_{t-K:t+K}; q_t, \lambda)$ typically Gaussian in $c_{t-K:t+K}$
  - called the trajectory HMM\(^3\)


Acoustic models

A bit more detail on the standard model

- unnormalized undirected graphical model
  \[ P(c|q, \lambda) = \prod_t \psi(c_{t-K:t+K}; q_t, \lambda) \]
  - not normalized
  - \( P(c|q, \lambda) \) factorizes over time with respect to \( q \)
  - factors \( \psi(c_{t-K:t+K}; q_t, \lambda) \) are Gaussian as for normalized undirected model
  - standard model used for speech synthesis

In fact

- even in standard approach we use the normalized model during synthesis
  - otherwise can’t talk about predictive distribution at all
- unnormalized model is just used during training

Note that in all three cases the overall distribution \( P(c|q, \lambda) \) is Gaussian
Outline

Overview of statistical speech synthesis

Acoustic models

Interlude – guess the fake

Effect of normalization

Synthesis

Summary
Interlude – guess the fake
Interlude – guess the fake

natural

undirected, mean

unnormalized undirected, mean
Interlude – guess the fake

undirected, mean

natural

directed, mean
Interlude – guess the fake

1.0 1.2 1.4 1.6 1.8 2.0
time / s

0.5
0.0
-0.5
-1.0

undirected, sampled
Interlude – guess the fake

undirected, sampled

natural

unnormalized undirected, sampled
Interlude – guess the fake
Interlude – guess the fake

undirected, sampled

natural

directed, sampled
Outline

Overview of statistical speech synthesis

Acoustic models

Interlude – guess the fake

Effect of normalization

Synthesis

Summary
Effect of normalization

How does normalization affect the trained models?
▶ plot the distribution over trajectories $P(c|q, \lambda)$ for some real utterances
▶ compare to natural trajectory

Technical details
▶ mcep6 (7th Mel-cepstral component)
▶ 1 second of speech
▶ synthesis given standard CMU ARCTIC phone-level transcription
▶ plot mean trajectory $\pm 1.5$ standard deviation, and natural trajectory
▶ (N.B. correlations over time not represented in this picture)
Effect of normalization

Unnormalized (standard HTS training)
Effect of normalization

Normalized (trajectory HMM)
Effect of normalization

Normalized (autoregressive HMM)
Effect of normalization

We can see

- the variance of the distribution over trajectories for the unnormalized model is too small (over-confident)
- the variance for the normalized models is larger, and looks more reasonable
- this is reflected in probabilities – log prob per frame of the natural trajectory is
  - 0.3 (unnormalized HMM)
  - 0.9 (trajectory HMM)
  - 0.9 (autoregressive HMM)
- normalization also changes the mean trajectory
  - at least for the trajectory HMM, improves naturalness of synthesized mean trajectories

---

Outline

Overview of statistical speech synthesis

Acoustic models

Interlude – guess the fake

Effect of normalization

Synthesis

Summary
Synthesis

Haven’t yet talked about how synthesis is done

- have an acoustic model $P(c|q, \lambda)$
- have estimated $\lambda$ or posterior distribution over $\lambda$
- want to synthesize acoustics $c$ for a new, unseen word sequence
Synthesis

Haven’t yet talked about how synthesis is done
- have an acoustic model $P(c|q, \lambda)$
- have estimated $\lambda$ or posterior distribution over $\lambda$
- want to synthesize acoustics $c$ for a new, unseen word sequence

One way (the right way)
- we have a two-level generative model
- so first sample segmentation from duration model
- then sample acoustics given this segmentation by sampling from acoustic model $P(c|q, \lambda)$
Synthesis

However, standard approach to synthesis:

- obtain segmentation by choosing the most likely duration for each segment
- generate $c$ by maximizing $P(c|q, \lambda)$
- so we choose the most like trajectory instead of sampling a random trajectory

Not well-justified theoretically:

- our stated goal was to imitate the original speaker exactly (to extend the training corpus without anyone realizing)
- our assumption during training is that the training corpus was generated by the speaker sampling (independently) from $P(c|q, \lambda)$ for each utterance

⇒ should really do synthesis by sampling trajectory

- also, implies the system says the same word sequence exactly the same every time, whereas the original speaker it’s trying to mimic doesn’t
Synthesis

The standard approach also means
- mean trajectories look very unrealistic – much too smooth

![Graph showing mean trajectories for natural and traj HMM mean comparisons.](image-url)
In my view

- the fact the mean trajectory sounds over-smoothed is not a sign of anything going wrong – we would probably expect the mean trajectory to be smoother than any given random trajectory
- the random part of the probability distribution over trajectories should be aiming to capture the speaker’s natural variability – the speaker says the same label sequence slightly differently each time they say it
Synthesis

Sampled trajectories certainly capture the characteristic roughness of natural trajectories.
Synthesis

Sampled trajectories from the normalized models we have currently

- look more like natural speech than mean trajectories
- have some nice properties
  - there is a standard hack used to boost global variance (GV) of trajectories
  - this is necessary since it is found that the GV of trajectories generated by the standard approach is smaller than the GV of natural trajectories
  - however sampled trajectories from normalized models automatically have the same distribution over GV as natural trajectories
- sound terrible (!)
  - undirected model with global variance hack
  - undirected model mean
  - undirected model sampled
Synthesis

Sampled trajectories from the normalized models we have currently

- look more like natural speech than mean trajectories
- have some nice properties
  - there is a standard hack used to boost global variance (GV) of trajectories
  - this is necessary since it is found that the GV of trajectories generated by the standard approach is smaller than the GV of natural trajectories
  - however sampled trajectories from normalized models automatically have the same distribution over GV as natural trajectories
- sound terrible (!)
  - undirected model with global variance hack
  - undirected model mean
  - undirected model sampled

⇒ existing models are not modelling something they should be modelling
Synthesis

Sampled trajectories from the normalized models we have currently

- look more like natural speech than mean trajectories
- have some nice properties
  - there is a standard hack used to boost *global variance (GV)* of trajectories
  - this is necessary since it is found that the GV of trajectories generated by the standard approach is smaller than the GV of natural trajectories
  - however sampled trajectories from normalized models automatically have the same distribution over GV as natural trajectories
- sound terrible (!)
  - undirected model with global variance hack
  - undirected model mean
  - undirected model sampled

⇒ existing models are not modelling something they should be modelling

- and it seems to be something *low-level* – instantly noticeable and uniform over the utterance, not some complicated contextual effect
Outline

Overview of statistical speech synthesis

Acoustic models

Interlude – guess the fake

Effect of normalization

Synthesis

Summary
Summary

To summarize

- standard model used during training is unnormalized
- normalization (trajectory HMM, autoregressive HMM) results in a better distribution over trajectories
  - theoretically more consistent
    - uses the same normalized model for training and synthesis
  - subjectively better
    - sampled trajectories from normalized models have many large rises and falls, just like natural trajectories, whereas sampled trajectories from the standard model are slightly too tame
    - the natural trajectory is massively outside the expected range less often with normalized models
  - objectively better
    - greatly increases test set log probability
need to **sample trajectories** to take full advantage of the better covariance present in normalized models

- theoretically the right thing to do
- generates much more natural looking trajectories
- sounds terrible (!)
- existing models (standard HMM, trajectory HMM, autoregressive HMM) are all failing to capture some important low-level aspect of speech
A unified view of current normalized models (optional)

Can distinguish two inter-related aspects to modelling $c|q$ well

1. model the random variation present for fixed $q$
   - imagine we fix the state sequence $q$ once and for all
   - just try to model the variability in the way speaker says the utterance
   - not necessarily easy!

2. model the way the overall distribution $P(c|q, \lambda)$ over $c$ depends on the individual states $q_t$ at each time $t$
   - expect state at time $t$ to have a local effect on trajectory – i.e. affect mainly $c_{t-L:t+L}$ for some $L$
   - the overlapping local effects of states near each other in the state sequence should interact in such a way that even unseen state sequences result in a sensible overall distribution $P(c|q, \lambda)$

How do current normalized models approach these two aspects?
A unified view of current normalized models (optional)

1. model the random variation present for fixed $q$

2. model the way the overall distribution over $c$ depends on the individual states $q_t$ at each time $t$
1. model the random variation present for fixed $q$
   ▶ assume $c|q$ is a Gaussian, i.e. $c|q \sim \mathcal{N}(\mu_q, \Sigma_q)$
   ▶ Gaussian is over time ($c$ is a $T$-dimensional vector)
   ▶ $\mu_q$ is mean trajectory

2. model the way the overall distribution over $c$ depends on the individual states $q_t$ at each time $t$
A unified view of current normalized models (optional)

1. model the random variation present for fixed $q$
   - assume $c|q$ is a Gaussian, i.e. $c|q \sim \mathcal{N}(\bar{\mu}_q, \bar{\Sigma}_q)$
     - Gaussian is over time ($c$ is a $T$-dimensional vector)
     - $\bar{\mu}_q$ is mean trajectory
2. model the way the overall distribution over $c$ depends on the individual states $q_t$ at each time $t$
   - define $\bar{P}_q = \bar{\Sigma}_q^{-1}$ (precision matrix) and $\bar{b}_q = \bar{P}_q\bar{\mu}_q$ ($b$-value)
   - assume the effect of the state $q_t$ at time $t$ is local in terms of the precision matrix and $b$-value
     - $q_t$ affects $(\bar{b}_q)_{t-K_L:t+K_R}$
     - $q_t$ affects $(\bar{P}_q)(t-K_L:t+K_R)(t-K_L:t+K_R)$
     - N.B. effect of $q_t$ on $\bar{\Sigma}_q$ and $\bar{\mu}_q$ typically lasts much longer than $K$ frames
   - $\bar{P}_q$ and $\bar{b}_q$ are the natural parameters of the Gaussian
In other words, $\bar{P}_q$ and $\bar{b}_q$ are built up from overlapping local contributions.

\[
\bar{P}_q = \begin{pmatrix}
\end{pmatrix}
\quad
\bar{b}_q = \begin{pmatrix}
\end{pmatrix}
\]

- the difference between the autoregressive HMM and trajectory HMM is in the form of the local contributions\(^5\)

---


