Synthesizing variation in prosody for Text-to-Speech

Iberspeech 2018
Rob Clark
What is this talk about

Thoughts and reflections on current TTS approaches

A tour of recent work in TTS and prosody at Google
Outline

Background of Prosody in TTS

Recent advances in TTS acoustic modelling

Recent advances in Prosody:

CHiVE

Style Tokens

Conclusions
What is TTS?

Typical Traditional TTS System

Text → Language Processing (Front end) → LF → Speech Production (Back end) → Sound

Tokenization → Text normalization → Phonemic analysis → Phrasing → Prosody → Acoustic modeling → Speech Generation → Post processing

NLP like stuff

Speech processing like stuff

Current TTS direction?

Text → Machine Learning → Sound
Wavenet

AR Wavenet

Output

Hidden Layer

Hidden Layer

Hidden Layer

Input

(van den Oord et al 2016)
Wavenet

(linguistic conditioning)

(conditioning stack)

(linguistic features)

Input

Output

Hidden Layer

Hidden Layer

Hidden Layer

(van den Oord et al 2016)
Parallel Wavenet

WaveNet Teacher

Linguistic features

Teacher Output
\[ P(x_i | x_{<i}) \]

Generated Samples
\[ x_i = g(z_i | z_{<i}) \]

WaveNet Student

Linguistic features

Student Output
\[ P(x_i | z_{<i}) \]

Input noise
\[ z_i \]

(van den Oord et al 2017)
Wavenet summary

- Traditional wavenet takes *Linguistic Features (LFs)* as input and produces a *waveform* as output.
- AR model is very slow.
- Parallel model is hard to train:
  - Quality can be less than AR model.
  - Difficult to train multi-speaker models.
Tacotron

Text Norm. → NLP, Lexicon → G2P → Duration Modeling → Pitch Prediction → Acoustic Prediction → Vocoding / Synthesis

Encoder → Attention → Decoder + Post-Net → Griffin-Lim

(Wang et al 2017)
Tacotron architecture

Encoder representation.

- CBHG
- Pre-net

Character embeddings

- A
- r
- e
- n
- t

Attention

Attention is applied to all decoder steps

<GO> frame

Decoder RNN

Attention RNN

Pre-net

Post-processing Net

Griffin-Lim reconstruction

Linear-scale spectrogram

Mel-spectrogram with r=3

Tacotron (Wang et al., 2017)
Tacotron 2

WaveNet MoL

mel spectrogram

Waveform samples

Bi-directional LSTM

3 Conv Layers

Character Embedding

Location Sensitive Attention

5 Conv Layer Post-Net

Linear Projection

2 LSTM Layers

2 Layer Pre-Net

My Name is...
Tacotron Summary

- Tacotron variants take one of either *characters* as input or *linguistic features*
- Tacotron (core) output is a spectrogram
  - Can either be converted to speech algorithmically
  - or with a Wavenet model built to take a spectrogram as input.
Plug and play toolbox

The linguistic input to the model

- End-to-end vs traditional front end

The structure of the model

- Wavenet full mode vs wavenet as a vocoder
Prosody

Traditionally Prosody in TTS has been:

- Phone Duration prediction, then F0 prediction
- Formant synthesis → Diphone synthesis → Unit selection synthesis → SPSS
Many problems

- Modelling duration independently
- Very English / Western European centric
  - F0 not always a primary correlate of prosody
  - No real account for lexical pitch accent
  - Average energy?
- Assumption is that we are dealing with isolated citation form sentences.
We don’t really understand prosody very well.

Speech is older than written language.

However, linguistics is traditionally about written language

Saussure: La Langue rather than La Parole

Chomsky: Competence rather than Performance

(Fred Cummins, Speech Prosody 2014)
Prosody example - Semantic variation

“John won at Mary’s”
John won at Mary’s
John won at Mary’s
John won at Mary's
John won at Mary’s
John won at Mary’s
John won at Mary's
Prosody example - Different style

(This is natural speech, not TTS!)
Recent prosody work at Google

Two approaches to address some of these problems.
Clockwork Hierarchical Variational Autoencoder (CHiVE)

Work by V. Wan, C. Chan & R. Clark
Contributions by J. Vit & I. Hodari

A structured approach to try to address:

1. The same text can be said in many different ways with many different styles and emotions.
2. Current model (LSTM) are local frame based models.
   - 100 frames of LSTM memory is 0.5s which is only a couple of words max.
   - Need globally consistent contours
     - Y/N-Questions Low - Low - High vs WH-Question High - High - Low
3. Joint modelling of f0, c0 and duration.
   - Prosody is more than just f0.
CHiVE overview

We need a model that can:

- Process multiple versions of an utterance without learning the average
- Synthesise different version of an utterance

- Use Latent variable to model this variation at the utterance level - Conditional Variational Autoencoder.
Decoder in more detail

Different layers in the model are clocked at different rates.

Rate relates to the linguistic structure
Unrolled phone decoder

Frame $F_0$ predictions (modelled per syllable)

Frame $c_0$ predictions (modelled per phone)

Phone duration predictions

Phone linguistics
Unrolled syllable decoder

- Frame $F_0$ predictions (modelled per syllable)
- Frame $c_0$ predictions (modelled per phone)
- Phone duration predictions

Phone linguistics

Syllable linguistics
Unrolled full decoder

Frame F₀ predictions
(modelled per syllable)

Frame c₀ predictions
(modelled per phone)

Phone duration predictions

Phone linguistics

Syllable linguistics

Word linguistics

Sentence linguistics

Prosody embedding
Inference Options

Take the mean of the sentence embedding distribution

- This is just the zero vector
Inference Options

Take the mean of the sentence embedding distribution
- This is just the zero vector

Sample from the sentence embedding distribution
- A range of natural prosody, but no obvious mapping to a given meaning or intention
Inference Options

Take the mean of the sentence embedding distribution

- This is just the zero vector

Sample from the sentence embedding distribution

- A range of natural prosody, but no obvious mapping to a given meaning or intention

Acoustic Prosodic Features

- Not useful on its own, but
- Can change the speaker id in a multi-speaker model
- Can encode with one sentence, and decode with another
Examples: sampling the embedding space
Examples: encoder-decoder prosody transfer

Natural speech -> TTS

speaker transfer
Examples: encoder-decoder prosody transfer

Natural speech

TTS

What's large, grey, and doesn't matter?

An irrelephant.

What's taken before you get it?

Your picture!
Examples: encoder-decoder prosody transfer

Natural speech

How robust is this transfer?

What's large, grey, and doesn't matter?

An irrelephant.

syl syl-syl

What's taken before you get it?

Your picture!
CHiVE summary

Pros

● Easy to train
● Models latent variation in intonation well

Cons

● Hard to choose right prosody when only given the text.
Style Tokens

There is more to prosody than intonation!

- Would like to control speaking style
- Specifically interested in long-form tasks like book reading

Can we learn unsupervised labels that account for style?
Tacotron with a reference encoder architecture

(Skerry-Ryan et al 2018)
Style token architecture

(Wang et al 2018)
Style selection

- Spectrogram Encoder
- Spectrogram Slices
- Prosody Embedding
- Fast Attention
- Trained Style Tokens
- Transcript Embeddings
- Character/Phone Embeddings
- to decoder
Style Token examples

There are several listings for gas station.

The forecast for San Mateo tomorrow is 61 degrees and mostly sunny.

Token A

Token B

Token C

Token D

Token E
Style transfer
Style transfer examples

“Pull his canoe home with your line, Fisherman.”

“You are not so important after all, Pau Amma,' he said.”
Non-Parallel style token transfer

Reference
natural speech

Tacotron
Synthesis

Without style tokens

With style tokens

“Something, however, happened this time that had not happened before; his stare into my face, through the glass and across the room, was as deep and hard as then, but it quitted me for a moment during which I could still watch it, see it fix successively several other things.”

“He was pale as smoke, and Harry could see right through him to the dark sky and torrential rain outside. "You look troubled, young Potter," said Nick, folding a transparent letter as he spoke and tucking it inside his doublet. "So do you," said Harry.”
Style token summary

Pros

● Style tokens are simple and powerful
● General technique for uncovering latent variation in speech data
  ○ Speaker ID, noise etc.

Cons

● Generally interpretable, but not always easy to isolate specific effects into a single style token
New challenges
Evaluating Prosody

Moving from naturalness to appropriateness

Need the rater to understand the situation the utterance is being spoken in.

- Describe the situation
  - Provide more detailed instructions
- Simulate the situation
  - Provide the text and audio for the previous discourse
Need for better metrics

Need research to ensure that we can perform useful evaluations.
Other challenges

Separating out different aspects of prosody

- E.g. style has a component of speaking rate

Lack of NLU!

Multi-speaker models

Multi-language models

Ethics of TTS
Conclusions

We’ve come a long way recently!

For TTS to continue to get better we need to be less isolated
Thank you!

Contributions from:

Yannis Agiomyrgiannakis, Igor Babuschkin, Eric Battenberg, Norman Casagrande, Chun-an Chan, Zhifeng Chen, Rob Clark, Luis C. Cobo, Sander Dieleman, George van den Driessche, Erich Elsen, Alex Graves, Dominik Grewe, Zack Hodari, Ye Jia, Nal Kalchbrenner, Koray Kavukcuoglu, Helen King, Yazhe Li, Edward Lockhart, Seb Noury, Aaron van den Oord, Ruoming Pang, Rif A. Saurous, Jonathan Shen, Joel Shor, Hanna Silen, Karen Simonyan, Daisy Stanton, Florian Stimberg, Fei Ren, RJ Skerry-Ryan, Andrew Senior, Oriol Vinyals, Jakub Vit, Tom Walters, Vincent Wan, Yuxuan Wang, Ron Weiss, Yonghui Wu, Ying Xiao, Zongheng Yang, Heiga Zen, Yu Zhang