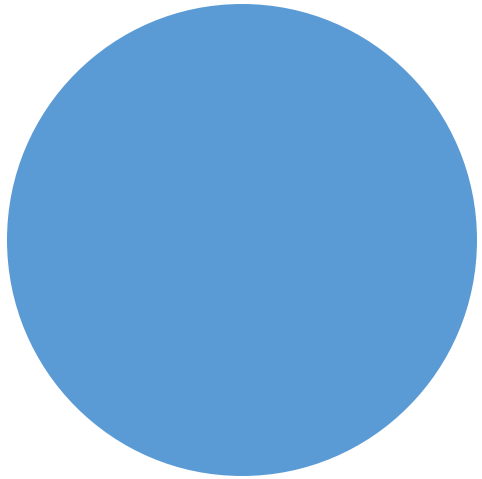
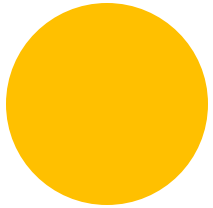



And then they began to speak

Towards end-to-end speech synthesis, and back again?

Markus Toman

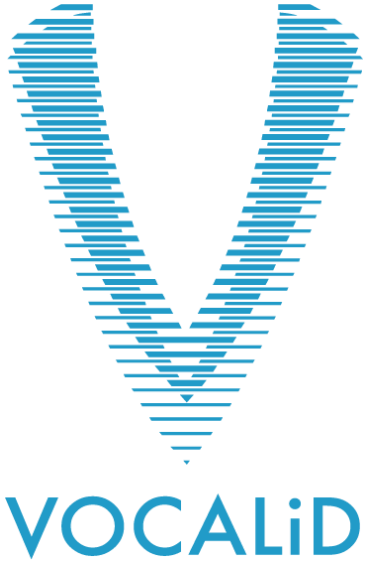
Motivation

- (Intro: https://www.youtube.com/watch?v=VCy4dBFIqH8)
- Synthetic speech quite decent nowadays
- Possibilities and threats
- Are we through the uncanny valley?
- Let's look at how the sausage is made

About me

- Computer Science background
- Doctorate speech synthesis, 8 years in the field (speech tech) now
- Working for VocaliD (https://vocalid.ai/)
- ~~Sword~~Brain for hire, i.e. running my own business (https://www.neuratec.com)
- Lecturer FH Wr. Neustadt
- Further endeavours (e.g. mygewo.at)

About me



- VocaliD, Boston, USA - https://vocalid.ai/
- Original mission to give voices to the voiceless
  - Oscar https://www.youtube.com/watch?v=Z0IBhUW_AJM
  - John https://www.youtube.com/watch?v=ji9cKNPgl-A
- Human Voice Bank with tens of thousands voice donors
- Now broader portfolio, including voice security and brand voices (later in this talk)

Traditional TTS Trinity

- **"Speech synthesis** is the artificial production of human speech."
- Also referred to as Text-To-Speech (TTS), although a bit more narrow in scope

- **Speech synthesis can be**
  - **Helpful**
  - **Fun**
  - **Dangerous**

  Samples for those aspects are sprinkled throughout the talk

Helpful

- Speak for those who can not speak
  - Oscar https://www.youtube.com/watch?v=Z0IBhUW_AJM
  - John https://www.youtube.com/watch?v=ji9cKNPgl-A
- Read for those who can not read
  - Fast speech for the blind https://wiki.inf.ed.ac.uk/CSTR/SalbProject
- Speak to those who need their eyes elsewhere
  - Car navigation, medical procedures, assistance robots
- Speak to those to those who don't want to read
  - Digital assistants, home automation...

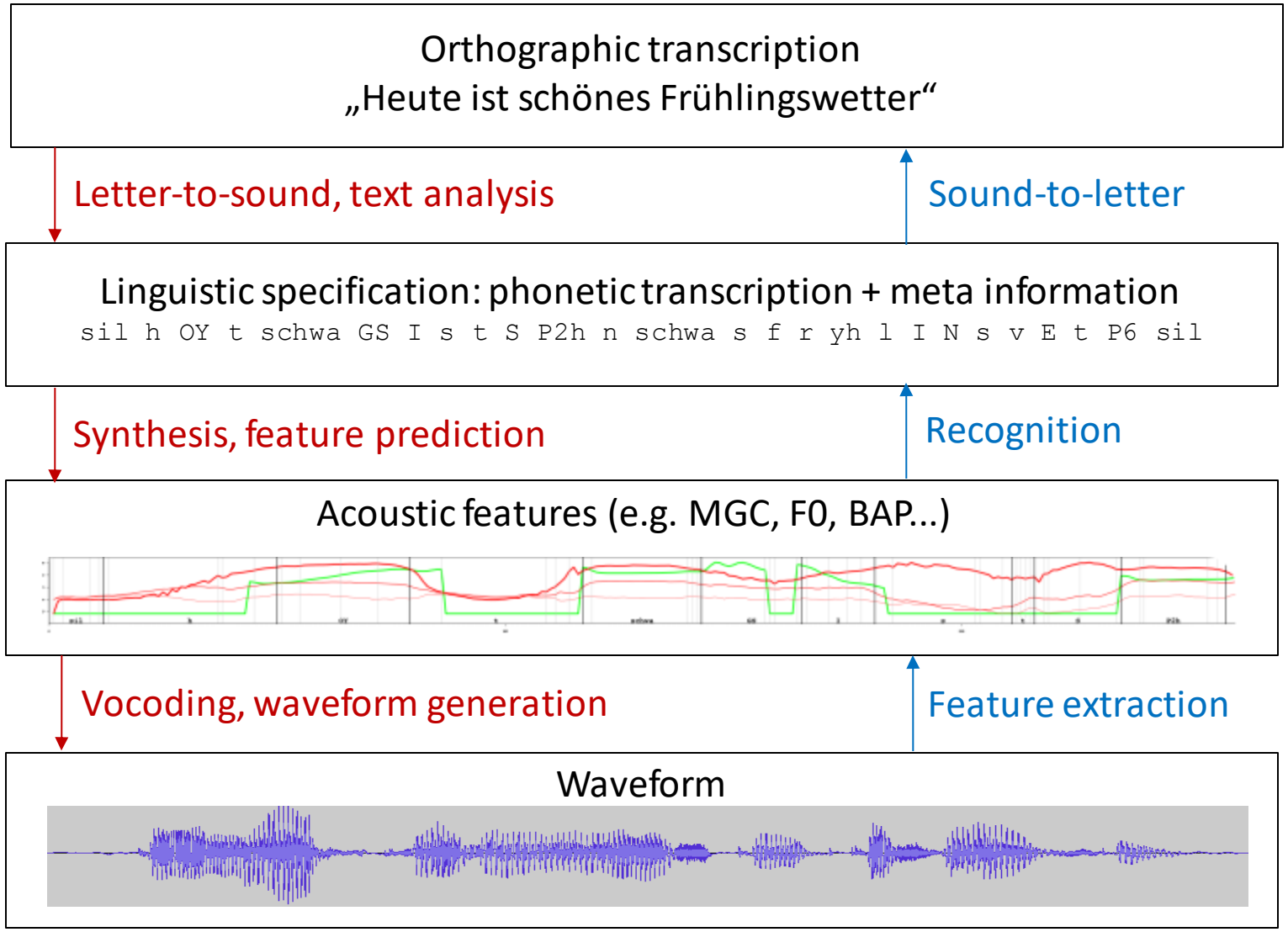
Traditional TTS Trinity

- **"Speech synthesis** is the artificial production of human speech."
- Also referred to as Text-To-Speech (TTS)

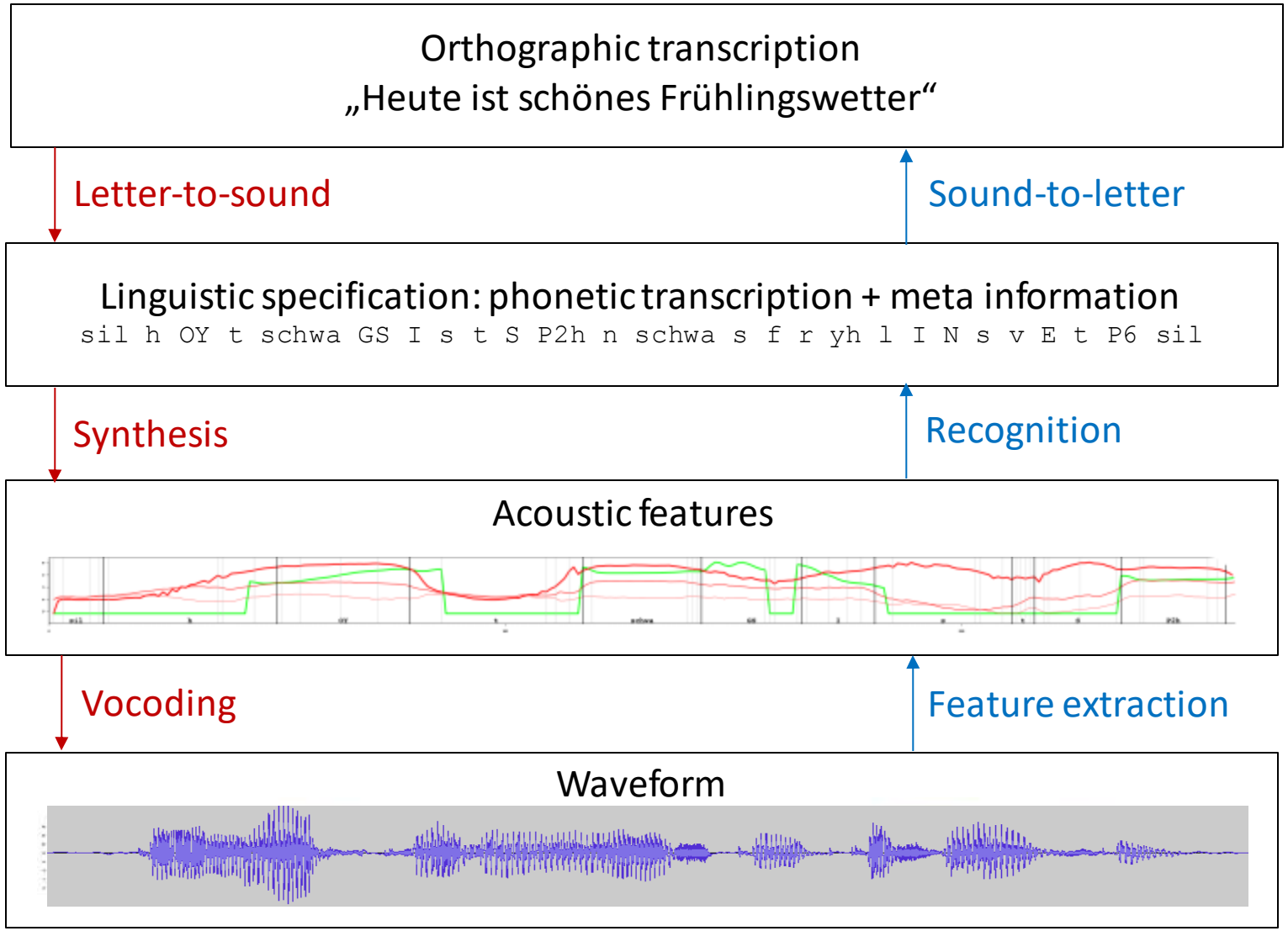
- **Common pattern to use 3 components:**
  - **Text analysis**
  - **Acoustic feature prediction**
  - **Waveform generation**

- We're discussing statistical parametric synthesis here, not touching e.g. unit selection, formant synthesis etc.

Traditional TTS Trinity



Traditional TTS Trinity Tools



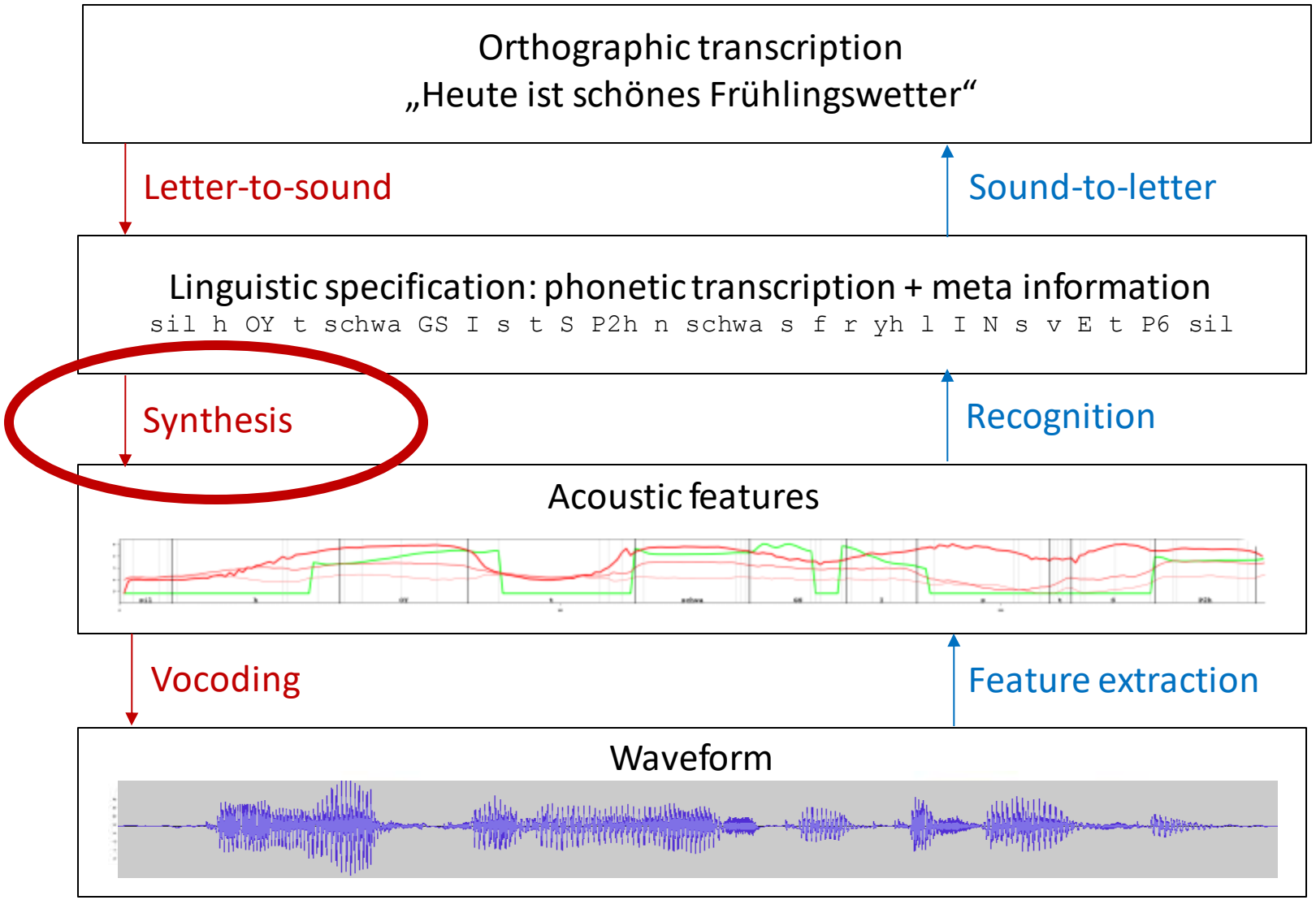
Festival
> 1000 source files
POS-Taggers, LTS-Trees,
pronunciation dictionaries,
ToBi endtone prediction,
phrasing models,
syllable stress prediction, ...

HTS/HTK
~ 120 source files
> 120k lines of code
~ 34 command line tools
with ~20 parameters
+ extra script files

STRAIGHT, WORLD, hts_engine
Small :)
Complicated :(

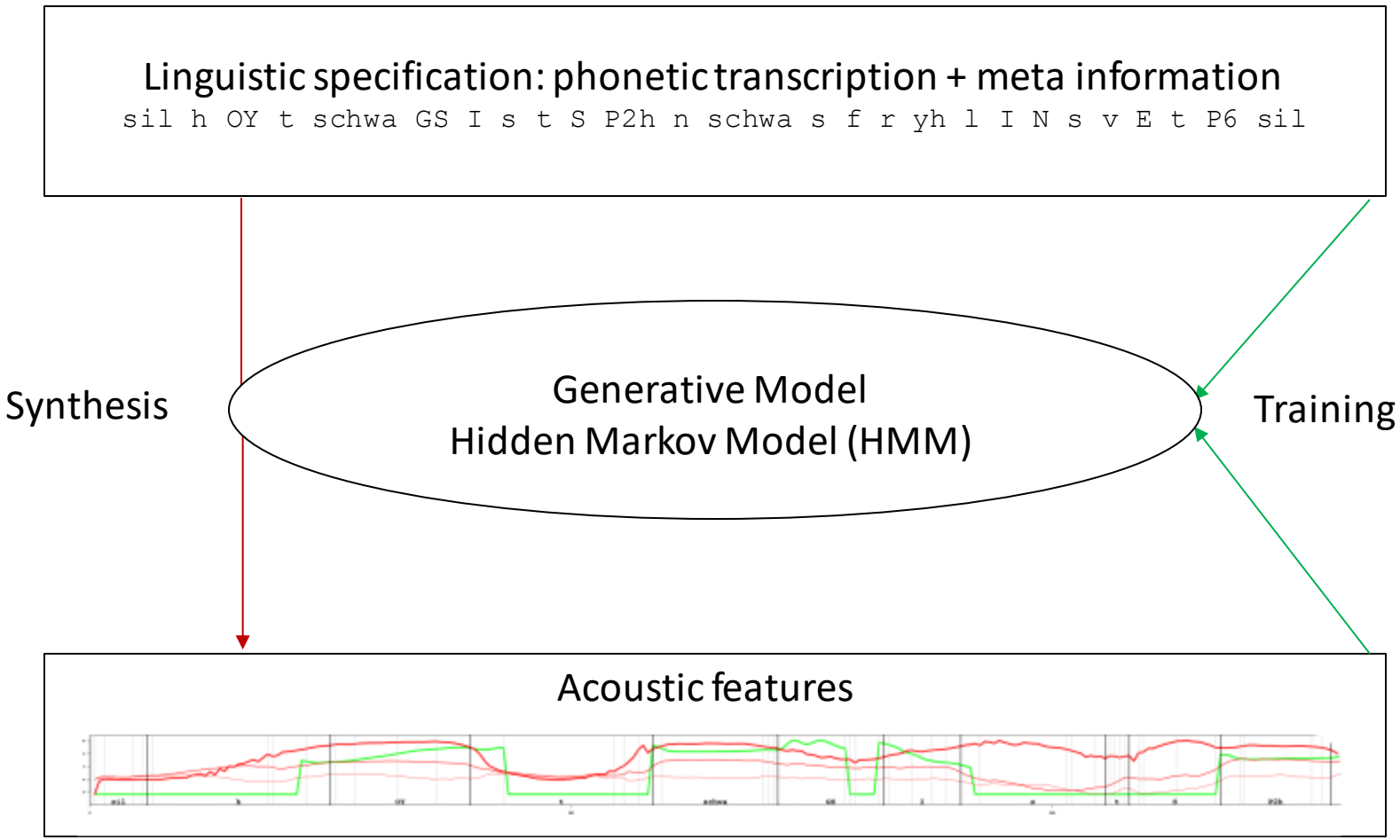
Tools grew over time, most of code specific to speech processing

Traditional TTS Trinity



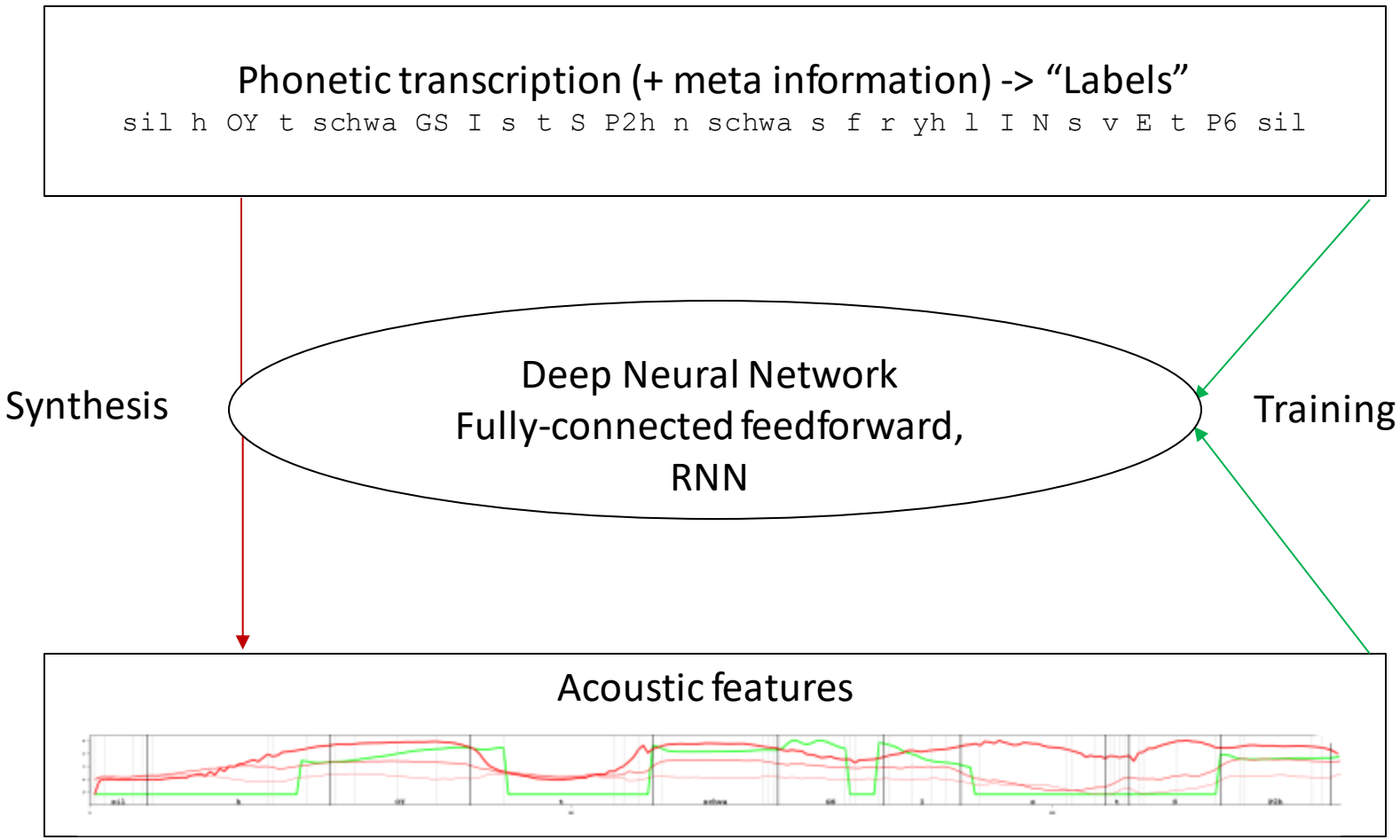
Classic machine learning techniques have been used for acoustic feature prediction for some time.

Traditional TTS Trinity



For example, HMMs with decision-tree-based clustering.
Adding dynamic features and Maximum Likelihood Parameter Generation (MLPG), Postfilter, Global Variance and lots of other tricks ...

Traditional TTS Trinity



Naturally the first place where DNNs were successfully introduced: to predict acoustic features from linguistic specification labels.

Still separate duration and acoustic models.

See Heiga Zen et al (2013): "STATISTICAL PARAMETRIC SPEECH SYNTHESIS USING DEEP NEURAL NETWORKS"

Towards end-to-end synthesis

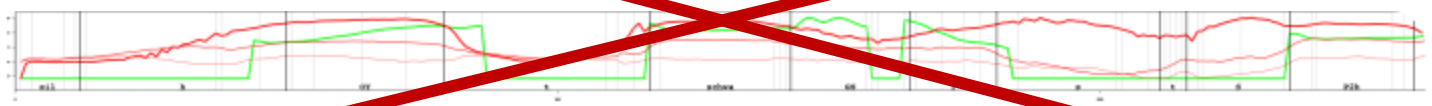
Orthographic transcription
„Heute ist schönes Frühlingswetter“

Letter-to-sound

Linguistic specification: phonetic transcription + meta information
sil h OY t schwa GS I s t S P2h n schwa s f r yh l I N s v E t P6 sil

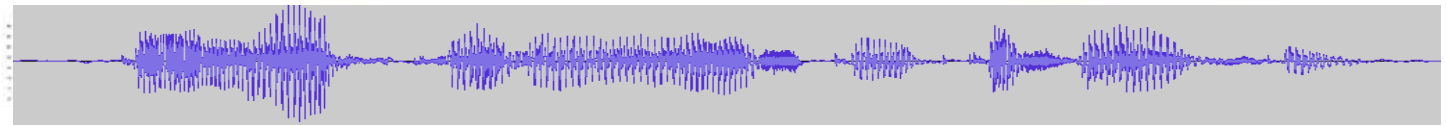
~~Synthesis~~

~~Acoustic features~~



~~Vocoding~~

Waveform



First Wavenet version
by Deepmind in 2016.
Predicting waveforms from
linguistic specification.

Towards end-to-end synthesis

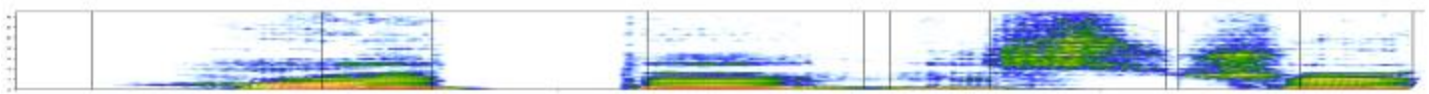
Orthographic transcription
„Heute ist schönes Frühlingswetter“

Letter-to-sound

Linguistic specification: phonetic transcription + meta information
sil h OY t schwa GS I s t S P2h n schwa s f r yh l I N s v E t P6 sil

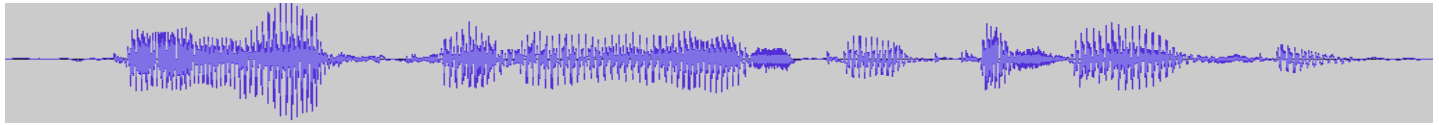
Synthesis

Acoustic features



Vocoding

Waveform



Tacotron (2017) predicts spectral features from plain text.
Does not require separate duration model or alignment

Towards end-to-end synthesis

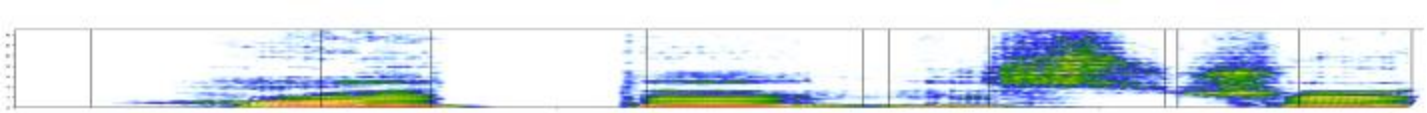
Orthographic transcription
„Heute ist schönes Frühlingswetter“

Letter-to-sound

Linguistic specification: phonetic transcription + meta information
sil h OY t schwa GS I s t S P2 h schwa s f r yh l I N s v E t P6 sil

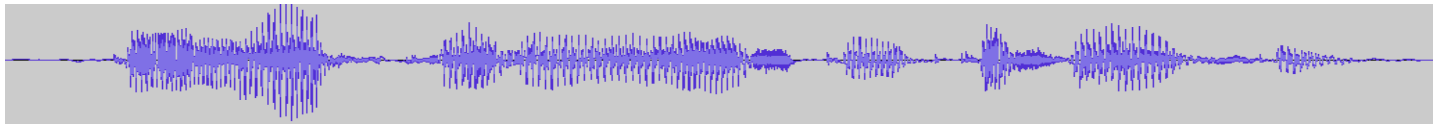
Synthesis

Acoustic features



Neural vocoding

Waveform



Modified Tacotron 1

WaveNet conditioned on mel spec

Tacotron 2 almost end-to-end

Traditional TTS Trinity Tools

The Emperor's New Clothes...

Festival
>1000 source files
> 43k lines of C++ code, even more scheme code
POS-Taggers, LTS-Trees, pronunciation dictionaries,
ToBi endtone prediction,phrasing models,
syllable stress prediction, ...

HTS/HTK
~ 120 source files
> 140k lines of C code
~ 34 command line tools
with ~20 parameters
+ extra script files

SPTK, EST for general signal and speech processing

Tacotron (nvidia implementation)
~2500 lines of Python - TTS-specific code
librosa for basic sound and signal processing (STFT, wav loading/writing)
The rest is in general purpose libraries, e.g. PyTorch, Tensorflow

A single model, trained with standard methods

Became much more accessible for experts from other fields – much more open source work

Fun 1

- Speak to entertain
- 'Twas the night before Christmas with VocaliD https://www.youtube.com/watch?v=Ov2A9wjoI_8
- Goodnight Moon https://www.youtube.com/watch?v=Jw02N9mYiCU
- Dialect interpolation http://mtoman.neuratec.com/thesis/interpolation/
- Adorno https://kutinkindlinger.com/le-parleur-radiopiece-52/

WaveNet

- Introduced by DeepMind in 2016
- 1D version of PixelCNN
- Convolutional neural network
  - Dilated convolutions to increase receptive field
  - Autoregressive
- Local conditioning to control output (e.g. linguistic specification)

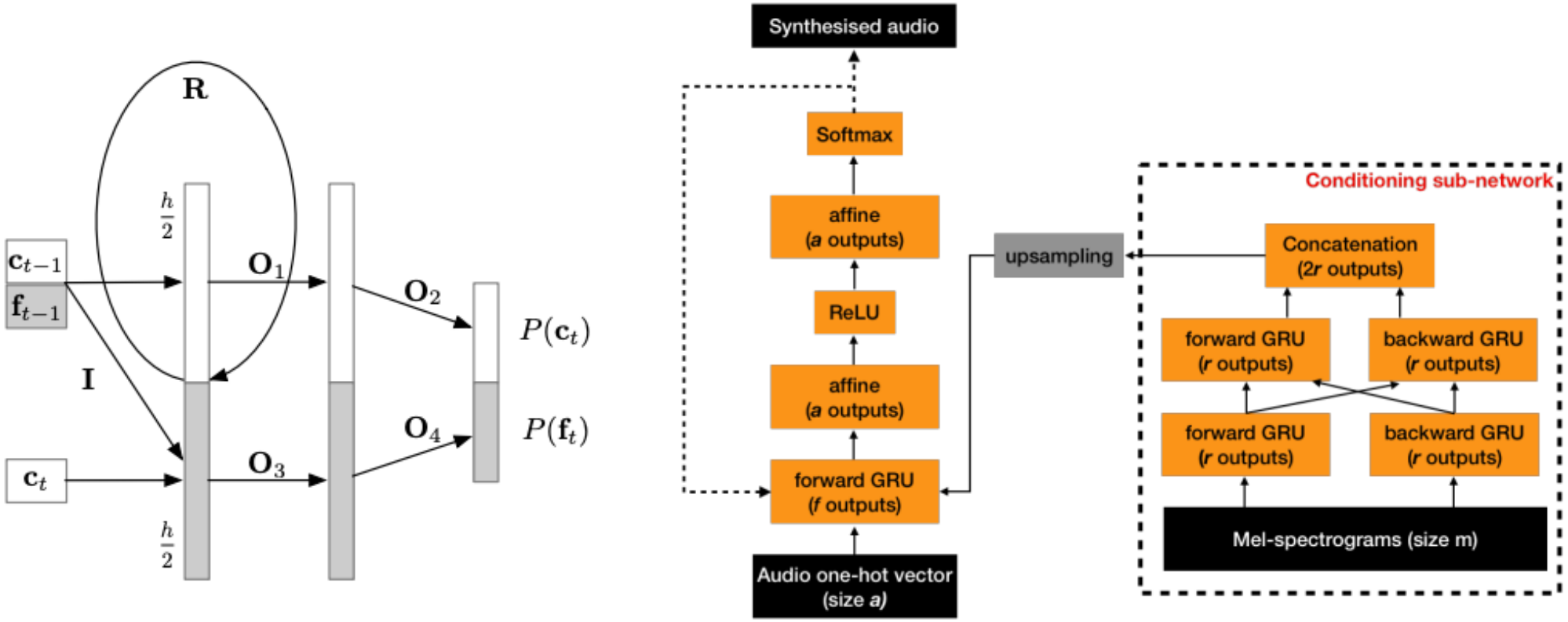
- Aäron van den Oord et al (2016): WAVENET: A GENERATIVE MODEL FOR RAW AUDIO
  - https://deepmind.com/blog/wavenet-generative-model-raw-audio/

WaveNet

- https://deepmind.com/blog/wavenet-generative-model-raw-audio/#gif-7
- Dilation factors grow exponentially with each layer
- Causal convolution avoids requiring "future" samples
- Outputs mu-law companion and quantization to 256 values, softmax distribution
  - Different output types used in adapted methods, e.g. Mixture of Logistics

WaveRNN

- RNN-based waveform modeling



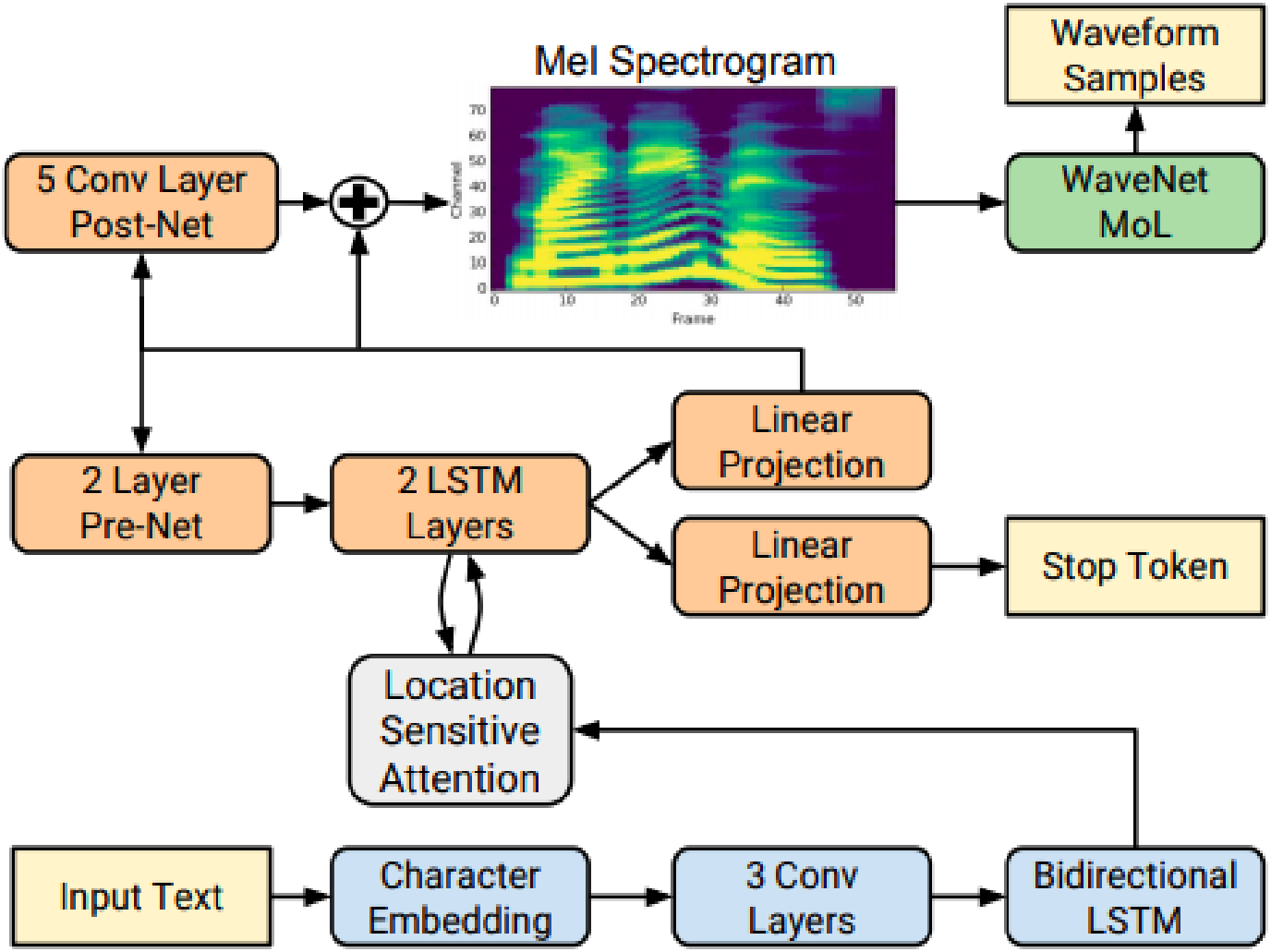
- Left: Nal Kalchbrenner et al (2018): *Efficient Neural Audio Synthesis* (https://arxiv.org/abs/1802.08435)
- Right: Jaime Lorenzo-Trueba et al (2018): *Robust Universal Neural Vocoding* (https://arxiv.org/abs/1811.06292)

Tacotron

- **Tacotron 1** uses **Attention**-based model to predict **spectral features from text** (character embeddings)
- **Tacotron 2** simplifies Tacotron 1, predicts **Mel-spectral** features and feeds them to **WaveNet used as neural vocoder**
  - Feature prediction network produces ground truth aligned features to train WaveNet
  - WaveNet conditioned on mel-spectral features predicts samples of waveform
- Except for a bit of text normalization and symbol selection, this process is mostly independent from the actual language
- No explicit duration model, learns alignment

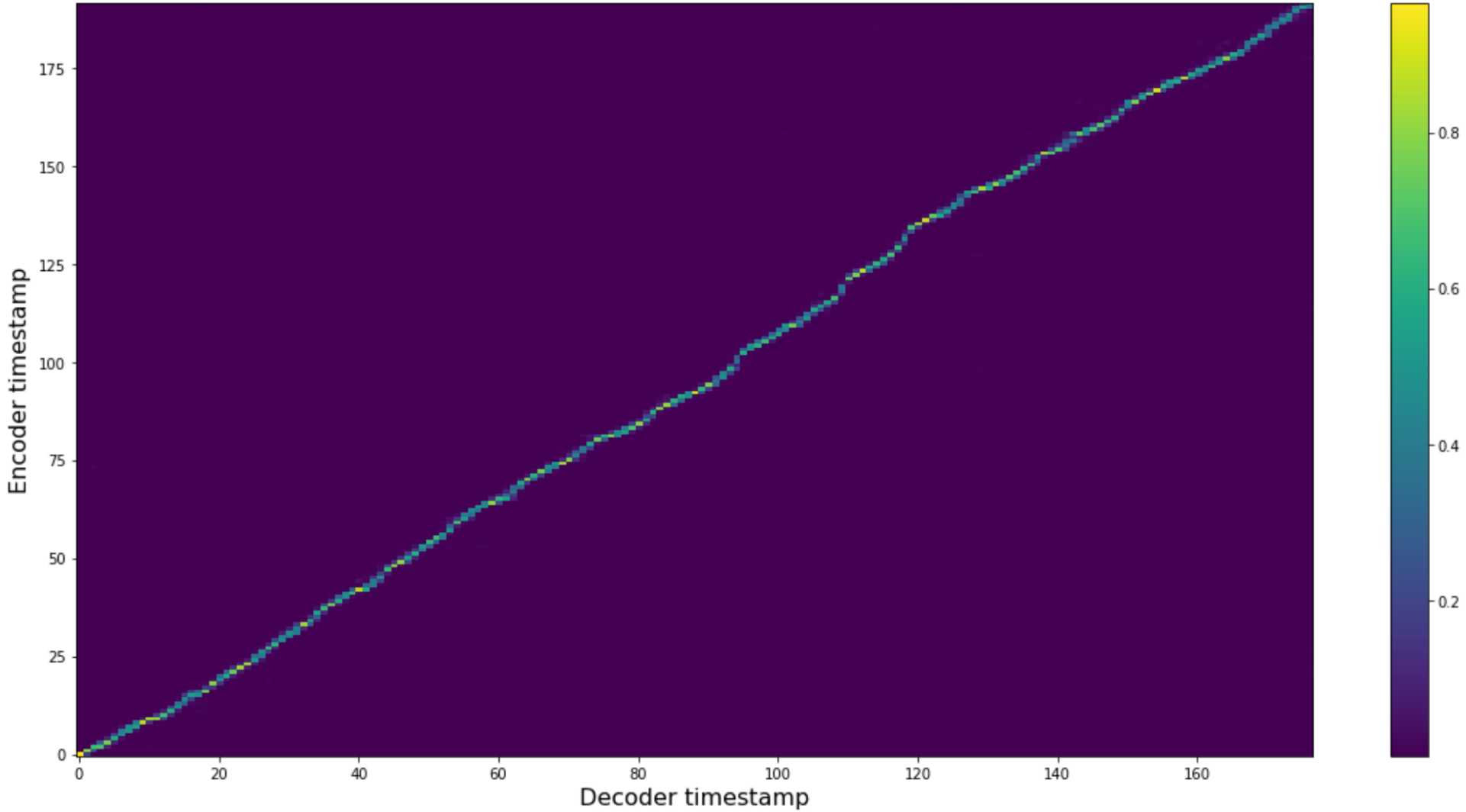
- Yuxuan Wang et al (2017): *Tacotron: Towards End-to-End Speech Synthesis*
- Jonathan Shen et al *(2017): Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*
- All the Google Tacotron papers and samples: https://google.github.io/tacotron/

Tacotron 2



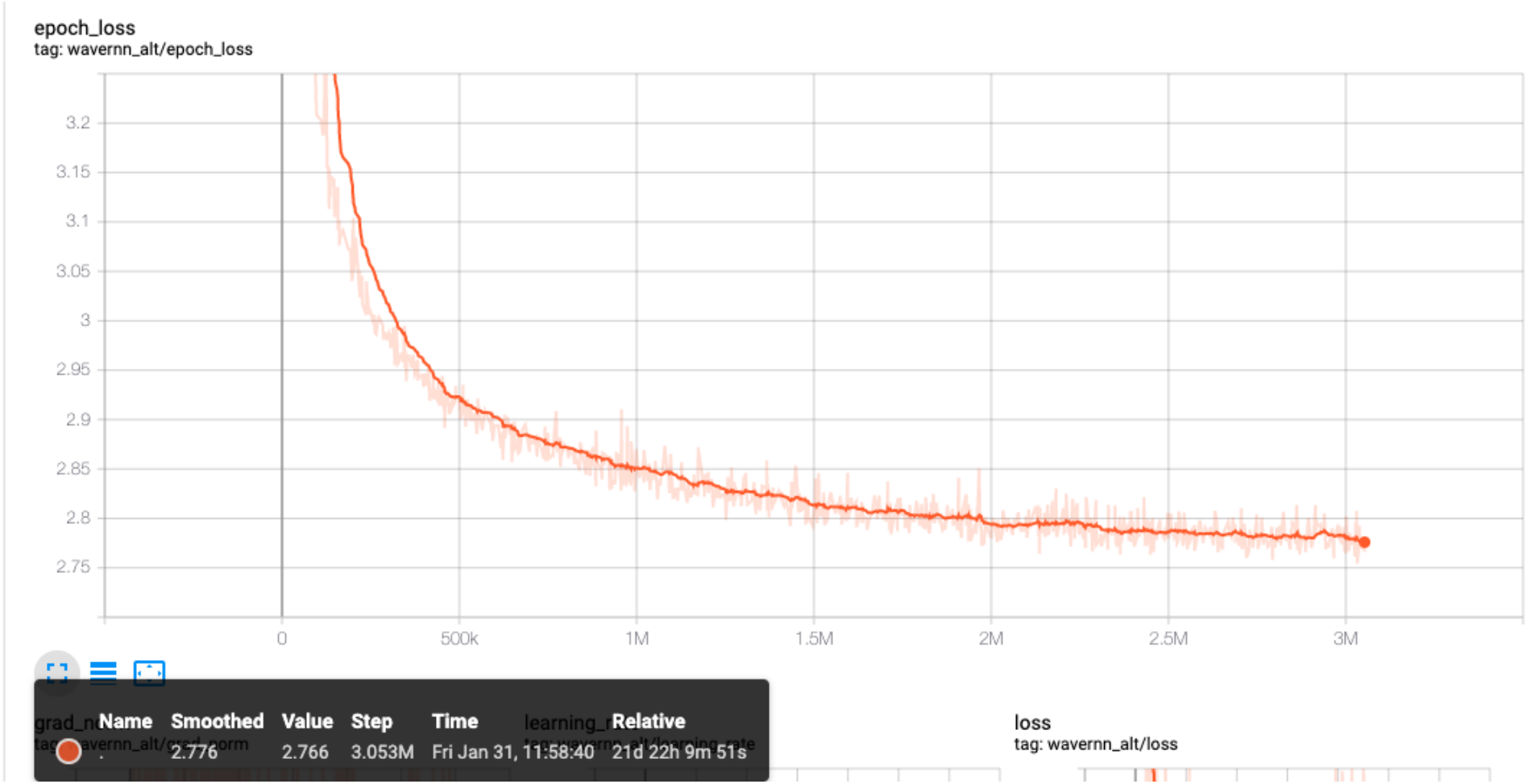
- Jonathan Shen et al *(2017): Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*

Tacotron



- Jonathan Shen et al *(2017): Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*

Deep learning



Patience is a virtue

(or fiddle with the learning rate and hope things don't blow up in your face)

Fun 2

- Speak to entertain
  - You will hear samples of brand voices, dialectal speech, bloopers...

Open Source

- Open Source Community needed more than a year after the WaveNet paper was published to produce roughly comparable results
  - https://github.com/keithito/tacotron
  - https://github.com/Rayhane-mamah/Tacotron-2
  - https://github.com/r9y9/wavenet_vocoder
- 2018, companies joined in
  - https://github.com/NVIDIA/tacotron2
  - https://devblogs.nvidia.com/nv-wavenet-gpu-speech-synthesis/
  - https://github.com/NVIDIA/nv-wavenet
  - https://github.com/mozilla/TTS

Proliferation

- SampleRNN
- Wav2Char
- WaveRNN
- Baidu Deep Voice 1, 2, 3
- Baidu ClariNet (uses a bridge network to interface with WaveNet to produce a single end-to-end network)
- NVidia WaveGlow
- Adobe FFTNet
- Facebook VoiceLoop
- Microsoft Transformer TTS, FastSpeech
- MelGAN, WaveGAN, ParallelWaveGAN, LPCNet, MelNet....

Personalization: Data in the wild

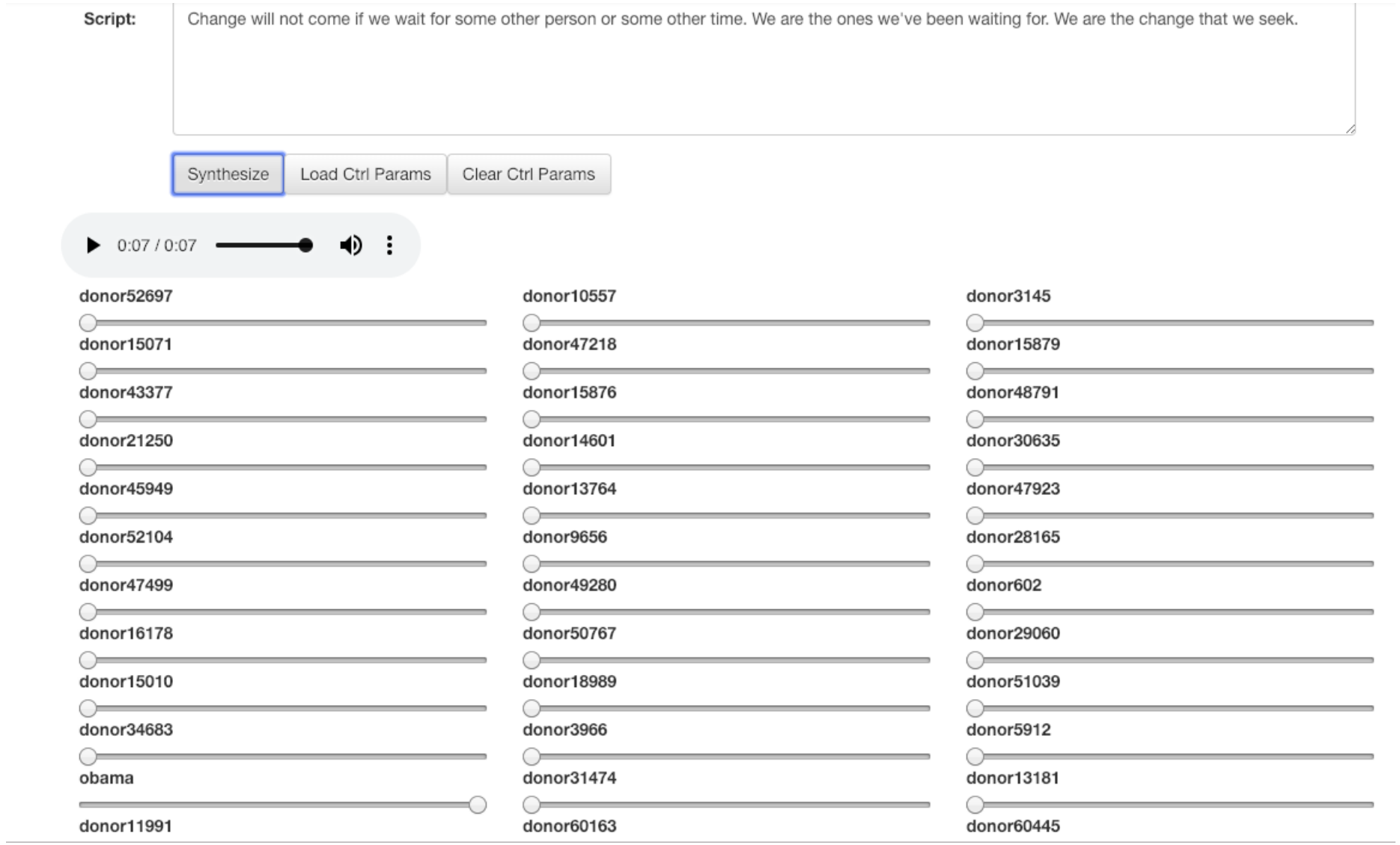
- Showcased voices trained from lots of polished data.
  - Studio recordings
  - Voice talents
  - Manual cleaning

- Most voice cloning solutions produce pretty awful results off-the-shelf given "real" data, e.g.
  - children talking into their headset microphones at home - dog barking, two washing machines running, lots of mispronunications
  - data scraped from interviews, videos, movies – other people talking, applause, lots of background noise, stopping mid-sentence, no good transcriptions available

Taming wild data

- Cleaning pipeline
  - noise removal, dereverberation, click removal, bad sentence tagger...
- Augmentation
  - concatenation, mixing with similar voices, slight transformations...
- Adaptation
  - e.g. finetuning - continue training with different dataset, potentially freezing parts of the network.
- Multispeaker models
  - Feature identity is also input to network - representation sharing between speakers. One-hot embedding, iVector, encoder network...

Multispeaker models



Dangerous

"Criminal deepfake attacks have claimed their first victims, with a British energy **defrauded of nearly a quarter-million dollars through a wire transfer ordered by what seemed from the voice to be a company executive**, The Washington Post reports. The company's managing director was phoned by what he thought was a company executive, according to representatives of French insurer Euler Hermes, and **though he felt the wire request was strange, he complied, thinking he was following instructions from his boss**. When the thieves made a second request, the managing director acted on his suspicions and called the executive. The fraudulent version phoned while that call was still in progress, exposing the fraud. Symantec researchers say they have discovered **at least three such incidents**, although it is unclear if that includes the above case. The **losses in one case exceeded a million dollars**."

https://www.biometricupdate.com/201909/deepfake-voice-technology-claims-first-fraud-victims

Dangerous

- Speak to fake, fraud, impersonate, break voice authentication systems
  - You will hear samples for impersonation

End-to-end and back again?

- Lacking control, dependency on attention model
- Most Tacotron-based methods use phones instead of characters again
- Microsoft FastSpeech uses explicit duration model
- IBM TTS uses various different networks for G2P, prosody modeling, duration modeling
- Tencent DurIAN uses explicit duration model

The Future

- End-to-end or not?
- Currently strong trend for more robustness – seq2seq often fragile
- New model types used in the field
  - GANs, Transformer, BERT, neural arithmetic logics units...
- More fine-grained control
  - Style, language varieties, emotion...
- Dealing with noisy data
- Unified models for synthesis, recognition, voice verification...?
- Ethical and security challenges – impersonation, deep fakes, breaking voice authentication systems and digital assistants

The End

Material

- Oscar's new VocaliD voice
  https://www.youtube.com/watch?v=Z0IBhUW_AJM
- John's new VocaliD voice
  https://www.youtube.com/watch?v=ji9cKNPgl-A
- Delaney cheers on her team
  https://www.youtube.com/watch?v=vxMzlB3gVBM
- 'Twas the night before Christmas with VocaliD
  https://www.youtube.com/watch?v=Ov2A9wjoI_8
- Goodnight Moon
  https://www.youtube.com/watch?v=Jw02N9mYiCU
- Dialect interpolation
  http://mtoman.neuratec.com/thesis/interpolation/
- Accent conversion
  http://mtoman.neuratec.com/thesis/transform-accent/
- Fast speech
  https://wiki.inf.ed.ac.uk/CSTR/SalbProject
- Adorno
  https://kutinkindlinger.com/le-parleur-radiopiece-52/