Introduction to modern and controllable speech synthesis

Excerpt from a presentation at the 2018 EACare retreat

Gustav Eje Henter

ghe@kth.se

Division of Speech, Music and Hearing (TMH)
KTH Royal Institute of Technology
Stockholm, Sweden

2018-04-20

Overview

1. Text to speech
2. Variation in natural speech
3. Synthetic speech
   - 3.1 Recreating variation
   - 3.2 . . . without extra information
   - 3.3 Recent breakthroughs
4. Outlook

Overview

A history of speech synthesis

Type	Formant	Unit selection	Statistical	Deep	Waveform
Era	1960s	1980s	2000s	2013–	2016–
Innovation	Machine speech	Use speech recordings	Statistical modelling	Deep learning	Predict waveform
Quality?	Poor	Great	Fair	Good	Excellent
Control?	Great	Poor	Fair	Good	In progress

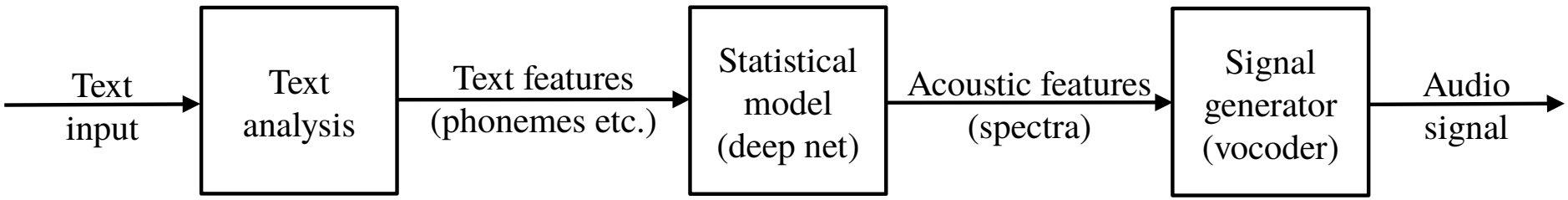
Statistical text-to-speech

Statistical speech synthesis (from 2000s onward):

1. Start from a database of text and corresponding speech audio
2. Create a statistical model that maps from text to audio
   - "Training"/"Learning"/"Optimisation"/"Parameter estimation" using the data
   - Not all processing steps are learned
3. Feed in new text and get new speech out

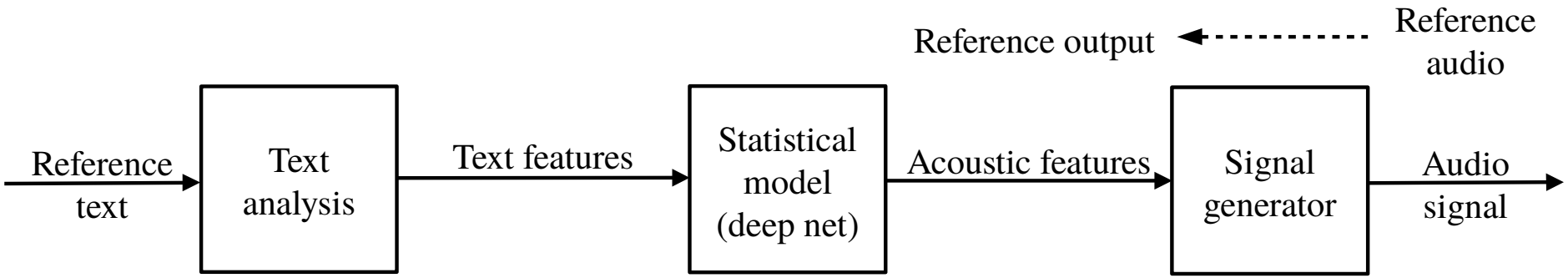
In pictures

Text is transformed into speech:



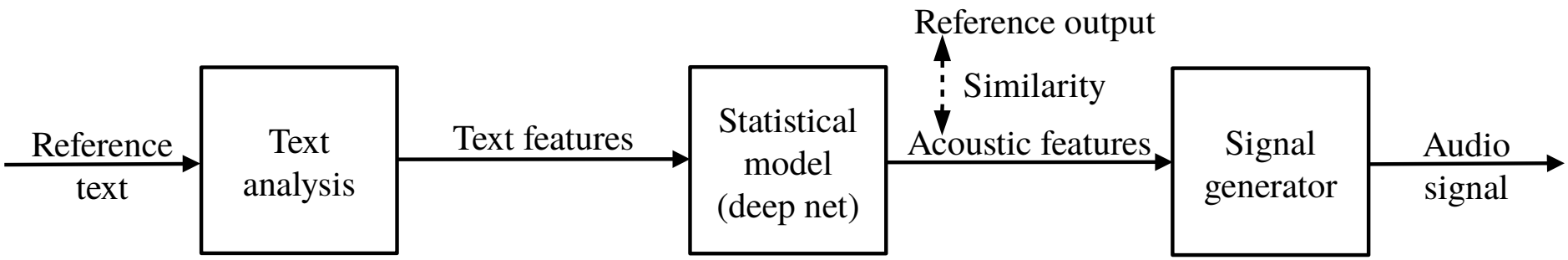
In pictures

We want output for reference input to be similar to examples:



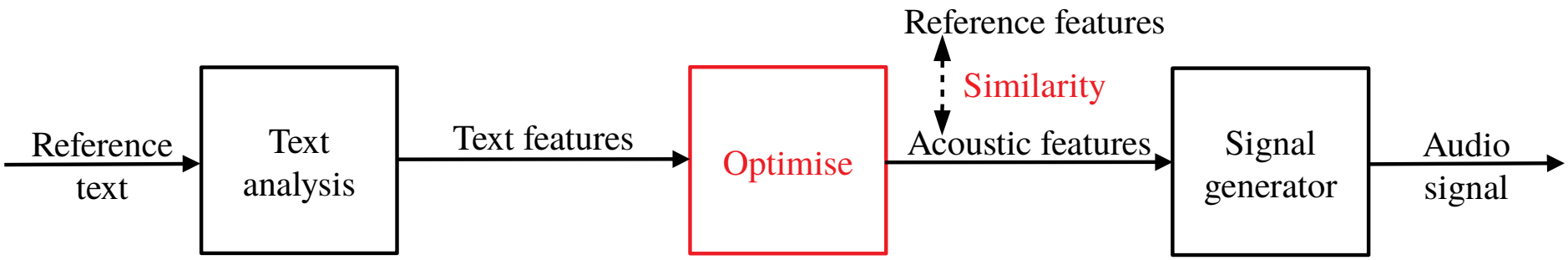
In pictures

Define a mathematical similarity measure between man and machine:



In pictures

Optimise learned parts for greatest similarity:



Text to speech

- In this approach, the way of speaking is determined by data
  - This gives better speech quality compared to manually designing the entire synthesiser
  - The operator is ceding control of speech expression to the algorithm
  - Later, we will look at how to re-enable control
    - Often possible, but easiest if built in in advance

Overview

1. Text to speech
2. Variation in natural speech
3. Synthetic speech
   - 3.1 Recreating variation
   - 3.2 . . . without extra information
   - 3.3 Recent breakthroughs
4. Outlook

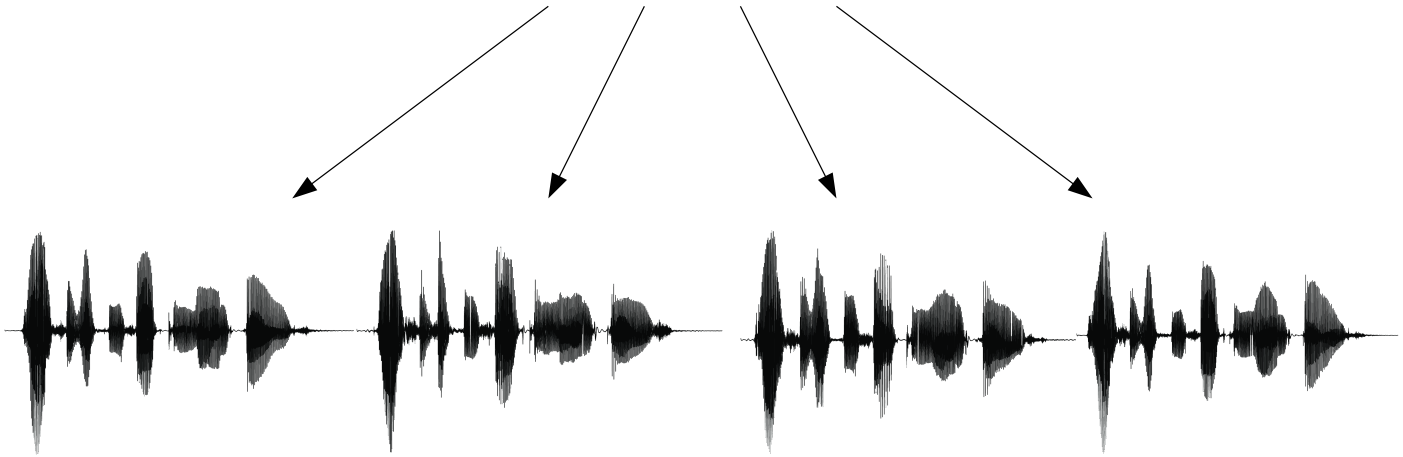
Variation matters

- What we say
  - Lexical variation
- How we say it
  - Variation is informative
  - Even non-informative variation is important

Acoustic variation

The same text can be realised in many different ways:

"Rice is often served in round bowls"



Paralinguistics

- Speech audio contains more information than just text:
  - Speaker identity and characteristics
  - Speaking style
  - Mood and emotion
  - Emphasis/intonation
  - Acoustic environment
  - etc.
- All of the above influence natural speech generation
  - Creating appropriate synthetic speech involves converting all of the above (not just text) into speech audio

Relevant publication

Content adapted from:

Oertel, C., Jonell, P., Kontogiorgos, D., Mendelson, J., Beskow, J., and Gustafson, J. (2017). Crowd-sourced design of artificial attentive listeners.
In *Proc. Interspeech*, pages 854–858

Feedback

- Speech feedback from listener to talker
  - “Okay”, “mh”, “mhm”, “yes”, “right”, etc.
- Gathered examples of feedback through crowd-sourcing
  - Simulated job interview
  - Video of robot or human actor as the applicant
  - Crowd-sourced participant playing the interviewer and asked to give feedback at specific points
    - Asked to be either sceptical, neutral, or supportive

Audio examples

- Examples of collected feedback tokens:

- Same lexical form
- Different information content (connotations)

Speech synthesis aspects

Type	Formant	Unit selection	Statistical	Deep	Waveform
Quality?	Poor	Great	Fair	Good	Excellent
Control?	Great	Poor	Fair	Good	In progress

Relevant publication

Content adapted from:

Henter, G. E., Merritt, T., Shannon, M., Mayo, C., and King, S.

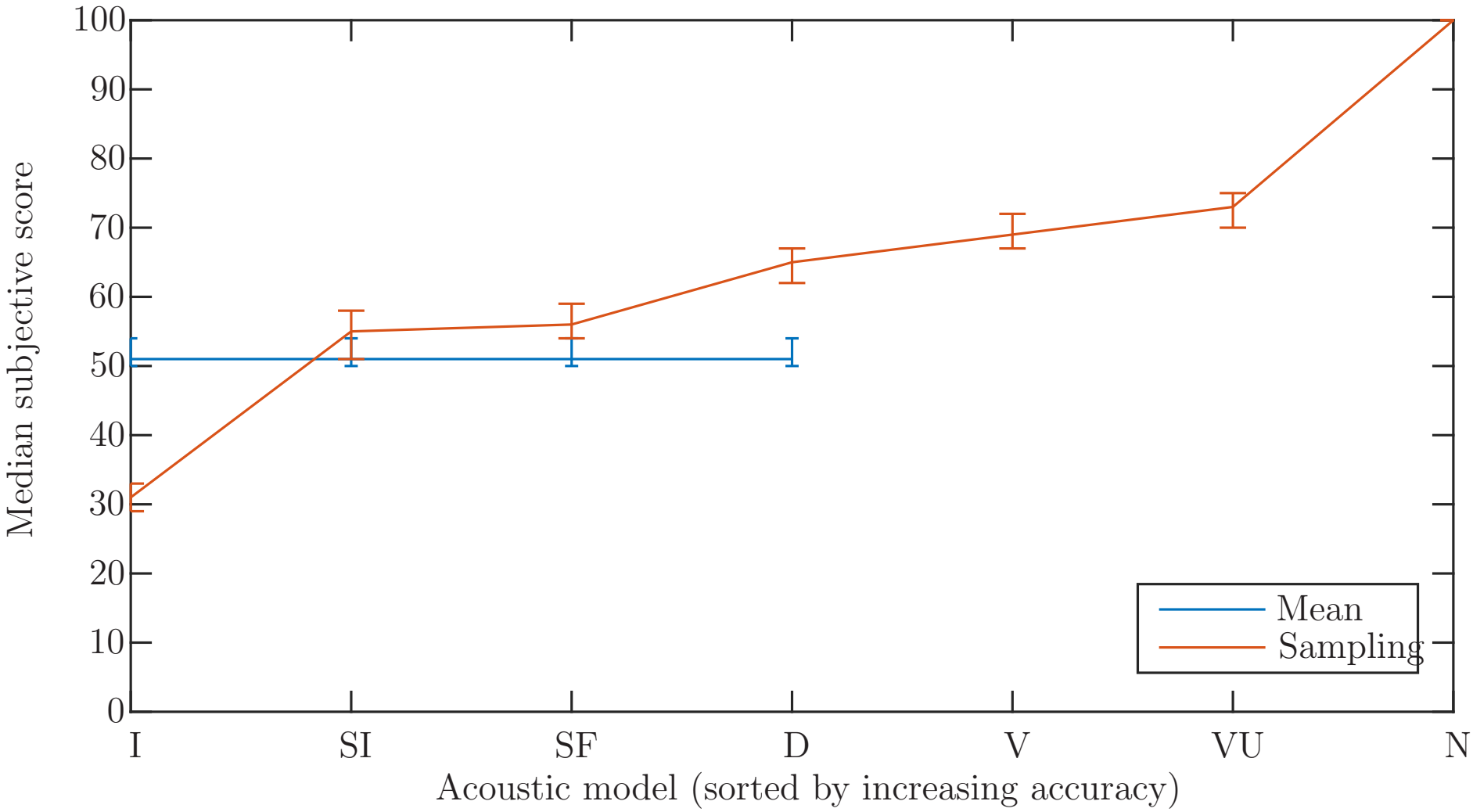
(2014). Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech.
In *Proc. Interspeech*, pages 1504–1508

Approach

- Recorded *repeated speech*: the same speaker reading the same text under similar circumstances
  - No intentional paralinguistic variation
  - Explores the inexorable variation in human speech
  - Independent examples from the "true model" of speech
- Used the natural repetitions to simulate the output of different speech models, beyond the accuracy attained by contemporary text-to-speech systems
  - Either generated as random draws from a model
  - Or using the average ("mean") model output – same every time
- Listeners were asked to rate the naturalness of the resulting speech stimuli in a side-by-side comparison

Experimental results

Different scores with and without randomness:



Take-home message

- Even with a single speaker heard speaking a sentence only once, variation is important
- Acoustic variation is an asset, not "noise"!
- Speech synthesis is more than "text to speech"
  - Ignoring speech variability creates a barrier to generating convincingly natural speech and speaking systems

Overview

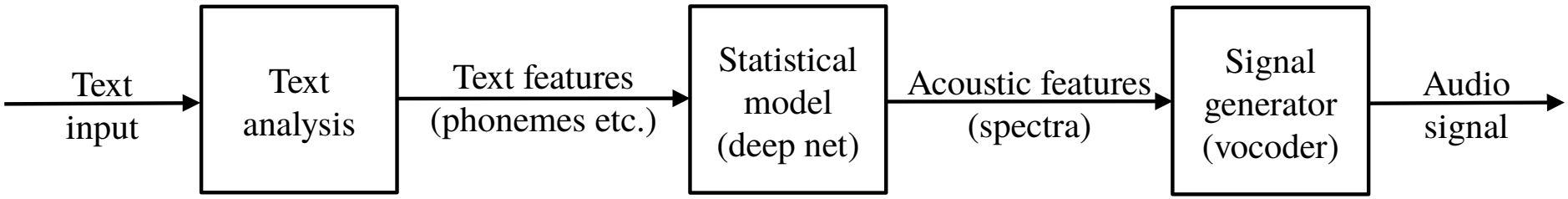
1. Text to speech
2. Variation in natural speech
3. Synthetic speech
   - 3.1 Recreating variation
   - 3.2 . . . without extra information
   - 3.3 Recent breakthroughs
4. Outlook

Speech synthesis aspects

Type	Formant	Unit selection	Statistical	Deep	Waveform
Quality?	Poor	Great	Fair	Good	Excellent
Control?	Great	Poor	Fair	Good	In progress

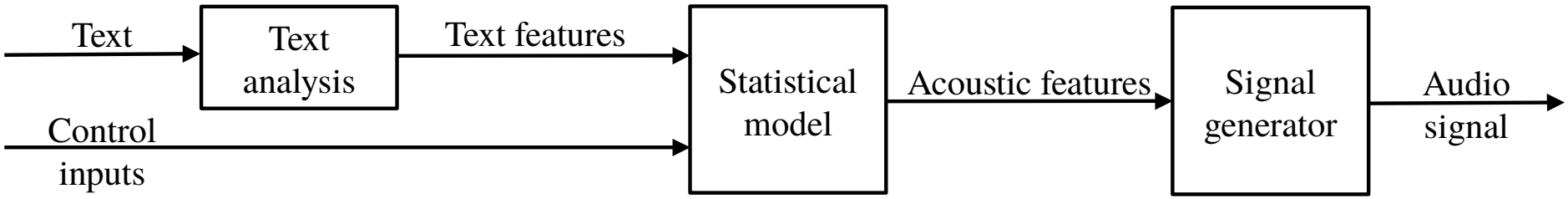
Controllable text-to-speech

To enable variability, add *control inputs* to influence how the text is spoken:



Controllable text-to-speech

To enable variability, add *control inputs* to influence how the text is spoken:



Relevant publication

Content adapted from:

Luong, H.-T., Takaki, S., Henter, G. E., and Yamagishi, J. (2017).

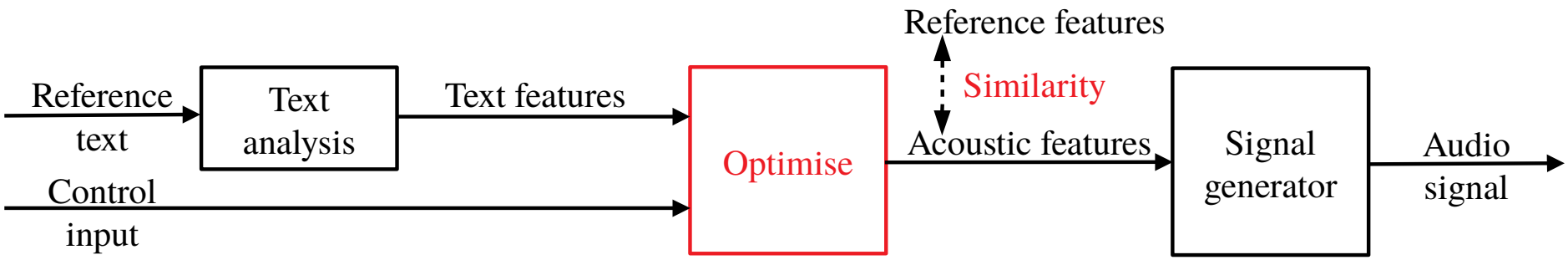
Adapting and controlling DNN-based speech synthesis using input codes.
In *Proc. ICASSP*, pages 4905–4909

Massively multispeaker synthesis

- Database of more than a hundred different speakers
  - Unusual – most speech synthesisers use a single voice
  - Wide age range across both genders
- For every sentence in the database we provided
  - Speaker ID (e.g., randomly assigned number)
  - Age
  - Gender
  - . . . as known control-input values, beside the text
- Leverage the extra input to better replicate training data
  - This works since the inputs are informative about the acoustics

Training a controllable system

Train system on reference data with reference control inputs:



Result

- The trained speech synthesiser was able to:
  - Speak in different voices from the database
  - Imitate ("adapt to") the speech of new speakers not in the database
  - Morph between different voices and attributes
    - By changing the control input values while speaking

Voice morphing

Demonstration of voice morphing while speaking:

	Example 1	Example 2
No manipulation	▶	▶
Gender (F→M)	▶	▶
Age (-50→200)	▶	▶

Relevant publication

Content adapted from:

Henter, G. E., Lorenzo-Trueba, J., Wang, X., Kondo, M., and Yamagishi, J. (2018). Cyborg speech: Deep multilingual speech synthesis for generating segmental foreign accent with natural prosody.
In *Proc. ICASSP*, pages 4799–4803

Goal

- Make a synthesiser speak that can speak in a foreign accent
  - Using only native speech data
  - With control over how the accent is realised
  - Having native-like rather than synthetic or non-native prosody (intonation etc.)
- Useful for research on foreign-accent perception
  - "What is the impact of isolated mispronunciations?"
  - Also another illustration of fine-grained synthesis control

Setup

- Idea: Synthesise speech with carefully controlled mispronunciations
  - Such speech is hard to create/elicit without synthesis
  - Build a multilingual synthesiser (uncommon)
  - Borrow pitch/emphasis from natural speech
    - "Text-and-speech-to-speech"
    - Native way of speaking
- Data: 2000 Japanese and 2000 US English sentences from a speaker native in both languages
- Interpolate from one language to another while speaking to simulate mispronunciations
  - The voice (speaker) remains the same

Example audio

- Natural, native speech examples:

US JP

- Synthetic American-accented Japanese by changing "r":
  JP /ɾ/ → US /ɹ/

	Non-accented	**Accented**
Example 1	▶	▶
Example 2	▶	▶

Result

- Listening test on 131 Japanese native speakers
  - Just mispronouncing this one sound increased the degree of foreign accent by one point on a seven-point scale!
  - The accent was judged as specifically American
- We have synthesised foreign accent from native data alone
  - Highly controllable pronunciation

Take-home message

- We have showed that the same basic technique that is used to learn to how to speak a text also can be used to learn how to control how it is spoken
  - For example
    - Voice, age, gender
    - Language, pronunciation
  - Very fine-grained control is possible

Overview

1. Text to speech
2. Variation in natural speech
3. Synthetic speech
   - 3.1 Recreating variation
   - 3.2 . . . without extra information
   - 3.3 Recent breakthroughs
4. Outlook

Unknown control inputs

- In all previous examples, we knew up front what the proper control input was
  - Training data was annotated (in addition to the text)
  - *Supervised* approach "$\boldsymbol{x} \rightarrow \boldsymbol{y}$"
- What if we only know the text, but not the input values that were used?
  - Very common in practice, since annotation is expensive
  - *Unsupervised* approach "$? \rightarrow \boldsymbol{y}$"

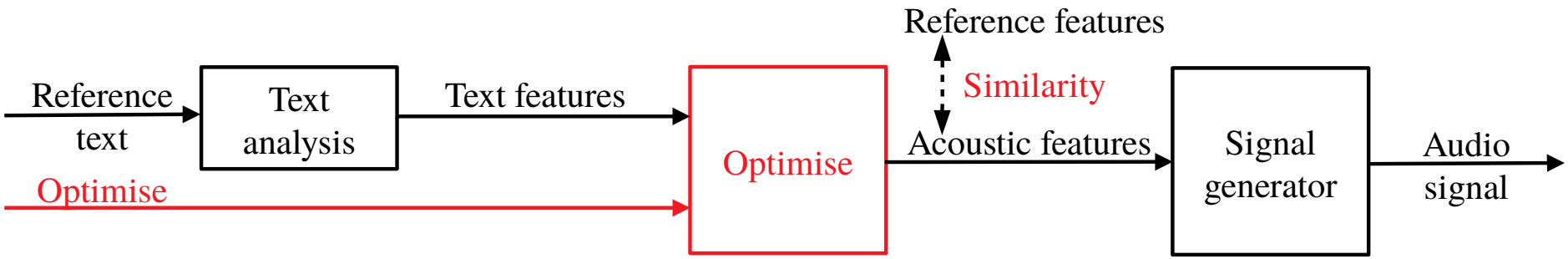
Relevant publication

Content adapted from:

Henter, G. E., Lorenzo-Trueba, J., Wang, X., and Yamagishi, J. (2017). Principles for learning controllable TTS from annotated and latent variation.
In *Proc. Interspeech*, pages 3956–3960

Learning input values

Find the hypothetical control inputs that "best explain" the observations in the database (i.e., give the most accurate prediction):



Expected outcome

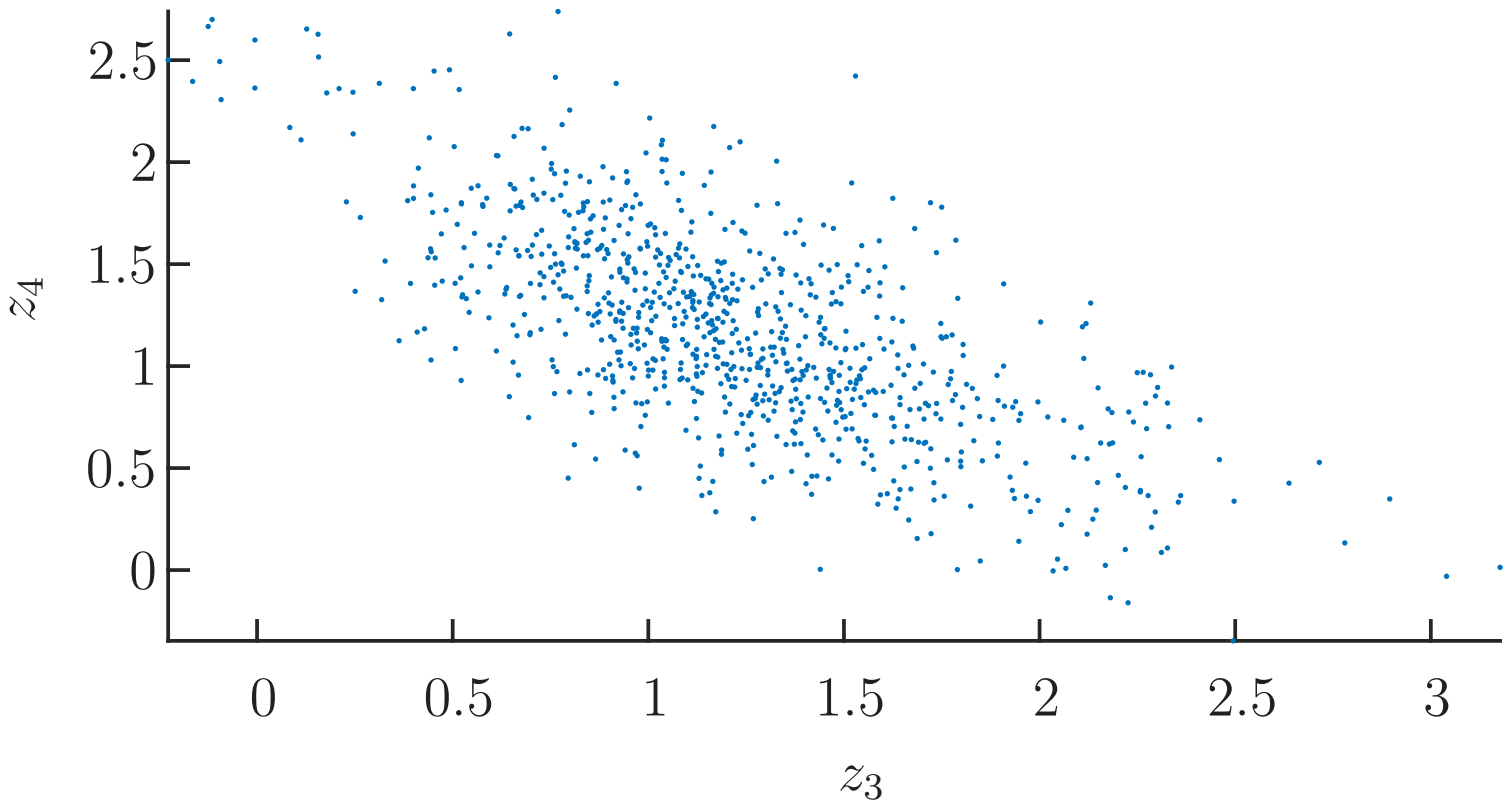
- The result will be a synthesiser with acoustically salient control
- However, we don't know *what* speech aspects it will control
  - Depends on what the most notable source of variation in the database is, according to our chosen similarity measure
  - The learned control knobs won't necessarily make sense to a human
  - "Whatever this button does, it's probably important"

Emotional speech application

- Database of a voice actress reading 8400 Japanese sentences
  - Each read in one of seven emotions
    - Neutral; happy and sad; excited and angry; calm and insecure
    - Tried to keep the expression as constant as possible
  - 1200 sentences for each emotion
- Can we learn nuances in this data beyond the annotated emotion?
  - Add two unknown inputs
    - Potentially different values for each sentence
    - Learn the per-sentence input values together with the synthesiser

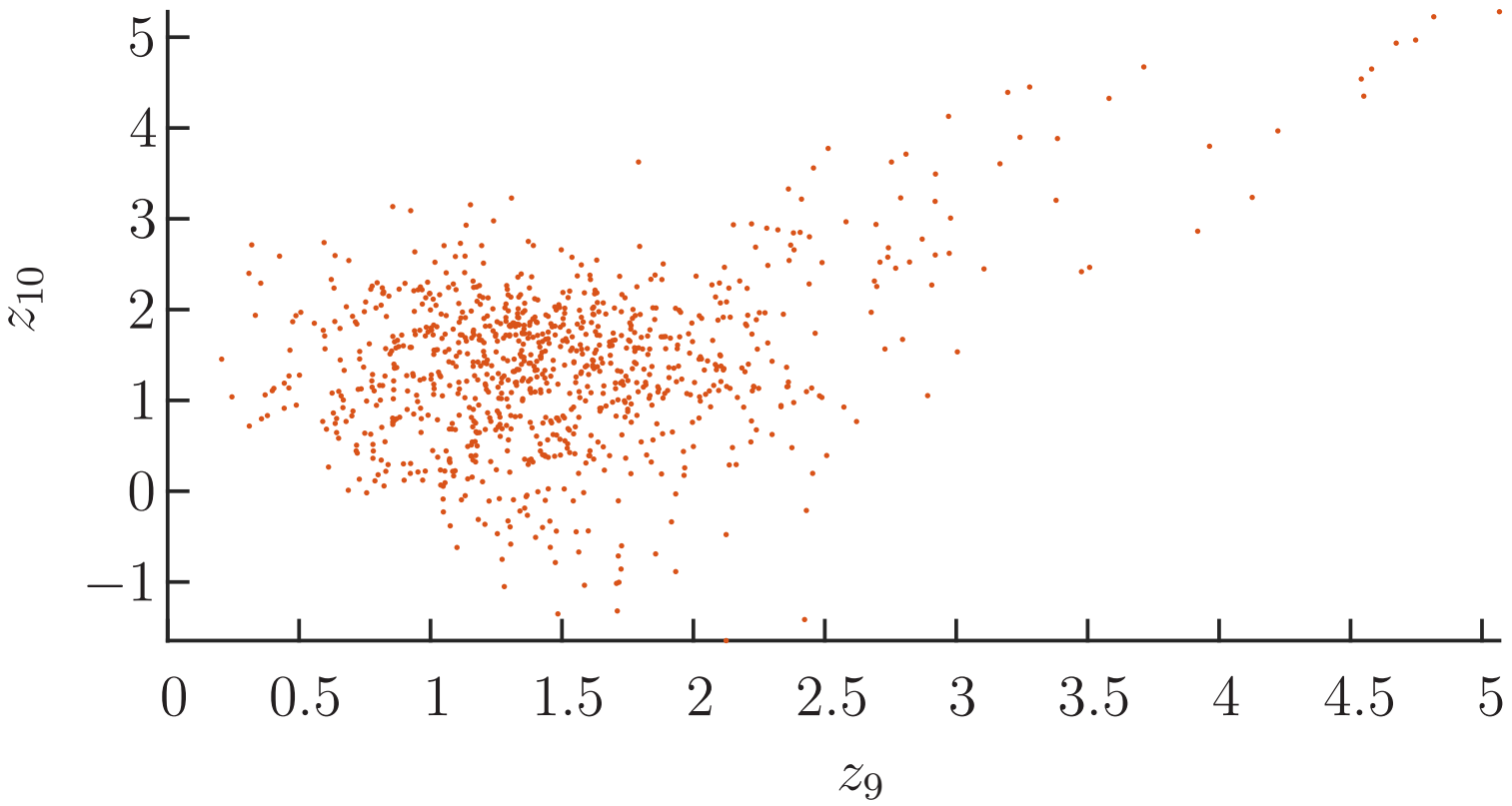
Control input values

Scatterplot of learned control input values (happy speech):



Control input values

Scatterplot of learned control input values (sad speech):



Evaluation

- Listening test with 75 Japanese listeners
  - Listeners could recognise the emotion expressed in synthetic speech nearly as well in the original speech recordings (67% vs. 77%)
  - Listeners could tell apart the same emotion synthesised with different control inputs
- Manipulating control inputs seemed to change the intensity of emotional expression
  - Not rigorously tested

Take-home message

- It is possible to learn to control speech variability and expression without annotating it
- However, it is not easy to tell (or influence) what aspects that the learned inputs will control
  - It is still much more efficient to annotate the control inputs afterwards, instead of the data in advance

Overview

1. Text to speech
2. Variation in natural speech
3. Synthetic speech
   - 3.1 Recreating variation
   - 3.2 . . . without extra information
   - 3.3 Recent breakthroughs
4. Outlook

Speech synthesis aspects

Type	Formant	Unit selection	Statistical	Deep	Waveform
Quality?	Poor	Great	Fair	Good	Excellent
Control?	Great	Poor	Fair	Good	In progress

Relevant publication

Content adapted from:

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K.

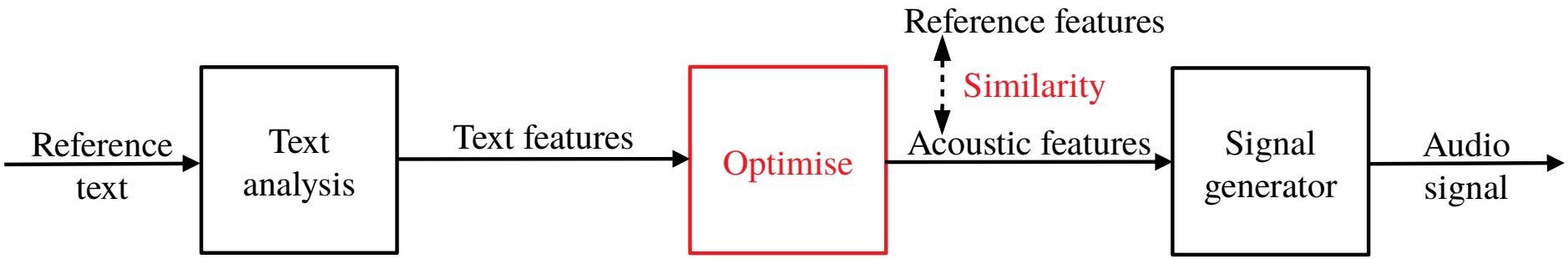
(2016). WaveNet: A generative model for raw audio.
arXiv preprint 1609.03499

WaveNet

- Breakthrough in synthetic speech quality
- First successful model of speech/audio waveforms
  - Not a new idea, but it has taken decades to make it work
- Slow to train and use
  - Much slower than real time
  - Has since been sped up significantly

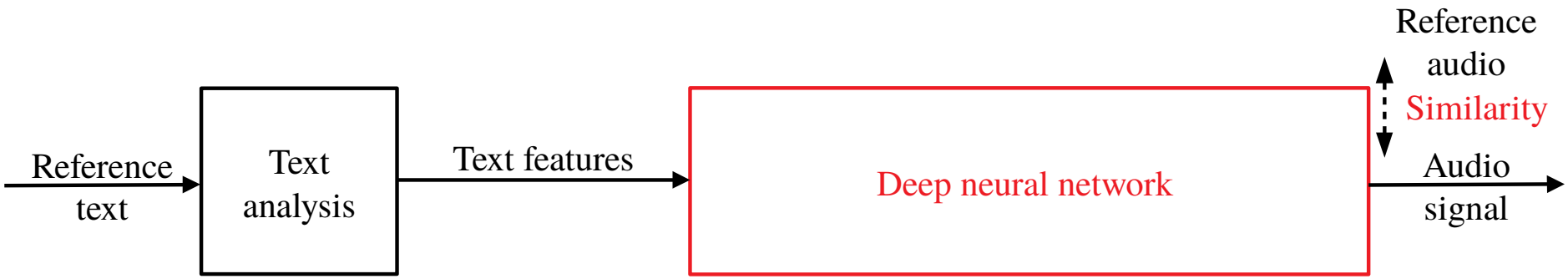
WaveNet idea

The main idea is to predict the waveform (audio signal values) directly:



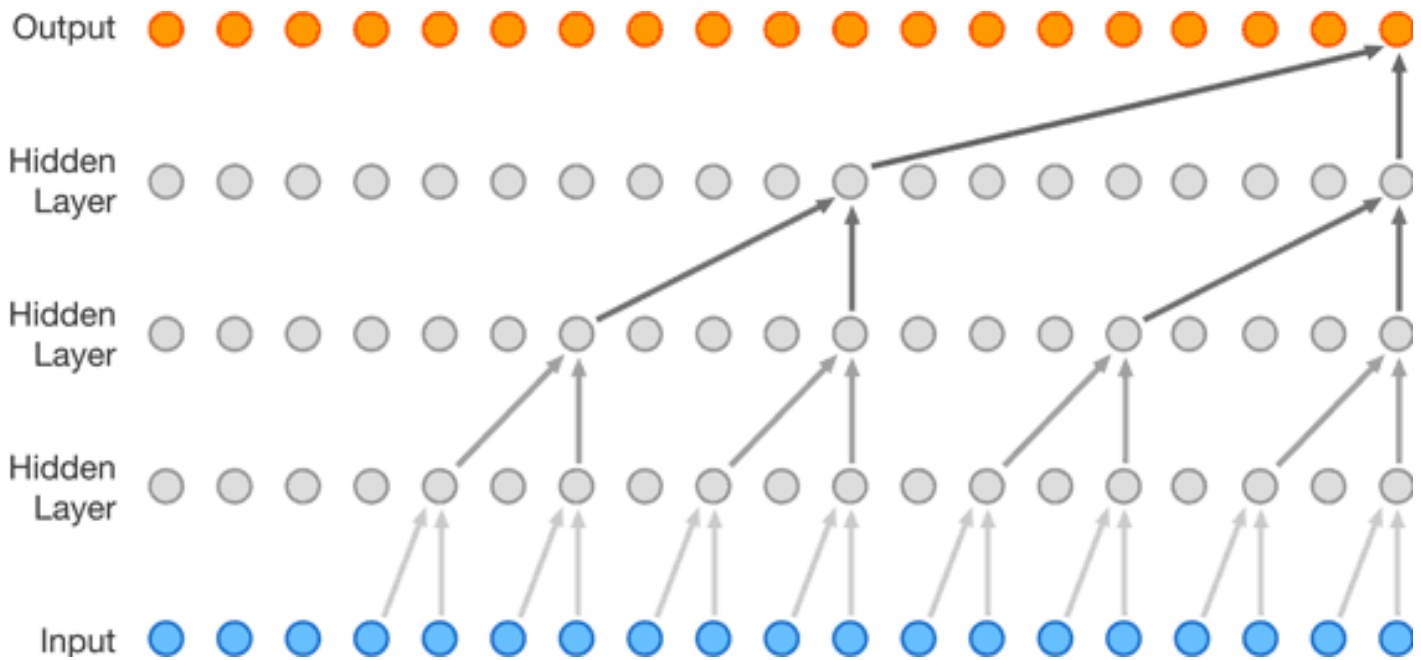
WaveNet idea

The main idea is to predict the waveform (audio signal values) directly:



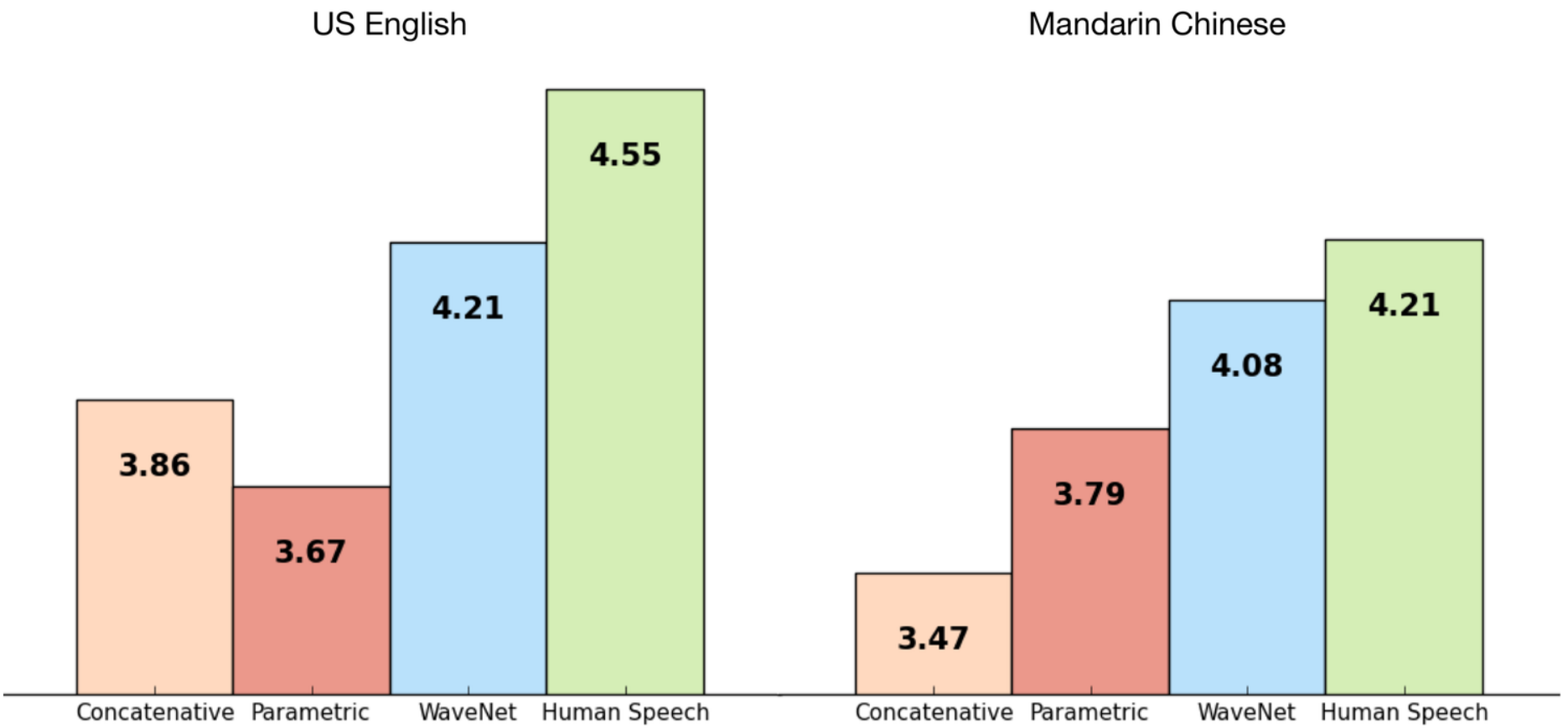
WaveNet model

- Iterated random generation of waveform values based on previously generated values
  - Many thousand iterations per second



Results

WaveNet led to a step change in speech quality ratings:



Example audio

- Comparison against other paradigms (if loudspeakers allow):

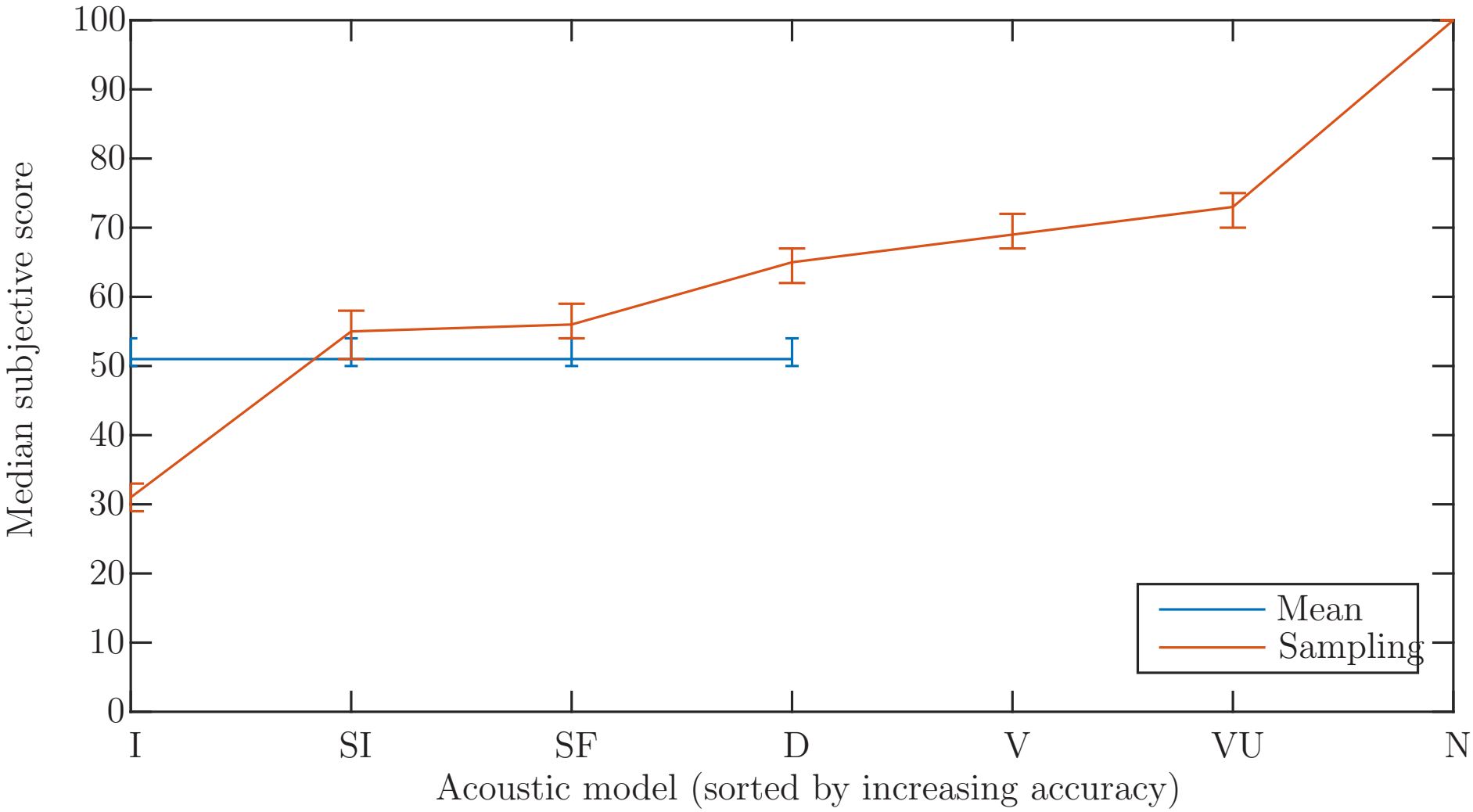
	Unit selection	**Deep parametric**	**WaveNet**
Example 1	▶	▶	▶
Example 2	▶	▶	▶

- Can speak in multiple voices if trained on multi-speaker data:

▶ ▶ ▶ ▶

Recall previous finding

First example of random samples sounding better than mean output:



Relevant publication

Content adapted from:

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. (2017).

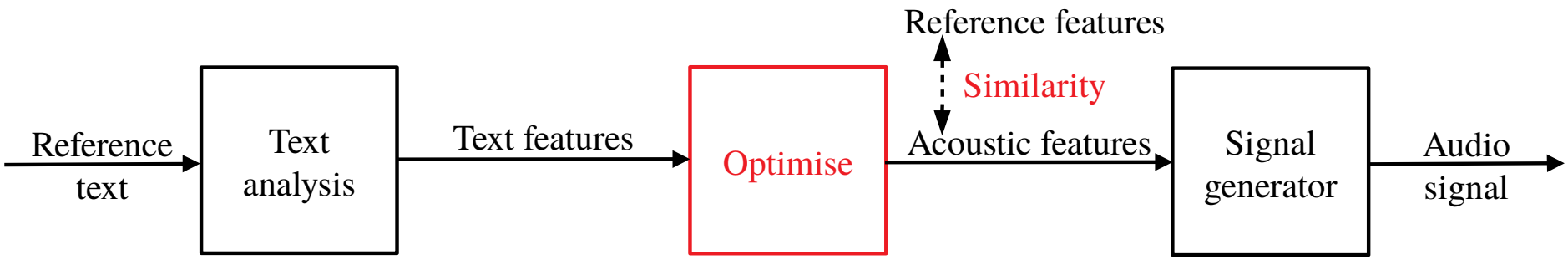
Tacotron: Towards end-to-end speech synthesis.
In *Proc. Interspeech*, pages 4006–4010

“Tacotron” (talk-o-tron)

- Breakthrough in text processing
- First successful example of learning to read text from letters alone, without language information
- Much richer prosody than previous text-to-speech systems

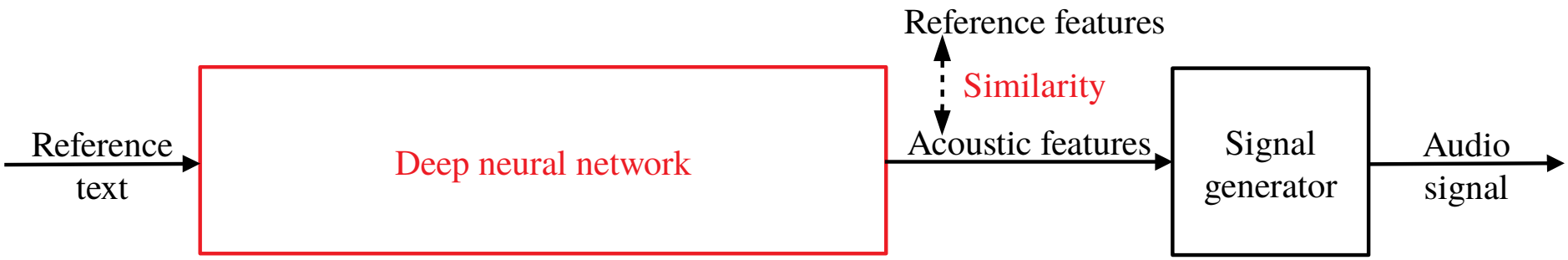
Tacotron idea

The main idea is to learn to speak directly from text characters:



Tacotron idea

The main idea is to learn to speak directly from text characters:



Example audio

- Tacotron correctly pronounced difficult, previously unseen words

▶ ▶

- Tacotron changes prosody based on punctuation
  - "This is your personal assistant[,] Google Home."

w/ comma **w/o comma**

▶ ▶

Relevant publication

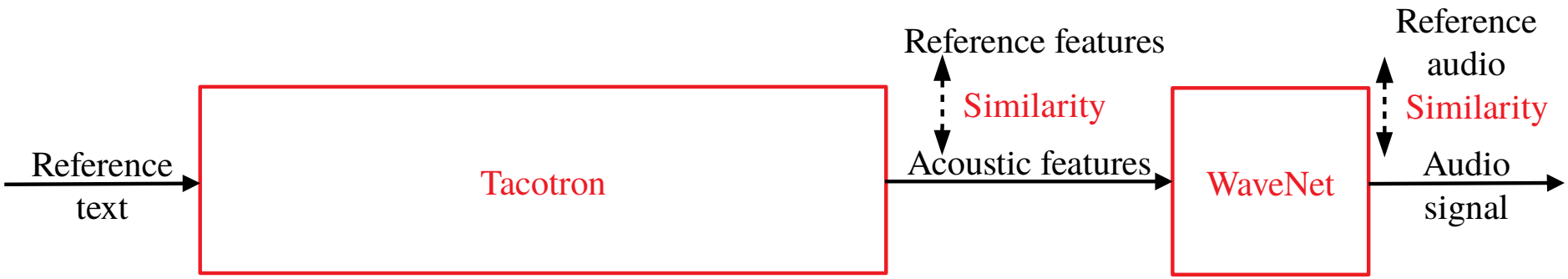
Content adapted from:

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A.,

Agiomyrgiannakis, Y., and Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions.
In *Proc. ICASSP*, pages 4799–4783

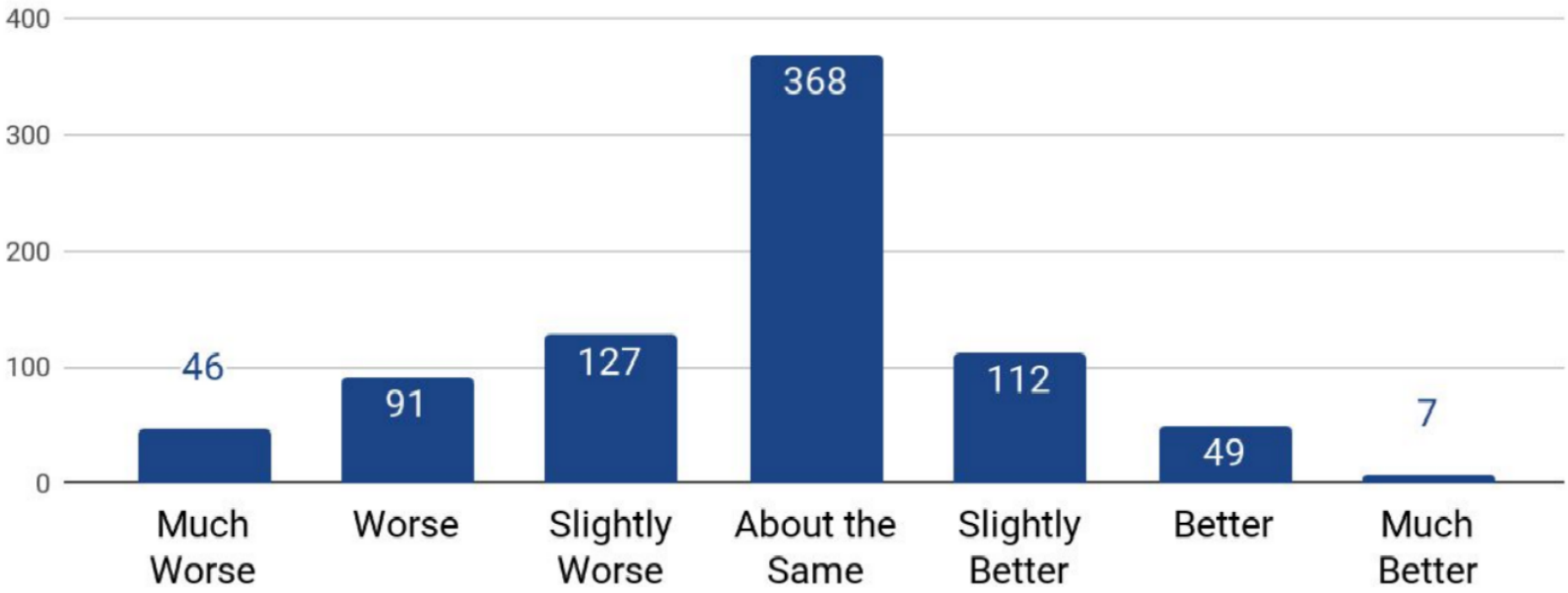
Tacotron 2

WaveNet and Tacotron can be combined for great prosody *and* excellent quality:



Result

- In a large listening test, Tacotron 2 synthesised sentences were almost indistinguishable from human recordings
  - Though only this particular voice (with very high-quality data) was tested



Speech synthesis aspects

Type	Formant	Unit selection	Statistical	Deep	Waveform
Quality?	Poor	Great	Fair	Good	Excellent
Control?	Great	Poor	Fair	Good	In progress

Hot research topic

- Combine breakthrough methods with the ability to learn control
  - The WaveNet paper showed that supervised control over speaker voice is possible
  - What about unsupervised control?

Relevant publication

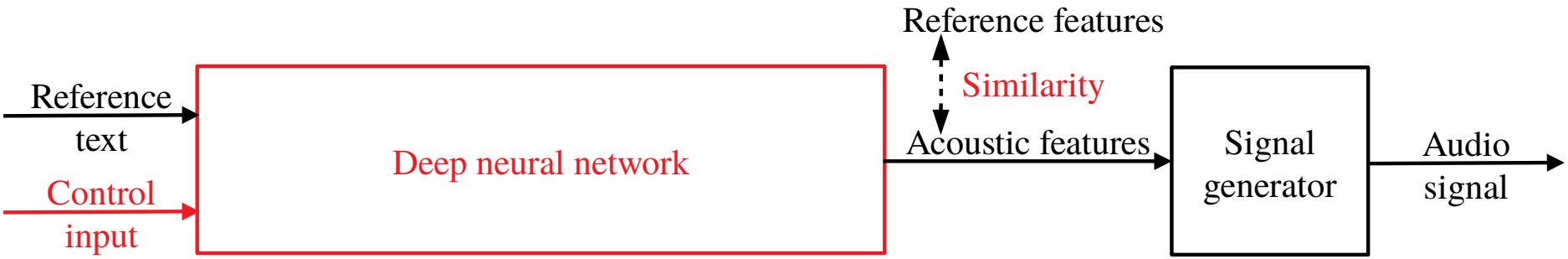
Content adapted from:

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E.,

Shor, J., Xiao, Y., Ren, F., Jia, Y., and Saurous, R. A. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis.
In *Proc. ICML*, pages 5180–5189

Tacotron with control

Learn unannotated control inputs together with the synthesiser, like before:



Example audio

- Application to 147 hours expressive speech from audiobooks
  - Seemed to learn a set of archetypical reading styles

▶ ▶ ▶ ▶ ▶

- Application to corrupted data
  - 90% had noise, reverb, music, or similar added
  - Different learned styles approximate to different corruptions

w/o control **“Noise”** **“Reverb”** **“Music”** **“Clean”**

▶ ▶ ▶ ▶ ▶

Take-home message

- State-of-the-art synthetic speech quality and intonation are now better than two years ago
  - More or less on par with recorded human speech
  - For read speech with large, high-quality databases
- Simple extensions of the latest systems can also achieve very successful output control
  - Per sentence; short time-scale control remains untested

Overview

1. Text to speech
2. Variation in natural speech
3. Synthetic speech
   - 3.1 Recreating variation
   - 3.2 . . . without extra information
   - 3.3 Recent breakthroughs
4. Outlook

Perspective

1. Solved problems
   - Speech signal quality (from high-quality data)
   - Learning control from annotated data
2. Problems being solved
   - Read speech prosody and pronunciation
   - Learning expressiveness from unannotated or partially annotated data
   - Coping with and leveraging speech variability
3. Problems yet to be solved
   - Choosing appropriate prosody and expression in communication
   - Exploring applications of truly convincing synthetic speech

The end

The end

Thank you for listening!

The end

Time for questions

References I

Henter, G. E., Lorenzo-Trueba, J., Wang, X., Kondo, M., and Yamagishi, J. (2018).
Cyborg speech: Deep multilingual speech synthesis for generating segmental foreign accent with natural prosody.
In *Proc. ICASSP*, pages 4799–4803.

Henter, G. E., Lorenzo-Trueba, J., Wang, X., and Yamagishi, J. (2017).
Principles for learning controllable TTS from annotated and latent variation.
In *Proc. Interspeech*, pages 3956–3960.

Henter, G. E., Merritt, T., Shannon, M., Mayo, C., and King, S. (2014).
Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech.
In *Proc. Interspeech*, pages 1504–1508.

Luong, H.-T., Takaki, S., Henter, G. E., and Yamagishi, J. (2017).
Adapting and controlling DNN-based speech synthesis using input codes.
In *Proc. ICASSP*, pages 4905–4909.

Oertel, C., Jonell, P., Kontogiorgos, D., Mendelson, J., Beskow, J., and Gustafson, J. (2017).
Crowd-sourced design of artificial attentive listeners.
In *Proc. Interspeech*, pages 854–858.

References II

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. (2018).
Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions.
In *Proc. ICASSP*, pages 4799–4783.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016).
WaveNet: A generative model for raw audio.
arXiv preprint 1609.03499.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. (2017).

Tacotron: Towards end-to-end speech synthesis.
In *Proc. Interspeech*, pages 4006–4010.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., and Saurous, R. A. (2018).
Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis.
In *Proc. ICML*, pages 5180–5189.