The NITech text-to-speech system for the Blizzard Challenge 2018

Kei Sawada[1,2], Takenori Yoshimura[1], Kei Hashimoto[1], Keiichiro Oura[1], Yoshihiko Nankaku[1], Keiichi Tokuda[1]

[1]Nagoya Institute of Technology (NITech)
[2]Microsoft Development Co., Ltd.

Blizzard Challenge 2018 Workshop on Sep. 8, 2018

Background

- **Text-to-speech (TTS) systems**
  - TTS systems are used in various applications
  - Demand for TTS systems is increasing
    - *High-quality, various speaking styles, various languages, etc.*
  - Success by introducing deep learning
    - *DNN, LSTM, WaveNet, Deep Voice, Char2Wav, Tacotron, etc.*
- **Evaluations of TTS systems**
  - Comparisons are difficult when the training corpus, task, and listening test are different
  - Blizzard Challenge [Black et al. ’05]
    - *In order to better understand and compare research techniques in constructing corpus-based TTS systems with the same data*
- **NITech TTS system for the Blizzard Challenge**
  - NITech have been submitting a statistical speech synthesis system to the Blizzard Challenge since 2005

Blizzard Challenge 2015-2018

Task

- Construct a TTS system from children’s audiobooks that is suitable for reading audiobooks to children

Data

- Children’s audiobooks were recorded by one female speaker
  - *2015 (pilot task year): 2 hours*
  - *2016: 5 hours*
  - *2017, 2018: 7 hours*
- Mismatches between speech data and text
  - *Misreading, onomatopoeia, etc.*
- Speech data includes various speaking styles
  - *Emotions, characters, singing voices, etc.*

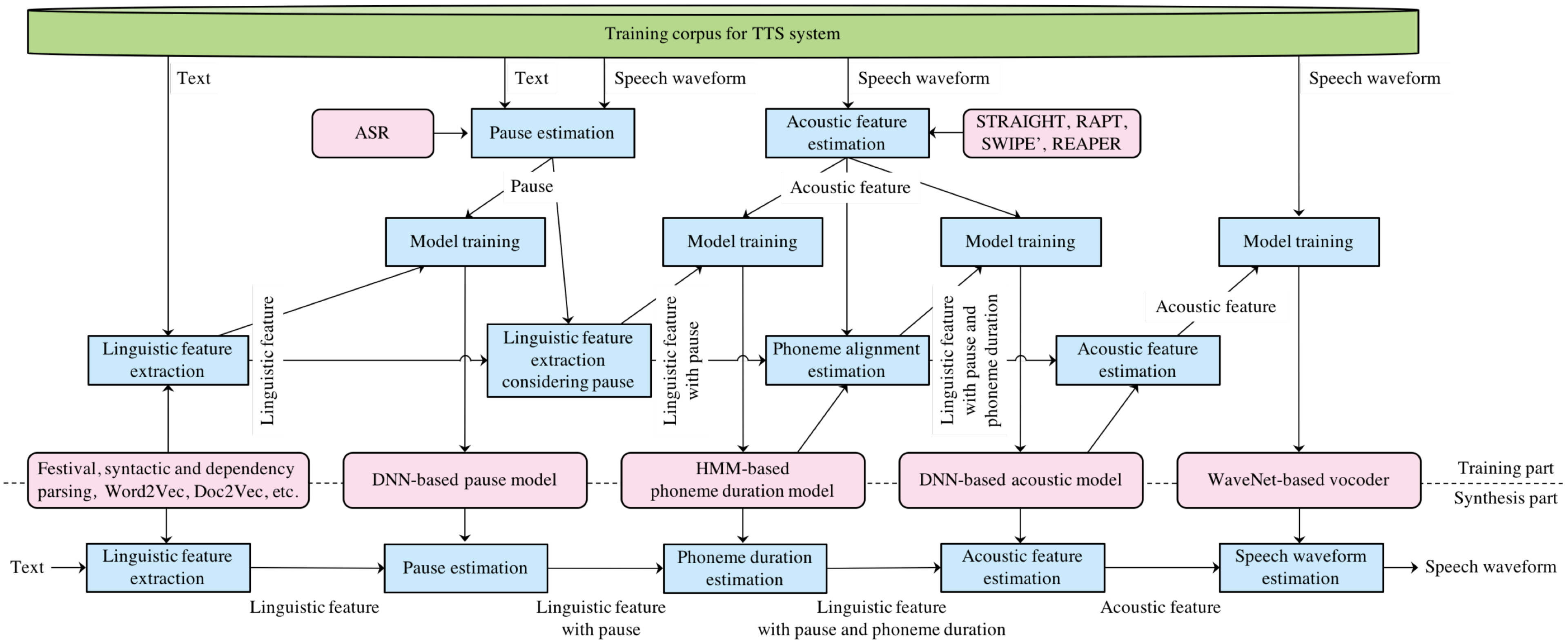
"I'm king of the jungle," roared Lion.
"I'm going to eat you all up."
"No!" cried the jungle animals.

Character1
Character2
Descriptive part

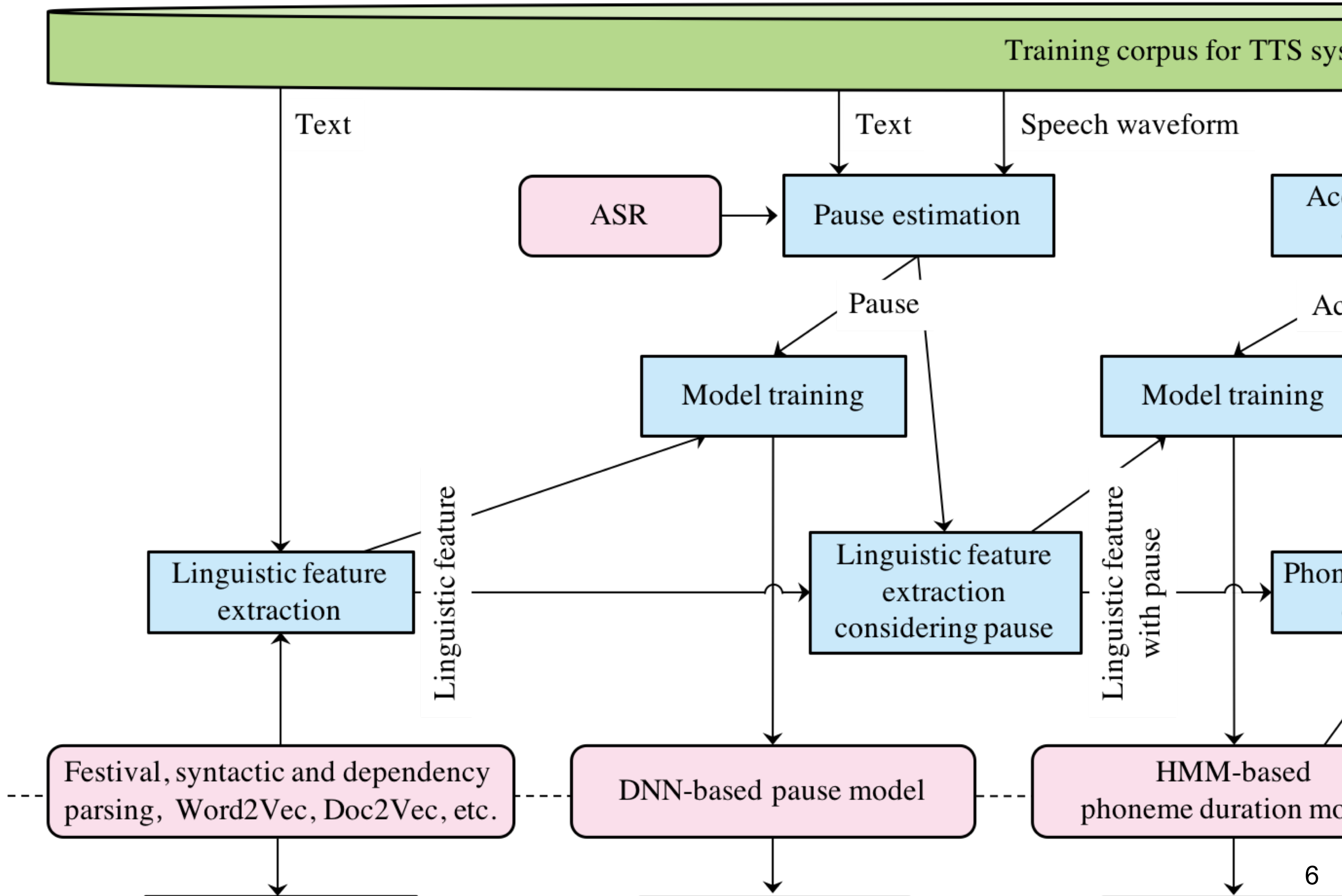
NITech 2015-2018 TTS systems

- **NITech 2015 TTS system**
  - Pruning of training data based on ASR
  - Introduction of linguistic features based on quotation marks
- **NITech 2016 TTS system**
  - Automatic construction of training corpus based on ASR
  - Introduction of linguistic features based on syntactic parsing
  - Introduction of DNN acoustic model considering GV trajectory
- **NITech 2017 TTS system**
  - Introduction of linguistic features which can predict and reproduce speaking style from text
  - Introduction of MDN acoustic model considering GV trajectory
- **NITech 2018 TTS system**
  - Introduce pause insertion model
  - Introduce WaveNet vocoder

NITech 2018 TTS system



NITech 2018 TTS system



Pause insertion model

Pause insertion

- Pause is used as one of emotional expressions
  ⇒ Introduce pause insertion model to reproduce pause insertion style of training corpus

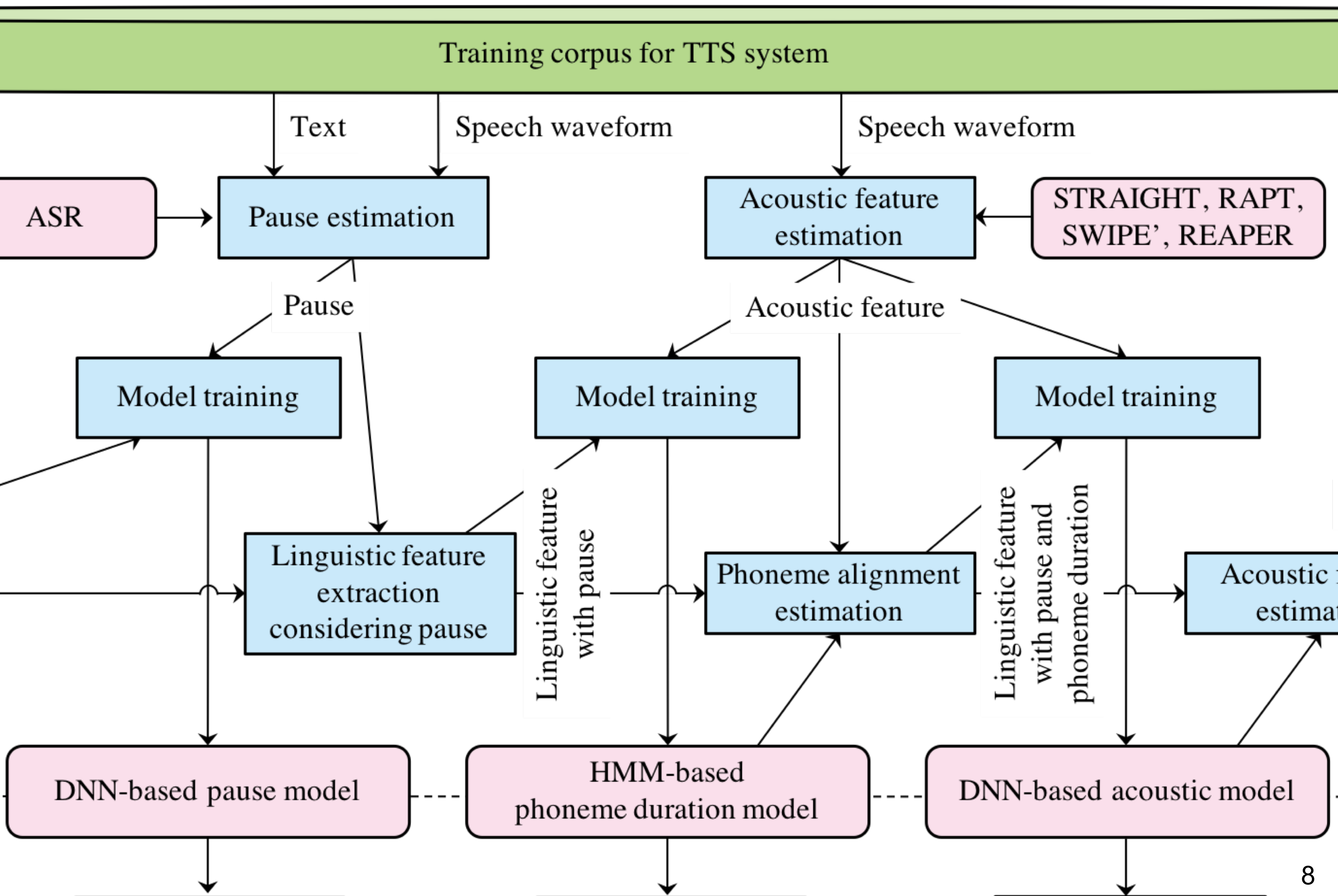
Pause estimation of training corpus

- Phoneme alignment estimation including short pause at all word boundaries
- Short pause model (HMM with state skip transition)
- Duration of estimated short pause is equal to or greater than threshold ⇒ word boundary contains a pause

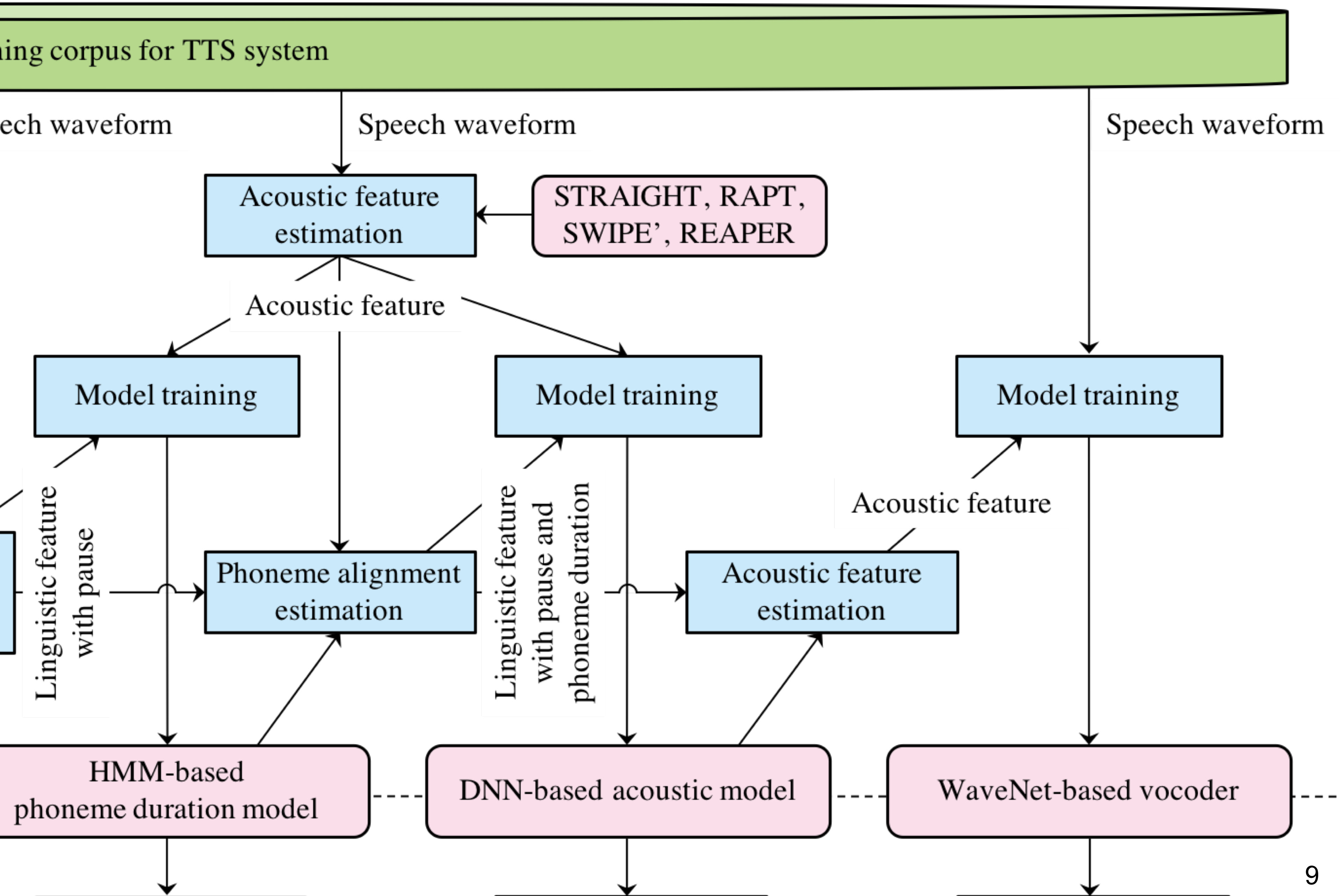
Pause insertion model

- Bi-directional gated recurrent unit (GRU)
- Input: linguistic features of word- and sentence-level
- Output: whether or not a pause is inserted after the word (0 or 1)

NITech 2018 TTS system



NITech 2018 TTS system



WaveNet-based vocoder

- **Frame-level vocoder**
  - Vocoder introduce degradation in speech quality
    ⇒ Introduce neural vocoder
- **WaveNet vocoder** [van den Oord et al.; ’16], [Tamamori et al.; ’17]
  - Directly modeling and generation speech waveform
  - Modeling speech waveform as classification problem
  - Quantization scheme introduces flat white noise
    ⇒ Introduce noise shaping quantization
    [Yoshimura et al.; ’18]
- **Mel-cepstrum-based quantization noise shaping**
  - Quantization noise considering human auditory
  - Apply mel-cepstrum-based prefilter to speech signals
    [Tokuda et al.; ’94]

Realize high-quality speech waveform generation

Experimental conditions (1/3)

Conditions of training corpus construction

Provided data	1258 pages
Acoustic features	12 dim. MFCC + Δ + ΔΔ
Acoustic model	3 state left-to-right tri-phone GMM-HMM
Language model	Tri-gram
Training corpus for TTS	924 pages

Conditions of pause insertion model

Input features	251 dim. linguistic features
Structure of DNN	Bi-directional gated recurrent unit, 3 hidden layers, 128 units, ReLU
Training algorithm	Adam, dropout rate 20%

Experimental conditions (2/3)

Conditions of phoneme duration model

Sampling rate	32 kHz
Acoustic features	64 dim. STRAIGHT mel-cepstrum, log F0, 32 dim. mel-cepstrum aperiodicity measure + Δ + ΔΔ
Number of questions	925 questions
Structure of HMM	5 state left-to-right MSD-HSMM

Conditions of acoustic model

Sampling rate	32 kHz
Acoustic features	64 dim. STRAIGHT mel-cepstrum, log F0 V/UV info., 32 dim. mel-cepstrum aperiodicity measure
Linguistic features	1685 dim.
Structure of DNN	Single-mixture density network, 3 hidden layers, 8000 units, sigmoid
Training algorithm	SGD, dropout rate 60%, Trajectory training considering GV

Experimental conditions (3/3)

Conditions of WaveNet vocoder

Sampling rate	32 kHz
Quantization	8 bit μ-law
Noise shaping parameters	γ =0.1, β=0.1
Structure of WaveNet	Dilation: [1, 2, 4, ..., 512] 3 stacks Dilation, residual, skip: 256
Condition features of WaveNet	98 dim. acoustic features
Training algorithm	Adam

Demo

The picture is quoted from the Usborne Publishing.

Evaluation results of sentence domain

Naturalness

MOS	ID
4.8	A
4.0	K
3.7	J
3.5	I
3.0	L, M
2.9	B
2.8	D
	⋮

Similarity

MOS	ID
4.5	A
3.9	K
3.6	J
3.5	I
3.4	L
3.2	B
3.0	M
	⋮

Intelligibility

WER	ID
11	I
14	E, O
15	D, G
16	K
17	N
18	J
20	F
	⋮

A: Natural speech
B, C, D, E: Benchmark TTS systems
I: NITech TTS system
Red line: Significant difference between NITech and other systems

Evaluation results of page domain

Overall impression

MOS	ID
48	A
38	K
34	J, I
29	B
28	L
	⋮

Pleasantness

MOS	ID
48	A
37	K
33	J, I
28	L, B
26	M
	⋮

Speech pause

MOS	ID
48	A
36	K, J
32	I
31	D, G
30	E
	⋮

Stress

MOS	ID
48	A
36	K
35	J
33	I
30	D, G
	⋮

Intonation

MOS	ID
48	A
37	K
35	J
33	I
28	D
	⋮

Emotion

MOS	ID
48	A
38	K
35	J, I
31	B
30	M
	⋮

Listening effort

MOS	ID
49	A
37	K
34	J
33	I
28	D
	⋮

Comparison of NITech 2017 and 2018

- **Introduction of WaveNet vocoder**
  - Improved naturalness and speaker similarity
  - Sometimes ambiguous pronunciation
    - *Multiple codecs and noise made WaveNet training difficult*
  - Reduced reproducibility of speaking styles
    - *Training data is insufficient to reproduce various speaking styles*

2017	2018

http://www.sp.nitech.ac.jp/~swdkei/syn/Blizzard_2018/index.html

Conclusion

- **NITech TTS system for the Blizzard Challenge 2018**
  - Introduce pause insertion model
  - Introduce WaveNet vocoder
  - Large-scale subjective listening tests
    - *Synthesized highly natural, similar, and intelligible speech*
  - Comparison of NITech 2017 and 2018 TTS systems
    - *Improved naturalness and speaker similarity*
    - *Insufficient accuracy of WaveNet vocoder*

- **Future work**
  - Generate expressive synthesized speech in neural vocoder
  - Introduce end-to-end approach