The NITech text-to-speech system for the Blizzard Challenge 2017

Kei Sawada, Kei Hashimoto,
Keiichiro Oura, Keiichi Tokuda

Nagoya Institute of Technology (NITech)

Blizzard Challenge 2017 Workshop on Aug. 25, 2017

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.*

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 parametric speech synthesis (SPSS) system to the Blizzard Challenge since 2005

Blizzard Challenge 2017 task

Blizzard Challenge 2017

- Task
  - *Construct a TTS system from children’s audiobooks that is suitable for reading audiobooks to children*
- Data
  - *7 hours speech data and text pairs*
  - *56 books were recorded by one female English speaker*
  - *Speech data includes various speaking styles, emotions, characters, etc.*
  - *Example of provided data*

> "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

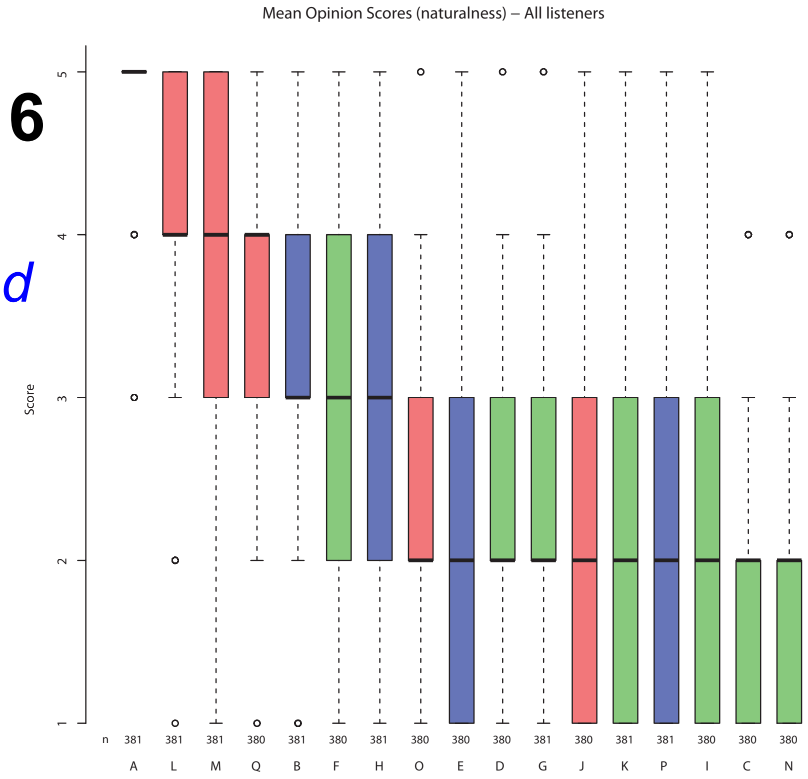
Review of Blizzard Challenge 2016

Blizzard Challenge 2016

- The task was almost same as the Blizzard Challenge 2017
- Difference was the amount of training data (2016: 5 hours, 2017: 7 hours)

Result of Blizzard Challenge 2016

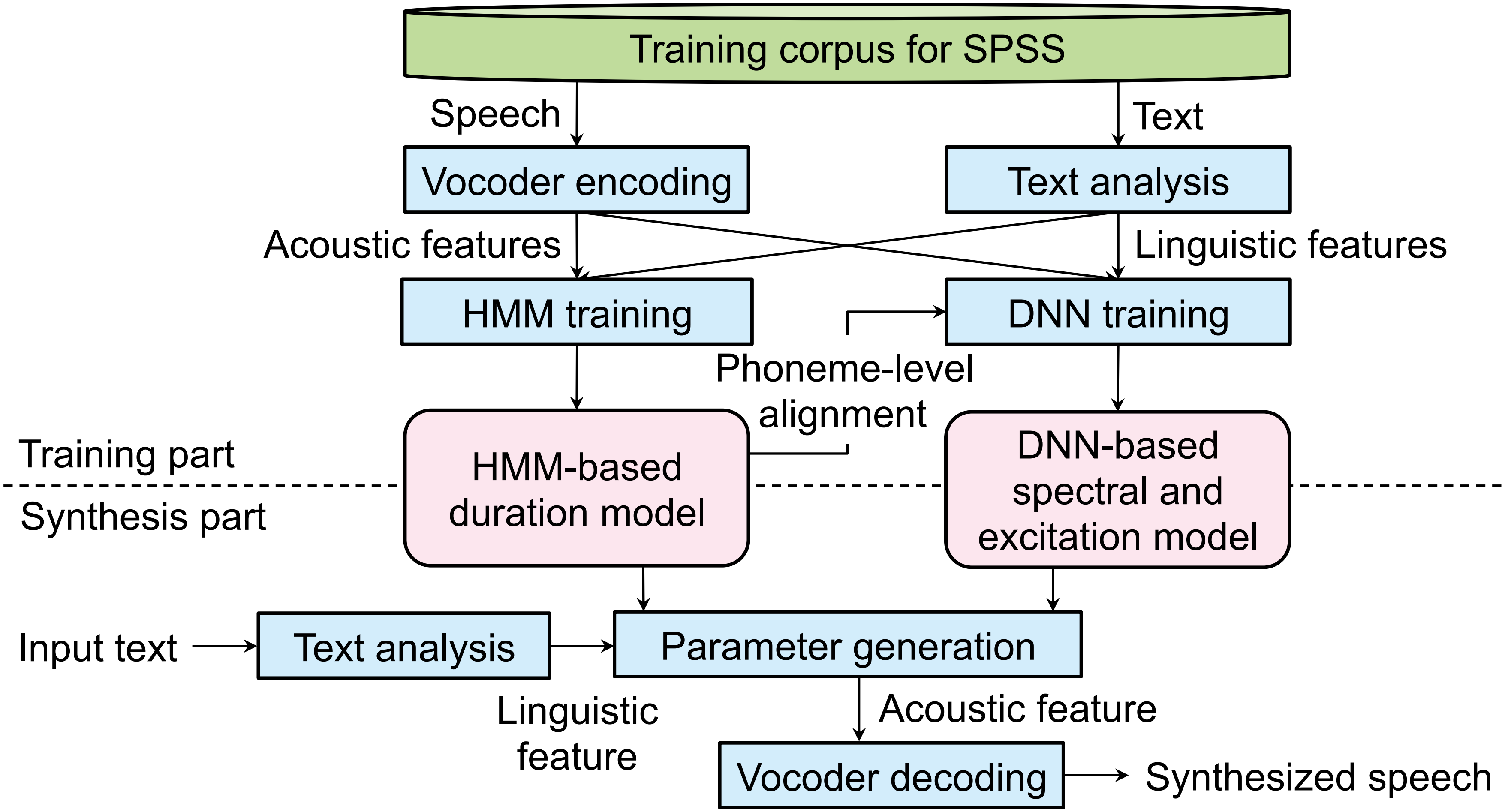
- MOS for naturalness
  - *MOS of SPSS systems is not so good*
- Why?
  - *Training corpus includes various speaking variations*
    ⇒ Modeling is difficult



Red: hybrid systems
Blue: unit selection systems
Green: SPSS systems

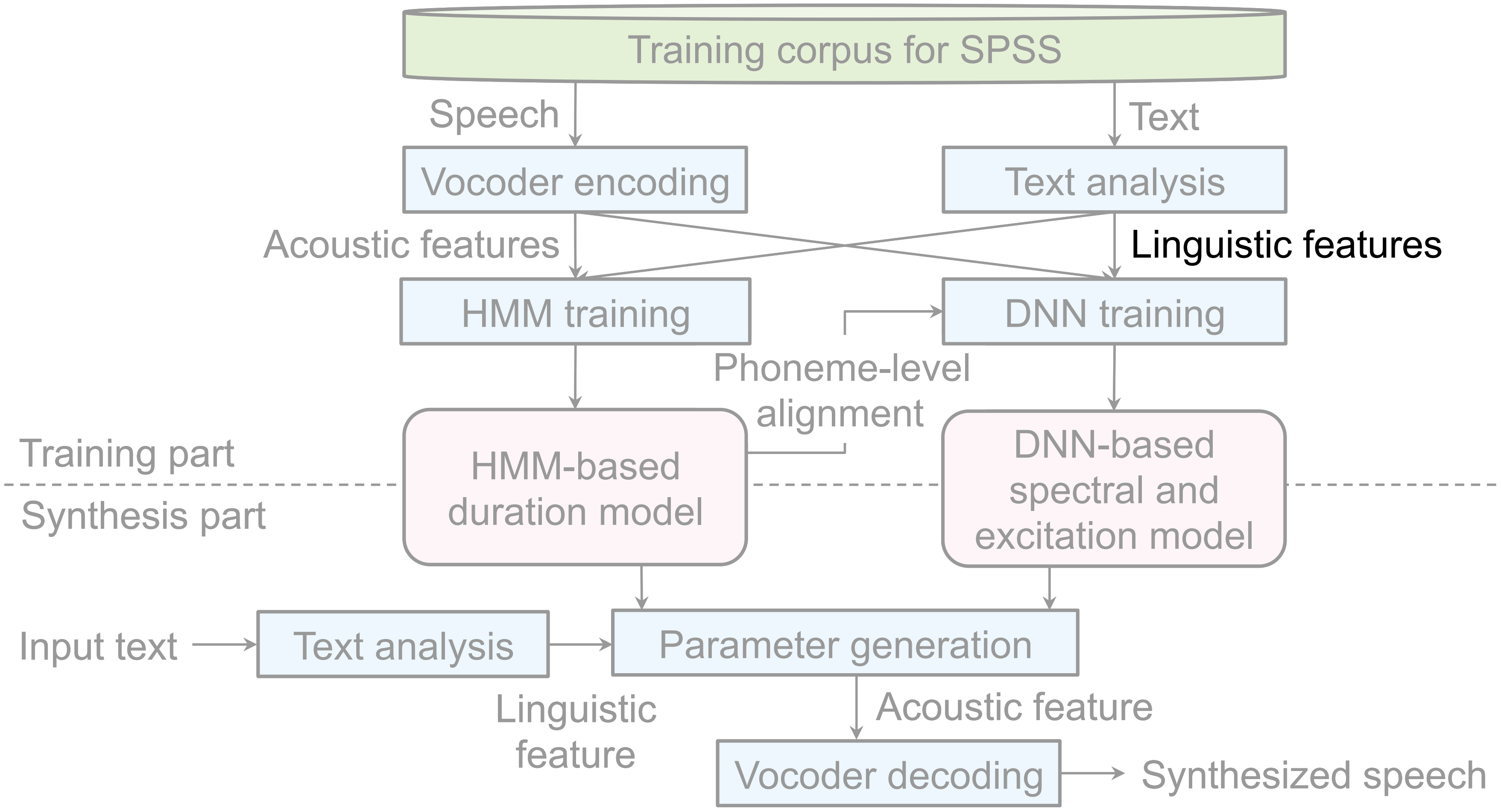
Redesign of linguistic features for audiobooks in SPSS

NITech TTS system



- Linguistic features for audiobooks in SPSS
- Trajectory training considering GV for mixture density networks

NITech TTS system



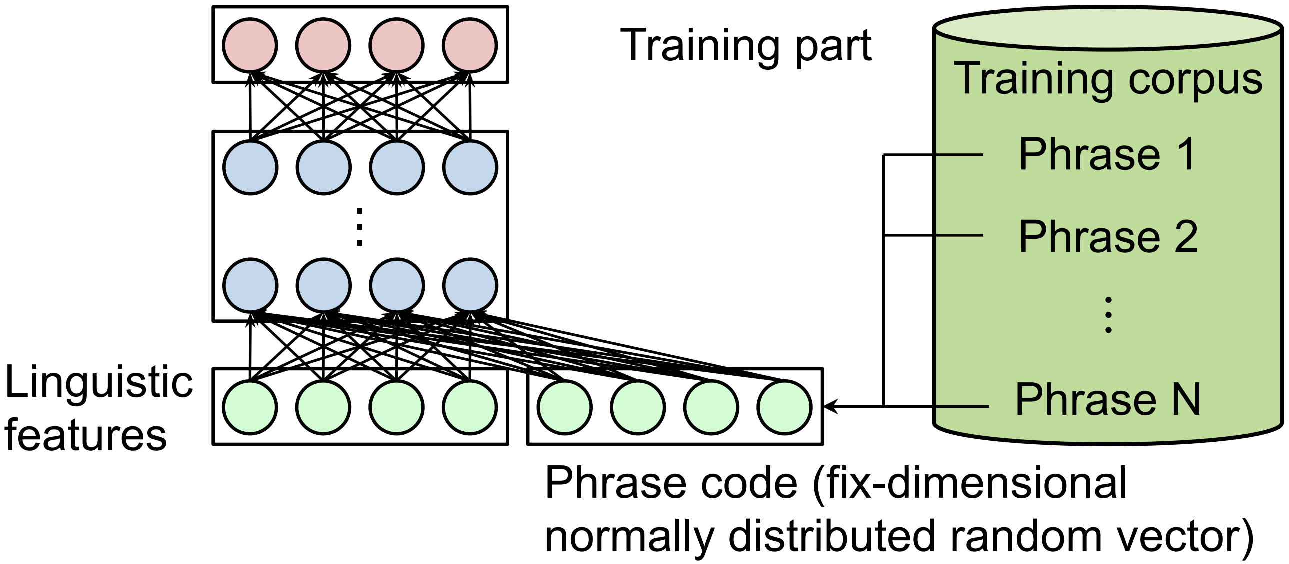
- Linguistic features for audiobooks in SPSS
- Trajectory training considering GV for mixture density networks

Linguistic features for audiobooks

- **Linguistic features**
  - Features obtained from texts express pronunciations
  - Suitable design for audiobooks is necessary
- **Additional linguistic features to HTS-2.3.1 demo script**
  - Page-level information
    - *Capture supra-sentential information*
  - Syntactic and dependency parsing information
    - *Capture sentence structure*
  - Type of sentence
    - *Distinguish different type of sentence*
  - Double quotes information
    - *Distinguish between descriptive and conversational parts*
  - Word and phrase codes
    - *Distinguish each word and phrase variation*

Phrase code

- **Training part**
  - A unique value (phrase code) is assigned to each phrase
  - Phrases are distinguished to represent speaking variation



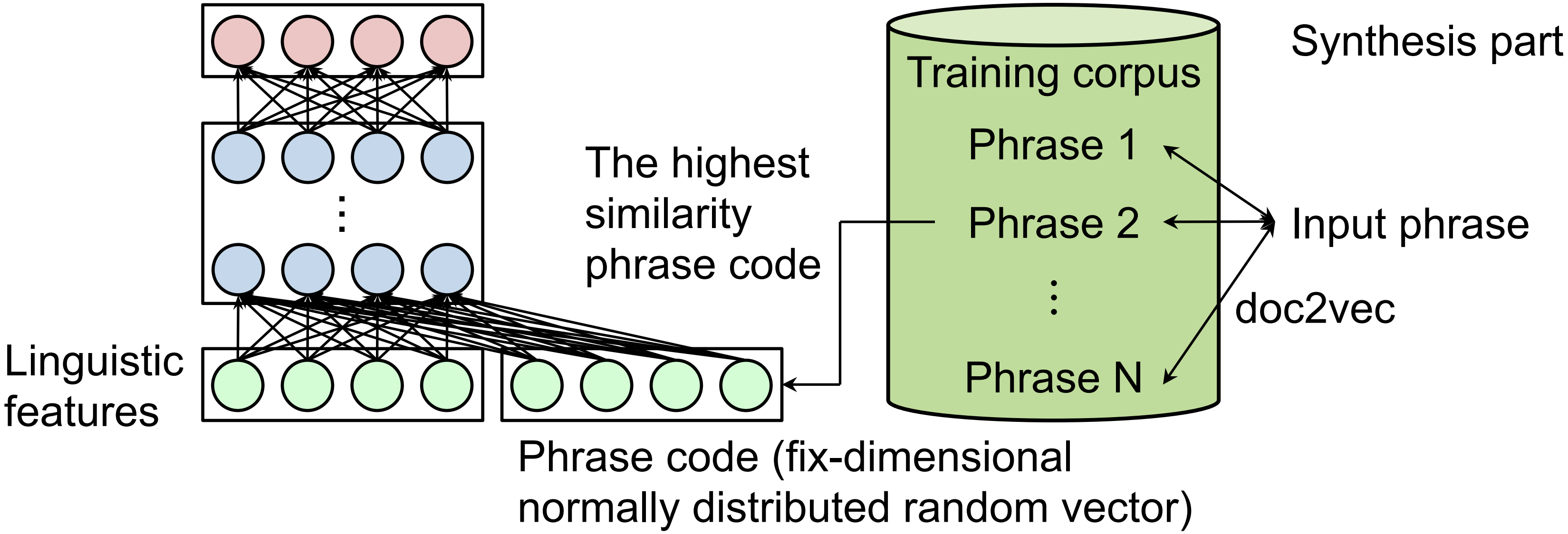
Phrase code

Training part

- A unique value (phrase code) is assigned to each phrase
- Phrases are distinguished to represent speaking variation

Synthesis part

- Phrase is vectorized by using doc2vec [Le et al. ’14]
- Phrase similarity between training and input phrases is calculated from vectorized ones
- Phrase code of the highest similarity phrase is used


Synthesis part
Training corpus
Phrase 1
Phrase 2
⋮
Phrase N
Input phrase
doc2vec
The highest similarity phrase code
Linguistic features
Phrase code (fix-dimensional normally distributed random vector)

Phrase code

Training part

- A unique value (phrase code) is assigned to each phrase
- Phrases are distinguished to represent speaking variation

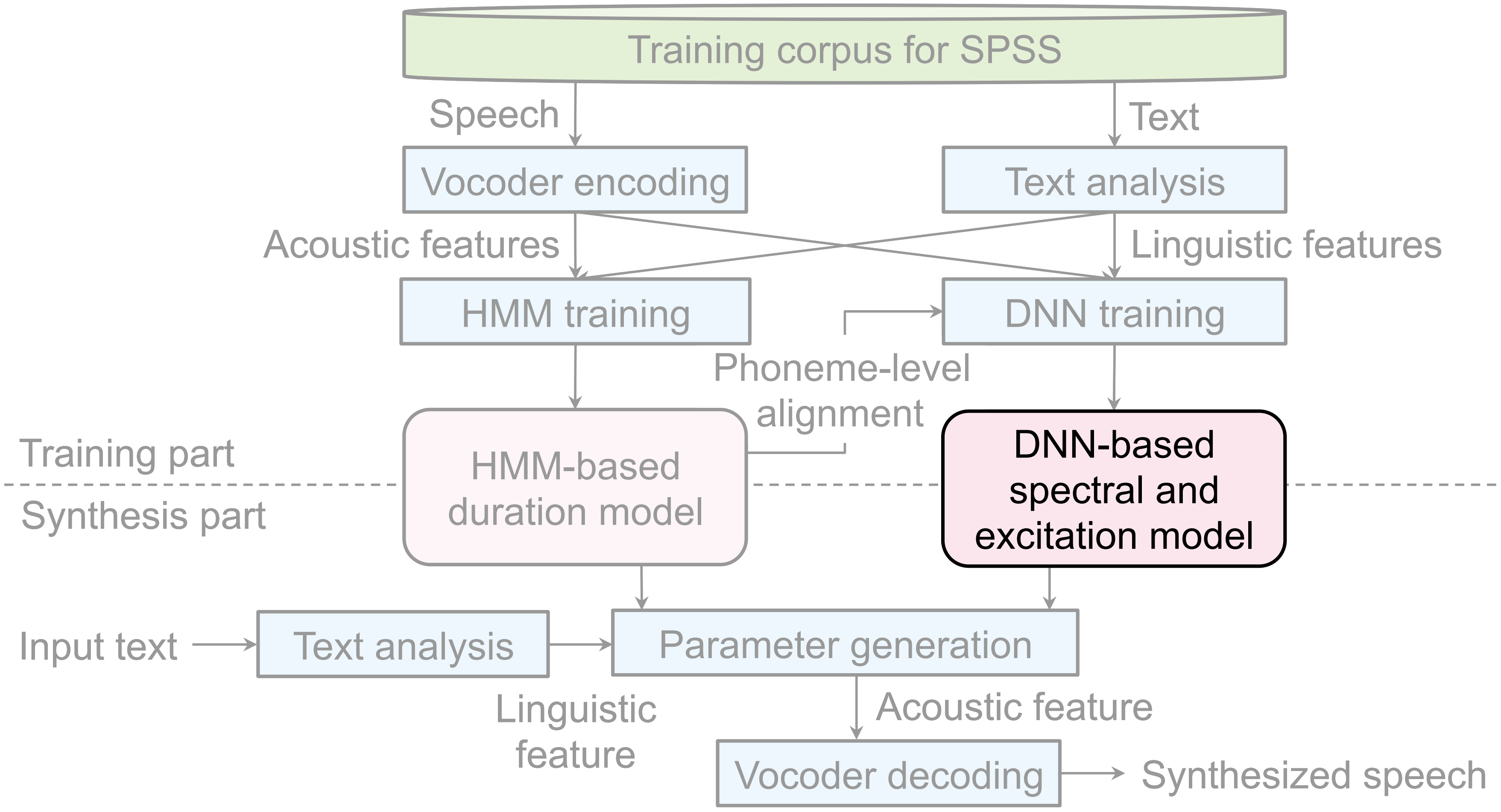
Synthesis part

- Phrase is vectorized by using doc2vec [Le et al. ’14]
- Phrase similarity between training and input phrases is calculated from vectorized ones
- Phrase code of the highest similarity phrase is used

Input text	Text of phrase adaptation	Synthesized speech
"I must tell Hamlet.”	Zero vector (average phrase)	🔊
	Come and see the friendly lion! 🔊	🔊
	"Who's been sitting in my chair?" 🔊	🔊
	"I must tell the King.” (highest similarity phrase) 🔊	🔊

Realize expressive speech synthesis

NITech TTS system

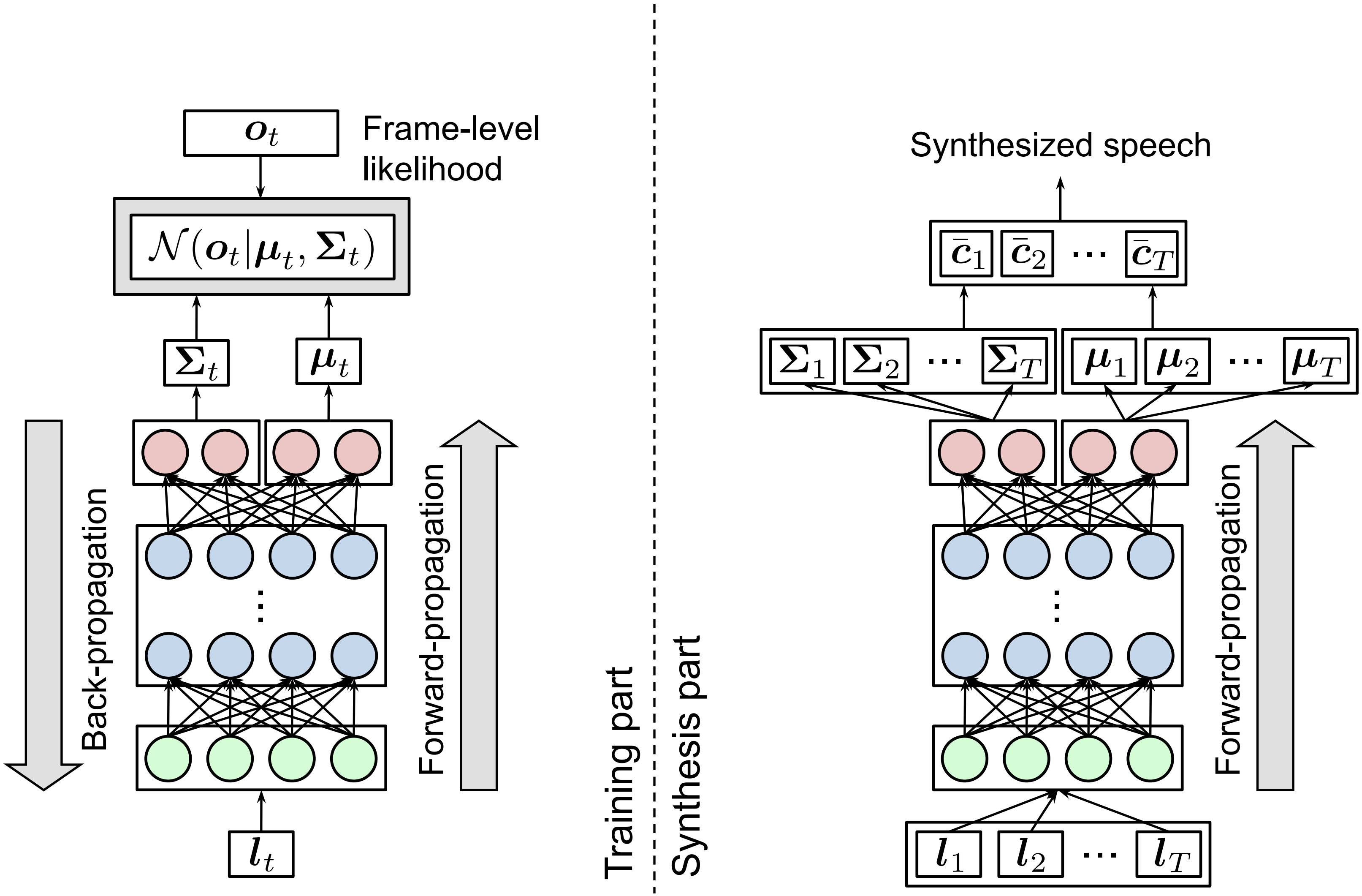


- Linguistic features for audiobooks in SPSS
- Trajectory training considering GV for mixture density networks

DNN-based SPSS

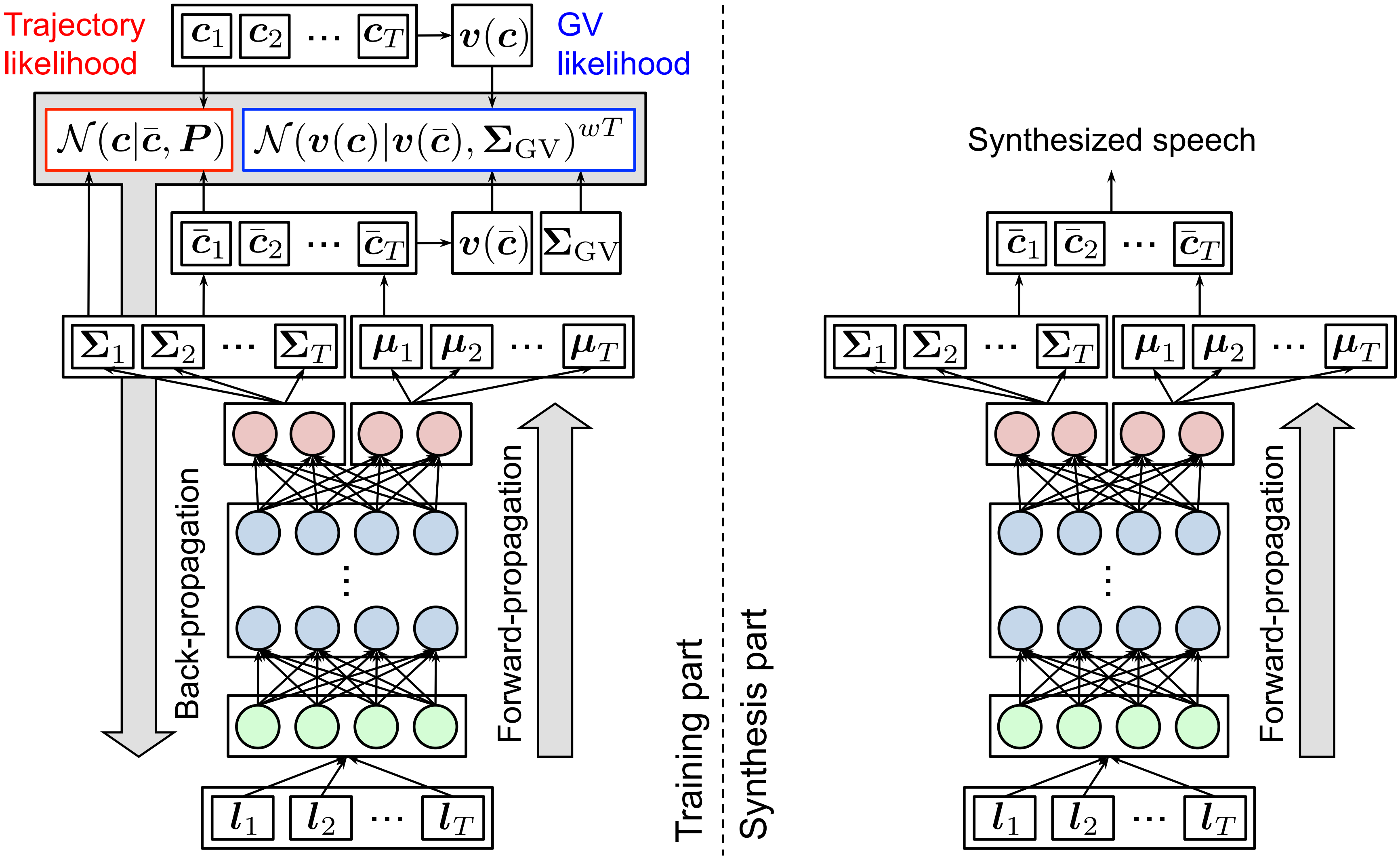
- **DNN-based SPSS [Zen et al. ’13]**
  - DNN is trained to represent a mapping function from linguistic features to acoustic features
  - Mixture density network (MDN)-based SPSS [Zen et al. ’14]
    - *DNN outputs provide Gaussian mixture model parameters*
  - Inconsistency in training and synthesis criteria
  - Over-smoothing on speech parameter trajectories
- **Trajectory training considering GV for DNN-based SPSS [Hashimoto et al. ’16]**
  - Can address inconsistency between training and synthesis
  - DNN is optimized considering GV
- **Trajectory training considering GV for MDN-based SPSS**
  - Expect high-quality acoustic model
  - Use a single MDN as the acoustic model

Frame-level training


$\boldsymbol{o}_t$
Frame-level likelihood
$\mathcal{N}(\boldsymbol{o}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$
$\boldsymbol{\Sigma}_t$
$\boldsymbol{\mu}_t$
Back-propagation
Forward-propagation
$\boldsymbol{l}_t$
Training part
Synthesis part
Synthesized speech
$\bar{\boldsymbol{c}}_1$ $\bar{\boldsymbol{c}}_2$ $\cdots$ $\bar{\boldsymbol{c}}_T$
$\boldsymbol{\Sigma}_1$ $\boldsymbol{\Sigma}_2$ $\cdots$ $\boldsymbol{\Sigma}_T$
$\boldsymbol{\mu}_1$ $\boldsymbol{\mu}_2$ $\cdots$ $\boldsymbol{\mu}_T$
Forward-propagation
$\boldsymbol{l}_1$ $\boldsymbol{l}_2$ $\cdots$ $\boldsymbol{l}_T$

$\boldsymbol{l}_t$	: linguistic feature vector	$\boldsymbol{c}_t$	: static-feature vector	$\boldsymbol{v}(\cdot)$	: GV vector
$\boldsymbol{\mu}_t$	: mean vector	$\bar{\boldsymbol{c}}_t$	: optimal static-feature vector	$\boldsymbol{\Sigma}_{\mathrm{GV}}$	: GV covariance matrix
$\boldsymbol{\Sigma}_t$	: covariance matrix	$\boldsymbol{P}$	: sequence covariance matrix	w	: GV weight

Trajectory training considering GV


Trajectory likelihood
GV likelihood
c_1 c_2 $\cdots$ c_T
$v(c)$
$\mathcal{N}(c|\bar{c}, P)$
$\mathcal{N}(v(c)|v(\bar{c}), \Sigma_{\mathrm{GV}})^{wT}$
$\bar{c}_1$ $\bar{c}_2$ $\cdots$ $\bar{c}_T$
$v(\bar{c})$
Σ_{GV}
Σ_1 Σ_2 $\cdots$ Σ_T
μ_1 μ_2 $\cdots$ μ_T
Back-propagation
Forward-propagation
l_1 l_2 $\cdots$ l_T
Training part
Synthesis part
Synthesized speech
$\bar{c}_1$ $\bar{c}_2$ $\cdots$ $\bar{c}_T$
Σ_1 Σ_2 $\cdots$ Σ_T
μ_1 μ_2 $\cdots$ μ_T
Forward-propagation
l_1 l_2 $\cdots$ l_T

l_t : linguistic feature vector

μ_t : mean vector

Σ_t : covariance matrix

c_t : static-feature vector

$\bar{c}_t$: optimal static-feature vector

P : sequence covariance matrix

$v(\cdot)$: GV vector

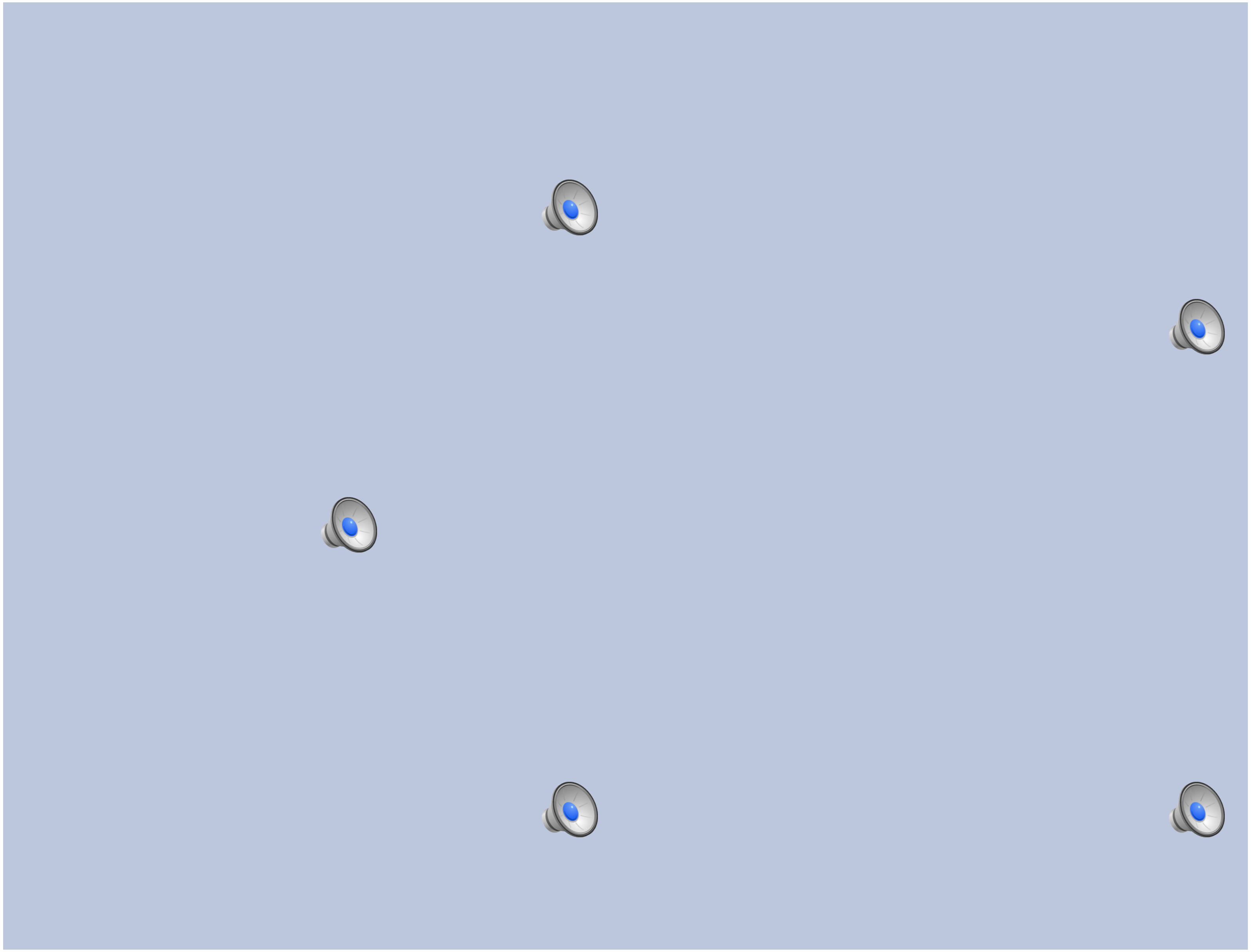
Σ_{GV} : GV covariance matrix

w : GV weight

TTS system conditions

Training corpus	921 pages
Sampling rage	44.1 kHz
Frame	window: F0-adapteve Gaussian, shift: 5 ms
HMM structure	5-state left-to-right MSD-HSMM
Acoustic features (HMM)	49-dim. STRAIGHT mel-cepstrum, 24-dim. mel-cepstrum aperiodicity measure, log F0, and Δ + ΔΔ
Number of questions	925 questions
MDN structure	3 hidden layers with 8000 hidden units, activation function: sigmoid (hidden), linear (output), dropout rate: 60%, GV weight: 0.001
Acoustic features (MDN)	69-dim. STRAIGHT mel-cepstrum, 34-dim. mel-cepstrum aperiodicity measure, interpolated log F0, voiced/unvoiced information
Linguistic features	925-dim. linguistic features for contexts, 10-dim. duration features, 150-dim. word code, 600-dim. phrase code

Synthesized speech samples



The picture is quote from the Usborne Publishing.

Experimental results

Experimental conditions

- 16 TTS system (+ 1 natural speech)
- Results of all participants

Page domain (60-point MOS)

Criterion	Overall impression	Pleasantness	Speech pause	Stress	Intonation	Emotion	Listening effort
MOS	31	30	31	31	31	33	31
Rank	4th	5th	4rd	4rd	4th	4th	3th

Sentence domain (5-point MOS)

Criterion	Naturalness	Similarity
MOS	3.6	3.0
Rank	3th	7th

Intelligibility test

WER	30%
Rank	1st

Highly natural synthesized speech

Experimental results

Experimental conditions

- 16 TTS system (+ 1 natural speech)
- Results of all participants

Page domain (60-point MOS)

Criterion	Overall impression	Pleasant ness	Speech pause	Stress	Intonation	Emotion	Listening effort
MOS	31	30	31	31	31	33	31
Rank	4th	5th	4rd	4rd	4th	4th	3th

Sentence domain (5-point MOS)

Criterion	Naturalness	Similarity
MOS	3.6	3.0
Rank	3th	7th

Intelligibility test

WER	30%
Rank	1st

Compared with MOS of naturalness, MOS of speaker similarity is low score

Experimental results

Experimental conditions

- 16 TTS system (+ 1 natural speech)
- Results of all participants

Page domain (60-point MOS)

Criterion	Overall impression	Pleasant ness	Speech pause	Stress	Intonation	Emotion	Listening effort
MOS	31	30	31	31	31	33	31
Rank	4th	5th	4rd	4rd	4th	4th	3th

Sentence domain (5-point MOS)

Criterion	Naturalness	Similarity
MOS	3.6	3.0
Rank	3th	7th

Intelligibility test

WER	30%
Rank	1st

Highly intelligible synthesized speech

Conclusion

- **NITech TTS system for the Blizzard Challenge 2017**
  - Linguistic features for audiobooks in SPSS
  - Trajectory training considering GV for MDN-based SPSS
  - Large-scale subjective listening tests
    - *Synthesized highly natural and intelligible speech*
    - *Should improve speaker similarity*

- **Future work**
  - Improve robustness of outliers
    - *ε-contaminated Gaussian loss [Zen et al. ’16]*
  - Introduce direct speech waveform prediction models
    - *WaveNet [van den Oord et al. ’16]*