The NITech text-to-speech system for the Blizzard Challenge 2016

Kei Sawada, Chiaki Asai, Kei Hashimoto, Keiichiro Oura, and Keiichi Tokuda

Nagoya Institute of Technology (NITECH)

Blizzard Challenge 2016 workshop on Sep. 16, 2016

Background

- Text-to-speech (TTS) systems
  - TTS systems are used in various applications
    - In-car navigation, smartphones, spoken dialogue systems, etc.
  - Demand for TTS systems is increasing
    - High speech quality, speaking styles, multilingual language, etc.
- TTS system based on big data
  - Quality of synthesized speech is improved by using big data
  - Speech data recorded with less noise and under same conditions are suitable for training
  - Recording a large amount of speech data requires a huge cost
- TTS system based on audiobooks
  - Audiobooks can be relatively easily used as a large amount of speech data and text pairs

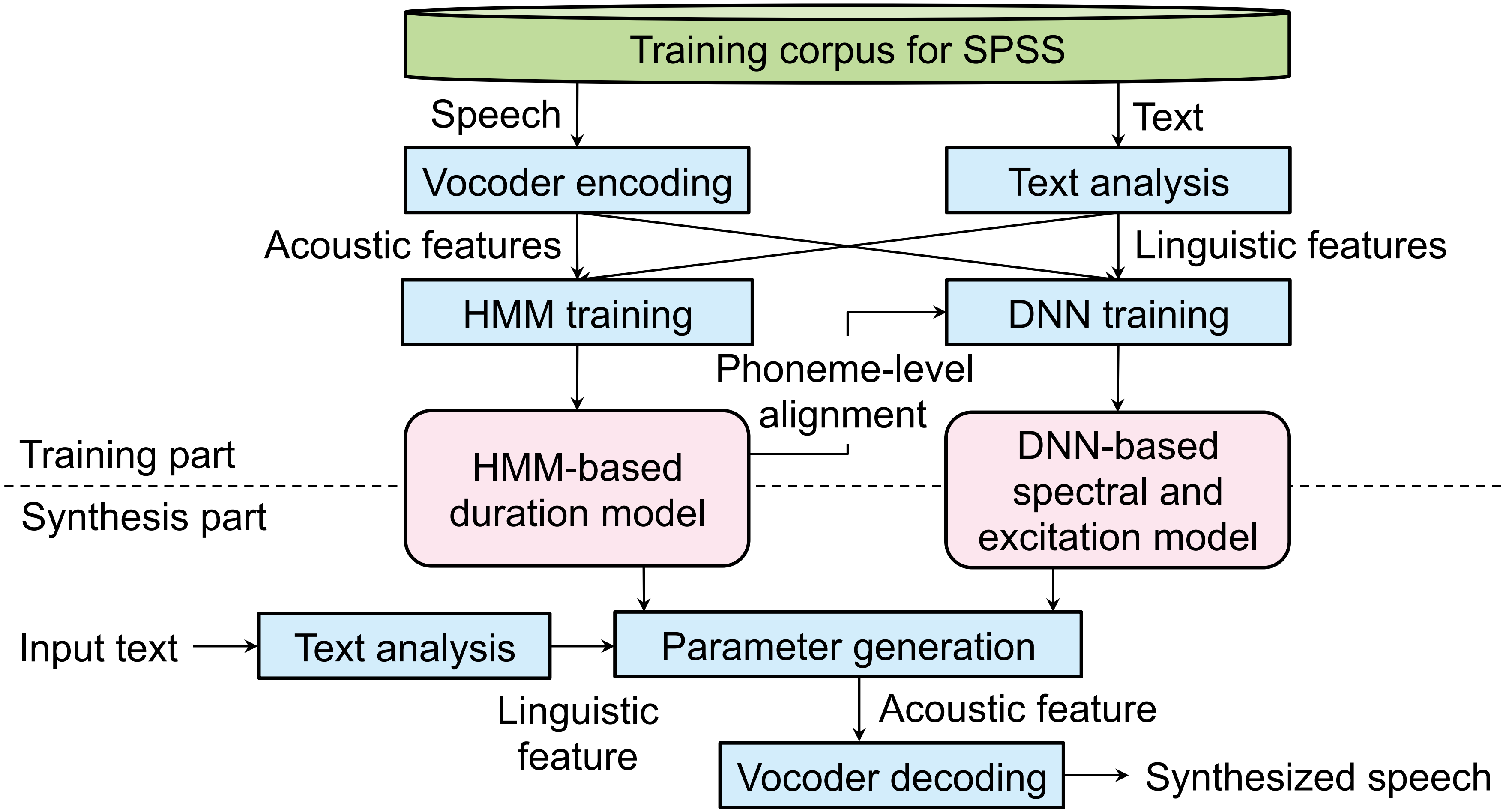
Blizzard Challenge 2016 task

- Blizzard Challenge [Black, *et al.*; ’05]
  - Blizzard Challenge was started in order to better understand and compare research techniques
- Blizzard Challenge 2016
  - Task is to construct a TTS system from children’s audiobooks
  - Five-hours speech data and text pairs are provided
  - All 50 books were recorded by one English female 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.

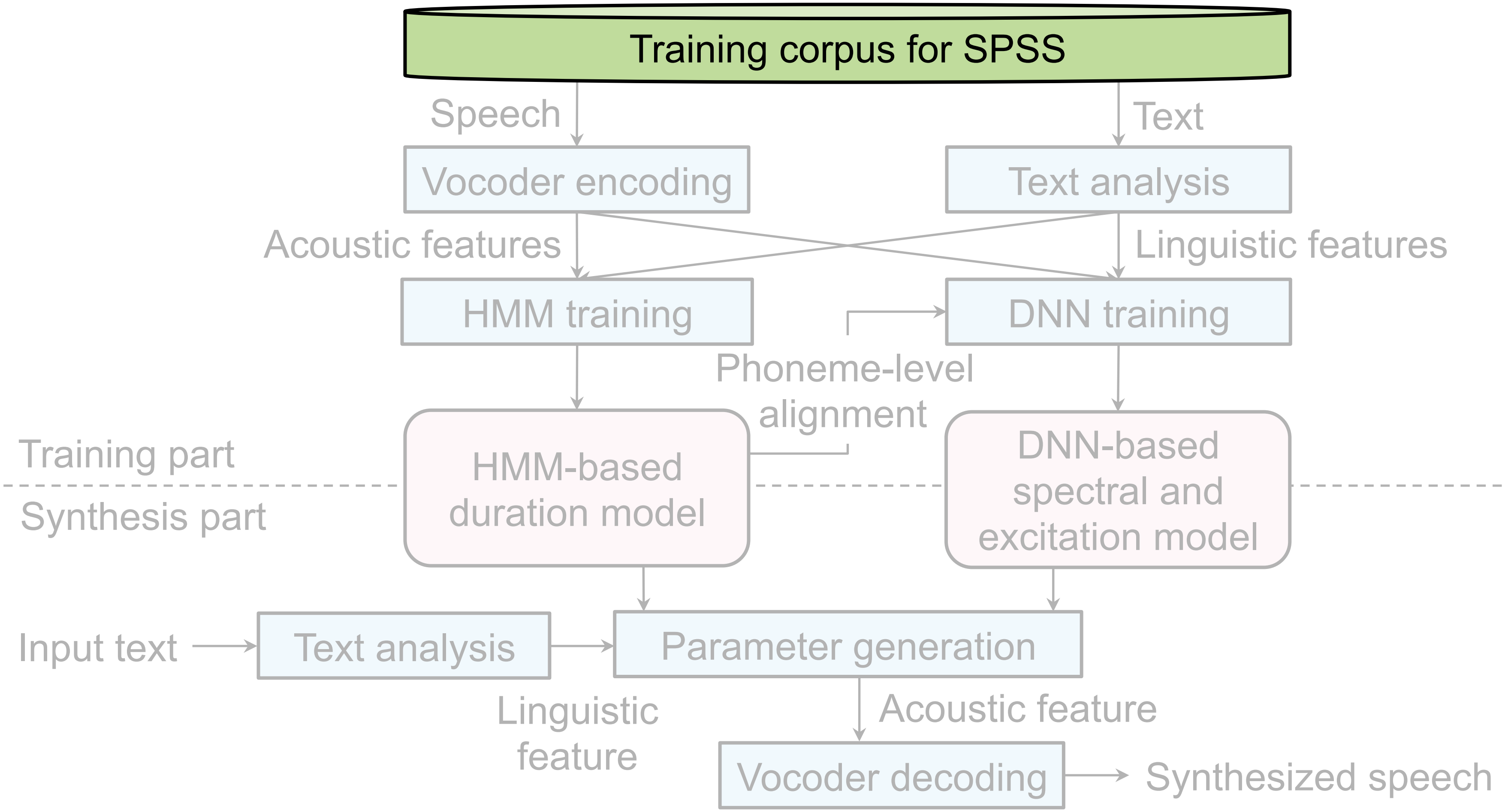
Character1
Character2
Descriptive part

NITech system



- Automatic construction of training corpus from audiobooks
- Design of linguistic features for SPSS based on audiobooks
- DNN-based SPSS

NITech system



- Automatic construction of training corpus from audiobooks
- Design of linguistic features for SPSS based on audiobooks
- DNN-based SPSS

Automatic construction of training corpus

- Mismatches are present in speech data and text
  - Substitution: misreading text
  - Deletion: unrecording text
  - Insertion: recording additional information, i.e., onomatopoeia

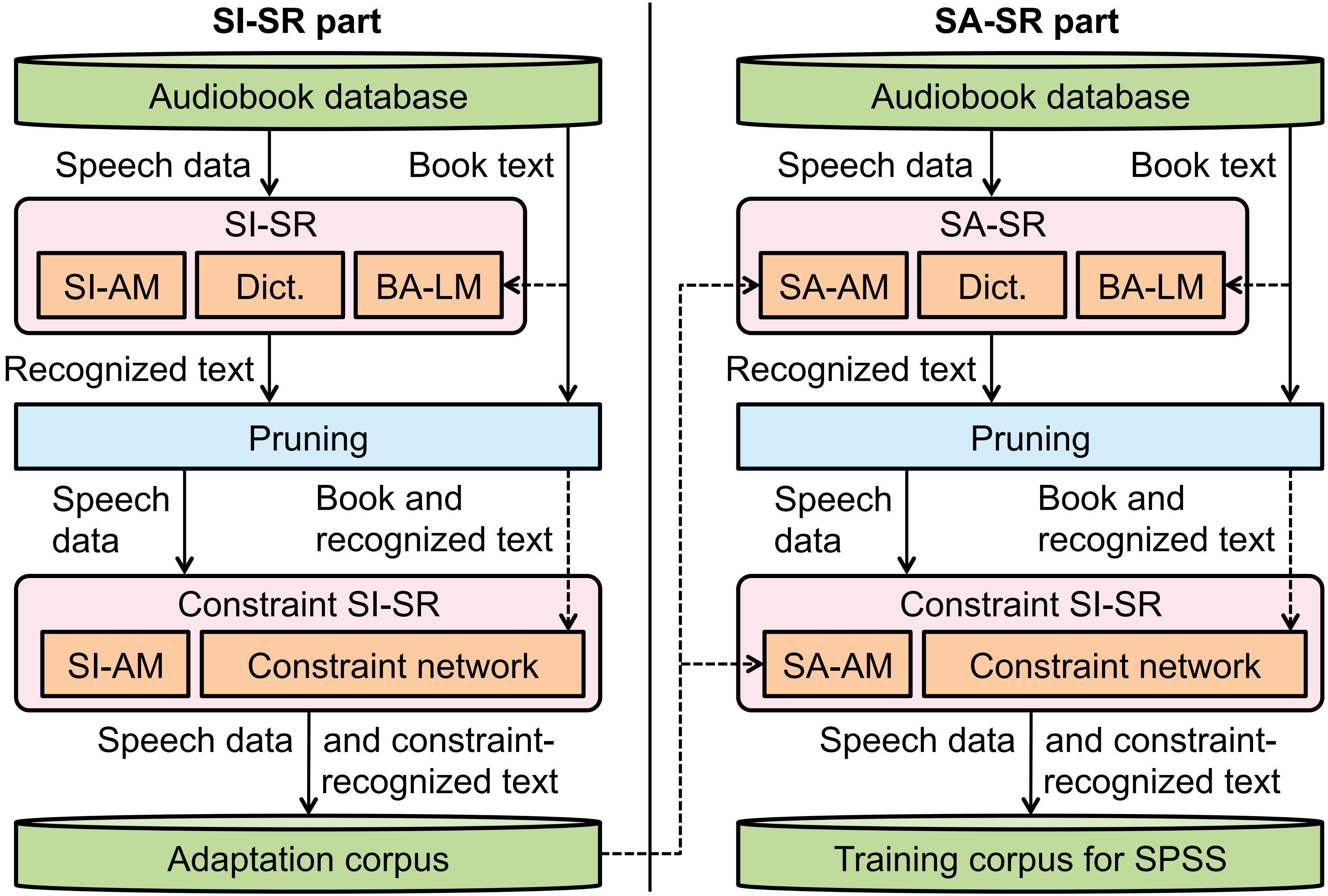
⇒ This will negatively affect an acoustic model of SPSS

- Training corpus construction using speech recognizer
  - Texts are estimated from speech data [Braunschweiler, *et al.*; ’10]
  - Texts may include speech recognition errors

Speech data	🔊
Correct text	he came to a cottage knock knock
Book text	he came to a cottage ________
Recognized text	she came to a cottage knock knock

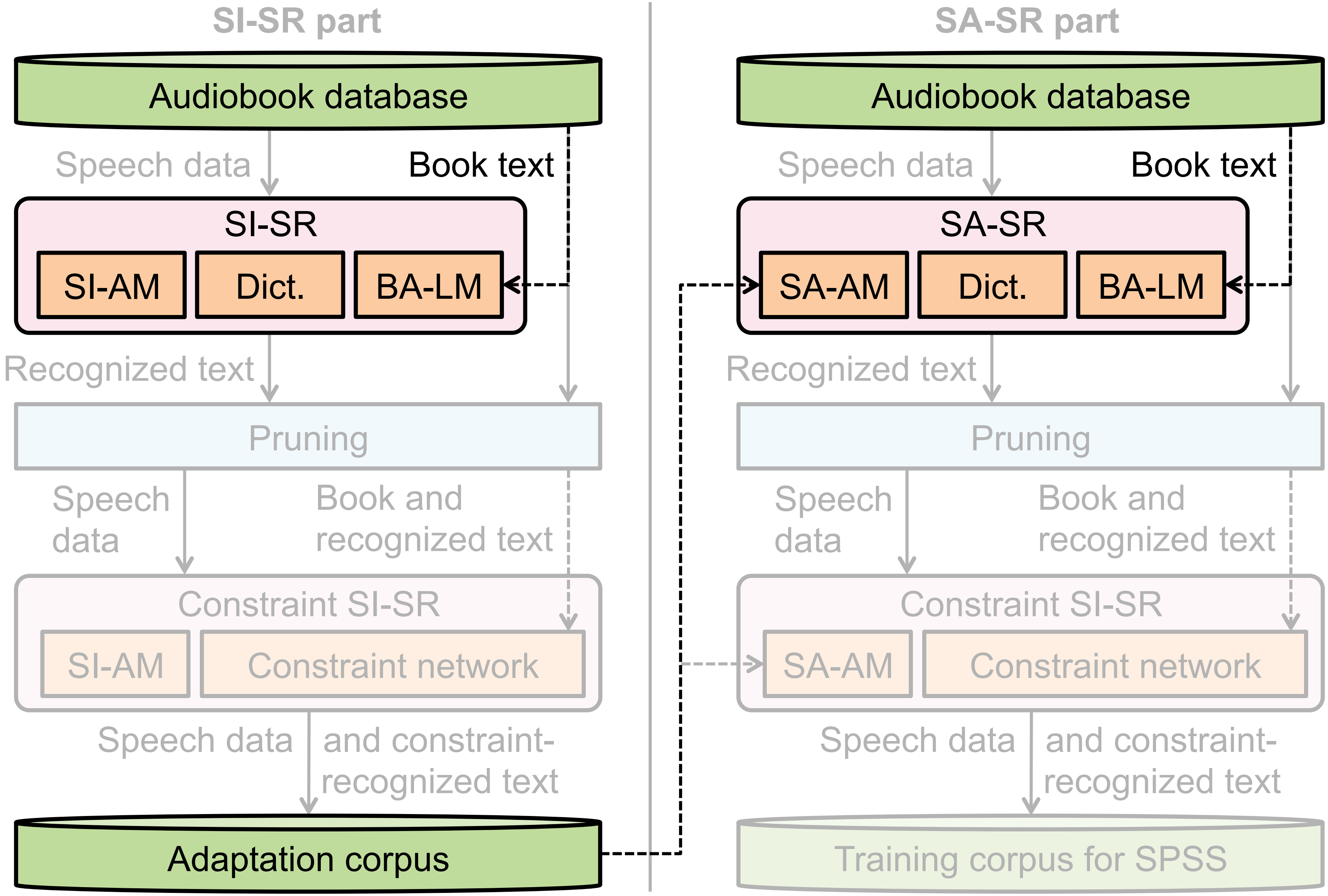
Speech recognition using a constrained word network is conducted

Overview of training corpus construction



SI: speaker-independent, SA: speaker-adapted, BA, book adapted,
SR: speech recognize, AM: acoustic model, LM: language model

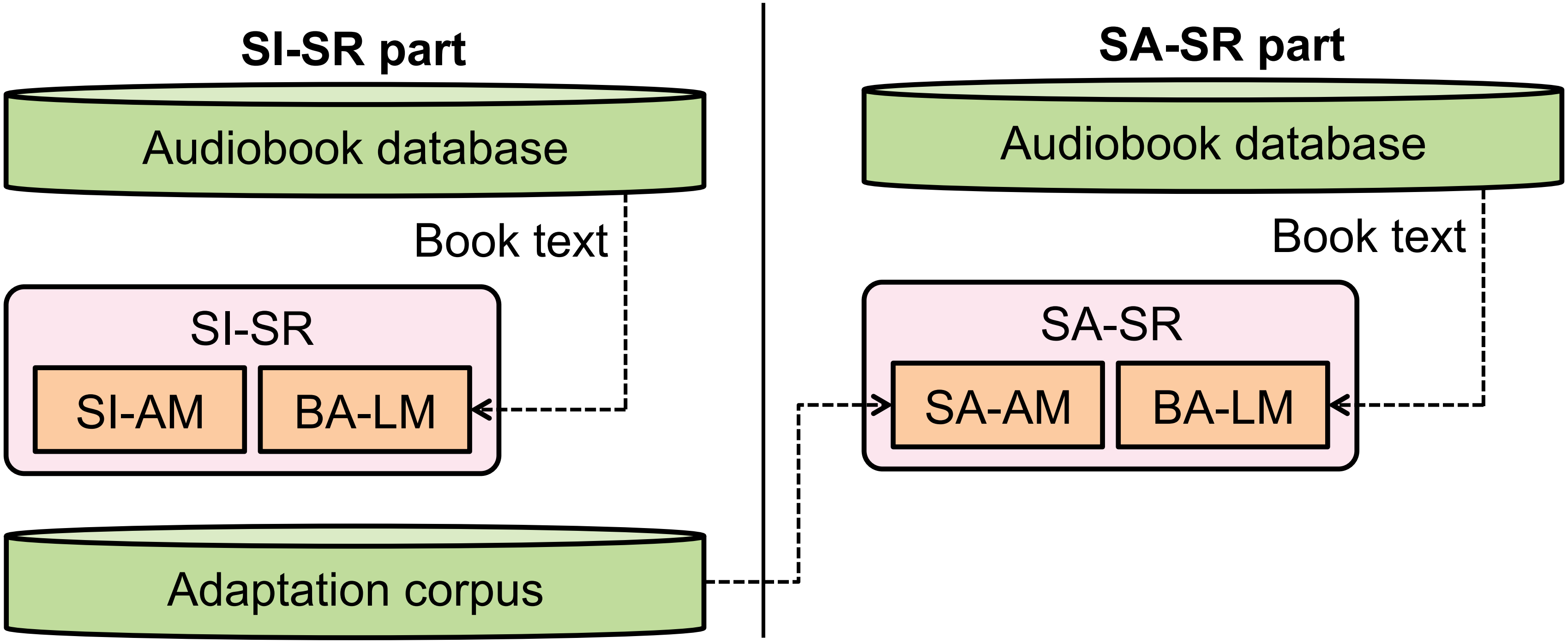
Overview of training corpus construction



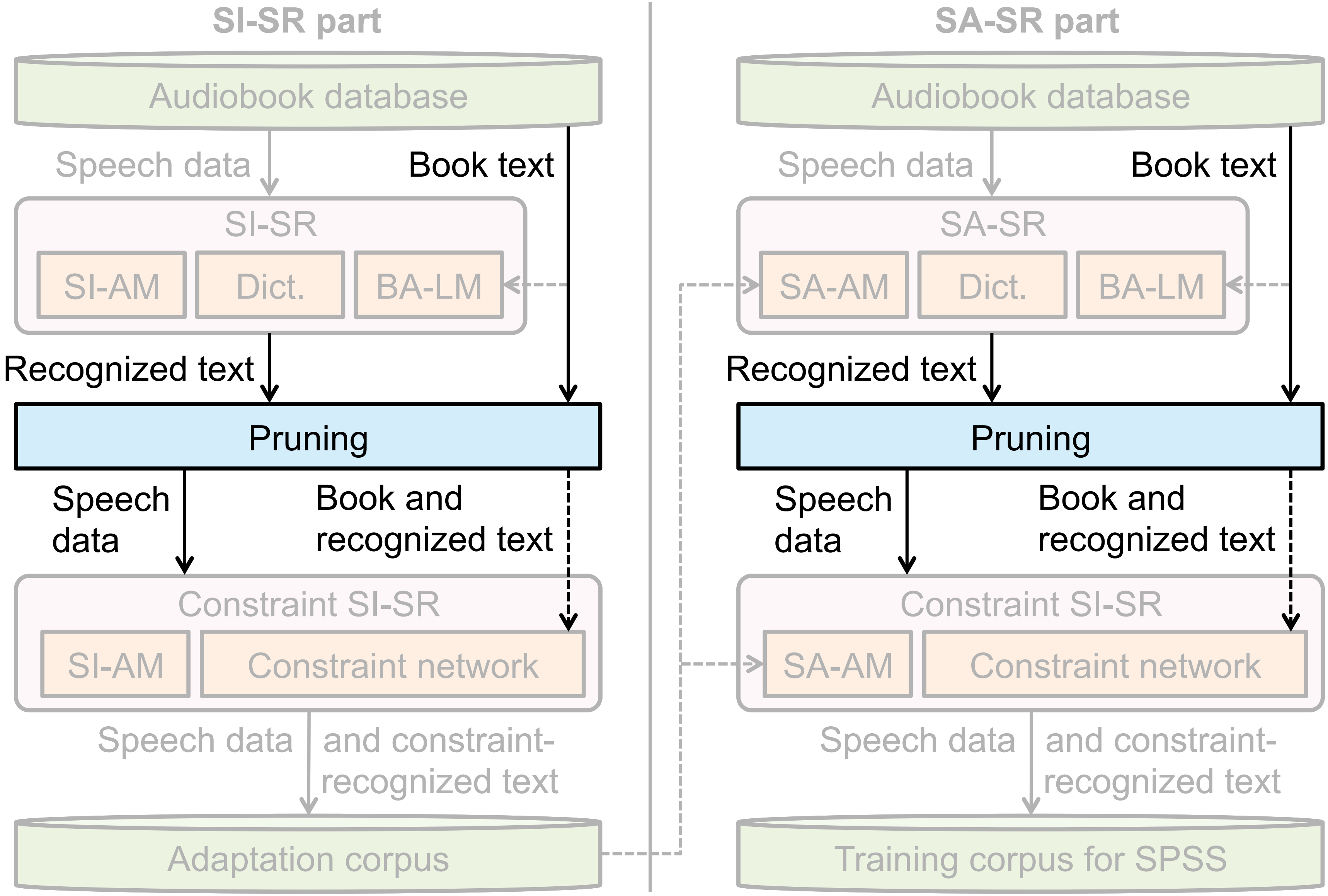
SI: speaker-independent, SA: speaker-adapted, BA, book adapted,
SR: speech recognize, AM: acoustic model, LM: language model

Adapted model

- Language model (LM)
  - Most book texts match speech data
  - LM based on book texts is useful for speech recognition
  - Book-adapted LM is used for speech recognition
- Acoustic model (AM)
  - Speaker-adapted AM is constructed by using SI-SR results



Overview of training corpus construction



SI: speaker-independent, SA: speaker-adapted, BA, book adapted,
SR: speech recognize, AM: acoustic model, LM: language model

Pruning

- Word-match accuracy
  - Concordance rate of book text and recognized text

Book text	he came to a cottage
Recognized text	she came to a cottage knock knock

word-match accuracy
= 57%

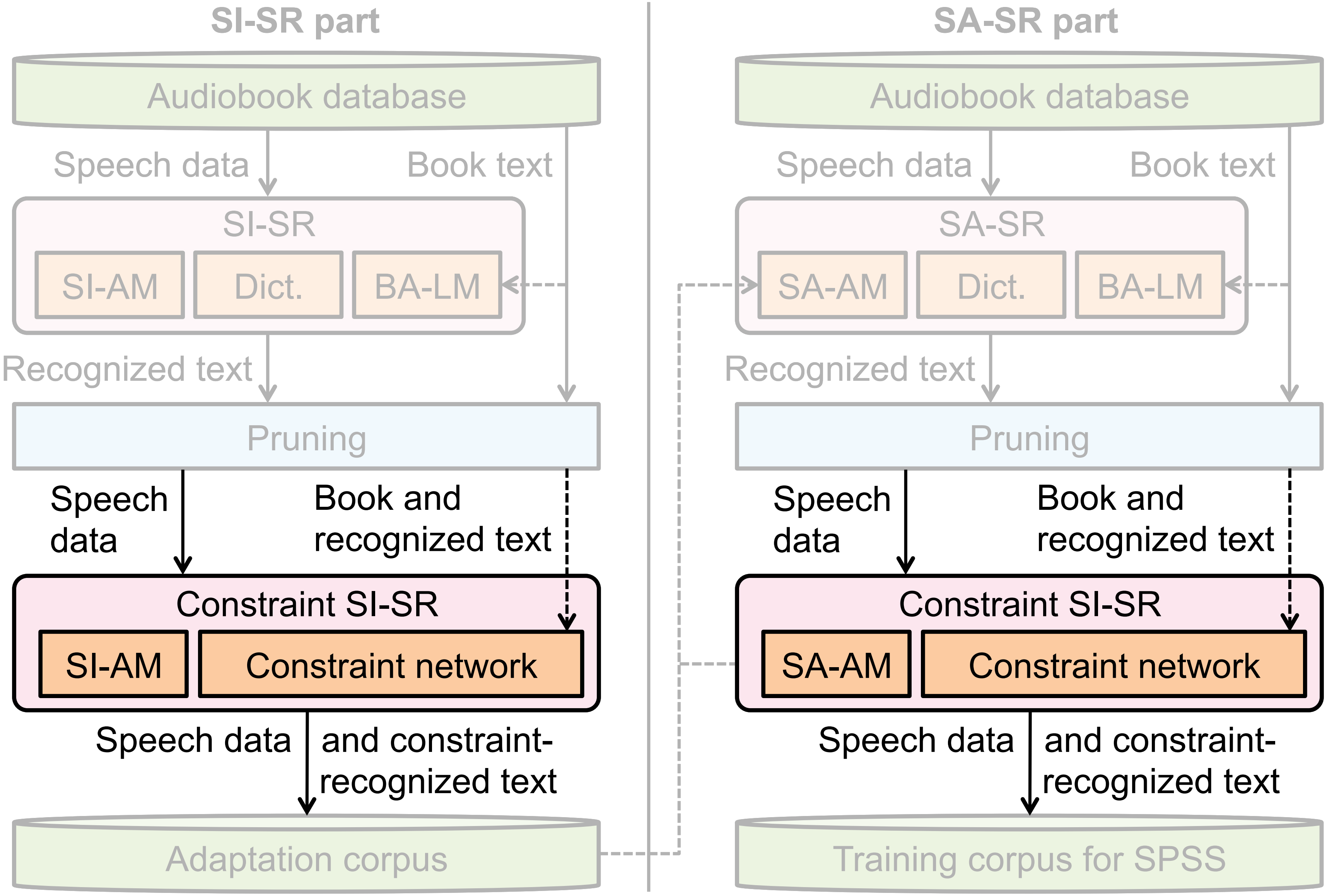
  - Low word-match accuracy

    ⇒ Reliability of text is lower

- Pruning of low word match accuracy
  - If word-match accuracy is not more than threshold, speech data and text pair is pruned from training corpus
  - Relation between quantity and quality of corpus is trade-off

Threshold	Large	Small
Quantity of corpus	Small	Large
Quality of corpus	High	Low

Overview of training corpus construction

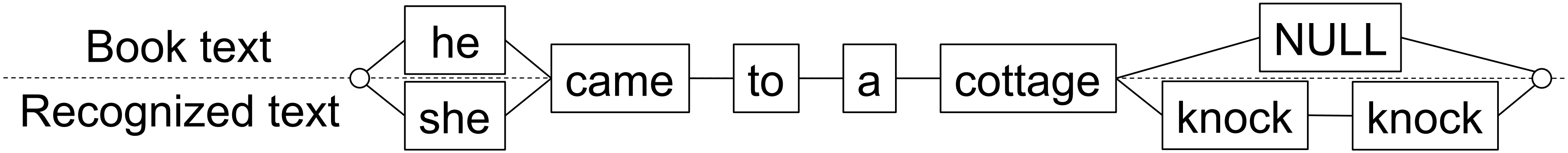


SI: speaker-independent, SA: speaker-adapted, BA, book adapted,
SR: speech recognize, AM: acoustic model, LM: language model

Constraint speech recognition

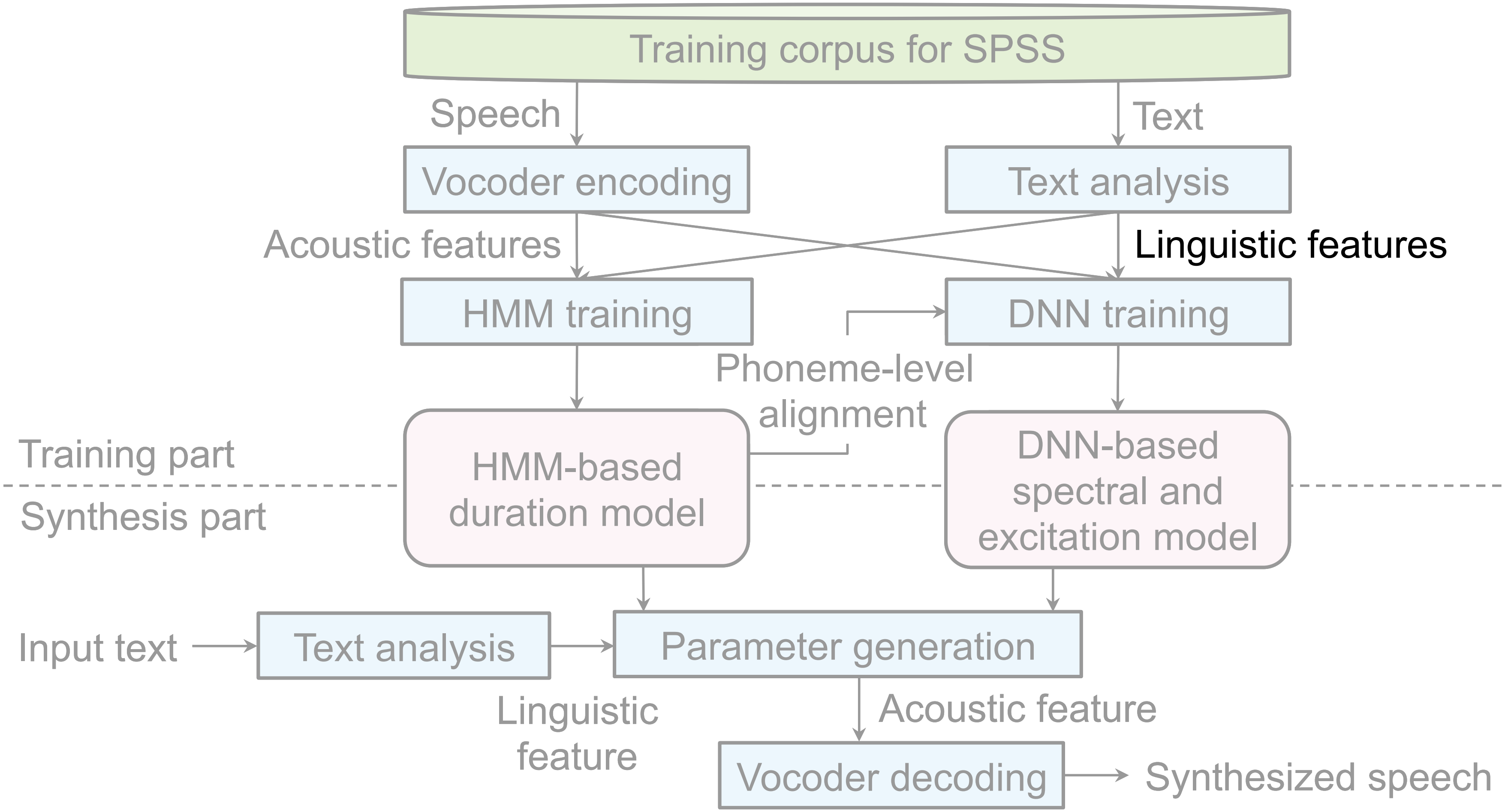
- Constraint-recognized text
  - Speech recognition using constrained word network consisting of book and recognized text
  - Path penalty
    - Book text is NULL: path penalty for book text
    - Otherwise: path penalty for recognized text
  - Speech recognizer with constrained word network without LM

Book text	he came to a cottage
Recognized text	she came to a cottage knock knock



- Contain text corresponding to additional speech information
- Reduce speech recognition errors

NITech system



- Automatic construction of training corpus from audiobooks
- Design of linguistic features for SPSS based on audiobooks
- DNN-based SPSS

Design of linguistic features (1/2)

- Linguistic features
  - Context-dependent model is used to capture contextual factors
  - Appropriate linguistic features design is needed to synthesize high-quality speech
  - Speech in conversational and descriptive parts of audiobook
    - Conversational part: emphatically, emotionally, etc.
    - Descriptive part: comparatively neutrally

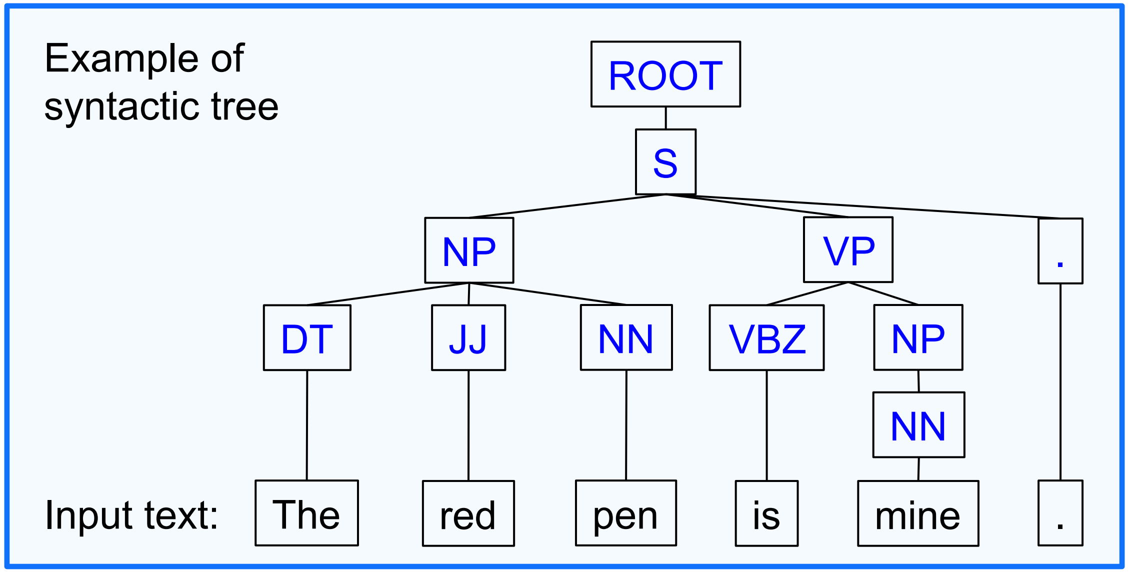
    ⇒ Double quotes are used to express conversational part
  - Prosodic information
    - Intonation, rhythm, etc.

    ⇒ Detailed parsing results are used to express prosodic information

- Linguistic feature using double quotes
  - Example of added linguistic features
    - Whether the current phoneme is enclosed by double quotes
    - The rate of word enclosed by double quotes in this page

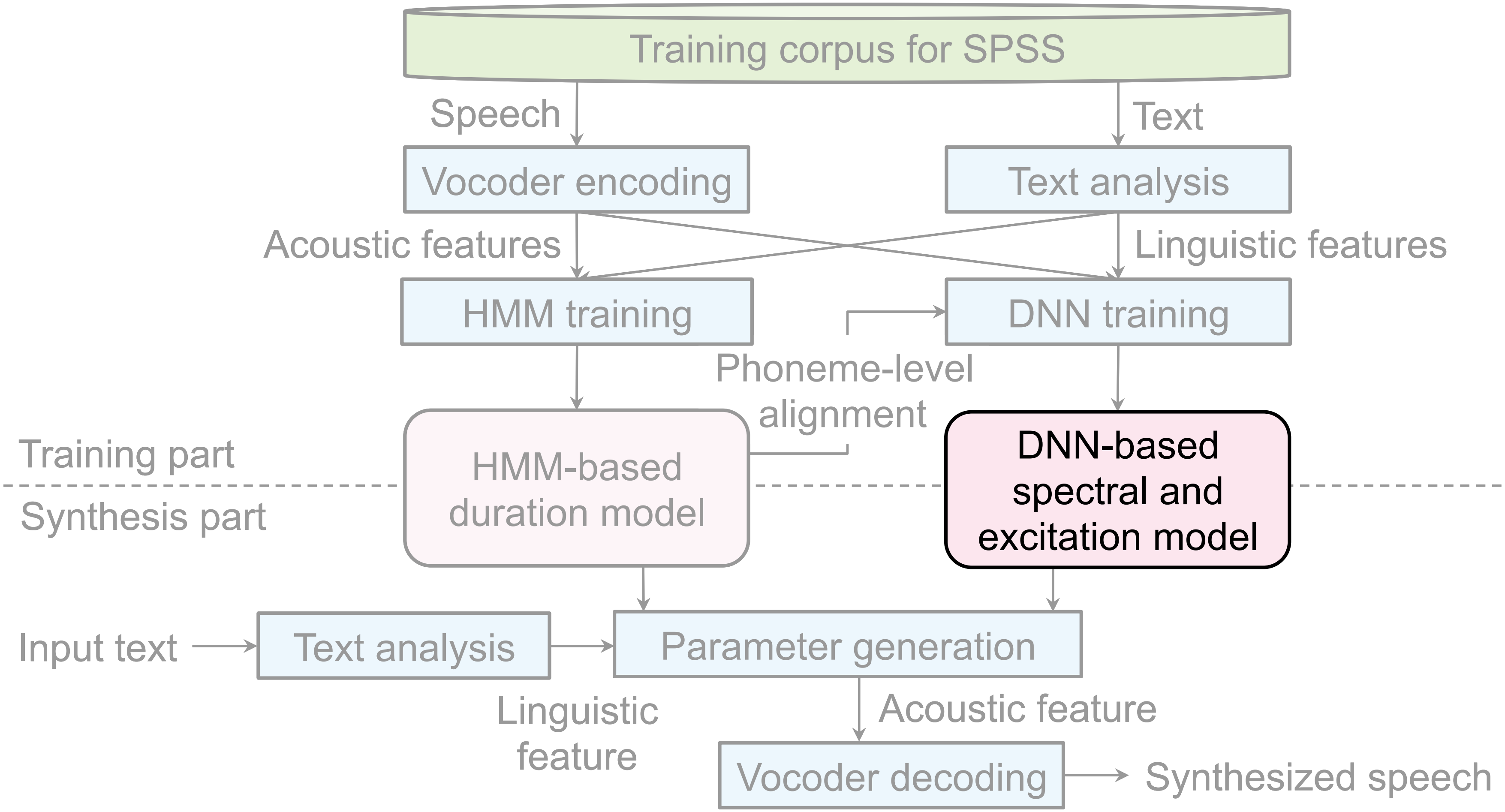
Design of linguistic features (2/2)

- Linguistic feature using detailed parsing results
  - Results of parsing is represented by syntactic tree



  - Example of added linguistic features
    - Guess part-of-speech of the parent of the current word
    - Distance on the syntactic tree between the current word and the previous word
    - Position of the current word in the parent of the current word
    - The number of phonemes in the parent of the current word

NITech system



- Automatic construction of training corpus from audiobooks
- Design of linguistic features for SPSS based on audiobooks
- DNN-based SPSS

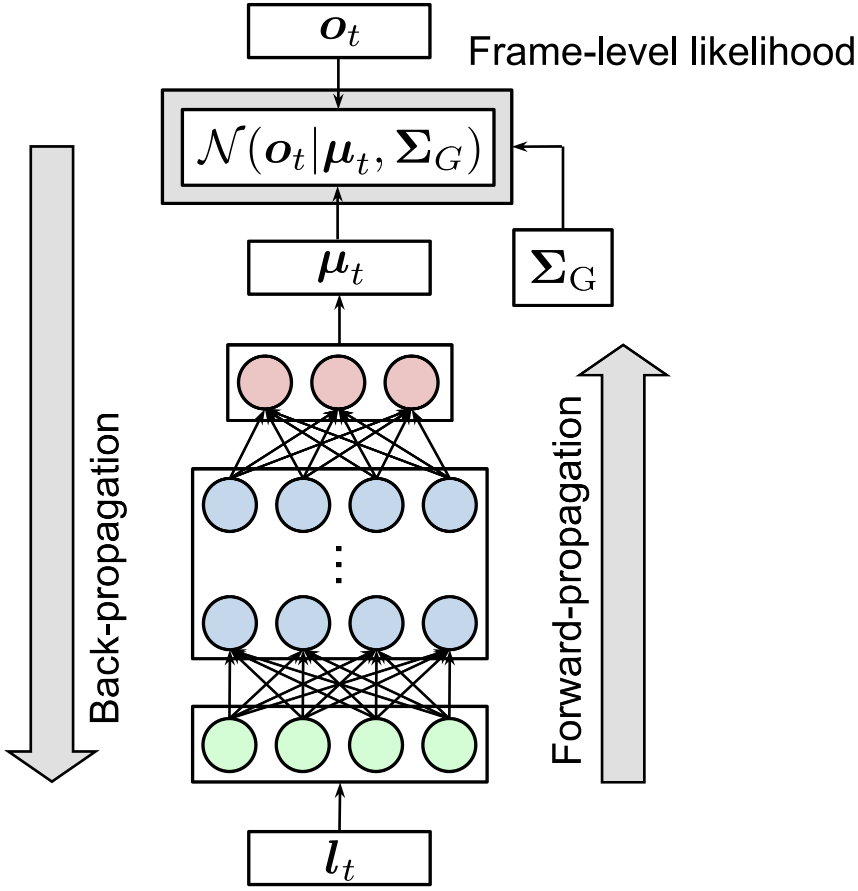
DNN-based SPSS

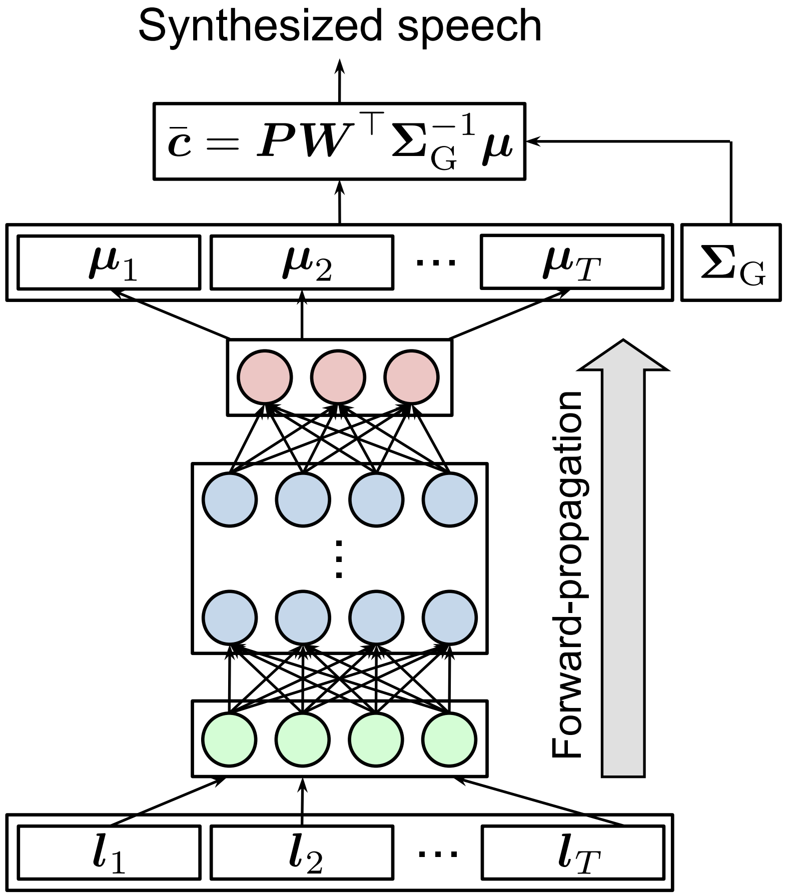
- DNN-based SPSS [Zen, *et al.*; ’12]
  - DNN is trained to represent a mapping function from linguistic features to acoustic features
  - DNN-based SPSS improves naturalness of synthesized speech
  - Inconsistency in training and synthesis criterion
  - Over-smoothing on speech parameter trajectories

Trajectory training considering global variance

DNN-based SPSS (1/3)

- Frame-level training





$\boldsymbol{l}$: linguistic feature vector

$\boldsymbol{\mu}$: mean vector

$\boldsymbol{\Sigma}_G$: globally tied covariance matrix

$\boldsymbol{o}$: speech parameter vector

$\boldsymbol{c}$: static-feature vector

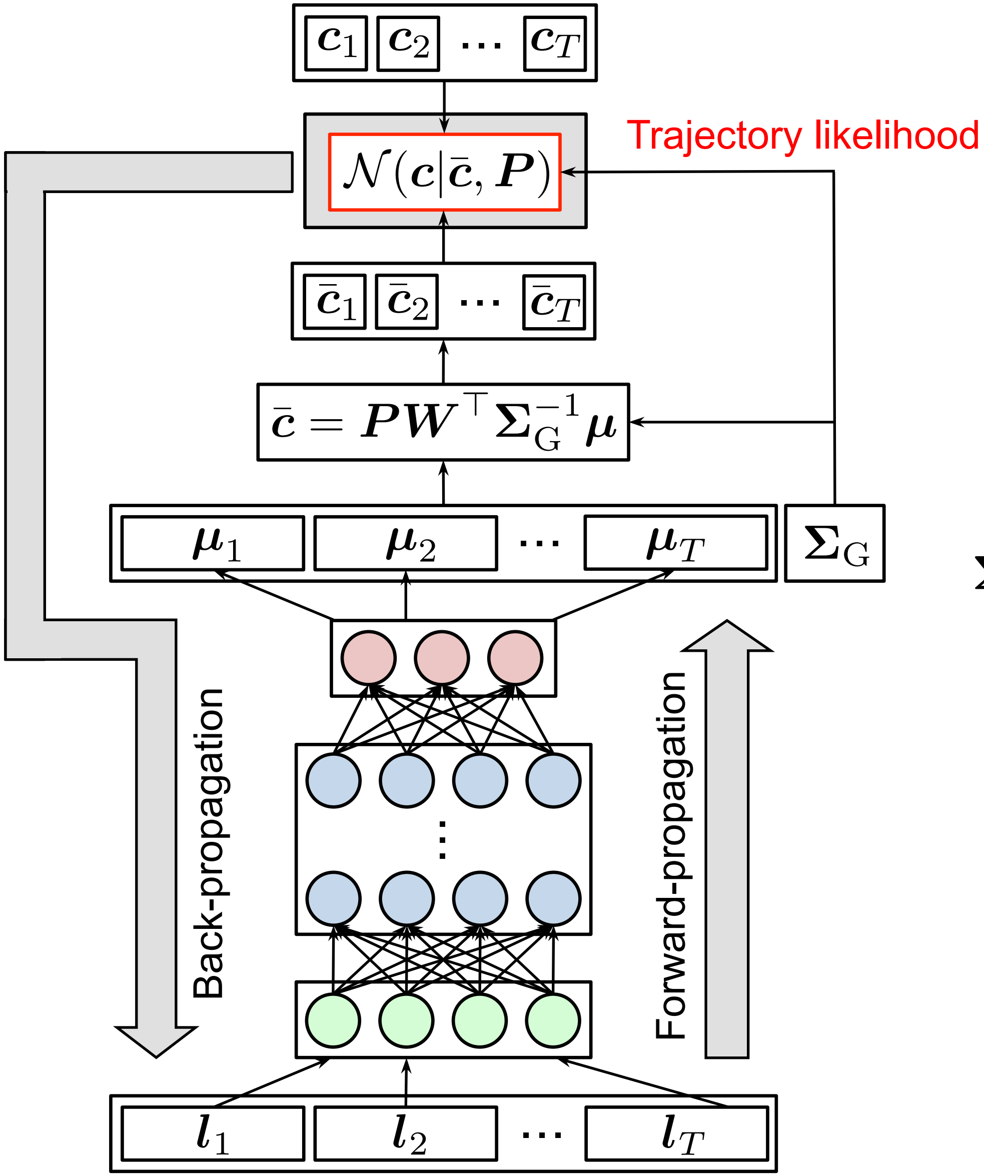
$\bar{\boldsymbol{c}}$: optimal static-feature vector

$$\boldsymbol{P} = \left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_\mathrm{G}^{-1}\boldsymbol{W}\right)^{-1}$$

$\boldsymbol{W}$: window matrix

DNN-based SPSS (2/3)

- Trajectory training [Hashimoto, *et al.*; ’16]



c_1 c_2 … c_T
$\mathcal{N}(c|\bar{c}, P)$
Trajectory likelihood
$\bar{c}_1$ $\bar{c}_2$ … $\bar{c}_T$
$\bar{c} = PW^\top \Sigma_G^{-1} \mu$
μ_1 μ_2 … μ_T
Σ_G
Back-propagation
Forward-propagation
l_1 l_2 … l_T

$\boldsymbol{l}$: linguistic feature vector

$\boldsymbol{\mu}$: mean vector

$\boldsymbol{\Sigma}_G$: globally tied covariance matrix

$\boldsymbol{c}$: static-feature vector

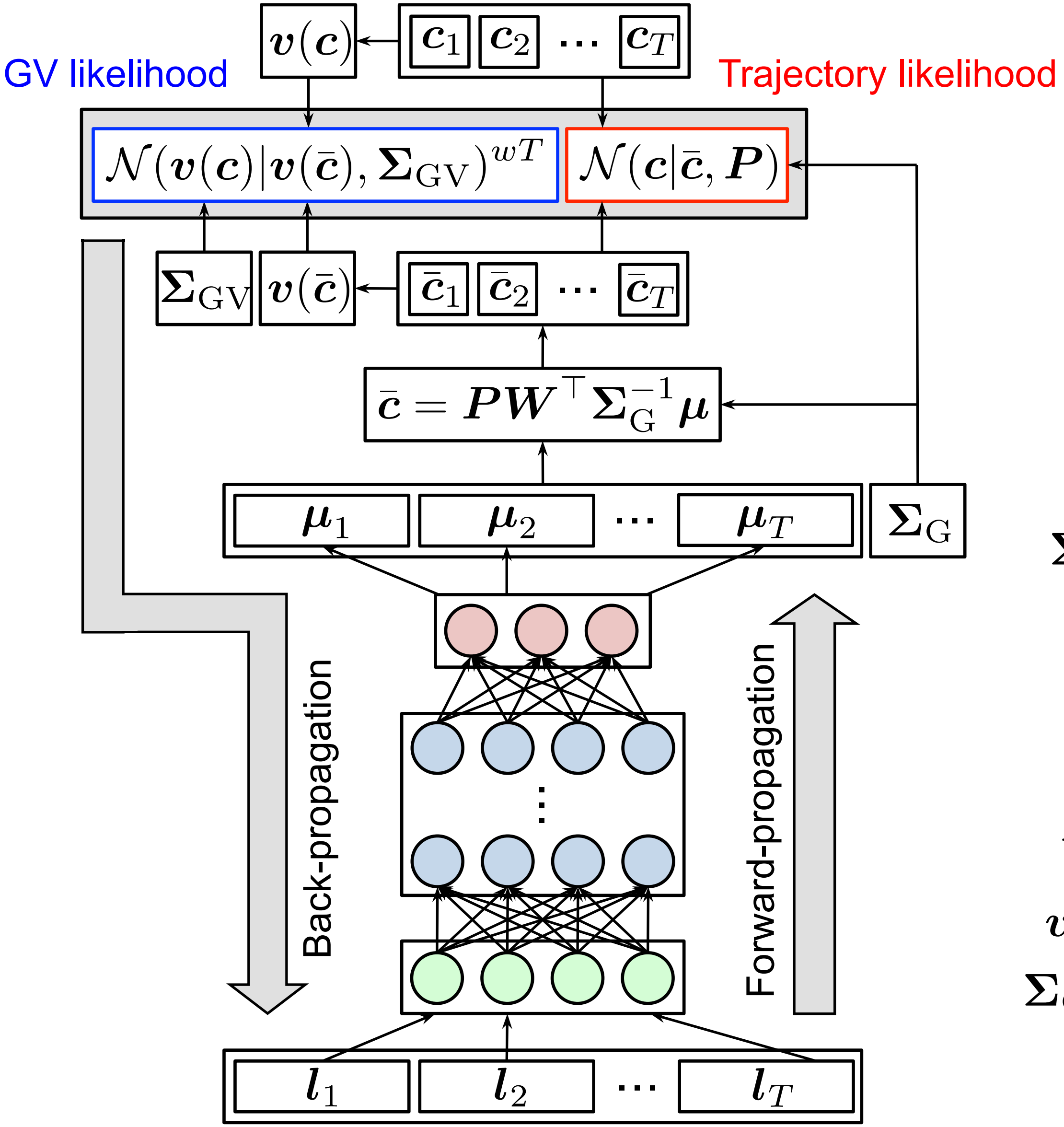
$\bar{\boldsymbol{c}}$: optimal static-feature vector

$$\boldsymbol{P} = \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_G^{-1} \boldsymbol{W}\right)^{-1}$$

$\boldsymbol{W}$: window matrix

DNN-based SPSS (3/3)

- Trajectory training considering GV [Hashimoto, *et al.*; '16]



GV likelihood
Trajectory likelihood
$\boldsymbol{v}(\boldsymbol{c})$
$\boldsymbol{c}_1$ $\boldsymbol{c}_2$ $\cdots$ $\boldsymbol{c}_T$
$\mathcal{N}(\boldsymbol{v}(\boldsymbol{c})|\boldsymbol{v}(\bar{\boldsymbol{c}}), \boldsymbol{\Sigma}_{\mathrm{GV}})^{wT}$
$\mathcal{N}(\boldsymbol{c}|\bar{\boldsymbol{c}}, \boldsymbol{P})$
$\boldsymbol{\Sigma}_{\mathrm{GV}}$
$\boldsymbol{v}(\bar{\boldsymbol{c}})$
$\bar{\boldsymbol{c}}_1$ $\bar{\boldsymbol{c}}_2$ $\cdots$ $\bar{\boldsymbol{c}}_T$
$\bar{\boldsymbol{c}} = \boldsymbol{P}\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\mathrm{G}}^{-1}\boldsymbol{\mu}$
$\boldsymbol{\mu}_1$ $\boldsymbol{\mu}_2$ $\cdots$ $\boldsymbol{\mu}_T$
$\boldsymbol{\Sigma}_{\mathrm{G}}$
Back-propagation
Forward-propagation
$\boldsymbol{l}_1$ $\boldsymbol{l}_2$ $\cdots$ $\boldsymbol{l}_T$

$\boldsymbol{l}$: linguistic feature vector

$\boldsymbol{\mu}$: mean vector

$\boldsymbol{\Sigma}_G$: globally tied covariance matrix

$\boldsymbol{c}$: static-feature vector

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

$$\boldsymbol{P} = \left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\mathrm{G}}^{-1}\boldsymbol{W}\right)^{-1}$$

$\boldsymbol{W}$: window matrix

$\boldsymbol{v}(\cdot)$: GV vector

$\boldsymbol{\Sigma}_{\mathrm{GV}}$: GV covariance matrix

Training corpus construction conditions

Children’s audiobook	50 books, 1090 pages
SR training corpus	WSJ0, WSJ1, TIMIT
Sampling rage	16 kHz
Frame	window: Hamming, length: 25 ms, shift: 10 ms
Acoustic-feature	12-dimensional MFCC + Δ + $\Delta\Delta$
Acoustic model	3-state left-to-right GMM-HMM
Language model	tri-gram
Pruning threshold	word-match accuracy: 90%

TTS system conditions

Training corpus	825 pages (constraint-recognition text)
Sampling rage	44.1 kHz
Frame	window: F0-adapteve Gaussian, shift: 5 ms
Acoustic feature	229-dimensional acoustic features (49-dimensional STRAIGHT mel-cepstrum, 24-dimensional aperiodicity measure + Δ + ΔΔ, log F0, voiced/unvoiced features)
Linguistic feature	426-dimensional linguistic features (423-dimensional binary and numerical features, three duration features)
HMM structure	5-sate left-to-right MSD-HSMM
DNN structure	3 hidden layers with 2048 hidden units, activation function: sigmoid, dropout: 50%, GV weight: 0.001

Experimental conditions of listening test

Participant	paid participants (104 native speakers)
Page domain	7 criteria, 60-point MOS
Sentence domain	2 criteria, 5-point MOS
Intelligibility test	semantically unpredictable sentence (SUS), word error rate (WER)
System	17 systems (1 natural speech, 16 TTS systems)

Speech samples

- Automatic construction of training corpus from audiobooks

Text \ Threshold	80%	90%	100%
Book text	🔊 🔊	🔊 🔊	🔊 🔊
Recognized text	🔊 🔊	🔊 🔊	🔊 🔊
Constraint-recognized text	🔊 🔊	🔊 🔊	🔊 🔊

- Design of linguistic features for SPSS based on audiobooks

Base	DQ	Parser	DQ + Parser
🔊 🔊	🔊 🔊	🔊 🔊	🔊 🔊

- DNN-based SPSS

HMM	DNN	Trajectory DNN	Trajectory GV DNN
🔊 🔊	🔊 🔊	🔊 🔊	🔊 🔊

Experimental results

- Page domain (60-point MOS)

Criterion	Overall impression	Pleasant ness	Speech pause	Stress	Intonation	Emotion	Listening effort
MOS	24	23	30	29	27	27	27
Rank	6th	7th	3rd	3rd	4th	5th	4th

- Sentence domain (5-point MOS)

Criterion	Naturalness	Similarity
MOS	3.0	2.6
Rank	5th	6th

- Intelligibility test

WER	12%
Rank	1st

Experimental results

Page domain (60-point MOS)

Criterion	Overall impression	Pleasant ness	Speech pause	Stress	Intonation	Emotion	Listening effort
MOS	24	23	30	29	27	27	27
Rank	6th	7th	3rd	3rd	4th	5th	4th

Sentence domain (5-point MOS)

Criterion	Naturalness	Similarity
MOS	3.0	2.6
Rank	5th	6th

Intelligibility test

WER	12%
Rank	1st

Page-level training and synthesis
⇒ High MOS of speech pause

Experimental results

Page domain (60-point MOS)

Criterion	Overall impression	Pleasant ness	Speech pause	Stress	Intonation	Emotion	Listening effort
MOS	24	23	30	29	27	27	27
Rank	6th	7th	3rd	3rd	4th	5th	4th

Sentence domain (5-point MOS)

Criterion	Naturalness	Similarity
MOS	3.0	2.6
Rank	5th	6th

Intelligibility test

WER	12%
Rank	1st

Linguistic features of parsing
and trajectory training
⇒ High MOS of stress and intonation

Experimental results

- Page domain (60-point MOS)

Criterion	Overall impression	Pleasantness	Speech pause	Stress	Intonation	Emotion	Listening effort
MOS	24	23	30	29	27	27	27
Rank	6th	7th	3rd	3rd	4th	5th	4th

- Sentence domain (5-point MOS)

Criterion	Naturalness	Similarity
MOS	3.0	2.6
Rank	5th	6th

- Intelligibility test

WER	12%
Rank	1st

Training corpus include various speaking style, emotion, character, etc.
⇒ Modeling is difficult
⇒ Low MOS of similarity

Experimental results

- Page domain (60-point MOS)

Criterion	Overall impression	Pleasant ness	Speech pause	Stress	Intonation	Emotion	Listening effort
MOS	24	23	30	29	27	27	27
Rank	6th	7th	3rd	3rd	4th	5th	4th

- Sentence domain (5-point MOS)

Criterion	Naturalness	Similarity
MOS	3.0	2.6
Rank	5th	6th

- Intelligibility test

WER	12%
Rank	1st

Linguistic features of double quotes
⇒ Can distinguish descriptive part
⇒ Intelligible synthesized speech

Conclusion

- TTS system developed for the Blizzard Challenge 2016
  - NITech team focused on:
    - Automatic construction of training corpus from audiobooks
    - Design of linguistic features for SPSS based on audiobooks
    - DNN-based SPSS
  - Large-scale subjective listening tests
    - Synthesized high natural and highest intelligible speech
    - Should improve speaker similarity
- Future work
  - Improving linguistic features
    - Adding linguistic features of book, page, sentence, etc. codes
  - Improving robustness of outliers
    - ε-contaminated Gaussian loss [Zen, *et al.*; ’16]

Page-level training and synthesis

- Explicit page-tuning sounds
  - Page-tuning sounds are not suited for training AM
  - GMM is trained to detect page-tuning sounds
  - Speech data are divided into page-by-page speech data
  - Page-level decoding, training, and synthesis are conducted

Design of linguistic features

- Linguistic feature using page information
  - The number of {phrases, sentences} in this page
  - position of the current sentence in this page
- Linguistic feature using double quotes
  - Whether the {previous, current, next} {phoneme, syllable, word, phrase} is enclosed by double quotes
  - The rate of {word, phrase} enclosed by double quotes in this page
- Linguistic feature using detailed parsing results
  - Guess part-of-speech of the parent of the current word
  - Distance on the syntactic tree between the current word and {the {previous, next} word, root of the syntactic tree, the {previous, next} content word}
  - Position of the current word in the parent of the current word
  - The number of {phonemes, syllables, words} in the parent of the current word

HTS benchmark system

- Text: INNOETICS + NII (shared)
- Speech: page-level speech data (shared)
- Differences from HTS STRAIGHT demo scripts
  - Page-level linguistic features (shared)
  - F0 extractor: RAPT, SWIPE’, PEAPER voting method
  - Flat start using DAEM algorithm without phoneme alignment
  - GV weight: 1.0 → 0.0001