Constructing a Deep Neural Network based Spectral Model for Statistical Speech Synthesis

Shinji Takaki, Junichi Yamagishi
National Institute of Informatics

Background (1/2)

Deep Neural Networks (DNNs)

- Many hidden layers, Non-linear transformation
- Great performance in speech research
  - Pre-training [Hinton; 07]
  - Noise robust [Maas et al.;12], Reverberant robust[Feng et al.; 14]
  - Speech enhancement [Lu et al.; 13],
  - Spectral binary coding [Deng et al.; 10]

Statistical parametric speech synthesis (SPSS)

- HMM-based speech synthesis [Tokuda et al.; 00]
- DNN-based speech synthesis [Zen et al.; 12]

Study on DNNs for other tasks in SPSS

Background (2/2)

Spectral modeling in SPSS

- HMM/DNN models low-dimensional spectral parameter
  - Mel-cepstrum, LSP
- Quality loss due to using lower spectral parameter

Feature extraction based on a deep auto-encoder

- Extracting efficient spectral parameter for SPSS in a non-linear, data-driven, unsupervised way

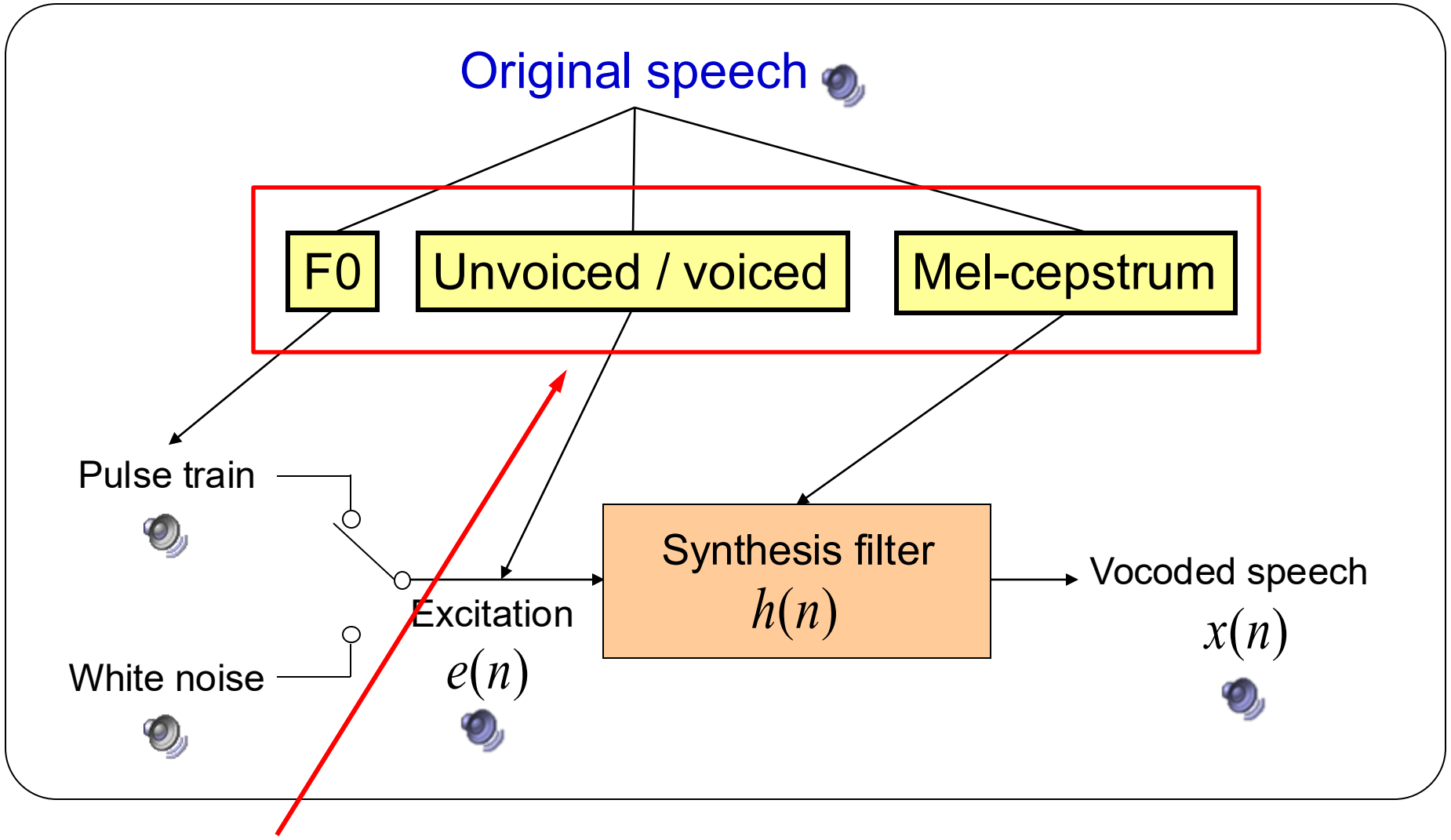
DNN-based direct spectral modeling

- Directly synthesize spectral amplitudes from linguistic features
- Function-wise pre-training using a deep auto-encoder and a DNN acoustic model

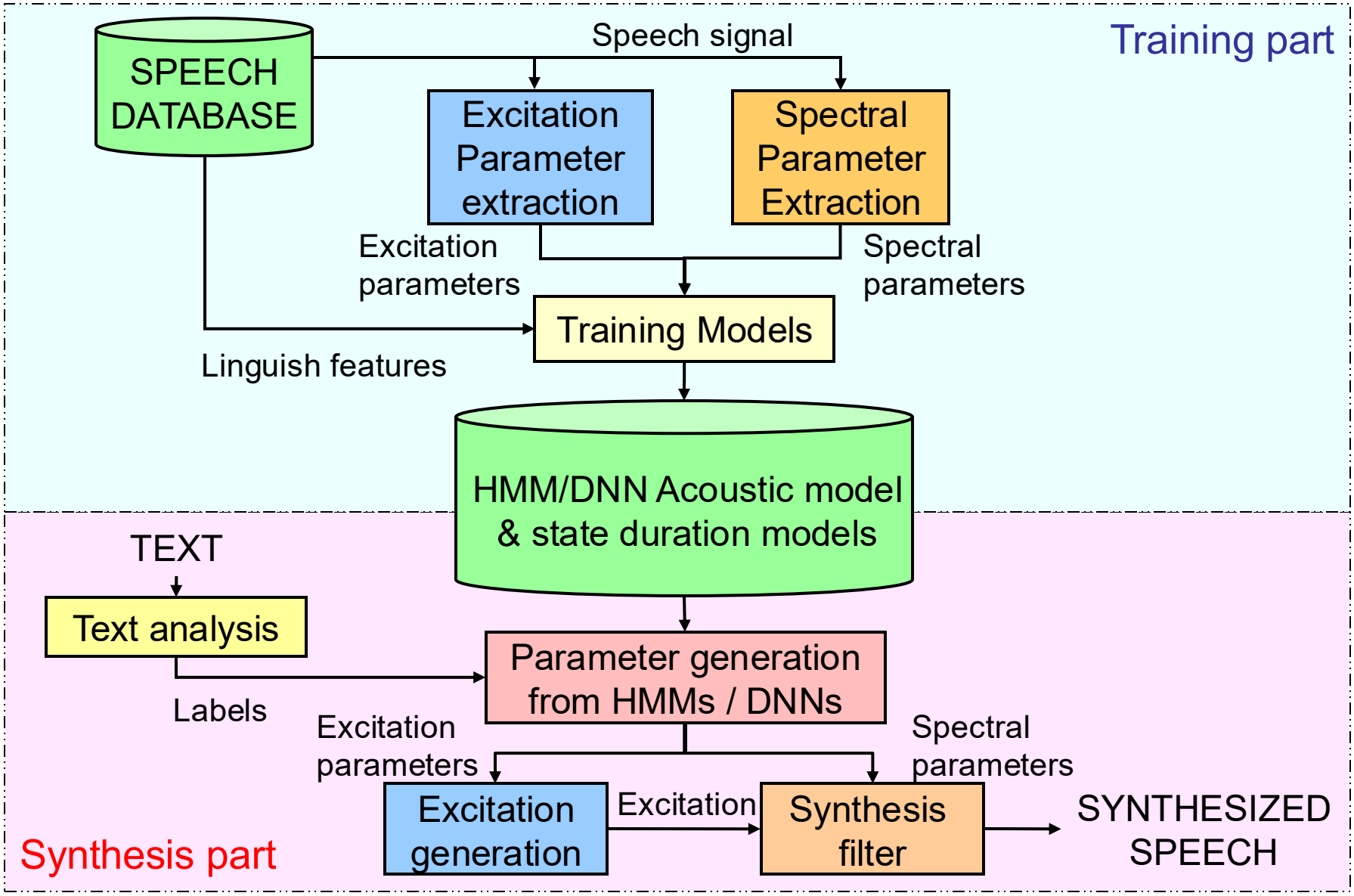
Outline

- Overview of DNN-based speech synthesis
- Spectral feature extraction based on deep auto-encoder
- DNN-based direct spectral modeling
- Text-to-speech synthesis experiment

Overview of speech vocoding

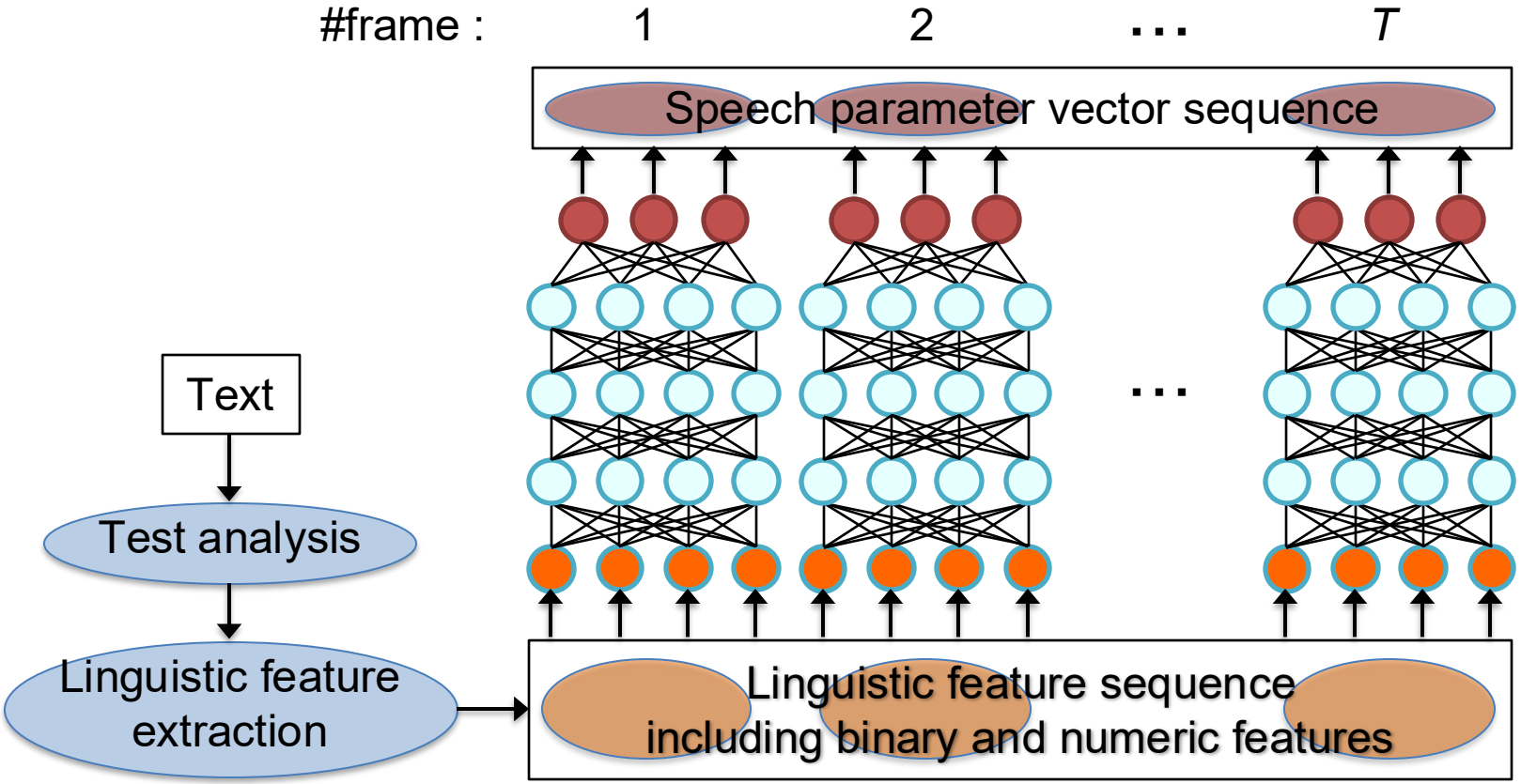


Statistical parametric speech synthesis



DNN-based speech synthesis [Zen et al; 12]

- Mapping linguistic features to speech parameter
- Context clustering part in HMM speech synthesis

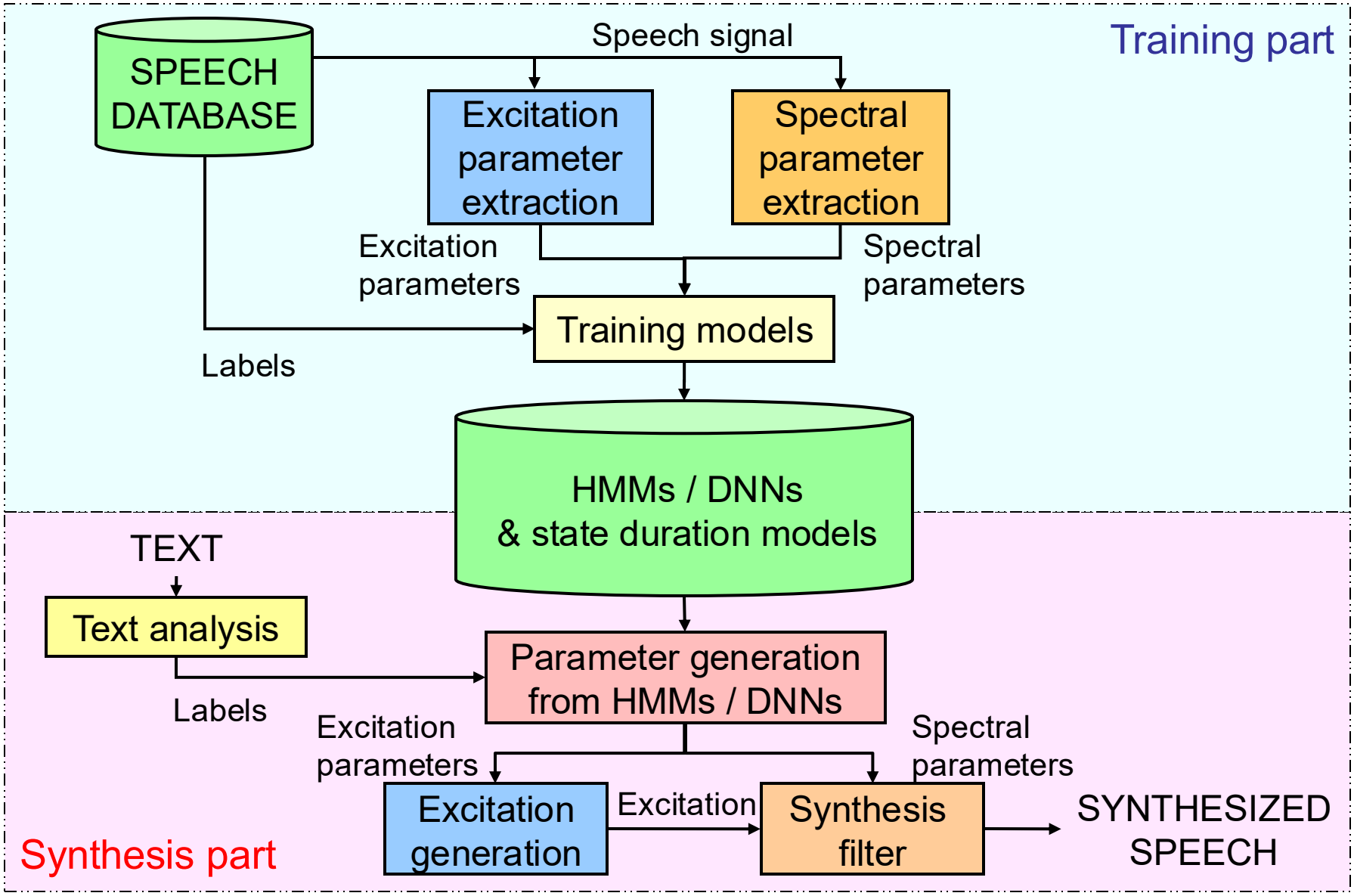


Linguistic features：Answers to questions about linguistic contexts,
Durations of the current phoneme

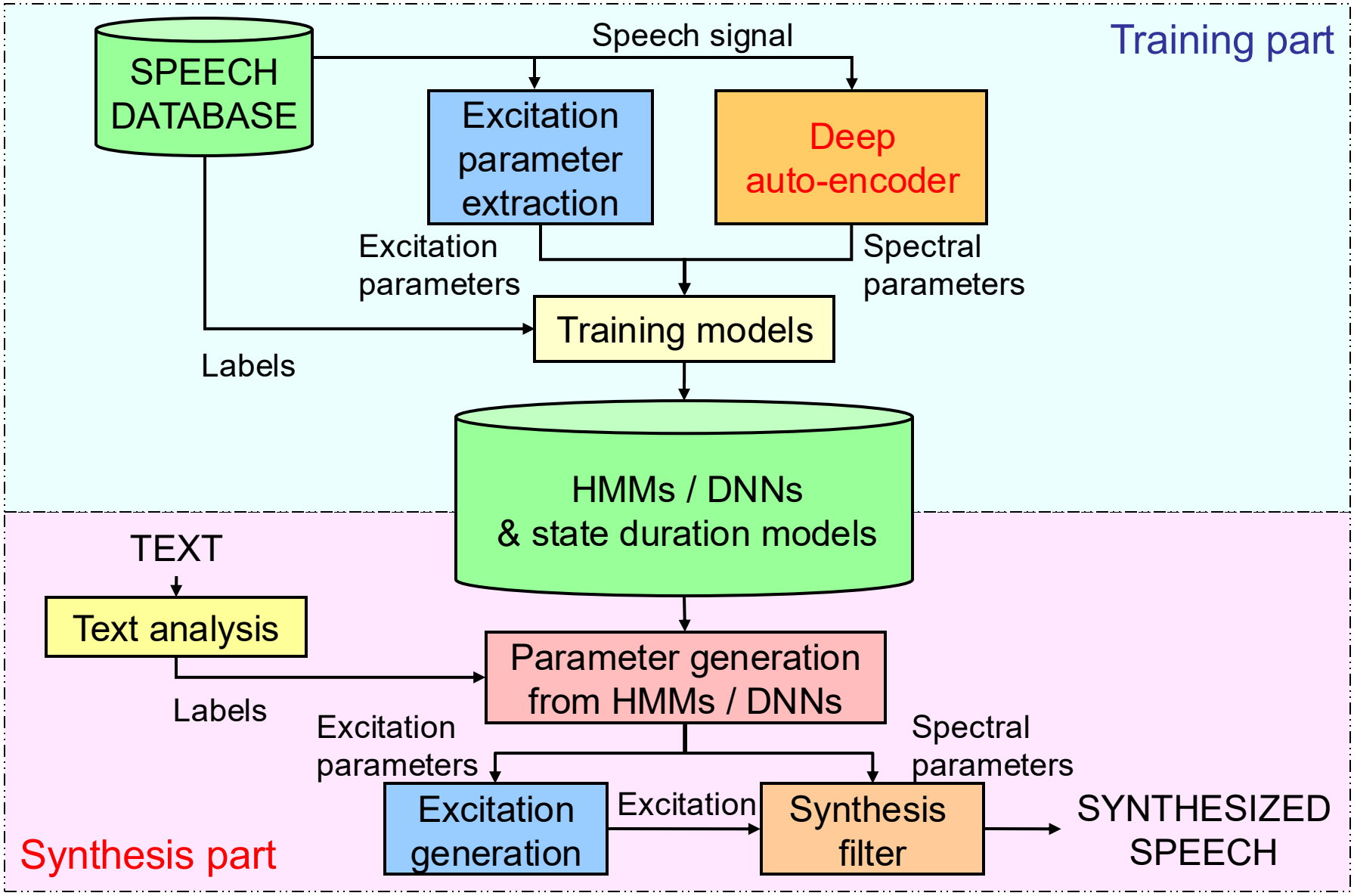
Outline

- Overview of DNN-based speech synthesis
- Spectral feature extraction based on deep auto-encoder 👈
- DNN-based direct spectral modeling
- Text-to-speech synthesis experiment

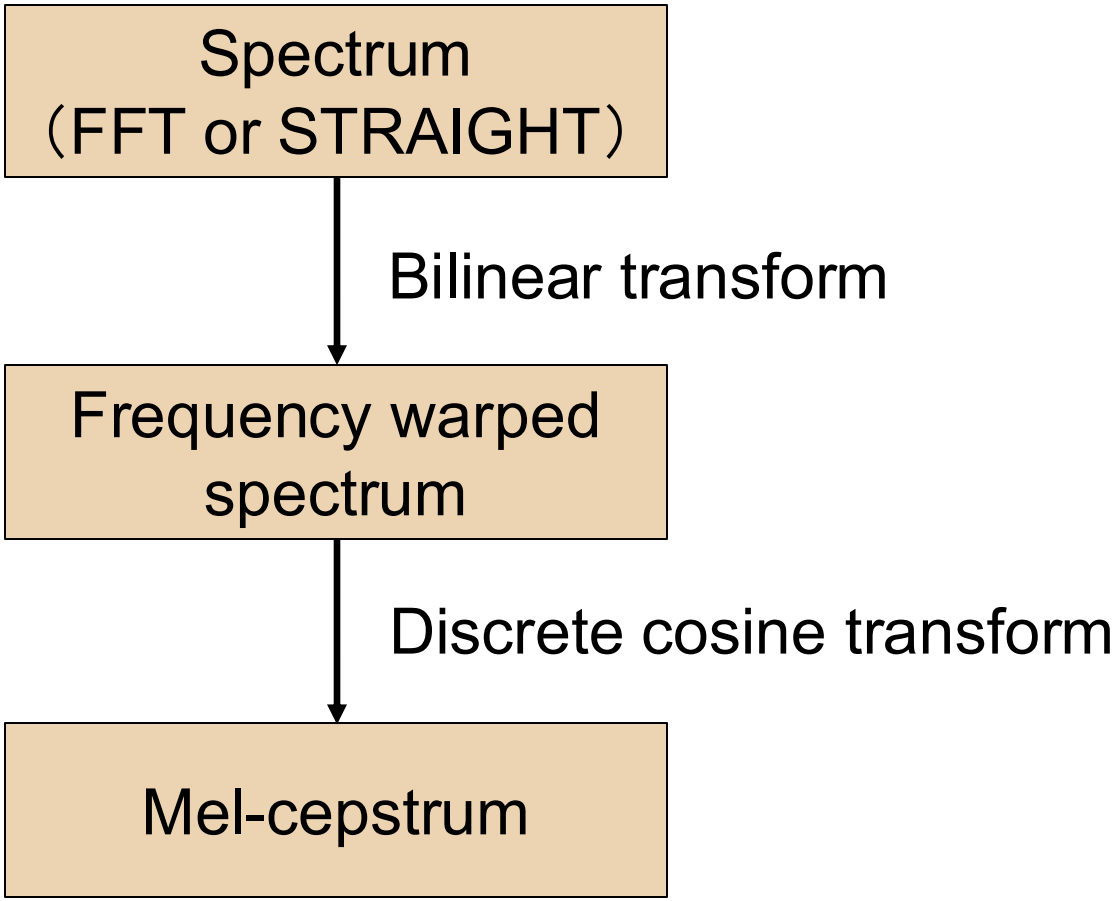
Statistical parametric speech synthesis



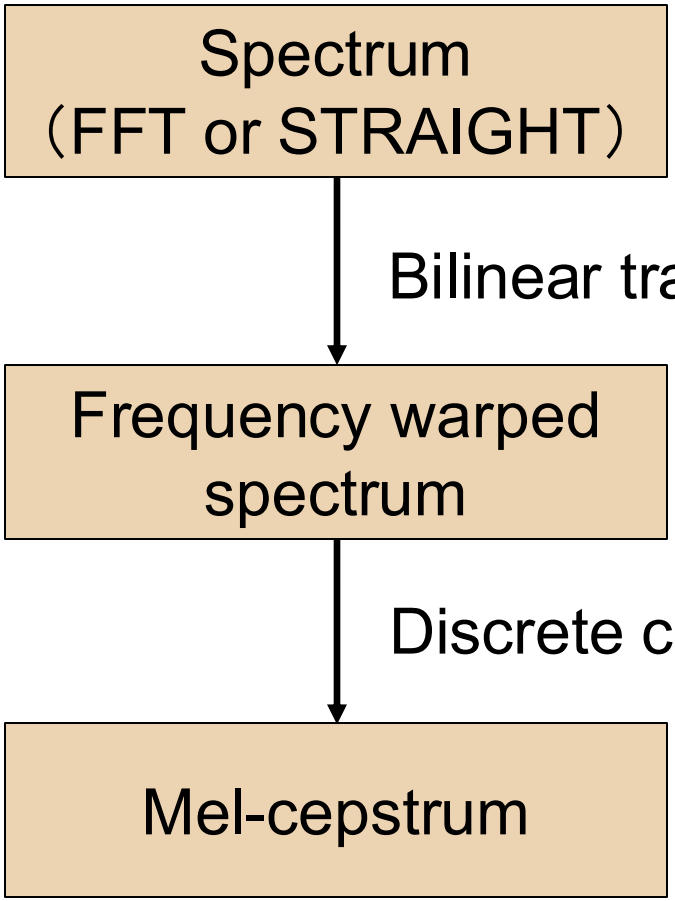
Statistical parametric speech synthesis



Spectral feature extraction



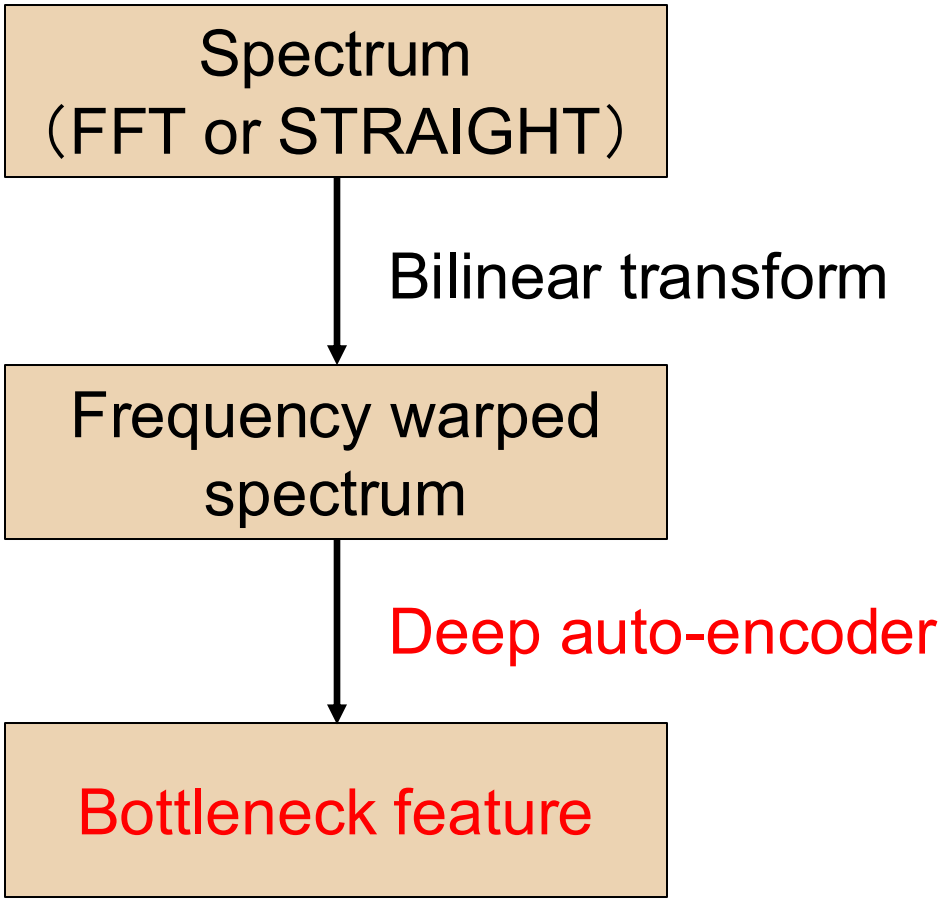
Spectral feature extraction



STRAIGHT Vocoder

- Spectral amplitude
- F0
- Apriodicity measure

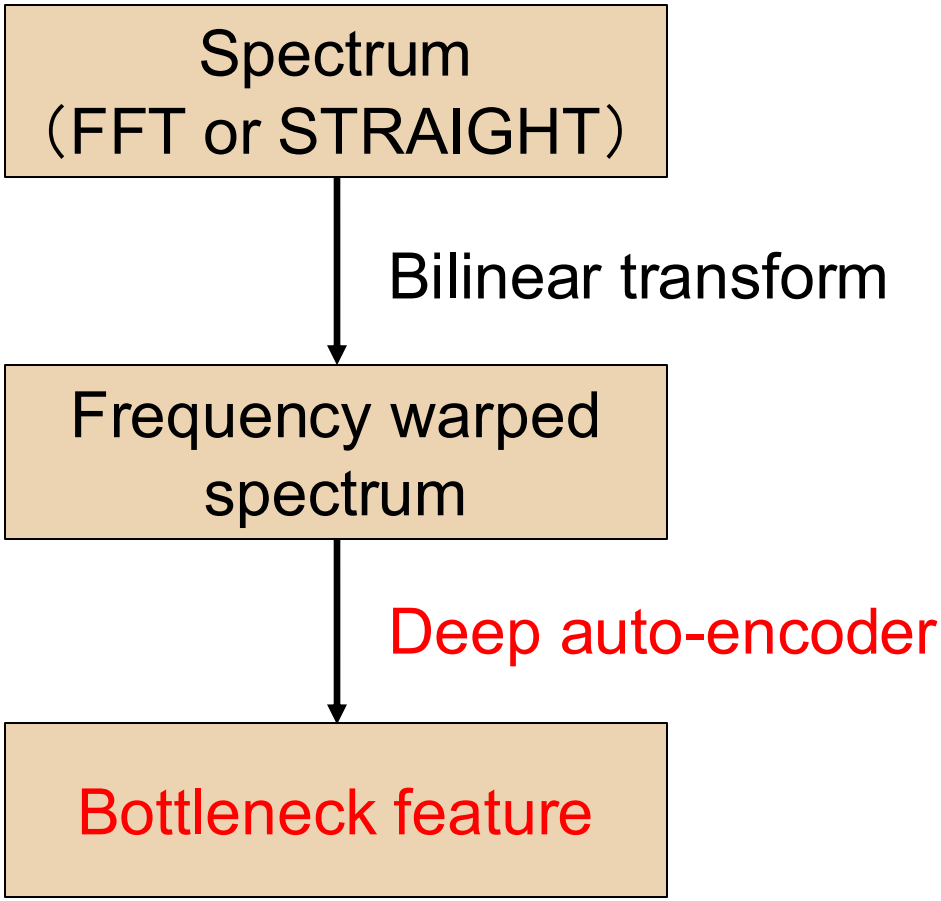
Spectral feature extraction



STRAIGHT Vocoder

- Spectral amplitude
- F0
- Apriodicity measure

Spectral feature extraction



STRAIGHT Vocoder

- Spectral amplitude
- F0
- Apriodicity measure

- Statistical approach
  - Data driven
  - Speaker dependent
- Non-linear
  - Vocal tract has a non-linear shape
- Deep
  - High potential

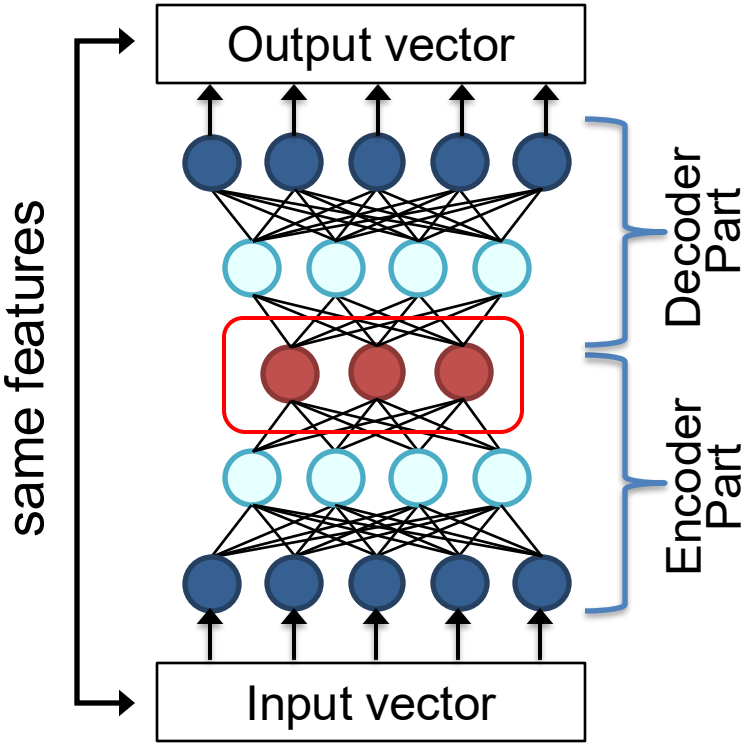
Deep Auto-encoder based feature extraction

Deep Auto-encoder（DAE）

– An artificial neural network

– Typical purpose is dimensionality reduction

– Same features are used as input and output vectors



Outputs of a encoder part can be used as dimensionality reduced features（Bottleneck features）

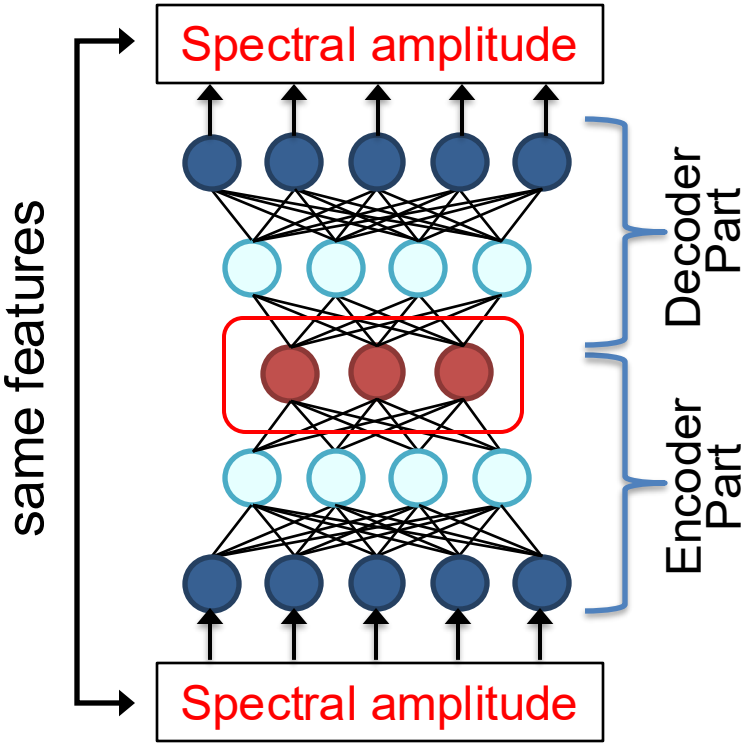
Deep Auto-encoder based feature extraction

Deep Auto-encoder（DAE）

– An artificial neural network

– Typical purpose is dimensionality reduction

– Same features are used as input and output vectors



Outputs of a encoder part can be used as dimensionality reduced features（Bottleneck features）

Extracting low-dimensional spectral parameter

Outline

- Overview of DNN-based speech synthesis
- Spectral feature extraction based on deep auto-encoder
- DNN-based direct spectral modeling 👈
- Text-to-speech synthesis experiment

DNN-based spectral modeling

Low-dimentional spectral parameter → Quality loss

Directly synthesize spectral amplitudes

– Catch the spectral fine structure

Difficulty of DNN training

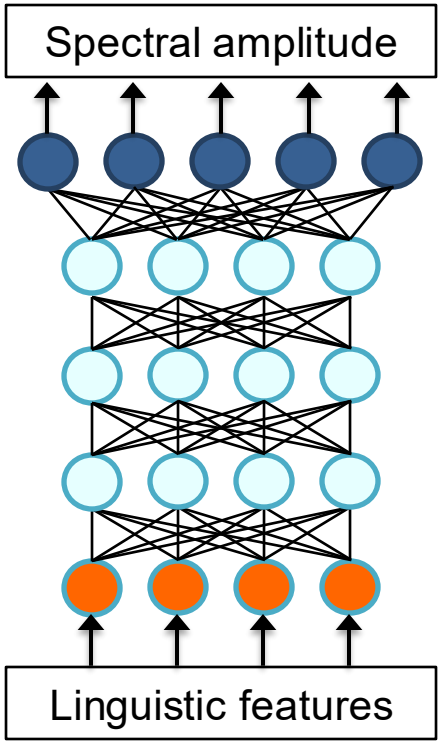
– Local maxima, Vanishing gradient

– High dimensionality

• Mel-cepstrum : 60 dims.

• Spectral amplitude : 2049 dims.

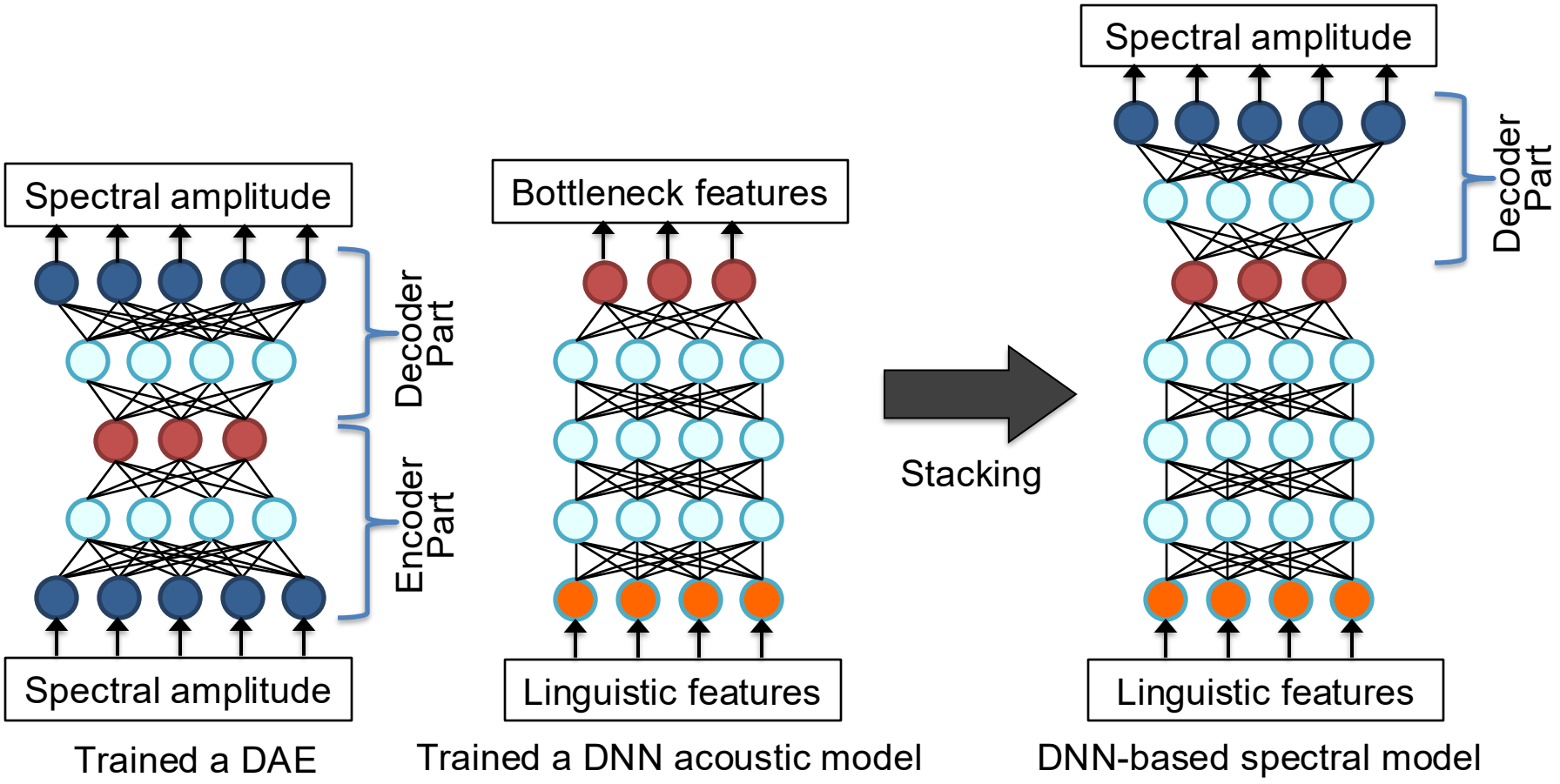
→ Demand efficient training technique



Pre-training using a DAE and a DNN acoustic model

Efficient training for a DNN-based spectral model

The general flow for constructing SPSS is included



Function-wise pre-training for DNN-based speech synthesis

Experimental condition (Text-to-speech)

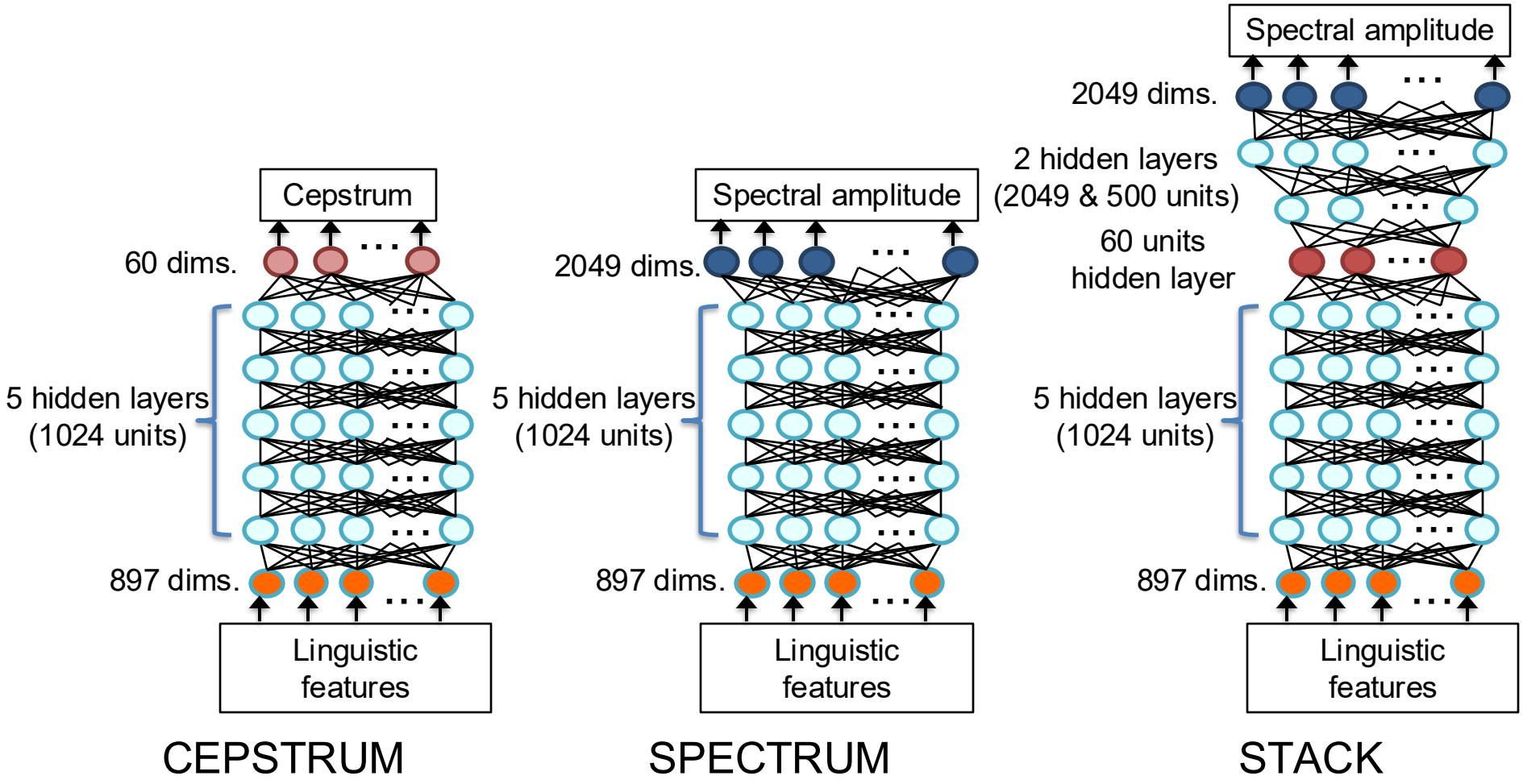
Database	Professional English female 4558 utterances
Sampling rate	48 kHz
FFT points	2049
Vocoder	STRAIGHT

Methods

CEPSTRUM	Mel-cepstrum
SPECTRUM	Spectral amplitude（w/o Pre-training）
STACK	Spectral amplitude（w/ Pre-training）

DNNs outputs only spectral parameters
HMMs synthesized other acoustic features

Systems

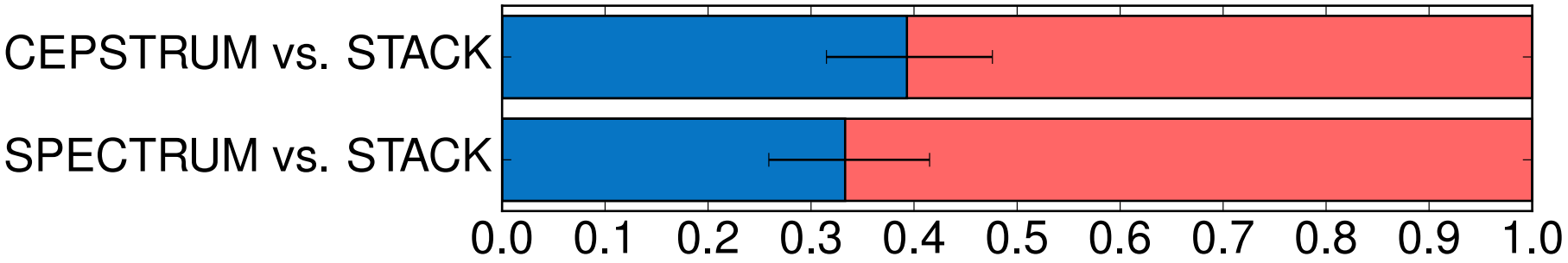


Time derivative features were not used

Experimental result

Subjective test（preference test）

– #subjects：7，each subject evaluate 30 test utterances



Synthesized sampls

CEPSTRUM	
SPECTRUM	
STACK	

Conclusions

DNN-based spectral modeling

- Deep Auto-encoder（DAE）based spectral parameter extraction
  - Statistical approach，deep, non-linear
- Directly synthesize spectral amplitudes using a DNN
  - Efficient training technique would be needed
  - Function-wise pre-training using a DAE and a DNN acoustic model

Future work

- Time derivative features
- Using a FFT spectrum

Thank you