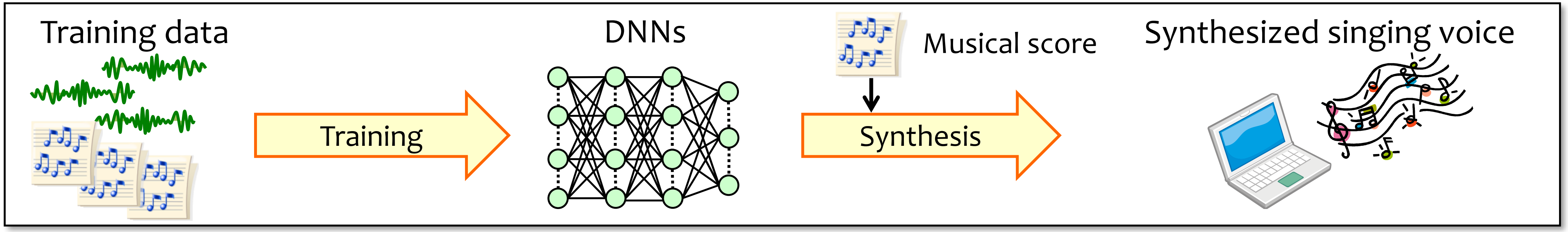
Fast and High-Quality Singing Voice Synthesis System Based on Convolutional Neural Networks

Kazuhiro Nakamura[1], Shinji Takaki[1,2], Kei Hashimoto[1,2], Keiichiro Oura[1,2], Yoshihiko Nankaku[2], and Keiichi Tokuda[1,2]

[1]Techno-Speech, Inc.,
[2] Nagoya Institute of Technology

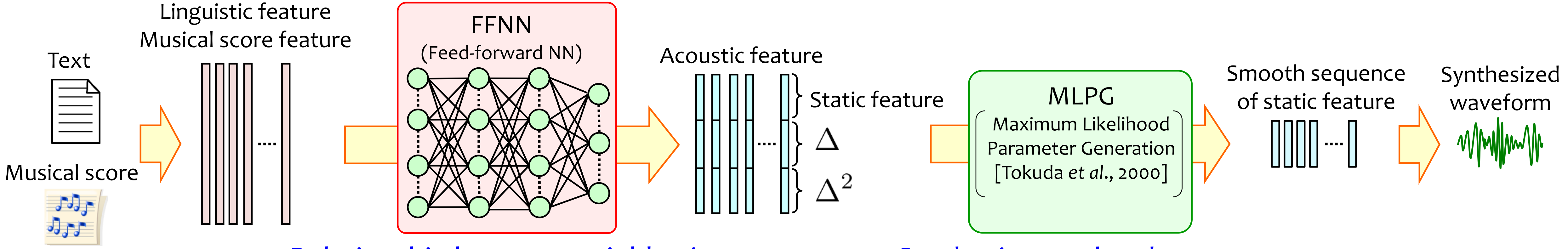
Background

- Deep neural network (DNN)-based acoustic modeling for speech synthesis
  - Represent complex dependencies between linguistic feature and acoustic feature
  - Synthesized speech is natural, but computational complexity is high
- Capturing long-term dependencies
  - Model correlations between neighboring frames (LSTM-RNN, trajectory training, ...)
    Long short-term memory recurrent NN
  - Generate smooth sequence of acoustic features (MLPG, special output layer, ...)
    Maximum likelihood parameter generation
- DNN-based Singing voice synthesis
  - Singing voices represent a rich form of expression
  - A powerful technique to model them accurately is required



Propose a fast and high-quality singing voice synthesis system based on convolutional NN

DNN-based speech synthesis



Relationship between neighboring frames is not considered

Synthesis speed and response are degraded

Conventional approaches

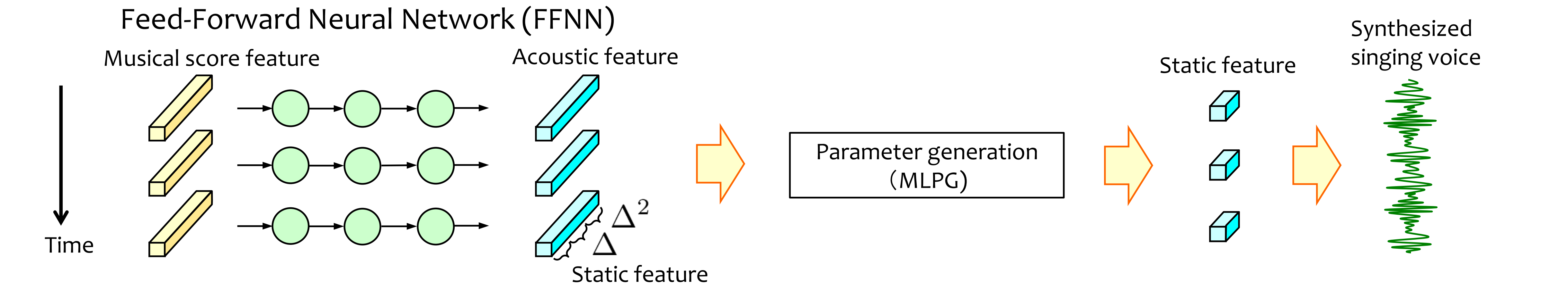
Model correlations between neighboring frames
- Recurrent structure [Fan *et al.*, 2014]
- Trajectory training [Hashimoto *et al.*, 2015]

Use output layer instead of MLPG
- Use recurrent output layer [Zen *et al.*, 2015]
- Use convolutional output layer [Wang *et al.*, 2017]

Proposed approach based on convolutional NN (CNN)

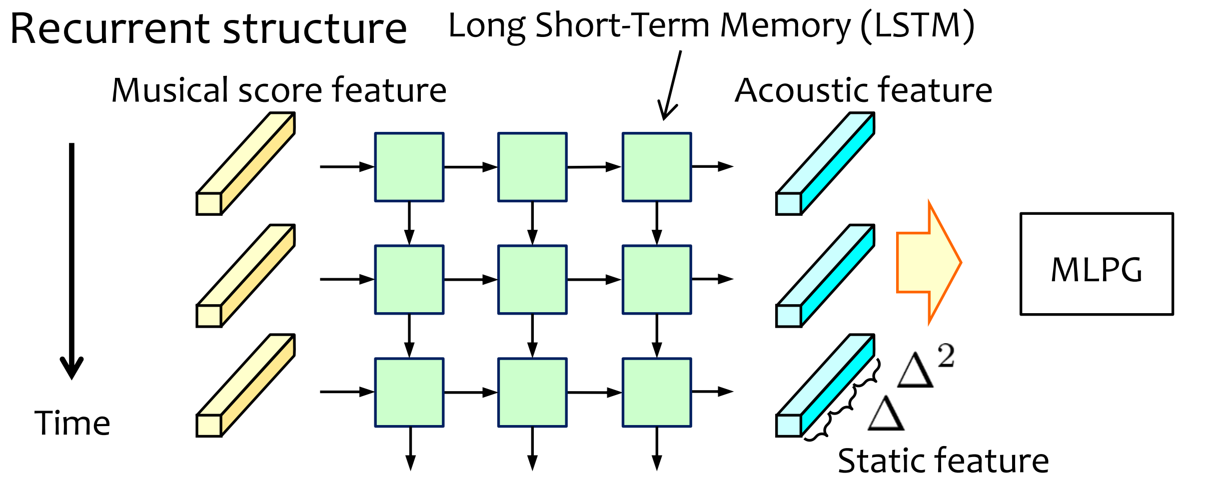
- Long-term dependencies of singing voices are modeled by CNNs
  ⇒ Represent a rich form of expression, easy to parallelize
- Parameter generation is included in the modeling algorithm
  ⇒ Natural trajectory is obtained without MLPG

[Conv.] Model correlations between neighboring frames

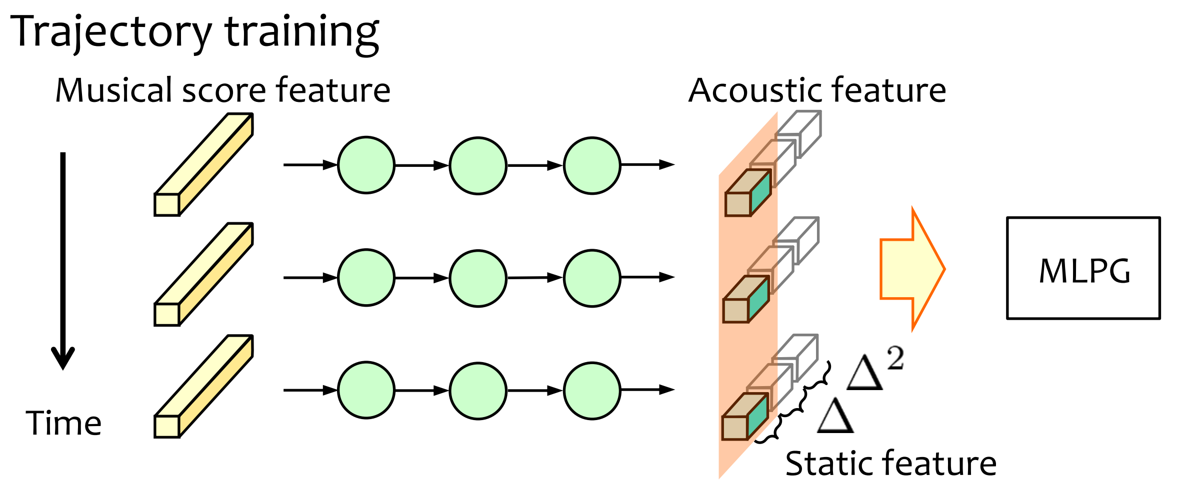


Assignment1: One-to-one mapping in each frame

Assignment2: Maximizing likelihood in each frame individually



Considering the relationship between neighboring frames

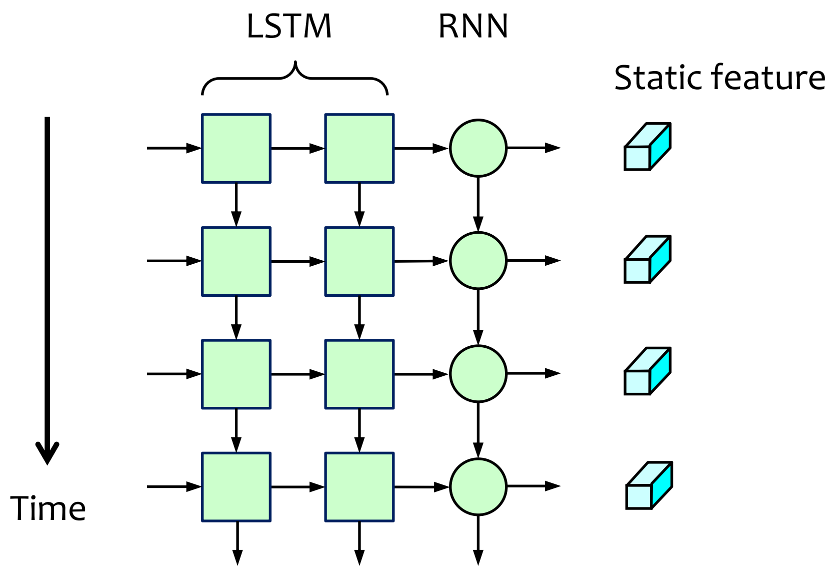


Maximize likelihood of entire sequence

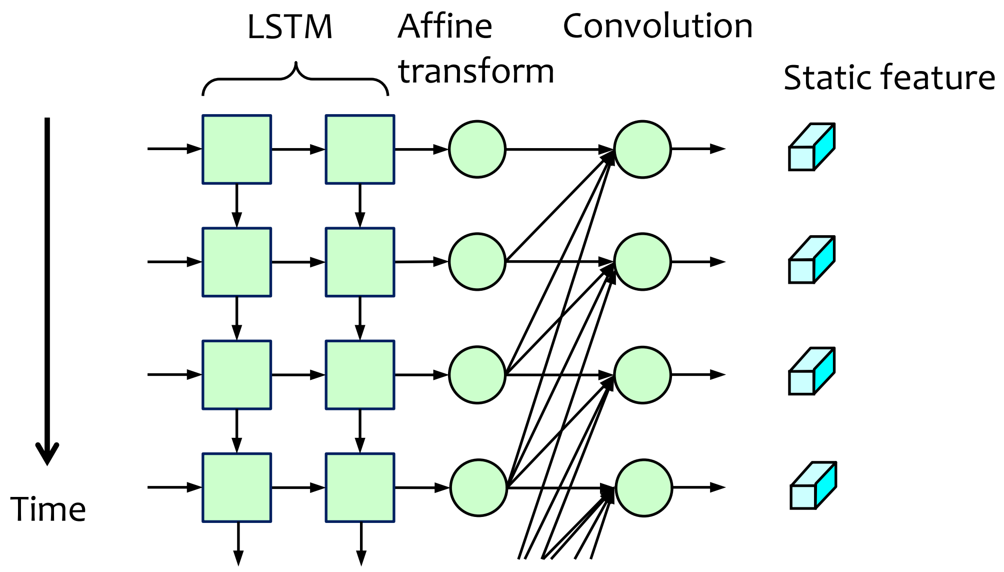
[Conv.] Generate smooth sequence of acoustic features

DNNs output static and dynamic features
⇒ Parameter generation considering relationship between static and dynamic features (MLPG)
Merit: Smooth static feature sequences are generated
Demerit: Synthesis speed and response are degraded

Use recurrent output layer



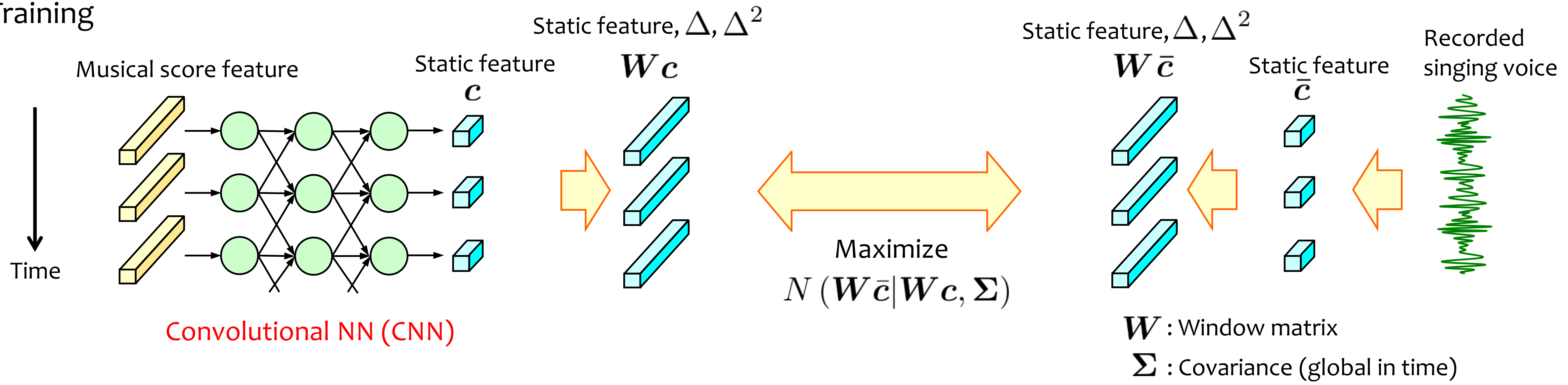
Use convolutional output layer



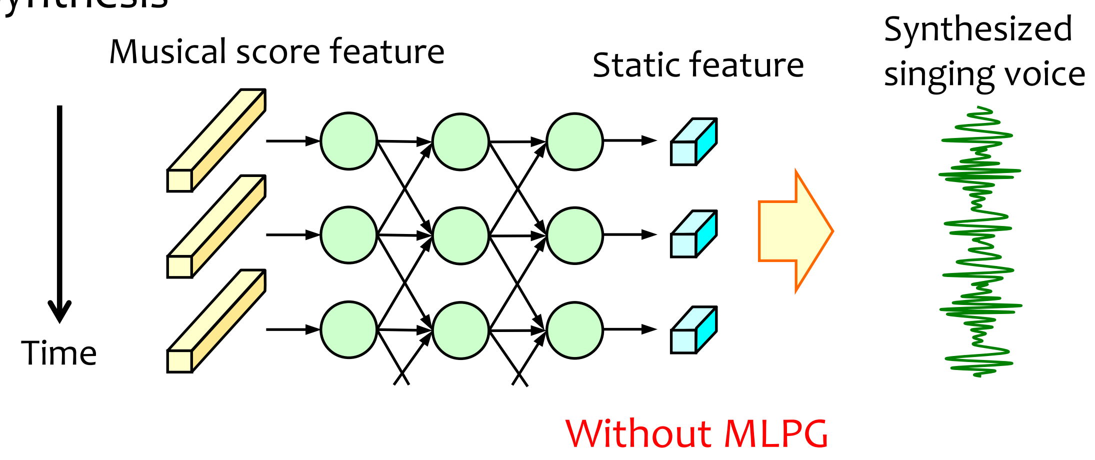
Synthesis speed and response are improved

CNN-based acoustic feature generation

Training



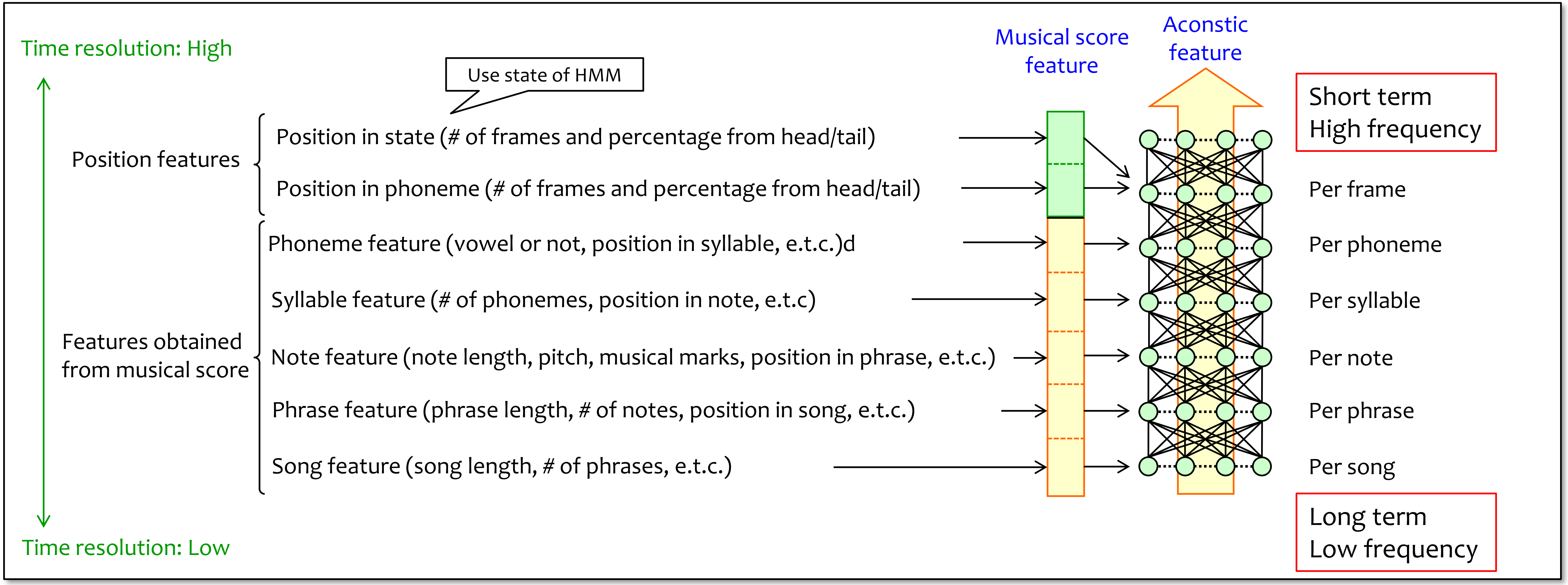
Synthesis



Long-term dependencies of singing voices are modeled by CNNs
⇒ Rich vocal expressions are represented

Trained to maximize likelihood of static and dynamic features
⇒ Natural trajectories are obtained without MLPG

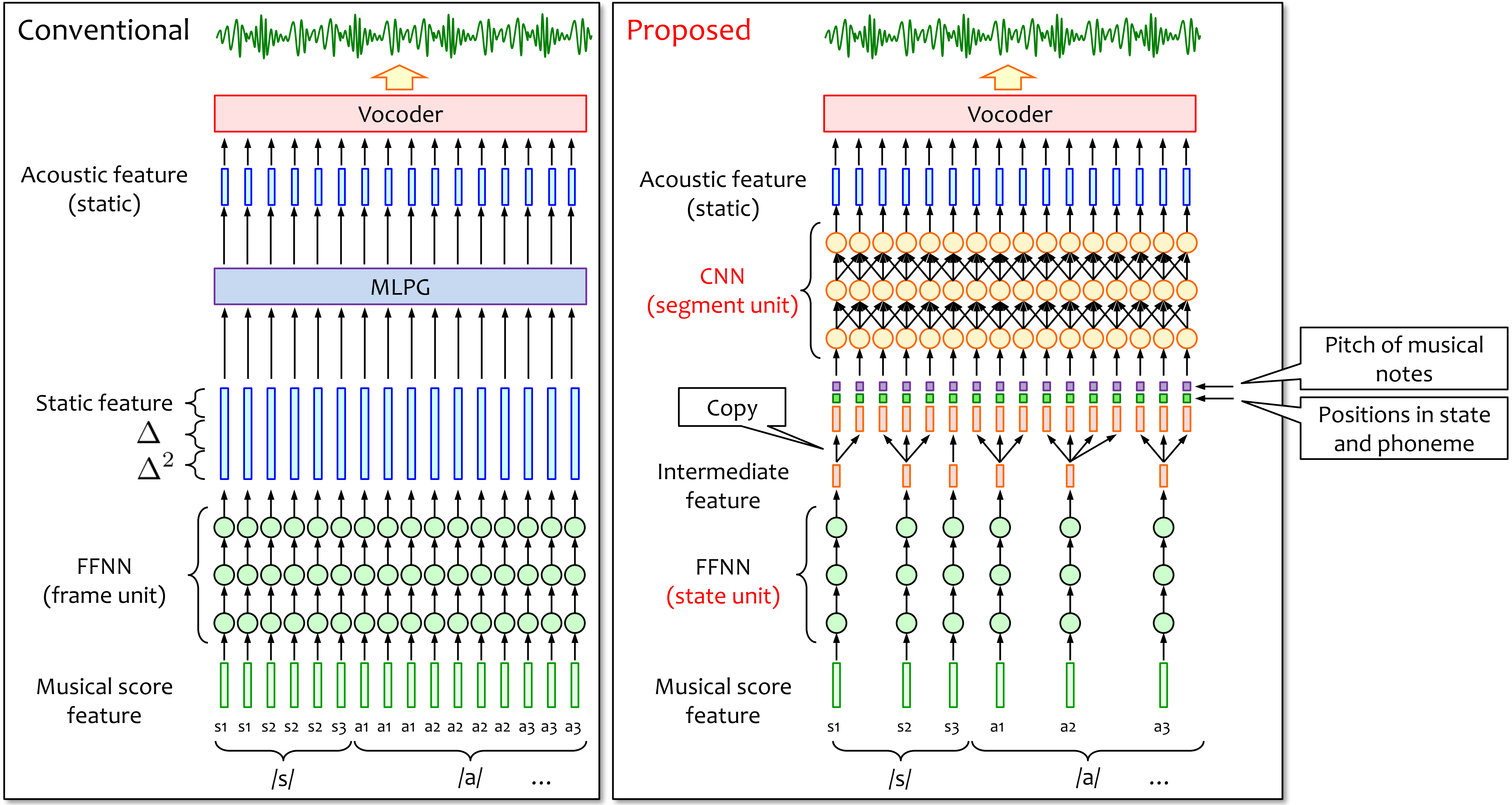
Temporal layer structure of input features



Features should be input in stages according to the temporal resolution

- Features obtained from musical score ⇒ Converted state-by-state
- Position features in phoneme and state ⇒ Converted frame-by-frame

Difference between conventional and proposed methods



Experimental condition of TEST1 (1/2)

Evaluate the quality of synthesized singing voices in cases of two types of vocoders

Database	Song DB by a female singer
Training / Test songs	55 Japanese children's songs and 55 J-POP songs / 5 J-POP songs
Sampling frequency	48 kHz
Frame shift	5 ms
Musical score feature	846 features (normalized from 0 to 1), 1 dimensional pitch in musical score (concatenated to the input of CNNs)
Acoustic feature	0-49 dimensional STRAIGHT mel-cepstral coefficients, log *F0* values + voiced/unvoiced flag, 22 dimensional aperiodicity measures, 2 dimensional vibrato parameters + with/without flag (normalized from 0.01 to 0.99)
Vocoder	- MLSA filter-based vocoder [Imai *et al*., 1983] - WaveNet vocoder [Oord *et al*., 2016, Tamamori *et al*., 2017] Dilation: 1, 2, 3, … , 512 x 3 times, 8 bit μ-law, noise shaping and prefiltering, # of channels: dilations=256, residual=512, skip-connections=256
MOS evaluation condition	5-point MOS 15 subjects x 4 methods x 10 phrases for each method

Experimental condition of TEST1 (2/2)

Conventional methods

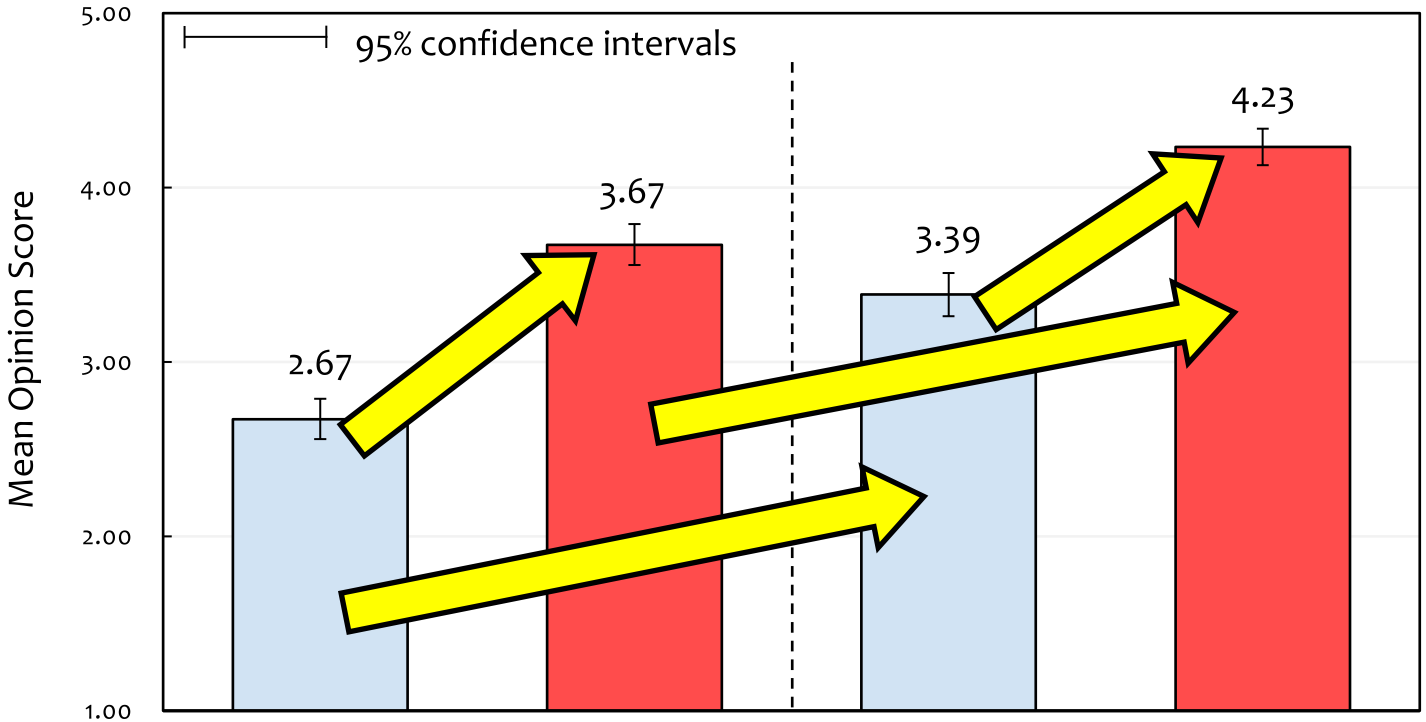
FFNN + MLSA	FFNN-based method (Output static feature, Δ, $\Delta^2 \Rightarrow$ MLPG) MLSA filter-based vocoder
FFNN + WaveNet	FFNN-based method (Output static feature, Δ, $\Delta^2 \Rightarrow$ MLPG) WaveNet vocoder

Proposed methods

CNN + MLSA	CNN-based method MLSA filter-based vocoder
CNN + WaveNet	CNN-based method WaveNet vocoder

State durations of the test songs were predicted by other FFNNs

Experimental result of TEST1

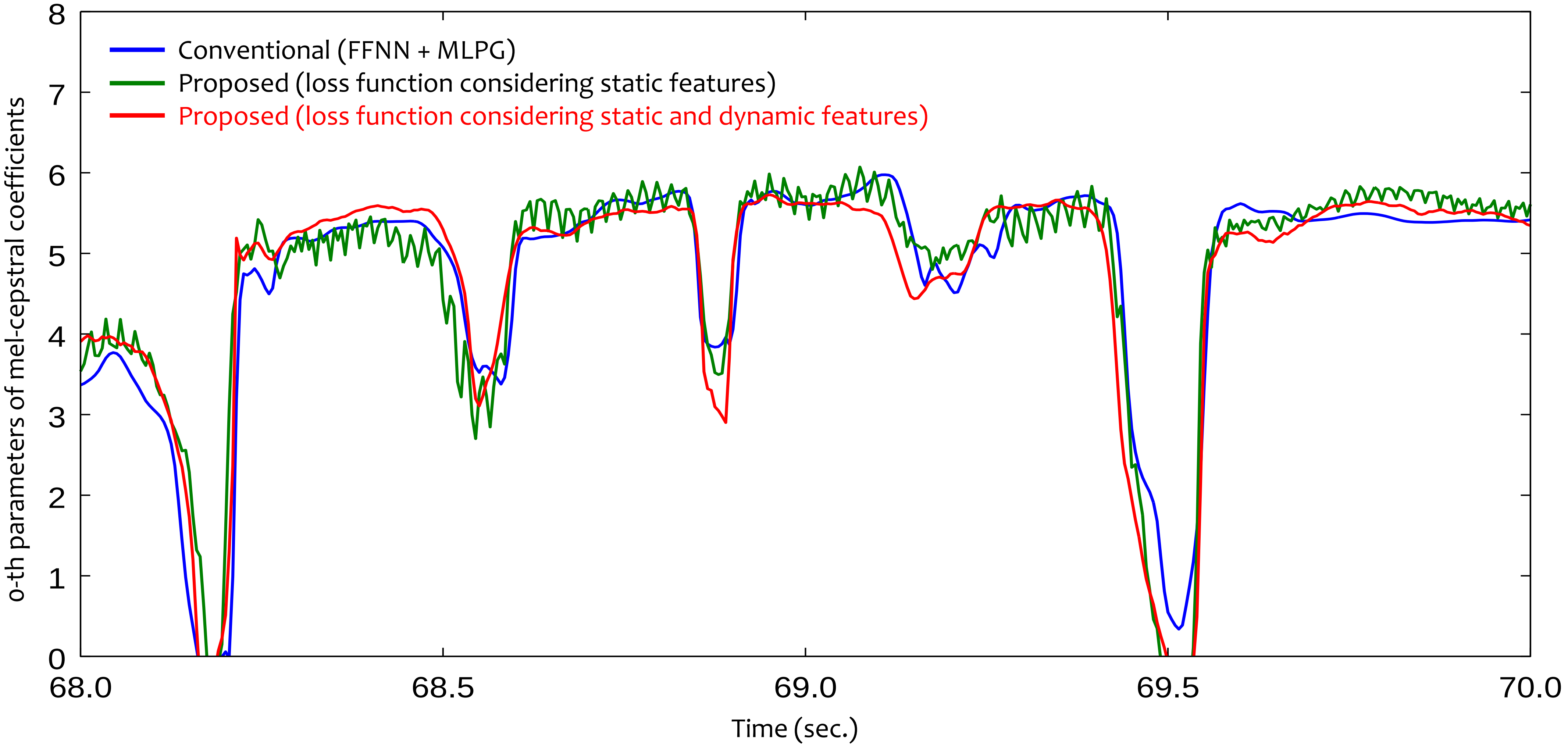


Method	FFNN	CNN	FFNN	CNN
Vocoder	MLSA filter-based vocoder		WaveNet vocoder	
Samples				

- Comparison about acoustic models
  ⇒ FFNN < CNN
- Comparison about vocoders
  ⇒ MLSA filter-based vocoder < WaveNet vocoder

Effect of loss function considering dynamic features

- Comparison of 0-th parameters of mel-cepstral coefficients



The loss of the dynamic features is effective to obtain a smooth parameter sequence

Experimental condition of TEST2 (1/2)

Evaluate the relationship between computational complexity and quality

Goal: Reduce computational complexity without degradation of naturalness

Database	Song DB by a female singer
Training / Test songs	55 Japanese children's songs and 55 J-POP songs / 5 J-POP songs
Sampling frequency	48 kHz
Frame shift	5 ms
Musical score feature	846 features (normalized from 0 to 1), 1 dimensional pitch in musical score (concatenated to the input of CNNs)
Acoustic feature	0-49 dimensional STRAIGHT mel-cepstral coefficients, log $F0$ values + voiced/unvoiced flag, 22 dimensional aperiodicity measures, 2 dimensional vibrato parameters + with/without flag (normalized from 0.01 to 0.99)
Vocoder	MLSA filter-based vocoder
Calculation time measurement condition	Core i7-6700 (single thread)
MOS evaluation condition	5-point MOS 16 subjects x 4 methods x 10 phrases for each method

Experimental condition of TEST2 (2/2)

Conventional method

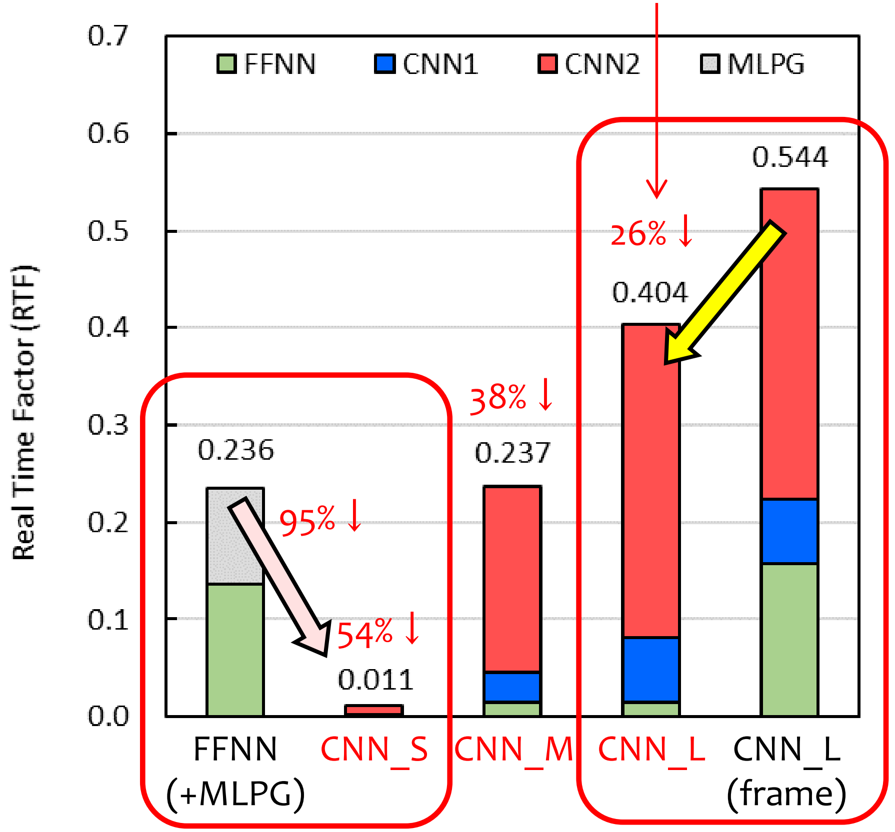
FFNN (+MLPG)	FFNN-based method (Output static feature, Δ, $\Delta^2 \Rightarrow$ MLPG) MLSA filter-based vocoder

Proposed methods

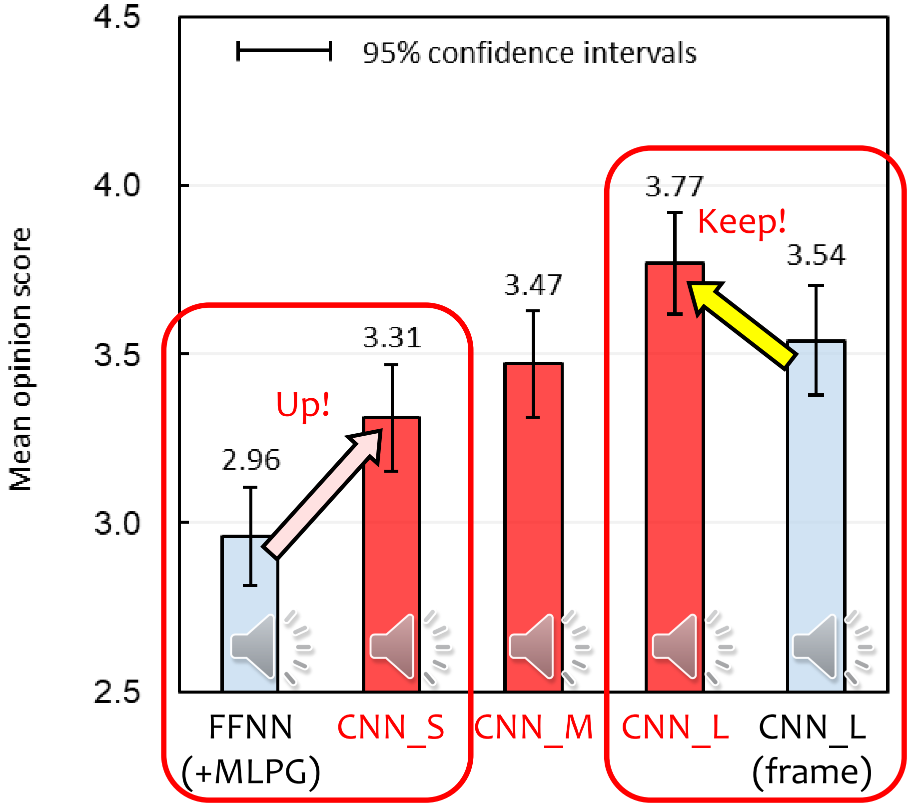
CNN_S	Computational time was about 5% of the conventional method	with computational complexity reduction technique
CNN_M	Computational time was about 100% of the conventional method	
CNN_L	Model size was same as CNN+MLSA in TEST1	
CNN_L (frame)	Same as CNN+MLSA in TEST1 (Frame-level CNN-based method)	without computational complexity reduction technique

State durations of the test songs were predicted by HSMMs

Experimental results of TEST2



Objective evalutioanl result of computational time



Subjective evaluation result of MOS

- Comparison between FFNN and CNN_S
  ⇒ Computational time was reduced by about 95 % and naturalness was improved
- Comparison between CNN_L (frame) and CNN_L
  ⇒ Computational time was reduced without degradation of naturalness

Conclusions

- CNN-based acoustic modeling technique for singing voice synthesis
  - Capturing long-term dependencies of singing voice
  - Loss function for obtaining smooth parameter sequence without MLPG

  ⇒ Generates more natural synthesized singing voices
- Model structure for fast synthesis
  - CNN-based method without recurrent structure ⇒ Easy to parallelize
  - Computational complexity reduction technique
    - Features obtained from musical score ⇒ Converted state-by-state
    - Position features in phoneme and state ⇒ Converted frame-by-frame

    ⇒ Computational time was reduced without degradation of naturalness
- Future work
  - Comparison with RNN-based method
  - Evaluation of this method on TTS
  - Parameter tuning

Demo (CNN + WaveNet)

https://www.techno-speech.com/news-20181214a-en

Synthesized singing voices with accompaniment
(without manual control)

Everytime (Britney Spears)

Japanese

English

Chinese


14th December 2018
Techno-Speech, Inc.

Reproducing high-quality singing voice
with state-of-the-art AI technology

Techno-Speech, Inc. and Nagoya Institute of Technology Speech and Language Processing Laboratory recently developed a singing voice synthesis technology that can reproduce human voice quality, unique characteristics, and singing style more precisely than ever.

Techno-Speech, Inc. and Nagoya Institute of Technology are collaborating on the research and development of speech/singing-voice synthesis technology. The technologies they have developed so far have already been applied in the commercial karaoke system "JOYSOUND," voice creation software "CeVIO Creative Studio," and elsewhere. In this research, a singing-voice database of about two hours of singing recorded by a specific singer is used to develop human voice quality, unique characteristics, and singing style by applying AI technology such as deep learning. When synthesizing, high-quality singing voices can be produced simply by entering any musical score with lyrics.

Languages: Japanese, English, Chinese
Samples: New technology (mix and a cappella)
Current technology (a cappella)
Input: Musical score with lyrics **that has not been manually adjusted**

* Singing voice database providers
- Japanese: CeVIO Project "Sato Sasara" http://www.cevio.jp/
- English: 1st PLACE co., Ltd. "IA" (Voice source: Lia) http://1stplace.co.jp/ia/world/

[Japanese] Diamonds
New (mix) - 00:00 00:21
New (a cappella) - 00:00 00:15
Current (a cappella) - 00:00 00:15

[Japanese] 瞳 (Hitomi)
New (mix) - 00:00 00:26
New (a cappella) - 00:00 00:21
Current (a cappella) - 00:00 00:21

[English] Rolling In The Deep
New (mix) - 00:00 00:22
New (a cappella) - 00:00 00:18
Current (a cappella) - 00:00 00:18

[English] Everytime
New (mix) - 00:00 00:35
New (a cappella) - 00:00 00:26
Current (a cappella) - 00:00 00:26

[Chinese] 愛情转移 (Ai Qing Zhuan Yi)
New (mix) - 00:00 00:25
New (a cappella) - 00:00 00:19
Current (a cappella) - 00:00 00:19