Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis

Shinji Takaki[1], Hirokazu Kameoka[2], Junichi Yamagishi[1]

[1] National Institute of Informatics
[2] NTT Communication Science Laboratories

Background（1/2）

Statistical parametric speech synthesis

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

Waveform generation for TTS

- High-quality vocoder (STRAIGHT, WORLD)
  - Quality deterioration such as buzziness
- Sinusoidal vocoder [Hu et al.; 15]
- Modeling complex spectra [Hu et al.; 16]
- Signal reshaping [Espic et al.; 16]
- Sample RNN [Mehri et al.; 17]
- WaveNet [van den Oord et al.; 16]

Text-to-speech synthesizer with neither the vocoder and computational explosion

Background（2/2）

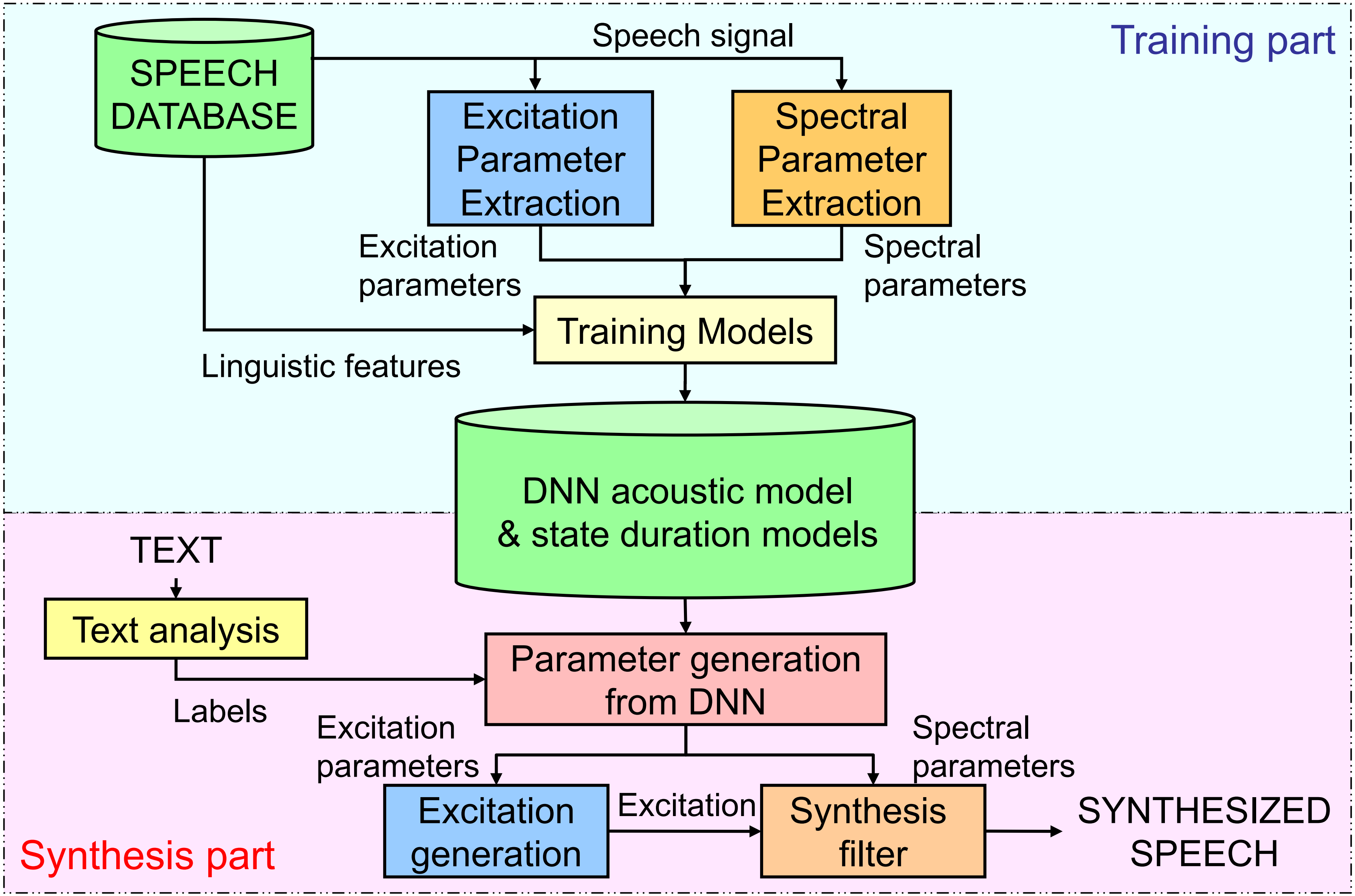
Direct modeling of frequency spectra

- Simple short-time Fourier transform (STFT)
- Spectral envelopes and harmonic structure are included
- Advantages of using STFT
  - The representation is much closer to original waveform
  - DNNs need to be used per frame instead of per sample
- Waveform generation based on phase recovery

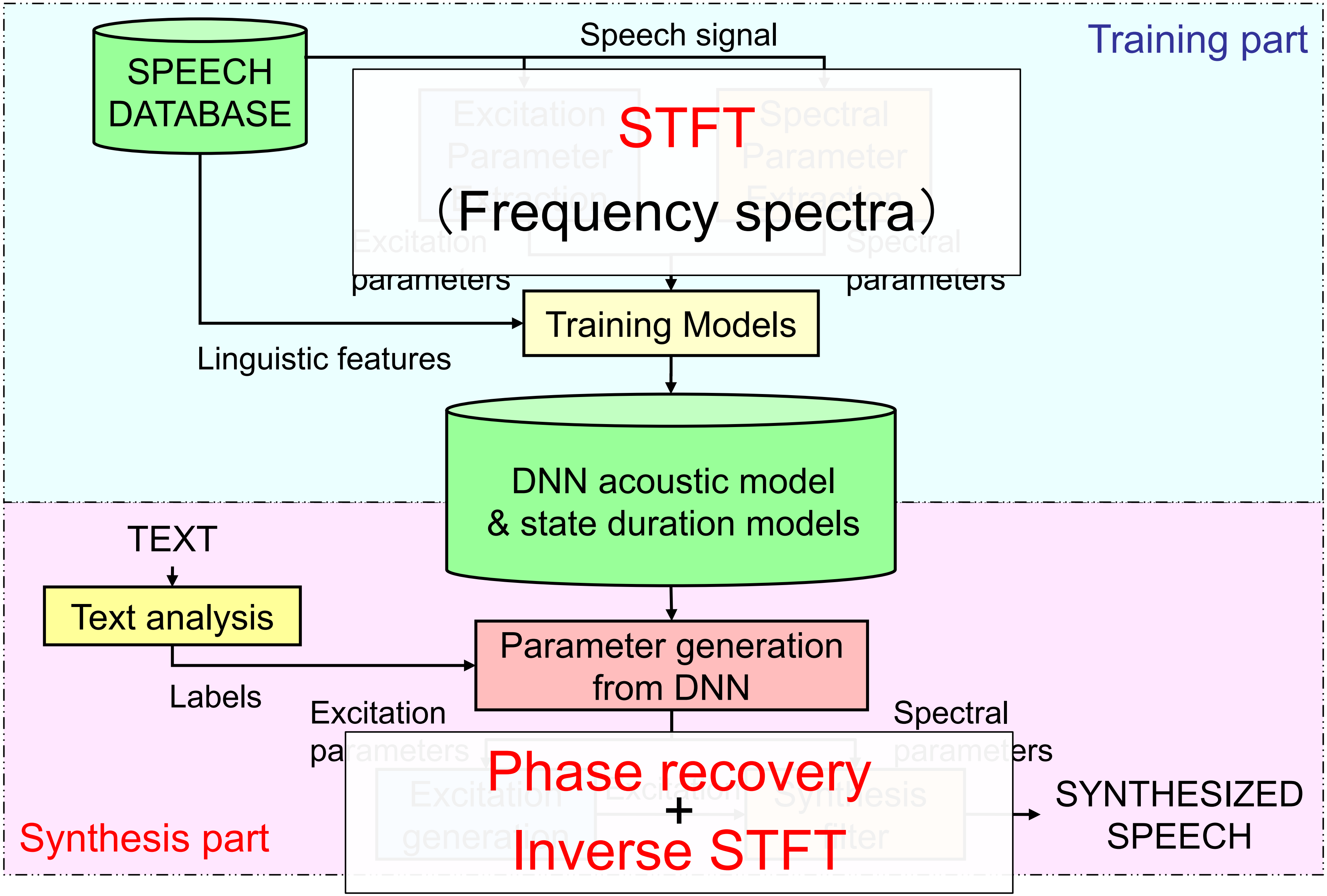
Prediction of STFT based on a DNN

1. The use of F0 information as well as linguistic feature
2. KLD-based objective criterion
3. Post-filtering of predicted STFT

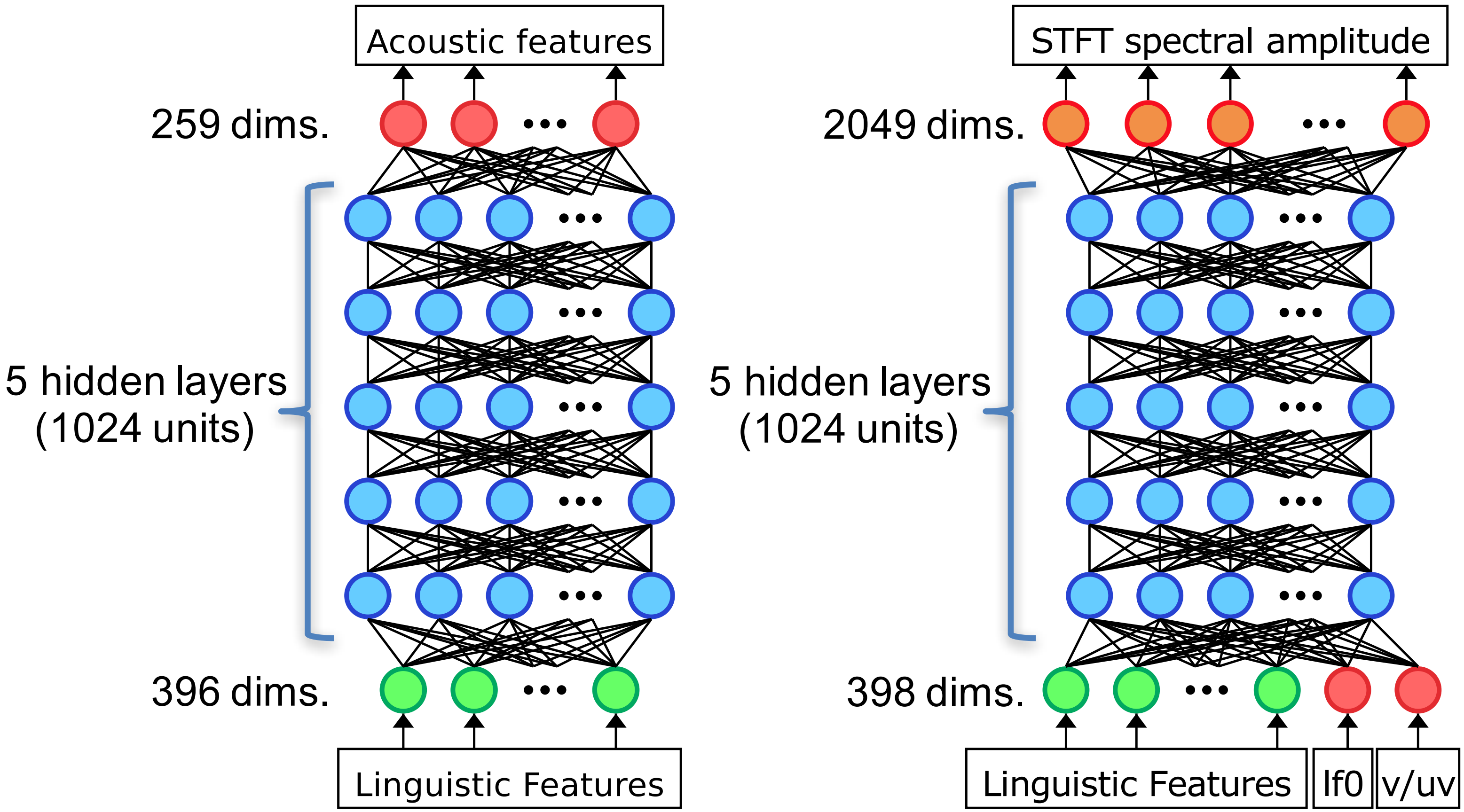
Statistical parametric speech synthesis



Statistical parametric speech synthesis



Architectures (Left conventional, Right proposed)



Vocoder-based conventional framework

STFT-based proposed framework

F0 information is explicitly used as inputs

STFT spectral amplitudes are the outputs

KLD-based training

Least square error (LSE)

$$E_{SE} = \frac{1}{2} \sum_{t=1}^{T} \sum_{d=1}^{D} (o_{t,d} - y_{t,d})^2$$

$o_{t,d}$: obs., $y_{t,d}$: DNN output, t : frame index, d : dim.

Kullback-Leibler divergence (KLD)

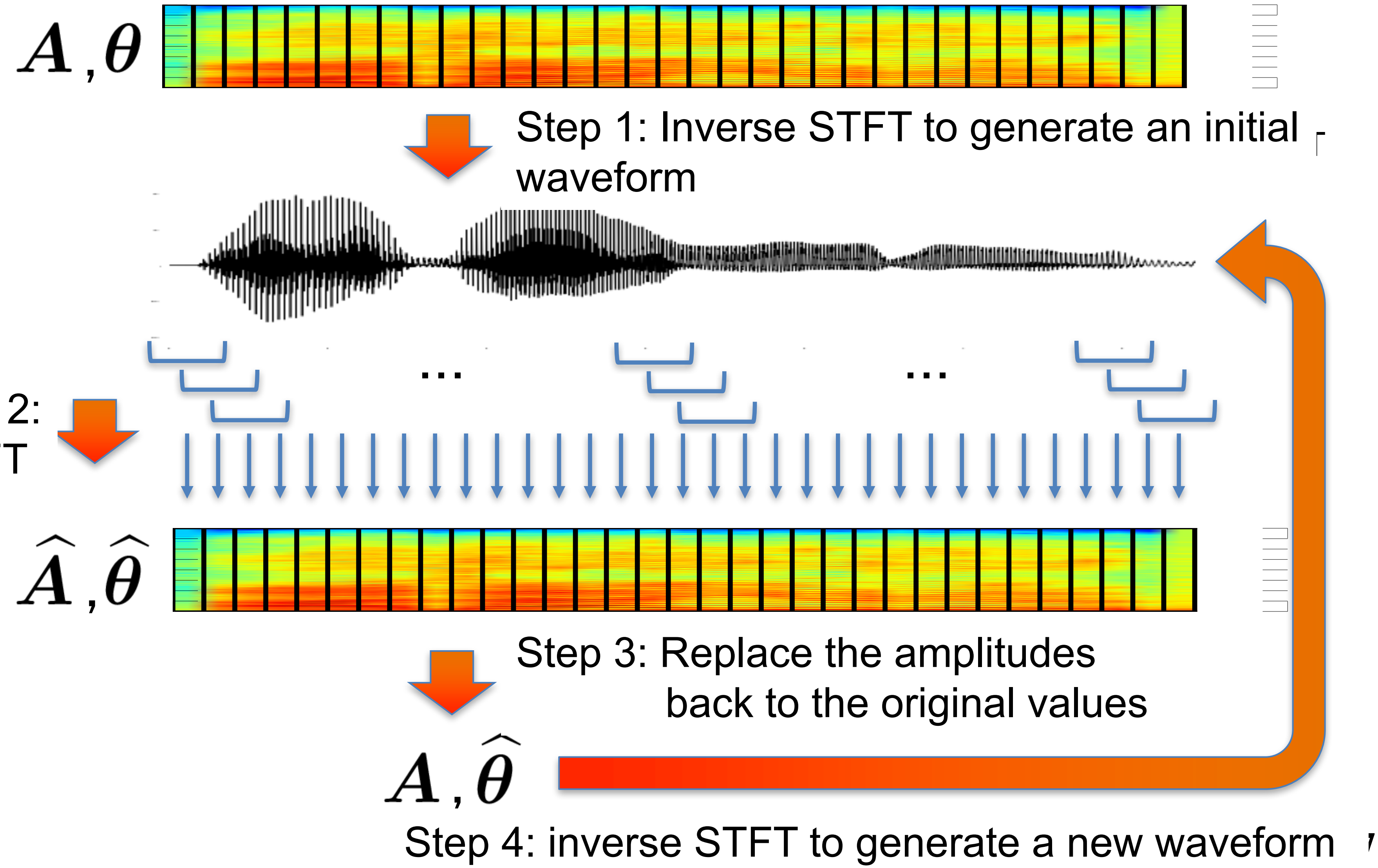
- KLD-based criterion has been successfully used for spectral-domain source separation with NMF
- The sigmoid is used for an output layer in this work

$$E_{KL} = \sum_{t=1}^{T} \sum_{d=1}^{D} o_{t,d} \log \frac{o_{t,d}}{\tilde{y}_{t,d}} - o_{t,d} + \tilde{y}_{t,d}\,, \qquad \tilde{y}_{t,d} = s_d y_{t,d} + b_d$$

$o_{t,d}$: linear spectrum, s_d , b_d : fixed values for unnormalization

Phase recovery from STFT amplitudes

Iterative framework to refine phase [Griffin and Lim; 84]



Experimental conditions (1/2)

Database	Blizzard Challenge 2011 Professional female, 12,085 utterances (17 hours)
Sampling frequency	48 kHz / 32 kHz
FFT points	4096 (2049-dim) / 2048 (1025-dim)
Feature vector （Conventional system）	59 mel-cepstrum $+\Delta + \Delta^2$ log F0 $+\Delta + \Delta^2$ Voiced/unvoiced parameter 25-band aperiodicity $+\Delta + \Delta^2$

Detailed information of the proposed system

- Training: lf0, v/uv obtained from natural speech
- Synthesis: lf0, v/uv synthesized from Baseline
- 100 iteration for phase recovery

Experimental conditions (2/2)

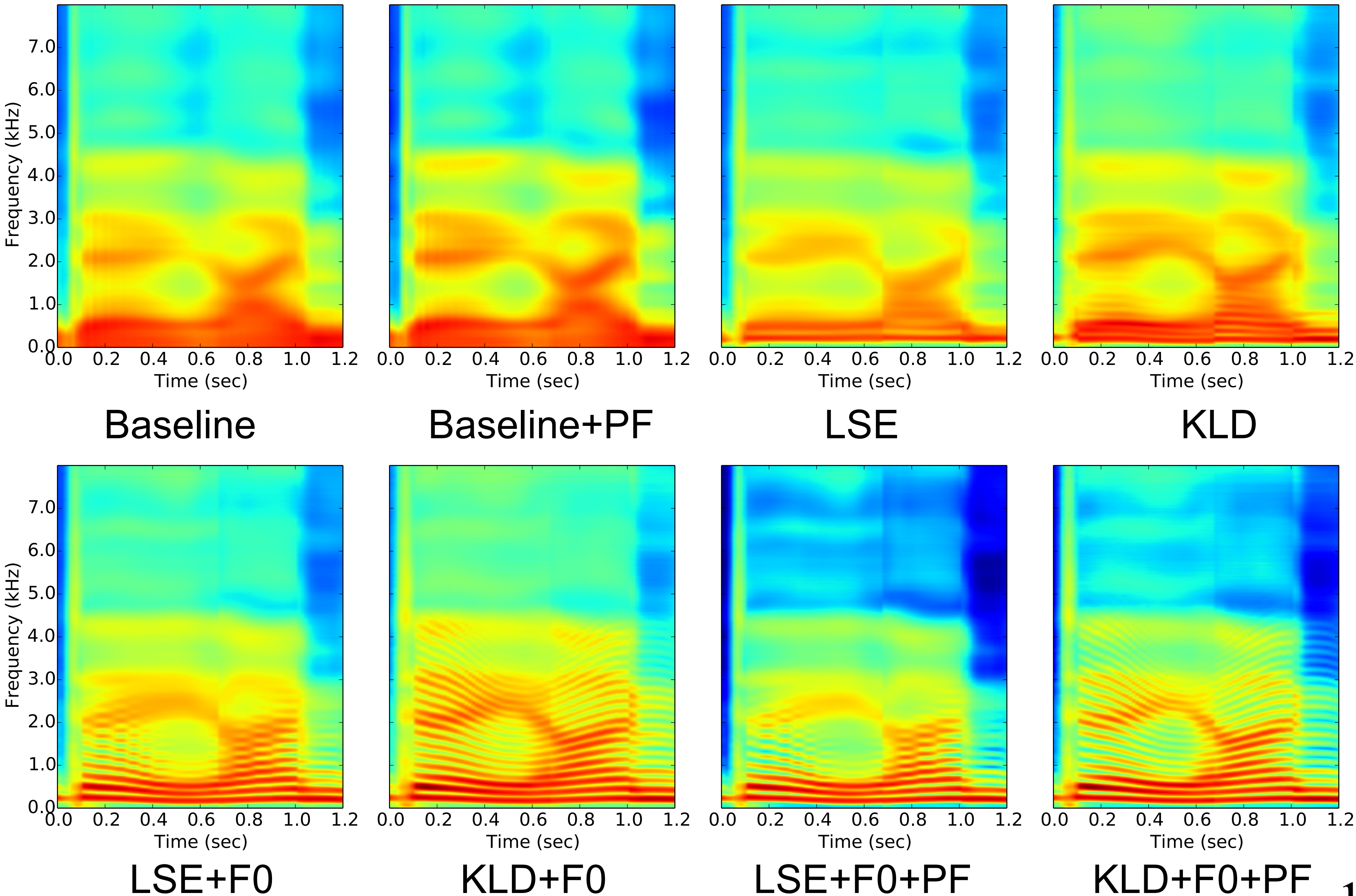
System name	Input	Output	Criterion	Post-filter	Generation
Baseline	Text	Vocoder para.	LSE		Vocoder
Baseline+PF	Text	Vocoder para.	LSE	√	Vocoder
LSE	Text	log STFT	LSE		Phase recovery
KLD	Text	STFT	KLD		Phase recovery
LSE+F0	Text, F0	log STFT	LSE		Phase recovery
KLD+F0	Text, F0	STFT	KLD		Phase recovery
LSE+F0+PF	Text, F0	log STFT	LSE	√	Phase recovery
KLD+F0+PF	Text, F0	STFT	KLD	√	Phase recovery

Signal processing post-filter for peak enhancement

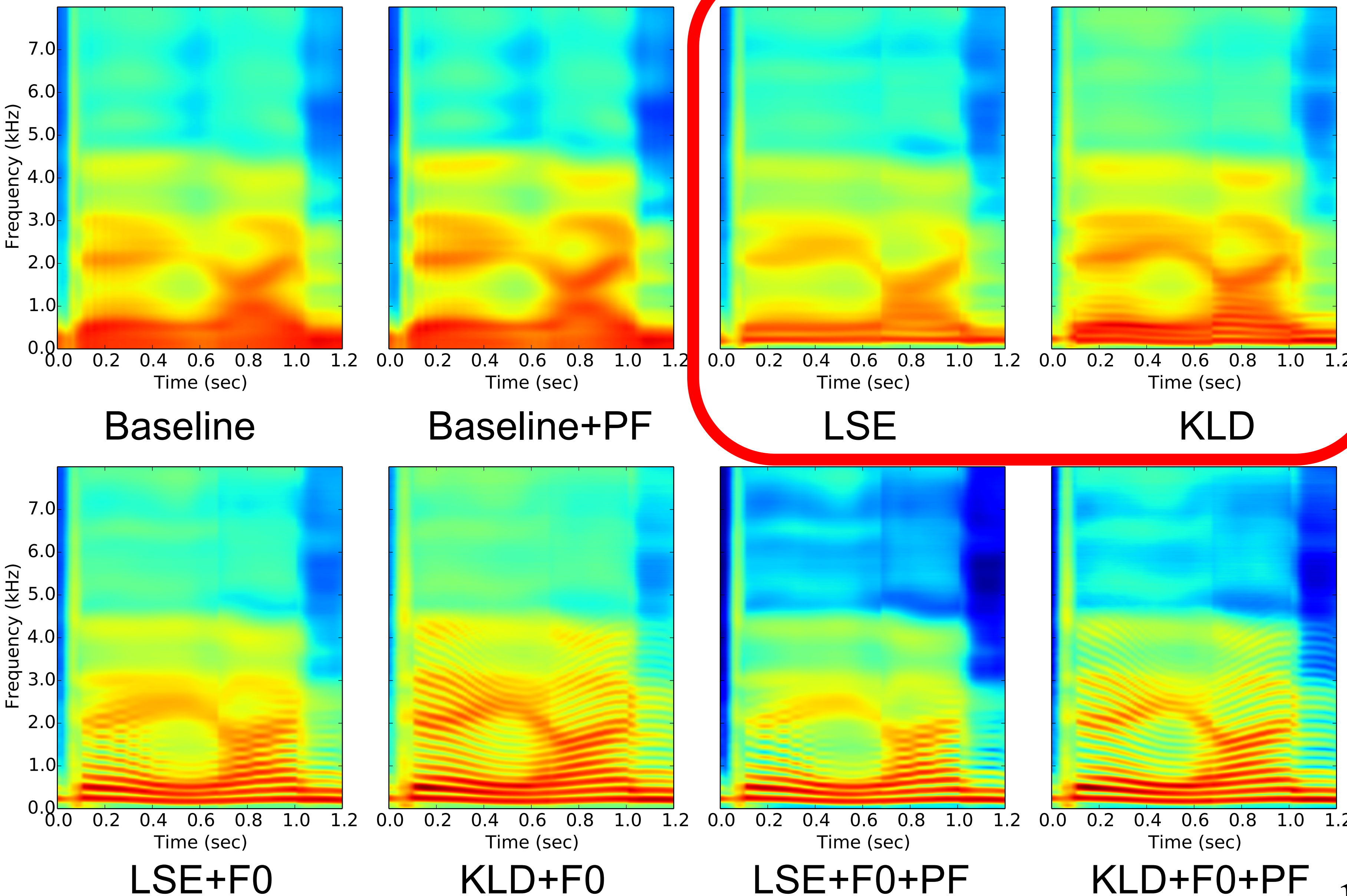
1. Predicted STFT are converted into linear-scale cepstrum
2. The post-filter is applied to the cepstrum
3. The post-filtered cepstrum is converted back into STFT

Wed-P-8-4-6 : GAN-based post-filter for STFT

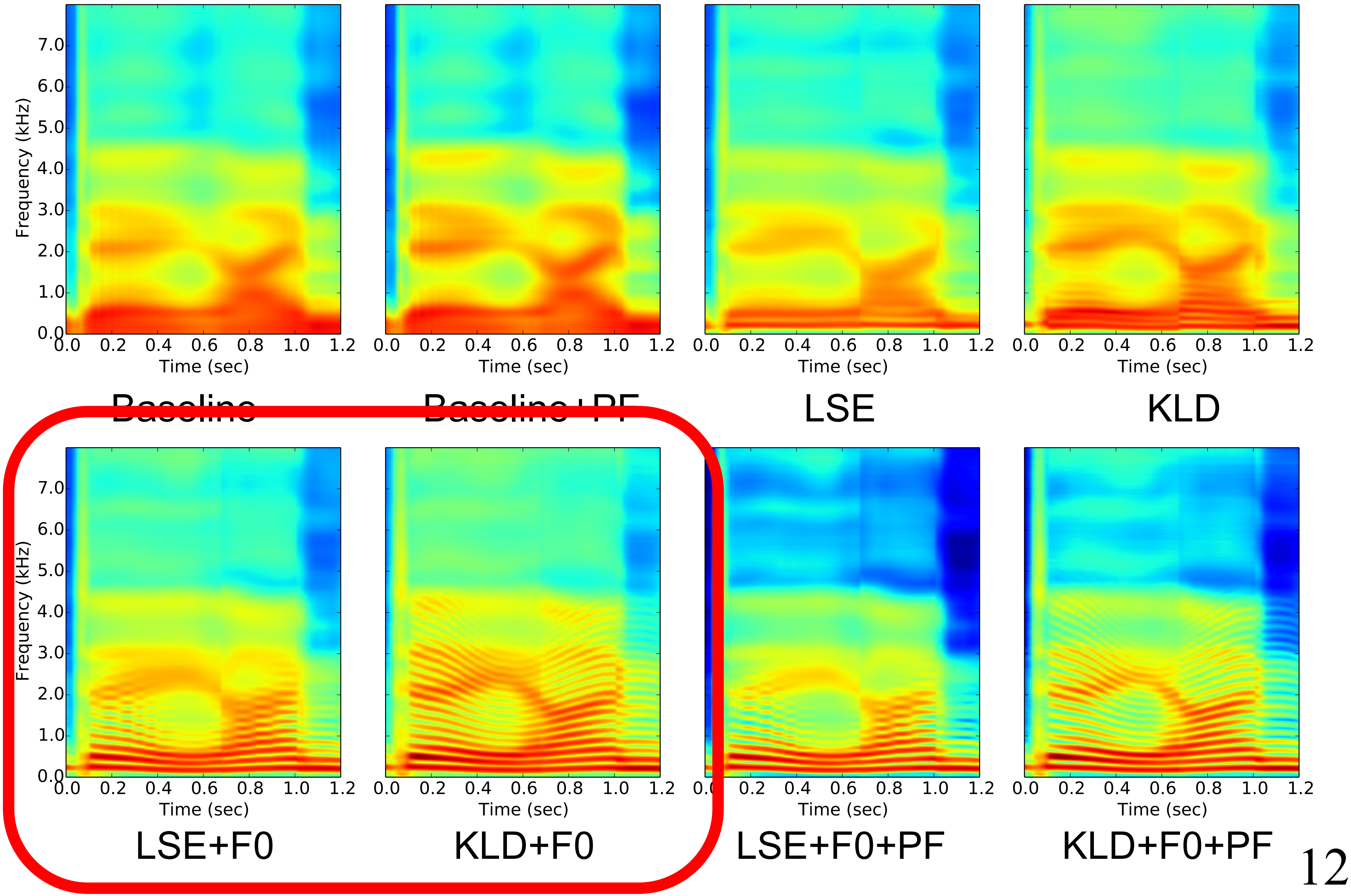
Synthetic spectra (Low-frequency parts)



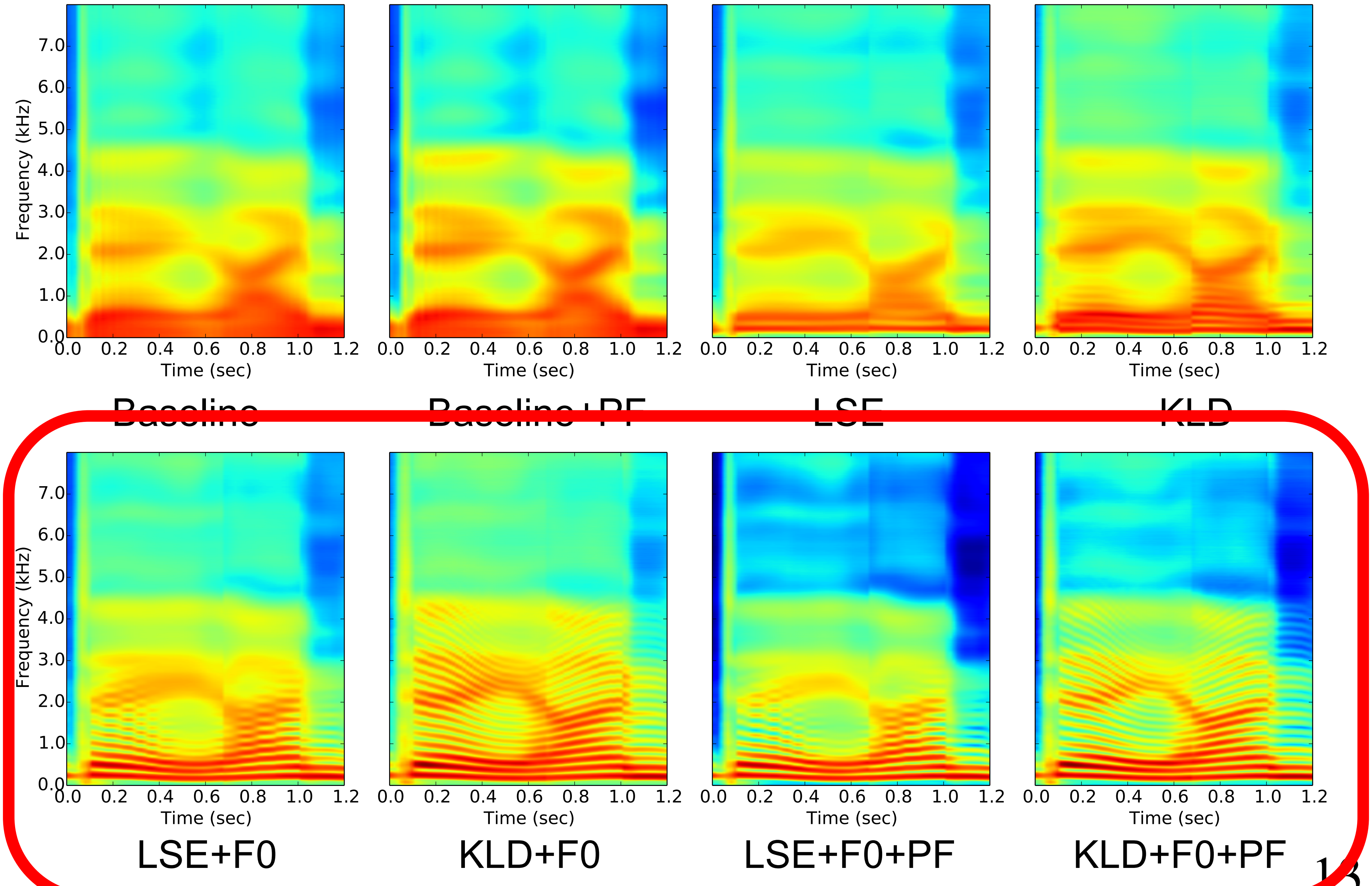
Synthetic spectra (Low-frequency parts)



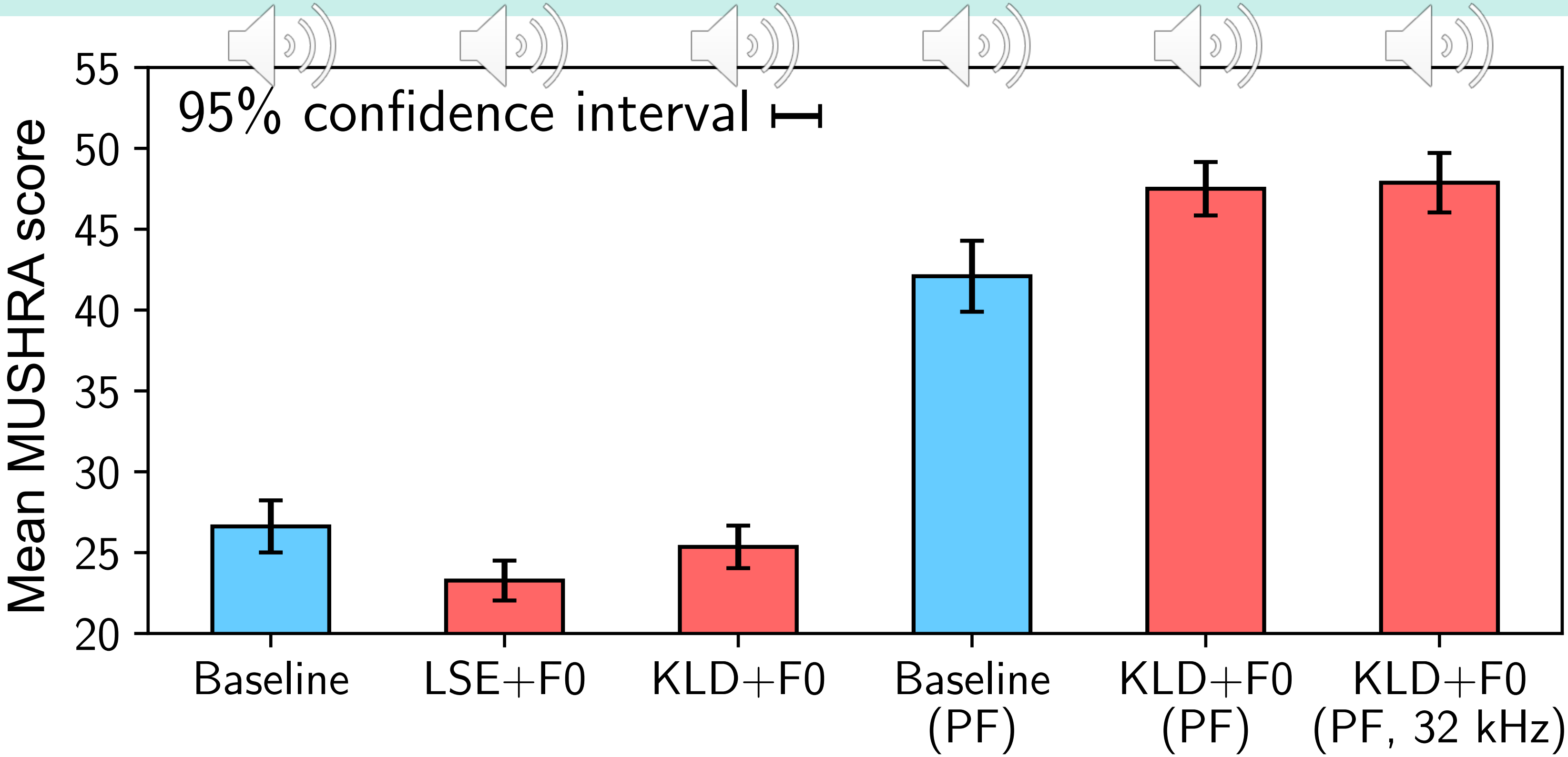
Synthetic spectra (Low-frequency parts)



Synthetic spectra (Low-frequency parts)



Subjective test (MUSHRA, 14 native participants)

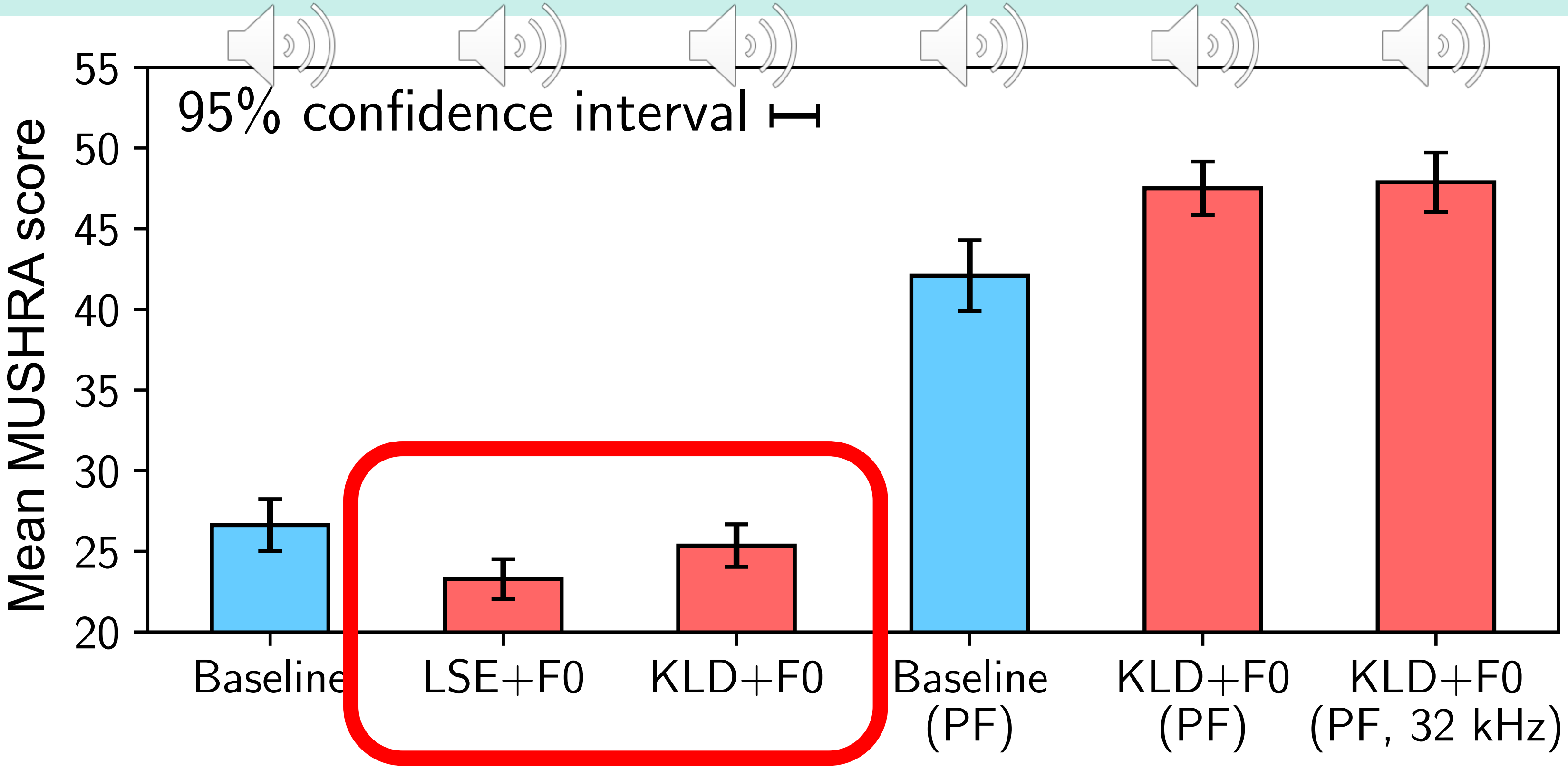


– KLD-based criterion was more appropriate

– Performance of STFT-based systems without post-filtering was insufficient

– The proposed systems with post-filtering outperformed the conventional DNN-based synthesizer

Subjective test (MUSHRA, 14 native participants)

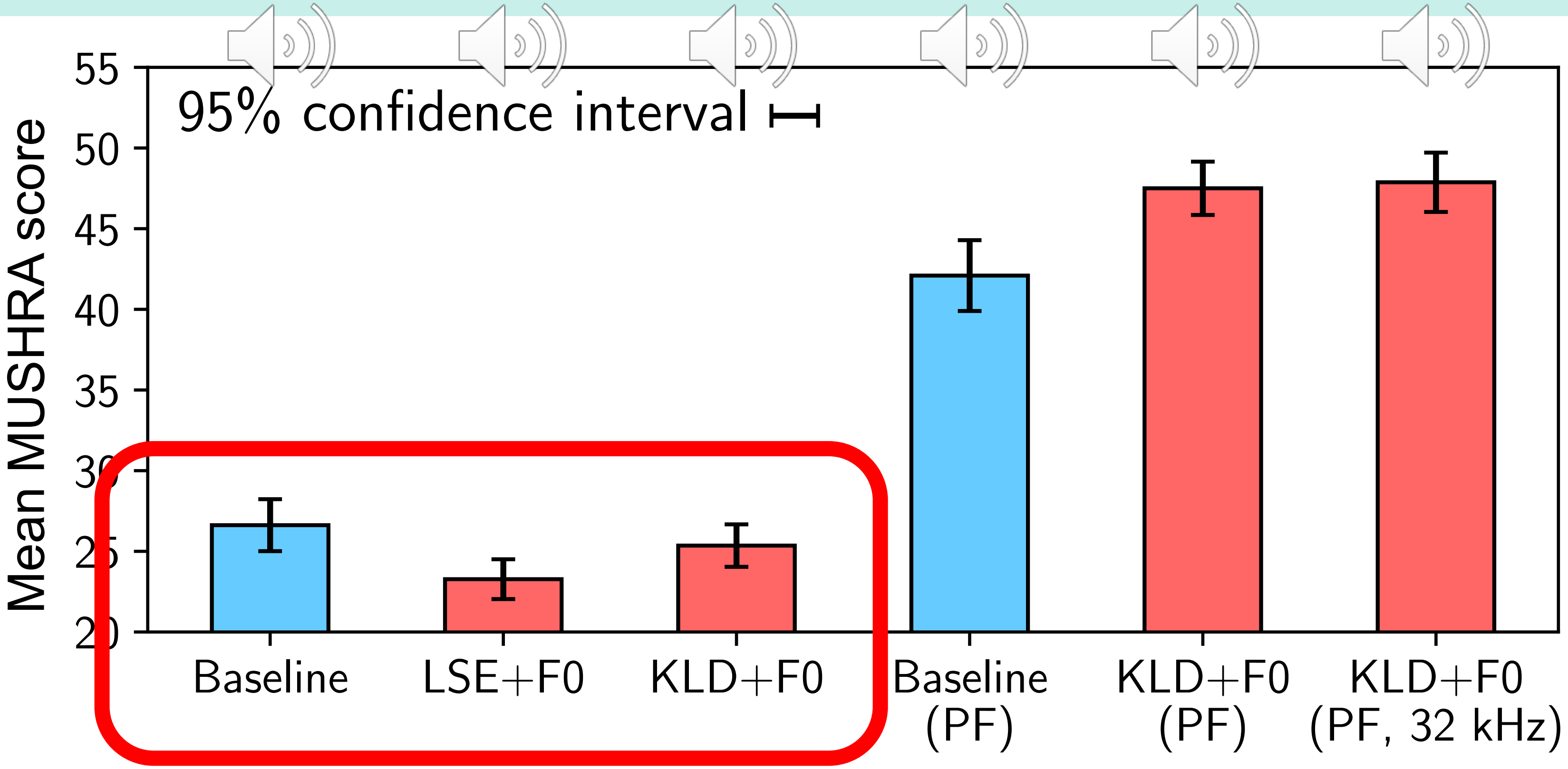


– KLD-based criterion was more appropriate

– Performance of STFT-based systems without post-filtering was insufficient

– The proposed systems with post-filtering outperformed the conventional DNN-based synthesizer

Subjective test (MUSHRA, 14 native participants)

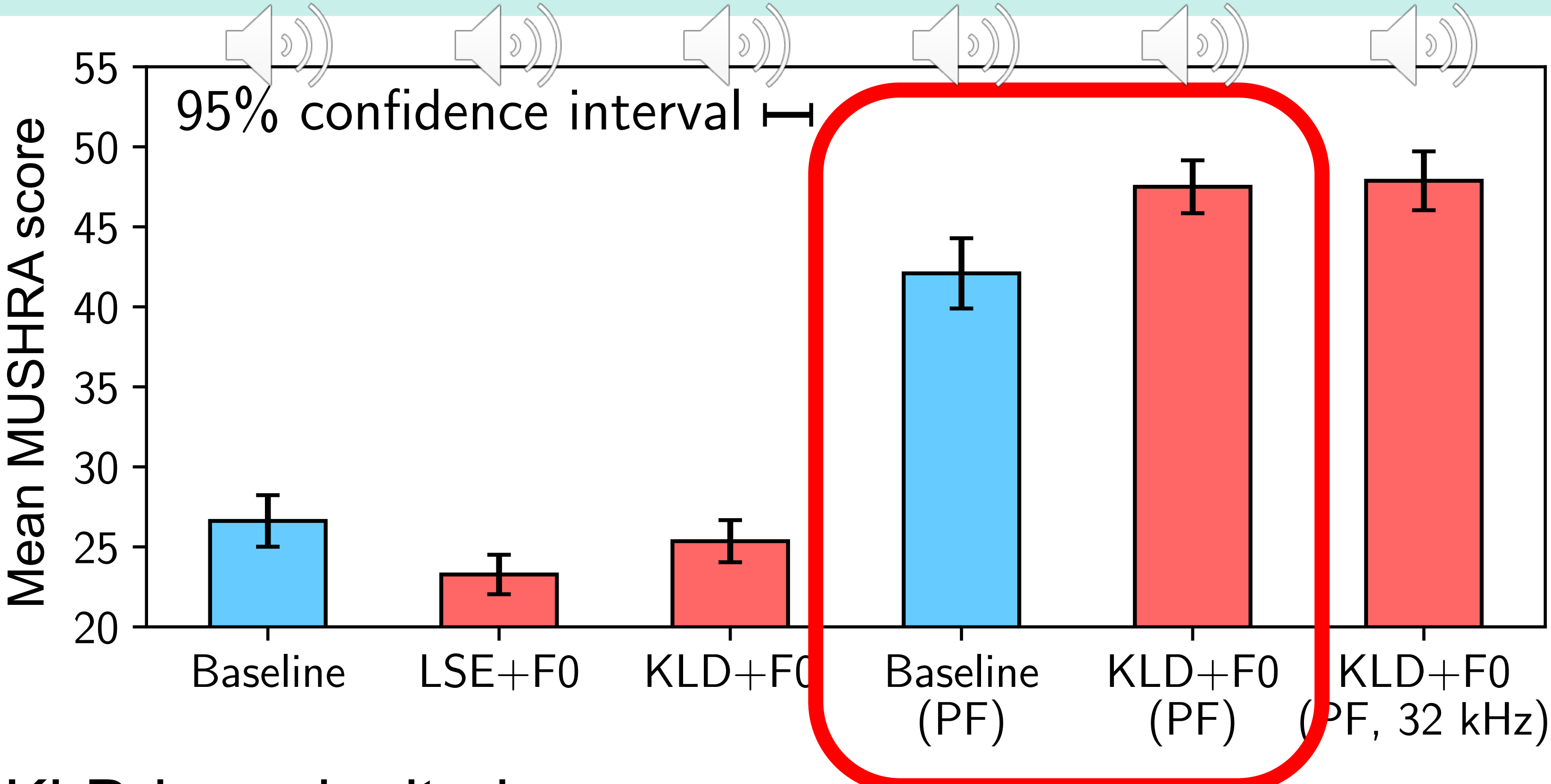


– KLD-based criterion was more appropriate

– Performance of STFT-based systems without post-filtering was insufficient

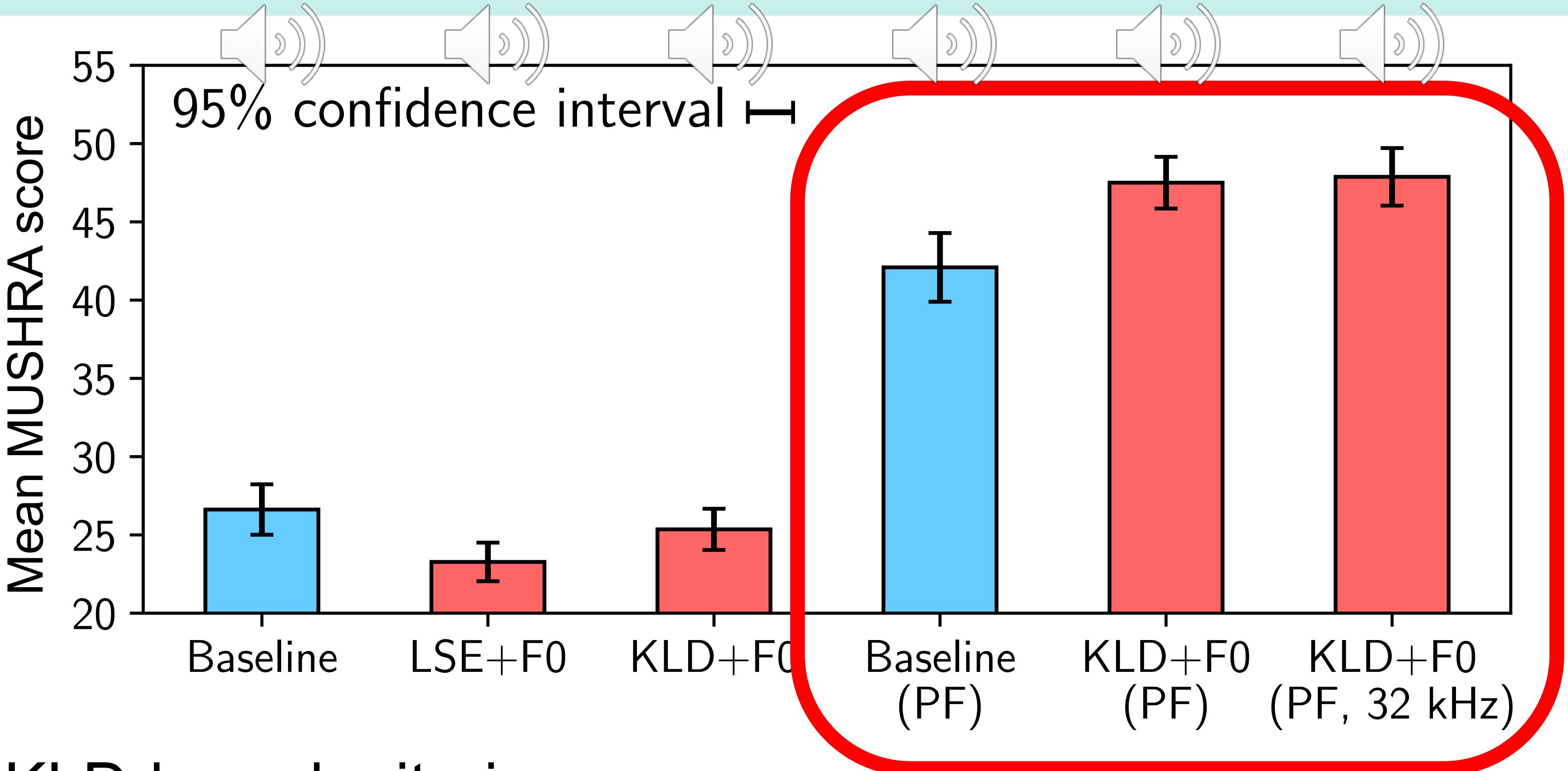
– The proposed systems with post-filtering outperformed the conventional DNN-based synthesizer

Subjective test (MUSHRA, 14 native participants)



- KLD-based criterion was more appropriate
- Performance of STFT-based systems without post-filtering was insufficient
- The proposed systems with post-filtering outperformed the conventional DNN-based synthesizer

Subjective test (MUSHRA, 14 native participants)



– KLD-based criterion was more appropriate

– Performance of STFT-based systems without post-filtering was insufficient

– The proposed systems with post-filtering outperformed the conventional DNN-based synthesizer

Conclusion

Direct modeling of frequency spectra

Waveform generation based on phase recovery

- These approaches were effective
  1. The use of F0 information as well as linguistic features
  2. KLD-based objective criterion
  3. Post-filtering of predicted STFT

Future work

- Wed-P-8-4-6: GAN-based post-filtering for STFT
- Phase modeling
- STFT-conditioned WaveNet, SampleRNN

Synthetic samples

Baseline	
Baseline+PF	
LSE+F0	
KLD+F0	
KLD+F0+PF	
KLD+F0+PF (32kHz)	

Details of KLD

Kullback-Leibler divergence (KLD)

– Representing parameters of Poisson distribution
  - Mean and variance are same

– Derivative

General

$$E_{KL} = \sum_{d=1}^{D} o_{t,d} \log \frac{o_{t,d}}{y_{t,d}} - o_{t,d} + y_{t,d},$$

$$\frac{\partial E_{KL}}{\partial y_{t,d}} = 1 - \frac{o_{t,d}}{y_{t,d}},$$

Our work

$$\begin{aligned} E_{KL} &= \sum_{d=1}^{D} o_{t,d} \log \frac{o_{t,d}}{\tilde{y}_{t,d}} - o_{t,d} + \tilde{y}_{t,d}, \\ &= \sum_{d=1}^{D} o_{t,d} \log \frac{o_{t,d}}{s_d y_{t,d} + b_d} - o_{t,d} + s_d y_{t,d} + b_d, \end{aligned}$$

$$\frac{\partial E_{KL}}{\partial y_{t,d}} = s_d \left(1 - \frac{o_{t,d}}{s_d y_{t,d} + b_d}\right),$$