Can Knowledge of End-to-end Text-to-speech Models Improve MIDI-to-audio Synthesis Systems?

Xuan Shi[1], Erica Cooper[2], Xin Wang[2], Junichi Yamagishi[2], Shrikanth Narayanan[1]

[1] University of Southern California
[2] National Institute of Informatics (NII), Japan

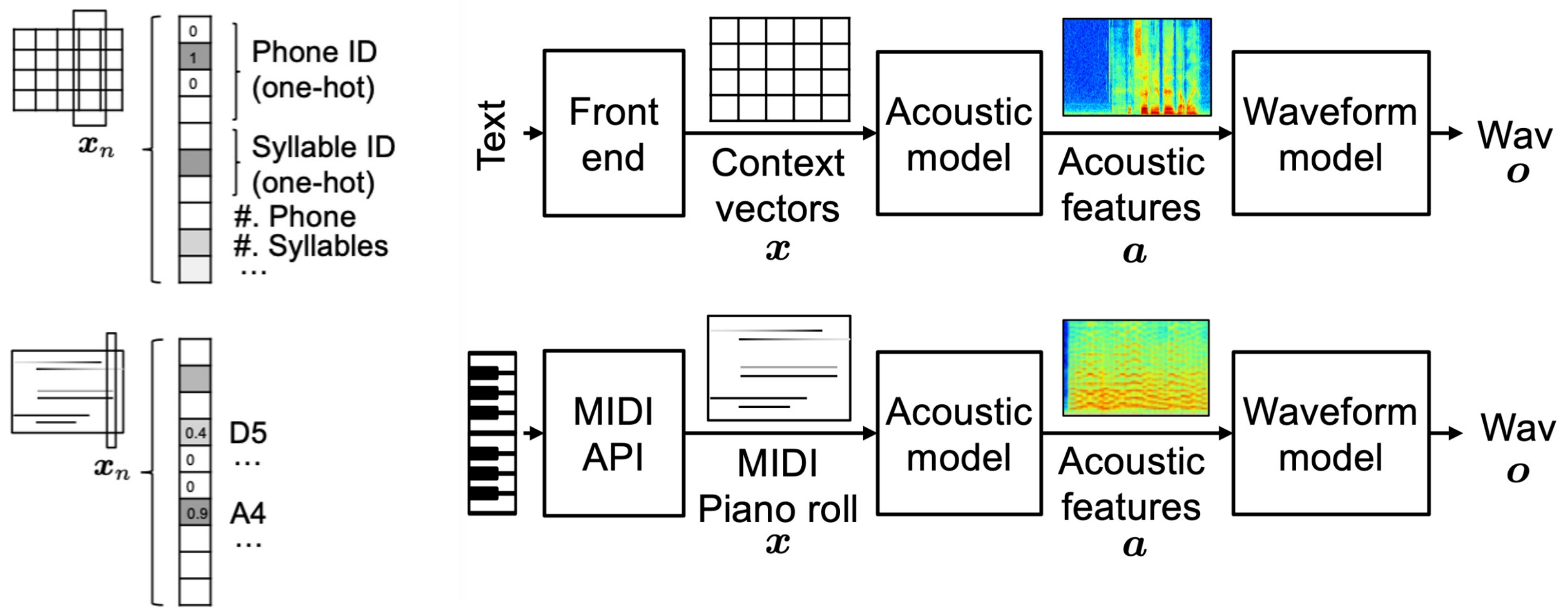
Motivation

- MIDI-to-Audio methods
  - Conventional
    - FluidSynth: pre-recording and resampling audio for synthesis
    - Pianoteq: constructing physical model for audio synthesis
    - …
  - Neural Network based
    - MIDI-DDSP: Multiple stages feature generation: Expression, Synthesis, and DDSP
    - Deep Performer: decomposing note attributes and synthesis music
    - …

[1] Wu, Yusong, et al. "MIDI-DDSP: Detailed control of musical performance via hierarchical modeling." *International Conference on Learning Representations (ICLR),* 2021.
[2] Dong, Hao-Wen, et al. "Deep performer: Score-to-audio music performance synthesis." *IEEE International Conference on Acoustics, Spee and Signal Processing (ICASSP)*. IEEE, 2022.

Motivation

- Text-to-Speech and MIDI-to-Audio

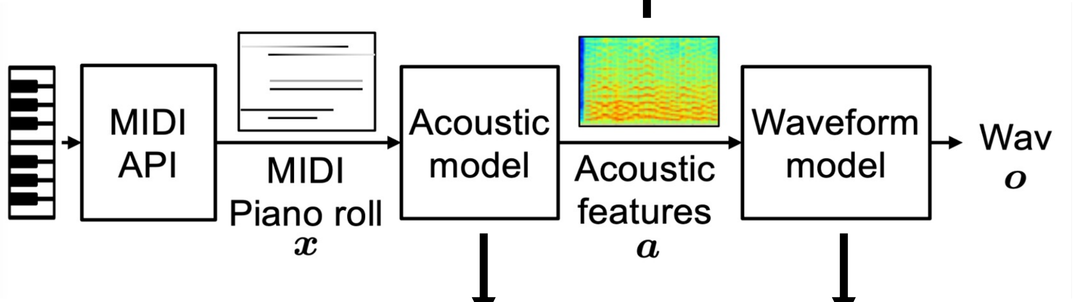


[3] Erica Cooper, Xin Wang, and Junichi Yamagishi, “Text-to-speech synthesis techniques for MIDI-to-audio synthesis.” SSW 11 (2021): 130–135

Motivation

- Previous work

MIDI filter-bank features



Tacotron2

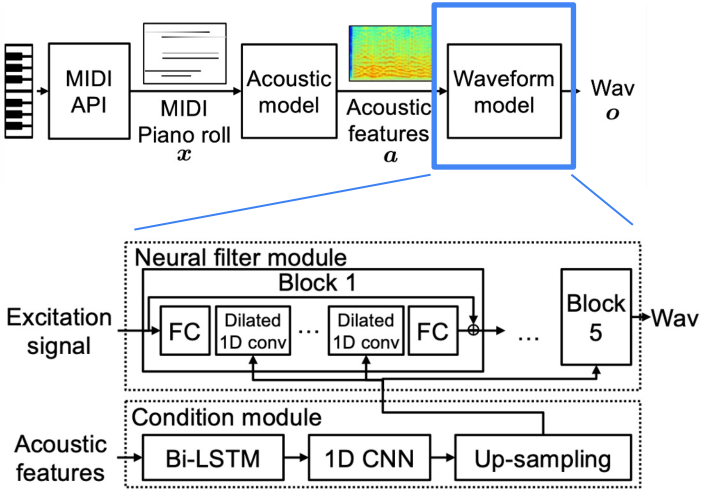
Neural source-filter (NSF) model

- Synthesized audio quality is limited — Q1: How to improve the synthesized audio **quality**
- Training & synthesis are time consuming — Q2: How to make the synthesis **efficient**

[3] Erica Cooper, Xin Wang, and Junichi Yamagishi, “Text-to-speech synthesis techniques for MIDI-to-audio synthesis.” SSW 11 (2021): 130–135
[4] Jonathan Shen, et.al. “Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions”, ICASSP 2018
[5] Wang, Xin, et.al.. "Neural source-filter waveform models for statistical parametric speech synthesis." *IEEE/ACM TASLP* 2019

Methods to improve synthesized audio **quality**

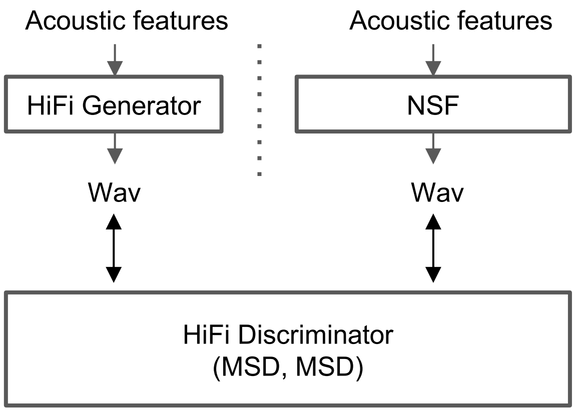
- Waveform Model with GAN



Previous model: NSF

Method 1:
HiFi-GAN

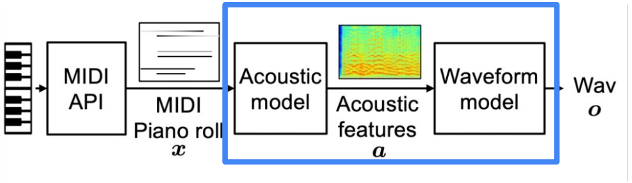
Method 2:
NSF + HiFi-GAN discri.



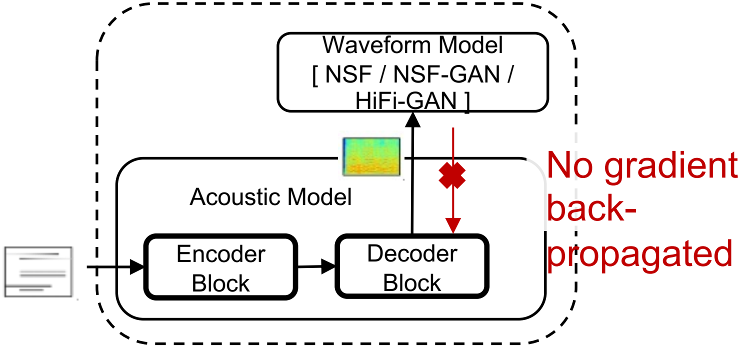
[5] Wang, Xin, et.al.. "Neural source-filter waveform models for statistical parametric speech synthesis." *IEEE/ACM TASLP* 2019
[6] Kong, Jungil, Jaehyeon Kim, and Jaekyoung Bae. "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesi" *Advances in Neural Information Processing Systems* 33 (2020): 17022-17033.

Methods to improve synthesized audio **quality**

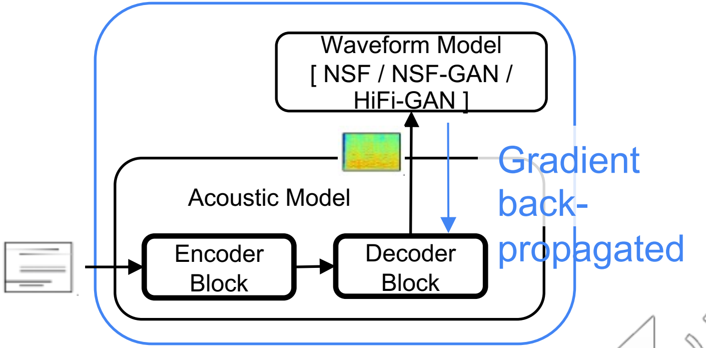
- Joint training of acoustic and waveform models



Previous model: separate training



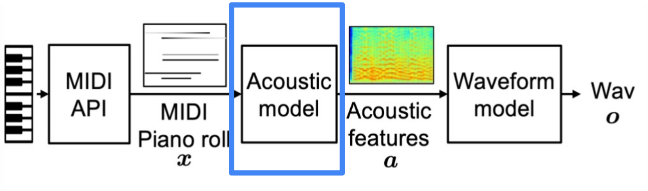
Improved model: joint training



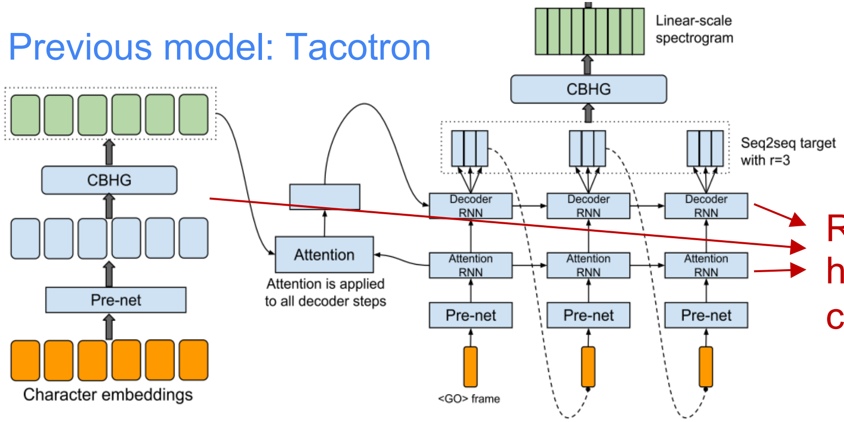
[4] Jonathan Shen, et.al. “Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions”, ICASSP 2018
[7] Kim, Jaehyeon, et.al.. "Conditional variational autoencoder with adversarial learning for end-to-end Text-to-speech." ICML, 2021.

Methods to improve synthesis **efficiency**

- Acoustic Model based on Transformer

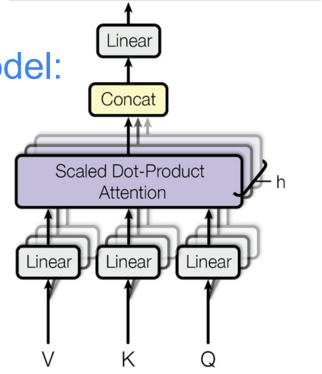


Previous model: Tacotron



RNN has high time complexity

Improved model: Transformer



Self-Attention

[8] Yasuda et al. "Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language." *ICASSP*, 2019.
[9] Li, Naihan, et al. "Neural speech synthesis with transformer network." Proc. *AAAI*. Vol. 33. No. 01. 2019.

Experiments – Conditions & Evaluation

- Database - MAESTRO
  - Train/Validation/Test: 159/19/20 hours
  - MIDI and audio alignment: < 3ms
  - Resampled to 24 kHz
  - Segmented to 800-frame pieces, around 10 seconds

- Subjective – crowdsourced subjective listening test
  - Mean Opinion Score (MOS), 1-5, the higher the better
  - **229** non-professional listeners
  - 510 samples per system are rated

- Objective (see results in the paper)
  - L2 distance on MIDI-Spectrogram, Chroma, Cross entropy on F0

[10] Hawthorne, Curtis, et al. "Enabling factorized piano music modeling and generation with the MAESTRO dataset." In International Conference on Learning Representations, 2019.

Experiments – Systems

- Baseline: Fluidsynth, Pianoteq
- Reference with "perfect acoustic model":
  - abs-*-* : use acoustic features extracted from test set audios

- Systems
  - Acoustic model: Tacotron or Transformer
  - Waveform model: NSF, NSF-GAN, HiFi-GAN
  - Training strategy: separate or joint training

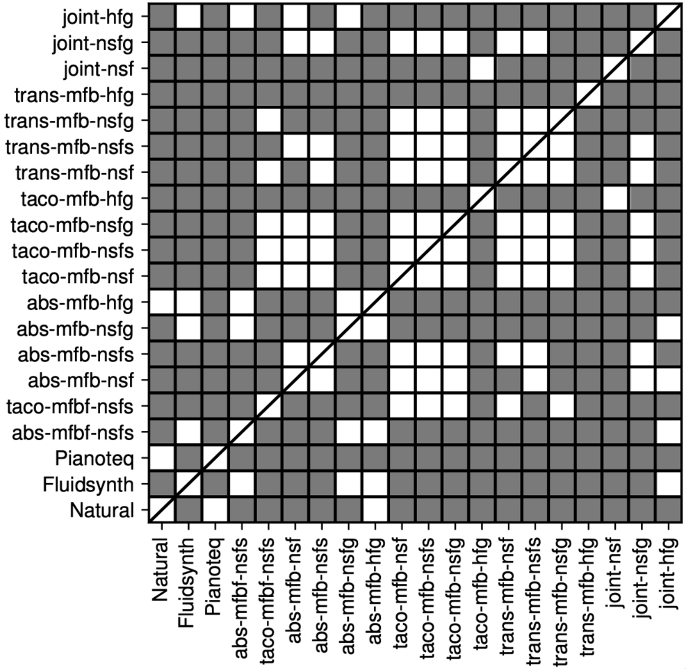
System ID	Acoustic model	Acoustic feature	Wave. model	Joint train
Natural	-	-	-	-
Software-based baselines				
Fluidsynth	Sample-based MIDI-to-audio s.w.			
Pianoteq	Physical-model MIDI-to-audio s.w.			
Synthesis system trained on flawed MIDI spectrogram				
abs-mfbf-nsfs	-	midi-fb-f	NSF [1]	-
taco-mfbf-nsfs	taco	midi-fb-f	NSF [1]	-
Waveform model trained on refined MIDI spectrogram				
abs-mfb-nsfs	-	midi-fb	NSF [1]	-
abs-mfb-nsf	-	midi-fb	NSF	-
abs-mfb-nsfg	-	midi-fb	NSF-GAN	-
abs-mfb-hfg	-	midi-fb	HiFi-GAN	-
Acoustic model trained on refined MIDI spectrogram				
taco-mfb-nsfs	taco	midi-fb	NSF [1]	-
taco-mfb-nsf	taco	midi-fb	NSF	-
taco-mfb-nsfg	taco	midi-fb	NSF-GAN	-
taco-mfb-hfg	taco	midi-fb	HiFi-GAN	-
trans-mfb-nsfs	trans	midi-fb	NSF [1]	-
trans-mfb-nsf	trans	midi-fb	NSF	-
trans-mfb-nsfg	trans	midi-fb	NSF-GAN	-
trans-mfb-hfg	trans	midi-fb	HiFi-GAN	-
Joint training of acoustic and waveform model				
joint-nsf	trans	midi-fb	NSF	✓
joint-nsfg	trans	midi-fb	NSF-GAN	✓
joint-hfg	trans	midi-fb	HiFi-GAN	✓

Note for join training:
- Stage 1: Pre-Train: separately train acoustic model and waveform model
- Stage 2: Joint-Train: load pre-trained model weights, jointly train acoustic & waveform model

Experiments – Subjective Evaluation Results

Table 1. Experimental systems and evaluation results.

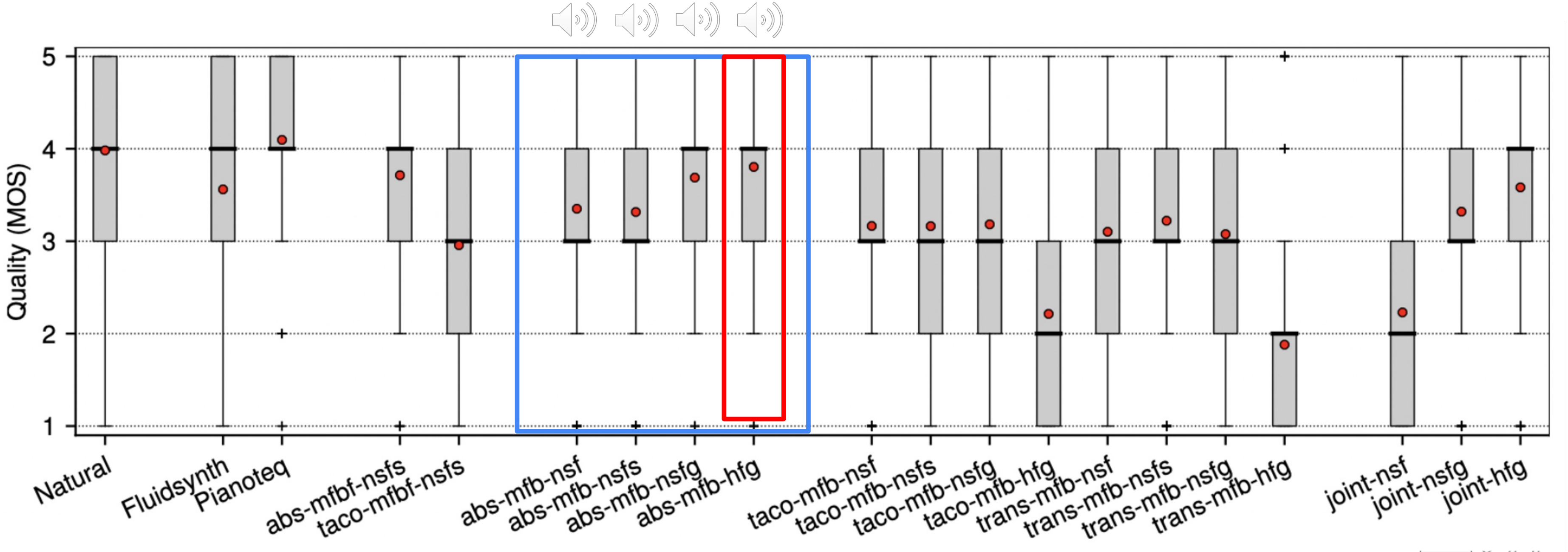
System ID	Acoustic model	Acoustic feature	Wave. model	Joint train	Obj. Eval. Pitch	Obj. Eval. Chroma	Obj. Eval. Spec	MOS (mean)
Natural	-	-	-	-	-	-	-	3.98
Software-based baselines								
`Fluidsynth`	Sample-based MIDI-to-audio s.w.				1.00	0.33	13.95	3.56
`Pianoteq`	Physical-model MIDI-to-audio s.w.				0.92	0.32	12.16	4.10
Synthesis system trained on flawed MIDI spectrogram								
`abs-mfbf-nsfs`	-	midi-fb-f	NSF [1]	-	1.01	0.31	6.60	3.71
`taco-mfbf-nsfs`	taco	midi-fb-f	NSF [1]	-	1.18	0.37	9.65	2.95
Waveform model trained on refined MIDI spectrogram								
`abs-mfb-nsfs`	-	midi-fb	NSF [1]	-	1.31	0.38	5.72	3.31
`abs-mfb-nsf`	-	midi-fb	NSF	-	1.37	0.39	7.20	3.35
`abs-mfb-nsfg`	-	midi-fb	NSF-GAN	-	1.26	0.34	5.14	3.69
`abs-mfb-hfg`	-	midi-fb	HiFi-GAN	-	1.16	0.31	4.69	3.80
Acoustic model trained on refined MIDI spectrogram								
`taco-mfb-nsfs`	taco	midi-fb	NSF [1]	-	1.19	0.37	9.70	3.16
`taco-mfb-nsf`	taco	midi-fb	NSF	-	1.29	0.40	11.78	3.16
`taco-mfb-nsfg`	taco	midi-fb	NSF-GAN	-	1.11	0.35	9.09	3.18
`taco-mfb-hfg`	taco	midi-fb	HiFi-GAN	-	1.58	0.56	10.07	2.21
`trans-mfb-nsfs`	trans	midi-fb	NSF [1]	-	1.33	0.41	9.41	3.22
`trans-mfb-nsf`	trans	midi-fb	NSF	-	1.42	0.44	10.94	3.10
`trans-mfb-nsfg`	trans	midi-fb	NSF-GAN	-	1.27	0.40	9.15	3.08
`trans-mfb-hfg`	trans	midi-fb	HiFi-GAN	-	1.83	0.60	9.95	1.88
Joint training of acoustic and waveform model								
`joint-nsf`	trans	midi-fb	NSF	✓	1.59	0.47	16.39	2.23
`joint-nsfg`	trans	midi-fb	NSF-GAN	✓	1.12	0.38	9.09	3.32
`joint-hfg`	trans	midi-fb	HiFi-GAN	✓	1.10	0.38	9.14	3.58



Results of two-sided Mann-Whitney U test with Holm-Bonferron correction. Grey block indicates statistically significant difference at α = 0.05

Experiments – Subjective Evaluation Results

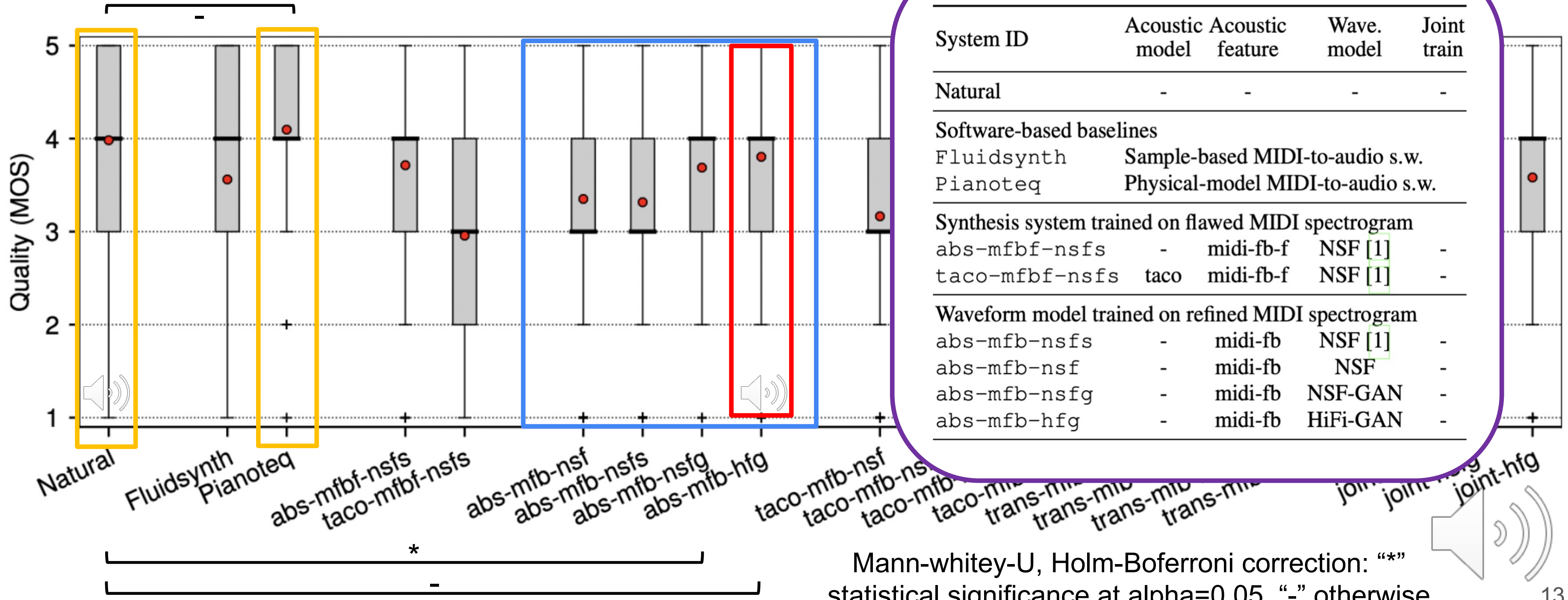
- Analysis-by-synthesis systems comparison



Boxplots of MOS per system. Red dots denote mean of MOS

Experiments – Subjective Evaluation Results

- Analysis-by-synthesis systems comparison

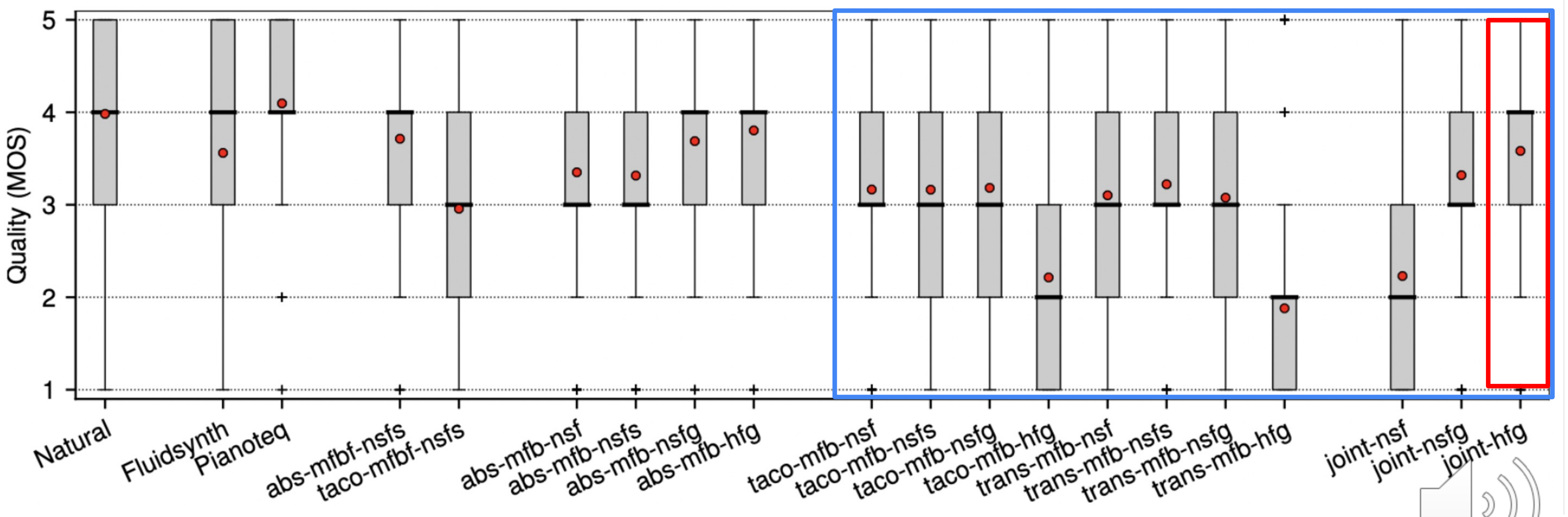


System ID	Acoustic model	Acoustic feature	Wave. model	Joint train
Natural	-	-	-	-
Software-based baselines				
`Fluidsynth`	Sample-based MIDI-to-audio s.w.			
`Pianoteq`	Physical-model MIDI-to-audio s.w.			
Synthesis system trained on flawed MIDI spectrogram				
`abs-mfbf-nsfs`	-	midi-fb-f	NSF [1]	-
`taco-mfbf-nsfs`	taco	midi-fb-f	NSF [1]	-
Waveform model trained on refined MIDI spectrogram				
`abs-mfb-nsfs`	-	midi-fb	NSF [1]	-
`abs-mfb-nsf`	-	midi-fb	NSF	-
`abs-mfb-nsfg`	-	midi-fb	NSF-GAN	-
`abs-mfb-hfg`	-	midi-fb	HiFi-GAN	-

Mann-whitey-U, Holm-Boferroni correction: “*” statistical significance at alpha=0.05, “-” otherwise

Experiments – Subjective Evaluation Results

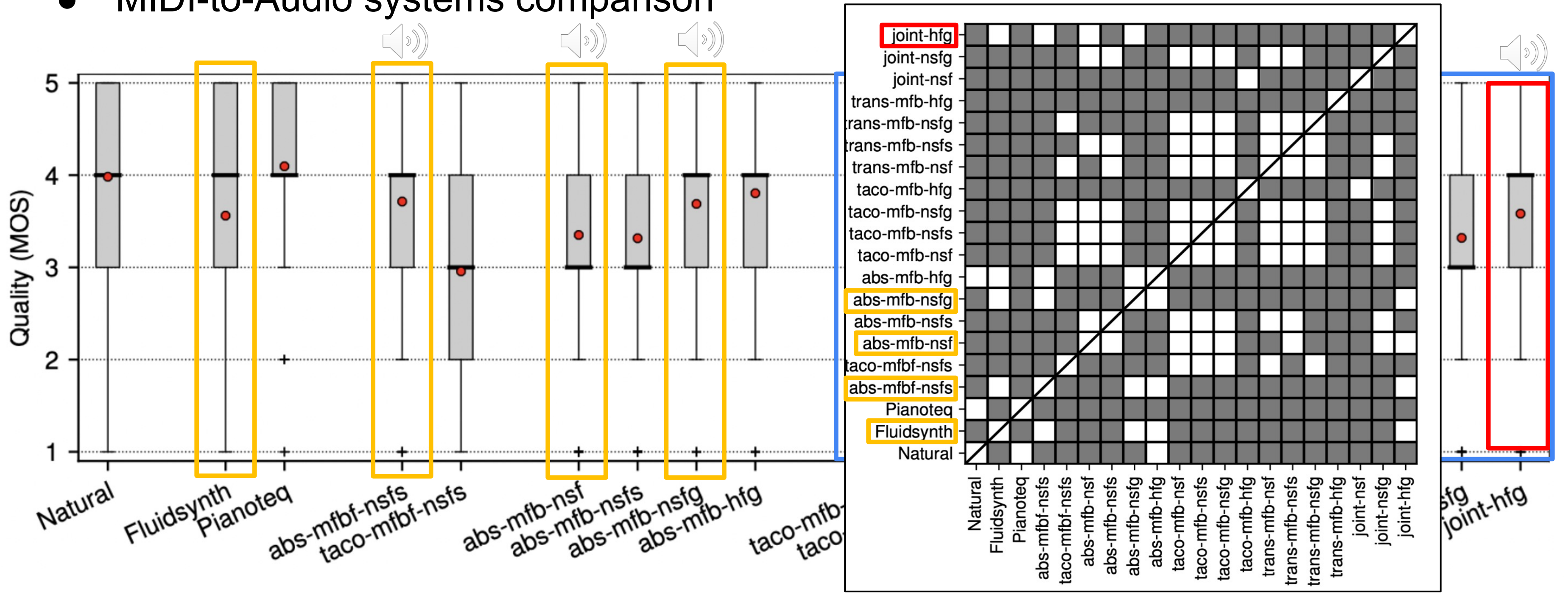
- MIDI-to-Audio systems comparison



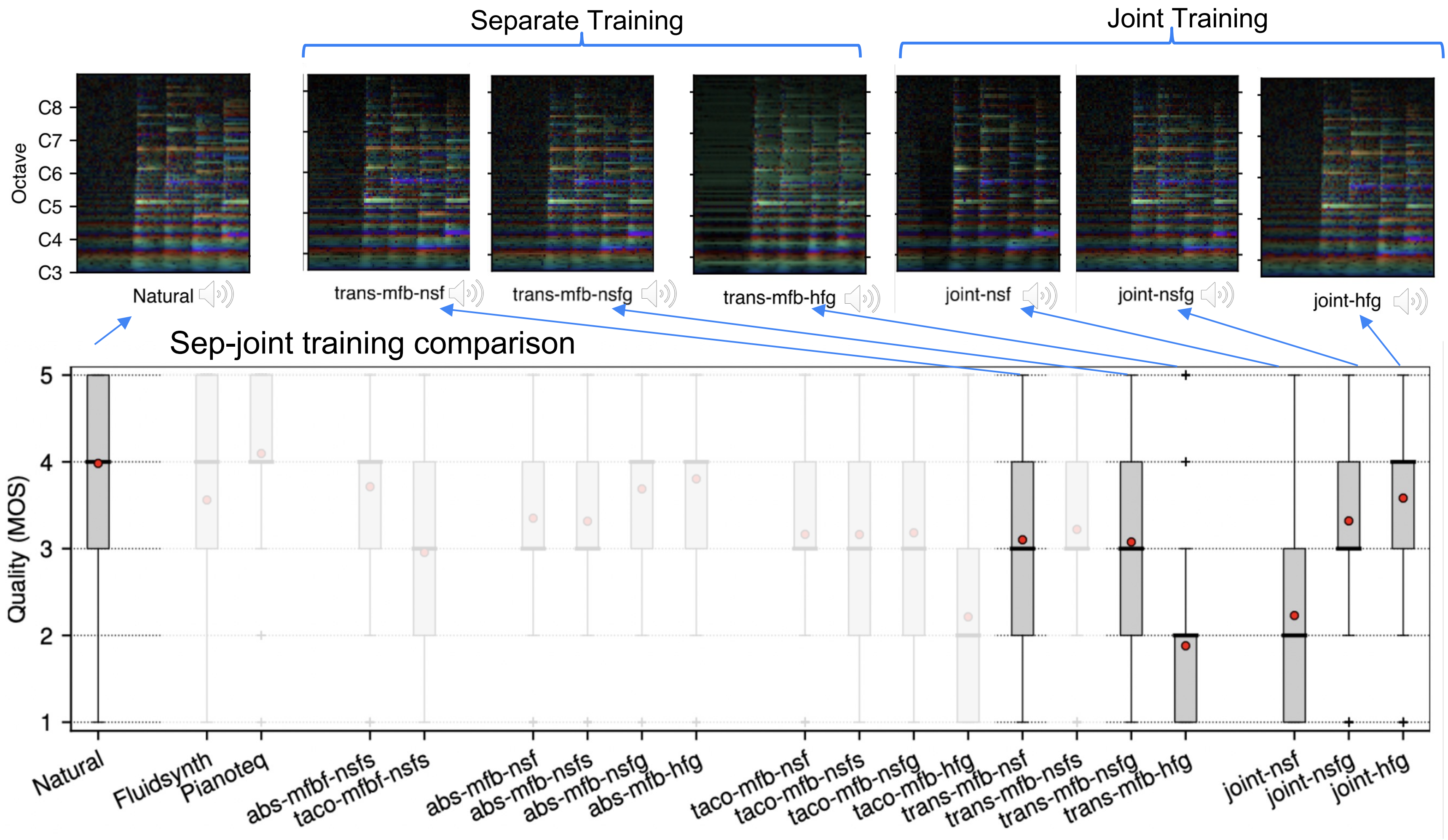
Boxplots of MOS per system. Red dots denote mean of MOS

Experiments – Subjective Evaluation Results

- MIDI-to-Audio systems comparison



Boxplots of MOS per system. Red dots denote mean of MOS



Separate Training
Joint Training
Octave
C8
C7
C6
C5
C4
C3
Natural
trans-mfb-nsf
trans-mfb-nsfg
trans-mfb-hfg
joint-nsf
joint-nsfg
joint-hfg
Sep-joint training comparison
Quality (MOS)
5
4
3
2
1
Natural
Fluidsynth
Pianoteq
abs-mfbf-nsfs
taco-mfbf-nsfs
abs-mfb-nsf
abs-mfb-nsfs
abs-mfb-nsfg
abs-mfb-hfg
taco-mfb-nsf
taco-mfb-nsfs
taco-mfb-nsfg
taco-mfb-hfg
trans-mfb-nsf
trans-mfb-nsfs
trans-mfb-nsfg
trans-mfb-hfg
joint-nsf
joint-nsfg
joint-hfg

Conclusion

- Can we improve the **quality** of the synthesized audio? If yes, how?
  - **Yes!**
  - TTS architecture + HiFi-GAN + joint training -> high-fidelity piano music
  - Best midi-to-audio system gets **MOS 3.58**.

- Can we improve the synthesis **efficiency** of the system? If yes, how?
  - **Yes!**
  - Transformer-based acoustic model improves efficiency while keeping performance.

Conclusion

- Can we improve the **quality** of the synthesized audio? If yes, how?
  - **Yes!**
  - TTS architecture + HiFi-GAN + joint training -> high-fidelity piano music
  - Best midi-to-audio system gets **MOS 3.58**.
- Can we improve the synthesis **efficiency** of the system? If yes, how?
  - **Yes!** Transformer-based acoustic model improves efficiency while keeping performance.
- What is the practical impact of the midi-to-audio synthesis?
  - Investigate more areas related to music synthesis, such as timbre transfer, multi-instrument audio synthesis, and performance generation in future work.