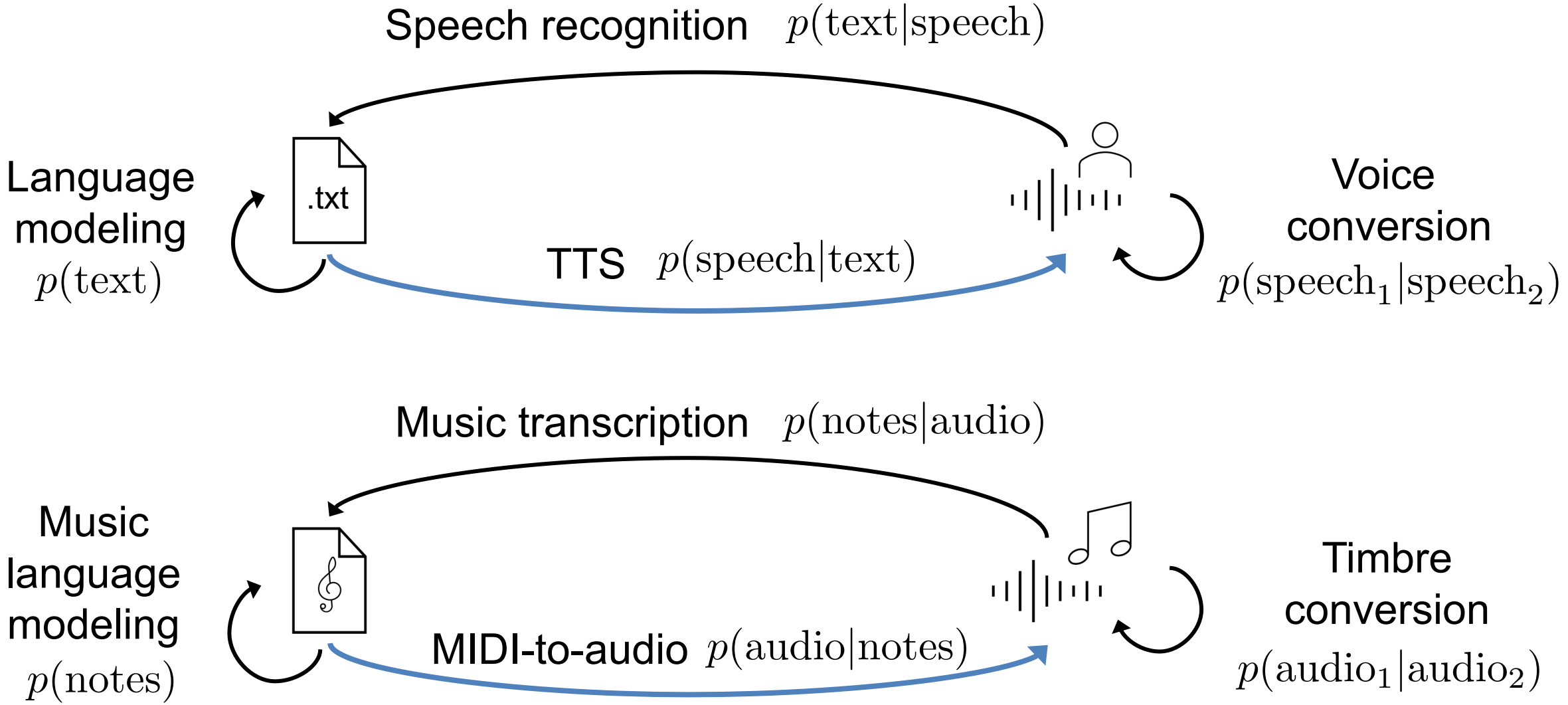
SSW11 Speech Synthesis Workshop
Aug 26-28 2021

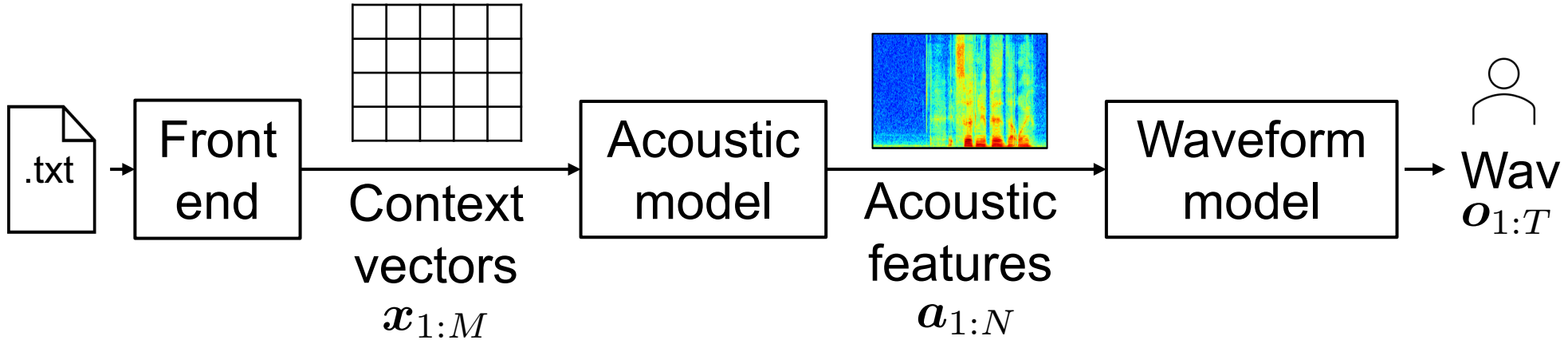
Text-to-Speech Synthesis Techniques for MIDI-to-Audio Synthesis

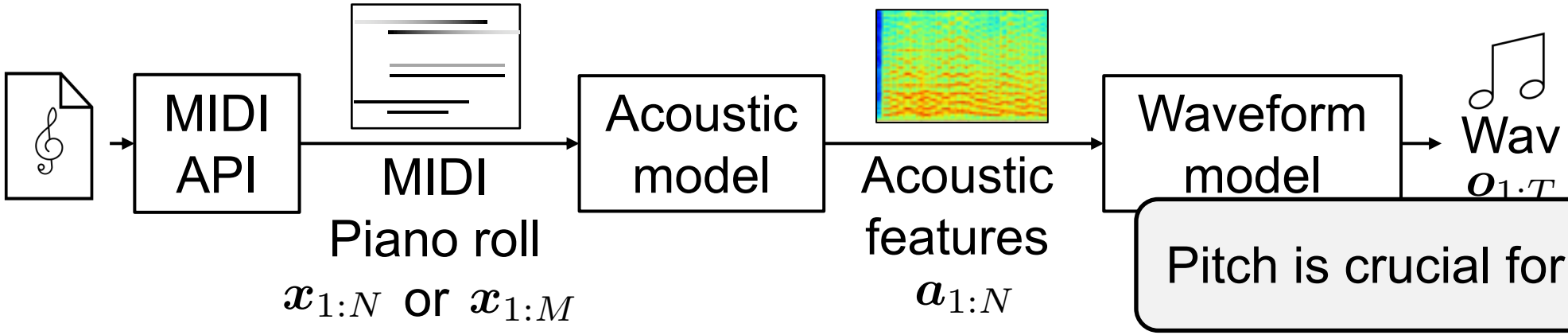
Erica Cooper, Xin Wang, Junichi Yamagishi
National Institute of Informatics, Japan

Motivation


Speech recognition $p(\text{text}|\text{speech})$
Language modeling $p(\text{text})$
.txt
TTS $p(\text{speech}|\text{text})$
Voice conversion $p(\text{speech}_1|\text{speech}_2)$
Music transcription $p(\text{notes}|\text{audio})$
Music language modeling $p(\text{notes})$
MIDI-to-audio $p(\text{audio}|\text{notes})$
Timbre conversion $p(\text{audio}_1|\text{audio}_2)$

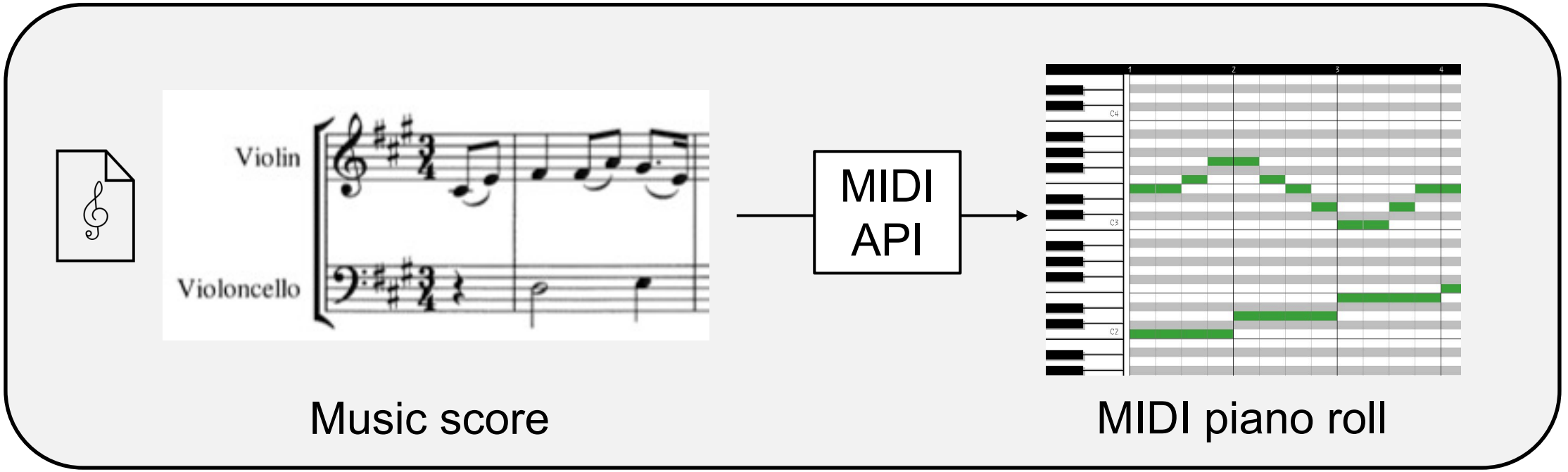
TTS & MIDI-to-Audio Synthesis

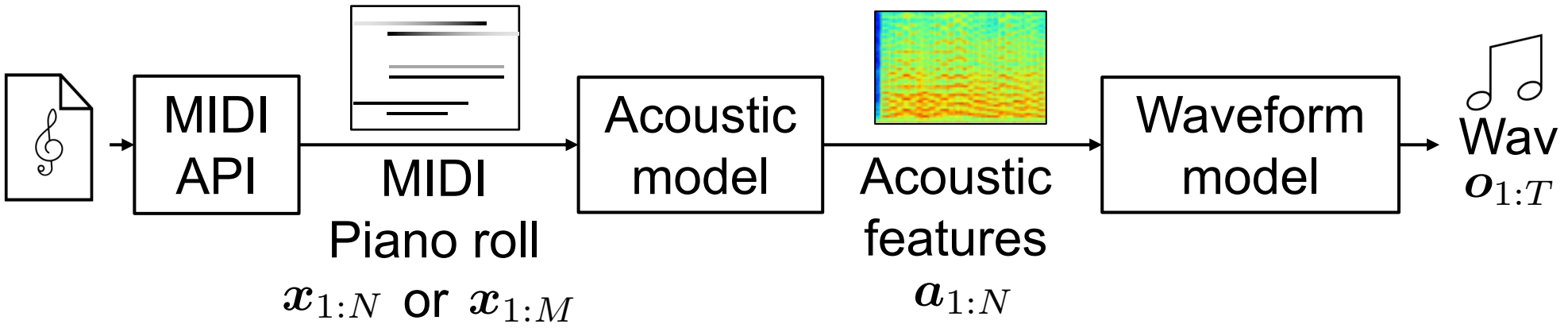




Pitch is crucial for music

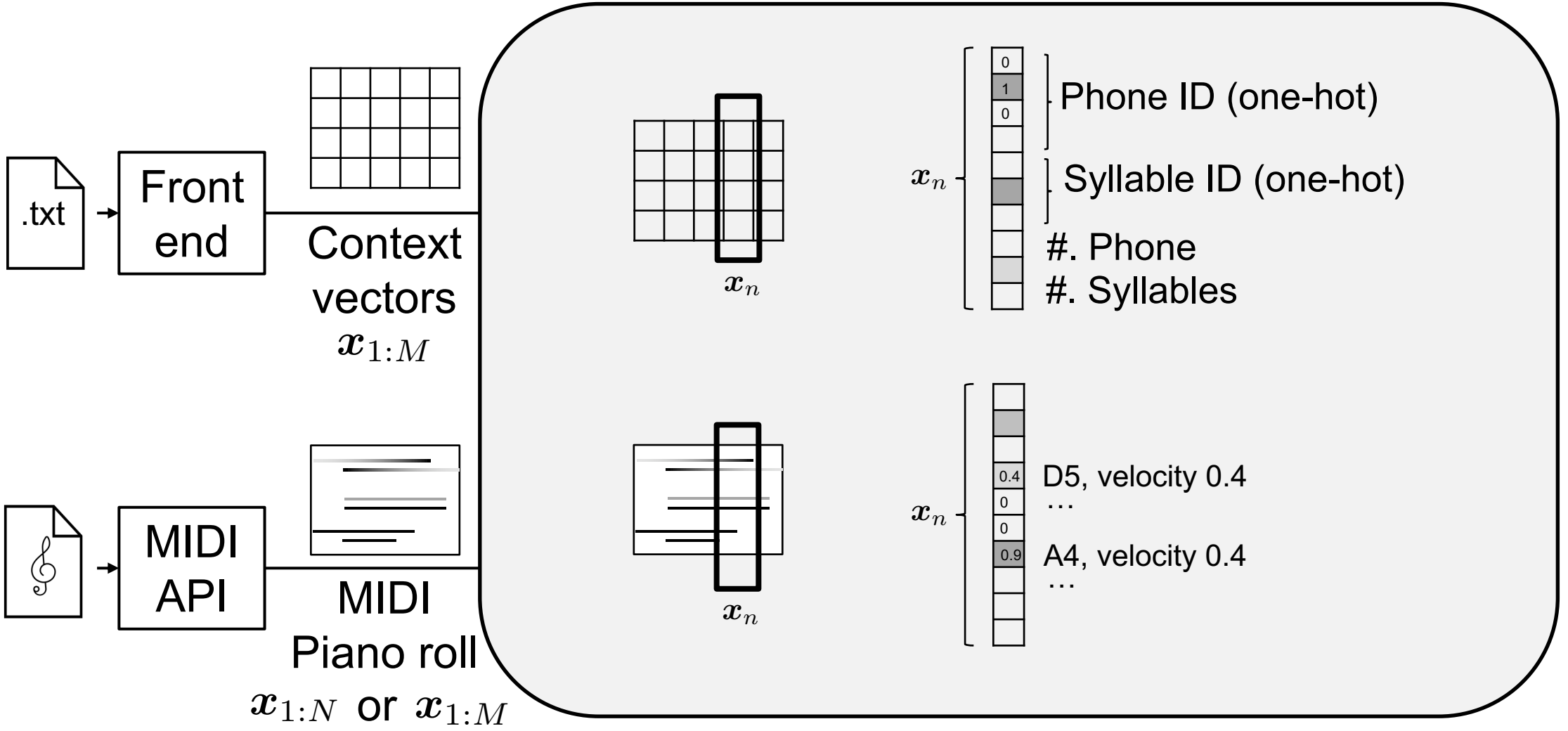
TTS & MIDI-to-Audio Synthesis



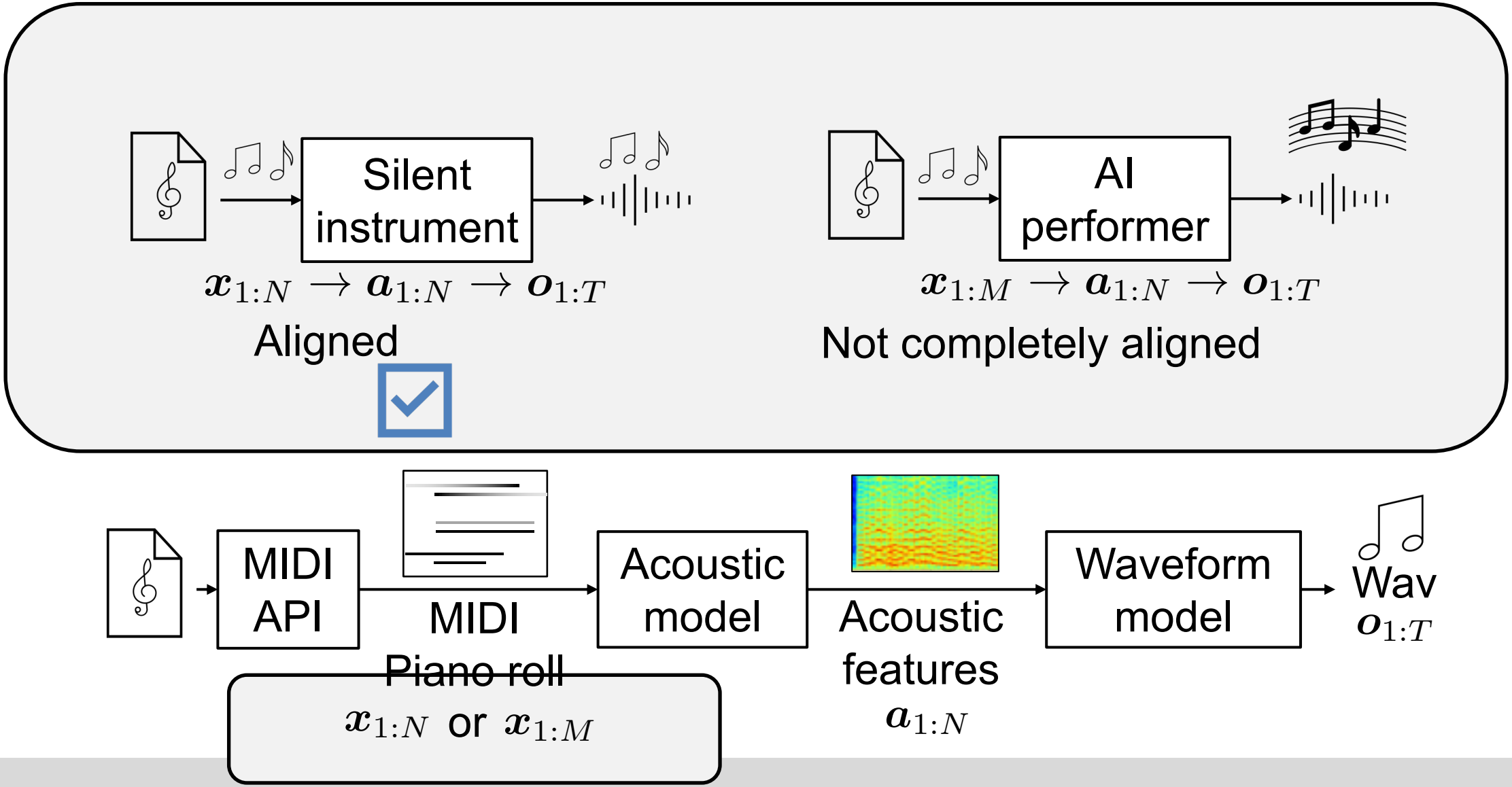


https://commons.wikimedia.org/wiki/File:Computer_music_piano_roll.png

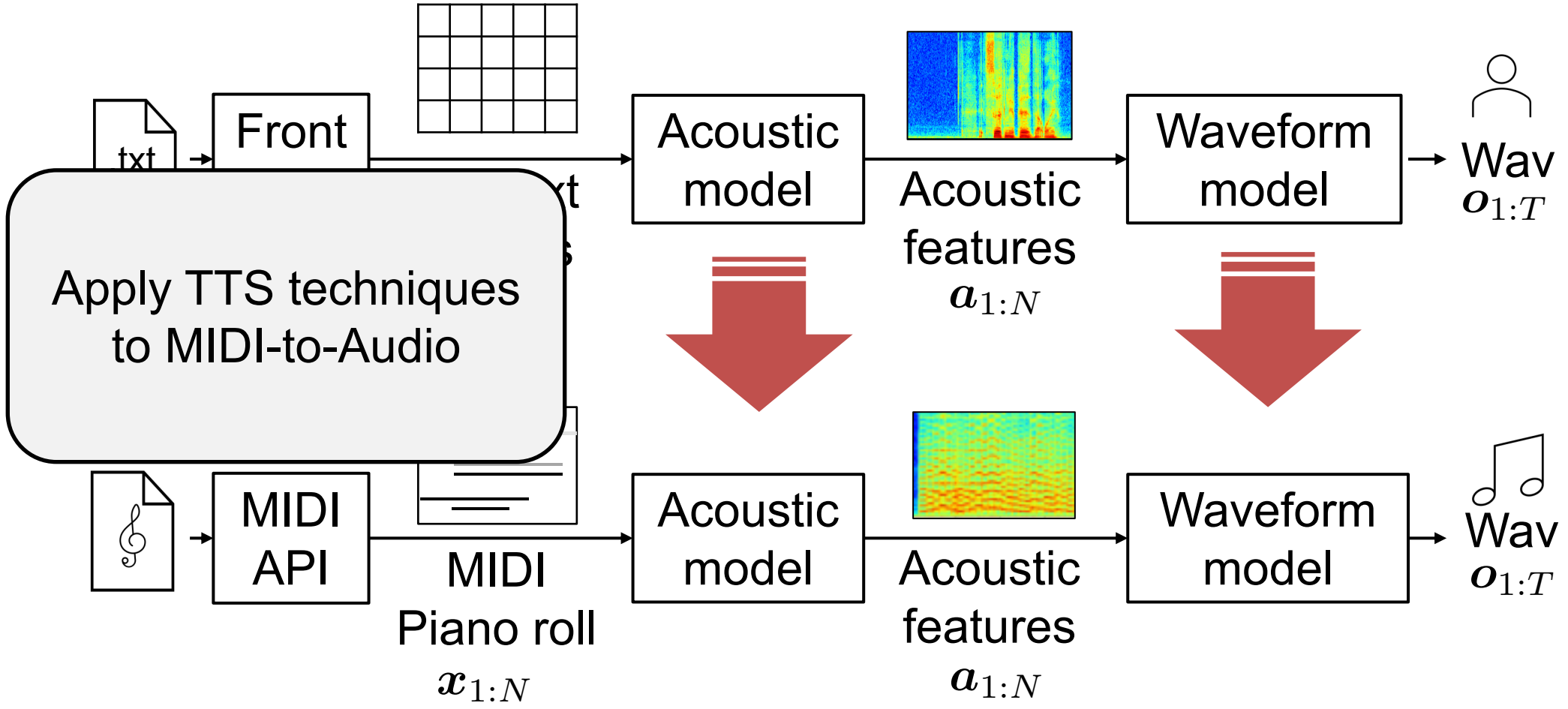
TTS & MIDI-to-Audio Synthesis



TTS & MIDI-to-Audio Synthesis

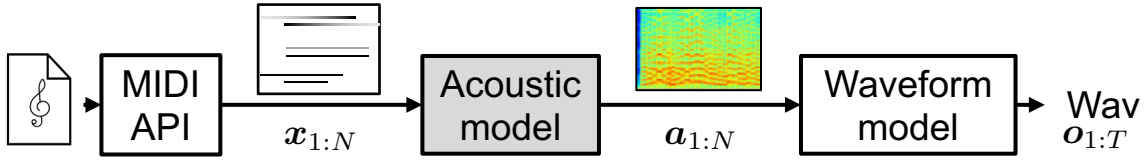


TTS & MIDI-to-Audio Synthesis

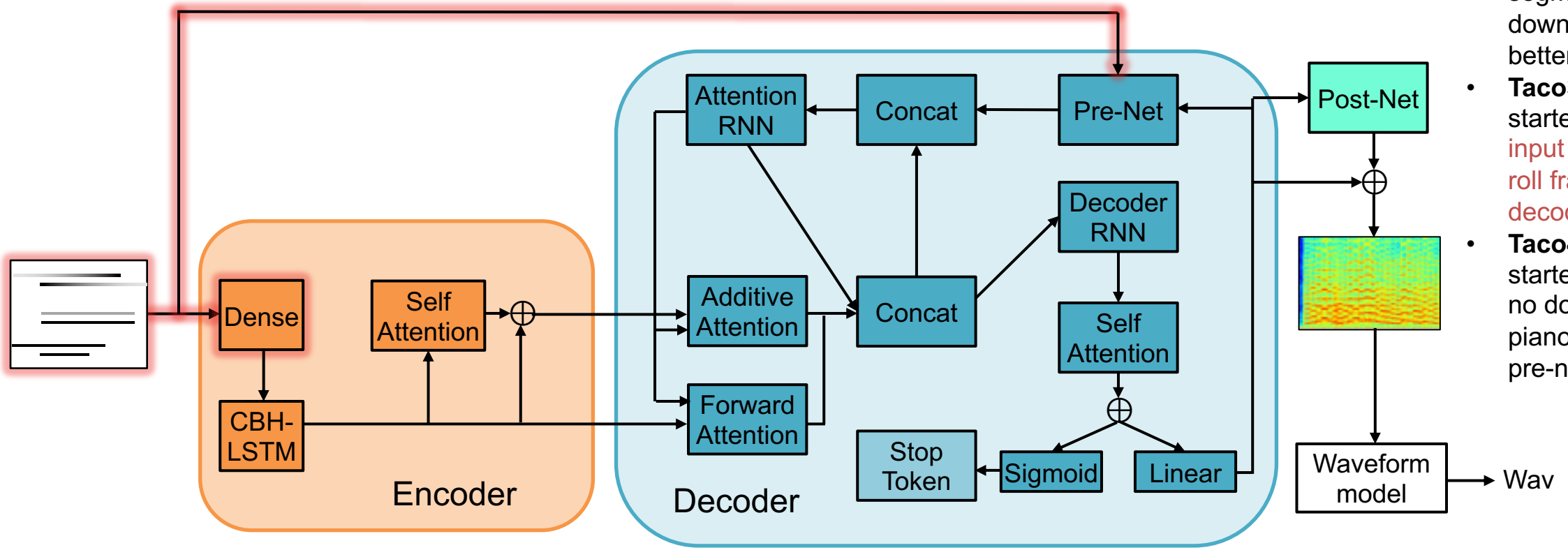


Methods

❑ Acoustic model



1. TTS model: Tacotron (Wang 2017, Yasuda 2019)

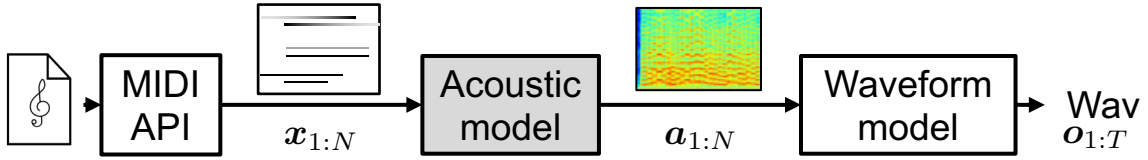


- **Taco2:** 800-frame segments; 4x downsampling for better alignments
- **Taco3:** Warm-started from taco2; input current piano-roll frame at the decoder pre-net
- **Taco4:** Warm-started from taco2; no downsampling or piano-roll input to pre-net

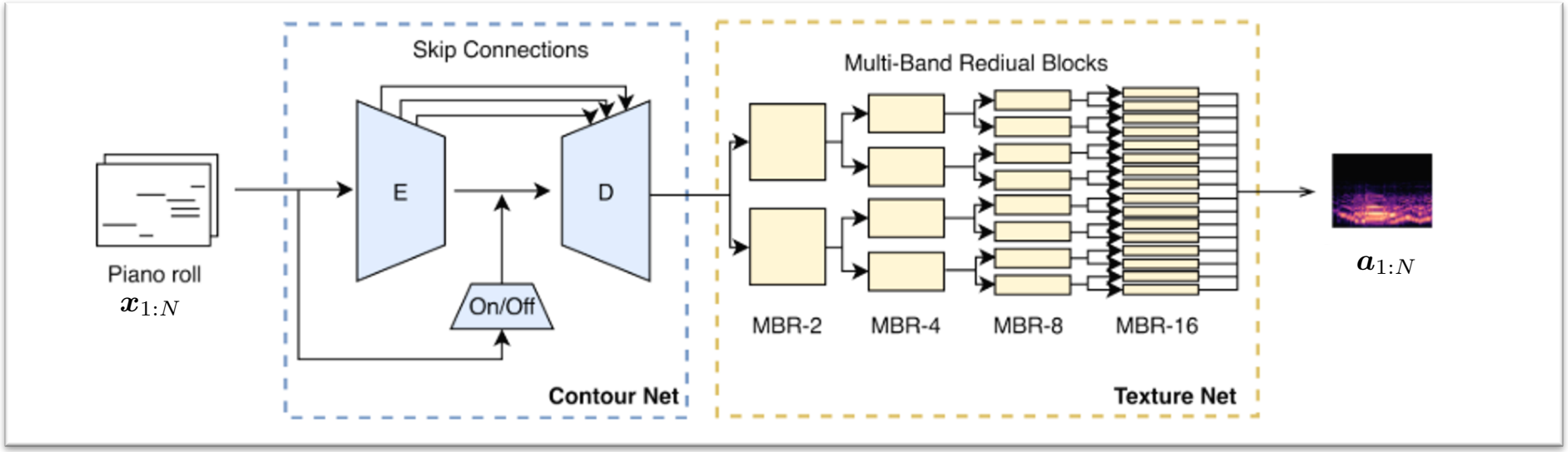
Wang, Y. *et al.* Tacotron: Towards End-to-End Speech Synthesis. in *Proc. Interspeech* 4006–4010 (2017).
Yasuda, Y., Wang, X., Takaki, S. & Yamagishi, J. Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language. in *Proc. ICASSP* 6905–6909 (2019).

Methods

❑ Acoustic model



2. Reference model from music field: PerformanceNet (Wang 2019)



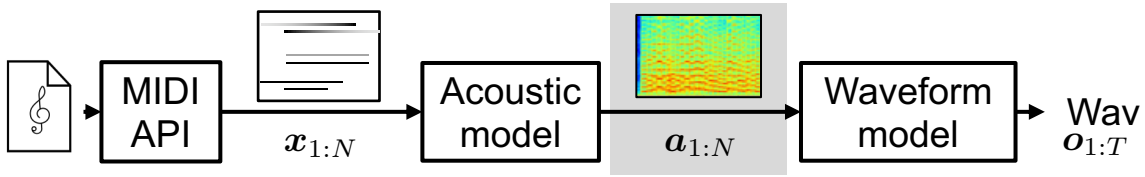
Adopted from figure 3 in (Wang 2019)

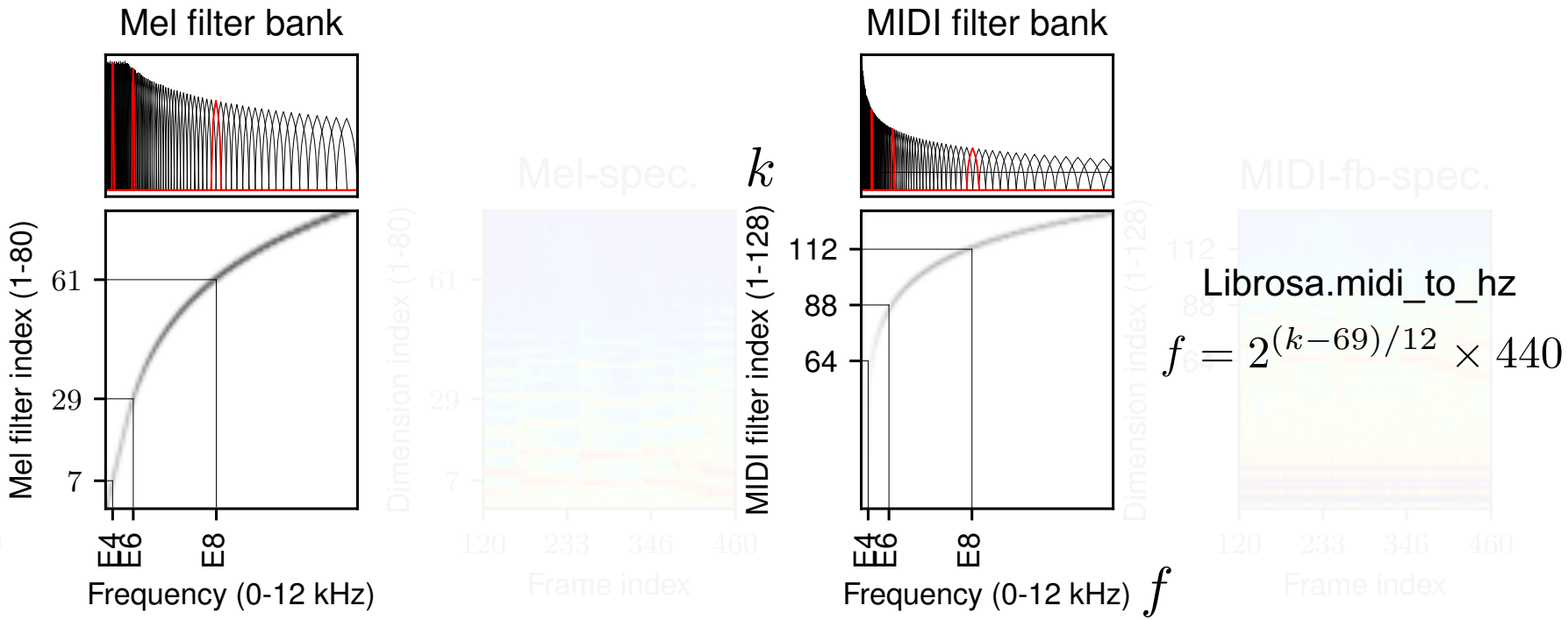
Wang, B. & Yang, Y.-H. PerformanceNet: Score-to-audio music generation with multi-band convolutional residual network. in *Proceedings of the AAAI Conference on Artificial Intelligence* vol. 33 1174–1181 (2019).

Methods

❑ Acoustic features

1. Mel-spectrogram
2. MIDI-filterbank-spectrogram



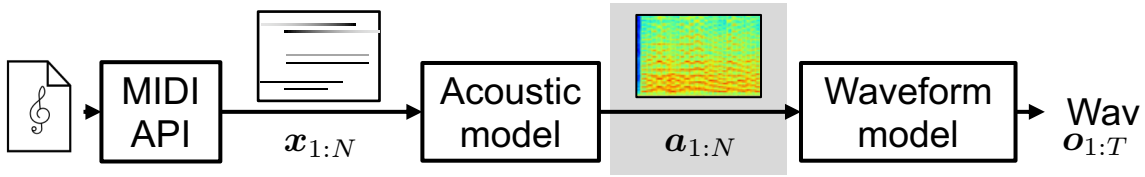


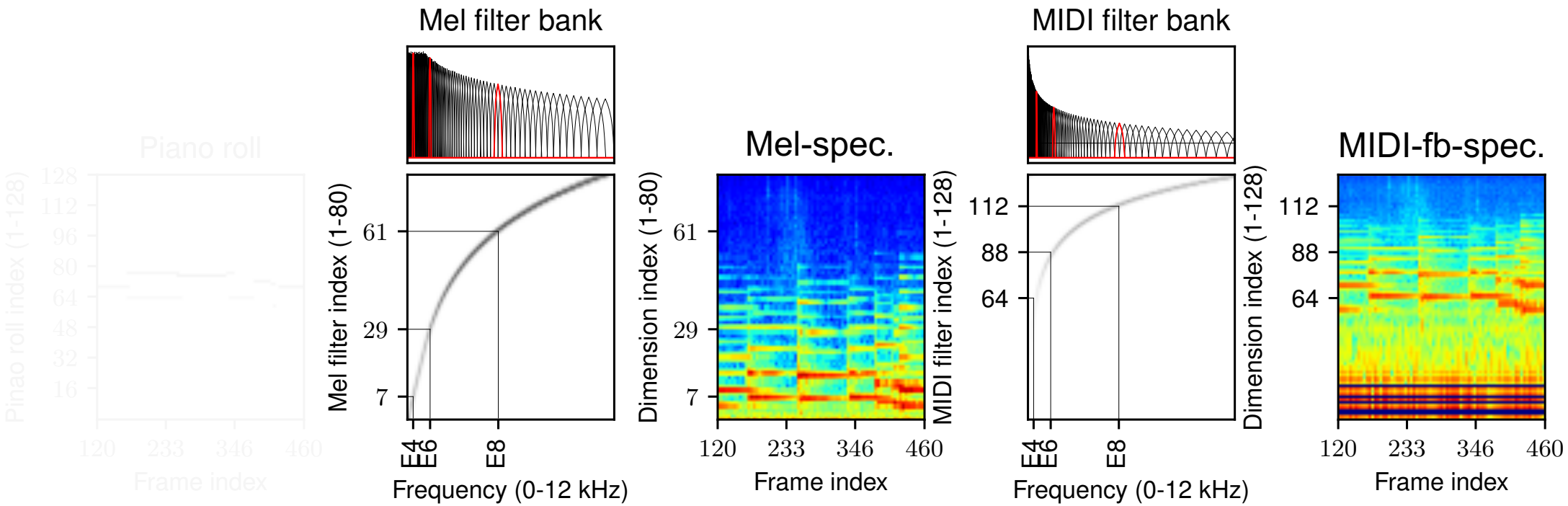
Librosa midi_to_hz: https://librosa.org/doc/0.7.0/generated/librosa.core.midi_to_hz.html

Methods

❑ Acoustic features

1. Mel-spectrogram
2. MIDI-filterbank-spectrogram

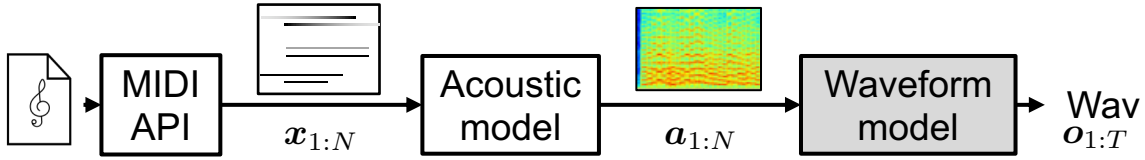




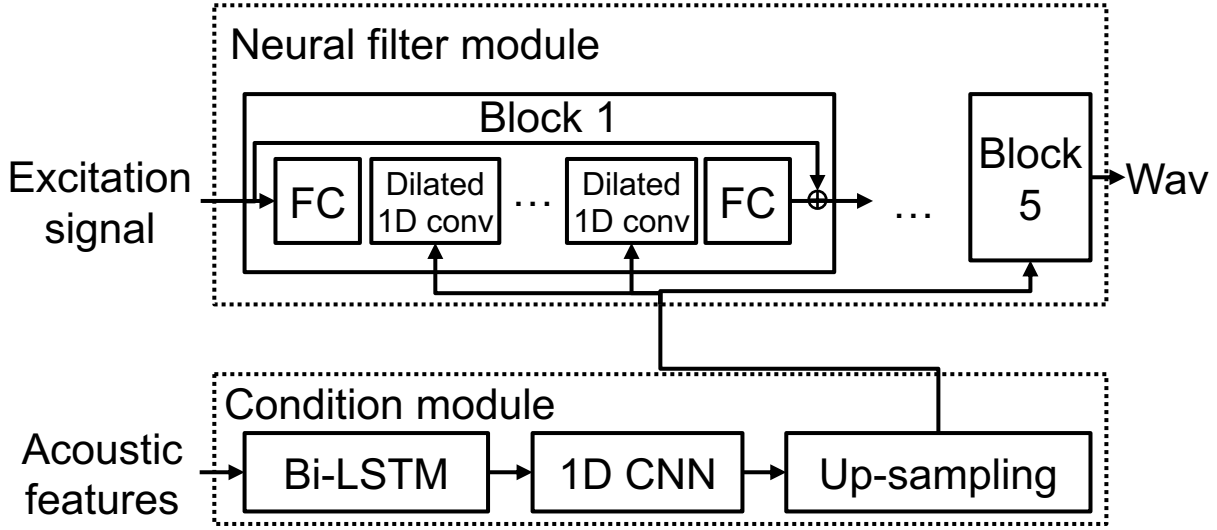
Blank lines in MIDI-fb-spec. due to frequency resolution of FFT (see appendix)

Methods

❑ Waveform model



- Based on music neural source-filter (NSF) model (Zhao 2020) but
  - No harmonic-plus-noise structure

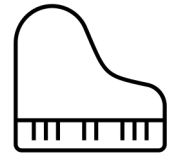


Zhao, Y., Wang, X., Juvela, L. & Yamagishi, J. Transferring neural speech waveform synthesizers to musical instrument sounds generation. in *Proc. ICASSP* 6269–6273 (IEEE, 2020). doi:10.1109/ICASSP40776.2020.9053047

Experiments

❑ Database: MAESTRO v2.0 (Hawthorne 2018)

- Real **piano** performances in International Piano-e-Competition
- MIDI was recorded simultaneously during performance
  - Aligned audio & piano roll



- For experiments:
  - Follow official data split
  - 24kHz, 16 bits PCM

Data split of MAESTRO

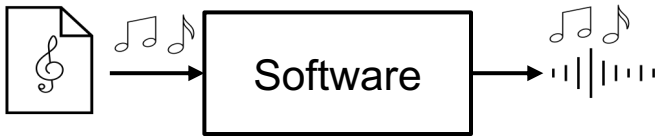
Split	Performances	Duration (hours)	Size (GB)	Notes (millions)
Train	967	161.3	97.7	5.73
Validation	137	19.4	11.8	0.64
Test	178	20.5	12.4	0.76
Total	**1282**	**201.2**	**121.8**	**7.13**

From https://magenta.tensorflow.org/datasets/maestro#v200

Hawthorne, C. *et al.* Enabling factorized piano music modeling and generation with the MAESTRO dataset. in *Proc. ICLR* (2018).
https://piano-e-competition.com/

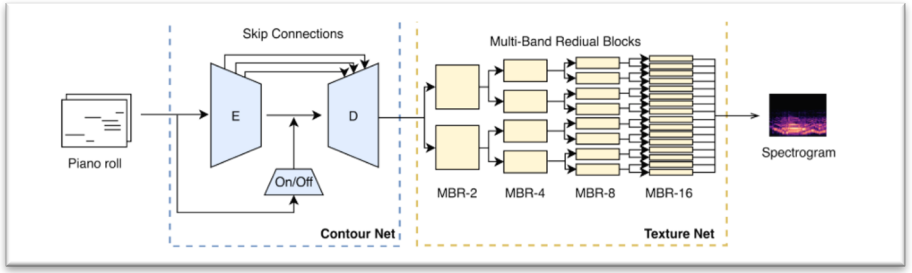
Experiments

❑ Models in comparison



System ID	Acoustic model	Acoustic feature	Excit. signal	Wave. model	Pitch mismatch note	Pitch mismatch chord	MOS (mean)	
Natural	-	-	-	-	-	-	4.04	
`Fluidsynth`	Sample-based MIDI-to-audio software				5.20	6.37	3.66	Software
`Pianoteq`	Physical-model MIDI-to-audio software				4.82	6.00	4.25	
abs-mfb-sin	-	midi-fb	sine	NSF	-	-	3.87	
abs-mfb-noi	-	midi-fb	noise	NSF	-	-	3.77	
abs-mel-sin	-	mel-spc.	sine	NSF	-	-	2.72	
abs-mel-noi	-	mel-spc.	noise	NSF	-	-	3.81	
taco2-mfb-sin	taco2	midi-fb	sine	NSF	4.61	6.34	2.97	
taco2-mfb-noi	taco2	midi-fb	noise	NSF	4.66	6.36	3.18	
taco3-mfb-sin	taco3	midi-fb	sine	NSF	4.78	6.48	3.19	
taco3-mfb-noi	taco3	midi-fb	noise	NSF	4.89	6.53	3.19	
taco4-mfb-sin	taco4	midi-fb	sine	NSF	4.86	6.39	2.98	
taco4-mfb-noi	taco4	midi-fb	noise	NSF	4.97	6.42	2.95	
pfnet-mfb-sin	PFNet	midi-fb	sine	NSF	5.59	7.14	3.10	
pfnet-mfb-noi	PFNet	midi-fb	noise	NSF	5.78	7.26	3.05	
pfnet-mel-sin	PFNet	mel-spec.	sine	NSF	5.66	7.17	1.82	
pfnet-mel-noi	PFNet	mel-spec.	noise	NSF	5.43	6.98	1.62	
`pfnet-spec-GL`	PFNet	spec.	-	GL				Open-source code
midi-sin-nsf	-	-	sine	NSF				
midi-noi-nsf	-	-	noise	NSF				

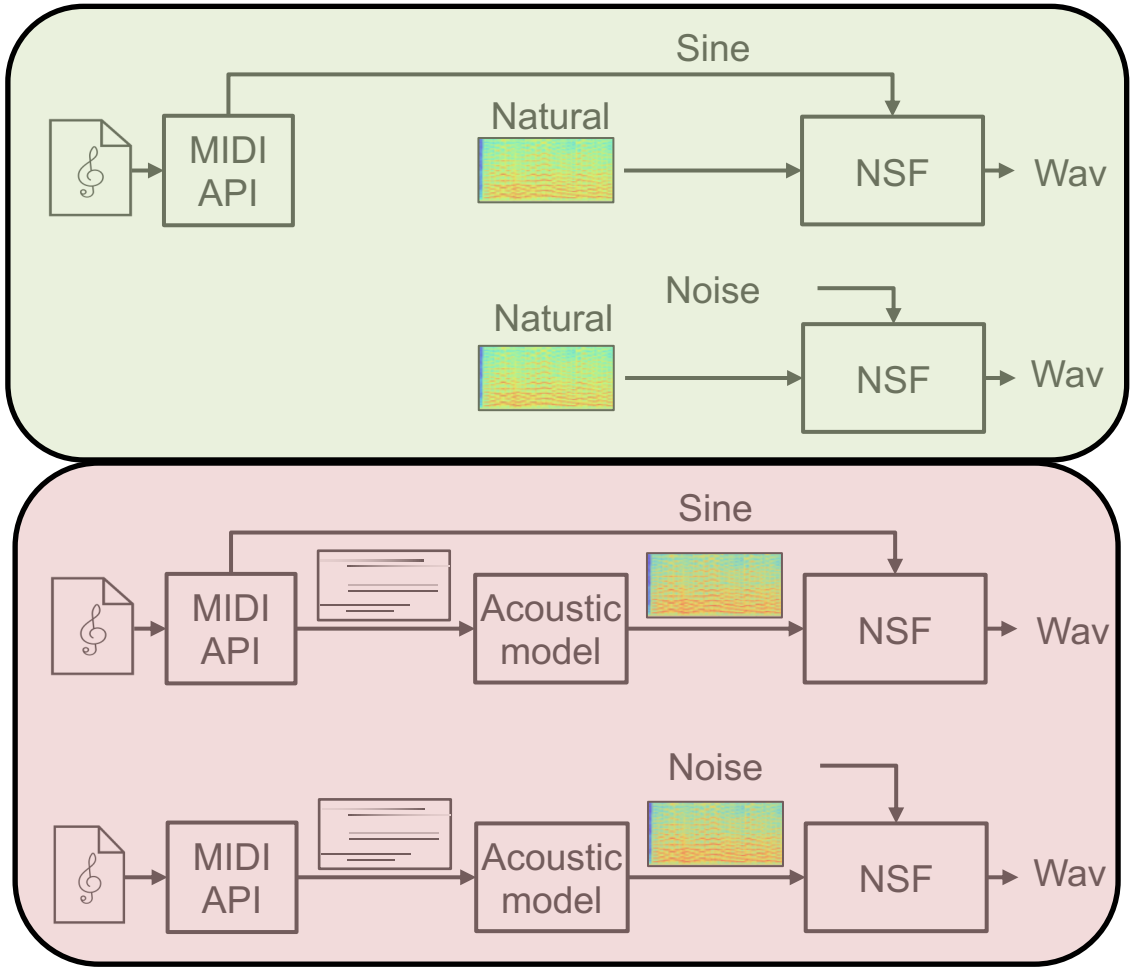
Original PerformanceNet



https://www.fluidsynth.org/
https://www.modartt.com/pianoteq
https://github.com/bwang514/PerformanceNet

Experiments

❑ Models in comparison



System ID	Acoustic model	Acoustic feature	Excit. signal	Wave. model	Pitch mismatch note	Pitch mismatch chord	MOS (mean)
Natural	-	-	-	-	-	-	4.04
Fluidsynth	Sample-based MIDI-to-audio software						
Pianoteq	Physical-model MIDI-to-audio software						
abs-mfb-sin	-	midi-fb	sine	NSF	-	-	3.87
abs-mfb-noi	-	midi-fb	noise	NSF	-	-	3.77
abs-mel-sin	-	mel-spc.	sine	NSF	-	-	3.82
abs-mel-noi	-	mel-spc.	noise	NSF	-	-	3.81
taco2-mfb-sin	taco2	midi-fb	sine	NSF	4.61	6.34	2.97
taco2-mfb-noi	taco2	midi-fb	noise	NSF	4.66	6.36	3.18
taco3-mfb-sin	taco3	midi-fb	sine	NSF	4.78	6.48	3.19
taco3-mfb-noi	taco3	midi-fb	noise	NSF	4.89	6.53	3.19
taco4-mfb-sin	taco4	midi-fb	sine	NSF	4.86	6.39	2.98
taco4-mfb-noi	taco4	midi-fb	noise	NSF	4.[illegible]	[illegible]	[illegible]
pfnet-mfb-sin	PFNet	midi-fb	sine	NSF	5.59	7.14	3.10
pfnet-mfb-noi	PFNet	midi-fb	noise	NSF	5.78	7.26	3.05
pfnet-mel-sin	PFNet	mel-spec.	sine	NSF	5.66	7.17	1.82
pfnet-mel-noi	PFNet	mel-spec.	noise	NSF	5.74	7.25	2.93
pfnet-spec-GL	PFNet	spec.	-	GL	5.43	6.98	1.62
midi-sin-nsf	-	-	sine	NSF			
midi-noi-nsf	-	-	noise	NSF			

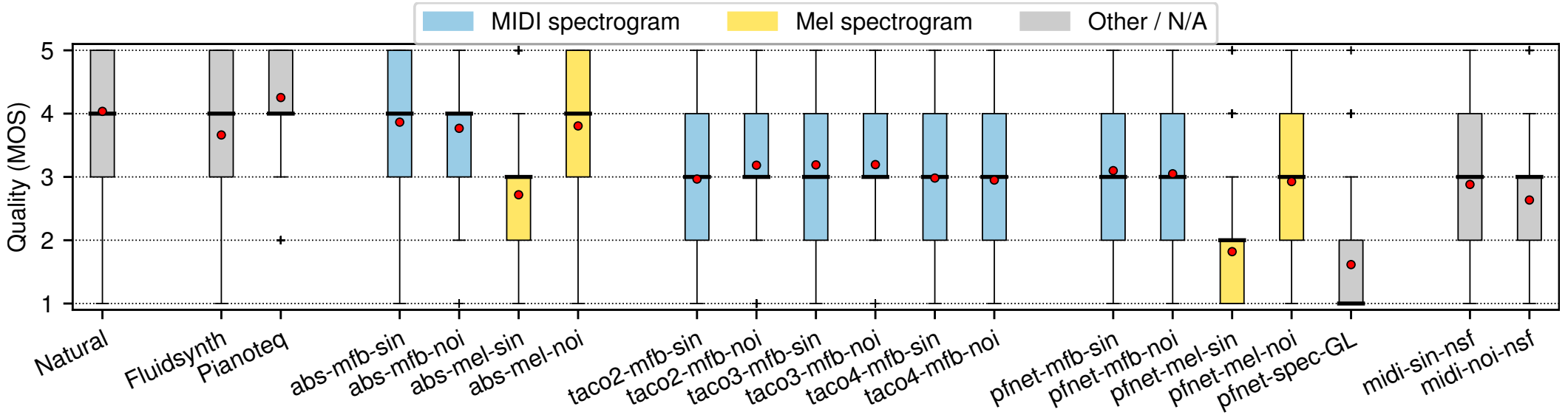
Copy-synthesis

MIDI-to-audio

https://www.fluidsynth.org/
https://www.modartt.com/pianoteq

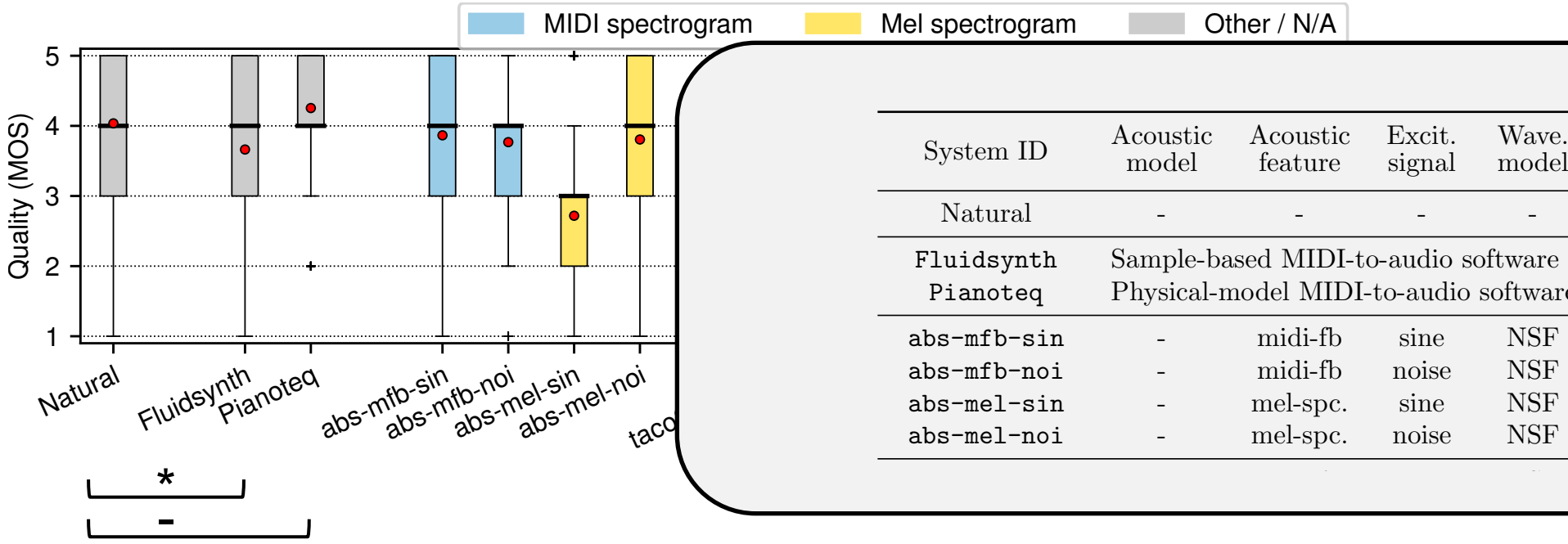
Experiments

- **Listening test:** crowd-sourcing, 224 amateur participants



Experiments

- **Listening test:** crowd-sourcing, 224 amateur participants

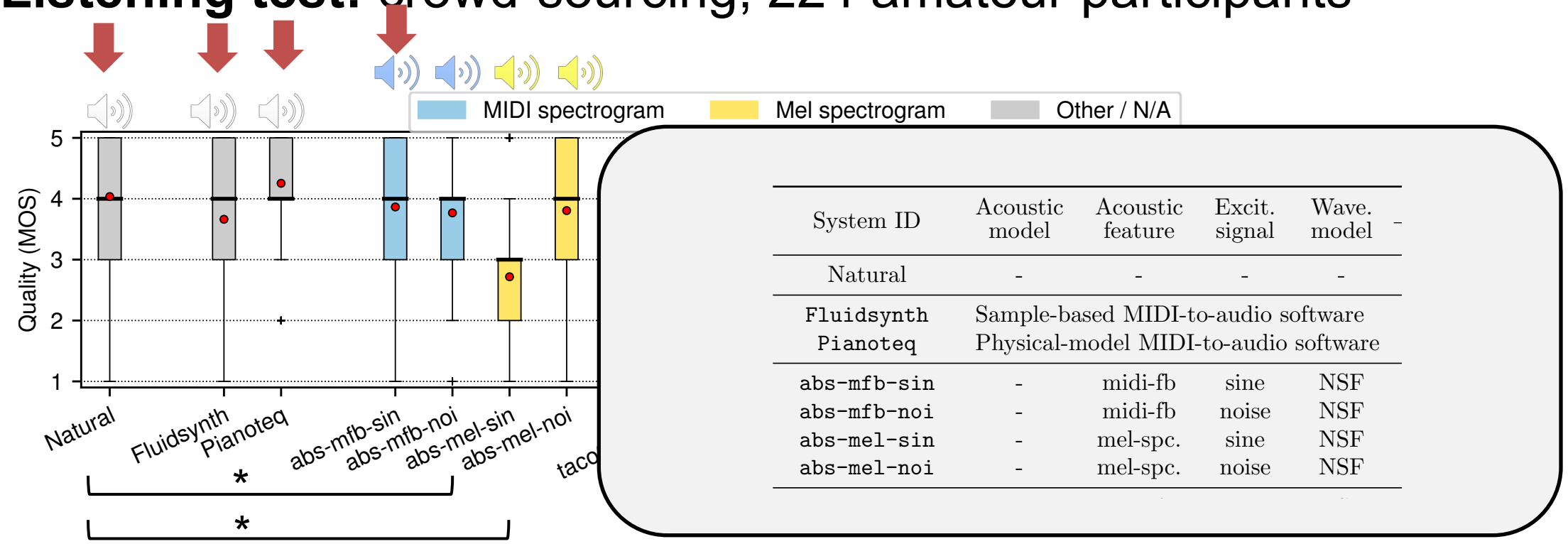


System ID	Acoustic model	Acoustic feature	Excit. signal	Wave. model
Natural	-	-	-	-
`Fluidsynth`	Sample-based MIDI-to-audio software			
`Pianoteq`	Physical-model MIDI-to-audio software			
`abs-mfb-sin`	-	midi-fb	sine	NSF
`abs-mfb-noi`	-	midi-fb	noise	NSF
`abs-mel-sin`	-	mel-spc.	sine	NSF
`abs-mel-noi`	-	mel-spc.	noise	NSF

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

Experiments

- **Listening test:** crowd-sourcing, 224 amateur participants



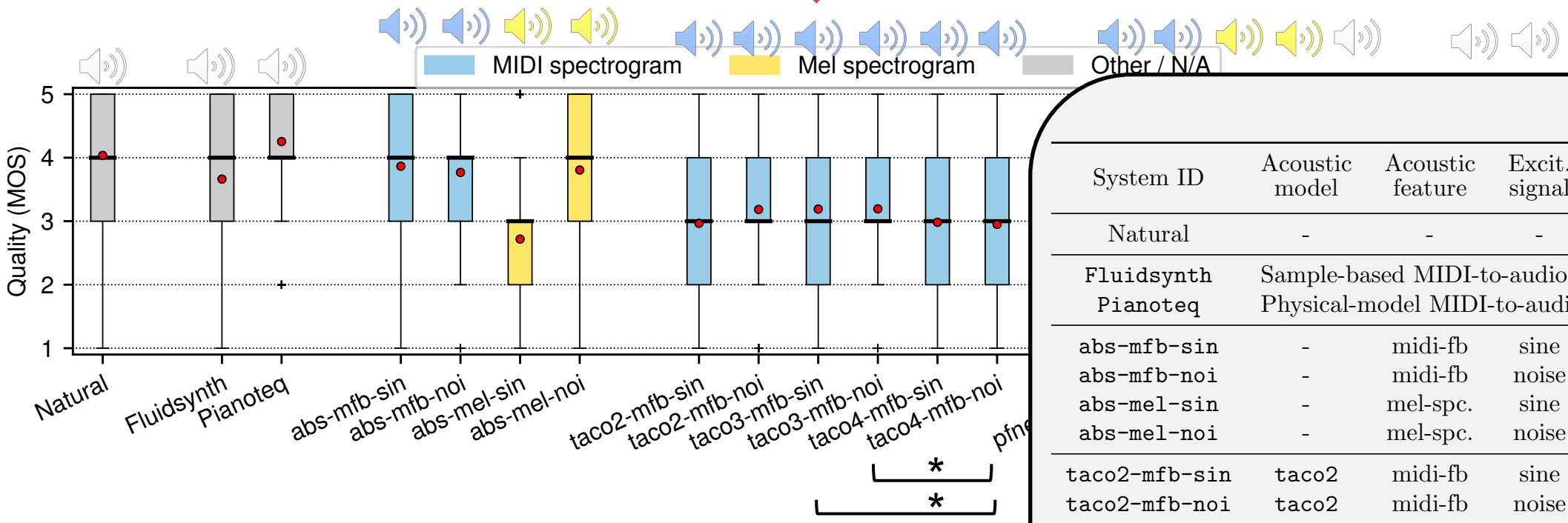
System ID	Acoustic model	Acoustic feature	Excit. signal	Wave. model
Natural	-	-	-	-
`Fluidsynth`	Sample-based MIDI-to-audio software			
`Pianoteq`	Physical-model MIDI-to-audio software			
`abs-mfb-sin`	-	midi-fb	sine	NSF
`abs-mfb-noi`	-	midi-fb	noise	NSF
`abs-mel-sin`	-	mel-spc.	sine	NSF
`abs-mel-noi`	-	mel-spc.	noise	NSF

- Neural waveform model works well in copy-synthesis condition

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

Experiments

- **Listening test:** crowd-sourcing, 224 amateur participants

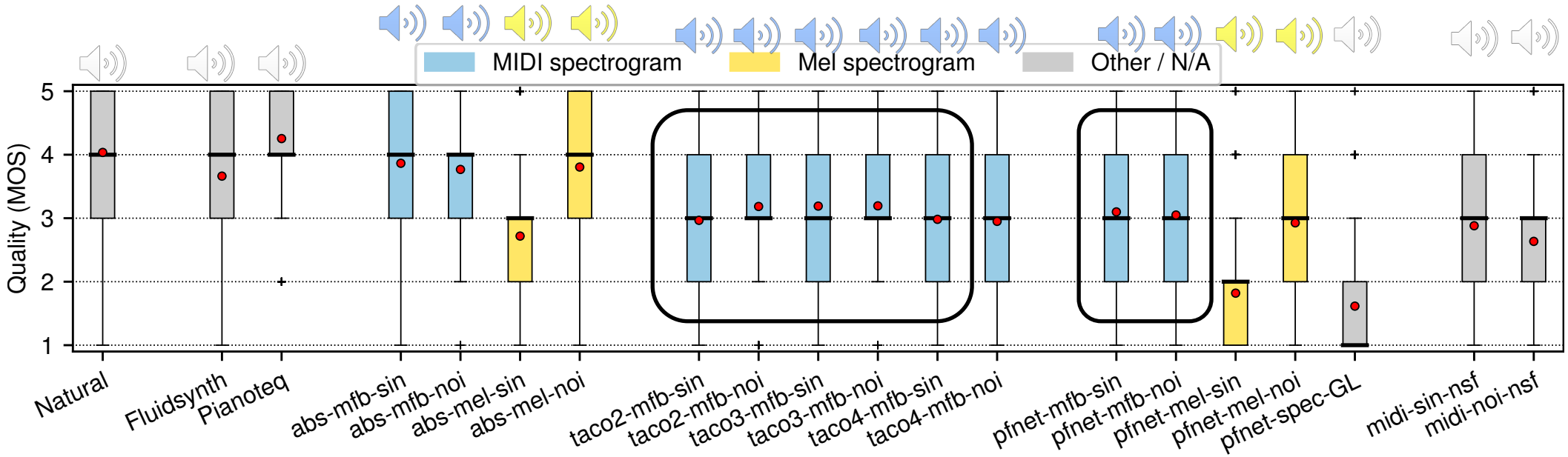


System ID	Acoustic model	Acoustic feature	Excit. signal	Wave. model
Natural	-	-	-	-
`Fluidsynth`	Sample-based MIDI-to-audio software			
`Pianoteq`	Physical-model MIDI-to-audio software			
`abs-mfb-sin`	-	midi-fb	sine	NSF
`abs-mfb-noi`	-	midi-fb	noise	NSF
`abs-mel-sin`	-	mel-spc.	sine	NSF
`abs-mel-noi`	-	mel-spc.	noise	NSF
`taco2-mfb-sin`	`taco2`	midi-fb	sine	NSF
`taco2-mfb-noi`	`taco2`	midi-fb	noise	NSF
`taco3-mfb-sin`	`taco3`	midi-fb	sine	NSF
`taco3-mfb-noi`	`taco3`	midi-fb	noise	NSF
`taco4-mfb-sin`	`taco4`	midi-fb	sine	NSF
`taco4-mfb-noi`	`taco4`	midi-fb	noise	NSF

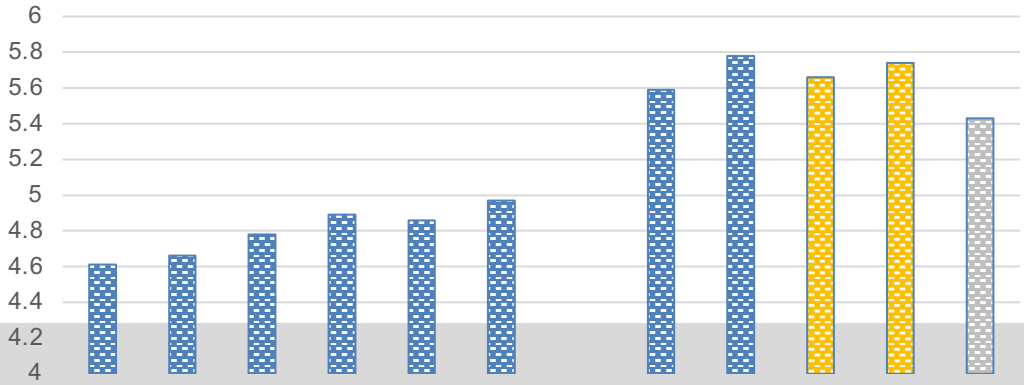
- **Taco2:** 800-frame segments; 4x downsampling for better alignments
- **Taco3:** Warm-started from taco2; input current piano-roll frame at the decoder pre-net
- **Taco4:** Warm-started from taco2; no downsampling or piano-roll input to pre-net

Experiments

- **Listening test:** crowd-sourcing, 224 amateur participants



- **Pitch distortion**
  - Natural audio as target
  - Evaluated on single notes
  - The lower the better



Messages

TTS & MIDI-to-audio

- Techniques can be shared: acoustic model, waveform model
- Performance bottleneck on acoustic model (and waveform model)

On waveform modeling

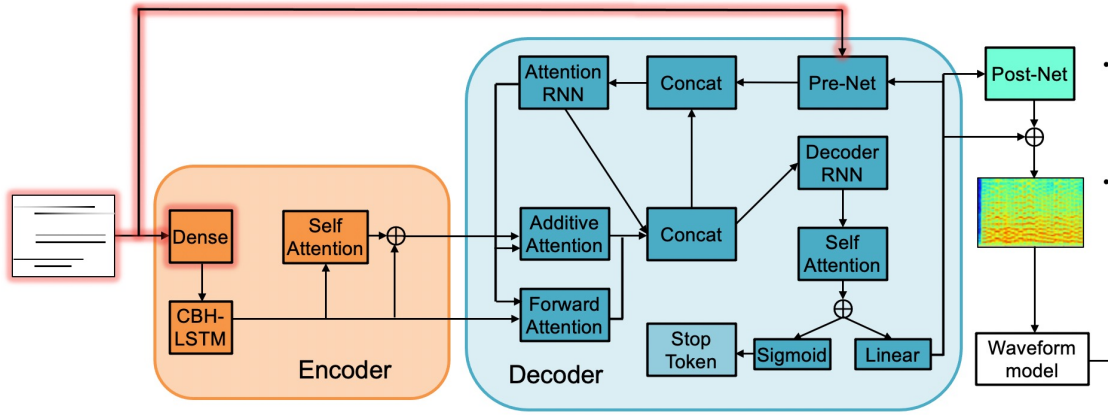
- Physical-model performs well but lacks reverberation effect
- Sample-based model replies on the sample database

- Non-AR waveform model is OK in copy-synthesis
  - Reverberation is captured
  - Noise excitation is OK

Messages

On acoustic model

- Obtaining good alignments for longer input sequences is challenging
- Inputting the piano-roll frame to the decoder prenet helps improve alignments
  - Acceptable for perfectly-aligned performance-MIDI
  - Have to consider other strategies for non-aligned score-MIDI



Thank you!

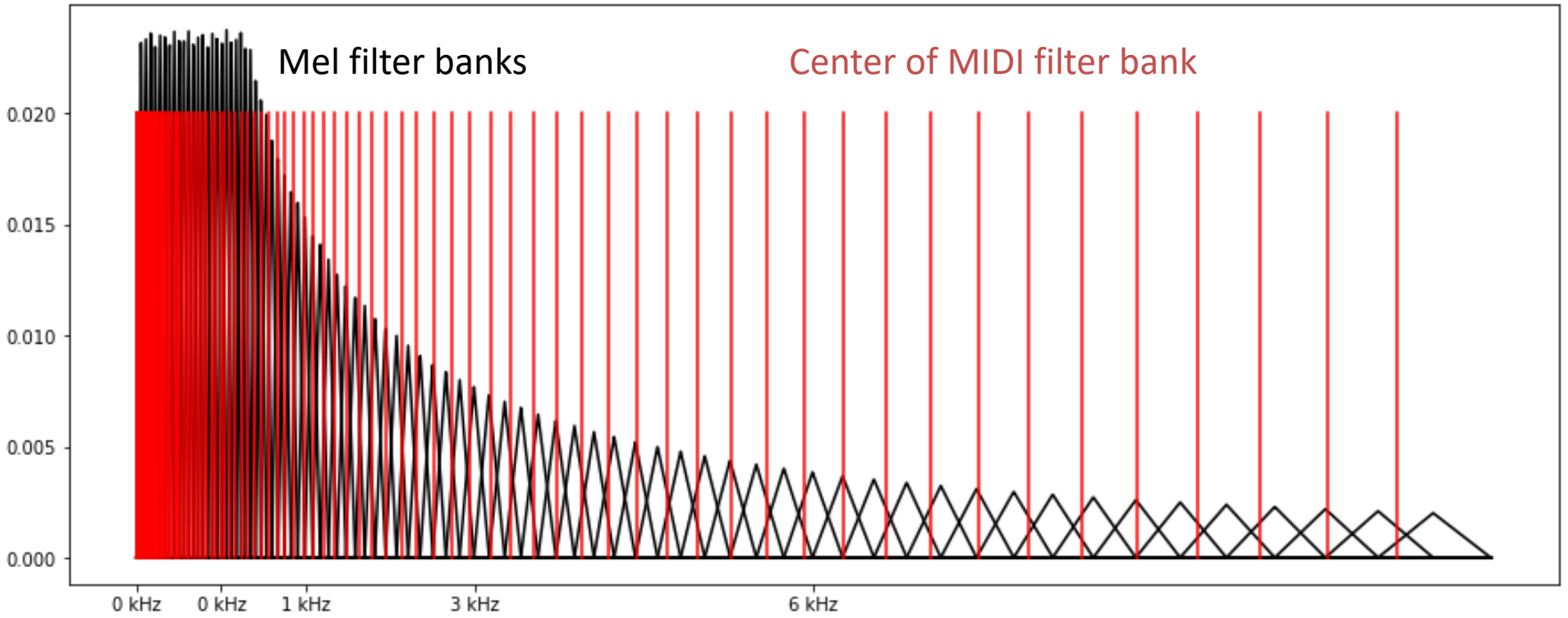
Tacotron code: https://github.com/nii-yamagishilab/self-attention-tacotron

NSF code: https://github.com/nii-yamagishilab/project-NN-Pytorch-scripts

Samples: https://nii-yamagishilab.github.io/samples-xin/main-midi2audio.html

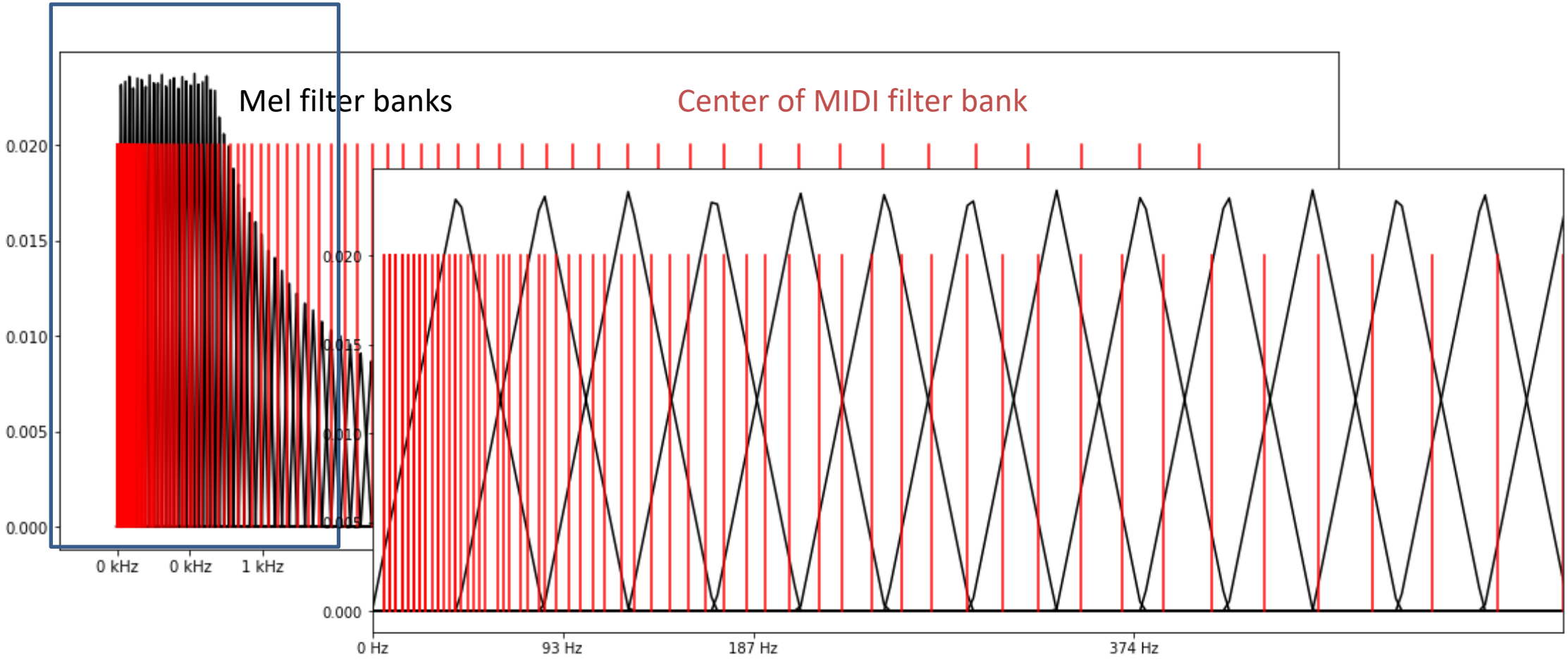
Appendix

- **On MIDI filterbank**



Appendix

- **On MIDI filterbank**



Appendix

❑ On MIDI filterbank