Transferring neural speech waveform synthesizers to musical instrument sounds generation

Yi Zhao*, Xin Wang*, Lauri Juvela†, Junichi Yamagishi*

*National Institute of Informatics, Tokyo, Japan
† Department of Signal Processing and Acoustics, Aalto University, Finland

Background

- Musical sound synthesis can benefit from neural generative models
  - Unconditional generation of continuous music [1]
  - Conditional synthesis of isolated musical notes [2, 3]
  - Approaches resembling the speech synthesis have been proposed musical instrument sounds generation [4] & singing synthesis [5] : a WaveNet generates audio from the Mel-spectrograms.

[1] Dieleman, Sander, Aaron van den Oord, and Karen Simonyan. "The challenge of realistic music generation: modelling raw audio at scale." *Advances in Neural Information Processing Systems*. 2018.
[2] Engel, Jesse, et al. "Neural audio synthesis of musical notes with wavenet autoencoders." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.
[3] Engel, Jesse, et al. "Gansynth: Adversarial neural audio synthesis." *arXiv preprint arXiv:1902.08710* (2019).
[4] Kim, Jong Wook, et al. "Neural music synthesis for flexible timbre control." ICASSP. IEEE, 2019.
[5] Blaauw, Merlijn, Jordi Bonada, and Ryunosuke Daido. "Data efficient voice cloning for neural singing synthesis." ICASSP. IEEE, 2019.

Motivations

- Two questions:
  - Is there a universal neural waveform synthesizer that is capable of synthesizing both speech (independent of speaker) and musical audio (independent of instrument)?
  - Can musical audio synthesis benefit from speech models?

- Motivation of this work :
  - Investigate transferring neural speech waveform synthesizers to musical instrument sounds generation

Neural Waveform Synthesizers

- Three types of neural waveform synthesizers in this paper
  - Neural autoregressive (AR) models

    WaveNet SampleRNN WaveRNN

    FFTNet ExcitNet LPCNet GlotNet …

  - Inverse AR flow-based models

    Parallel WaveNet ClariNet

    WaveGlow FloWaveNet …

  - Non AR nor inverse AR flow models

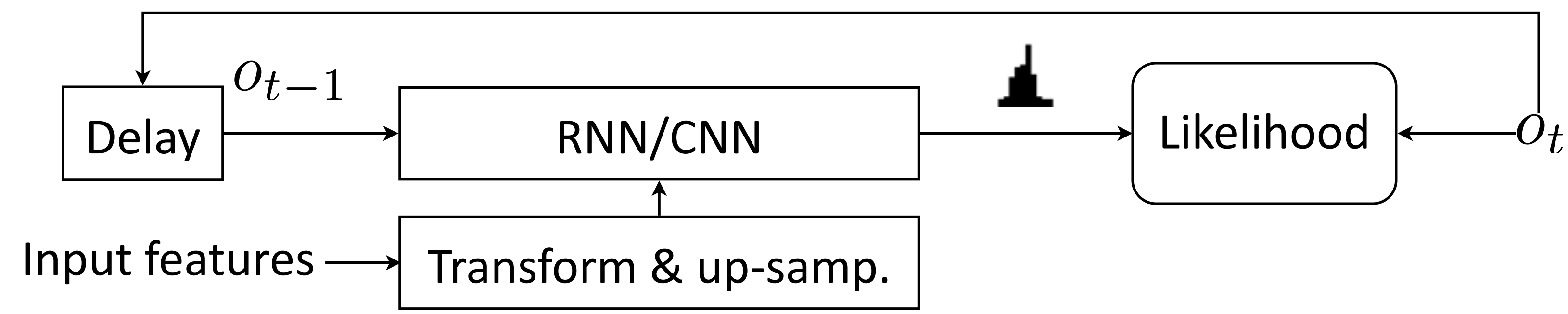
    Multi-head CNN GELP Hi-Net

    Neural source-filter waveform (NSF) models …

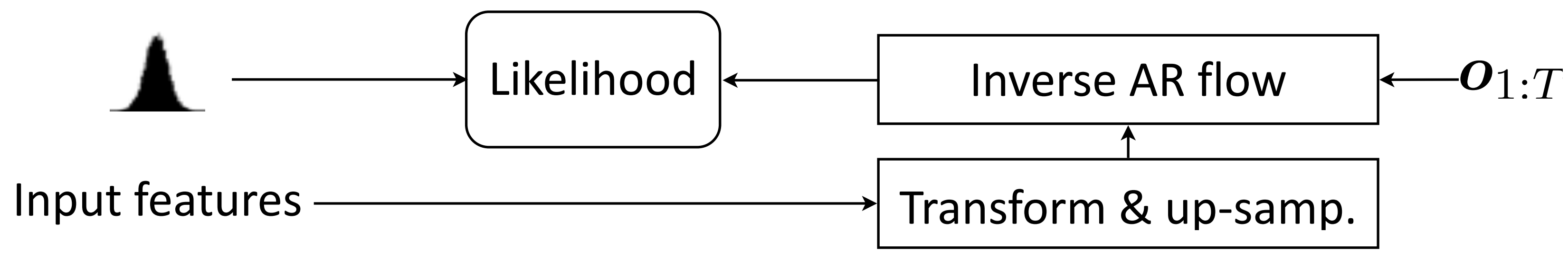
Tutorial on neural waveform models: https://www.slideshare.net/jyamagis/neural-waveform-modeling

Neural Waveform Synthesizers

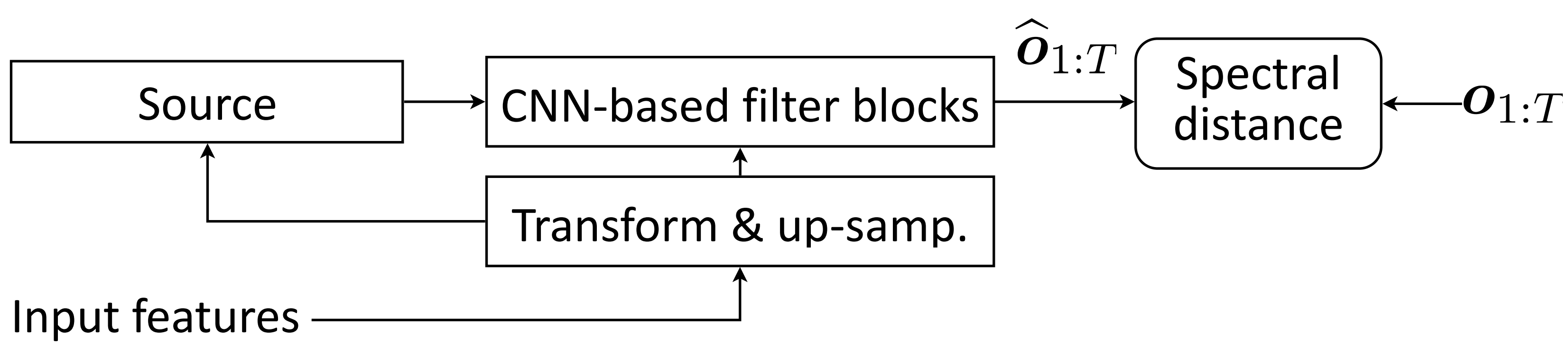
- Three types of neural waveform synthesizers in this paper
  - Neural autoregressive (AR) models



  - Inverse AR flow-based models



  - Non AR nor inverse AR flow models

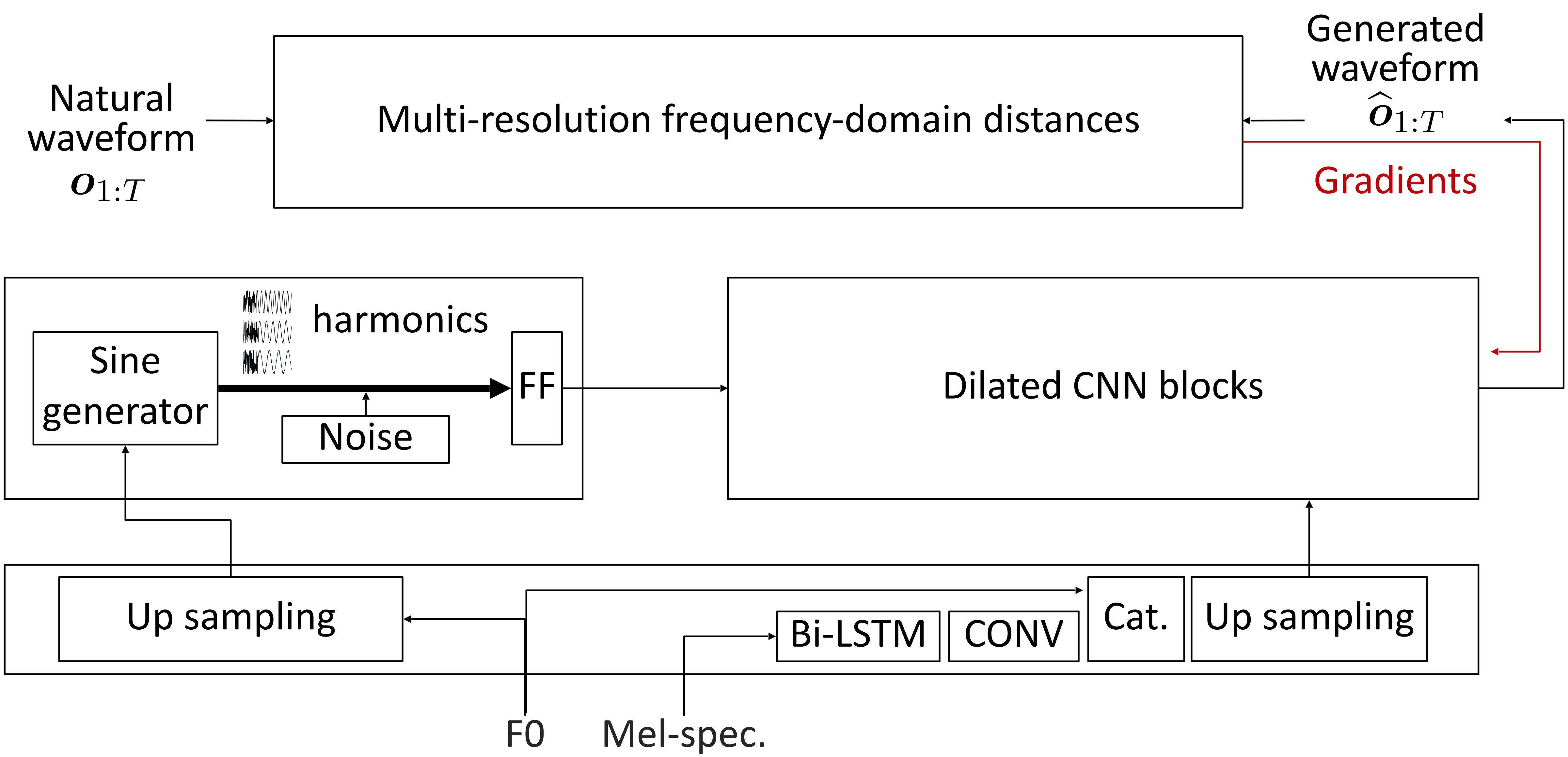


$\boldsymbol{o}_{1:T}$ is natural waveform of length T, $\widehat{\boldsymbol{o}}_{1:T}$ is generated waveform

Tutorial on neural waveform models: https://www.slideshare.net/jyamagis/neural-waveform-modeling

Neural Waveform Synthesizers

- NSF models [1]



[1] Wang, X., Takaki, S., & Yamagishi, J. (2020). Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis. IEEE/ACM TASLP, 28, 402–415.

Neural Waveform Synthesizers

- We selected three neural waveform synthesizers: WaveNet [1], WaveGlow [2], and Neural source-filter model (NSF) [3]

  - WaveNet : auto-regressive (AR) neural approach
  - WaveGlow: inverse-AR-flow-based approach
  - NSF: neither AR nor inverse AR, sine-based excitation

[1] van den Oord, A., Dieleman, S., Zen, H., et al. (2016). WaveNet: A generative model for raw audio. ArXiv Preprint ArXiv:1609.03499.

[2] Prenger, R., Valle, R., & Catanzaro, B. (2019). WaveGlow: A Flow-based Generative Network for Speech Synthesis. Proc. ICASSP, 3617–3621.

[3] Wang, X., Takaki, S., & Yamagishi, J. (2020). Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis. IEEE/ACM TALSP, 28, 402–415.

Methods

- Three neural waveform synthesizers in this paper
  - WaveNet
    - 10 bits μ-law companded waveform values
    - 30 dilated CNN blocks
    - 2.4 million model parameters
  - WaveGlow
    - NVIDIA implementation
    - Waveform in float numbers
    - 87 million model parameters
  - NSF
    - Harmonic-plus-noise NSF with trainable *sinc* filters [1]
    - Waveform in float numbers
    - 1.2 million model parameters
    - Details in appendix

[1] Wang, X., & Yamagishi, J. (2019). Neural Harmonic-plus-Noise Waveform Model with Trainable Maximum Voice Frequency for Text-to-Speech Synthesis. Proc. SSW, 1–6.

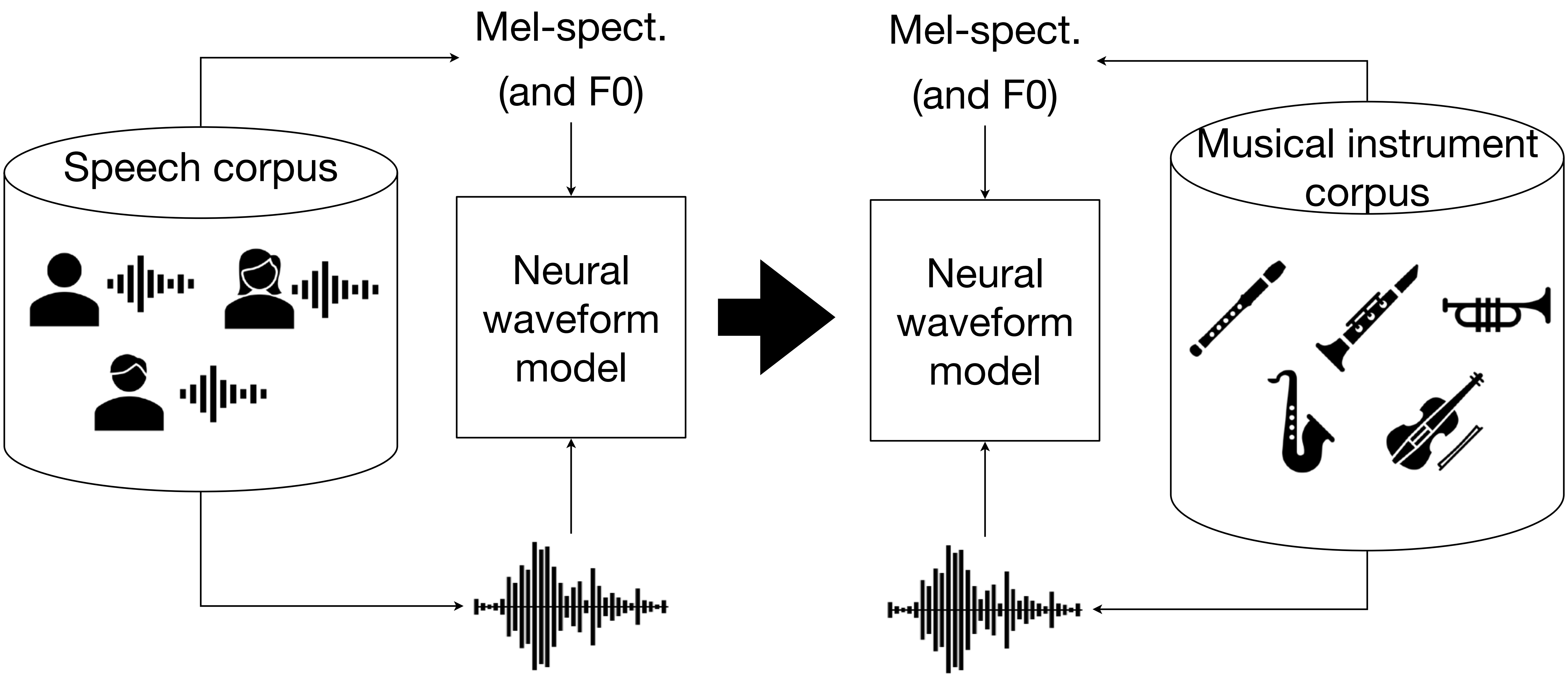
Methods

- Training scenarios

	Training from scratch (TS)	Zero-shot learning (ZS)	Adaptation by fine tuning (FT)
Pre-train	-	Speech corpus	Speech corpus
Fine-tuning	Musical instrument corpus	-	Musical instrument corpus

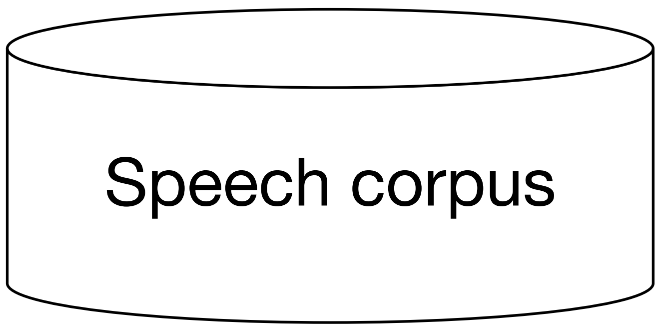
Methods

- Training scenario: adaptation by fine-tuning (FT)

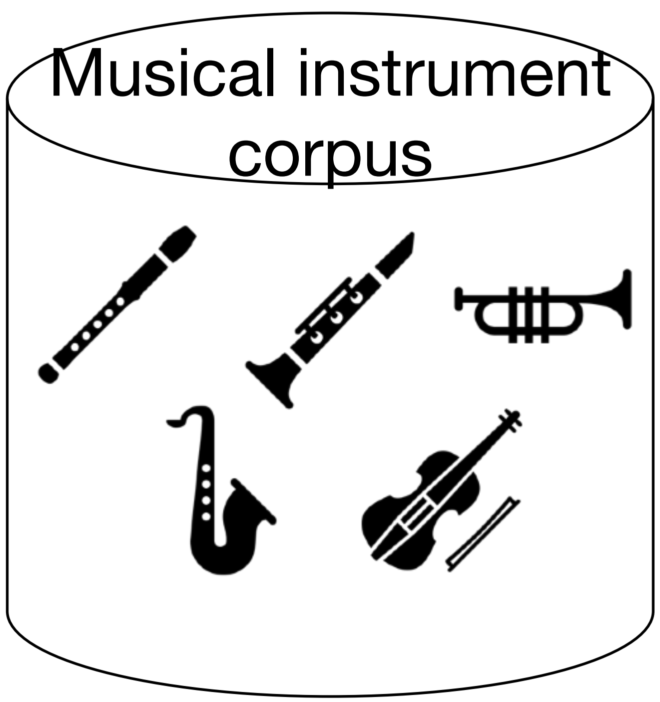


Experiments

- ## Data corpora



- VCTK speech corpus [1]
- 87 speakers
- 17.4k utterances, 21 hours



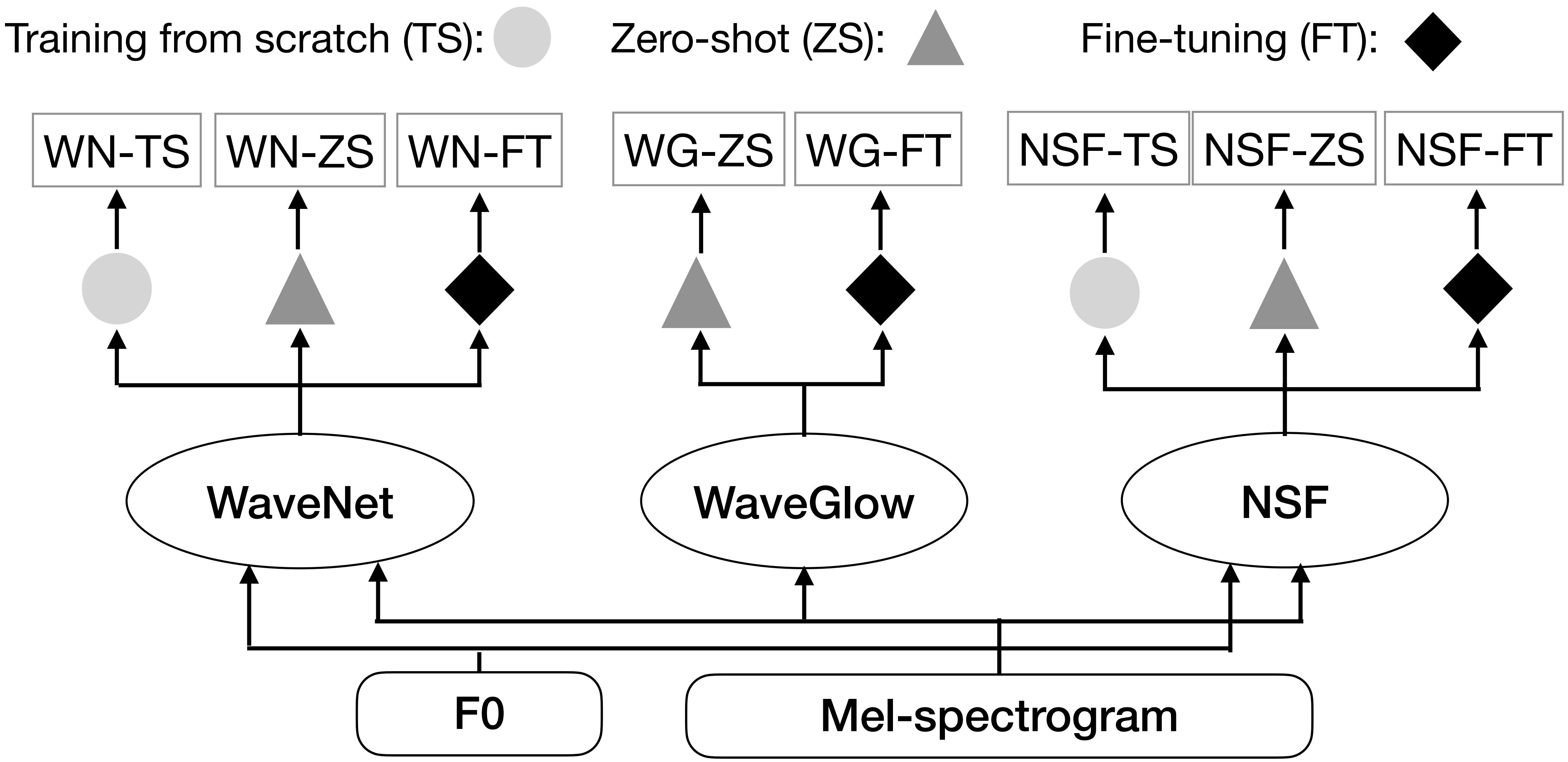
- University of Rochester Multi-Modal Music Performance (URMP) Dataset [2]
- 13 instruments: 4 strings, 5 woodwind, 4 brass
- 149 individual **monophonic** pieces, 4.5 hours
- Ground-truth F0 available
- Non-professional music performers

[1] Veaux, C., Yamagishi, J., MacDonald, K., & others. (2016). Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. University of Edinburgh

[2] Li, B., Liu, X., Dinesh, K., Duan, Z., & Sharma, G. (2018). Creating a Multitrack Classical Music Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications. *IEEE Transactions on Multimedia*, *21*(2), 522–535.

Experiments

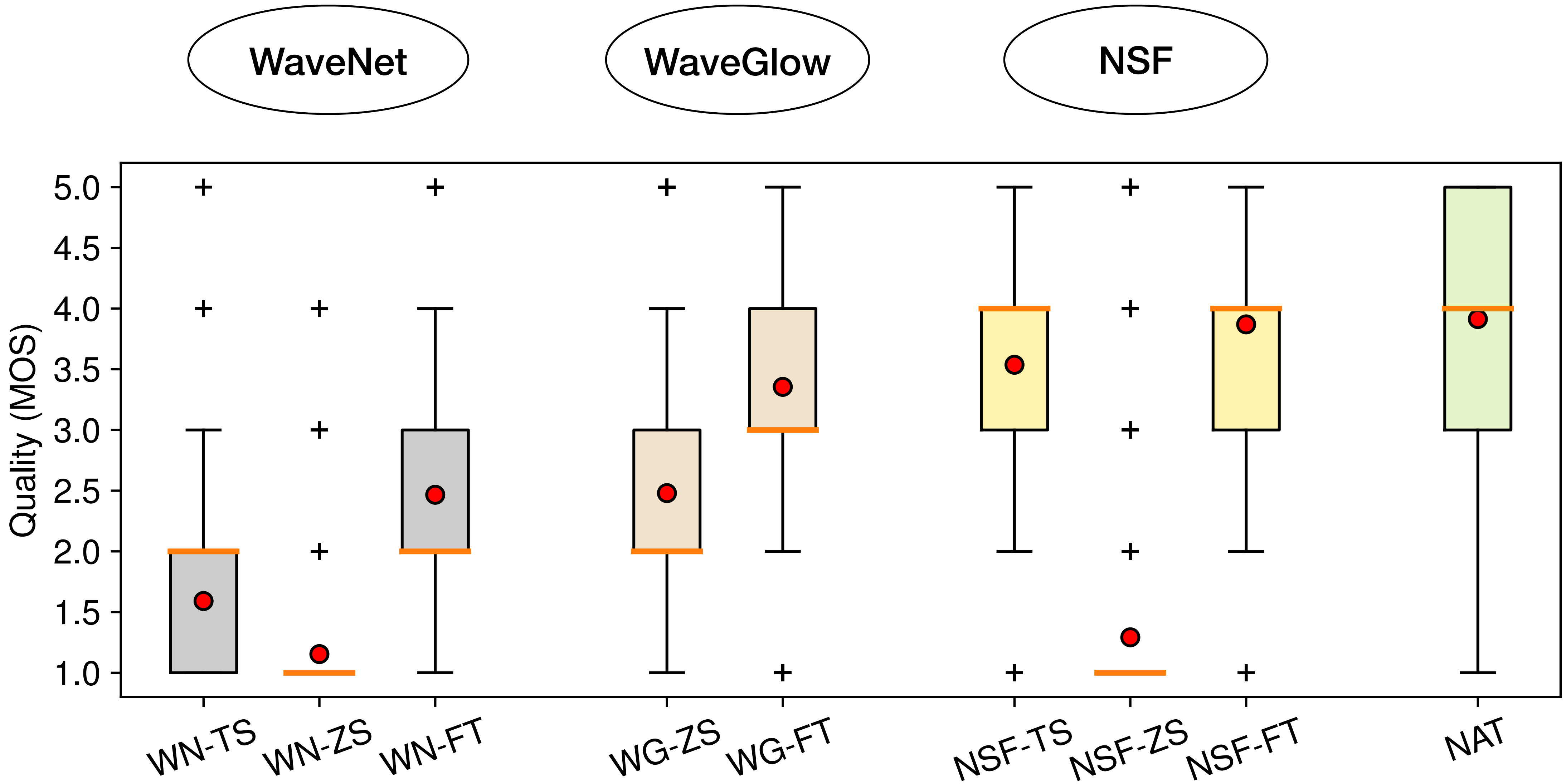
- Systems



- Evaluation
  - Subjective evaluation on quality and similarity
  - 136 non-professional listeners

Experiments

- Results: quality

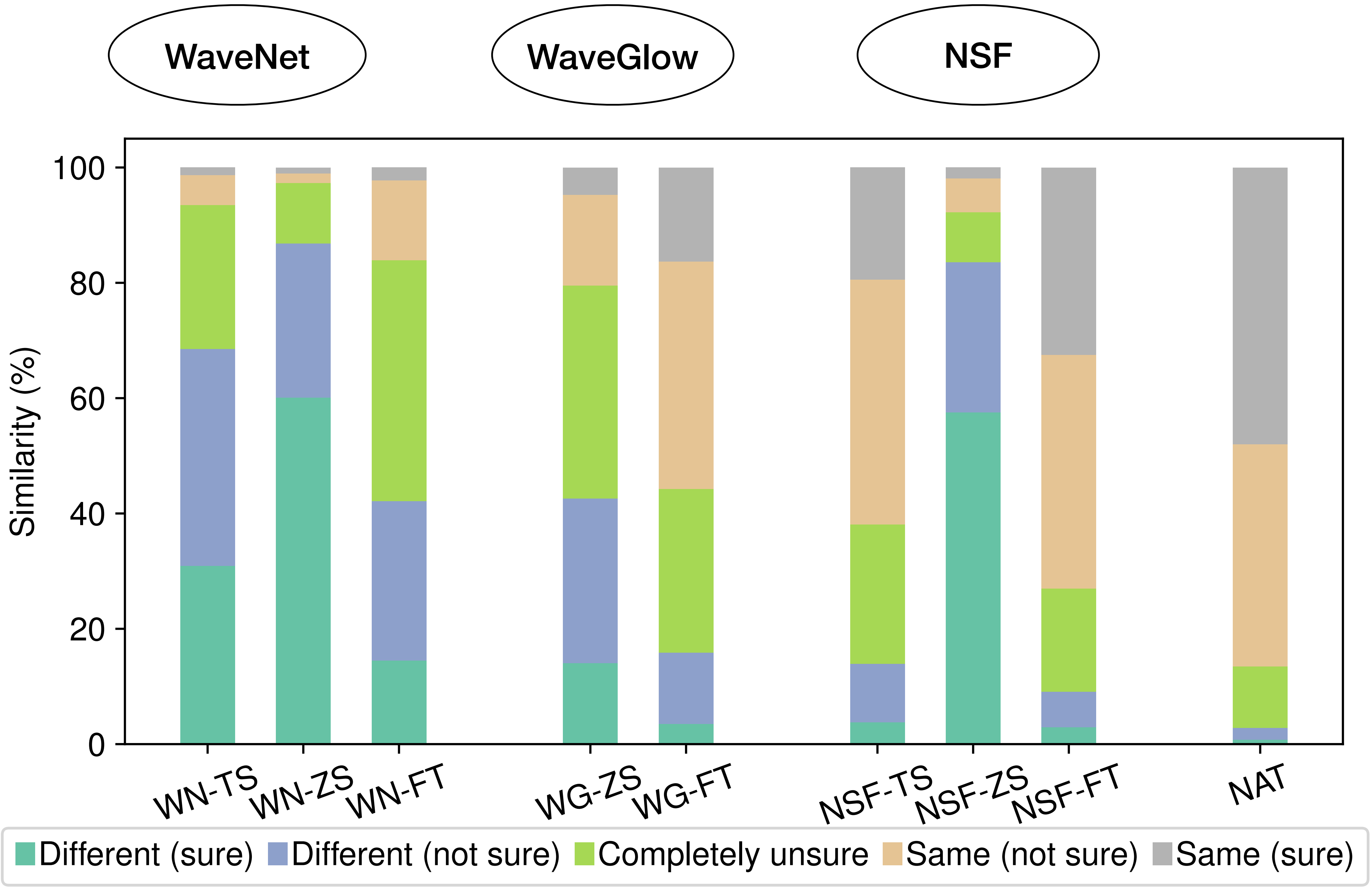


https://nii-yamagishilab.github.io/samples-nsf/neural-music.html

Training from scratch (TS) Zero-shot (ZS) Fine-tuning (FT)

Experiments

- **Results: similarity**



Training from scratch (TS) Zero-shot (ZS) Fine-tuning (FT)

Experiments

- **Results: quality**

Comparable to or even slightly better than natural sounds

	WaveNet			WaveGlow			NSF			Natural
	FS	ZS	FT	FS	ZS	FT	FS	ZS	FT	
Violin	1.78	1.19	2.50	-	2.26	3.50	3.58	1.12	3.69	4.09
Cello	1.13	1.30	2.11	-	2.72	3.51	3.11	1.43	3.42	4.26
Viola	1.33	1.81	2.05	-	2.62	3.29	3.40	2.90	3.83	3.62
Flute	1.57	1.00	2.74	-	2.45	3.65	3.84	1.09	4.00	4.03
Oboe	1.38	1.00	2.51	-	1.74	3.25	3.47	1.11	3.92	3.89
Clarinet	1.64	1.34	2.57	-	2.86	3.76	3.88	1.07	4.09	4.29
Saxophone	1.73	1.13	2.47	-	2.87	3.13	3.40	1.10	3.75	3.70
Trumpet	1.94	1.04	2.84	-	2.49	3.60	3.69	1.11	4.17	4.07
Horn	1.70	1.04	2.74	-	2.87	3.37	3.57	1.09	3.87	3.61
Trombone	1.07	1.31	1.93	-	2.69	3.02	3.37	1.04	3.67	3.54
Tuba	1.06	1.19	1.13	-	1.72	1.69	2.70	2.59	3.41	3.20
Overall	1.48	1.21	2.33	-	2.48	3.25	3.46	1.42	3.80	3.85

Training from scratch (TS) Zero-shot (ZS) Fine-tuning (FT)

Audio samples

Natural

NSF-FT

More samples can be found here:
https://nii-yamagishilab.github.io/samples-nsf/neural-music.html

Conclusions

- Knowledge from the speech domain can be transferred to music signal modeling.
- NSF performed best in both the fine-tuning and training-from-scratch scenarios.
- WaveGlow showed great potential in zero-shot cross-domain adaptation

Future work

- Improve sounds of cello and violin
- Evaluate the neural waveform synthesizers on other corpora.
- Measure the impact of music data size on the training of neural waveform synthesizers
- Improve neural waveform synthesizers for polyphonic sounds
- Predict Mel-spectrograms from MIDI for end-to-end synthesis

Final demonstration:
Waltz, 3 instruments, generated
by the common NSF-FT model

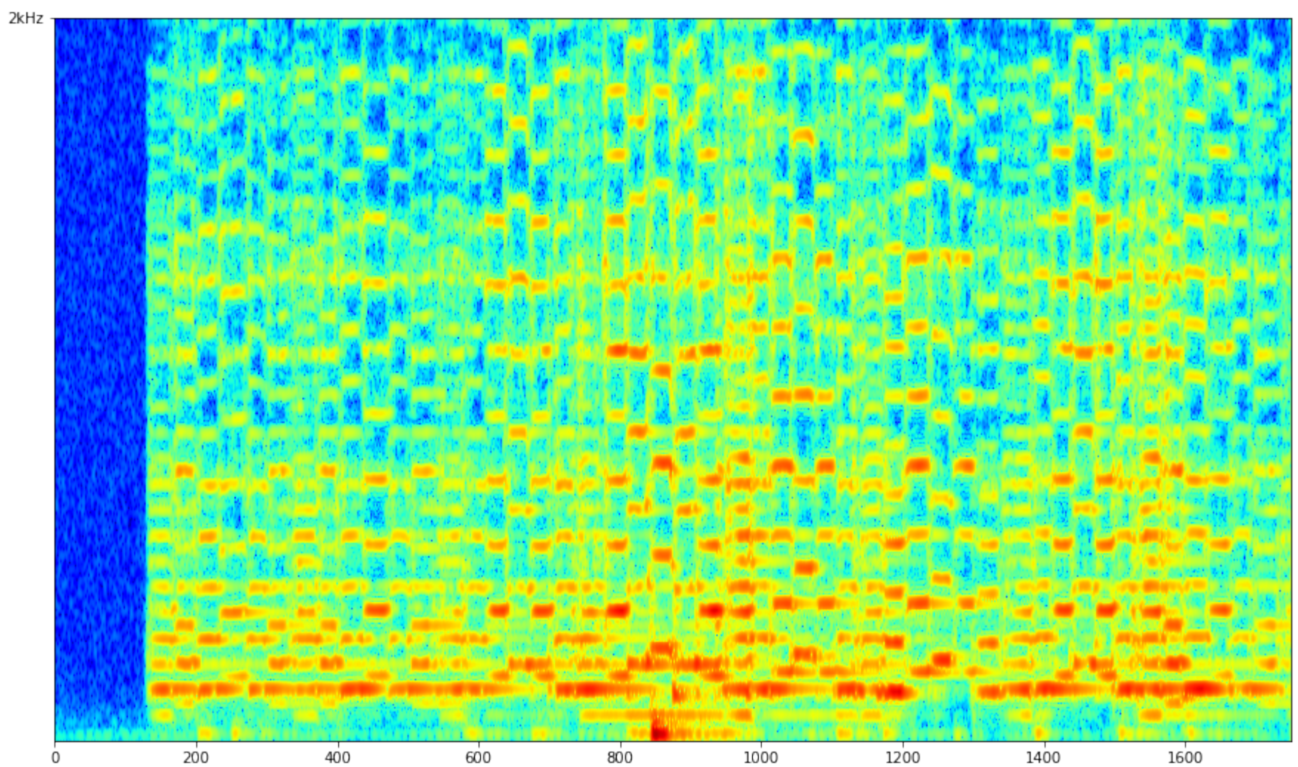
Thanks

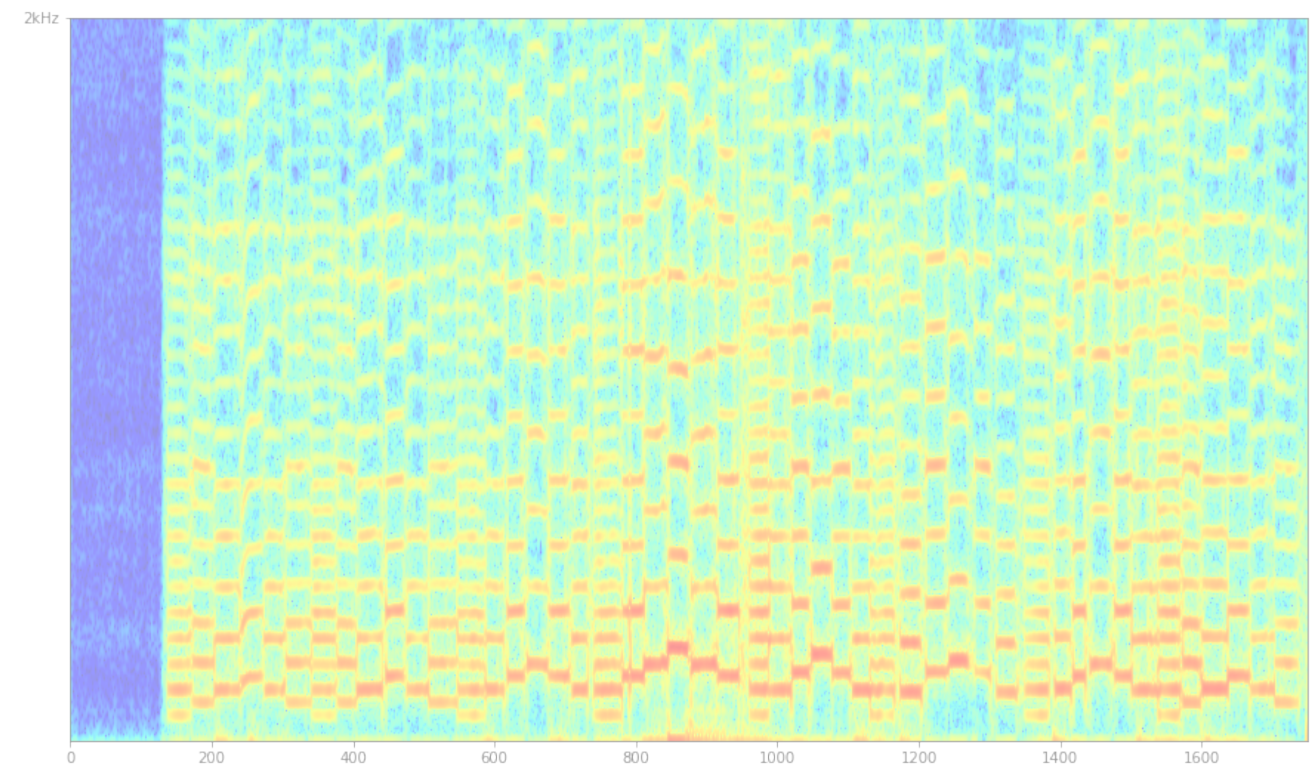
https://nii-yamagishilab.github.io/samples-nsf/neural-music.html

Future work

- improve quality for cello

Natural

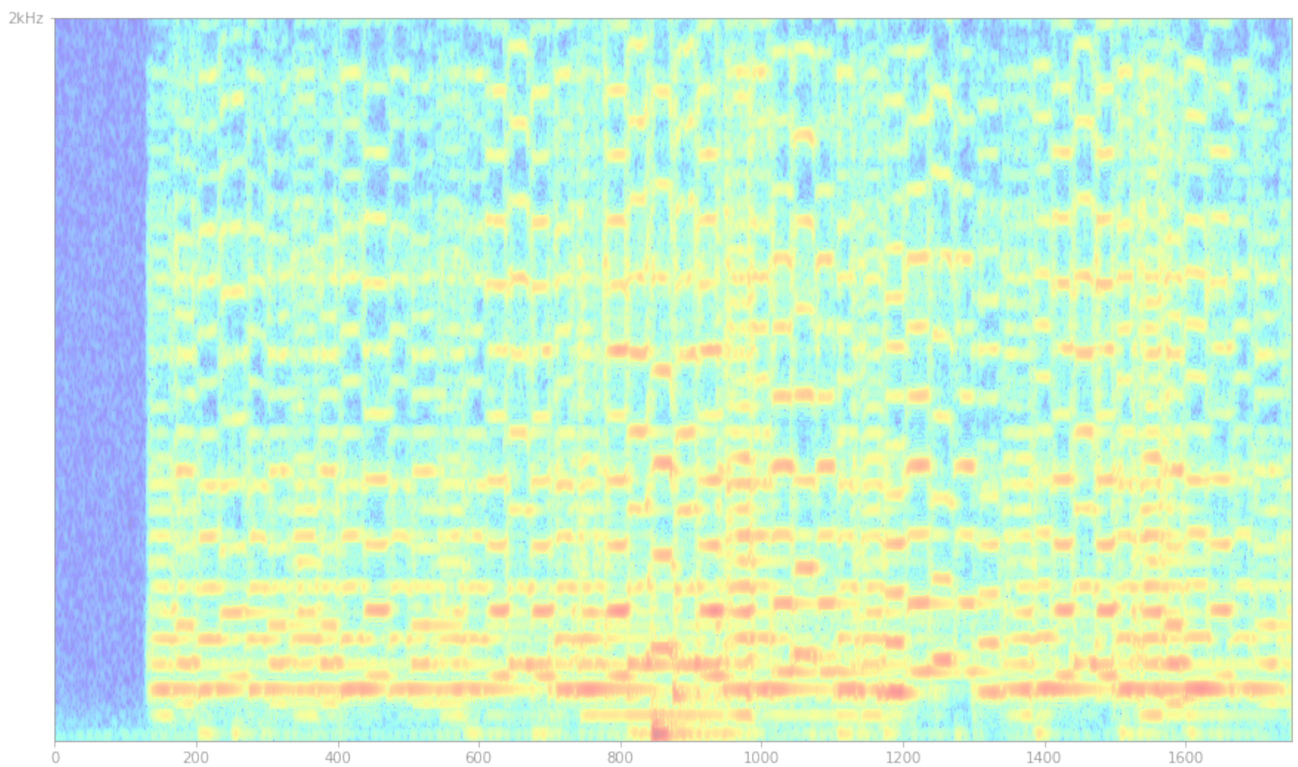


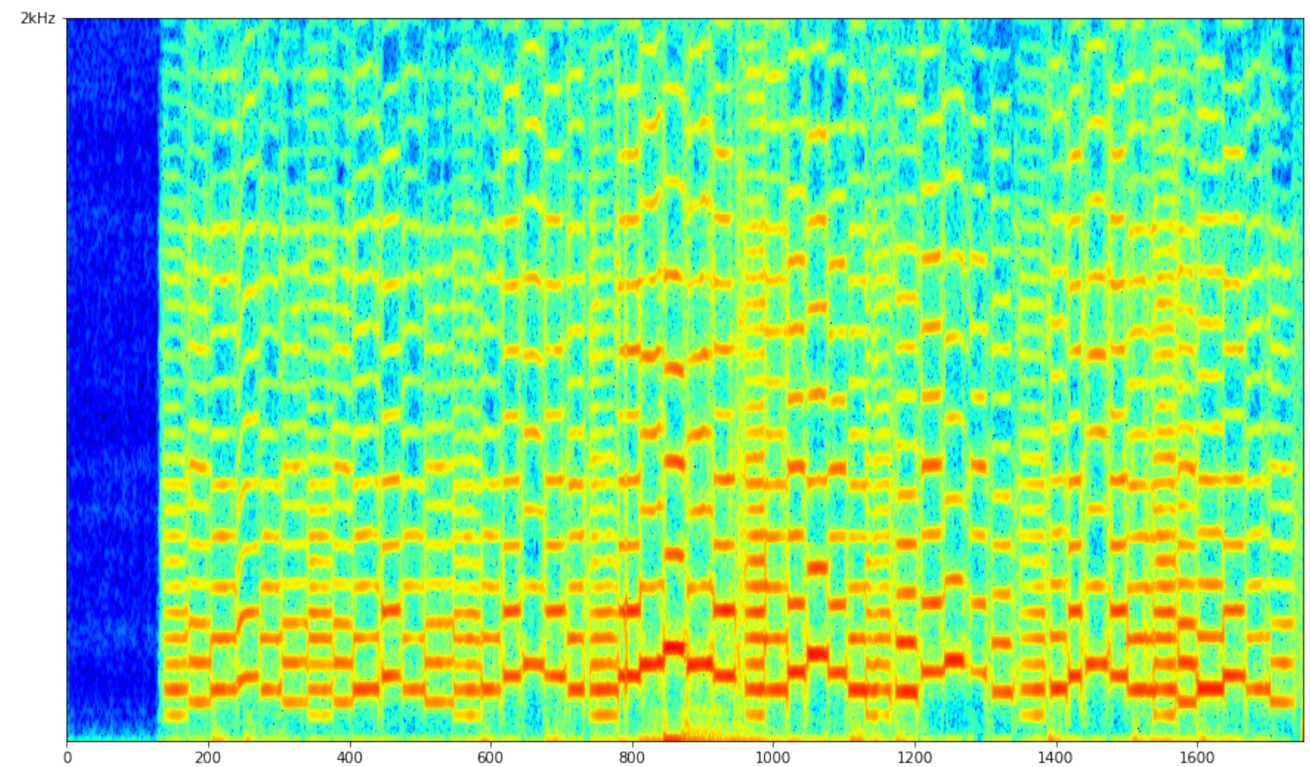


Future work

- improve quality for cello

NSF



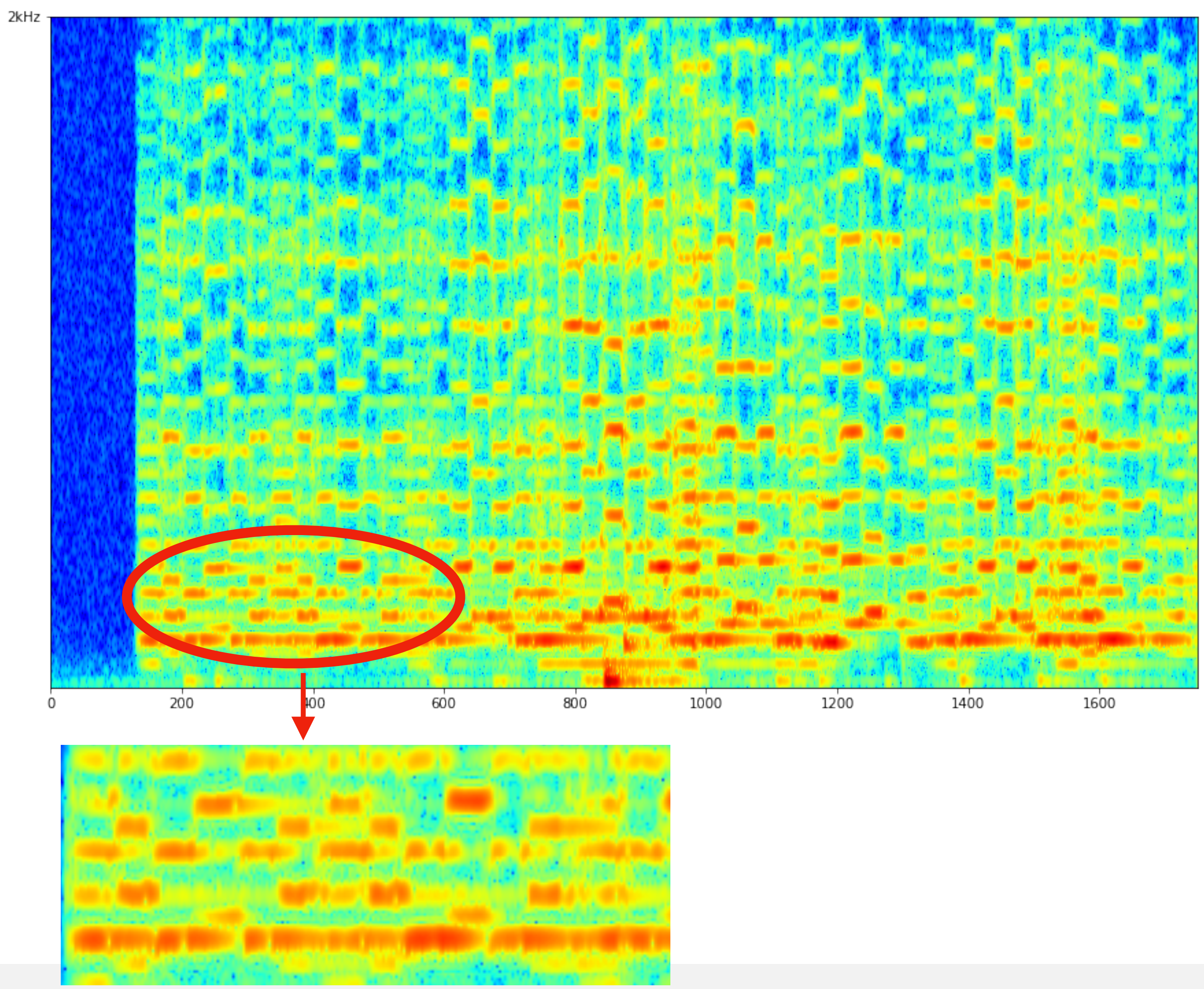


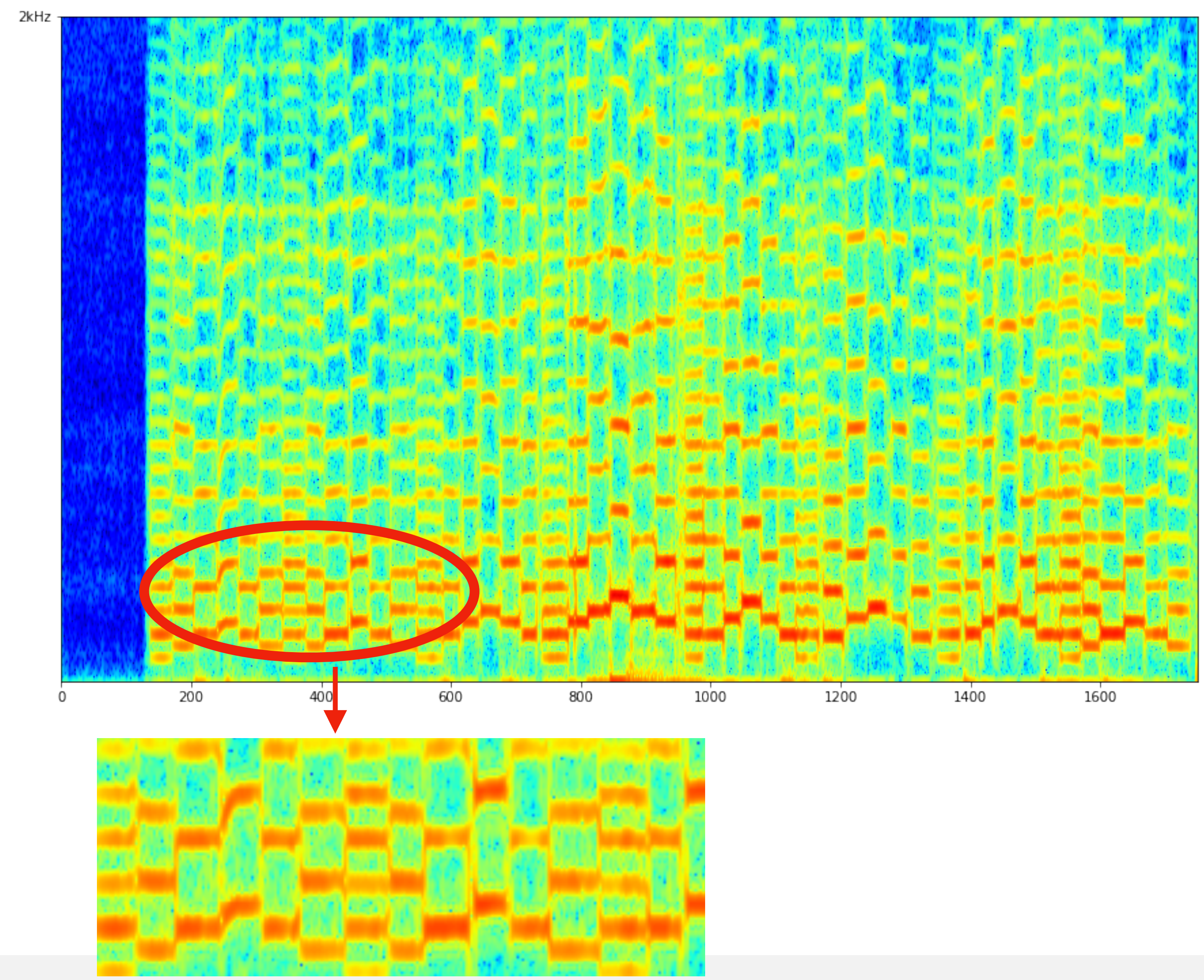
Future work

- improve quality for cello

Natural

NSF





Appendix

Protocol using URMP

Test set pieces

No.	Distinct Abbreviation	Full Piece Name	Type	Duration (mm:ss)	Instrumentations		
08	Spring	Spring from the Four Seasons	Duet	00:35	Fl.	Vn.	
10	March	March from Occasional Oratorio	Duet	01:43	Tpt.	Sax.	
11	Maria	Ave Maria	Duet	01:44	Ob.	Vc.	
13	Hark	Hark the herald angels sing	Trio	00:47	Vn.	Vn.	Va.
14	Waltz	Waltz from sleeping Beauty	Trio	01:33	Fl.	Fl.	Cl.

Vn.	Violin
Va.	Viola
Vc.	Cello
Db.	Double Bass
Fl.	Flute
Ob.	Oboe
Cl.	Clarinet
Sax.	Saxophone
Bn.	Bassoon
Tpt.	Trumpet
Hn.	Horn
Tbn.	Trombone
Tba.	Tuba

http://www2.ece.rochester.edu/projects/air/projects/URMP/URMP_doc.pdf

Appendix

Protocol using URMP

Training & development sets (grey)

No.	Distinct Abbreviation	Full Piece Name	Type	Duration (mm:ss)	Instrumentations		
01	Jupiter	Jupiter	Duet	01:03	Vn.	Vc.	
02	Sonata	Sonata	Duet	00:46	Vn.	Vn.	
03	Dance	Dance of the Sugar Plum Fairy	Duet	01:37	Fl.	Cl.	
04	Allegro	Allegro for Musical Clock	Duet	02:37	Fl.	Fl.	
05	Entertainer	The Entertainer	Duet	01:27	Tpt.	Tpt.	
06	Entertainer	The Entertainer	Duet	01:28	Sax.	Sax.	
07	GString	Air on the G string	Duet	04:30	Tpt.	Tbn.	
09	Jesus	Jesus Bleibet Meine Freude	Duet	03:19	Tpt.	Vn.	
12	Spring	Spring from the Four Seasons	Trio	02:11	Vn.	Vn.	Vc.
15	Surprise	Theme from Surprise Symphony	Trio	00:53	Tpt.	Tpt.	Tbn.
16	Surprise	Theme from Surprise Symphony	Trio	00:53	Tpt.	Tpt.	Sax.
17	Nocturne	Nocturne	Trio	01:36	Vn.	Fl.	Cl.
18	Nocturne	Nocturne	Trio	01:36	Vn.	Fl.	Tpt.
19	Pavane	Pavane	Trio	02:13	Cl.	Vn.	Vc.
20	Pavane	Pavane	Trio	02:13	Tpt.	Vn.	Vc.
21	Rejouissance	La Rejouissance	Trio	01:16	Cl.	Tbn.	Tba.

Appendix

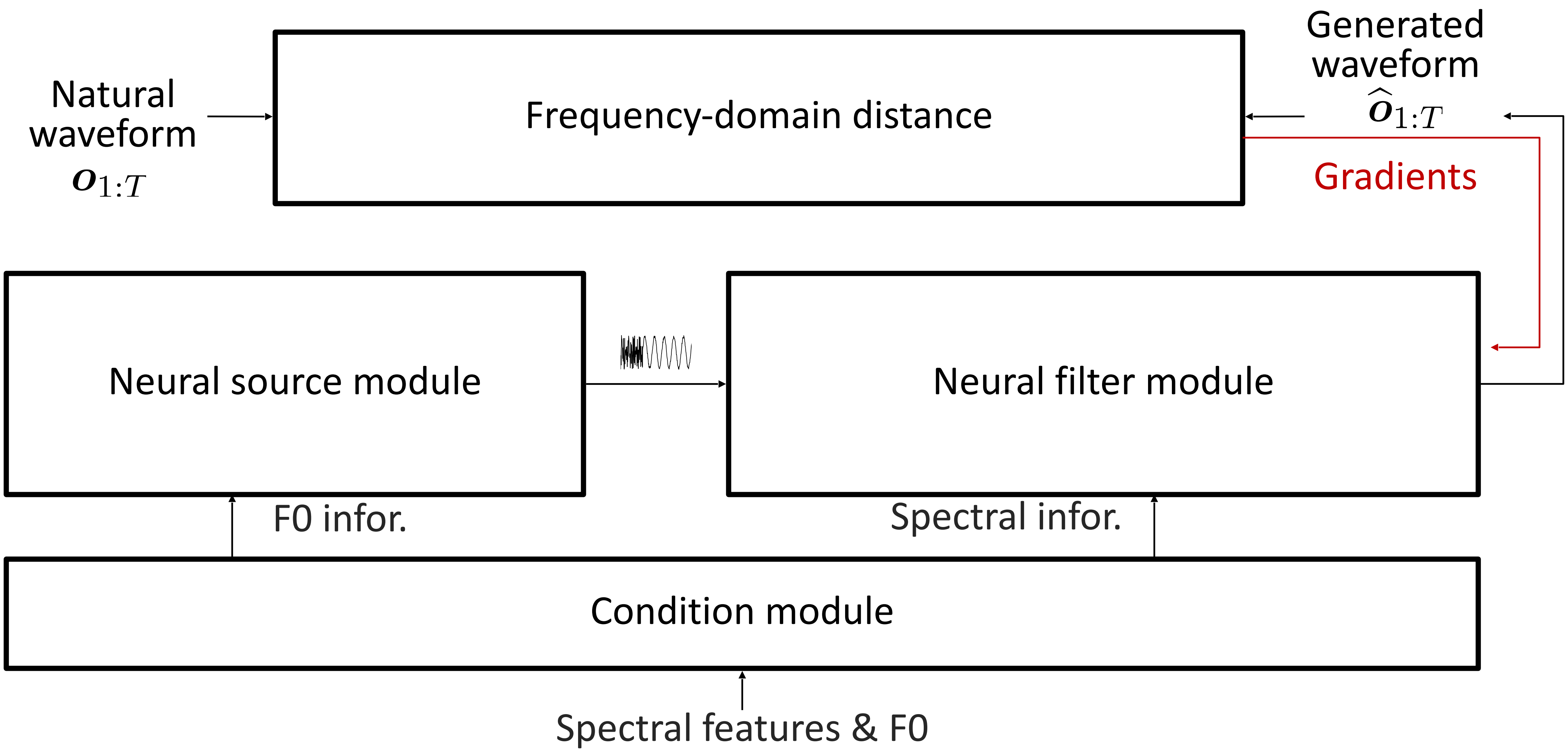
Protocol using URMP

Training & development sets

22	Rejouissance	La Rejouissance	Trio	01:16	Sax.	Tbn.	Tba.		
23	Rejouissance	La Rejouissance	Trio	01:16	Cl.	Sax.	Tba.		
24	Pirates	Pirates of the Aegean	Quartet	00:50	Vn.	Vn.	Va.	Vc.	
25	Pirates	Pirates of the Aegean	Quartet	00:50	Vn.	Vn.	Va.	Sax.	
26	King	In the Hall of the Mountain King	Quartet	01:25	Vn.	Vn.	Va.	Vc.	
27	King	In the Hall of the Mountain King	Quartet	01:25	Vn.	Vn.	Va.	Sax.	
28	Fugue	The Art of the Fugue	Quartet	02:55	Fl.	Ob.	Cl.	Bn.	
29	Fugue	The Art of the Fugue	Quartet	02:52	Fl.	Fl.	Ob.	Cl.	
30	Fugue	The Art of the Fugue	Quartet	02:52	Fl.	Fl.	Ob.	Sax.	
31	Slavonic	Slavonic Dance	Quartet	01:22	Tpt.	Tpt.	Hn.	Tbn.	
32	Fugue	The Art of the Fugue	Quartet	02:54	Vn.	Vn.	Va.	Vc.	
33	Elise	Fur Elise	Quartet	00:48	Tpt.	Tpt.	Hn.	Tbn.	
34	Fugue	The Art of the Fugue	Quartet	02:54	Tpt.	Tpt.	Hn.	Tbn.	
35	Rondeau	Rondeau from Abdelazer	Quartet	02:08	Vn.	Vn.	Va.	Db.	
36	Rondeau	Rondeau from Abdelazer	Quartet	02:08	Vn.	Vn.	Va.	Vc.	
37	Rondeau	Rondeau from Abdelazer	Quartet	02:08	Fl.	Vn.	Va.	Cl.	
38	Jerusalem	Jerusalem	Quintet	01:59	Vn.	Vn.	Va.	Vc.	Db.
39	Jerusalem	Jerusalem	Quintet	01:59	Vn.	Vn.	Va.	Sax.	Db.
40	Miserere	Miserere Mei Deus	Quintet	00:40	Fl.	Fl.	Ob.	Cl.	Bn.
41	Miserere	Miserere Mei Deus	Quintet	00:40	Fl.	Fl.	Ob.	Sax.	Bn.
42	Arioso	Arioso	Quintet	04:15	Tpt.	Tpt.	Hn.	Tbn.	Tba.
44	K515	String Quintet K515	Quintet	03:45	Vn.	Vn.	Va.	Va.	Db.

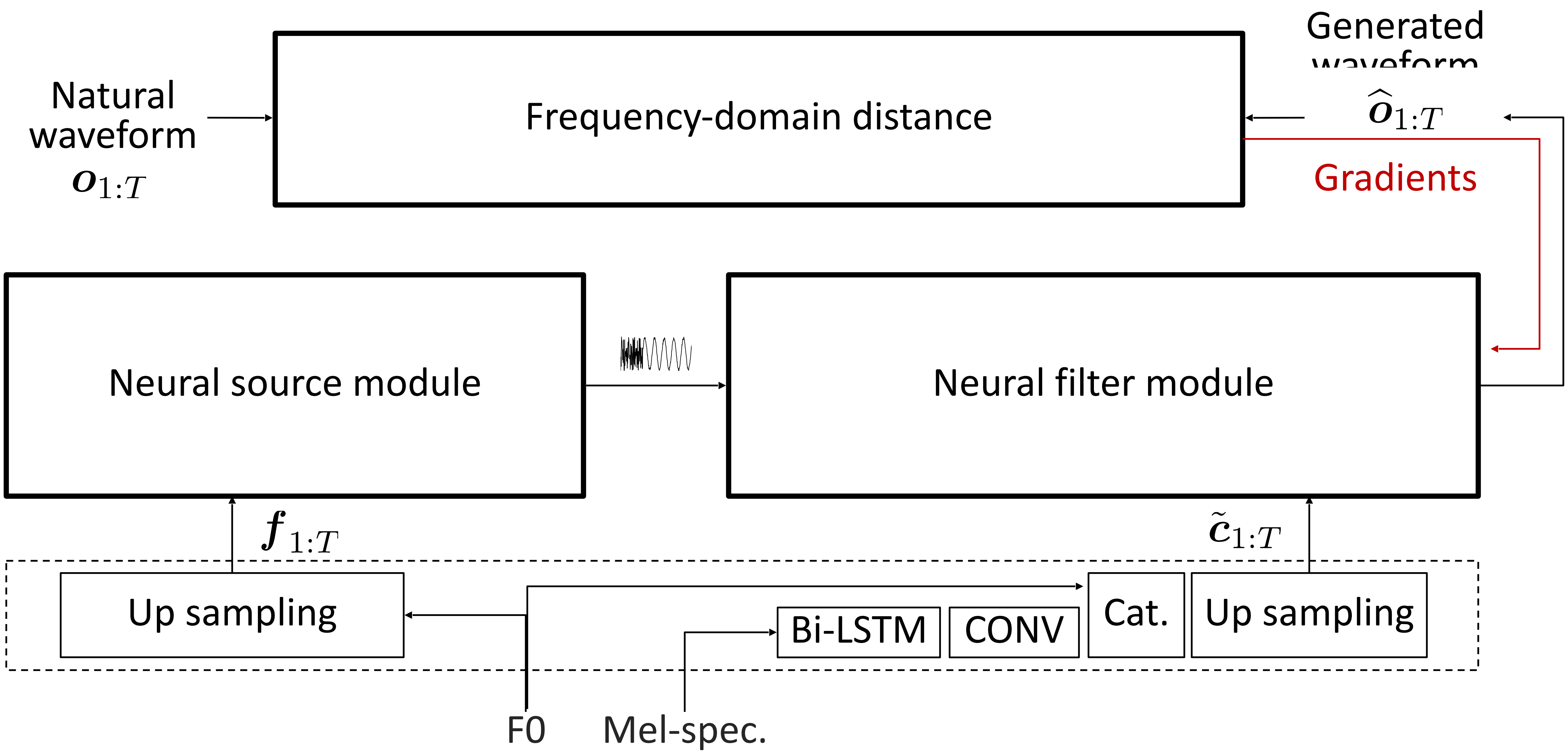
Appendix

Neural source-filter waveform models



Appendix

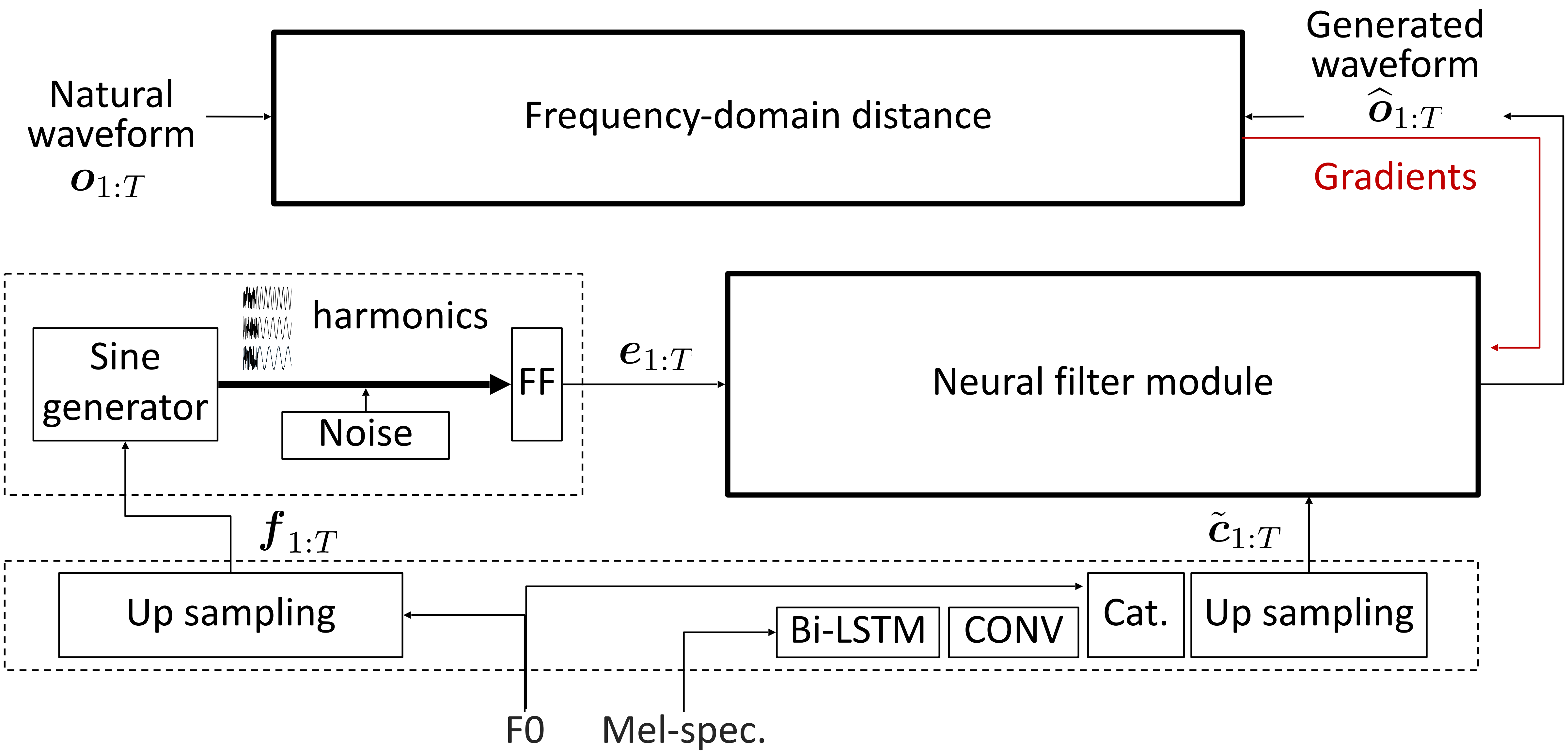
Neural source-filter waveform models



Condition module: change feature dimension & do up-sampling

Appendix

Neural source-filter waveform models

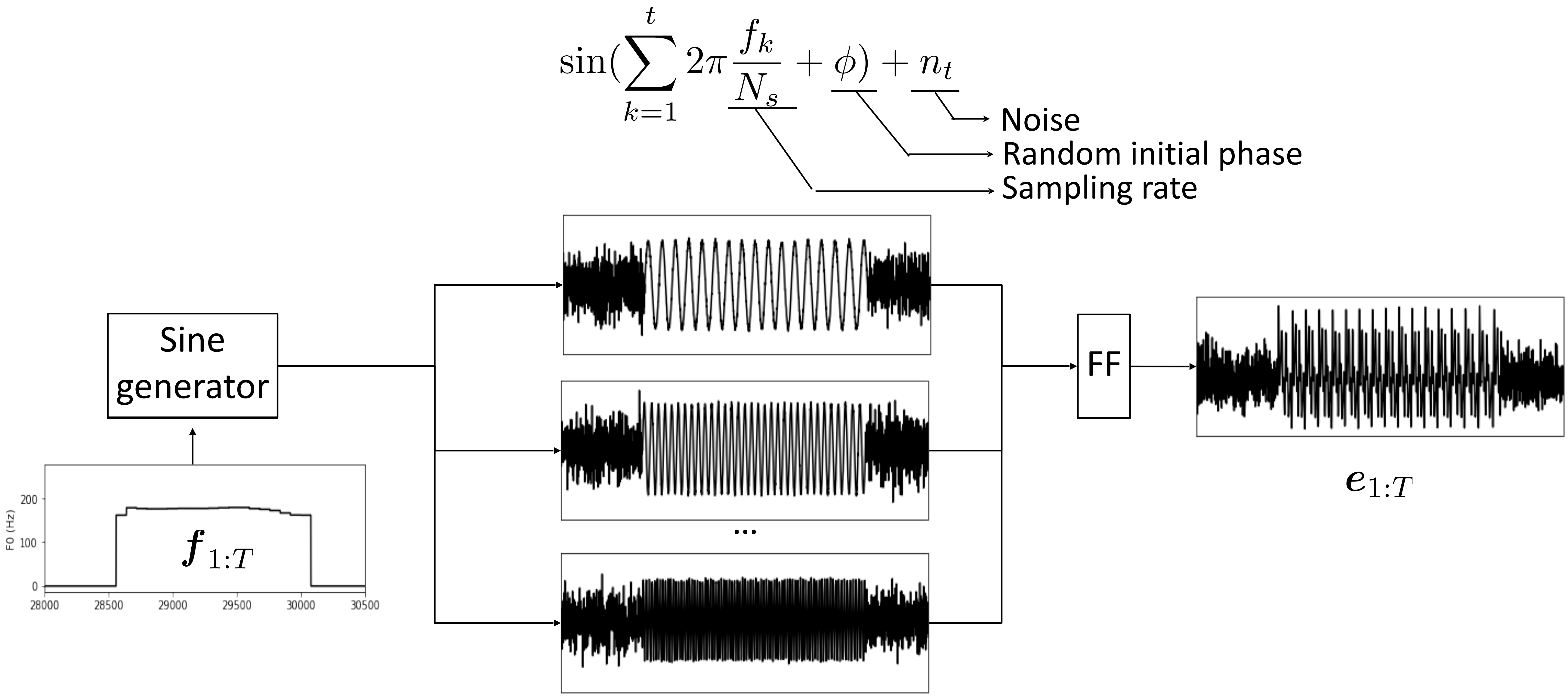


Sine-based source signal $e_t \in \mathbb{R}, \forall t \in \{1, \cdots, T\}$ given F0 $f_t \in \{0\} \cup \mathbb{R}^+$

- FF: feedforward layer with Tanh

Appendix

Neural source-filter waveform models

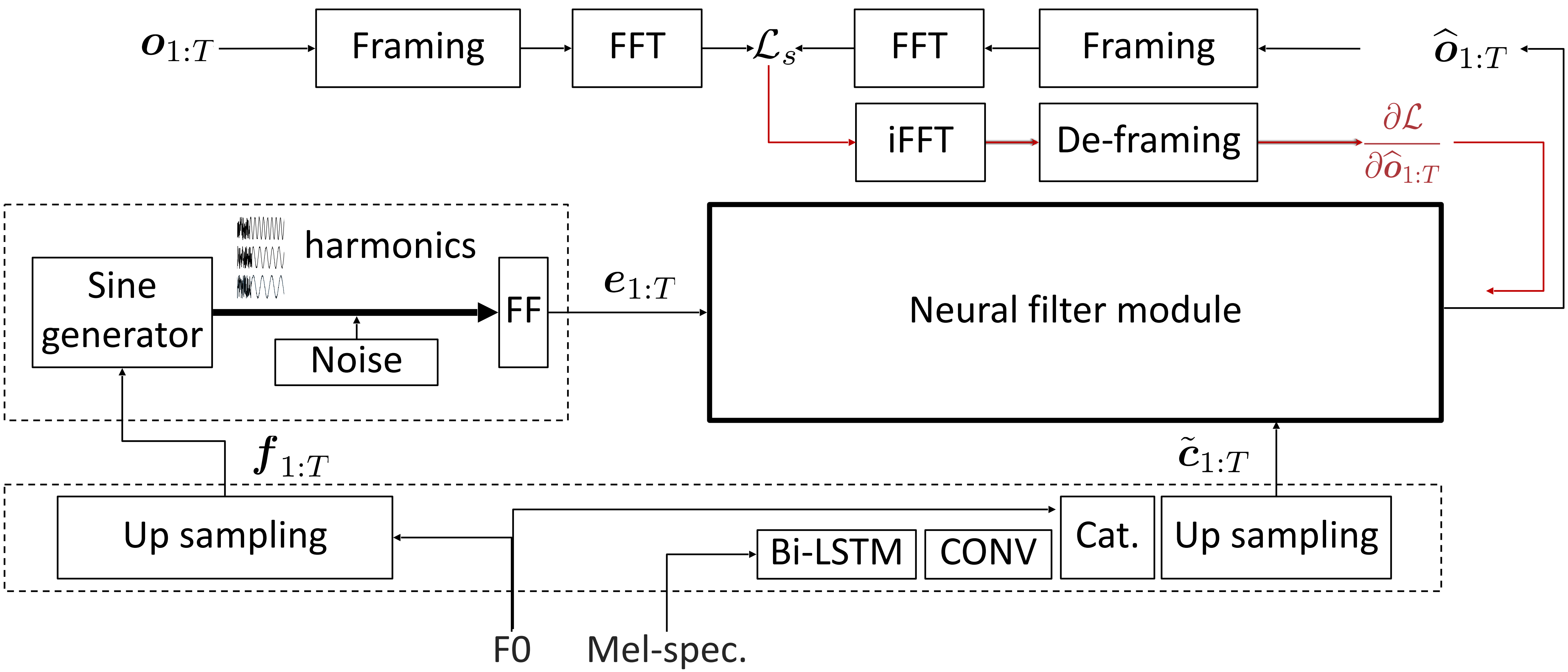


Sine-based source signal $e_t \in \mathbb{R}, \forall t \in \{1, \cdots, T\}$ given F0 $f_t \in \{0\} \cup \mathbb{R}^+$

- FF: feedforward layer with Tanh

Appendix

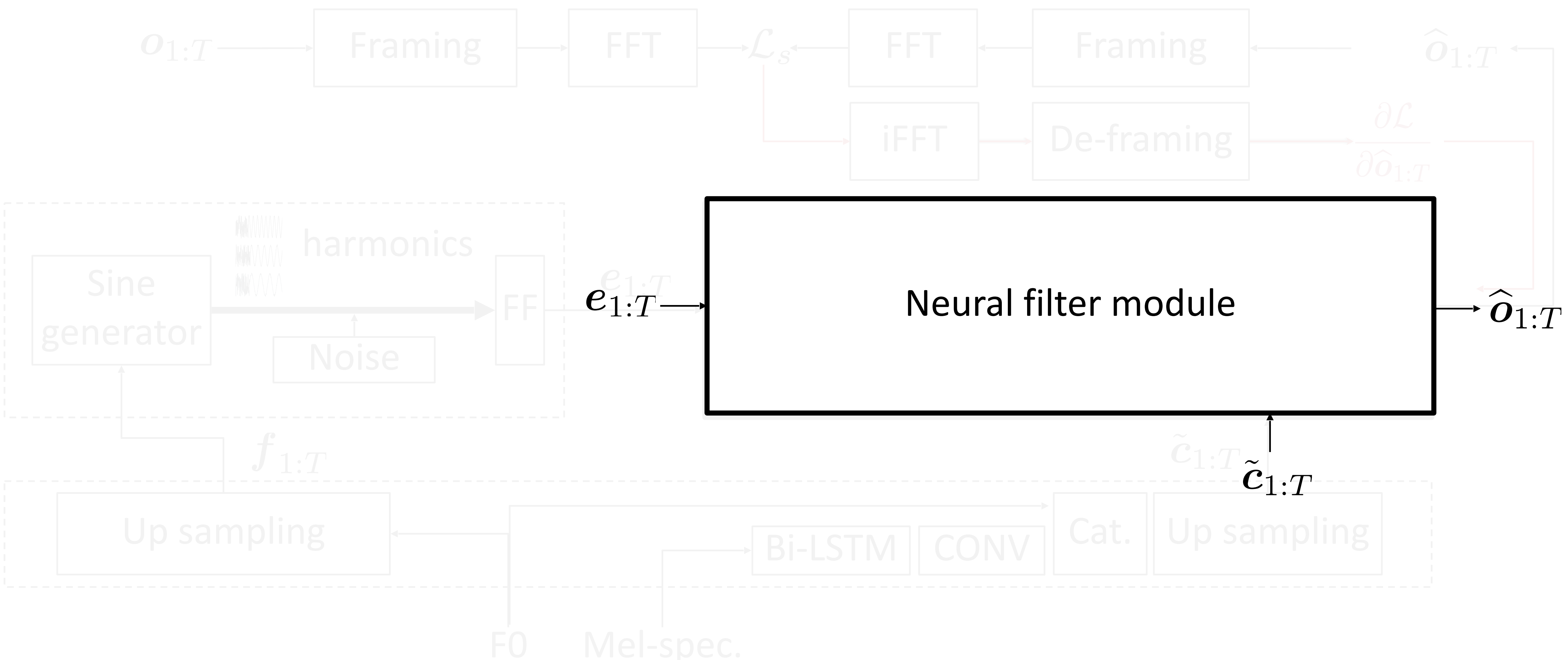
Neural source-filter waveform models



Multiple L_s : different frame shift & length

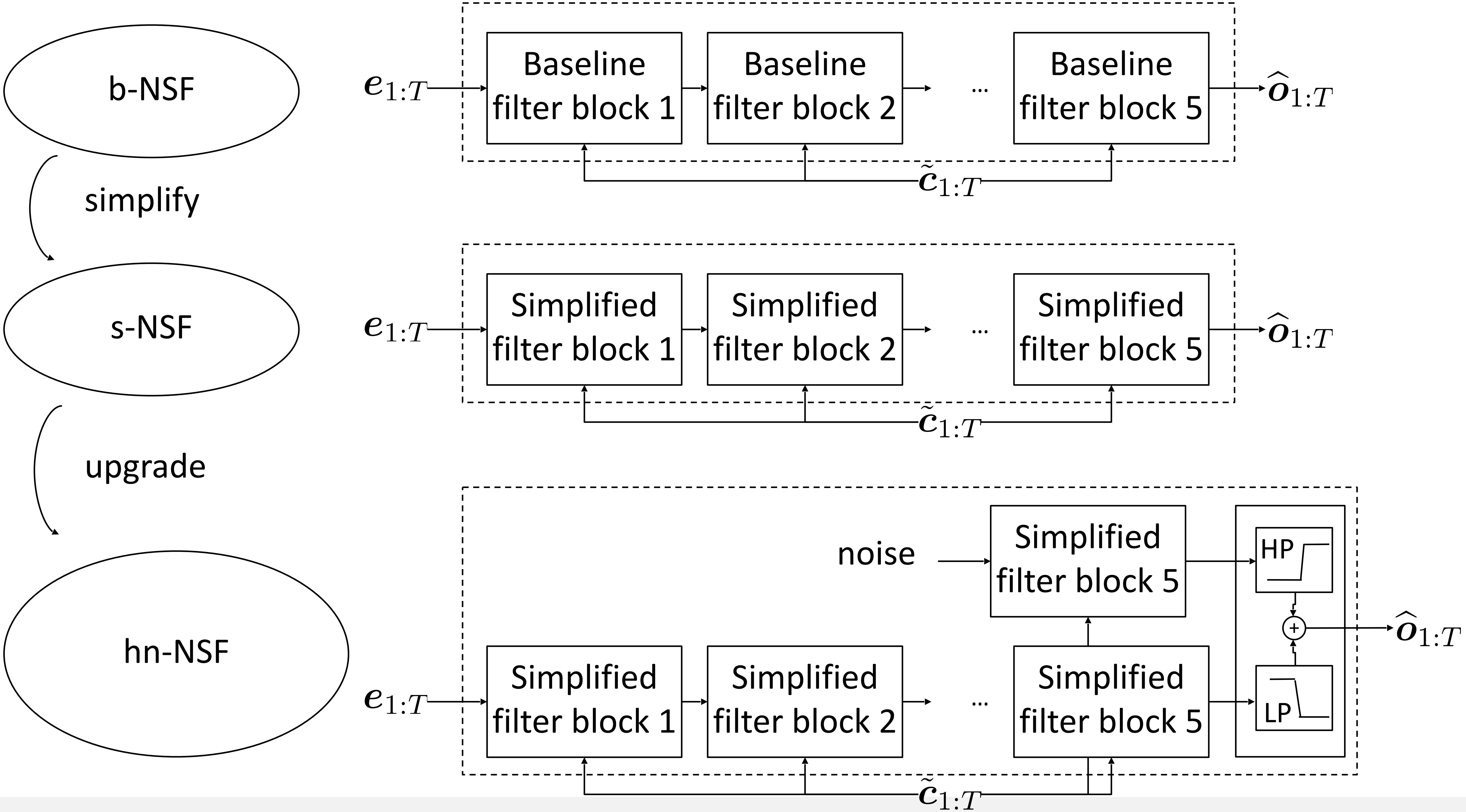
Appendix

Neural source-filter waveform models



Appendix

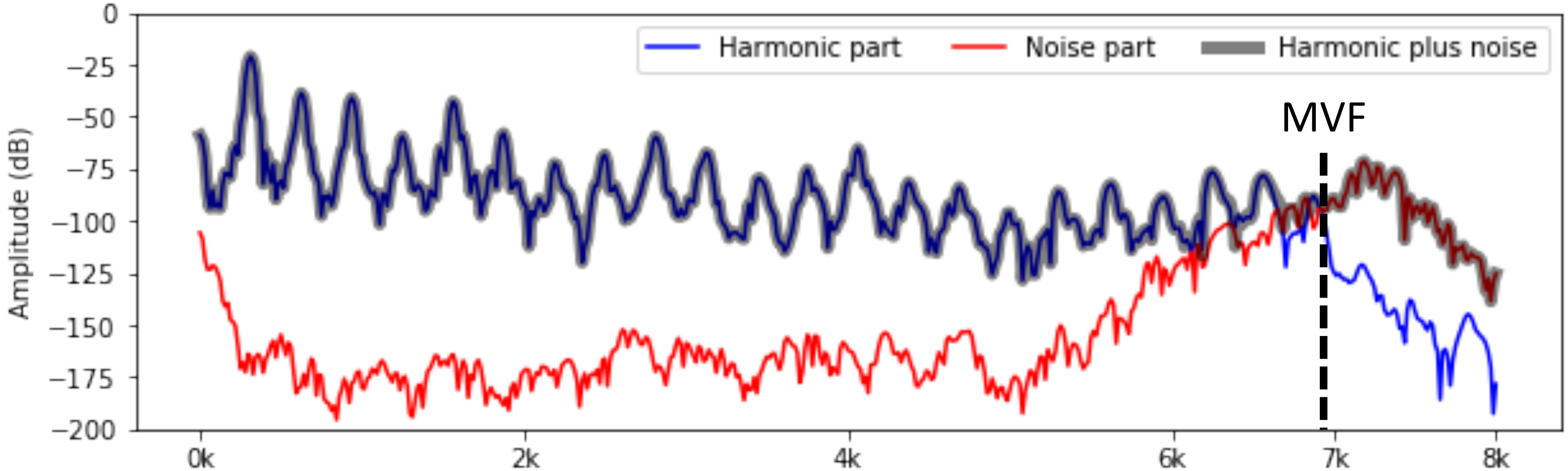
Neural source-filter waveform models



- HP, LP: high- and low-pass finite-impulse-response (FIR) filter

Appendix

Neural source-filter waveform models



Maximum voiced frequency is predicted from input features