IEEE ICASSP 2023
Paper ID: 3191
Session: SLT-P26
Type: Poster
Date: Wed., June 7
Time: 15:35 ~ 17:05

NNSVS: A Neural Network-Based Singing Voice Synthesis Toolkit

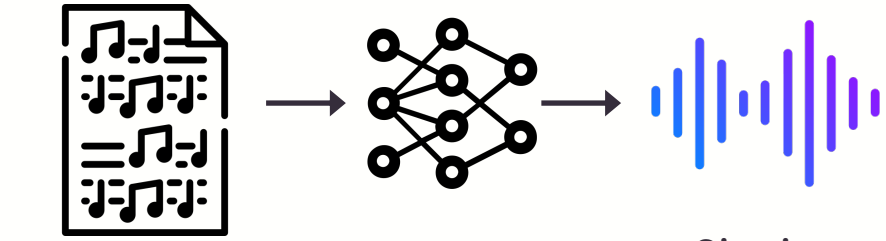
https://github.com/nnsvs/nnsvs

Ryuichi Yamamoto [1,2], Reo Yoneyama [2], Tomoki Toda [2]
[1] LINE Corp., Japan.
[2] Nagoya University, Japan.

Samples

What is NNSVS?

Neural-Network-based Singing Voice Synthesis toolkit for research



Features

- Everything is open-source
- Complete recipes for reproducible research [Watanabe+2018]
- High naturalness

[Watanabe+2018] “ESPnet: End-to-End Speech Processing Toolkit”, in Proc. Interspeech, 2018.

Why we need a new toolkit for SVS?

Sinsy (2002 ~ current) [Hono+2021]

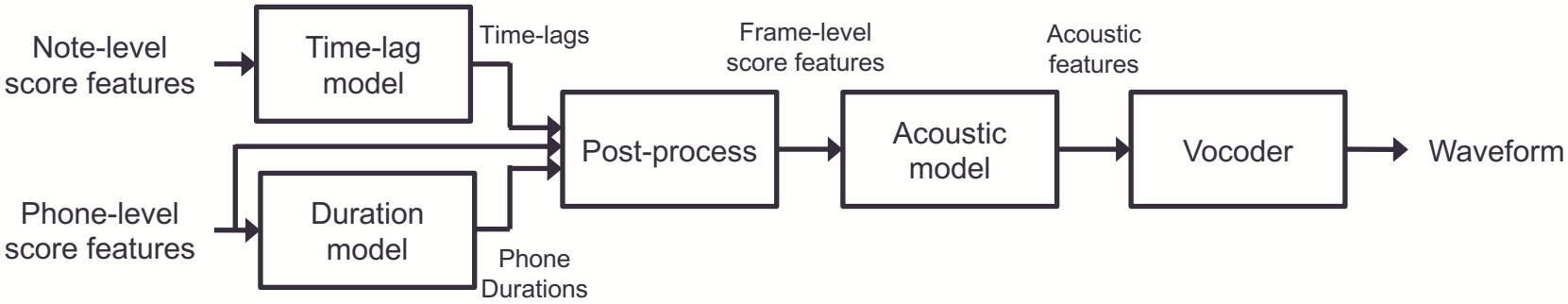
- Limited functionality
  - Public version still relies on the traditional parametric method based on HMMs
  - New DNN version is not publicly available
- Open-source version is outdated
  - Last public release was at December, 2015.

Muskits (2021 ~ current) [Shi+2022]

- Towards end-to-end systems
- It does not support parametric models such as Sinsy

[Hono+2021] “Sinsy: a deep neural network-based singing voice synthesis system”, IEEE/ACM Trans. on Audio, Speech, and Lang. Process., 29:2803–2815, 2021.
[Shi+2022] “Muskits: an End-to-End Music Processing Toolkit for Singing Voice Synthesis”, in Proc. Interspeech, 2022.

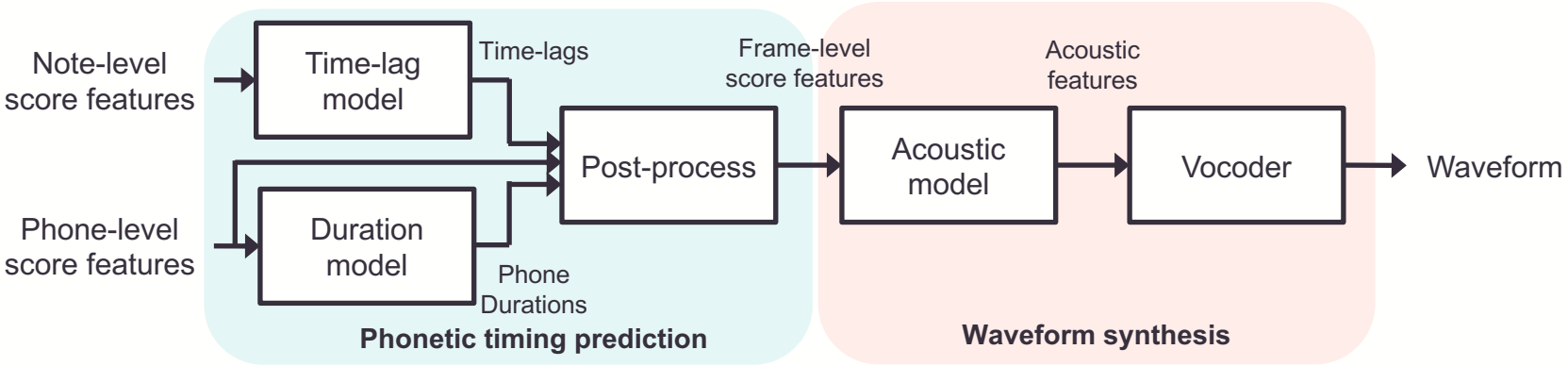
Overview of NNSVS's SVS system



A pipeline system that mimics the traditional statistical parametric speech synthesis (SPSS)

- Score/acoustic features are used as intermediate features
- Each module can be flexibly configured by design

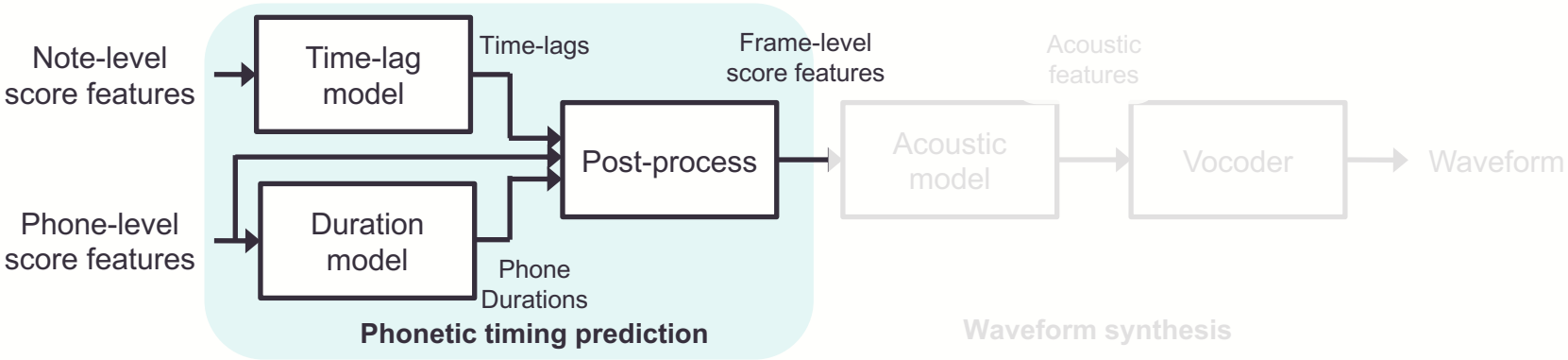
Overview of NNSVS's SVS system



A pipeline system that mimics the traditional statistical parametric speech synthesis (SPSS)

- Score/acoustic features are used as intermediate features
- Each module can be flexibly configured by design

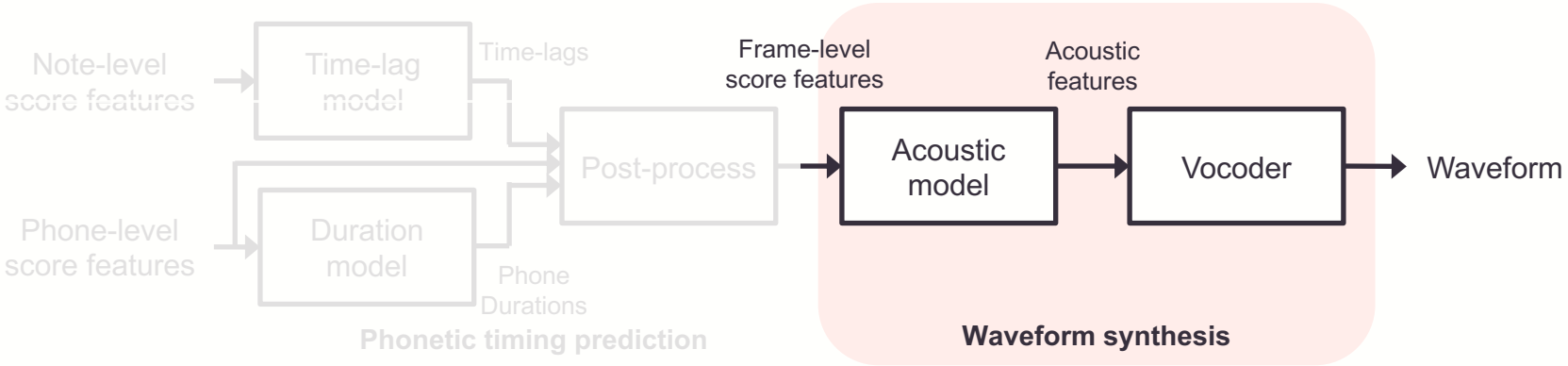
Overview of NNSVS's SVS system



A pipeline system that mimics the traditional statistical parametric speech synthesis (SPSS)

- Score/acoustic features are used as intermediate features
- Each module can be flexibly configured by design

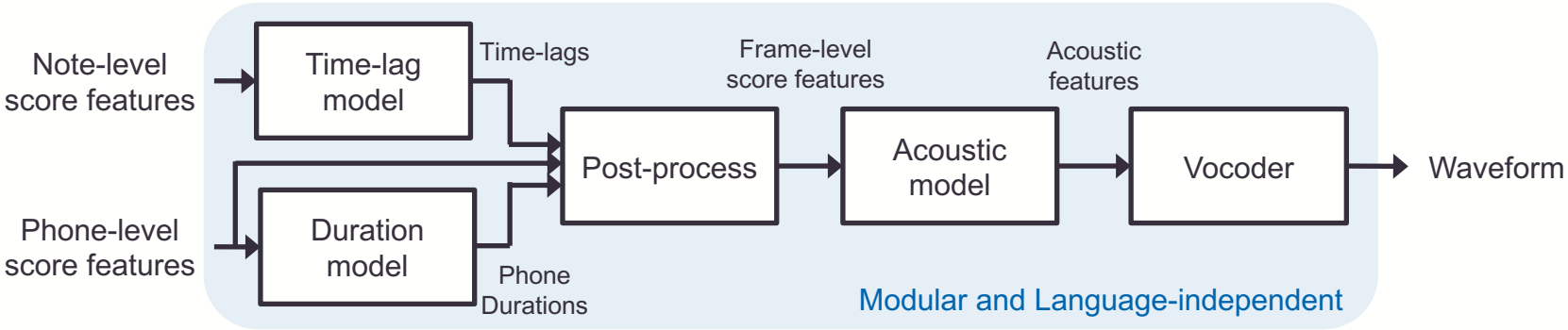
Overview of NNSVS's SVS system



A pipeline system that mimics the traditional statistical parametric speech synthesis (SPSS)

- Score/acoustic features are used as intermediate features
- Each module can be flexibly configured by design

Highlights



Design

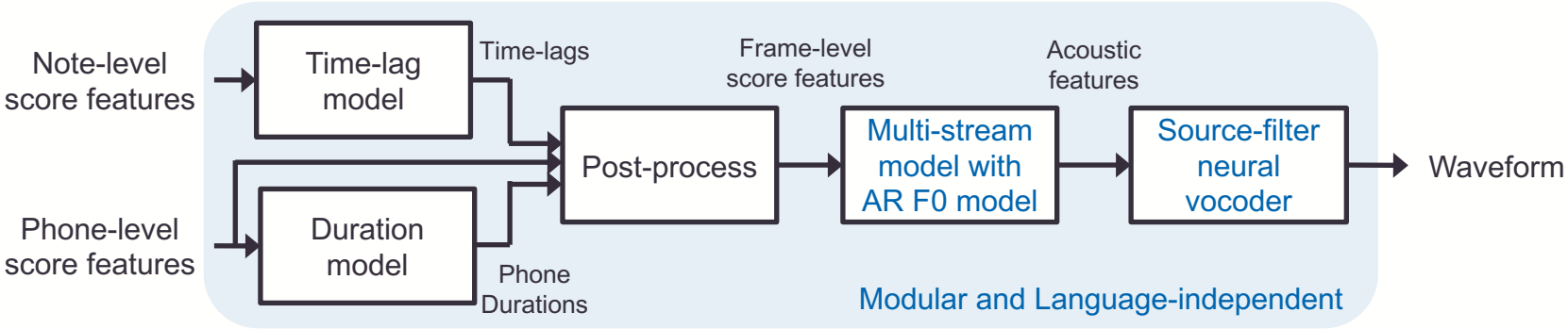
- Modular
- Language-independent

Models

- Multi-stream acoustic model
- Autoregressive(AR) F0 models
- Source-filter neural vocoders (hn-uSFGAN) [Yoneyama+2022]

[Yoneyama+2022] “Unified Source-Filter GAN with Harmonic-plus-Noise Source Excitation Generation”, Interspeech, 2022.

Highlights



Design

- Modular
- Language-independent

Models

- Multi-stream acoustic model
- Autoregressive(AR) F0 models
- Source-filter neural vocoders (hn-uSFGAN) [Yoneyama+2022]

[Yoneyama+2022] “Unified Source-Filter GAN with Harmonic-plus-Noise Source Excitation Generation”, Interspeech, 2022.

Experimental conditions

Database

- Namine Ritsu
- 110 songs, 4.35 hours (silence excluded)

Acoustic features

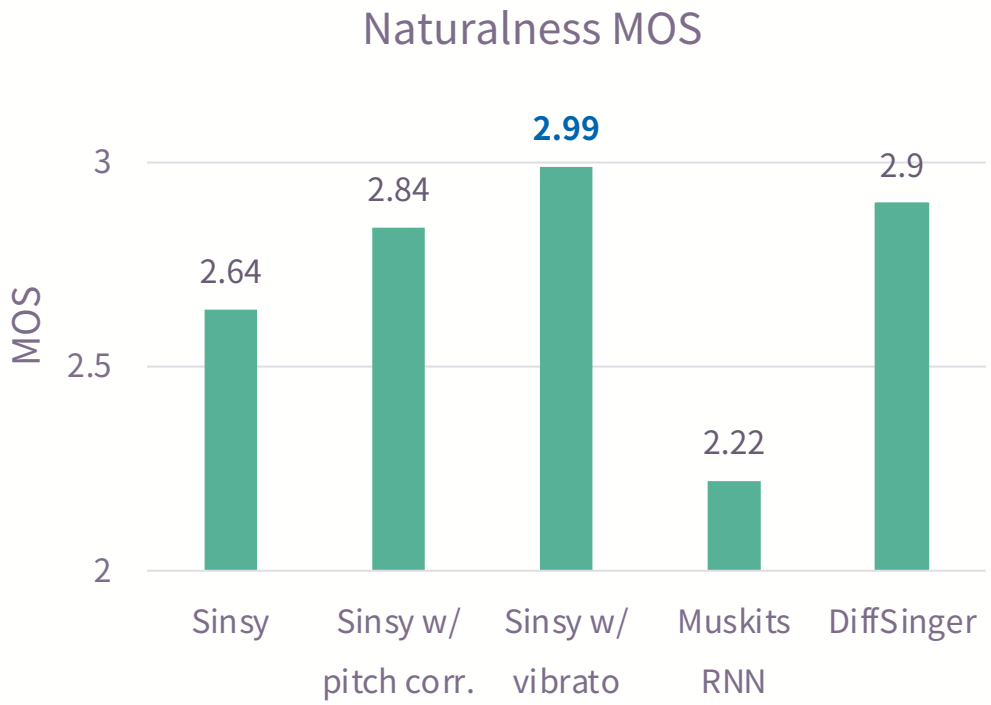
- Mel-spectrogram (MEL): 80-dim
- WORLD-features: [MGC, LF0, VUV, BAP] that consists of 67-dim ([60, 1, 1, 3]) features

Baseline systems

System	Acoustic features	Multi-stream Architecture	Autoregressive Streams	Vocoder
Sinsy	MGC, LF0, VUV, BAP	No	-	hn-uSFGAN
Sinsy w/ pitch correction	MGC, LF0, VUV, BAP	No	-	hn-uSFGAN
Sinsy w/ vibrato modeling	MGC, LF0, VUV, BAP, VIB	No	-	hn-uSFGAN
Muskits RNN [1]	MEL	No		HiFi-GAN
DiffSinger [2]	MEL,LF0,VUV	Yes	-	hn-HiFi-GAN

[1] https://github.com/SJTMusicTeam/Muskits
[2] https://github.com/MoonInTheRiver/DiffSinger

Naturalness MOS test results for baseline systems



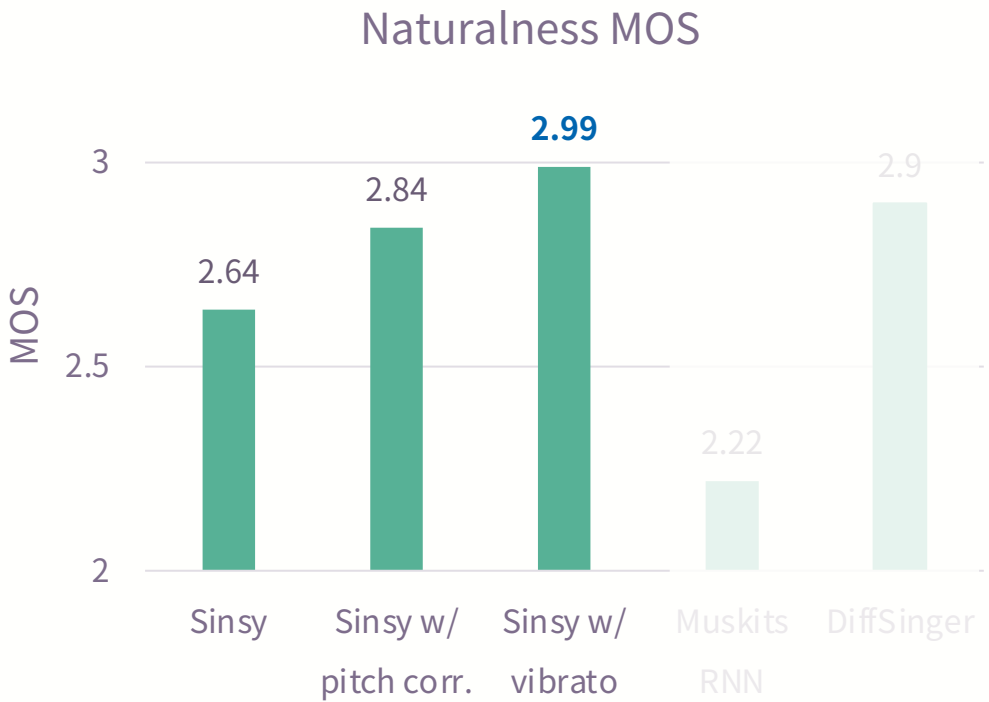
Sinsy with vibrato modeling performed best among three Sinsy systems

→ Demonstrated the importance of F0 modeling

Our reproduction of Sinsy performed comparable to

→ Parametric SVS can still achieve good results

Naturalness MOS test results for baseline systems



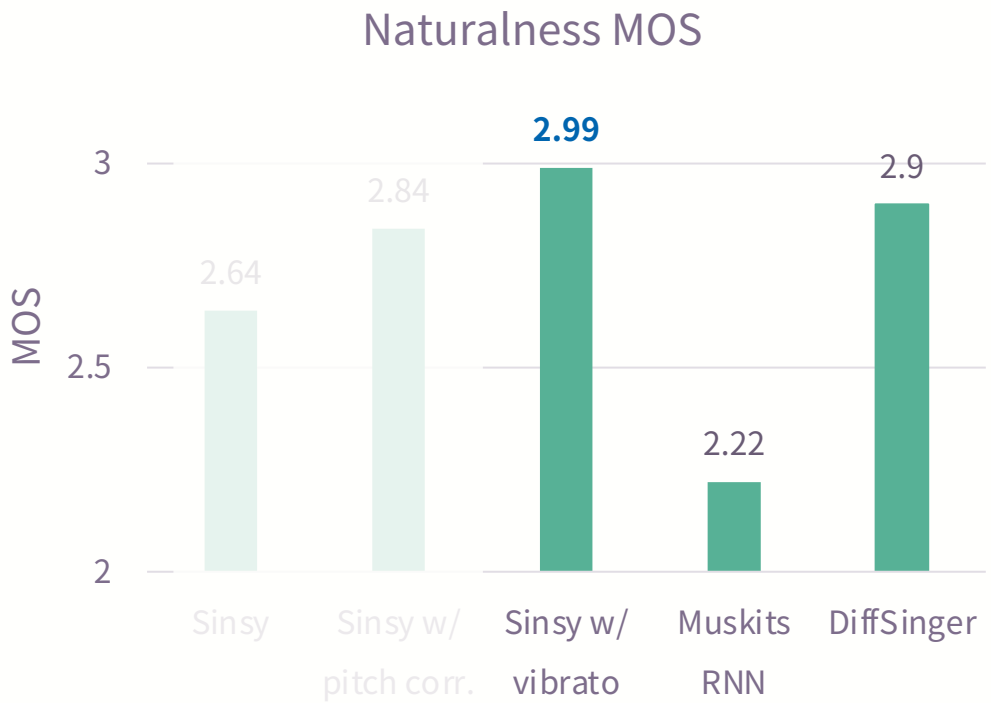
Sinsy with vibrato modeling performed best among three Sinsy systems

→ Demonstrated the importance of F0 modeling

Our reproduction of Sinsy performed comparable to

→ Parametric SVS can still achieve good results

Naturalness MOS test results for baseline systems



Sinsy with vibrato modeling performed best among three Sinsy systems

→ Demonstrated the importance of F0 modeling

Our reproduction of Sinsy performed comparable to

→ Parametric SVS can still achieve good results

NNSVS systems

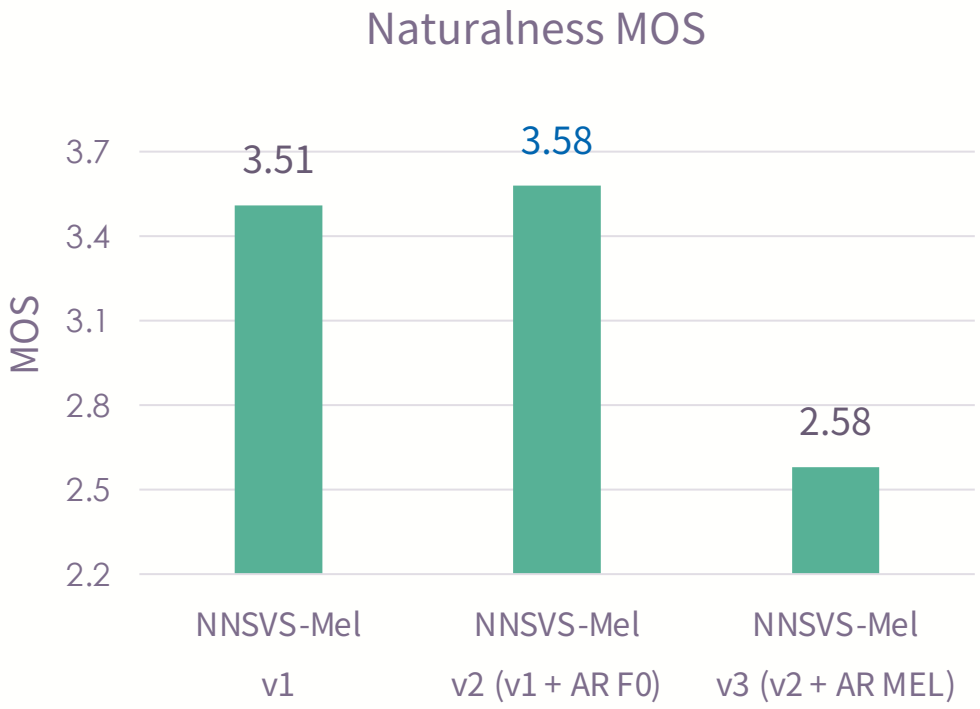
Acoustic features

- Mel-spectrogram (MEL): 80-dim
- WORLD-features: [MGC, LF0, VUV, BAP] that consists of 67-dim ([60, 1, 1, 3]) features

NNSVS systems

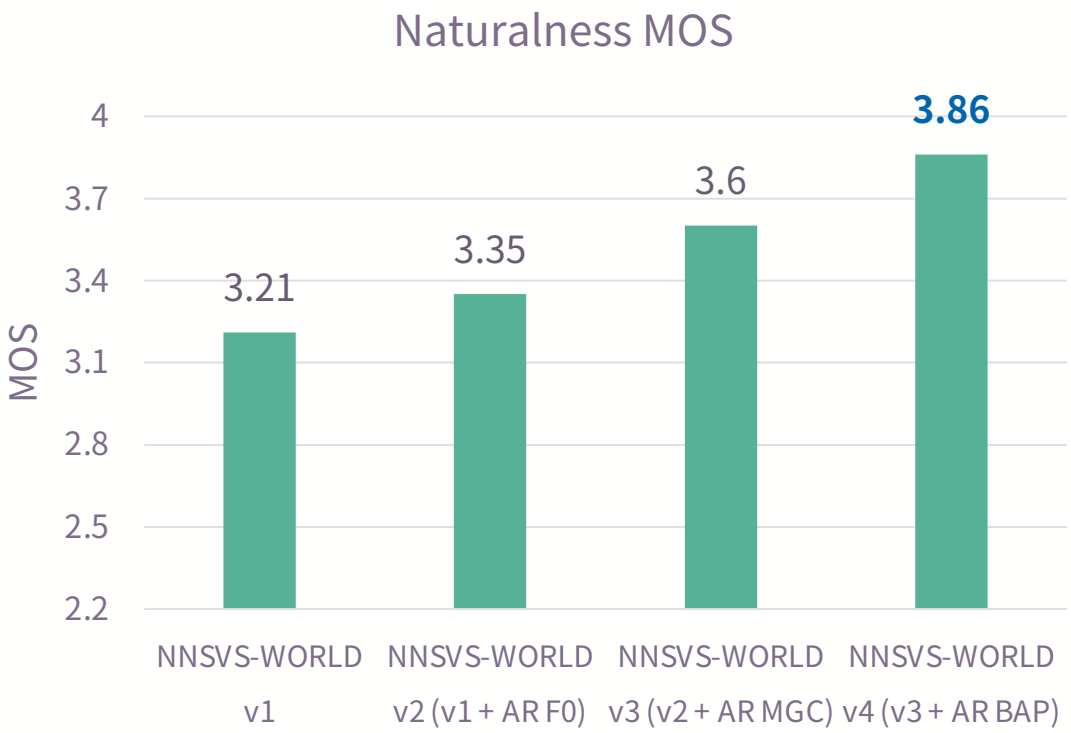
System	Acoustic features	Multi-stream Architecture	Autoregressive Streams	Vocoder
NNSVS-Mel v1	MEL, LF0, VUV	Yes	-	hn-uSFGAN
NNSVS-Mel v2	MEL, LF0, VUV	Yes	LF0	hn-uSFGAN
NNSVS-Mel v3	MEL, LF0, VUV	Yes	Mel, LF0	hn-uSFGAN
NNSVS-WORLD v1	MGC, LF0, VUV, BAP	Yes	-	hn-uSFGAN
NNSVS-WORLD v2	MGC, LF0, VUV, BAP	Yes	LF0	hn-uSFGAN
NNSVS-WORLD v3	MGC, LF0, VUV, BAP	Yes	MGC, LF0	hn-uSFGAN
NNSVS-WORLD v4	MGC, LF0, VUV, BAP	Yes	MGC,LF0, BAP	hn-uSFGAN

Naturalness MOS test results for NNSVS systems (1/2)



- AR F0 > Non-AR F0
- AR model for mel-spectrogram didn't work well possible due to exposure bias issues

Naturalness MOS test results for NNSVS systems (2/2)



- AR > non-AR
- Modeling MGC, LF0, BAP with autoregressive models obtained the best score of 3.86.

→ Demonstrated the effectiveness of multi-stream AR models for WORLD features

Conclusions

NNSVS: neural network-based singing voice synthesis toolkit

New features are available at GitHub

- Diffusion-based acoustic model
- SiFi-GAN [Yoneyama+2023]
- Mandarin SVS recipes using Opencpop [Yu+2022]

https://github.com/nnsvs/nnsvs

Samples

[Yoneyama+2023] “Source-Filter HiFi-GAN: Fast and Pitch Controllable High-Fidelity Neural Vocoder”, in Proc. ICASSP 2023.
[Yu+2022] “Opencpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis”, in Proc. Interspeech, 2022.