DDSP-based Singing Vocoders: A New Subtractive-based Synthesizer and A Comprehensive Evaluation

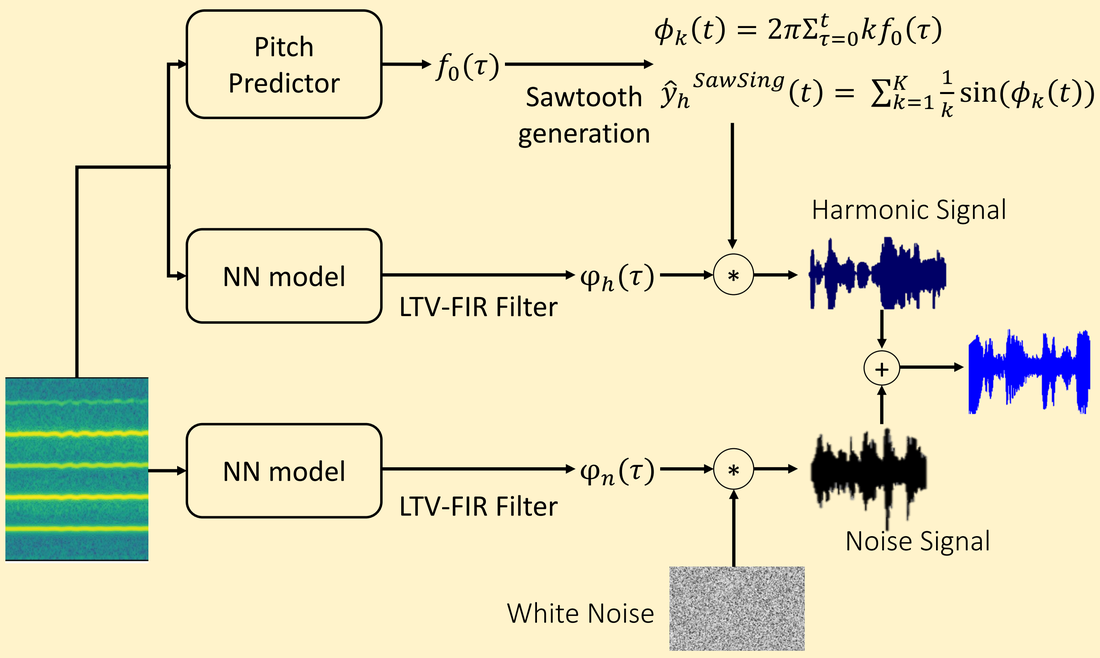
Da-Yi Wu, Wen-Yi Hsiao, Fu-Rong Yang, Oscr Friedman,
Warren Jackson, Scott Bruzenak, Yi-Wen Liu, Yi-Hsuan Yang
{ericwudayi2, s101062219, fjbcrs34}@gmail.com

INTRODUCTION

- SawSing synthesizes the harmonic part of singing voices by filtering a sawtooth source signal with a linear time-variant finite impulse response.
- Evaluation shows that SawSing converges faster and better than state-of-the-art vocoders in a resource-limited scenario with only 3 training recordings and a 3-hour training time.

INSIGHT & DIAGRAM

- Vocoder: Reconstruct the audio samples from acoustic feature like mel-spectrogram. In general, vocoder is trained with pairs of mel-spectrogram and waveforms. Some typical vocoders like : Parallel WaveGAN, Hifi-GAN, FastDiff, …
- DDSP: Build sine wave with pre-extracted f_0 as excitation signal and mix-up with filtered noise to re-construct audio from mel-spectrogram.
- SawSing:
  - Extract f_0 with pitch predictor, trained jointly with model.
  - Use *sawtooth signal* as our excitation signal.
  - Use subtractive-based method that convolving excitation signal with LTV-FIR Filter generated from NN model.

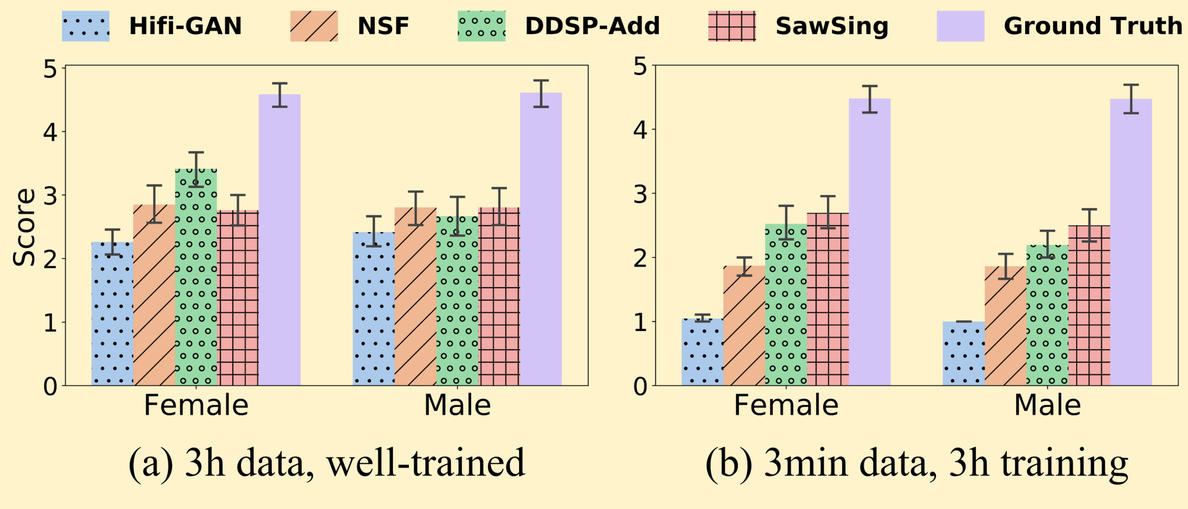


EVALUATION

- SawSing performs the best in MSSTFT and FAD across both scenarios and both
- For scenario (b), DDSP-based model perform much better than pure neural network like HiFi-GAN, PWG.
- Among the evaluated models, the performance gap between (a) and (b) is the smallest in the result of SawSing.

Model	Para-meters	RTF	MSSTFT↓ Female (a)	MSSTFT↓ Female (b)	MSSTFT↓ Male (a)	MSSTFT↓ Male (b)	FAD↓ Female (a)	FAD↓ Female (b)	FAD↓ Male (a)	FAD↓ Male (b)
FastDiff	15.3M	0.017	14.5	17.9	11.1	16.9	2.29	7.40	3.53	10.0
HiFi-GAN	13.9M	0.004	7.13	16.7	7.82	18.9	0.59	3.50	0.51	10.5
PWG	1.5M	0.007	7.39	13.0	7.83	14.8	0.36	6.15	2.56	6.29
NSF	1.2M	0.006	7.51	10.9	10.2	13.4	0.49	3.73	2.08	4.83
DDSP-Add	0.5M	0.003	7.61	9.29	8.37	12.1	0.56	0.92	1.06	2.09
DWTS	0.5M	0.019	7.72	9.75	8.83	13.0	0.60	2.98	0.36	8.58
SawSing	0.5M	0.003	**6.93**	**8.79**	**7.76**	**11.7**	**0.12**	**0.38**	**0.22**	**0.59**

- We had 2 sets of questionnaires, one for the female and the other for the male singer. For each singer, we prepared 8 clips from the 3 testing unseen, each clip corresponding to the singing of a full sentence.
- The result of SawSing reveals that its output contains an audible electronic noise, or “buzzing” artifact, notably when singers emphasize the airflow with breathy sounds and for unvoiced consonants such as /s/ and /t/. We address this issue by post-processing and trying to improve the harmonic filter.



(a) 3h data, well-trained (b) 3min data, 3h training

DATA & SCENARIO

Our data is from MPop600 [31], a set of free Mandarin singing recordings. We

(a)Regular [3h data, well-trained]: full training data, and train the vocoders for each singer for up to 2.5 days.

(b)Resource-limited [3min data, 3h training]: randomly picked 3 recordings from the training set per singer, using 3-hour training time.

CONCLUSION

- We presented SawSing, a new DSP-based vocoder.
- We conduct evaluations on several DSP-based model compare favorably with SOTA neural vocoders such as HiFi-GAN and FastDiff.
- Open source code and demo page: