Paper ID: 3410
ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

A Synthetic Corpus Generation Method for Neural Vocoder Training

Zilin Wang[1], Peng Liu[2], Jun Chen[1], Sipan Li[1,5], Jinfeng Bai[3], Gang He[3], Zhiyong Wu[1,4,6], Helen Meng[1,6]

[1] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China [2] Transsion, Shanghai, China [3] TAL Education Group, Beijing, China
[4] Peng Cheng Lab, Shenzhen, China [5] Tencent AI Lab, Shenzhen, China [6] The Chinese University of Hong Kong, Hong Kong SAR, China

1. Introduction

1.1 Background

- Vocoder
  - ✓ Synthesizes the final waveform from acoustic information
    - Input: acoustic feature, *e.g.,* MFCC, Mel-Spectrogram
    - Learning target: speech or music waveforms
  - ✓ Applications:
    - Text-To-Speech (TTS)
    - Voice Conversion (VC)
    - Singing Voice Synthesis
- Neural Vocoder
  - ✓ Deep Neural Network-based
  - ✓ Offer state-of-the-art speech synthesis quality

1.2 Motivation

- Data scarcity
  - ✓ The collection of training data is challenging
    - High-quality and noise-free audio is scarce
    - Require the speakers to vocalize for a long time in a professional recoding environment
  - ✓ Demand for generalization
    - An ideal vocoder should be able to work for various audios in various environments
    - Call for more audio data with more diversified forms
    - Further push up the demand for high-quality corpus

2. Methodology

We present a synthetic corpus generation method to tackle the real data scarcity problem during neural vocoder training

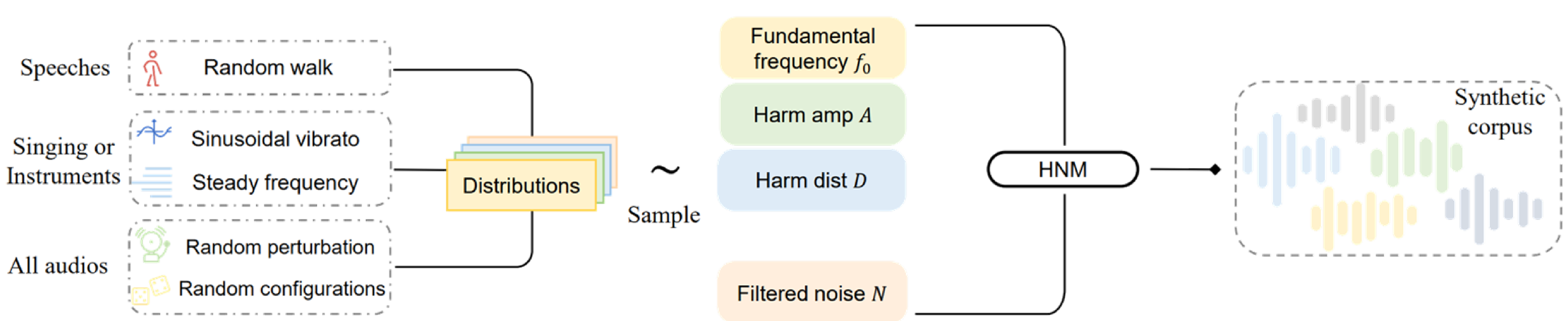
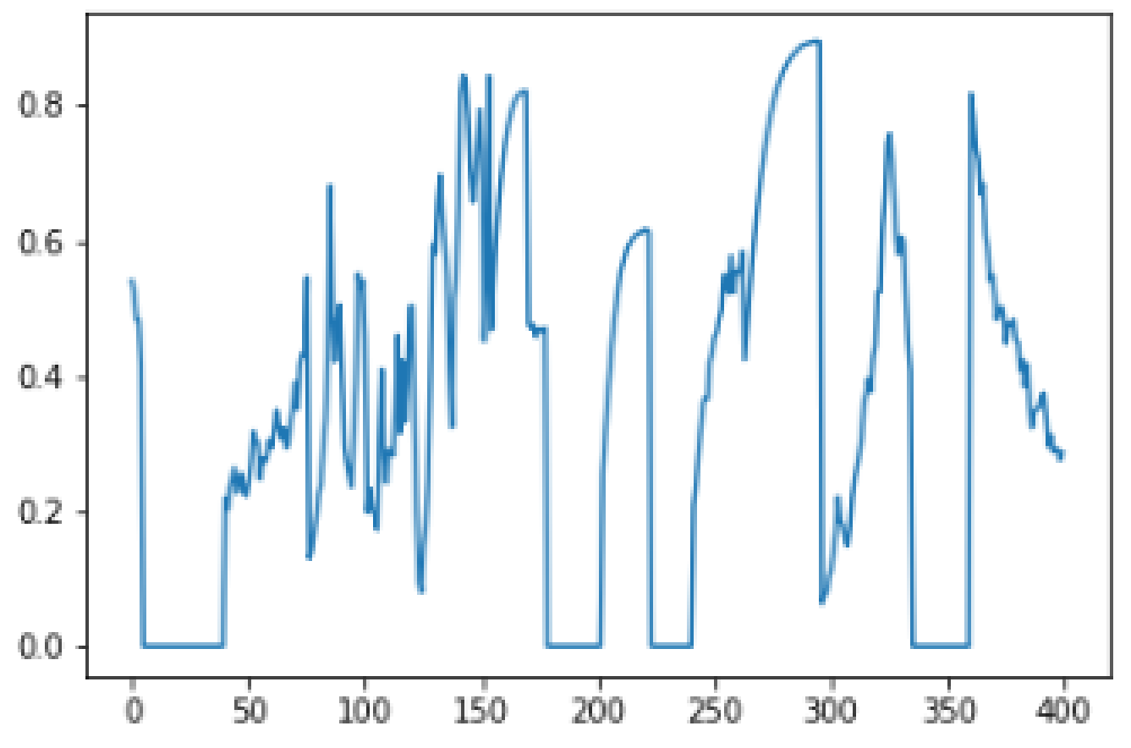


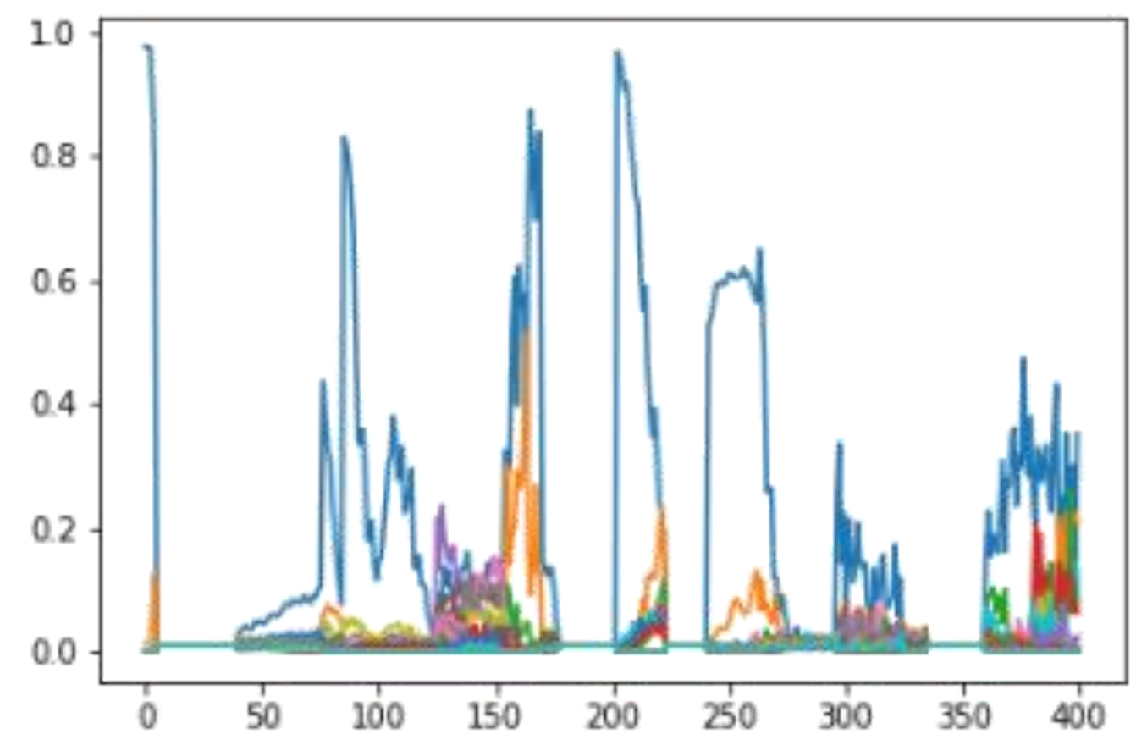
Figure 1. The pipeline of our proposed method. At first, based on the prior knowledge from different target audio domains, we model the distributions of acoustic features. Then we sample the acoustic features, including fundamental frequency f_0, harmonic amplitude A, harmonic distribution D, and time-varying filtered noise signal N from corresponding distributions, respectively. Lastly, audio is synthesized based on the sampled acoustic features by Harmonic-plus-Noise Model and forms the synthetic corpus.

3.1 Overall pipeline

- Model the distributions of acoustic features
- Sample a bulk of acoustic features from the modeled distributions
- Synthesize corpus according to the Harmonic-plus-Noise Model (HNM)



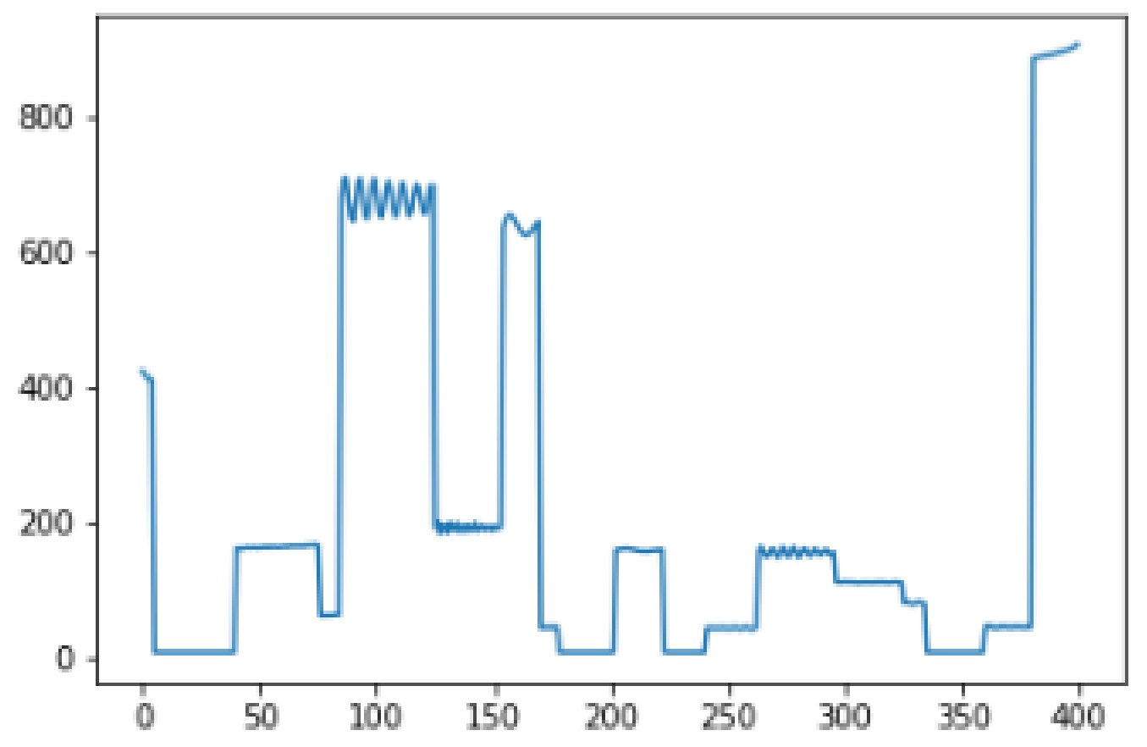
(a) Harmonic amplitude



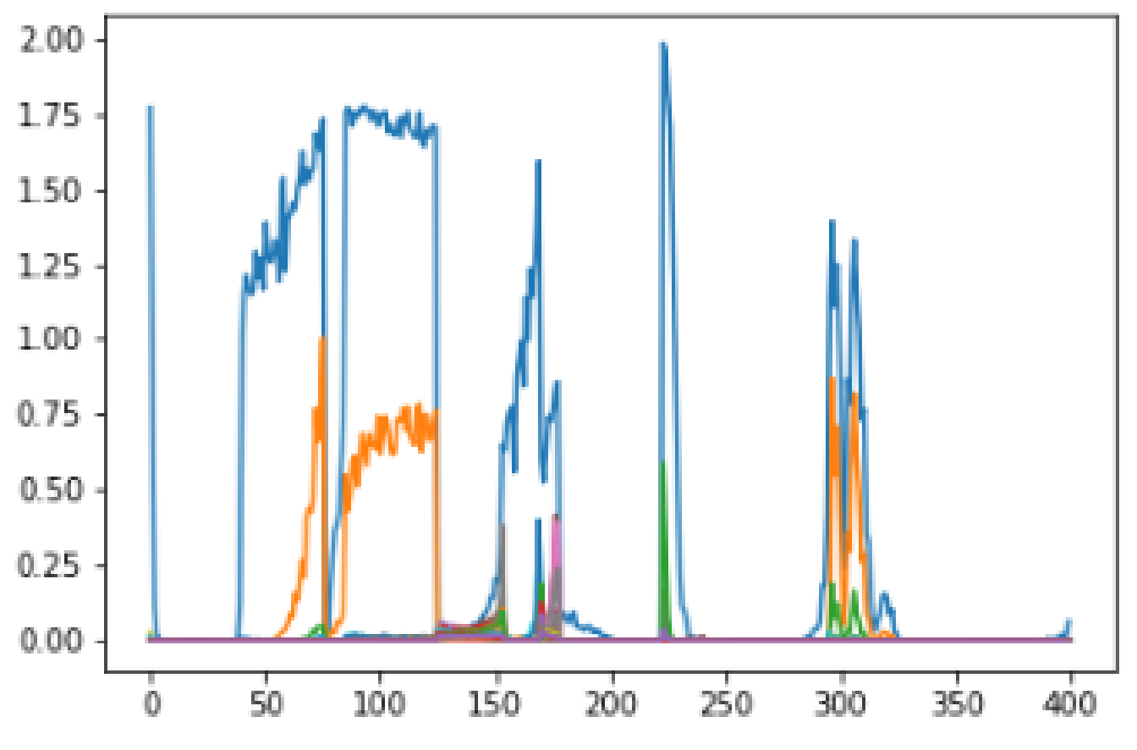
(b) Harmonic distribution

3.2 Acoustic features generation

- Target scenarios:
  - ✓ Speeches
  - ✓ Musical pieces (singing voices and instrument pieces)
- Random Walk (RW) model to fit with uncertainty in speeches
- Sine constituents in fundamental frequency to simulate vibrato in music pieces
- Uniform constituents in fundamental frequency to simulate the steady long tones in music pieces
- An example is visualized and shown in right figures



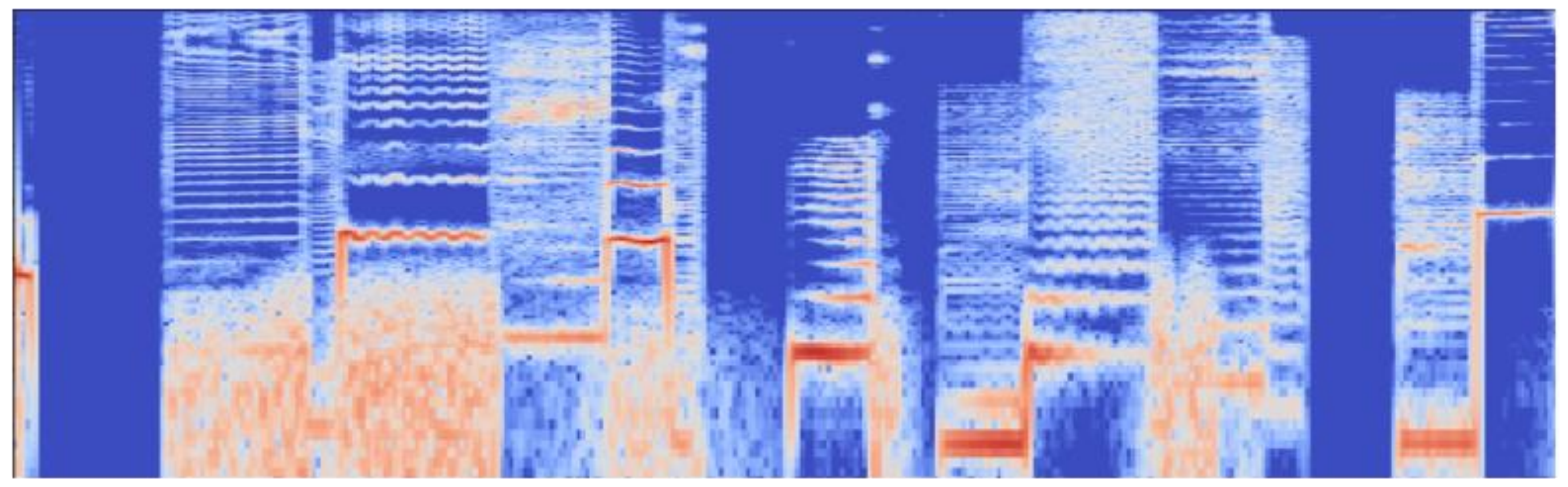
(c) Harmonic fundamental frequency



(d) Filtered noise

3.3 Harmonic-plus-Noise Model recap

- In HNM, audio signal can be represented as the sum of the harmonic and noise components.
- For the voiced part, the signal can be approximated by superimposing a series of harmonic components.
- For the unvoiced part, a FIR filter and a white noise signal are used for noise modulation.



(e) Spectrogram of a synthetic audio

3. Experiments

3.1 Experiment settings

- DDSP is employed as the implementation of HNM
- 1,000,000 pieces of synthetic audios are synthesized
- Sample rate is 24,000 Hz
- Vocoders for comparison:
  - ✓ Neural Vocoder (1M): The HiFi-GAN vocoder trained with all the 1,000,000 pieces of synthetic audio
  - ✓ Griffin-Lim: A traditional waveform synthesis method that also does not rely on real data
  - ✓ Neural Vocoder (Real): The HiFi-GAN vocoder trained with real corpus, the LJSpeech, a popular open-source dataset has a total length of approximately 24 hours
  - ✓ Neural Vocoder (10K): The HiFi-GAN vocoder trained with 10,000 pieces of synthetic audio, which has the nearly same total length as the LJSpeech
- Metrics
  - ✓ Mean Opinion Score (MOS) of audio quality as a subjective metric
  - ✓ Perceptual Evaluation of Speech Quality (PESQ) and Short-Term Objective Intelligibility (STOI) as two objective metrics
- A total of twenty-seven people participated in the MOS test

3.2 Experimental results of the full synthetic corpus (1M)

- Speeches
  - ✓ Audio clips from VCTK dataset
  - ✓ Average audio quality for male speeches
  - ✓ Not bad results on female speeches but defeated by real corpus (overfitting)
- Singing voices
  - ✓ Audio clips from Opencpop dataset
  - ✓ Achieves the highest scores in all subjective and objective metrics
  - ✓ Outstanding effectiveness for singing voices
- Instrumental pieces
  - ✓ Audio clips from URMP dataset
  - ✓ Significantly outperforms all other vocoders
  - ✓ Excels for instrumental pieces synthesis tasks

3.3 Investigation of the small synthetic corpus (10K)

- Beaten by the real corpus in speeches and singing voices for both objective and subjective metrics
- Usable in female speeches and musical pieces
- MOS score of 3.91 for instrumental pieces synthesis
- Hard to beat high-quality real data with the same total length
- Nearly no cost to synthesized

Table 1. Subjective evaluation results (MOS values). "Neural Vocoder (Real)" denotes the vocoder trained with real corpus LJSpeech. "Neural Vocoder (10K)" denotes the vocoder trained with 10,000 pieces of synthetic audio and "Neural Vocoder (1M)" denotes the vocoder trained with 1,000,000 pieces of synthetic audio. "CI" denotes the confidence interval. † denotes the neural vocoder trained with synthetic corpus generated by our proposed method.

Model	speeches (male)		speeches (female)		singing voices		instrumental pieces	
	MOS	95% CI	MOS	95% CI	MOS	95% CI	MOS	95% CI
Ground Truth	4.76	± 0.06	4.64	± 0.07	4.91	± 0.03	4.68	± 0.07
Griffin Lim	2.03	± 0.08	1.73	± 0.07	1.64	± 0.07	1.54	± 0.07
Neural Vocoder (Real)	3.05	± 0.10	**4.47**	± 0.07	3.28	± 0.09	2.77	± 0.10
Neural Vocoder† (10K)	2.66	± 0.10	3.37	± 0.10	3.04	± 0.09	3.91	± 0.09
Neural Vocoder† (1M)	**3.18**	± 0.09	3.81	± 0.08	**4.20**	± 0.08	**4.00**	± 0.08

Table 2. Objective evaluation results (PESQ and STOI values). "Neural Vocoder (Real)" denotes the vocoder trained with real corpus LJSpeech. "Neural Vocoder (10K)" denotes the vocoder trained with 10,000 pieces of synthetic audio and "Neural Vocoder (1M)" denotes the vocoder trained with 1,000,000 pieces of synthetic audio. † denotes the neural vocoder trained with synthetic corpus generated by our proposed method.

Metric	Model	speeches (male)	speeches (female)	singing voices
PESQ	Griffin-Lim	1.989	1.382	0.517
	Neural Vocoder (Real)	3.050	**3.577**	2.891
	Neural Vocoder† (10K)	2.909	3.074	3.013
	Neural Vocoder† (1M)	**3.121**	3.272	**3.226**
STOI	Griffin-Lim	79.29	77.11	56.96
	Neural Vocoder (Real)	83.77	87.43	79.35
	Neural Vocoder† (10K)	84.43	84.93	78.38
	Neural Vocoder† (1M)	**88.28**	**88.40**	**82.06**

Listen to Samples

Contact:
wangzl21@mails.tsinghua.edu.cn