INVERTIBLE VOICE CONVERSION WITH PARALLEL DATA

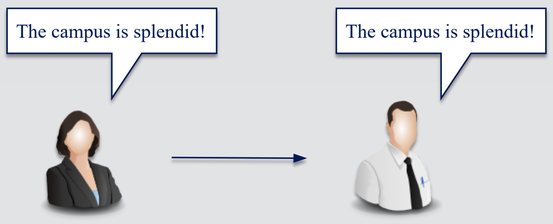
Zexin Cai[1], *Ming Li*[1,2]

[1]Department of Electrical and Computer Engineering, Duke University, Durham, NC, United States

[2]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Duke Kunshan University, Kunshan, China

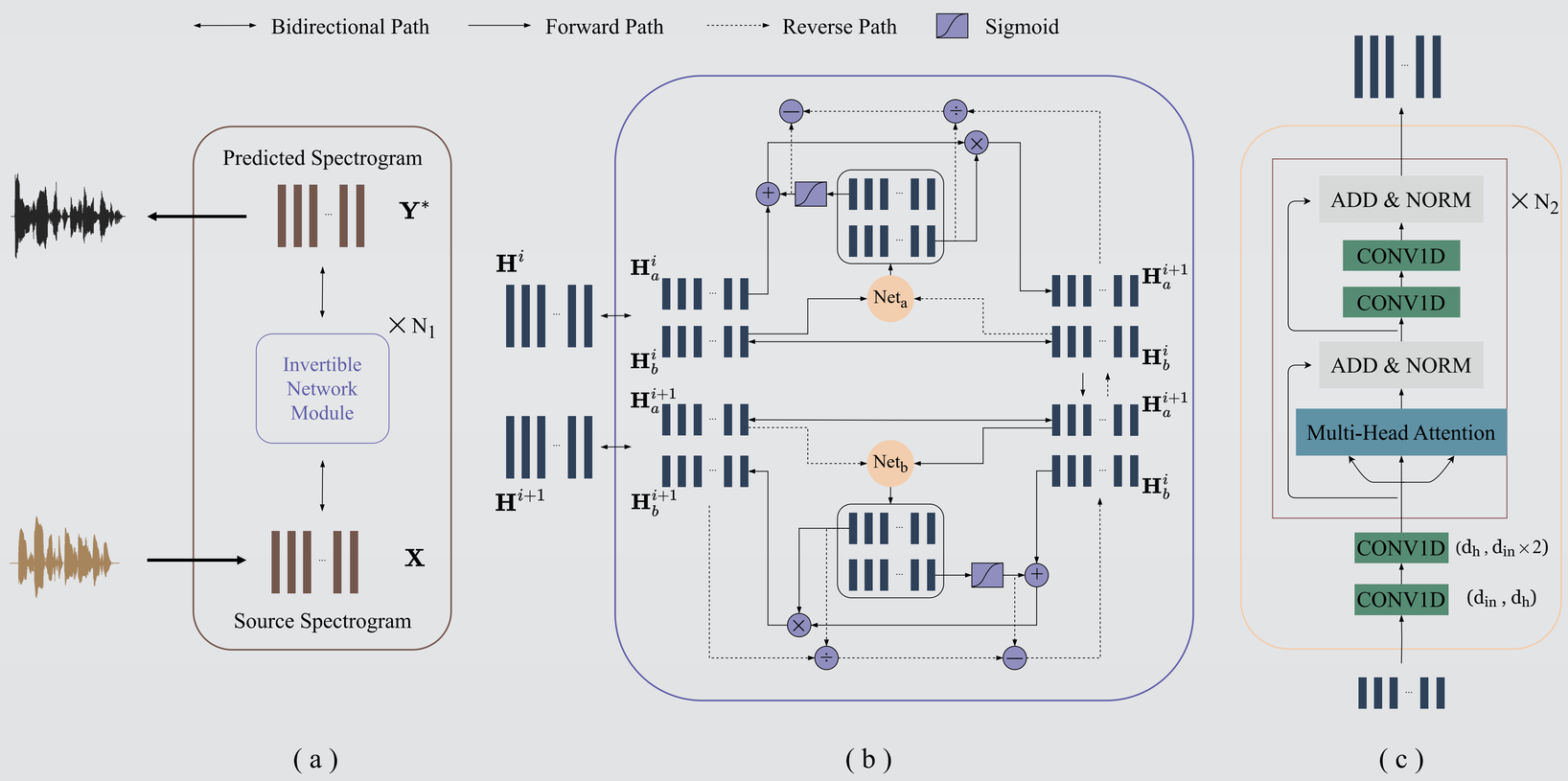
INTRODUCTION

- **We introduced an innovative deep voice conversion framework to elevate the security and reliability of voice conversion**
- **Specifically, we present a model that allows for the retrieval of source voice**
- Voice Conversion (VC)
  - aims to alter voice without changing the linguistic content
  - has been advanced by deep learning models and vocoders, enabling the generation of high-fidelity voices with impressive similarity [1]
- Nevertheless …
  - poses threat to societal security and voice biometric authentication [2,3]
  - could lead to breaches of privacy and misrepresentation



- Current countermeasures aimed at **discerning** whether an audio signal is synthetic, while they are **unable to trace** the origin of the fraudulent activity or **identify** the true speaker behind the converted audio
- Thus, we propose to design a **reliable** conversion system that possesses the ability to **reverse** the conversion process

METHODS



(a) (b) (c)

a. General overview of our proposed model

b. Structure and dataflow of the Invertible Network Module, consists of blocks of consecutive affine coupling layers [7]

c. Structure of the nonlinear network component 'Net', which could be any network structure while we adopt a transformer-based for conversion

FORWARD	REVERSE
$\mathbf{H}_a^i, \mathbf{H}_b^i = \text{SPLIT}(\mathbf{H}^i)$	$\mathbf{H}_a^{i+1}, \mathbf{H}_b^i = \text{SPLIT}(\mathbf{H}')$
$\mathbf{U}, \mathbf{B} = \text{SPLIT}(Net_a(\mathbf{H}_b^i))$	$\mathbf{U}, \mathbf{B} = \text{SPLIT}(Net_a(\mathbf{H}_b^i))$
$\boldsymbol{S} = \sigma(\mathbf{U} + \epsilon)$	$\boldsymbol{S} = \sigma(\mathbf{U} + \epsilon)$
$\mathbf{H}_a^{i+1} = \boldsymbol{S} \odot \mathbf{H}_a^i + \mathbf{B}$	$\mathbf{H}_a^i = (\mathbf{H}_a^{i+1} - \mathbf{B}) \oslash \boldsymbol{S}$
$\mathbf{H}' = \text{CONCAT}(\mathbf{H}_a^{i+1}, \mathbf{H}_b^i)$	$\mathbf{H}^i = \text{CONCAT}(\mathbf{H}_a^i, \mathbf{H}_b^i)$

EXPERIMENTS

- Dataset
  - CMU ARCTIC (English, parallel)
  - 4 speakers: '**bdl**' (male), '**rms**' (male), '**slt**' (female), '**clb**' (female)
  - 1000 utterances for training, 132 for evaluation
- Vocoder
  - HiFiGAN [4]
- Metrics
  - Mean Opinion Scores (MOS), scaled from 1 to 5, where 1 indicates poor performance and 5 signifies excellent performance
- **Systems Trained**
  - **Transformer-VC (Non-invertible)**
    transformer-based (Fastspeech-based [5]) voice conversion system
  - **CycleGAN-VC3 [6] (Non-invertible)**
    a generative modelling VC approach using GAN
  - **Invertible VC**
    our proposed invertible voice conversion model

RESULTS

Speakers		Naturalness ↑			Similarity ↑		
source	target	Invertible VC	Transformer-VC	CycleGAN-VC3	Invertible VC	Transformer-VC	CycleGAN-VC3
bdl	clb	3.84±0.23	**4.01±0.19**	3.71±0.23	**4.13±0.18**	4.1±0.18	3.38±0.22
	rms	**4.21±0.18**	4.17±0.17	3.98±0.2	**4.12±0.19**	4.1±0.18	3.47±0.21
	slt	3.75±0.19	**4.02±0.19**	3.77±0.2	4.22±0.17	**4.24±0.17**	3.85±0.21
clb	bdl	3.35±0.22	3.2±0.24	**3.53±0.23**	3.83±0.22	**4.12±0.18**	3.48±0.22
	rms	3.81±0.21	**3.98±0.23**	3.39±0.23	4.03±0.18	**4.18±0.18**	2.47±0.2
	slt	3.31±0.24	3.93±0.22	**4.1±0.2**	3.83±0.2	**4.23±0.19**	4.22±0.19
rms	bdl	3.01±0.23	**3.11±0.25**	2.69±0.21	3.76±0.22	**3.82±0.19**	3.17±0.21
	clb	3.44±0.23	**3.47±0.24**	2.82±0.26	3.93±0.21	**3.95±0.19**	1.91±0.2
	slt	3.24±0.22	**3.47±0.2**	3.21±0.22	3.91±0.18	**4.03±0.2**	3.0±0.22
slt	bdl	3.21±0.23	**3.39±0.23**	3.36±0.23	3.97±0.19	**4.02±0.2**	3.72±0.22
	clb	4.02±0.21	4.08±0.2	**4.27±0.17**	4.35±0.18	**4.48±0.16**	4.27±0.18
	rms	4.01±0.2	**4.17±0.18**	3.58±0.21	4.05±0.19	**4.15±0.17**	2.75±0.2
All		3.59±0.07	**3.78±0.06**	3.52±0.07	4.01±0.06	**4.12±0.05**	3.31±0.07
p-values		-	6.2×10^{-5}	0.154	-	6.43×10^{-3}	$< 10^{-5}$

Samples

DISCUSSIONS

Current Limitation

- **Restricted** to utterances synthesized by the invertible VC model
- Invertibility is only available at the **spectrogram level**
- Use **Parallel data**

Future

- **Non-parallel** Invertible VC
- Invertibility at the **Waveform level**

ICASSP 2024

References

[1] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li, "An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning," IEEE Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 132–157, 2020.

[2] Tomi Kinnunen, Zhi-Zheng Wu, Kong Aik Lee, Filip Sedlak, Eng Siong Chng, and Haizhou Li, "Vulnerability of Speaker Verification Systems against Voice Conversion Spoofing Attacks: The Case of Telephone Speech," in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4401–4404.

[3] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, "Spoofing and Countermeasures for Speaker Verification: A Survey," Speech Communication, vol. 66, pp.130–153, 2015.

[4] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," Proc. of NeurIPS 2020, vol. 33, pp. 17022–17033.

[5] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in International Conference on Learning Representations, 2021.

[6] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-Spectrogram Conversion," in Proc. Interspeech 2020, pp. 2017–2021.

[7] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio, "Density Estimation Using Real NVP," in 5th International Conference on Learning Representations, 2017.

Duke