DreamVoice: Text-Guided Voice Conversion

Author: Michał Krępa
ミハウ クレパ

Introduction

IE: Why is it interesting?

Development of Virtual/Augmented Reality and virtual spaces

- Enhanced Personalisation and Identity
- Intuitivity in creating one's own voice in virtual spaces

Social Aspects

- Gender Dysphoria
- Speech Impairments

Content Creation

- Amateur creators
- Tiktok, Youtube etc.

Keywords of the research: Audio Processing, Voice Generation, Voice Conversion, Prompt Based

Task Adressed

Current Challenges in Voice Conversion (VC):

- Traditional VC models, also known as **one-shot VC**, require audio samples from the target voice to work.

There are available models that allow for text-to-speech conversion, however:

- Datasets are either small or
- Provide low quality information about the voices
- Mainly restricted access

Solution and Contributions

To approach and overcome those challenges the authors proposed 2 solutions:

- DreamVoiceDB: Dataset with voice timbre data from 900 speakers, created with voice-acting experts , to support detailed

- DreamVC: An end-to-end model that uses diffusion probabilistic models (DPM) and Classifier-Free Guidance (CFG) to synthesize voices based on text prompts, creating high-quality, text-aligned voices.

- DreamVG: A plugin that employs that into a working ui, allowing also already existing models.

How does it work?

1

Input source file and Text prompt

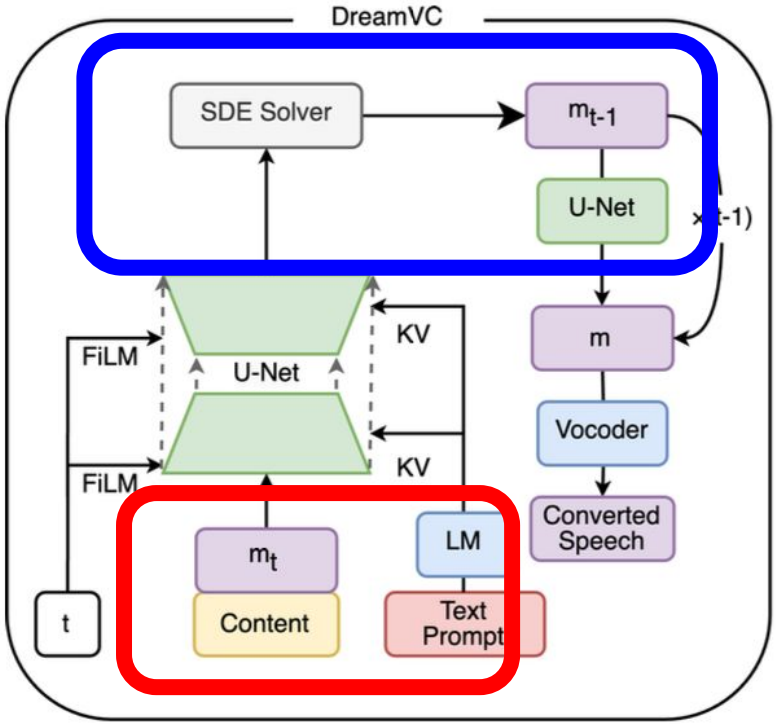
Generate noise and apply that to the source, we end up with -> 2

2

Two process pipeline

Front: Sample the data

Backward: recover the data from the samples, apply it to the target



Demonstration

Sample 1: Feminine voice on a podcast

Prompt used: Authoritative sounding person, who is gender-ambiguous and adult.

Sample 2: Old blues song (a cappella)

Prompt used: A teenage girl's voice that is smooth, warm, and attractive, perfect for captivating storytelling.

Sample 3: Polish audiobook

Prompt used: A mature male voice, bright and engaging, good for client and public interaction.

Results

DreamVC outperformed other VC models in:

- Consistency with text prompts
- Quality and Naturalness

Table 1: *Comparison of Objective scores: Word Error Rate (WER), Phoneme Error Rate (PER), Relative Inference Speed (RIS), and Mean Opinion Scores (MOS) with their 95% confidence intervals (CI): Q-Quality, N-Naturalness, C-Prompt-Voice-Consistency.*

Method	Text-Guided VC	WER ↓	PER ↓	RIS↑	MOS-Q ↑	MOS-N ↑	MOS-C ↑
Grount-Truth	/	/	/	/	4.42 ± 0.11	4.26 ± 0.11	4.12 ± 0.13
FreeVC	×	6.37	9.79	/	4.09 ± 0.12	3.98 ± 0.13	/
ReDiffVC	×	3.45	8.26	/	3.67 ± 0.14	3.76 ± 0.13	/
DreamVC	✓	**4.10**	**8.08**	1.00x	3.62 ± 0.14	3.61 ± 0.14	$\mathbf{3.72 \pm 0.15}$
DreamVG+FreeVC	✓	7.58	10.05	2.71x	$\mathbf{3.90 \pm 0.13}$	$\mathbf{3.85 \pm 0.14}$	3.43 ± 0.16
DreamVG+ReDiffVC	✓	5.11	8.65	1.08x	3.80 ± 0.14	3.70 ± 0.13	3.66 ± 0.15

Future Work

Results are promising but:

- **Inference speed**
- **Voice Quality Issues**:

- **My own observation:** seems like the conversion works fine only for slow paced, calm monologue, anything else than that creates distortion, or artifacts. Not to mention different accents.

Conclusion

Thanks!

Sources:

Paper:

https://arxiv.org/pdf/2406.16314v1

Code and examples:

https://github.com/myshell-ai/DreamVoice

Project Website:

https://haidog-yaqub.github.io/dreamvoice_demo/

Forked repo with the samples used here:

https://github.com/Darkmik70/DreamVC/tree/master

Citations:

[1] - Hai, J., Thakkar, K., Wang, H., Qin, Z., & Elhilali, M. (2024). DreamVoice: Text-Guided Voice Conversion. *arXiv preprint arXiv:2406.16314*.