Towards human like voice interaction

Zhuo Chen

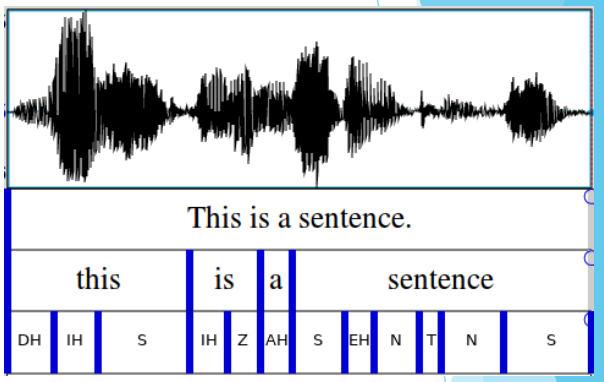
Bytedance

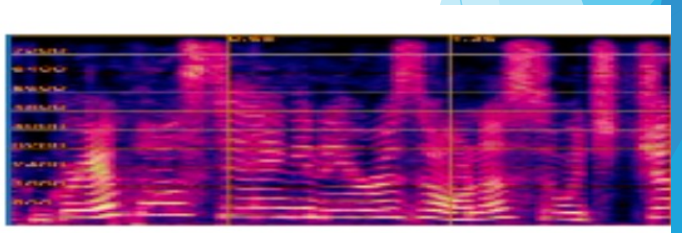
Content

- Two general architectures in recent audio generation and their limitations
- Representation and alignment improvement
- An end to end perspective

Two perspectives for speech generation

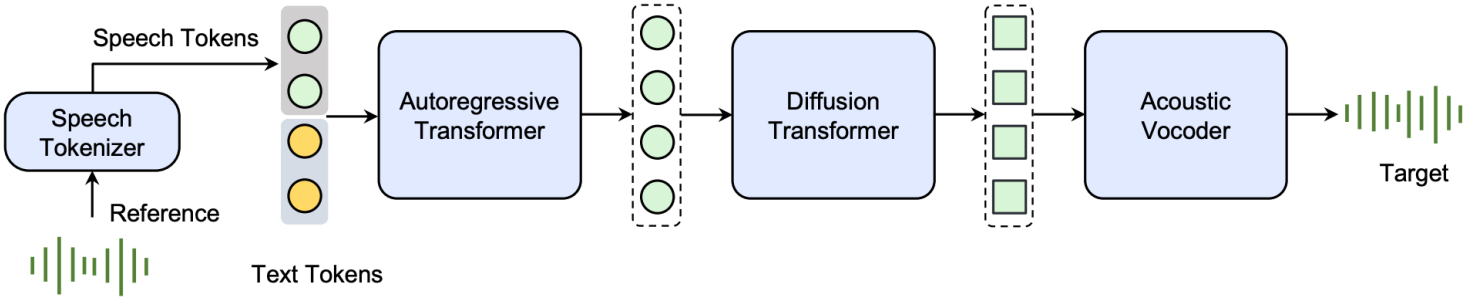
- Text sequence
  - Monotonic mapping
  - Semantic information
  - Friendly for auto-regressive generation
  - AudioLM, Valle, Cosyvoice1,2 etc.
- Image patch
  - One to many mapping
  - “Low rank” structure
  - Naturally benefit from image generation
  - Voice box, F5-TTS etc.





[SIL HH HH AW AW AW

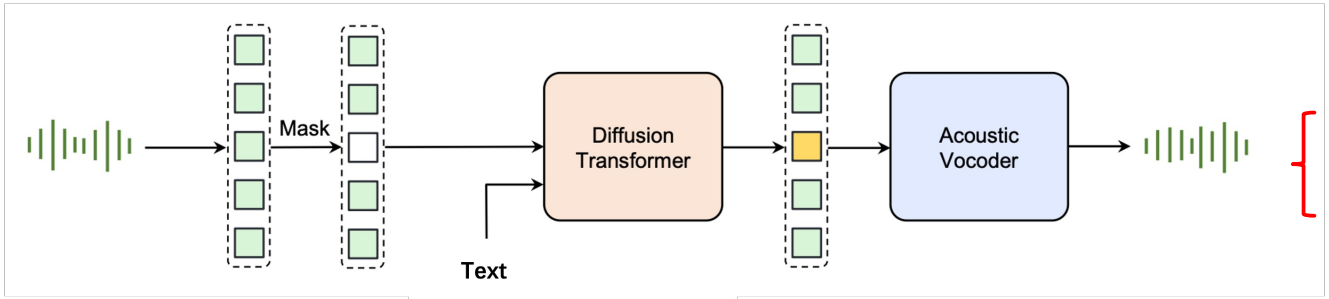
Seed-TTS



- Audio Tokenizer
- Language model
- Diffusion
- Vocoder

- On language modeling side:
  - Discrete sematic token ensures the general architecture of the speech
  - Causal diffusion decoder further polish details
  - Friendly to streaming processing

Seed-TTS-DIT



- Diffusion
- Vocoder

- From image generation side:
  - One latent diffusion transformer, much simpler pipeline
  - No pre-estimated alignment
  - Excels in editing tasks
  - Block online or offline processing

Evaluation

- Both systems can achieve high quality speech generation
  - Surpass the recording and vocoder reconstruction in ASR/ASV metrics
  - Large diversity and good scaling property
  - Interestingly, the two architecture have very similar performance in all subjective and objective tests

System	Lang.	WER (↓)	SIM (↑)
Human	EN	2.143	0.730
Vocoder resynthesized	EN	2.165	0.702
Seed-TTS$_{\text{ICL}}$	EN	2.249	0.762
Seed-TTS$_{\text{DiT}}$	EN	**1.733**	**0.790**
Human	ZH	1.254	0.750
Vocoder resynthesized	ZH	1.342	0.733
Seed-TTS$_{\text{ICL}}$	ZH	**1.115**	0.796
Seed-TTS$_{\text{DiT}}$	ZH	1.178	**0.809**


More audio demo can be found in https://bytedancespeech.github.io/seedtts_tech_report/

Applications in Audio and Music generation

- Seed Music
  - Condition on text description, singing prompt, or midi
  - Full song generation and editing
  - Reinforcement learning to improve the generation quality
  - Demo can be found in https://team.doubao.com/en/special/seed-music

- Seed Foley
  - Audio effect generation conditioned on text or video input
  - Multi-source generation and event synchronization

Demo for Seed Foley

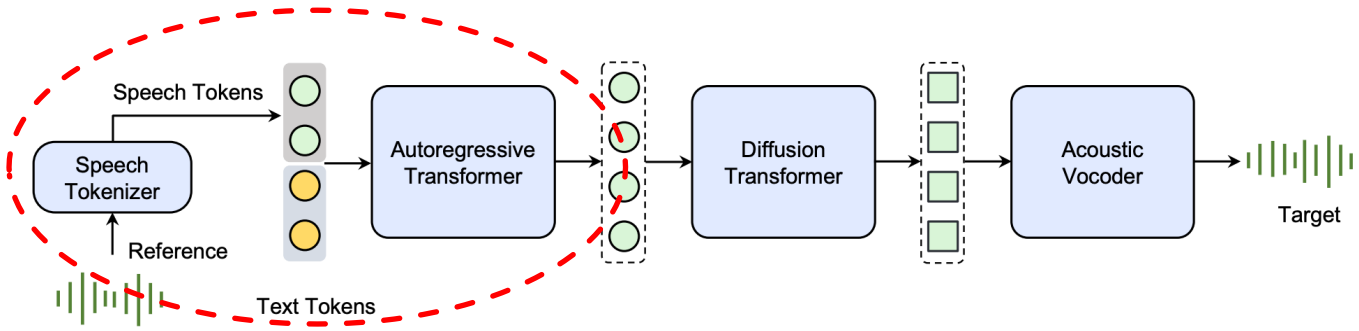
A bit recap

- Speech and Audio signal shares properties from both text and image. From either perspective, we can build high quality generation models with different advantages, and apply them to different applications.

- It seems that these systems can compose a rather complete story, what's next?
  - Do we just need to work on data optimization?
  - Are there systematic limitations in these systems

- Unfortunately, there are always more problems than solutions
  - Two obvious limitations
    - Representation
    - Alignment
  - And more…

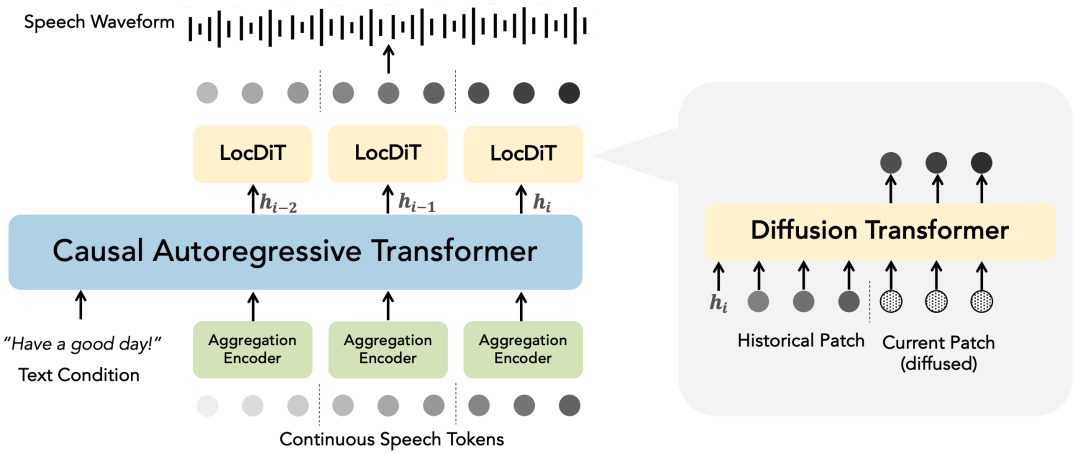
Limitation in representation



- Representation limitation
  - Discrete speech tokens suffer from the information loss
    - High frequency rate results in long sequence
    - Difficulties in extend to video/image signal
    - Hallucination from the diffusion decoder
  - Meanwhile, NAR architecture is hard to joint model with LLM and streaming processing
    - Major LLMs are auto-regressive based
- Can we bring the benefit from both sides?

DiTAR: Diffusion Transformer Autoregressive Modeling for Speech Generation

- A patch based auto-repressive model
  - Directly build on the continuous VAE latent
  - Auto-regressive backbone for streaming language modeling
  - Local diffusion head for diverse generation
  - Continuous representation ensures high quality, low frame rate wave generation



DiTAR results

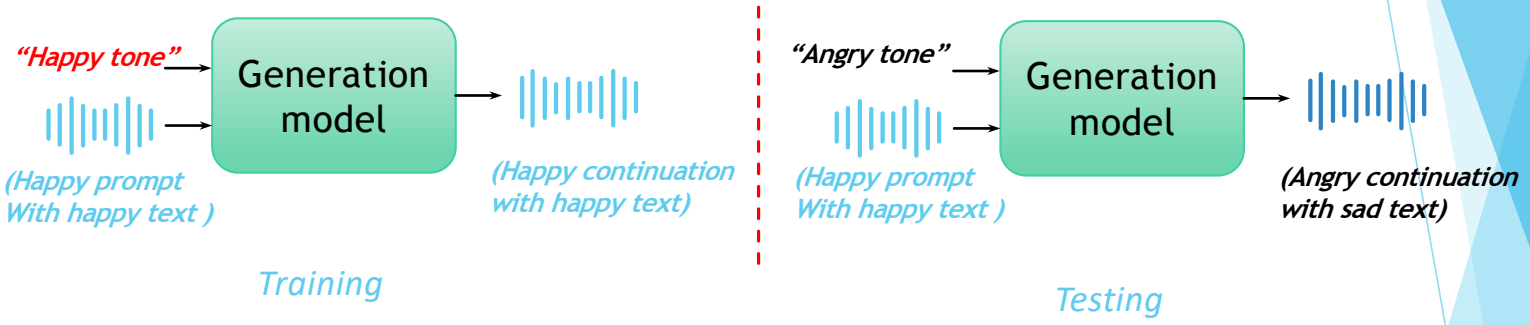
System	Seed-EN		Seed-ZH	
	WER(%)↓	SIM↑	WER(%)↓	SIM↑
Human	2.06	0.73	1.254	0.750
Seed-TTS$_{\text{DiT}}$	<u>1.733</u>	**0.790**	<u>1.178</u>	**0.809**
CosyVoice	4.29	0.609	3.63	0.723
CosyVoice 2	2.57	0.652	1.45	0.748
CosyVoice 2-S	2.38	0.654	1.45	0.753
FireRedTTS	3.82	0.46	1.51	0.63
MaskGCT	2.623	0.717	2.273	<u>0.774</u>
E2TTS	2.19	0.71	1.97	0.73
F5TTS	1.83	0.67	1.56	0.76
DiTAR	**1.685**	<u>0.735</u>	**1.023**	0.753

Limitation in alignment

- Conflicts between multi-source condition generation
  - In zero shot setting, can we generate an angry voice from a happy prompt?
  - Mismatch in prompt continuation.

- Difficulty in generalizable sematic alignment
  - Emergence in voice command control
  - Context aware tone switching

Improving condition alignment through RL



- Training/Testing mismatch results in unstable control
- With simple RL algorithm, we can largely boost the control stability
  - Various RL algorithms works, e.g. reinforce, DPO, PPO etc.

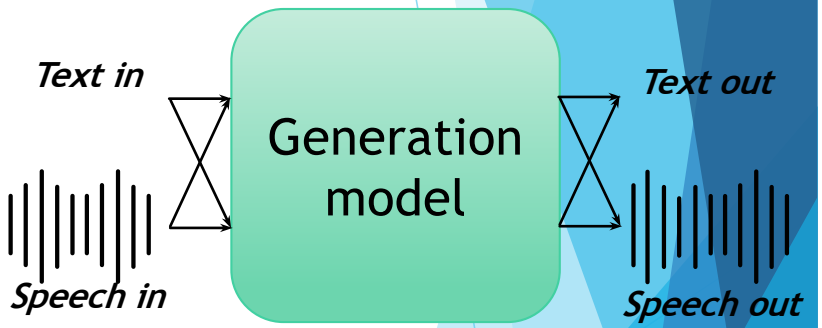
System	Angry	Happy	Sad	Surprise
Seed-TTS$_{\text{ICL}}$	0.46	0.44	0.53	0.13
Seed-TTS$_{\text{RL-SER}}$	**0.91**	**0.8**	**0.78**	**0.82**

Table 9. Comparison of the emotion control accuracy (↑) between Seed-TTS$_{\text{RL-SER}}$ and Seed-TTS$_{\text{ICL}}$ in the zero-shot scenario using the emotion set from subsection 3.2.

A more end to end perspective

- Even with an improved controllability, the model still struggles in
  - Generalization to unseen commands
  - Unseen natural sound behavior, e.g. cough, yawn
  - Context aware tone switching

- GPT-4o first demonstrate the potential of joint model
  - We explore this direction and push the performance on E2E audio-LLM



Better alignment through joint optimization

- Generalization to unseen command
  - An emerging funny drunken voice
  - Multi-turn adjustment
  - Character playing
- Context aware tone switching
  - Storytelling
  - Singing



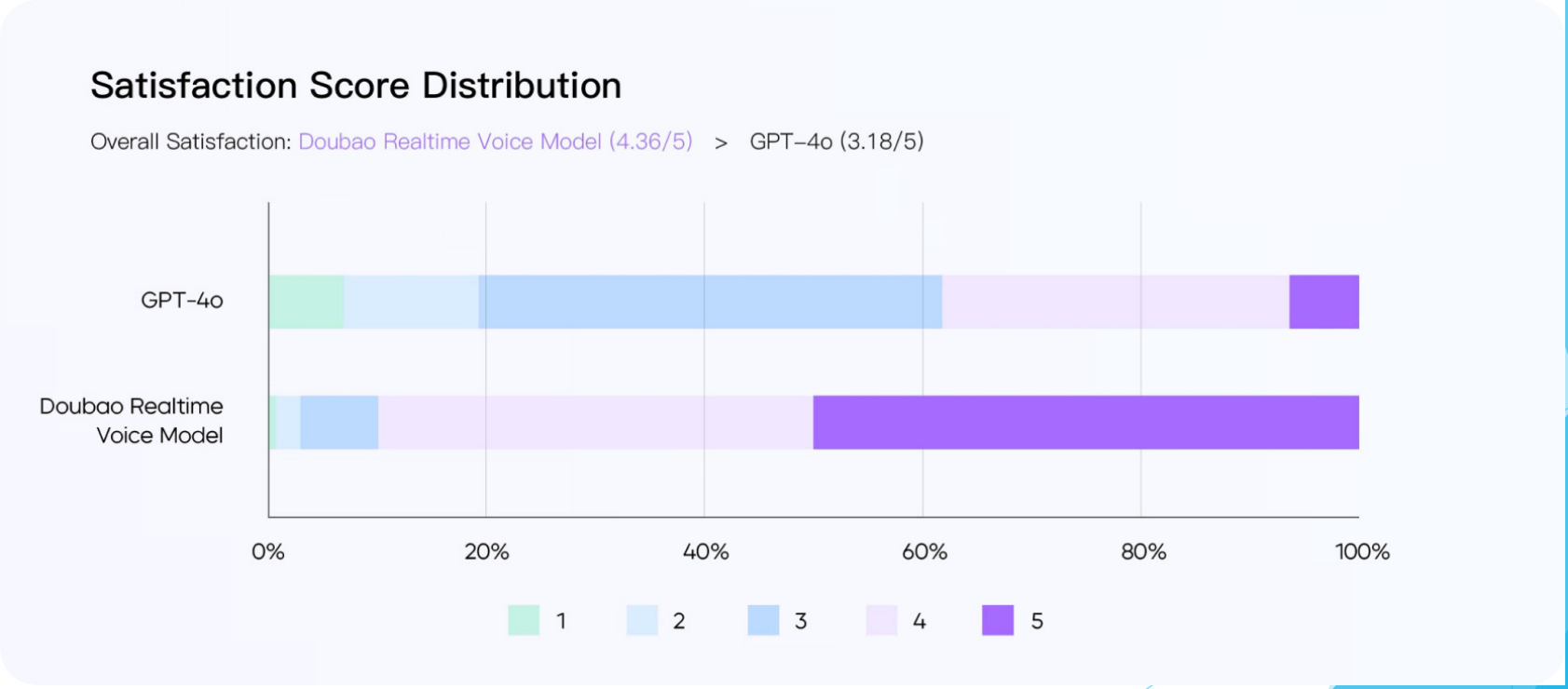

Demo for emotion control

More demo can be found in https://team.doubao.com/en/special/realtime_voice

Subjective evaluation

- We recruited 27 external participants, providing 810 conversations across 270 topics
- Participants were from 10 cities in China, including 9 males and 18 females, all aged between 21 and 33
- Among the participants, 11.11% had never used Doubao, 70.37% were light users (1-2 days per week), 14.81% were more frequent users (3-5 days per week), and 3.7% used Doubao every day.

Subjective evaluation result



A bit recap

- The LM or DIT model both produce realistic audio generation, but suffers from limitations in representation and alignment
- By using the DiTAR architecture, we benefit from both ends
- Reinforcement learning helps the condition controllability. End to end model further improve the model alignment
- What’s next?

Towards human like voice agent

- Converse like real human in everyway in all major languages
- Usefulness, responsiveness and empathy
- Long term memory, reasoning and personalization
- Further multi-agent, multi-modality fusion
- And more…