Lessons From the Autoregressive/Non-autoregressive Battle in Speech Synthesis

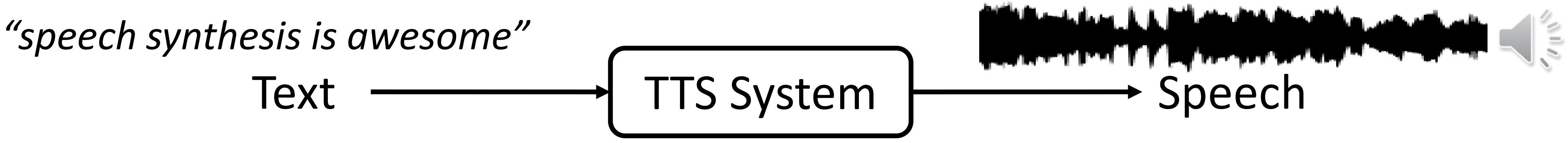
Xu Tan
Microsoft Research Asia
xuta@microsoft.com

About Me

- Xu Tan (谭旭)
  - Principal Researcher and Research Manager @ Microsoft Research Asia
- Research interests
  - Speech: FastSpeech 1/2, NaturalSpeech 1/2, UniAudio (https://speechresearch.github.io/)
  - Music: Muzic project (https://github.com/microsoft/muzic)
  - Avatar: GAIA project (https://microsoft.github.io/GAIA/)
  - Large language models
- Homepage
  - https://www.microsoft.com/en-us/research/people/xuta/
  - https://scholar.google.com/citations?user=tob-U1oAAAAJ

Text-to-Speech Synthesis

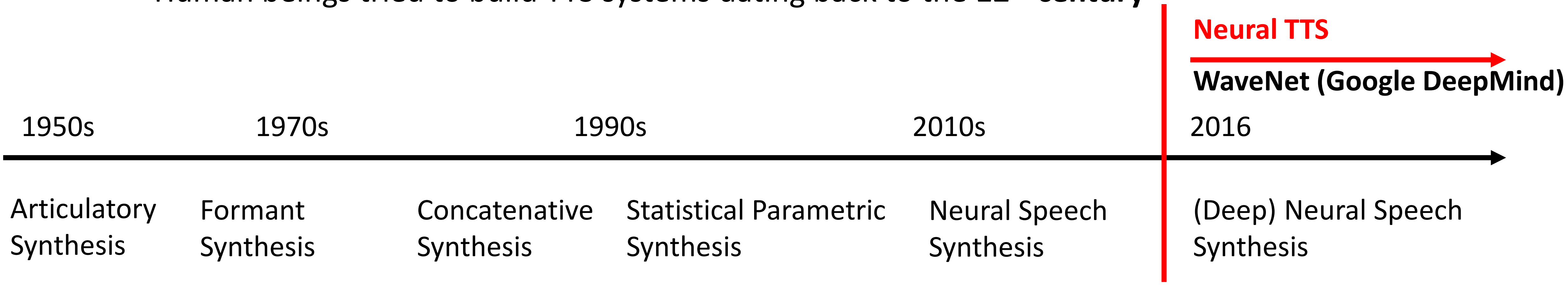
- Text-to-speech (TTS): generate intelligible and natural speech from text



“speech synthesis is awesome”

Text → TTS System → Speech

- Enabling machine to speak is an important part of AI
  - **TTS (speaking)** is as important as **ASR (listening), NLU (reading), NLG (writing)**
  - Human beings tried to build TTS systems dating back to the **12th century**



Neural TTS

WaveNet (Google DeepMind)

1950s — Articulatory Synthesis

1970s — Formant Synthesis

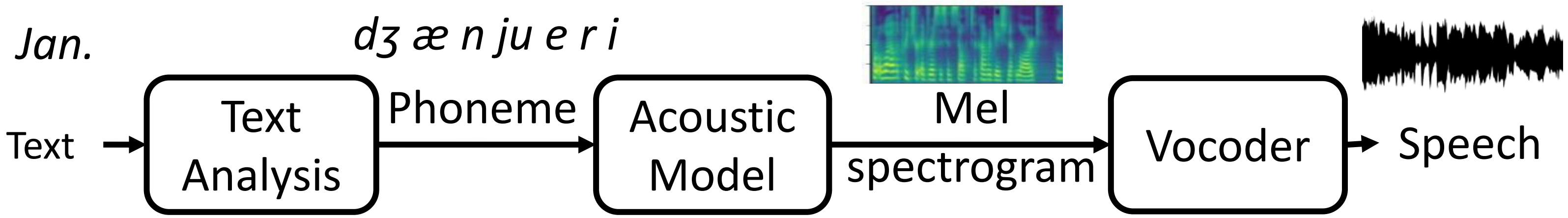
1990s — Concatenative Synthesis; Statistical Parametric Synthesis

2010s — Neural Speech Synthesis

2016 — (Deep) Neural Speech Synthesis

Typical Neural TTS Pipeline

- Text analysis, acoustic model, and vocoder



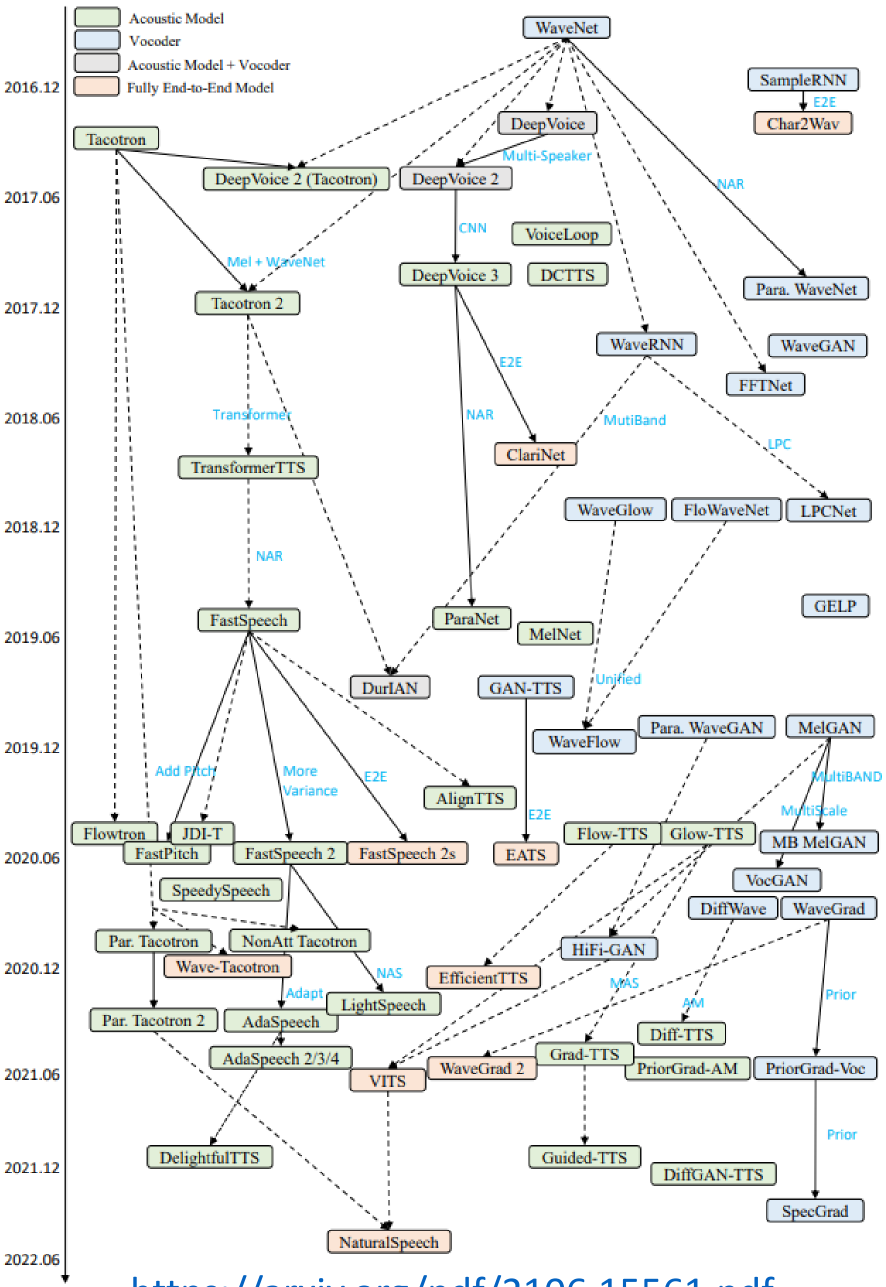
Jan.

dʒ æ n ju e r i

Text → Text Analysis → Phoneme → Acoustic Model → Mel spectrogram → Vocoder → Speech

- Text analysis: text → linguistic features
- Acoustic model: linguistic features → acoustic features
- Vocoder: acoustic features → speech

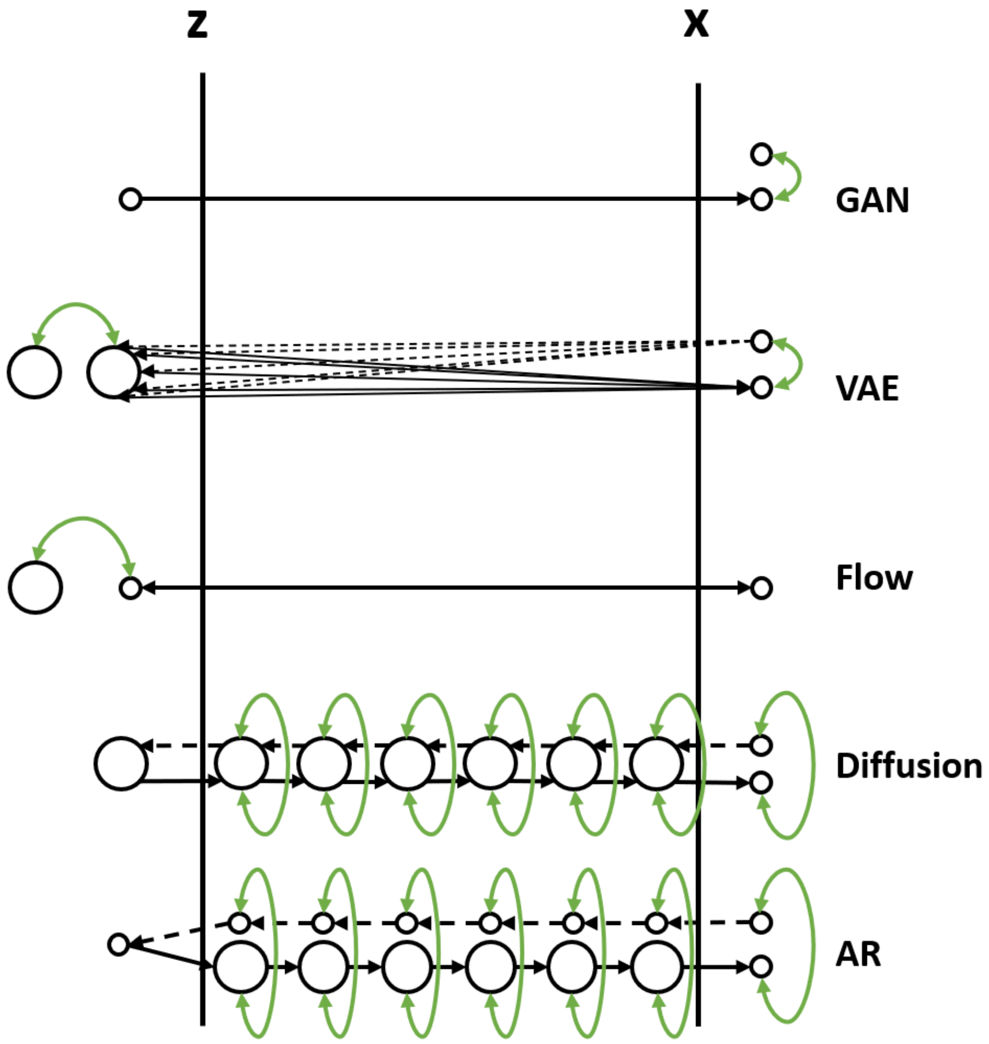
- Linguistic features: phoneme, prosody features
- Acoustic features: **mel-spectrogram, discrete token, latent vector**



https://arxiv.org/pdf/2106.15561.pdf

AR vs NAR in Neural TTS

- Generative models can be classified in AR/NAR
  - AR: Autoregressive
  - NAR: GAN, VAE, Flow, Diffusion (Flow Matching)
- Difference between AR and NAR (Diffusion)
  - How to factorize data?
    - AR: along time axis
    - NAR: along noise level
  - How to determine alignment/duration?
    - AR: implicitly
    - NAR: explicitly
  - Iteration steps
    - AR: sequence length
    - NAR: flexible



https://zhuanlan.zhihu.com/p/591881660

The Battle Between AR and NAR

	Acoustic Model		Vocoder	
Timeline	AR	NAR	AR	NAR
2017.06	Tacotron		WaveNet	
2017.12	Tacotron 2, DeepVoice 3			Par. WaveNet
2018.06			WaveRNN	
2018.12	Transformer TTS		LPCNet	WaveGlow
2019.06		FastSpeech		
2019.12				MelGAN, Par. WaveGAN
2020.06		FastSpeech 2, Glow-TTS		
2020.12				DiffWave, WaveGrad, HiFiGAN
2021.06		GradTTS		
2021.12		VITS		**SoundStream**
2022.06		NaturalSpeech		BigVGAN
2022.12	**AudioLM**			**EnCodec**
2023.06	**VALL-E, SPEAR-TTS**	**NaturalSpeech 2**		
2023.12	**UniAudio**	**SoundStorm, VoiceBox**		

The Battle Between AR and NAR

	Acoustic Model		Vocoder	
Timeline	AR	NAR	AR	NAR
2017.06	Tacotron		WaveNet	
2017.12	Tacotron 2, DeepVoice 3			Par. WaveNet
2018.06			WaveRNN	
2018.12	Transformer TTS		LPCNet	WaveGlow
2019.06		FastSpeech		
2019.12				MelGAN, Par. WaveGAN
2020.06		FastSpeech 2, Glow-TTS		
2020.12				DiffWave, WaveGrad, HiFiGAN
2021.06		GradTTS		
2021.12		VITS		**SoundStream**
2022.06		NaturalSpeech		BigVGAN
2022.12	**AudioLM**			**EnCodec**
2023.06	**VALL-E, SPEAR-TTS**	**NaturalSpeech 2**		
2023.12	**UniAudio**	**SoundStorm, VoiceBox**		

The Battle Between AR and NAR

	Acoustic Model		Vocoder	
Timeline	AR	NAR	AR	NAR
2017.06	Tacotron		WaveNet	
2017.12	Tacotron 2, DeepVoice 3			Par. WaveNet
2018.06			WaveRNN	
2018.12	Transformer TTS		LPCNet	WaveGlow
2019.06		FastSpeech		
2019.12				MelGAN, Par. WaveGAN
2020.06		FastSpeech 2, Glow-TTS		
2020.12				DiffWave, WaveGrad, HiFiGAN
2021.06		GradTTS		
2021.12		VITS		**SoundStream**
2022.06		NaturalSpeech		BigVGAN
2022.12	**AudioLM**			**EnCodec**
2023.06	**VALL-E, SPEAR-TTS**	**NaturalSpeech 2**		
2023.12	**UniAudio**	**SoundStorm, VoiceBox**		

Trends of the AR/NAR Battle

- Trend 1: NAR dominates Vocoder (Codec)

- Trend 2: NAR shows advantage in acoustic model before the LLM era

- Trend 3: LLMs revive the AR/NAR battle

Explanation of Trend 1&2

- Target-Target (T-T) vs Target-Source (T-S) dependency
  - T-T: dependency among target tokens
  - T-S: dependency on source tokens
- Difficulty of AR/NAR
  - If T-T > T-S → more information is needed from target tokens → NAR is more difficult
  - Vice versa
- Connection to multi-modality
  - Multi-modality: P(x|y) is not single-modal, not one-one mapping
    - e.g., “Thank You” → “Vielen Dank” or “Danke”
  - If T-S dominates, P(x|y) is more single-modal, a source token will have one definite mapping
  - If T-T dominates, P(x|y) is multi-modal, a source token will have multiple mappings

T-S Dependency

Type of T-S Dependency	Task	Alignment
Target has correspondence with source	Speech Enhancement	**Inherent** alignment
	Voice Conversion	
	Text to Speech	**Duration/Attention** alignment
	Singing Voice Synthesis	**MusicScore** alignment
	Speech Recognition	**CTC/Transducer/Attention** alignment
Target is a minor change of source	Text Error Correction	Locate the minor changes
	Text Style Transfer	Content not changes but style changes
Target is a translation of source	Machine Translation	**Attention** alignment
Target is implicitly correlated to source	Dialogue Generation	**Semantic** alignment
	Image Generation	**Semantic** alignment

T-T Dependency

Type of T-T Dependency	Task	Description
Text	Machine Translation	Discrete tokens in languages are **contextualized**, explained mutually. **Strong mutual dependency**
	Text Summarization	
	Text Error Correction	
	Text Style Transfer	
	Dialogue Generation	
	Speech Recognition	
Speech and Image	Text to Speech	For continuous signal like speech/sound/image, they depends on the concept, like speech frames depend on a word, image pixel depend on a class. **Weaker mutual dependency**
	Singing Voice Synthesis	
	Image Generation	

T-T/T-S Dependency and NAR Difficulty

Modality	Task	Source	Target	T-T vs T-S	Difficulty of NAR
Text Generation	Machine Translation	Source language	Target language	≈	*****
	Text Summarization	Long text	Short Summarization	≈	*****
	Dialogue Generation	Dialogue	Response	>	******
	Text Error Correction	Error Text	Correct Text	<	***
	Text Style Transfer	Source Text	Target text	<	***
	Speech Recognition	Speech	Text	≤	****
Speech Generation	**Text to Speech**	**Text**	**Speech**	<	*******
	Singing Voice Synthesis	Score	Singing Voice	<	**
	Voice Conversion	Source Voice	Target Voice	≪	*
	Speech Enhancement	Noisy Speech	Clean Speech	≪	*
Image Generation	Pixel Generation	Class ID	Image Pixel	-	*
	Discrete Token Generation		Image Token	-	**

Explanation of Trend 1&2

- Trend 1: NAR dominates Vocoder (Codec)
- Trend 2: NAR shows advantage in acoustic model before the LLM era

	TTS (Overall)	Vocoder	Acoustic Model
Target	Signal Not Symbol	Continuous Signal (Perceptual)	Content/Prosody/Timbre/Acoustic (Semantic)
T-T Dependency	Weaker Than Text	Short-term, Low-level, Local	Long-term, High-level, Global
T-S Dependency	1-1 Correspondence	Frame-level alignment	Duration/Attention Alignment
NAR Difficulty	**Easier than ASR/NMT**	**Very Easy, NAR Dominates**	**Easy, NAR Shows Advantage Before LLM Era**

Lessons Learned From Trend 1&2

- **Lesson 1**: To generate low-level perceptual details, **NAR is preferred**. If T-S has strong dependency, NAR is the best choice.
  - Audio (speech/music/sound): vocoder, codec
  - Image: VAE/VQ-VAE/VQ-GAN
  - Image/audio super-resolution/enhancement
- **Lesson 2**: To generate high-level semantic information, **AR is preferred**. If T-S has no strong correspondence, AR is the best choice.
  - LLMs for text generation
  - Non-autoregressive NMT is a great lesson
- **Lesson 3**: To generate mid-level semantic/acoustic information, **NAR has advantages**, if T-S has strong dependency, and speed/robustness are considered
  - NAR-based acoustic model in TTS, speed/robustness are better than AR-based
  - e.g., FastSpeech 2 vs Transformer TTS

Iterative NAR can also do well in modeling T-T dependency!

Trend: LLMs Revive the AR/NAR Battle

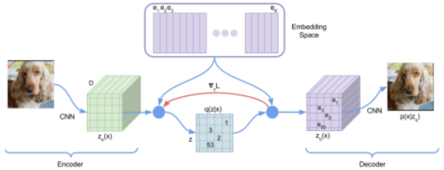
	Acoustic Model		Vocoder	
Timeline	AR	NAR	AR	NAR
2017.06	Tacotron		WaveNet	
2017.12	Tacotron 2, DeepVoice 3			Par. WaveNet
2018.06			WaveRNN	
2018.12	Transformer TTS		LPCNet	WaveGlow
2019.06		FastSpeech		
2019.12				MelGAN, Par. WaveGAN
2020.06		FastSpeech 2, Glow-TTS		
2020.12				DiffWave, WaveGrad, HiFiGAN
2021.06		GradTTS		
2021.12		VITS		**SoundStream**
2022.06		NaturalSpeech		BigVGAN
2022.12	**AudioLM**			**EnCodec**
2023.06	**VALL-E, SPEAR-TTS**	**NaturalSpeech 2**		
2023.12	**UniAudio**	**SoundStorm, VoiceBox**		

Lesson 4: The Goal/Paradigm of TTS Has Shifted In the New Era

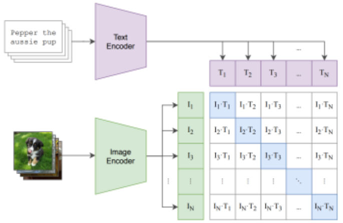
- Original goal: synthesize intelligible and natural speech
  - Intelligible: **achieved**
  - Natural: quality on limited styles/speakers/languages, **achieved**
- Goal now: natural and human-like
  - Diverse styles/speakers/languages
  - Huge effort to cover so many varieties
    - Prosody/emotion/style: **unlimited** variety
    - Speaker/timbre: **billions** of speakers in the world
    - Content/language: **thousands** of languages
- The paradigm to achieve the new goal
  - **Pre-train** on large-scale/diverse data
  - **Fine-tune** on specific style/speaker/language
  - **Zero-shot/in-context learning** on novel styles/speakers/languages

Recap LLM-based TTS

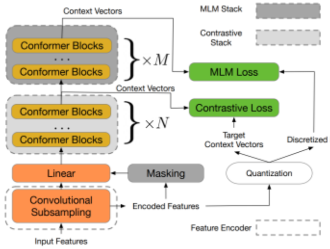
- Neural data representation/tokenization



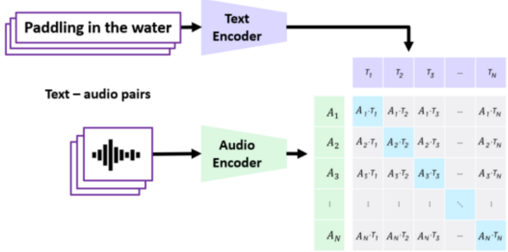
(Image)
VQ-VAE



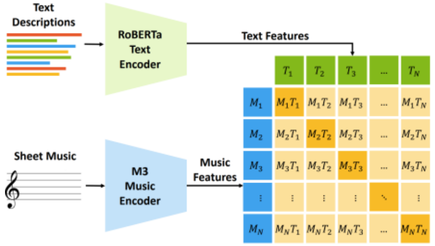
(Text-Image)
CLIP



(Audio)
w2v-BERT



(Text-Audio)
CLAP

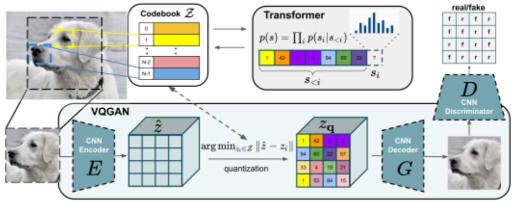


(Text-Audio)
CLaMP

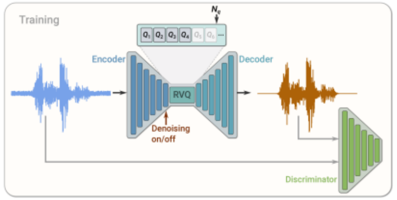
2020.12 2021.7 2021.12 2022.8

2017.11 2021.3 2021.8 2022.6 2023.4

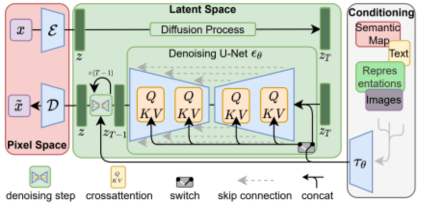
VQ-GAN
(Image)



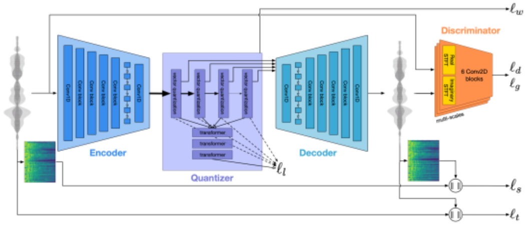
SoundStream
(Audio)



Stable Diffusion
(Image)

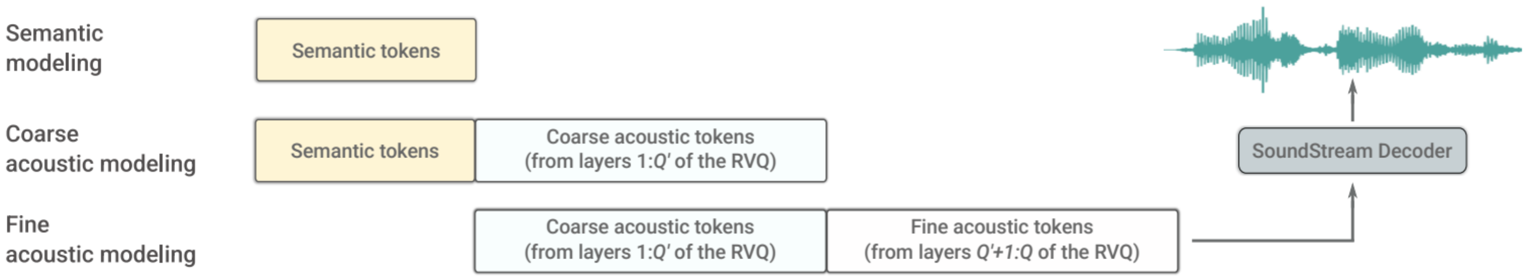


Encodec
(Audio)



Recap LLM-based TTS

- Transformer and decoder-only based LLMs
  - **AudioLM**: 1) Semantic, 2) Semantic→Coarse Acoustic, 3) Coarse Acoustic→ Fine Acoustic
  - **SPEAR-TTS**: 1) Text → Semantic Tokens, 2) Semantic → Acoustic
  - **VALL-E**: 1) Text→Acoustic 1st, 2) Acoustic 1st→Acoustic 2nd -8th (NAR)



Semantic modeling

Semantic tokens

Coarse acoustic modeling

Semantic tokens

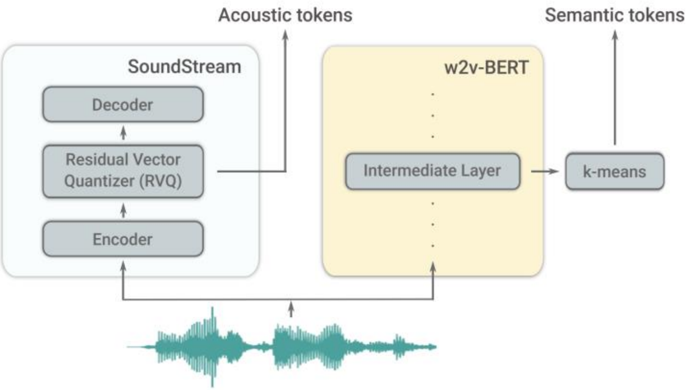
Coarse acoustic tokens (from layers 1:Q′ of the RVQ)

Fine acoustic modeling

Coarse acoustic tokens (from layers 1:Q′ of the RVQ)

Fine acoustic tokens (from layers Q′+1:Q of the RVQ)

SoundStream Decoder



Acoustic tokens

Semantic tokens

SoundStream

Decoder

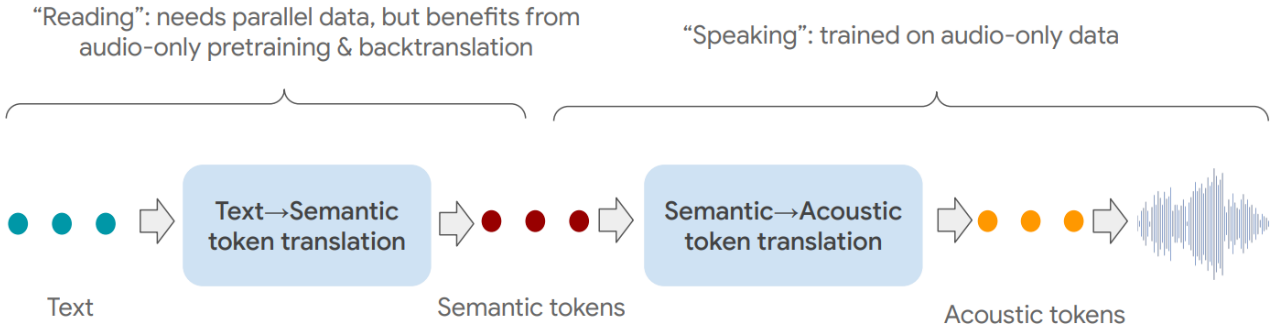
Residual Vector Quantizer (RVQ)

Encoder

w2v-BERT

Intermediate Layer

k-means



“Reading”: needs parallel data, but benefits from audio-only pretraining & backtranslation

“Speaking”: trained on audio-only data

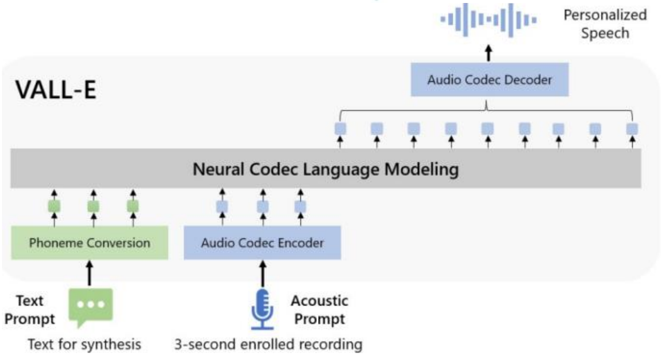
Text→Semantic token translation

Semantic→Acoustic token translation

Text

Semantic tokens

Acoustic tokens



Personalized Speech

VALL-E

Audio Codec Decoder

Neural Codec Language Modeling

Phoneme Conversion

Audio Codec Encoder

Text Prompt

Text for synthesis

Acoustic Prompt

3-second enrolled recording

Lesson 5: Data/Model Scaling (Out)Weigh Domain Knowledge

- With LLMs and data/model scaling, AR show competitiveness against with NAR
  - Prior domain knowledge (**duration alignment**) show advantages before the LLM era
  - Simple data/modeling scaling (**hundreds of thousands or millions of hours**) weigh or outweigh
  - **Inspirations from other areas (i.e., LLMs) can bring new variables in the battle that was originally going to be lost**
- Perspective
  - Practitioners in TTS: research or product
  - Practitioners in language/speech, audio domain, multimodality

Lesson 6: The AR/NAR Battle Is Not A Zero-Sum Game

	AR	NAR
Models	- AudioLM - VALL-E - SPEAR-TTS - UniAudio	- NaturalSpeech 2 - SoundStorm - Mega-TTS - VoiceBox
Pros	- Stand on the shoulder of LLMs (e.g., in-context learning, scalability) - Diverse/expressive (sampling)	- Stable/Robust - Fast inference - Control/Disentangle
Cons	- Not stable/robust (severe in 0-shot) - Slow inference - Long sequence (complex pipeline)	- Over-smoothness (fidelity, prosody) and less diversity - Complicated alignment process

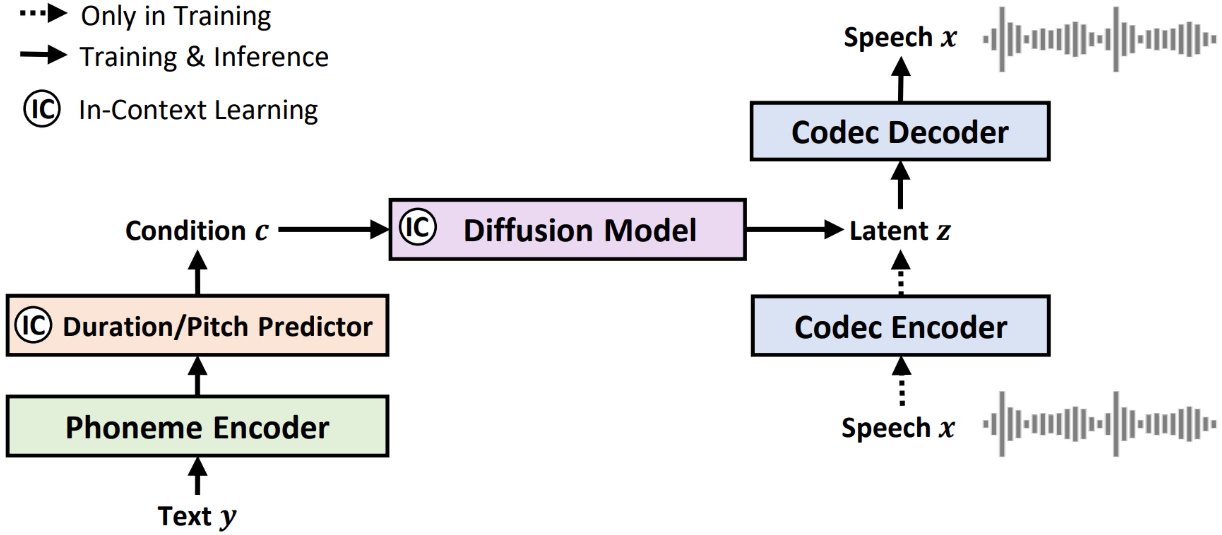
Difference	AR	NAR	Impact
Data Factorization	Along time axis	Along noise level	
Alignment/duration	Implicitly	Explicitly	Stable/Robust, Flexible
Iteration steps	Sequence length	Flexible	Fast

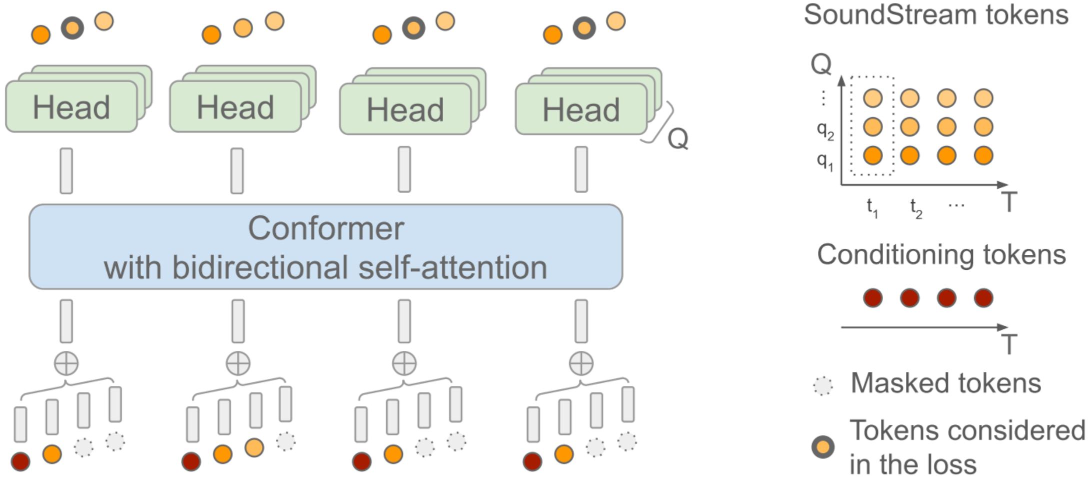
Lesson 6: The AR/NAR Battle Is Not A Zero-Sum Game

- AR/LLM-based and NAR-based TTS models have different application scenarios
  - AR-based has better **diversity, prosody, expressiveness, and flexibility** than NAR model
  - NAR is better in **speed and robustness**
  - After single-speaker finetuning, **AR models also has few bad cases**, although loses zero-shot capabilities
  - NAR is better in **disentanglement and control** (timbre, prosody, etc)
  - Combine AR and NAR: **semantic-level AR + perceptual-level NAR**

Lesson 7: Tokenization/Sampling Is Critical for Diversity/Expressiveness

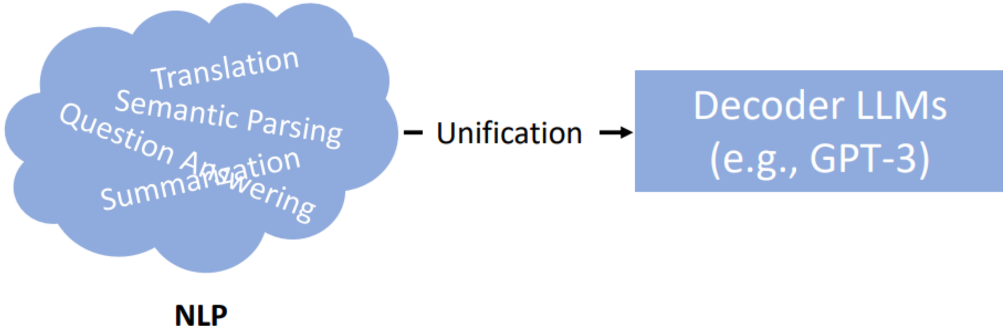
- Tokenization: softmax and cross-entropy
  - Classification to model diverse distribution and support sampling, instead of regression (GAN, VAE, Flow, Diffusion)
  - **Not only benefit for AR but also NAR** (NAR can model discrete tokens)
  - e.g., NaturalSpeech 2 (latent diffusion model with L_{ce_rvq} loss) and SoundStorm

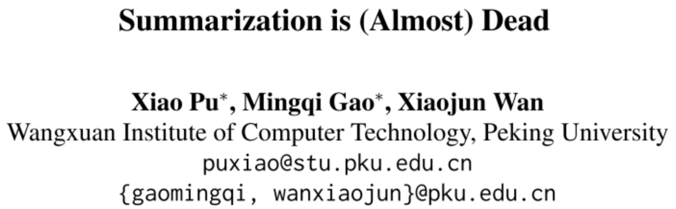




Lesson 8: Think Outside The Box: The Real Competition May Not Come From Within The Field

- The advantage of LLMs is **scalability and flexibility**, instead of perfect performance on every single task
  - Do not care winning or losing battles but care the war!
- A lesson from NLP

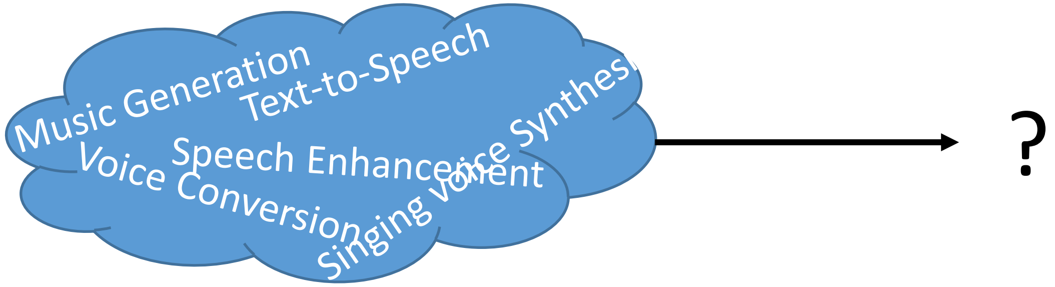




Summarization is (Almost) Dead

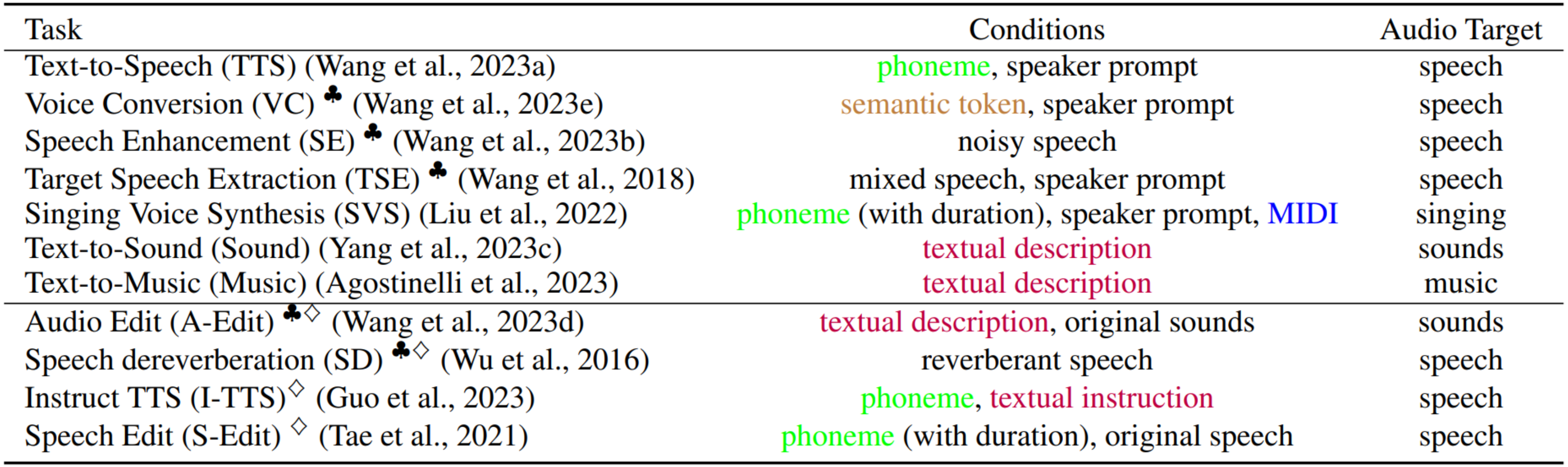
Xiao Pu*, Mingqi Gao*, Xiaojun Wan
Wangxuan Institute of Computer Technology, Peking University
puxiao@stu.pku.edu.cn
{gaomingqi, wanxiaojun}@pku.edu.cn

- How about speech/audio?



Lesson 8: Think Outside The Box: The Real Competition May Not Come From Within The Field

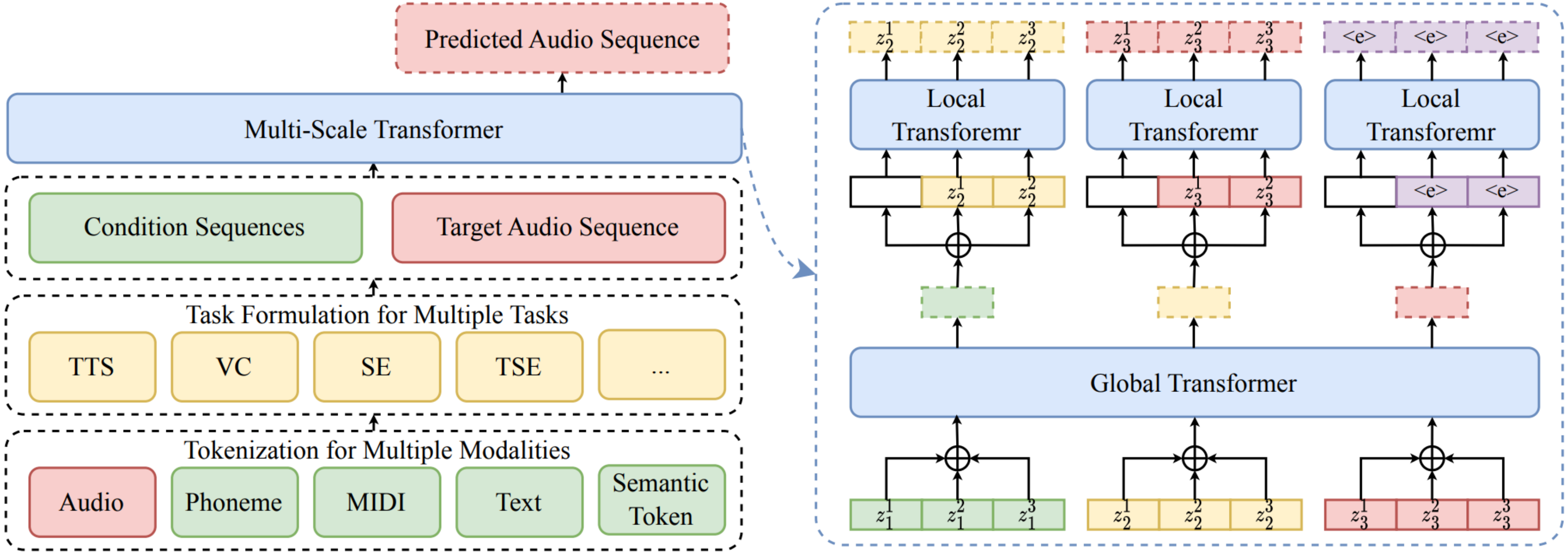
- UniAudio: Unify all audio (speech, singing, music, sound) generation tasks in a single LLM
  - Task formulation: concatenate condition-target as a single sequence
  - e.g., <start> <audio_task> <text_start> text_sequence <text_end> <audio_start> audio_sequence <audio_end> <end>



Task	Conditions	Audio Target
Text-to-Speech (TTS) (Wang et al., 2023a)	phoneme, speaker prompt	speech
Voice Conversion (VC) ♣ (Wang et al., 2023e)	semantic token, speaker prompt	speech
Speech Enhancement (SE) ♣ (Wang et al., 2023b)	noisy speech	speech
Target Speech Extraction (TSE) ♣ (Wang et al., 2018)	mixed speech, speaker prompt	speech
Singing Voice Synthesis (SVS) (Liu et al., 2022)	phoneme (with duration), speaker prompt, MIDI	singing
Text-to-Sound (Sound) (Yang et al., 2023c)	textual description	sounds
Text-to-Music (Music) (Agostinelli et al., 2023)	textual description	music
Audio Edit (A-Edit) ♣◇ (Wang et al., 2023d)	textual description, original sounds	sounds
Speech dereverberation (SD) ♣◇ (Wu et al., 2016)	reverberant speech	speech
Instruct TTS (I-TTS)◇ (Guo et al., 2023)	phoneme, textual instruction	speech
Speech Edit (S-Edit) ◇ (Tae et al., 2021)	phoneme (with duration), original speech	speech

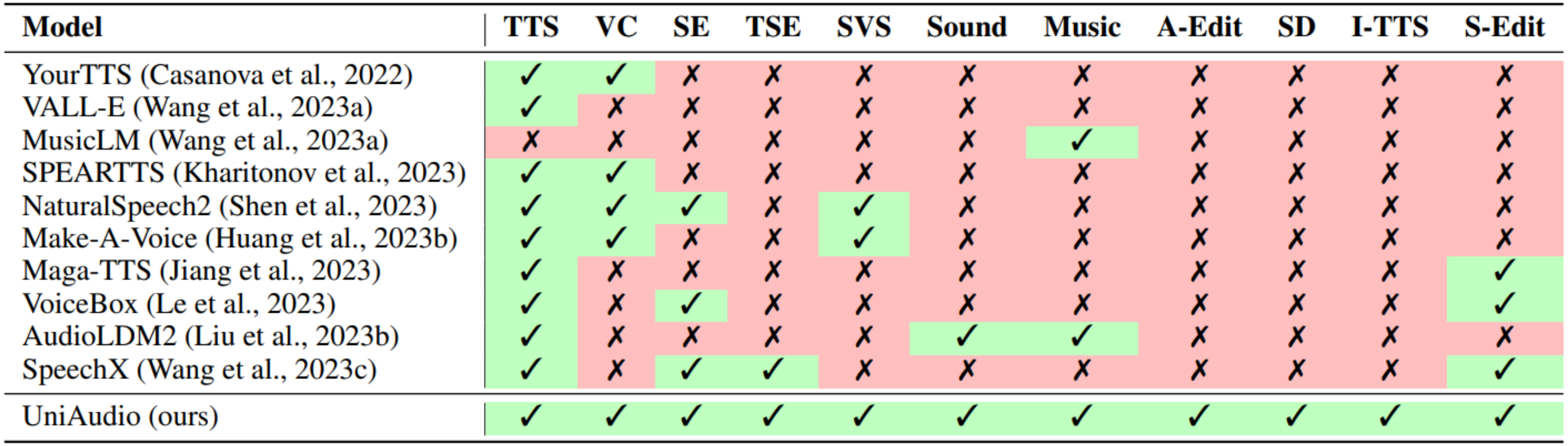
Lesson 8: Think Outside The Box: The Real Competition May Not Come From Within The Field

- UniAudio: Unify all audio (speech, singing, music, sound) generation tasks in a single LLM



Lesson 8: Think Outside The Box: The Real Competition May Not Come From Within The Field

- UniAudio: Unify all audio (speech, singing, music, sound) generation tasks in a single LLM



Model	TTS	VC	SE	TSE	SVS	Sound	Music	A-Edit	SD	I-TTS	S-Edit
YourTTS (Casanova et al., 2022)	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
VALL-E (Wang et al., 2023a)	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
MusicLM (Wang et al., 2023a)	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
SPEARTTS (Kharitonov et al., 2023)	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
NaturalSpeech2 (Shen et al., 2023)	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗
Make-A-Voice (Huang et al., 2023b)	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗
Maga-TTS (Jiang et al., 2023)	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
VoiceBox (Le et al., 2023)	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓
AudioLDM2 (Liu et al., 2023b)	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗
SpeechX (Wang et al., 2023c)	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✓
UniAudio (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Lesson 8: Think Outside The Box: The Real Competition May Not Come From Within The Field

- The advantage of LLMs is **scalability and flexibility**, instead of perfect performance on every single task
  - Do not care winning or losing battles but care the war!
  - UniAudio-like work will dominate the whole audio tasks, not merely TTS or generation
  - Universal task support (speech/singing/music/sound, understanding/generation), next word prediction, scaling law, in-context learning, prompting

Lessons Learned

- **Lesson 1**: To generate low-level perceptual detail, NAR is preferred. If T-S has strong dependency, NAR is the best choice
- **Lesson 2**: To generate high-level semantic information, AR is preferred. If T-S has no strong dependency, AR is the best choice
- **Lesson 3**: To generate mid-level semantic/acoustic information, NAR has advantages, if T-S has strong dependency, and speed/robustness are considered
- **Lesson 4**: The goal/paradigm of TTS has shifted in the new era
- **Lesson 5**: Data/model scaling (out)weigh domain knowledge
- **Lesson 6**: The AR/NAR battle is not a zero-sum game
- **Lesson 7**: Tokenization/sampling is critical for diversity/expressiveness
- **Lesson 8**: Think outside the box: the real competition may not come from within the field

Tips From These Lessons

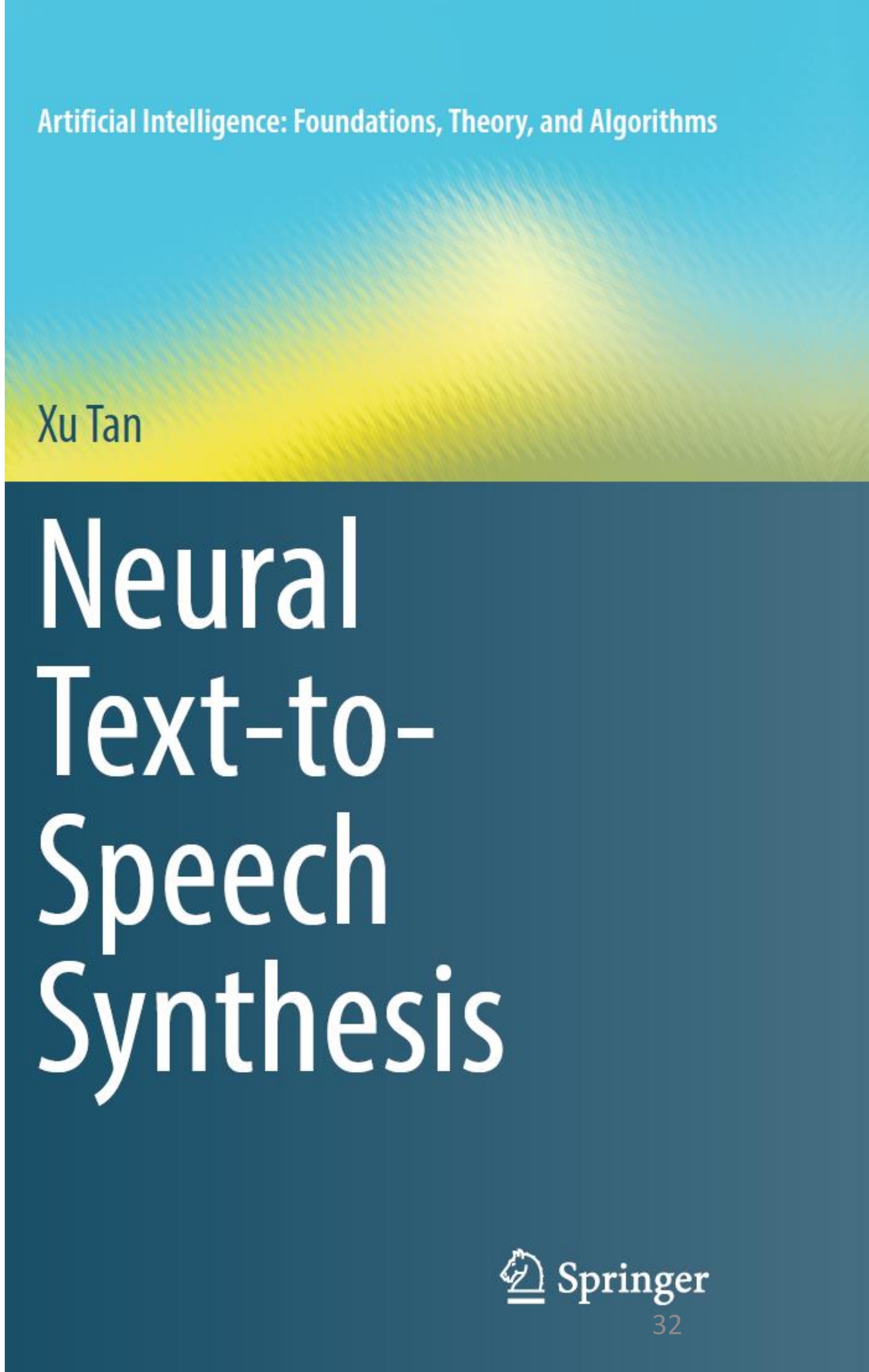
- **Tip 1**: **Choose AR/NAR** according to your scenarios (more signal/perceptual or semantic/contextual, fast inference, streaming, high-quality single speaker, zero-shot, stableness, scalability?)
- **Tip 2**: **Exploit NAR**, e.g., tokenization/sampling, disentanglement/control, stable zero-shot
- **Tip 3**: **Explore AR**, beyond speech synthesis, ChatGPT moment in audio domain
- **Tip 4**: **Scale** data/model/task, explore the unknow

Thanks

A book on “*Neural Text-to-Speech Synthesis*”

published by Springer!

https://link.springer.com/book/9789819908264



Artificial Intelligence: Foundations, Theory, and Algorithms

Xu Tan

Neural Text-to-Speech Synthesis

Springer

Thank You!

https://www.microsoft.com/en-us/research/people/xuta/
https://speechresearch.github.io/
tan-xu.github.io