

高等机器学习

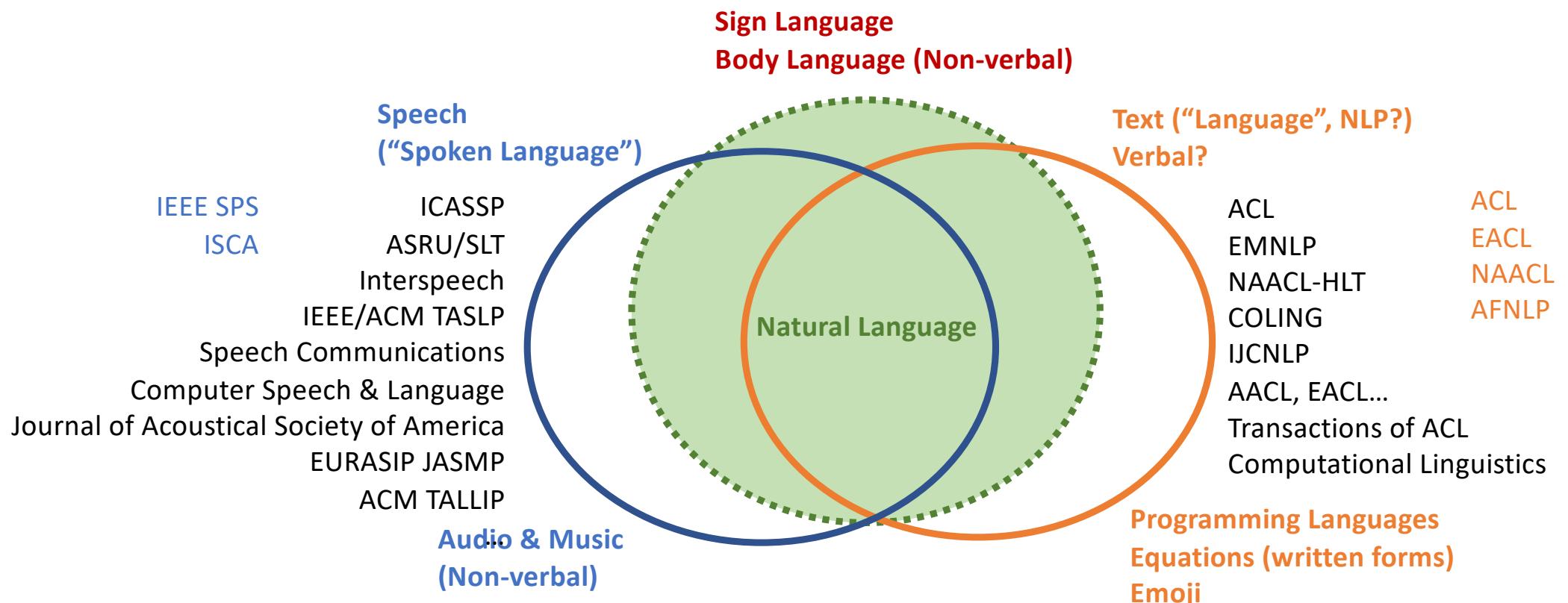
# 语音处理(上)

张超 (图信所)  
清华大学电子工程系



清华大学  
Tsinghua University

# Speech vs Text: An ML Perspective



# Speech vs Text: A Scientific Perspective

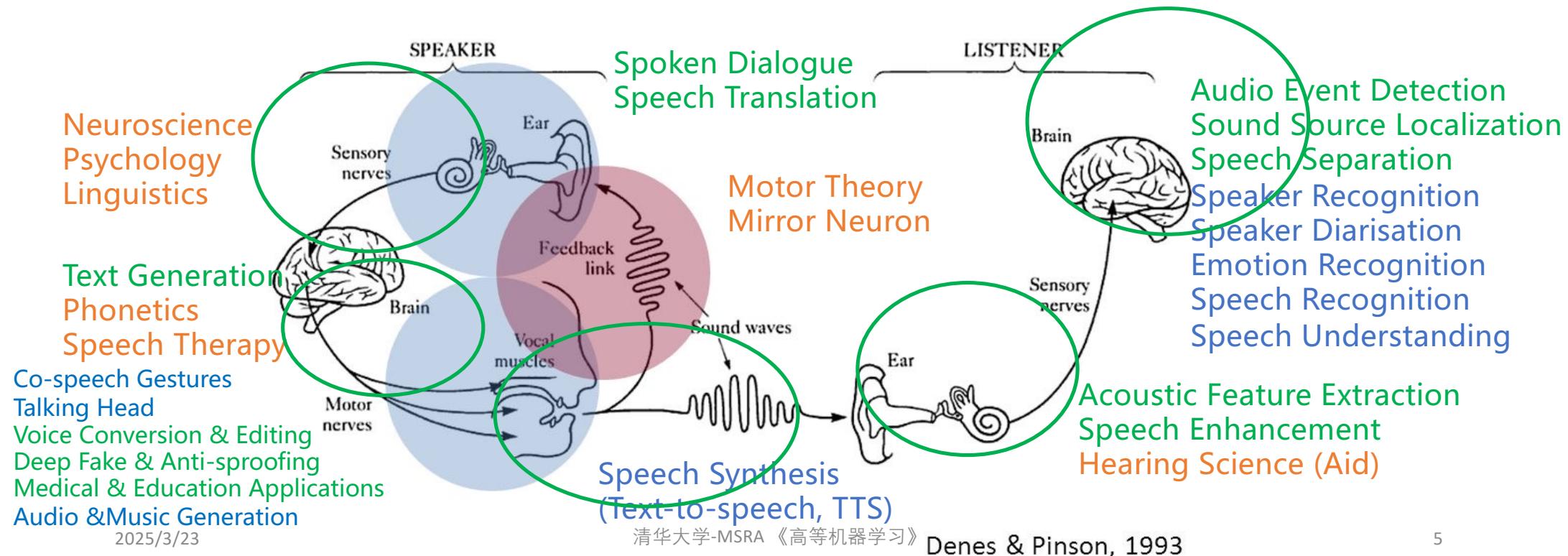
- Speech
    - Speech comprehension (complex): A defining ability of human beings
    - More natural: Developed in evolution & used in language acquisition
      - Include more information: Pronunciations, time, prosody, speaker etc.
      - Used more casually & frequently
  - Text
    - An invention, school education needed
    - A simplification of spoken language
      - More efficient & abstractive: Faster perception, used for formal languages
      - Used more formally, much cheaper (**10 times!**)
- Streaming, multi-party,  
audio-video (audio-visual)      7,100 (spoken) languages  
in the world. Many don't  
have written forms!

# The King's Speech

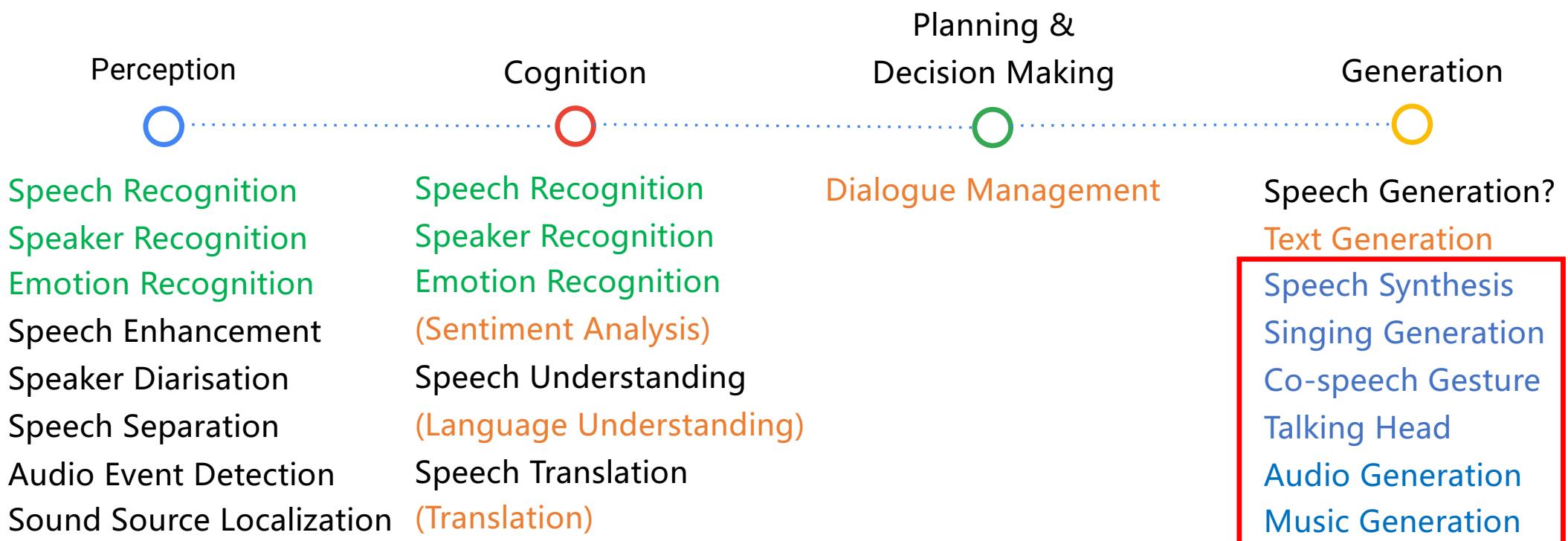


# Speech Science & Technology

## The Speech Chain

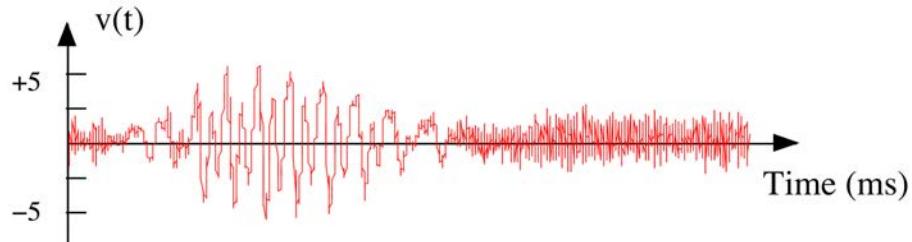


# Speech vs Text: An AI Perspective



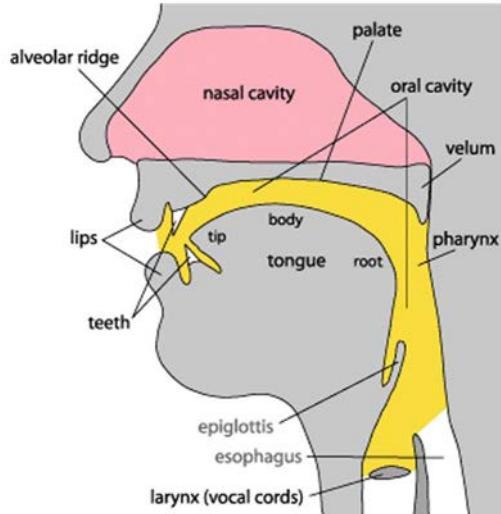
# The Speech Waveform: One-dimensional, High-frequency, Sequential, Variable-length

- The speech waveform (electrical signal from a microphone representing the speech pressure waveform) is **non-stationary** and containing a mix of **pseudo-periodic** and **random** components

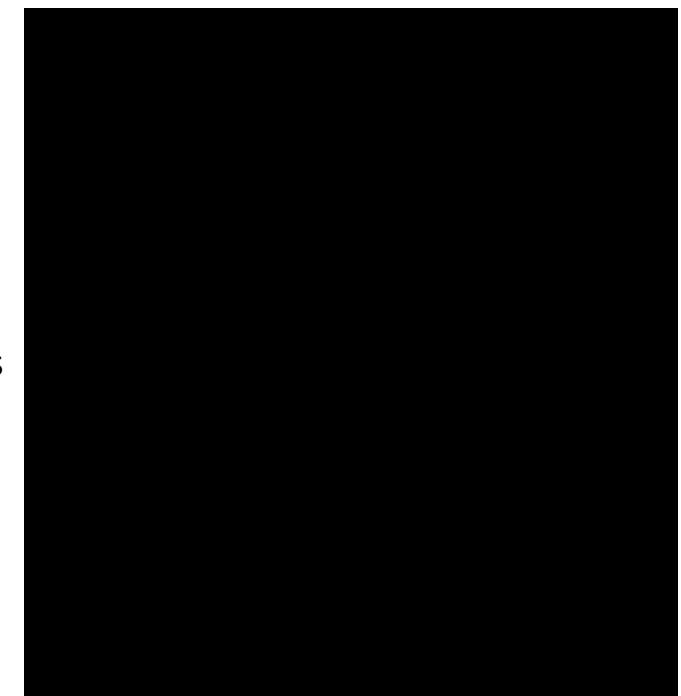


- Different classes of speech sounds have properties that depend on how they were produced.
- Two main components to human speech production mechanism:
  - A variably-shaped **acoustic tube**
  - An **excitation source** for the tube
- Some broad distinctions in speech sound are due to the type of excitation and detailed sounds are due to the shape of the tube.

# Speech Production: Vocal Tract



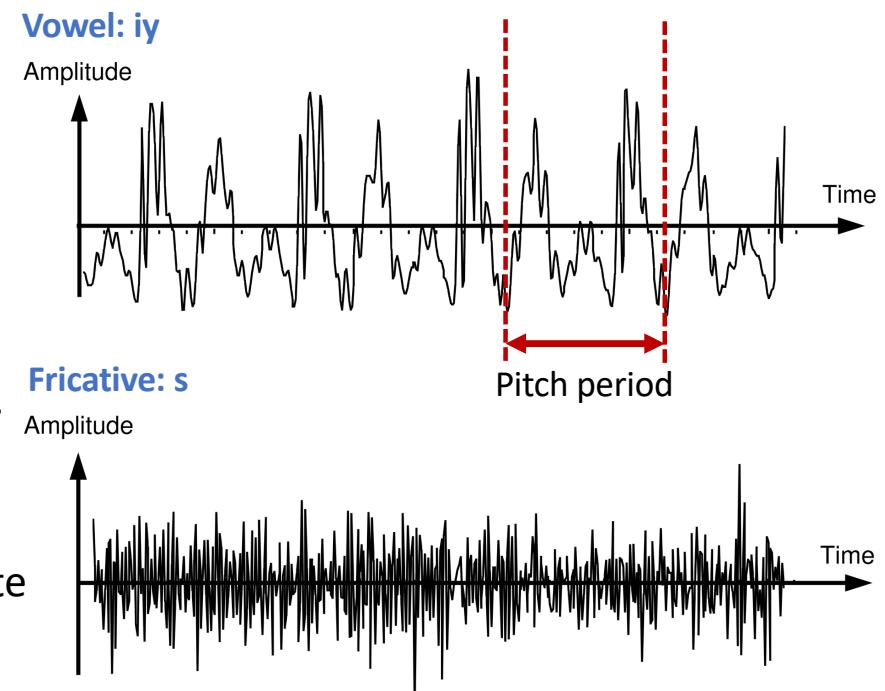
- Lungs push air through **larynx**
- **Vocal folds** vibrate to release air in puffs, creating a periodic excitation for voiced speech
- Unvoiced sounds from turbulence sound source in **vocal tract (fricatives)** or build up and release (**stops**)
- Articulators (**tongue, jaw** etc.) move to position for particular speech sound: move continually as sounds produced.
- Note neighbouring sounds effect realisation of each sound: **co-articulation**
- Resonances in the **vocal tract (formants)** alter the signal to shape the sound formed
- Side branch used in **nasal** sounds
- Sound pressure waves leave mouth and travel through air



MRI | Sehnsucht nach dem Frühlinge (Mozart) |  
Anna-Maria Hefele  
<https://www.youtube.com/watch?v=YIUVX7hebBA>

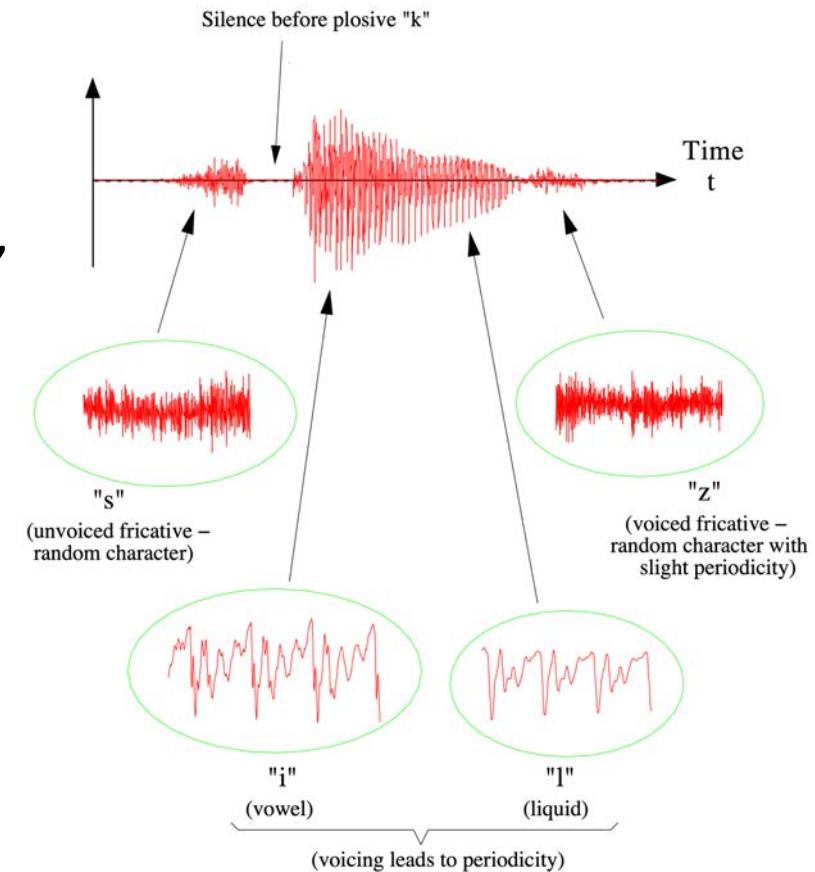
# Speech Production: Excitation Sources

- The acoustic tube has three sources of excitation.
- 1. Vocal cords which vibrate when air from the lungs is forced through them.
  - Leads to **voiced** sounds as in “feel”, “hit. The sounds are quasi-periodic at the **pitch** frequency.
- 2. Turbulence caused by forcing air through constrictions formed by raising the tongue to narrow the acoustic tube.
  - Leads to **fricative** sounds which appear random in the time-domain as in “feel”, “shoe”, etc.
- 3. Turbulence caused by release of air following a complete closure of the acoustic tube.
  - Leads to **stops** sounds as in “take”. Sounds may also have mixed excitation as for example in “zoo”.



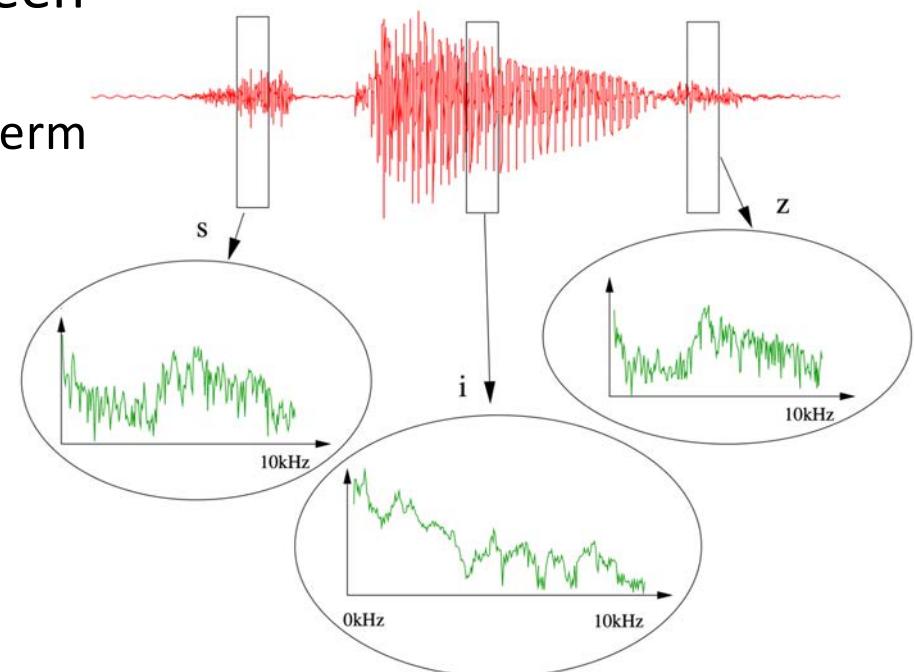
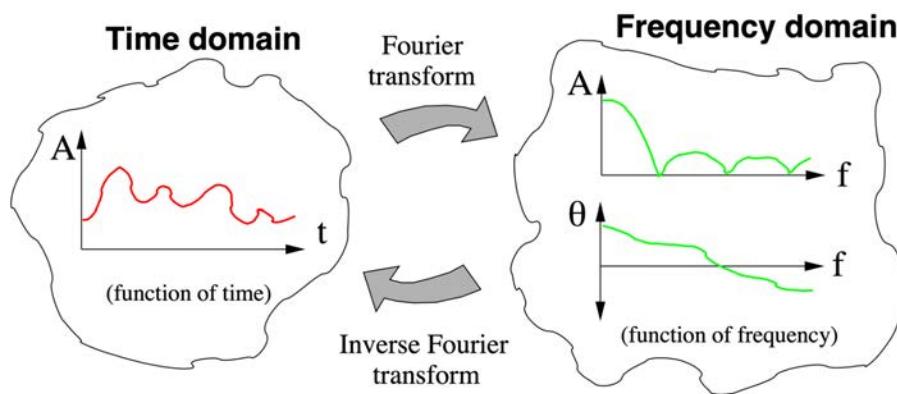
# The Sounds of Speech: “skills”

- Convenient to view speech as being composed of a sequence of sounds called phones (associated with basic units of speech, the phonemes).
  - 40 phones are often used for English: **Consonants & vowels.**
  - Initial / Finals (including zero initials) are often used for Putonghua.
  - Finals are compound vowels (1-3 vowels).
  - Phones (& IPA) are often not good enough to characterize pronunciations in practice .

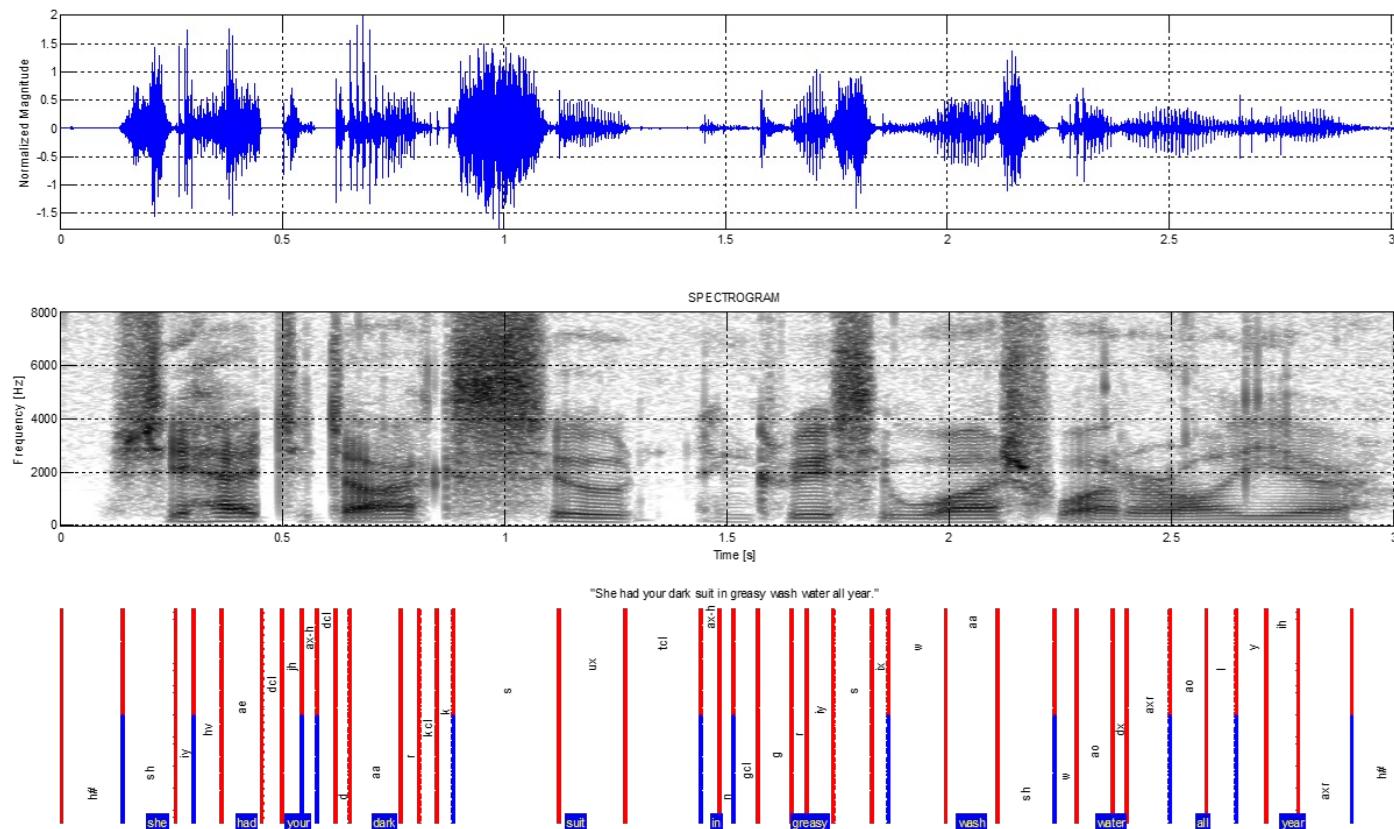


# Speech Representations: Spectrograms

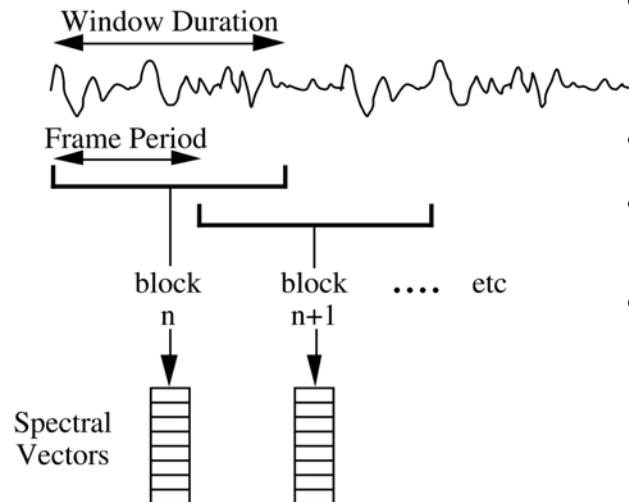
- Fourier Transform allows to move between **time domain & Frequency domain**
  - Divide speech into 10ms segments (short term stationary )
  - Spectral “envelope” is most important for describing speech sounds



# Speech Representations: Spectrograms



# Speech Representations: Acoustic features

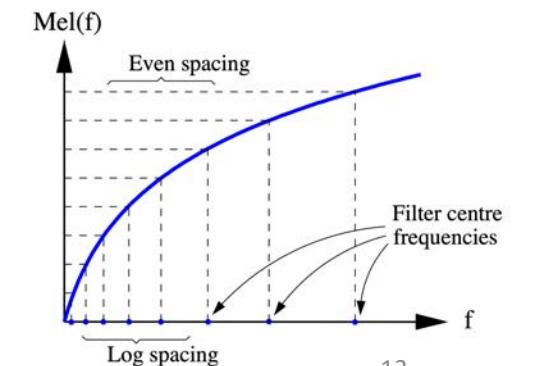
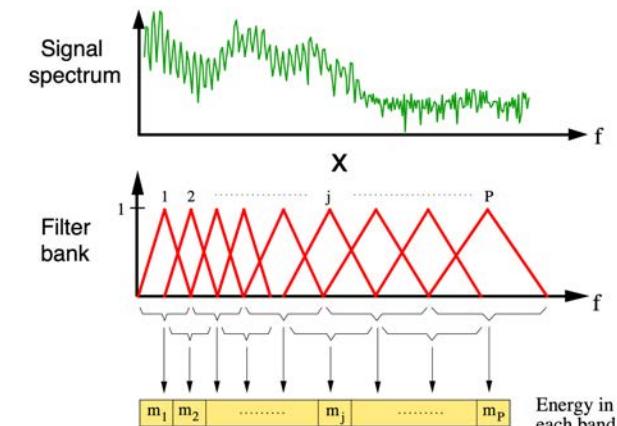


- Mel-scale Filterbanks

- Divide speech into **25ms window duration & 10ms frame period** (**short-term stationary assumption**)
- Drop **phase information**.
- The energy in each frequency band is computed using DFT.
- The spacing of the centre frequencies is based on the Mel-scale.

$$\text{Mel}(f) = 2595 \log_{10} \left( 1 + \frac{f}{700} \right)$$

- Reduce frequency resolution and analysis to model **ears spectral resolution**.
- MFCC, LPCC, PLP etc.

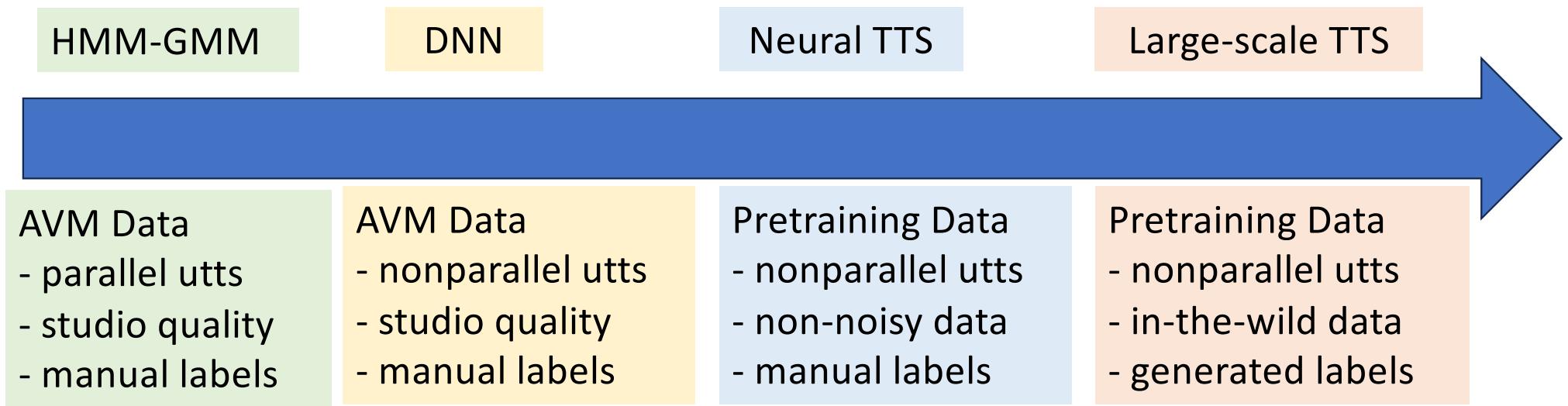


# Text-to-Speech Synthesis (TTS)

# Speech Synthesis

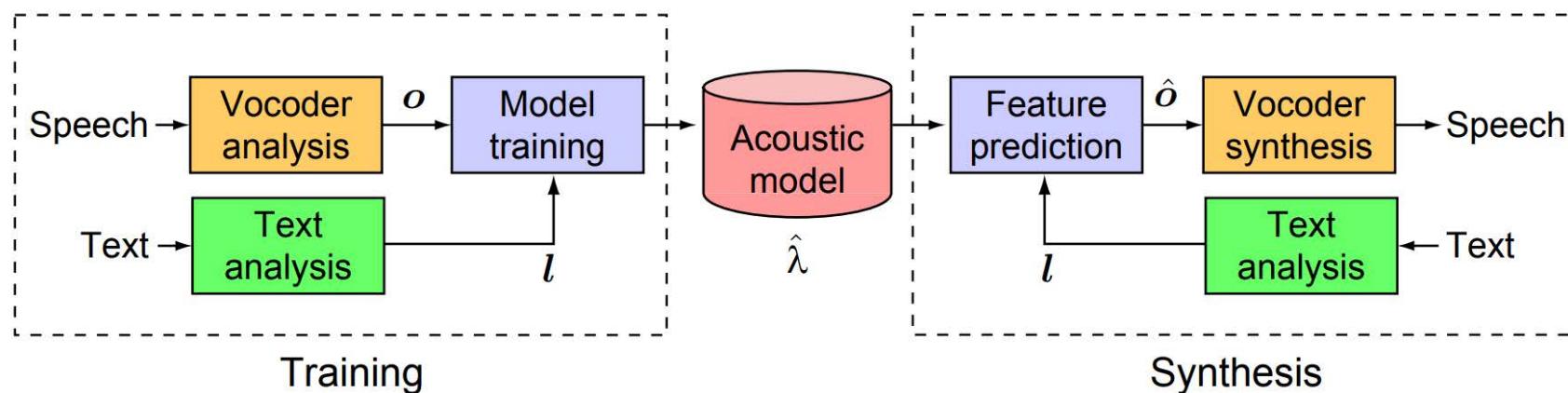
- Text (words for speech content) in, speech waveform out
- Evaluation: Intelligibility, naturalness, expressiveness
- Metrics
  - Objective: Mel-cepstral distortion, F0 frame error, Prosody std. dev., WER ...
  - Subjective: [Mean opinion scores](#) (MOS) – Automatic MOS being studied
- History
  - Concatenative speech synthesis
  - Statistical parametric speech synthesis (SPSS): HMM-based, DNN-based
  - End-to-end(?) speech synthesis
  - Neural codec-based speech synthesis

# Overview of Speech Synthesis Paradigms



# Statistical Parametric Speech Synthesis

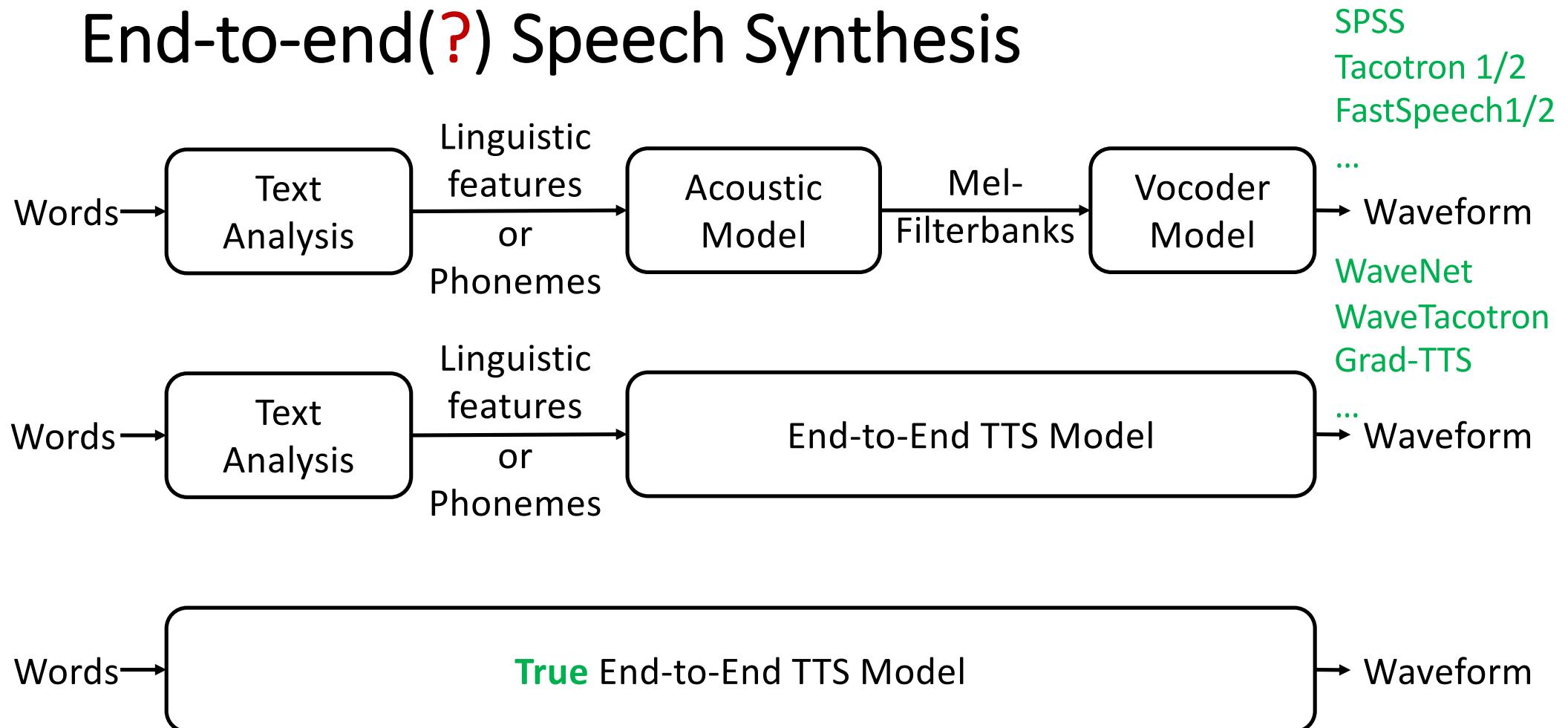
- Text analysis, acoustic model (HMM/DNN), and vocoder analysis/synthesis



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

H. Zen et al., “Statistical Parametric Speech Synthesis”, 2009.  
H. Zen et al., “Statistical Parametric Speech Synthesis using Deep Neural Networks”, 2013.  
H. Zen et al., “The HMM-based Speech Synthesis System (HTS) Version 2.0”, 2007.  
Z. Wang, “Improved Hidden Markov Model in Speech Recognition”, 1988.

# End-to-end(?) Speech Synthesis



# Acoustic Model

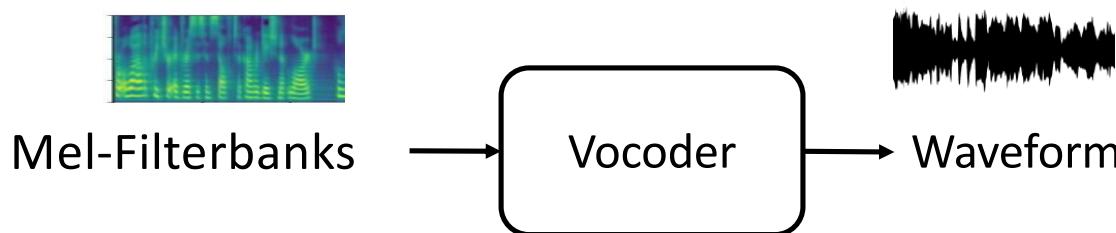
- Predict acoustic features from linguistic features



- HMM, BLSTM, Seq2Seq (LSTM, CNN, Transformer)
- The requirements for acoustic model
  - More context information (input)
  - Model correlation between frames (output)
  - Combat over-smoothing prediction
  - Alignment between linguistic and acoustic features

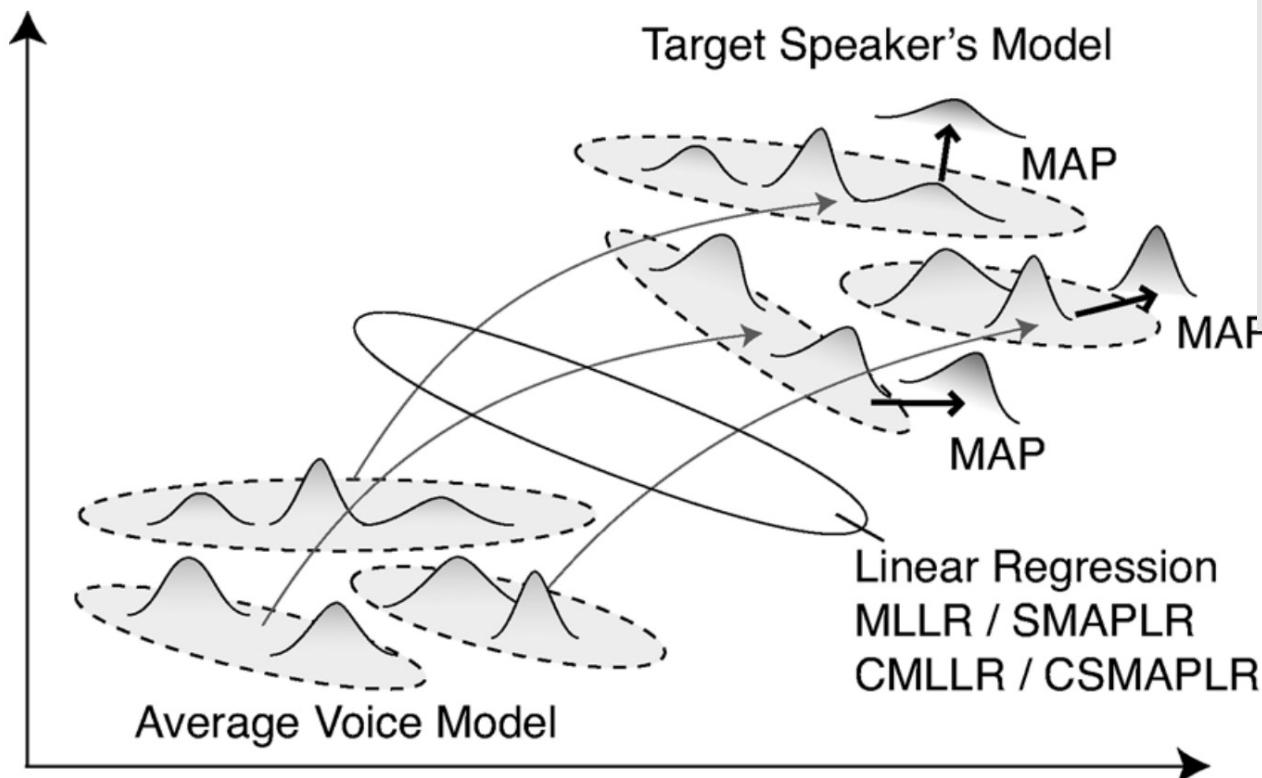
# Vocoder

- Model-free: Griffim Lim, WORLD etc.
- (Neural) Model-based:

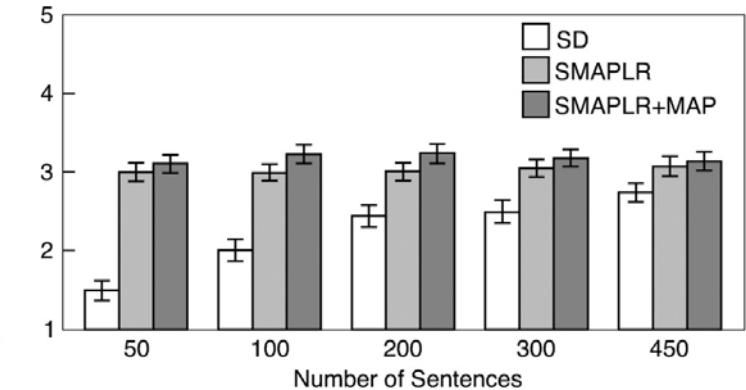


- WaveNet, ParallelWaveNet
- SampleRNN, WaveRNN, LPCNet
- GAN-based
- Flow-based
- Diffusion-based

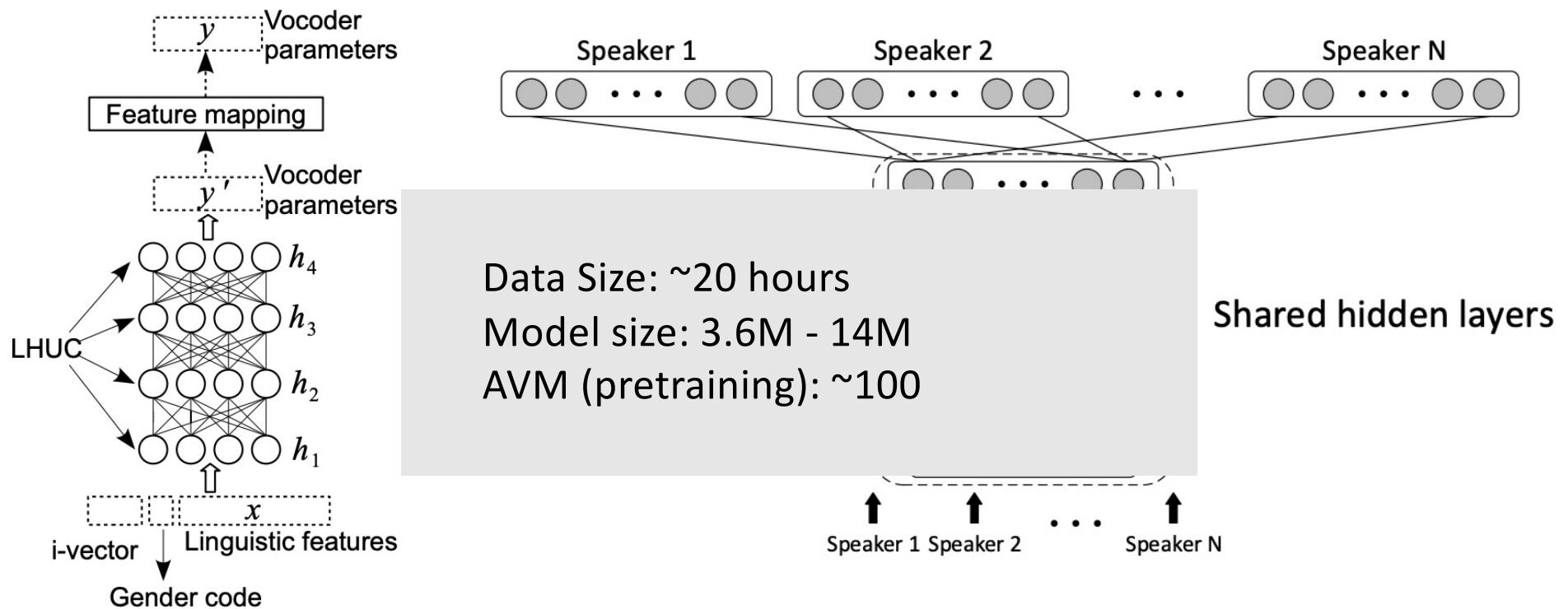
# Adaptation in GMM-HMM-based TTS



Data Size: ~5 hours (AVM)  
Model size: 1M - 7M  
AVM speakers: ~8 (4F, 4M)

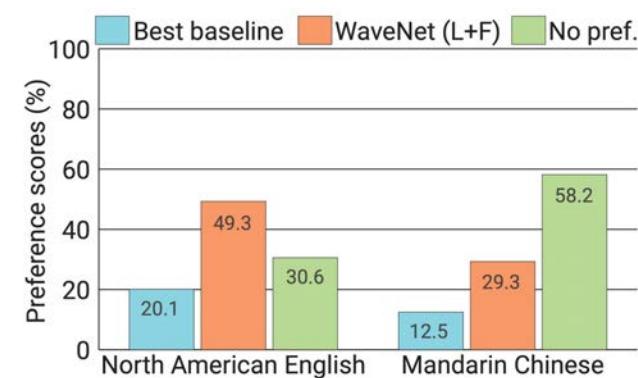
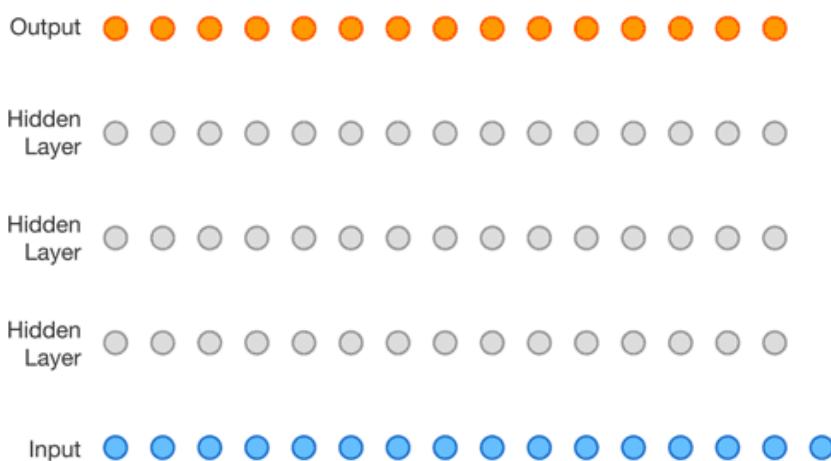


# Adaptation in DNN-HMM-based SPSS



# WaveNet

- First TTS directly generating raw waveform
    - Linguistic features in, raw waveform out.
    - Fully probabilistic & per-sample-level autoregressive (dilated causal CNNs)
    - Generating more natural speech than ever

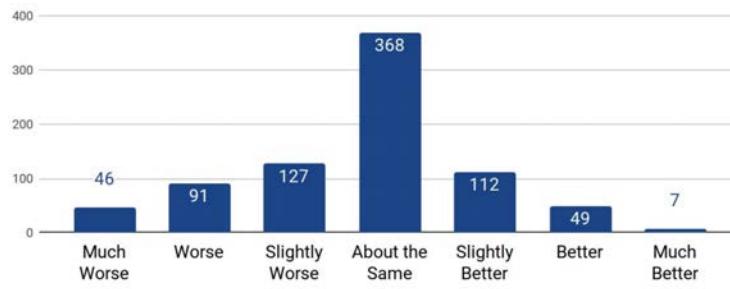


A. van den Oord et al., "WaveNet: A Generative Model for Raw Audio", 2016.

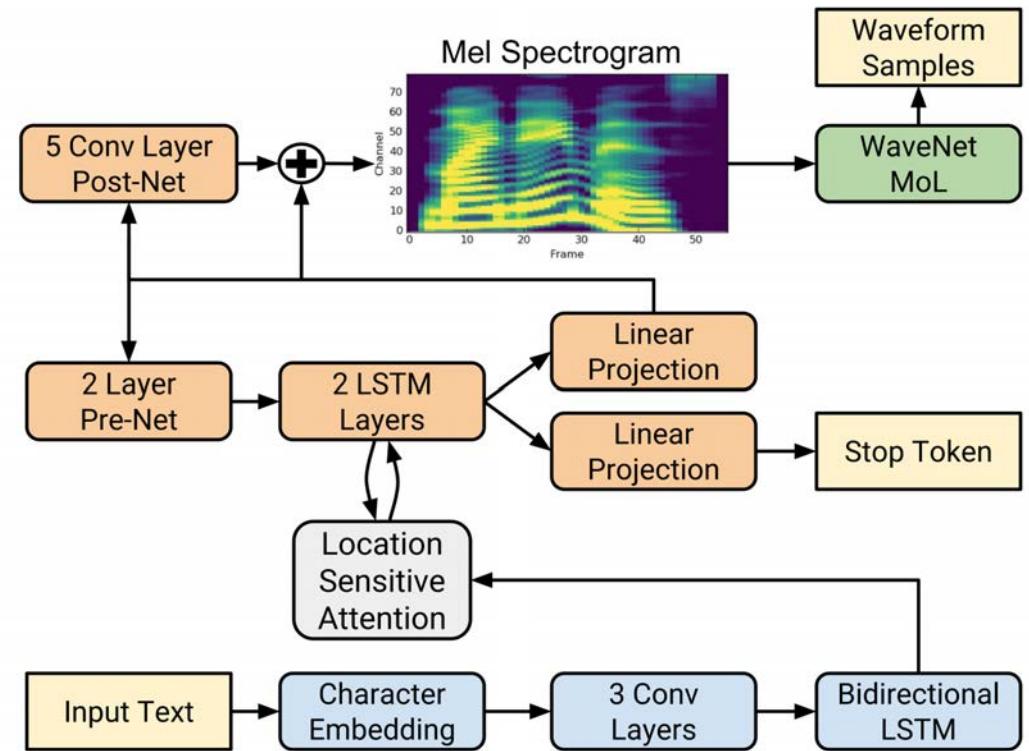
- Concatenative
- SPSS
- WaveNet

# Tacotron 2

- WaveNet
  - Can be **overly slow** due to per-sample-level autoregressive.
  - Requires complex text analysis.
- Tacotron
  - Uses Griffin-Lim for end-to-end training.
- Tacotron 2
  - Uses a WaveNet vocoder.
  - Achieved recording-level MOS.

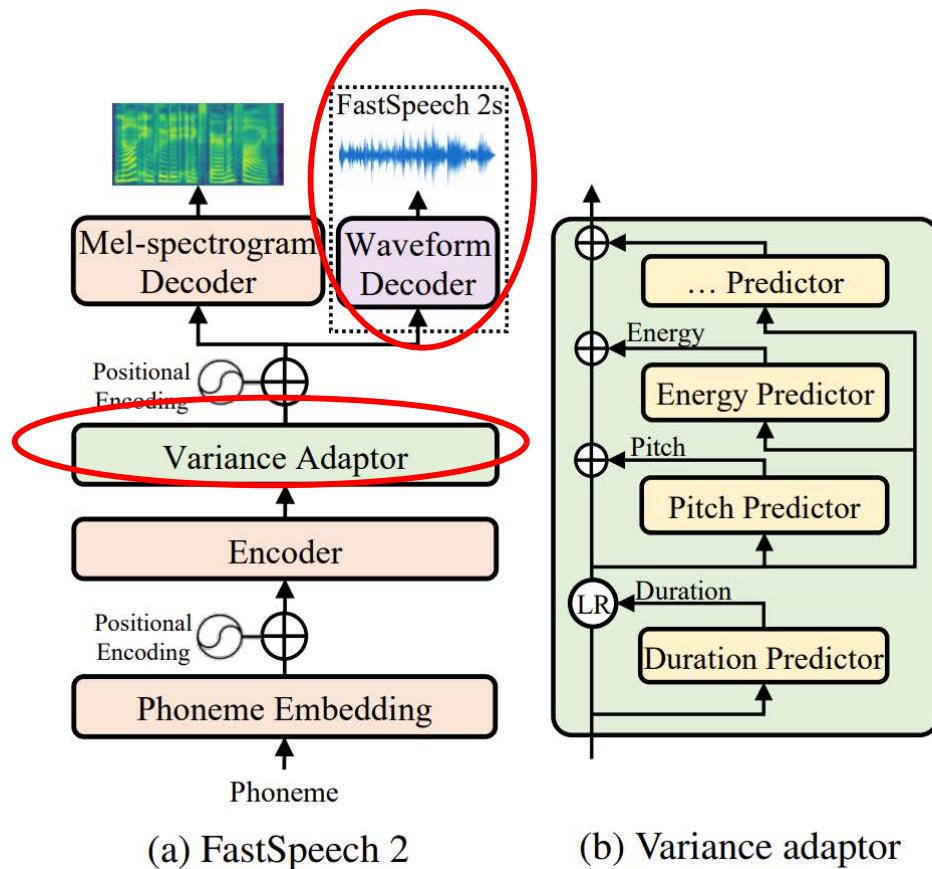


Tacotron 2  
Human  
Tacotron 2



Y. Wang et al., "Tacotron: Towards End-to-End Speech Synthesis", 2017.  
J. Shen et al., "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions", 2018.

# FastSpeech 2

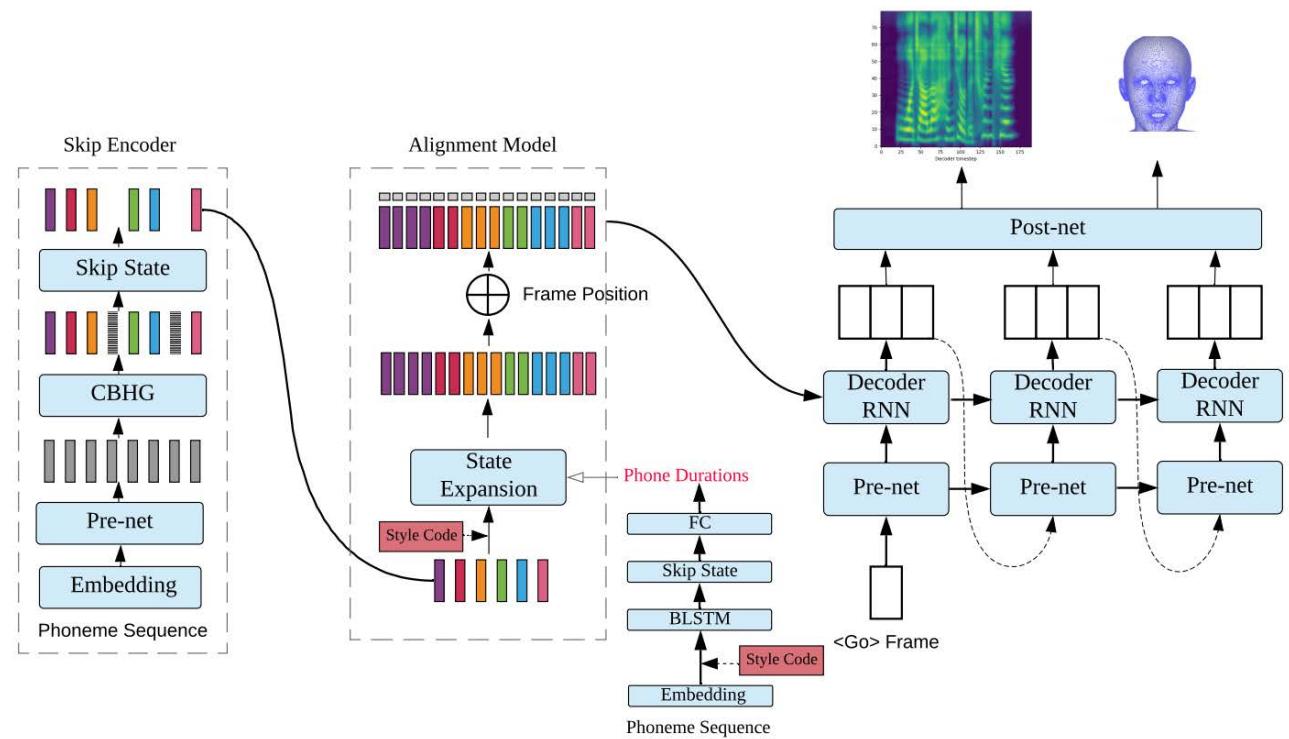


- **Problems:** Tacotron 2 suffers from
  - **Slow speed:** autoregressive decoder
  - **Not robust:** words skipping and repeating
- Planning-based: To predict duration, pitch, energy etc. (**SPSS with end-to-end planning**)
- **FastSpeech (1&) 2**
  - Generate Mel-spectrogram in parallel (**speed**)
  - Remove the text-speech attention (**robust**)
  - Can manipulate outputs of predictors (**control**)
  - Requires **forced-alignment** for duration **in training**

Y. Ren et al., “Fasan tSpeech: Fast, Robust, and Controllable Text to Speech”, 2019.  
Y. Ren et al., “FastSpeech 2: Fast and High-Quality End-to-End Text-to-Speech”, 2021.

# DurlAN: Duration Informed Attention Network

- A hybrid of SPSS (enc.) & Tacotron (dec.)
  - Use an alignment model to avoid the instability with attention
  - **Forced alignment required in training** for phoneme durations
  - Empirically can be as expressive as Tacotron



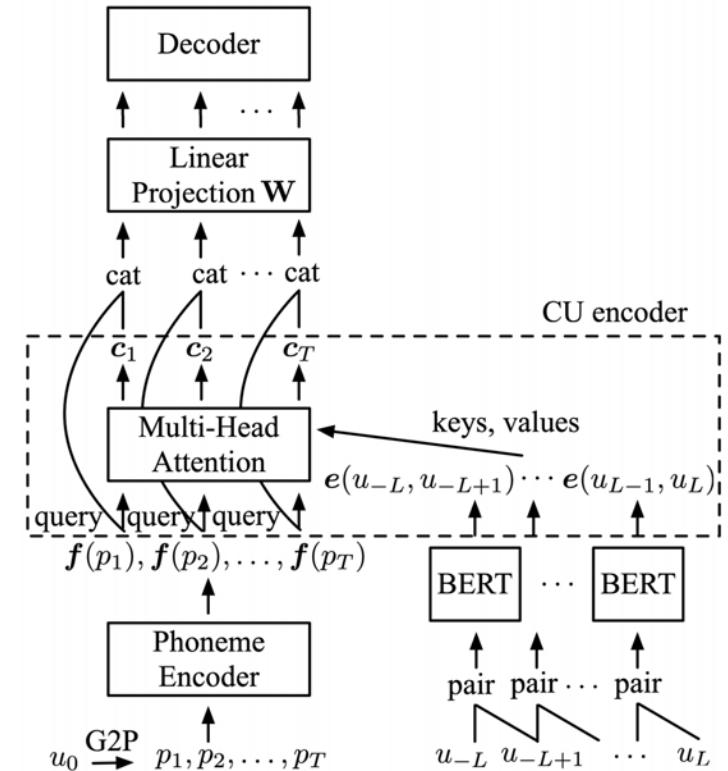
C. Yu et al., "DurlAN: Duration Informed Attention Network for Speech Synthesis", 2020.

# Cross-Utterance TTS

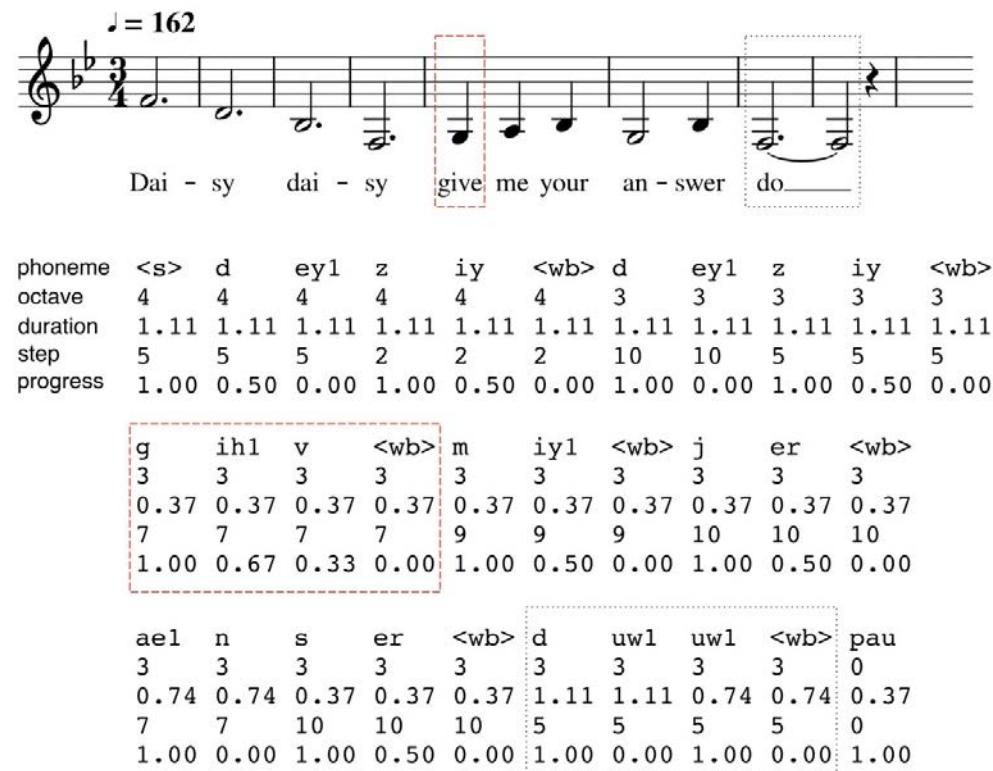
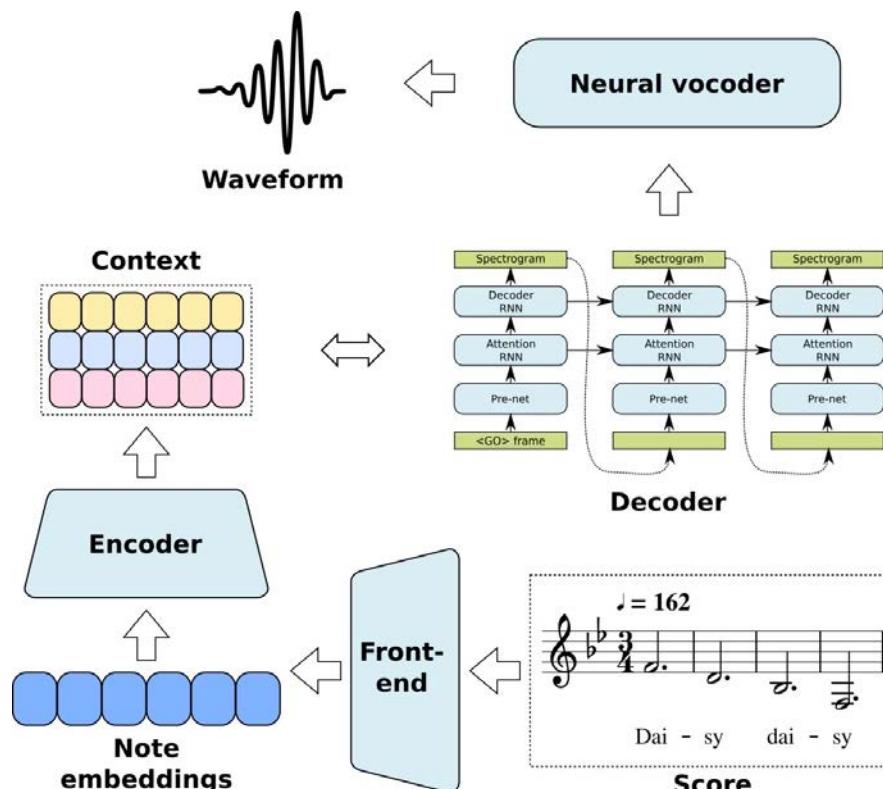
- Use words from context utterances to improve prosody of current utterance
  - Uses an extra attention module to query the cross-utterance contexts every decoding step.
  - Use BERT to encode context utterance words to **generalise better** to unseen texts.
  - **Improves stability** (Tacotron 2) when training on difficult (**more natural & expressive**) data.
  - Generate **more reasonable prosody** with context words.
  - Also applied to other structures & previous acoustic features.  
with Griffin-Lim vocoder

Tacotron 2

Cross-Utterance



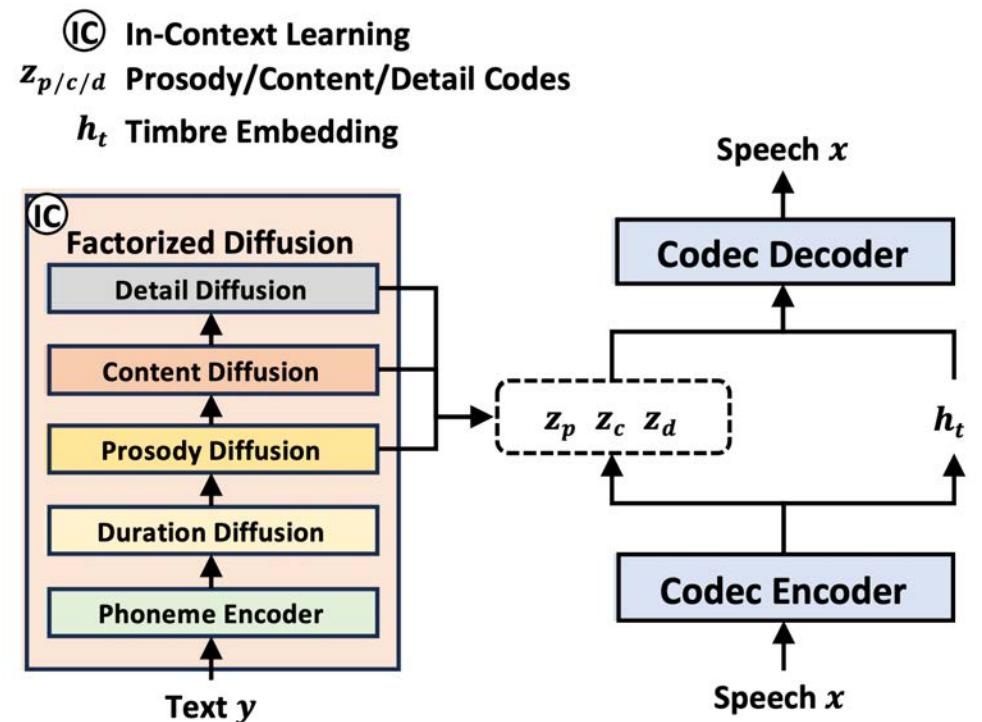
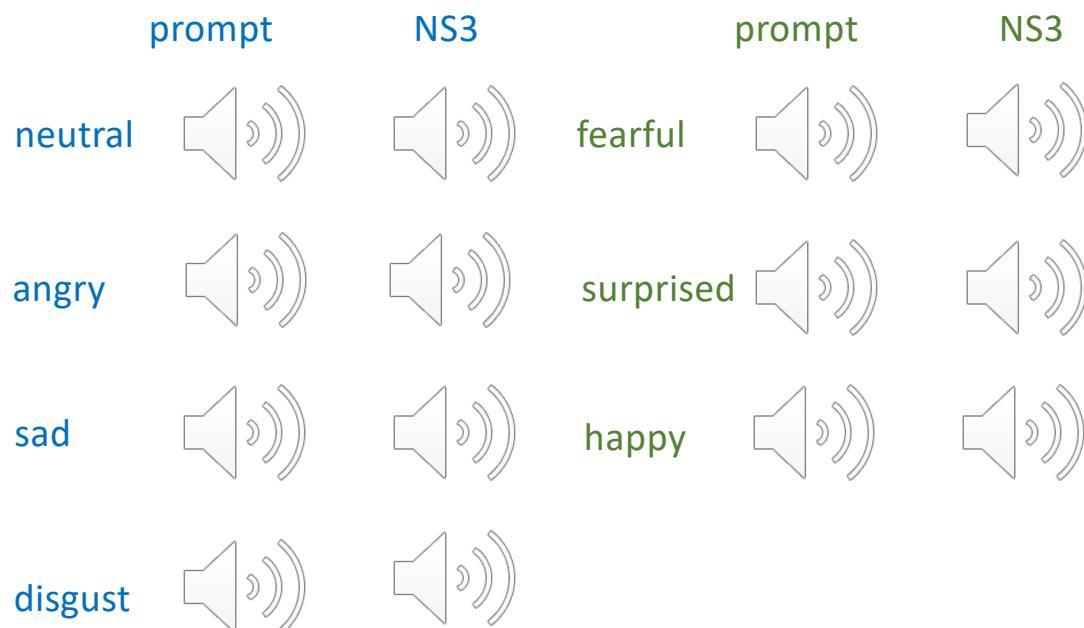
# Singing Voice Synthesis with Seq-to-Seq



O. Angelini et al., "Singing Synthesis: With A Little Help from My Attention", 2020.

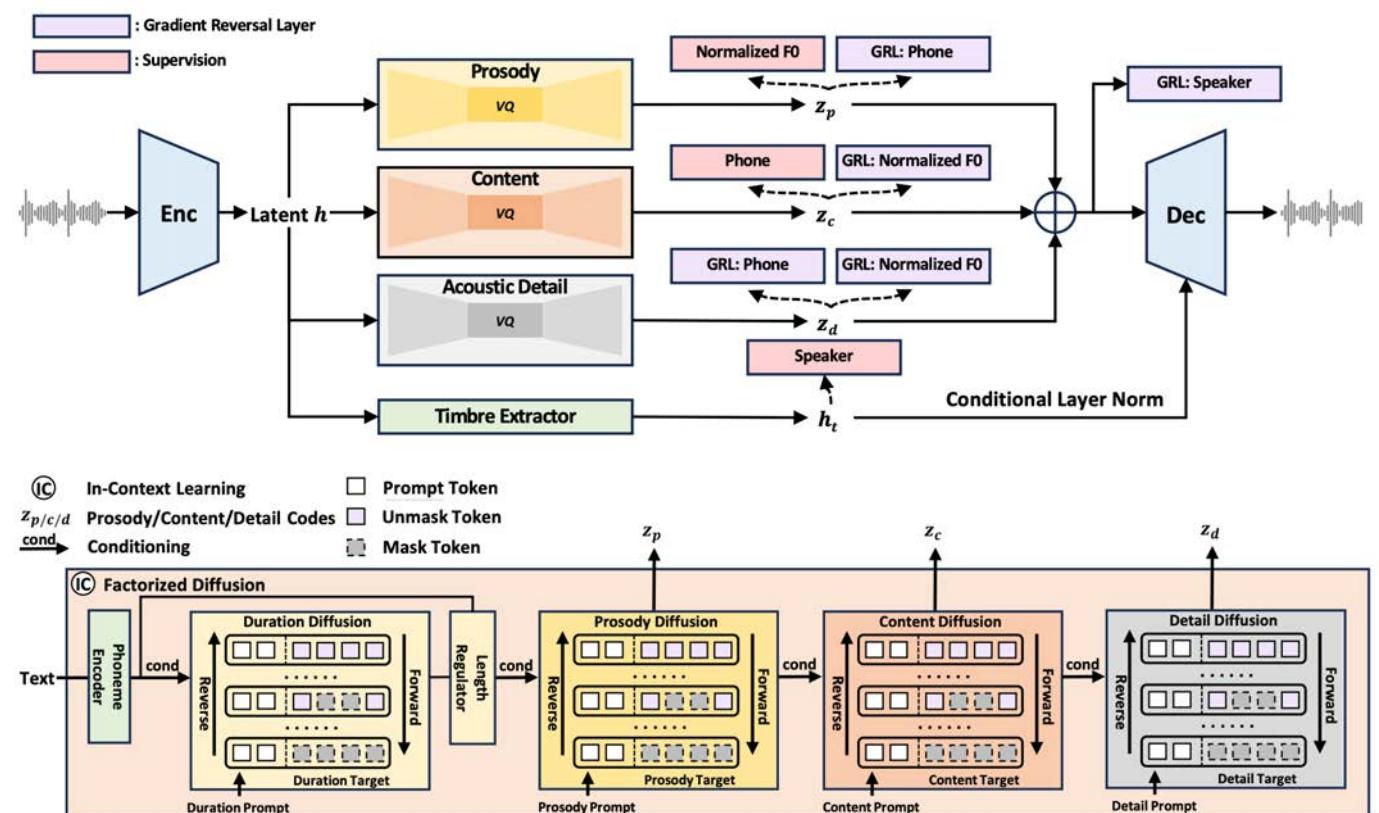
# Nature Speech 3

- An approach proposed by MSRA etc. six days ago



# Nature Speech 3

- FACodec: Factorises speech into prosody, content, acoustic detail and timbre
- Factorized diffusion: Four cascaded diffusion models to predict prosody/content/acoustic detail codes



2025/3/23

# OpenAI's Voice Engine

- OpenAI announced its first (actually third) speech project yesterday
- Promoted as to clone a person's voice with 15 second prompt
- With good disentanglement between language and speaker

reference    biology    reading



current    previous    generated

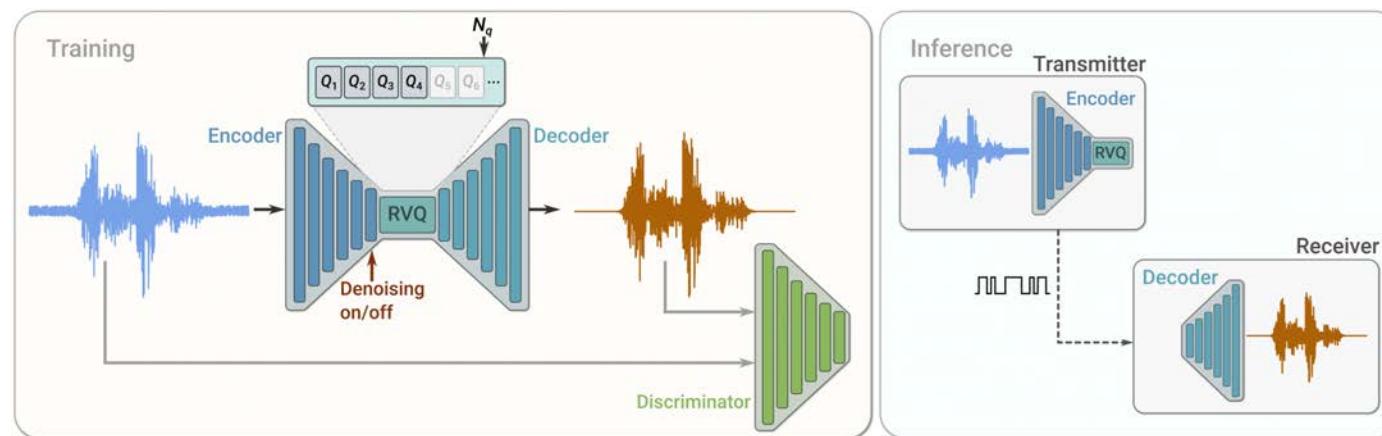


reference    Spanish    French    Japanese    Mandarin



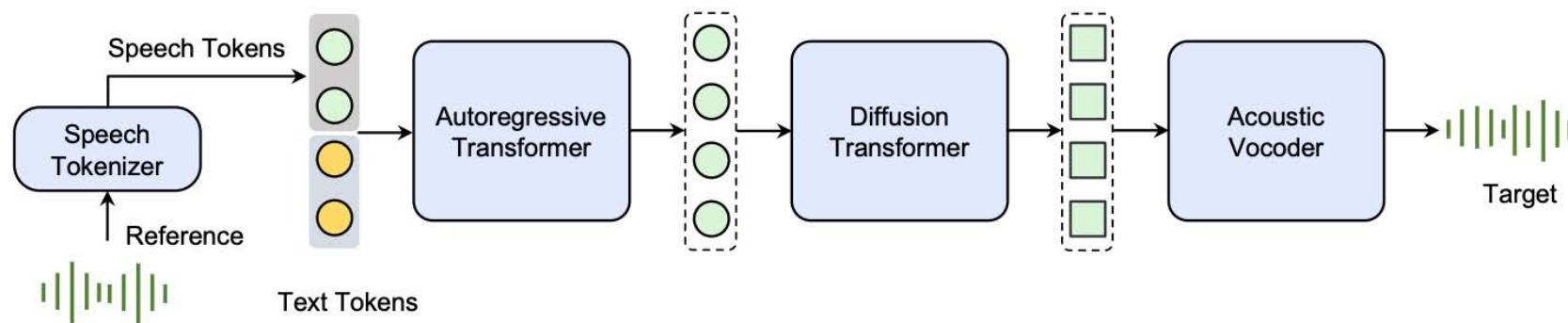
# Neural Audio Codec: SoundStream

- Developed for low-bitrate speech using residual vector quantization
  - Scalable codec with quantizer dropout
  - A codec-based vocoder released



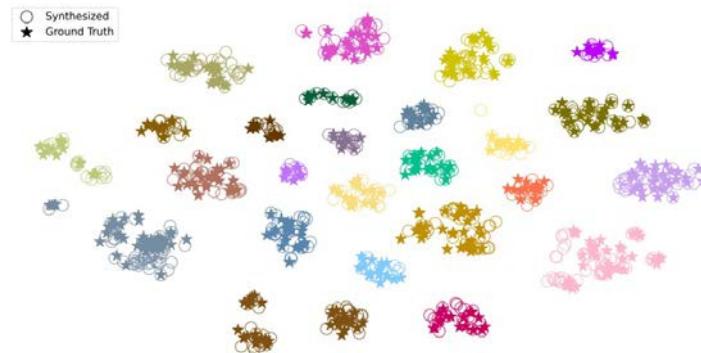
N. Zeghidour et al., "SoundStream: An End-to-End Neural Audio Codec", 2021.

# Seed-TTS



System	Lang.	Objective set		Subjective set CMOS ( $\uparrow$ ) vs. Human
		WER ( $\downarrow$ )	SIM ( $\uparrow$ )	
Seed-TTS	EN	2.249	<b>0.762</b>	-0.07
Vocoder resynthesized	EN	2.165	0.702	-
Human	EN	<b>2.143</b>	0.730	-
Seed-TTS	ZH	<b>1.115</b>	<b>0.796</b>	-0.08
Vocoder resynthesized	ZH	1.342	0.733	-
Human	ZH	1.254	0.750	-

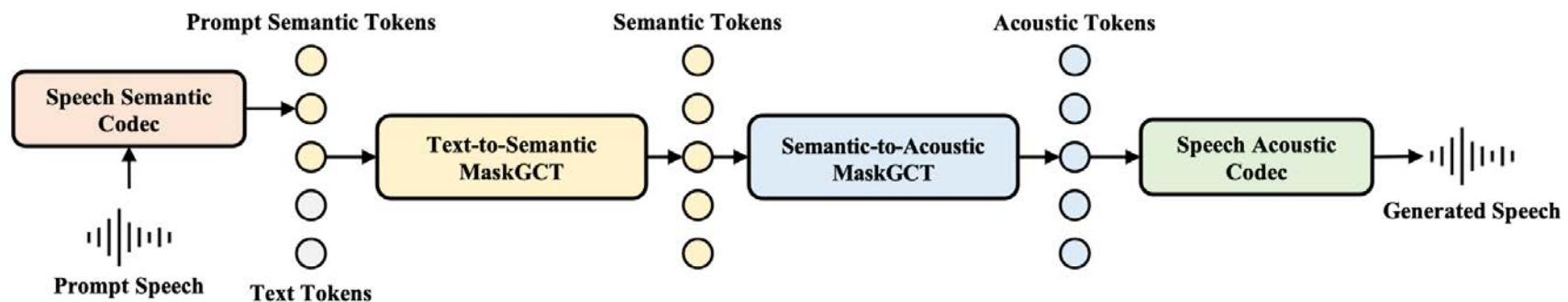
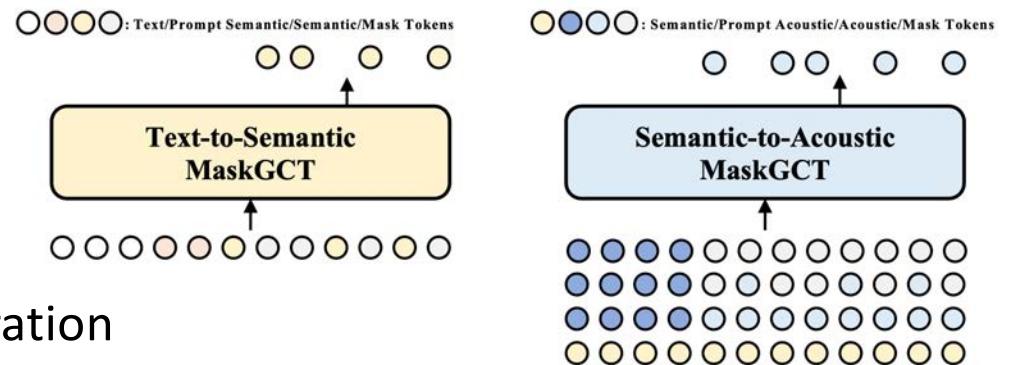
**Table 1:** Evaluation results of Seed-TTS against resynthesized and real human speech.



**Figure 3.** t-SNE visualization of speaker embeddings from the VoxCeleb1 test set (25 speakers) on synthesized and ground truth speech.

# Mask GCT

- Inspired by MaskGIT from image generation
- Generating speech codec in a non-autoregressive way, masking low-quality codecs and iteratively regenerating the speech codecs

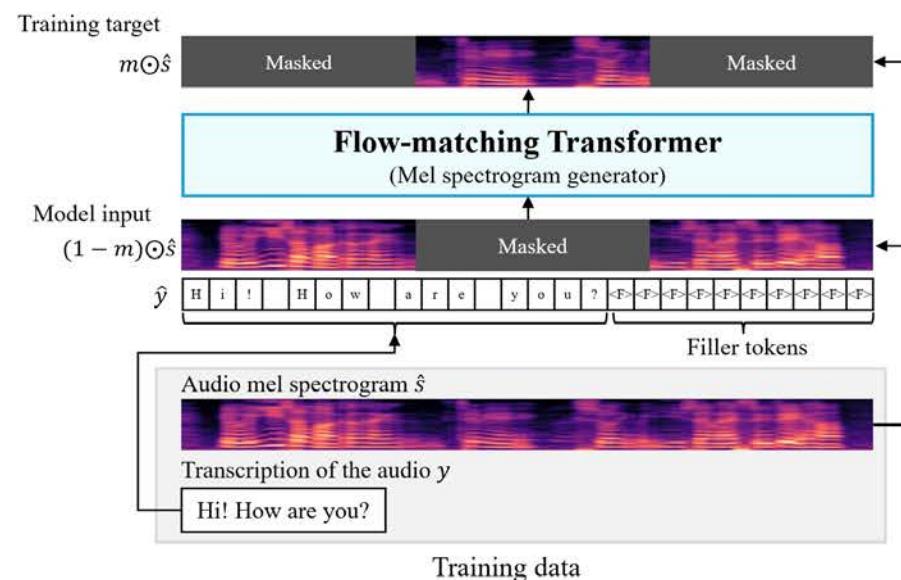


# Mask GCT

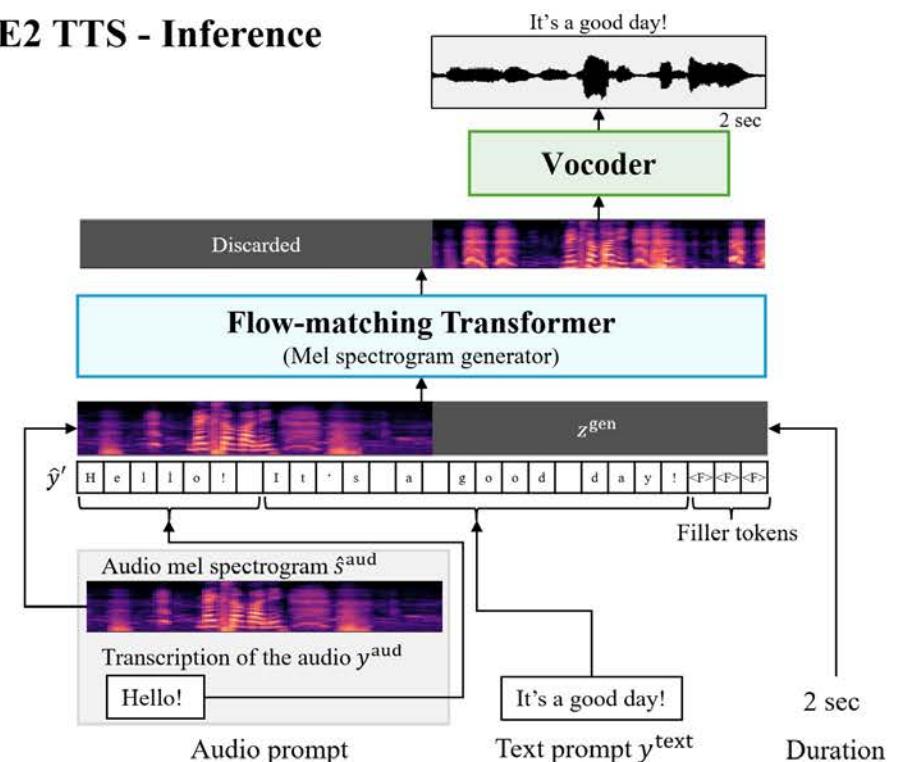


# E2 TTS

## E2 TTS - Training



## E2 TTS - Inference

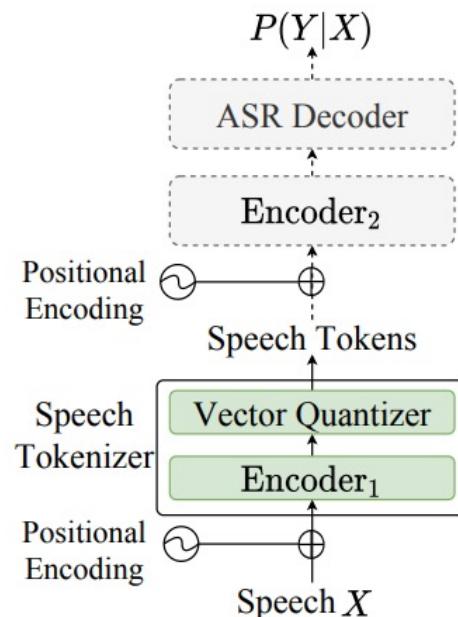


I enjoyed a day in **Besiktas, Istanbul.**

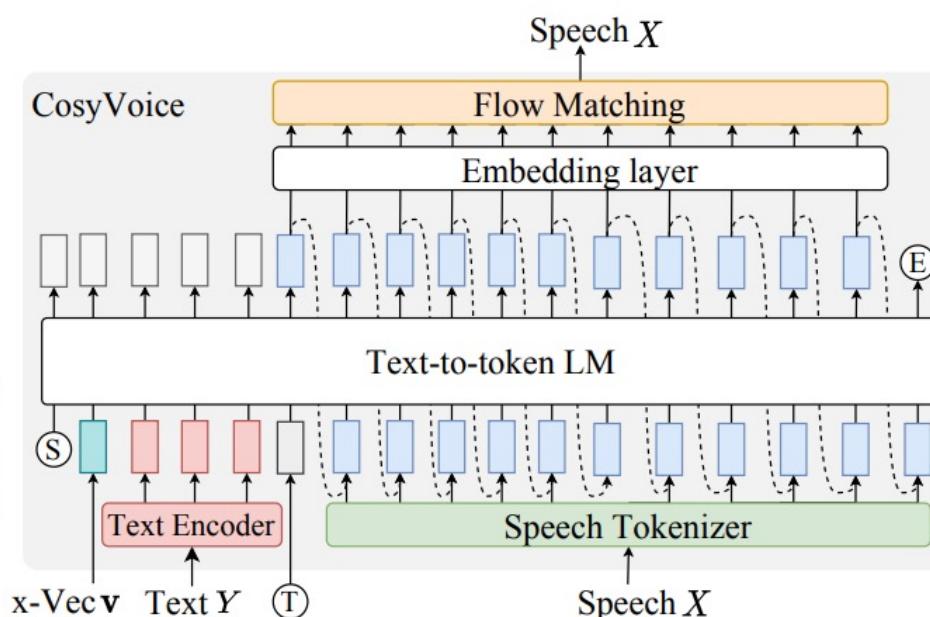
I enjoyed a day in (**B EH1 SH IH0 K T AA0 SH**), (**IH0 S T AA1 N B UH0 L**).

# CosyVoice

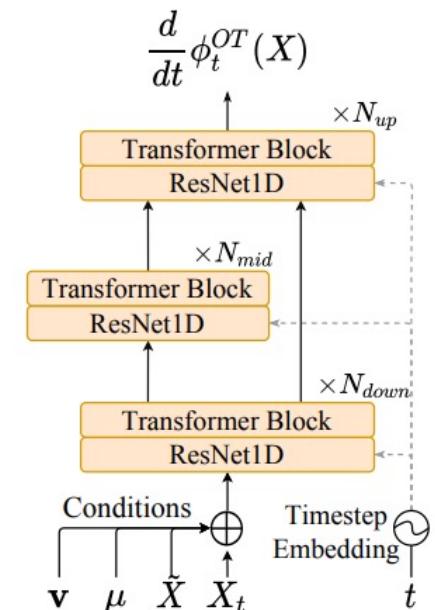
- 170k hours of speech data, 300M model parameters



(a) Supervised speech tokenizer



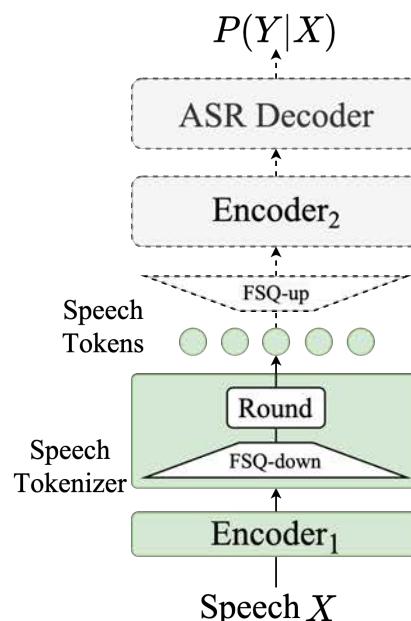
(b) An overview of the proposed CosyVoice LM



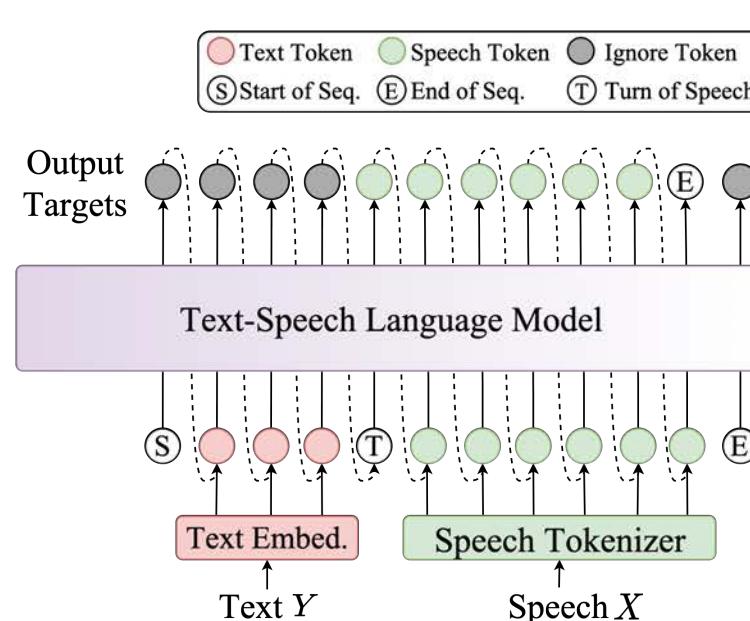
(c) Conditional flow matching

# CosyVoice 2

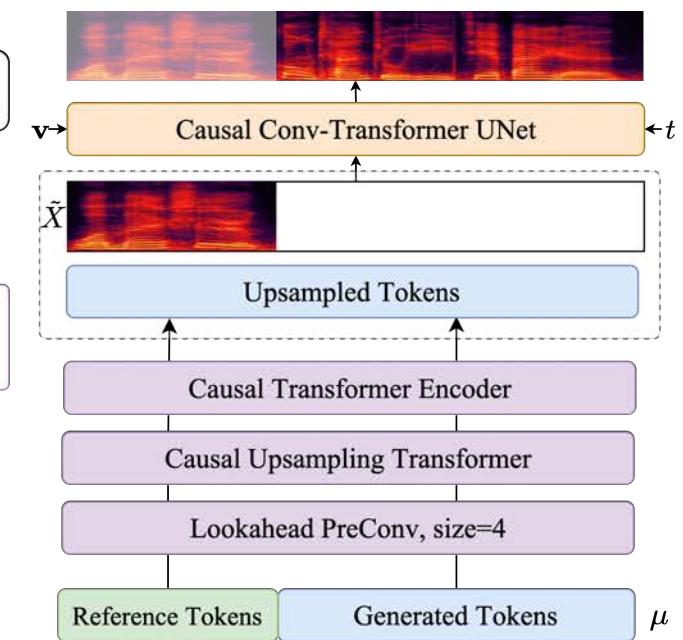
- 



(a) Supervised Speech Tokenizer



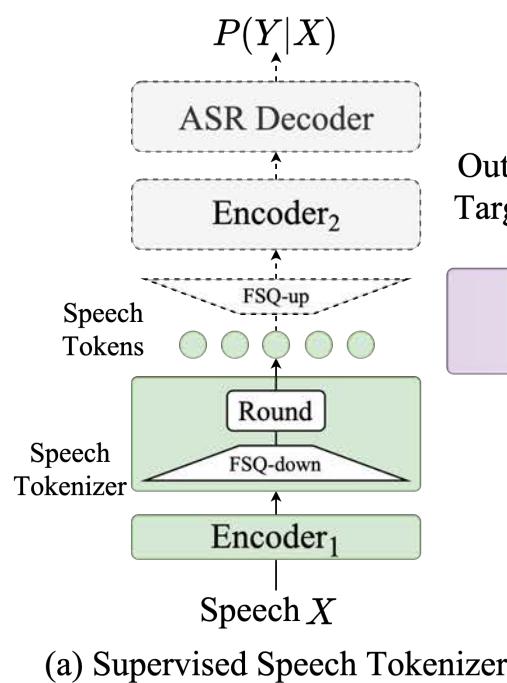
(b) Text-Speech Language Model



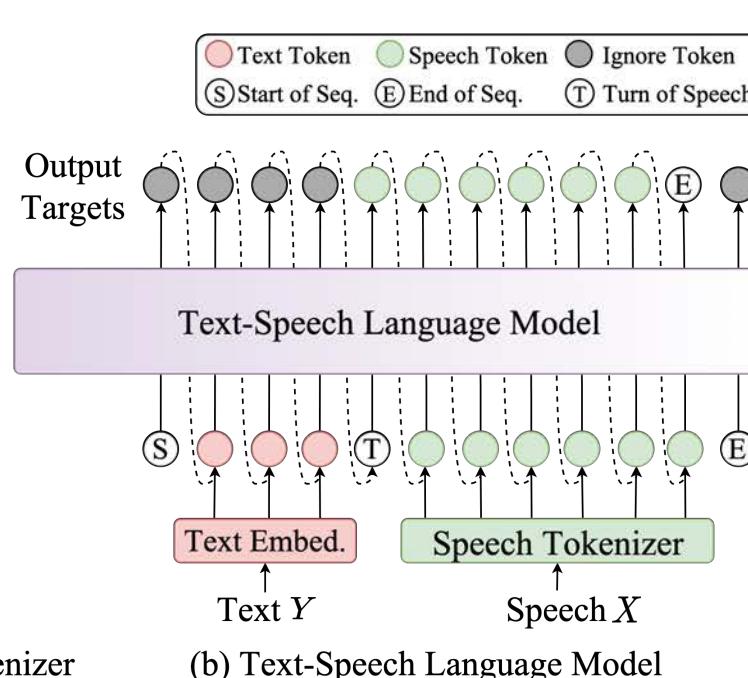
(c) Causal Flow Matching

# CosyVoice 2

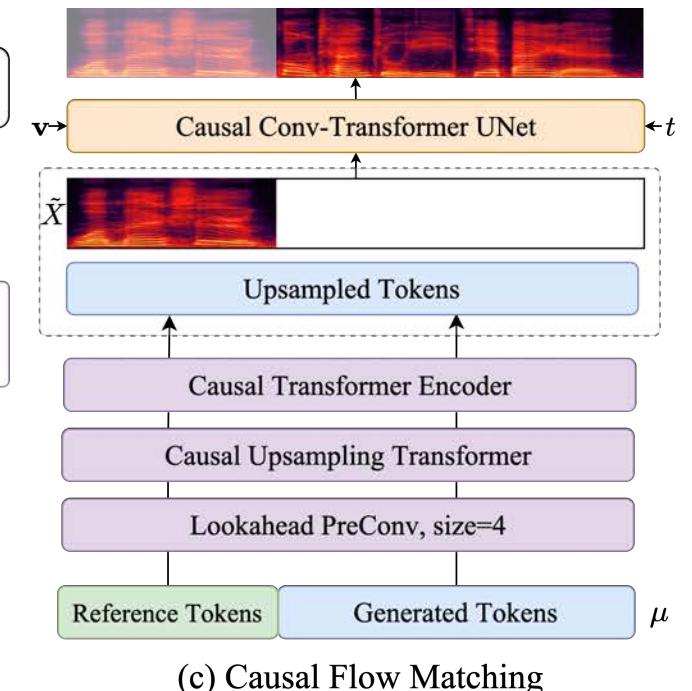
- With a 0.5B LLM to better understand text prompt



(a) Supervised Speech Tokenizer



### (b) Text-Speech Language Model



### (c) Causal Flow Matching

# CosyVoice 2



Speech  
Prompt



Generated Speech



Speech  
Prompt



Resting is important,  
but you shouldn't  
let your guard down  
too much.



休むことも大事だ  
けど、気を抜きすぎ  
てはダメよ。



어떤 것에 강한 감정을  
느낀다면, 목소리를 내고  
행동해야 합니다. 그것이  
제가 사는 철학입니다.



Speech  
Prompt



CosyVoice1



CosyVoice 2



Speech  
Prompt



CosyVoice1



CosyVoice 2

你昨天的 presentation がよかったです  
ので, 오늘도 좋은 피드백을 받을  
거예요。

# CosyVoice 2



黑化肥发灰，灰化肥发黑，  
黑化肥挥发会发黑，灰化肥挥发会发灰。化肥会挥发，  
灰化肥挥发发黑会挥发，黑化肥挥发发灰会挥发。



煢煢子立 沆瀣一氣 蹤蹠  
獨行 醍醐灌頂 綿綿瓜瓞  
奉為圭臬 龍行龕龕 犄角  
旮旯 嬌嬌嫋嫋 淚泗滂沱  
呶呶不休 不稂不莠。



我等不及等不及等不及等不及等去去去去玩水，  
我站站站站在沙上沙上沙上沙上的时候，脚快脚快脚快脚快烧起来了。



神秘<|endofprompt|>那座古老的城堡笼罩在神秘的雾气中，吸引着冒险者前去探索。



模仿机器人风格  
<|endofprompt|>在人工智能技术的支持下...



一个忧郁的诗人，言语中总是透露出一丝哀愁和浪漫。  
<|endofprompt|>月光下的一切都是那么宁静...



郑州话<|endofprompt|>  
这阵子我在听豫剧，虽然有些地方唱词听不太明白，但音乐一响，耳朵就被吸引了。



天津话<|endofprompt|>  
今儿个去逛古文化街，那些个手工艺品五花八门，特别是杨柳青年画，色彩鲜艳得让人眼花缭乱。

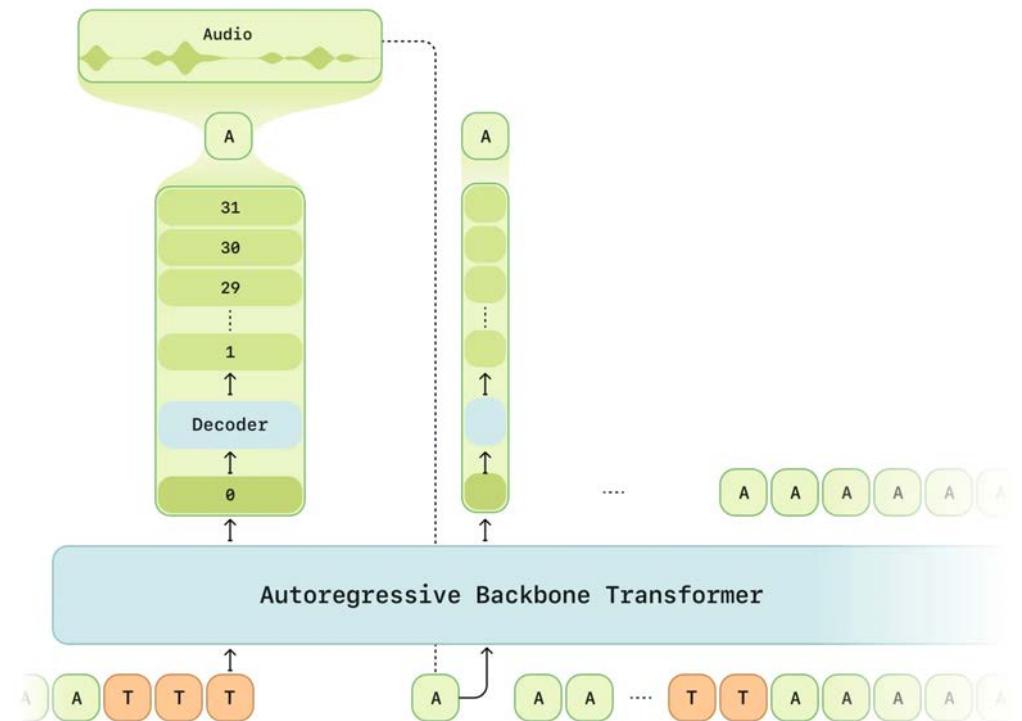
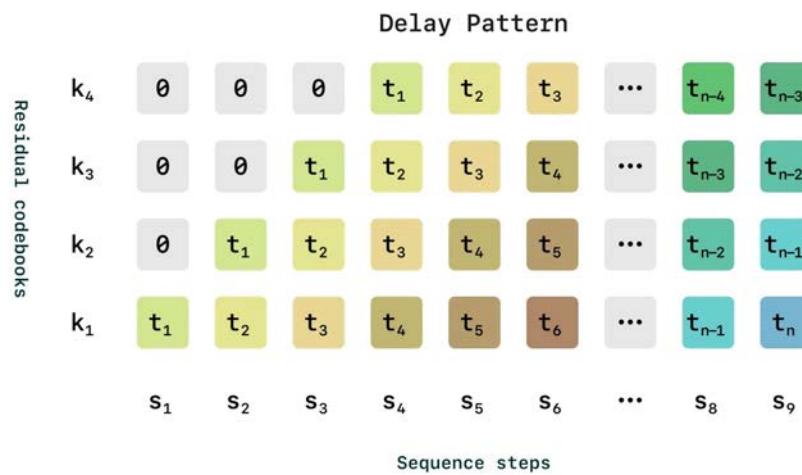


在他讲述那个荒诞故事的过程中，他突然[laughter]停下来，因为他自己也被逗笑了[laughter]。

# Sesame



Conversations with multiple speakers



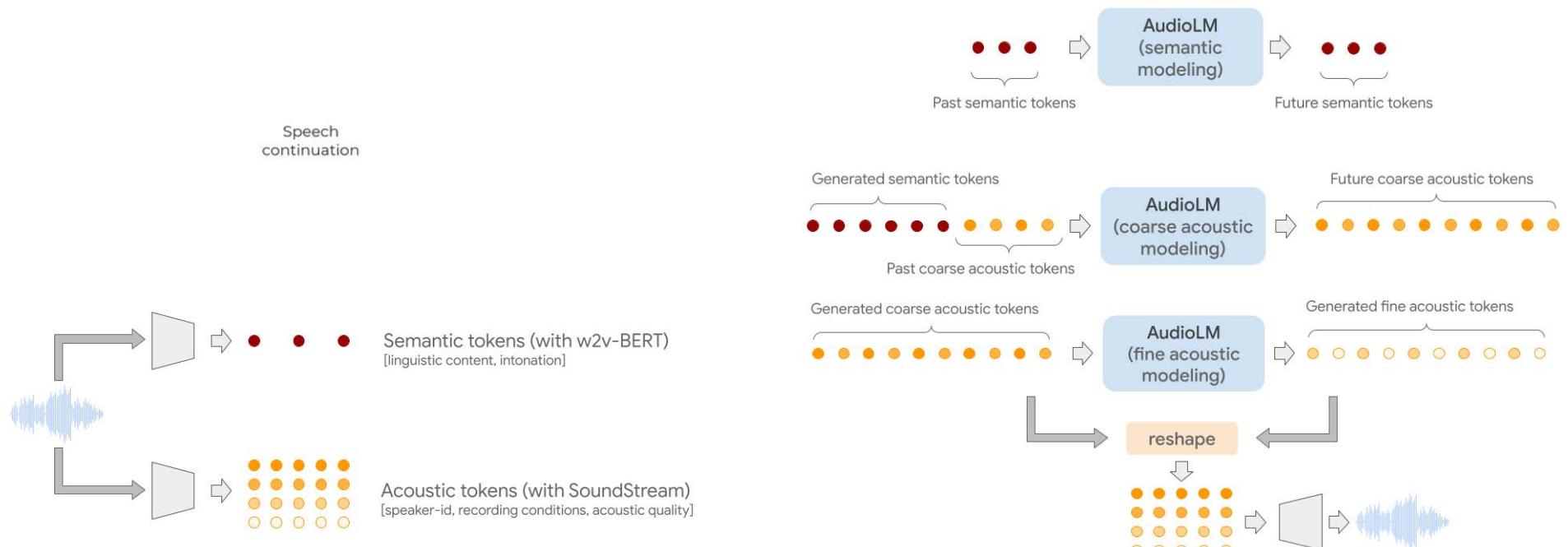
# Audio, Music, Co-Speech Gesture, Facial Expression Generations

# Neural Audio Codec

- SoundStream → AudioLM
- EnCodec → AudioGen, VALL-E
- DAC
- X-Codec → YuE

# Audio LM

- Audio language modelling using SoundStream codec

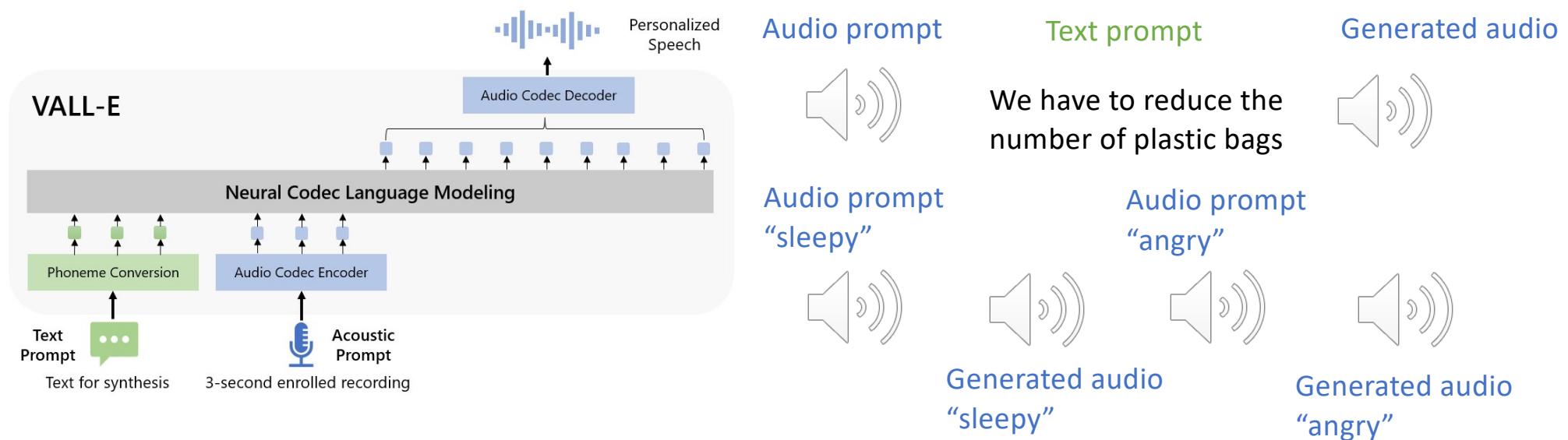


Z. Borsos et al., “AudioLM: A Language Model Approach to Audio Generation”, 2022.

清华大学-MSRA 《高等机器学习》

# VALL-E

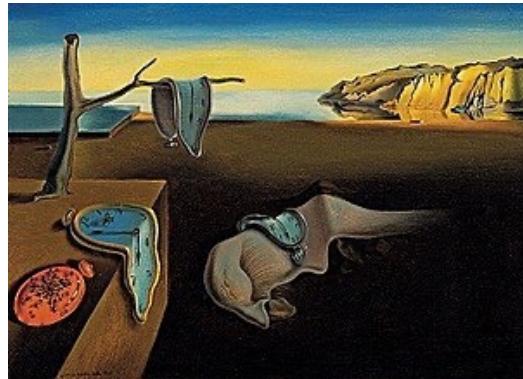
- A neural codec (by FAIR) language model with **audio** and **text** prompts



C. Wang et al., “VALL-E: Neural Codec Language Models are Zero-shot Text to Speech Synthesizers”, 2023.

# Music LM

- A model for generating high-fidelity music from text descriptions



达利 “记忆的永恒”

他融化的时钟图像嘲笑计时时间的僵硬。手表本身看起来像软奶酪——事实上，根据 Dali 自己的说法，他们的灵感来自于吃了卡门培尔奶酪后的幻觉。在图片的中央，在其中一只手表的下面，是一张扭曲的人脸。盘子里的蚂蚁代表腐烂。



路易大卫 “拿破仑穿越阿尔卑斯”

这幅作品展示了拿破仑和他的军队于 1800 年 5 月通过大圣伯纳德山口穿越阿尔卑斯山的真实穿越的强烈理想化景象。



A. Agostinelli et al., "MusicLM: Generating Music from Text", 2023.

清华大学-MSRA 《高等机器学习》

2025/3/23



街机游戏的主要配乐。它节奏快且乐观，带有朗朗上口的电吉他即兴重复段。音乐是重复的，容易记住，但有意想不到的声音，如铙钹撞击声或鼓声。



带有长笛和吉他的冥想歌曲，平静而舒缓。音乐很慢，重点是营造一种和平与安宁的感觉。

使用文本或  
音频提示：



Instrument



guitar solo

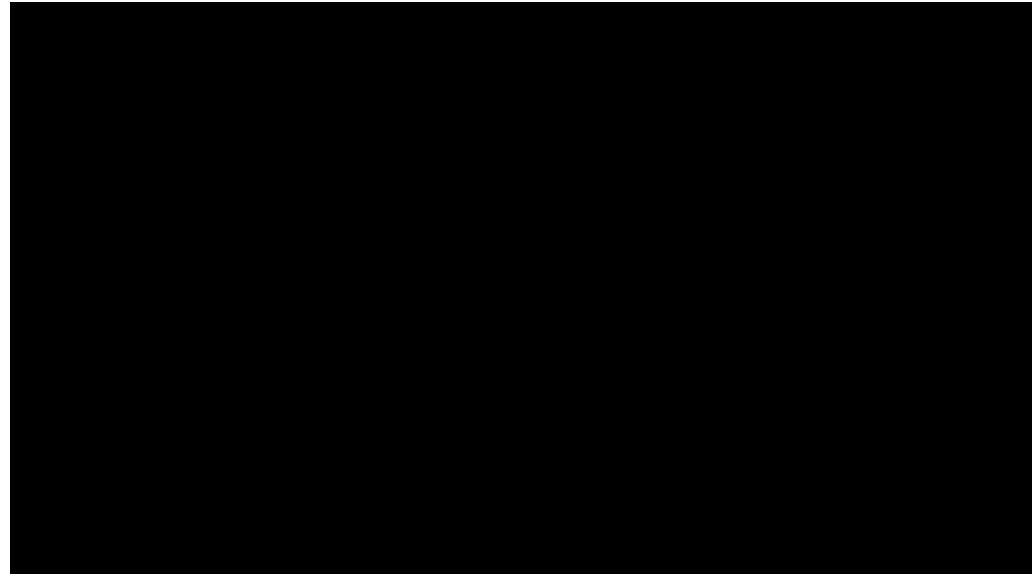
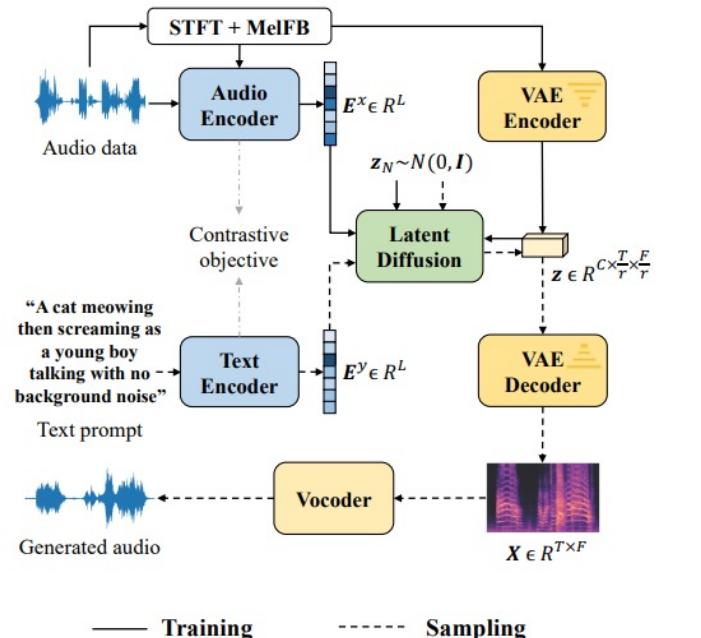


ode to joy  
humming<sup>47</sup>

bella ciao -  
humming

# AudioLDM & AudioLDM 2

- Text-to-audio in contrastive language-audio pretraining latent space



H. Liu, et al., "AudioLDM: Text-to-audio Generation with Latent Diffusion Models", 2023.

# Suno V3

- <http://www.suno.ai>
- The first AI that can generate broadcast level music.



# Suno V3

- Let's combine GPT-4 with Suno V3...

ChatGPT 4 ▾



You

你好，我是清华大学电子工程系的一名教师，我一会儿要给研究生上一门和微软亚洲研究院合开的《高等机器学习课》，我想给这门课作一首说唱风的歌曲，请你给这首歌创作歌词



ChatGPT

你好！这个想法真是太酷了，我很高兴能帮助你创作一首与《高等机器学习》相关的说唱歌词。让我们开始吧！

---

(Verse 1)

从清华到微软的桥梁，我们跨界合作，风起云涌，  
电子工程的精英，机器学习的探索，始于好奇心的涌动。  
在代码的海洋，我们航行，算法是我们的罗盘，方向，  
深度学习的奥秘，等待我们揭晓，知识的边界，正待我们拓宽。

(Chorus)

这是高等机器学习，不仅仅是理论的学习，  
在实验中找答案，用数据说话，证明我们的观点。  
清华和微软，携手共创未来，  
让机器智能，照亮人类的明天，这是我们的使命和愿望。

(Verse 2)

支持向量机，随机森林，神经网络跳动的节奏，  
每一个模型，都藏着无限可能，等你我去探索。  
误差反向传播，优化算法的步伐，不停地迭代，  
直到找到最佳，这是科学也是艺术，追求的境界。

(Bridge)

在数据的世界里，我们是探险家，  
每个问题，都是未知的挑战，激发我们的灵感。  
高维空间的旅行，寻找隐藏的规律，  
机器学习的魅力，是无穷的，让人着迷。

(Chorus)

这是高等机器学习，让复杂的问题变得简单，  
在挑战中成长，用智慧开启新的篇章。  
清华和微软，共同书写历史，  
用技术的力量，造福人类，这是我们的梦想和希望。

希望这首歌能够为你的课程增添活力和灵感，激发学生们的热情！

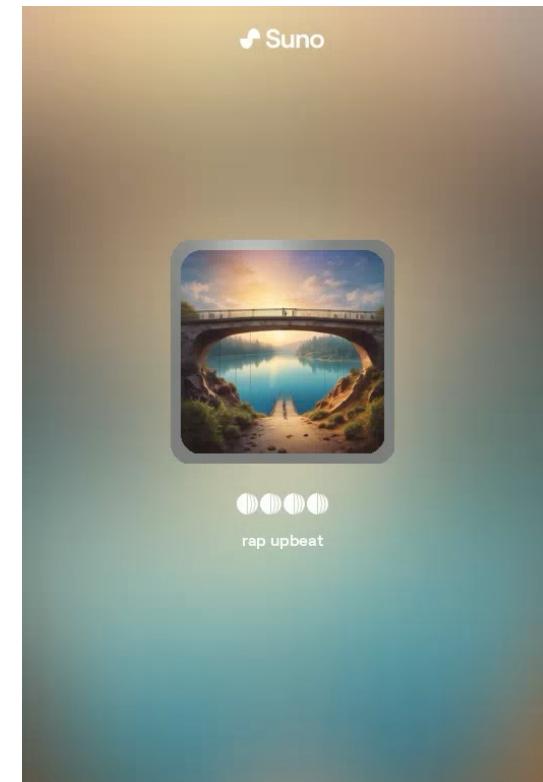
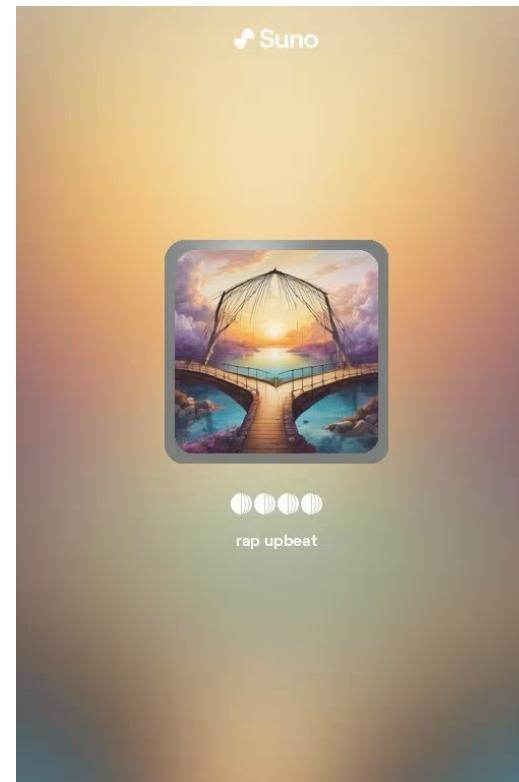
# Suno V3

(Verse 1)

从清华到微软的桥梁，我们跨界合作，风起云涌，  
电子工程的精英，机器学习的探索，始于好奇心的涌动。  
在代码的海洋，我们航行，算法是我们的罗盘，方向，  
深度学习的奥秘，等待我们揭晓，知识的边界，正待我们拓宽。

(Chorus)

这是高等机器学习，不仅仅是理论的学习，  
在实验中找答案，用数据说话，证明我们的观点。  
清华和微软，携手共创未来，  
让机器智能，照亮人类的明天，这是我们的使命和愿望。



# Suno V3

(Verse 2)

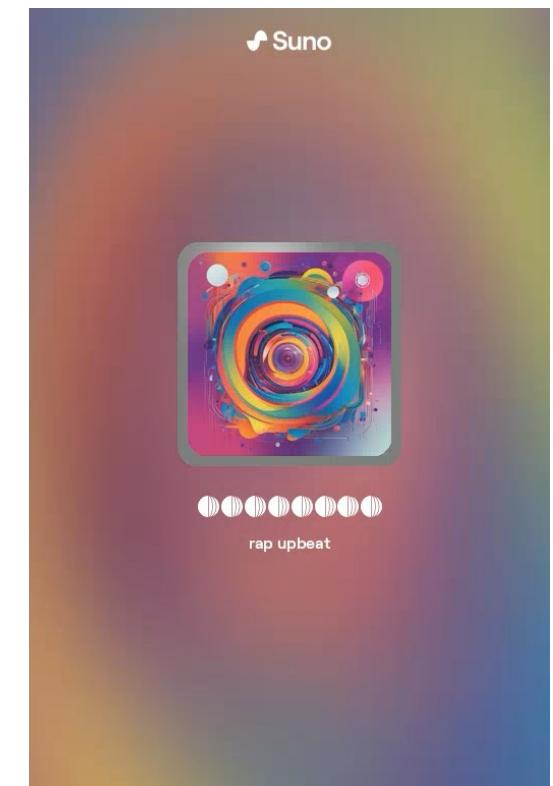
支持向量机，随机森林，神经网络跳动的节奏，  
每一个模型，都藏着无限可能，等你我去探索。  
误差反向传播，优化算法的步伐，不停地迭代，  
直到找到最佳，这是科学也是艺术，追求的境界。

(Bridge)

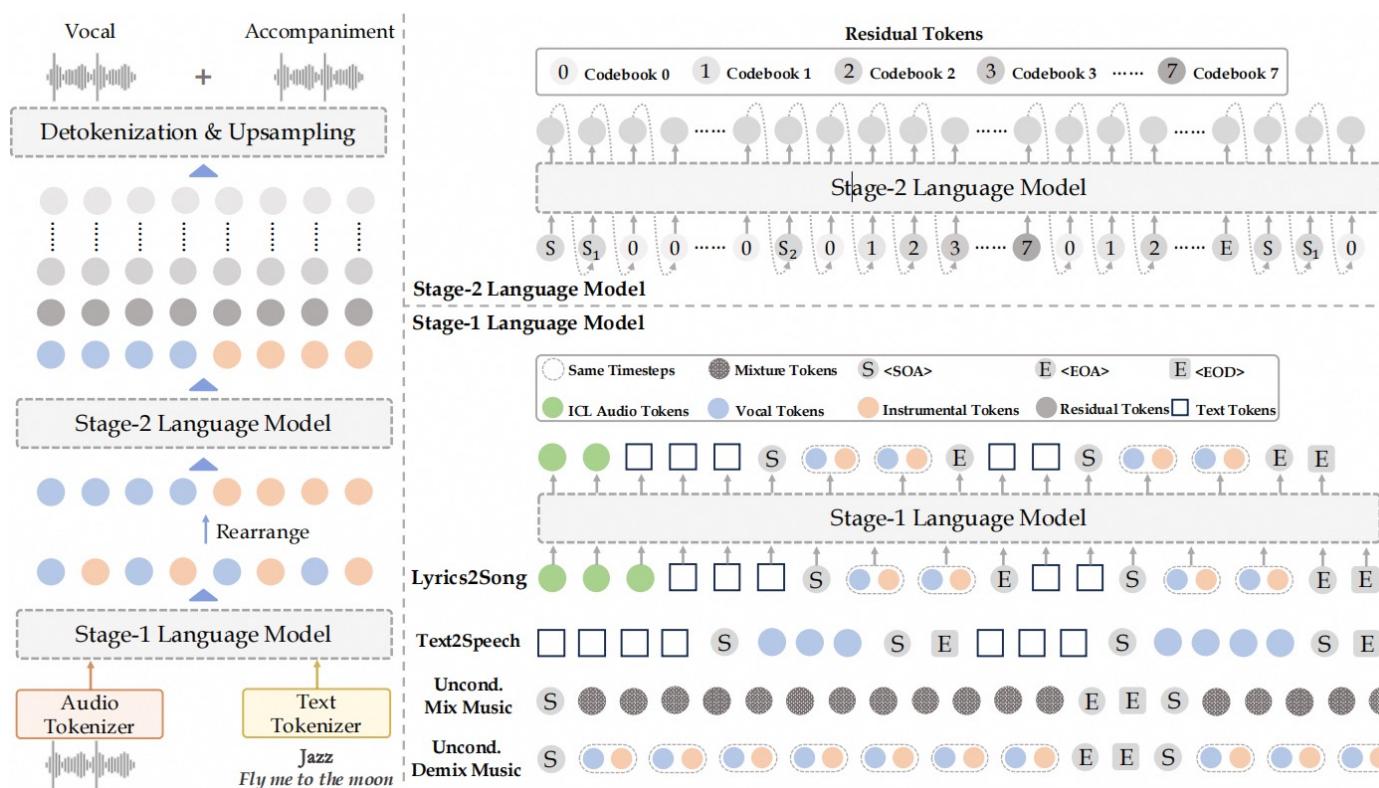
在数据的世界里，我们是探险家，  
每个问题，都是未知的挑战，激发我们的灵感。  
高维空间的旅行，寻找隐藏的规律，  
机器学习的魅力，是无穷的，让人着迷。

(Chorus)

这是高等机器学习，让复杂的问题变得简单，  
在挑战中成长，用智慧开启新的篇章。  
清华和微软，共同书写历史，  
用技术的力量，造福人类，这是我们的梦想和希望。



# YuE



# Co-Speech Gesture & Facial Expression

Speech-to-Gesture (ground-truth mouth facial keypoints)

Mark Kubinec  
(chemistry lecture)

DisCo: Co-Speech Gesture Generation



S. Ginosar et al., "Learning Individual Styles of Conversational Gesture", 2019.

H. Liu et al., "Disentangled Implicit Content and Rhythm Learning for Diverse Co-speech Gesture Synthesis", 2022.

Thanks for your listening!