表现力语音合成可控性研究

刘 瑞

2022-12-27

01 语音合成

02 语音情感/口音表达

03 情感可控语音合成

04 口音可控语音合成

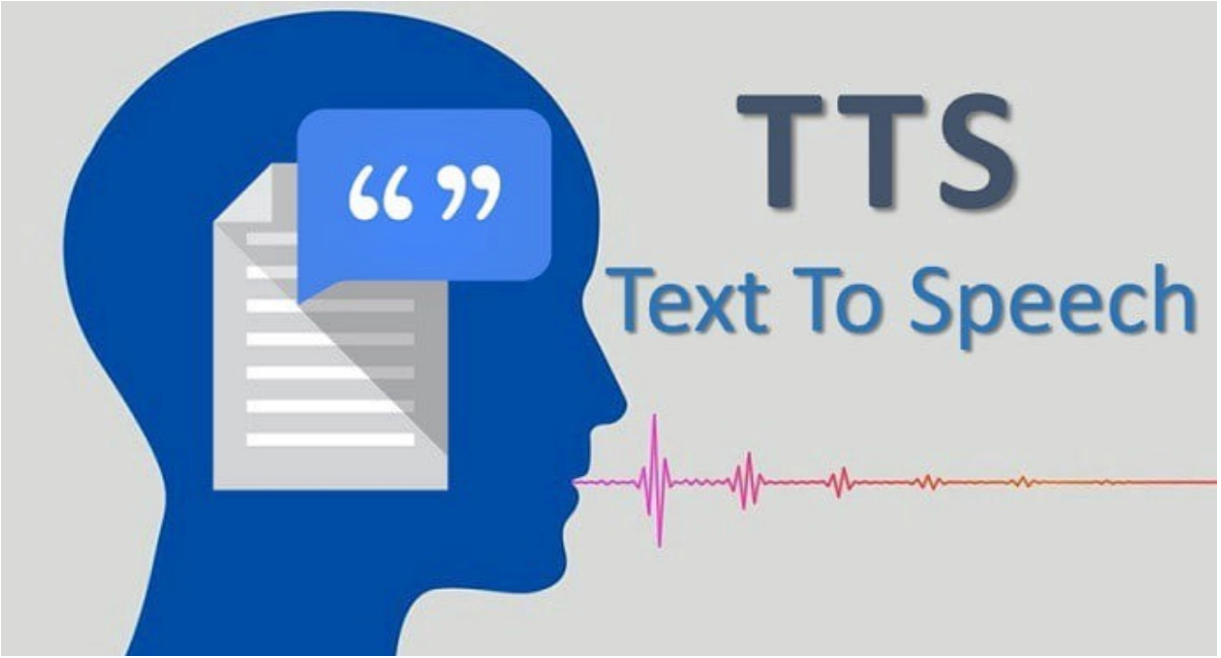
01 语音合成

02 语音情感/口音表达

03 情感可控语音合成

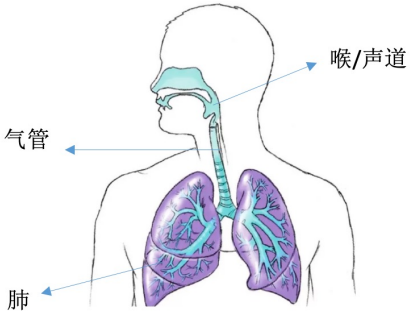
04 口音可控语音合成

语音合成

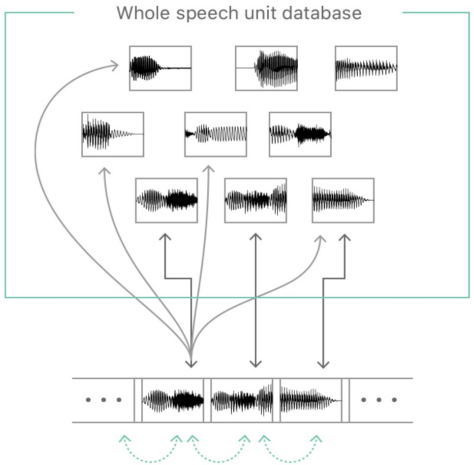


语音合成

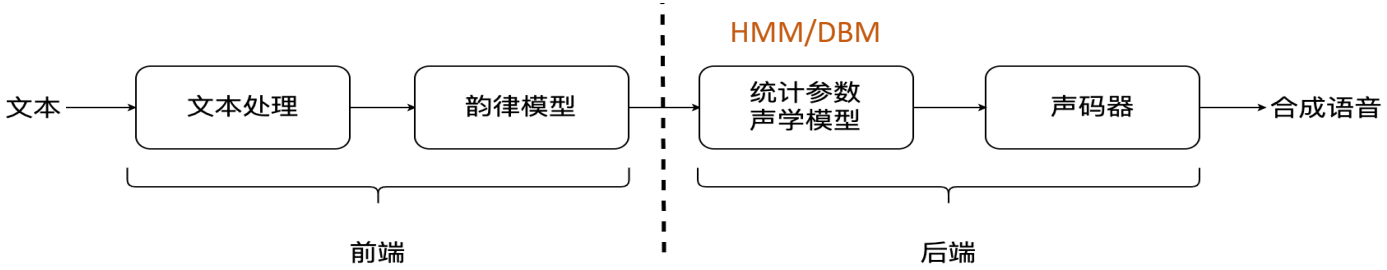
- 机械式语音合成方法
- 电子式语音合成方法



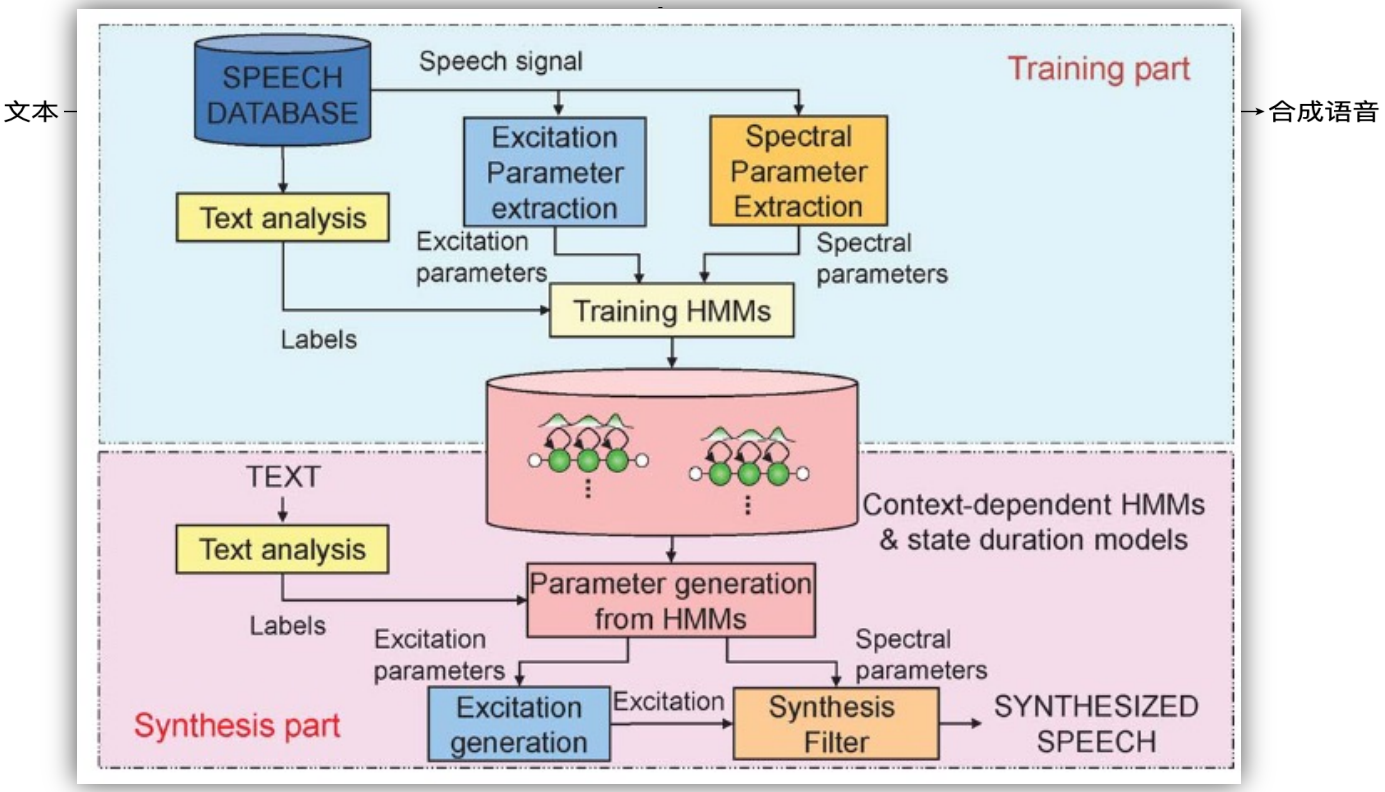
- 波形拼接（单元挑选）语音合成方法



统计参数语音合成

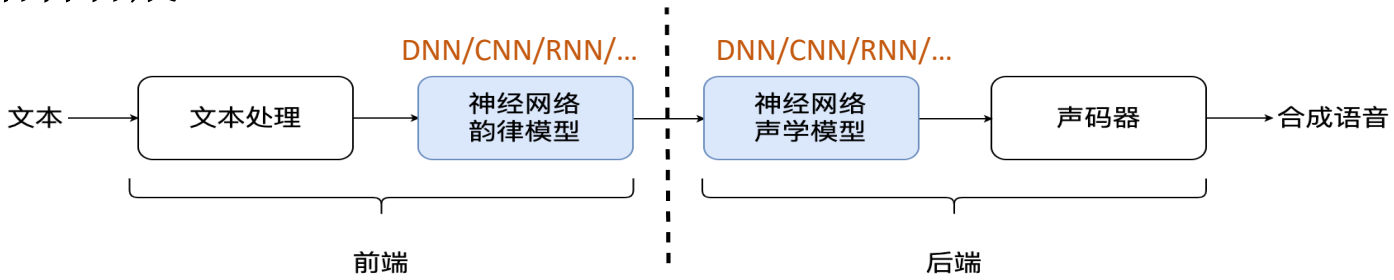


统计参数语音合成

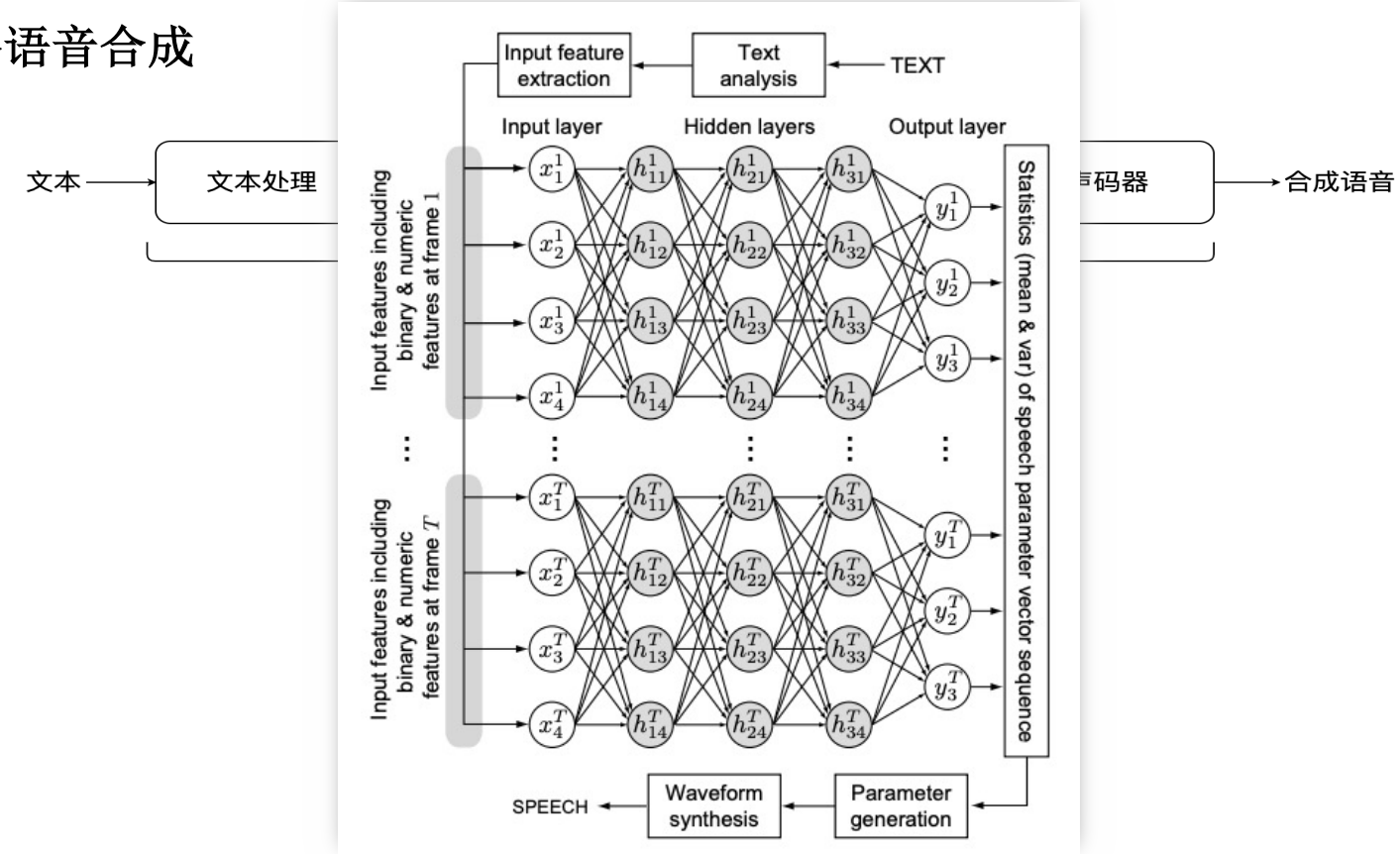


K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi and K. Oura, "Speech Synthesis Based on Hidden Markov Models," in Proceedings of the IEEE, vol. 101, no. 5, pp. 1234-1252, May 2013, doi: 10.1109/JPROC.2013.2251852.

神经网络语音合成

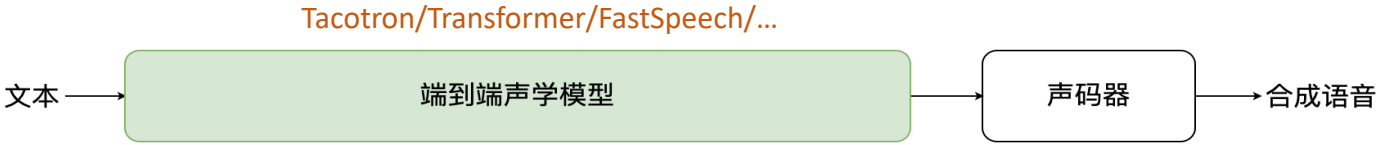

DNN/CNN/RNN/...
DNN/CNN/RNN/...
文本
文本处理
神经网络
韵律模型
神经网络
声学模型
声码器
合成语音
前端
后端

神经网络语音合成

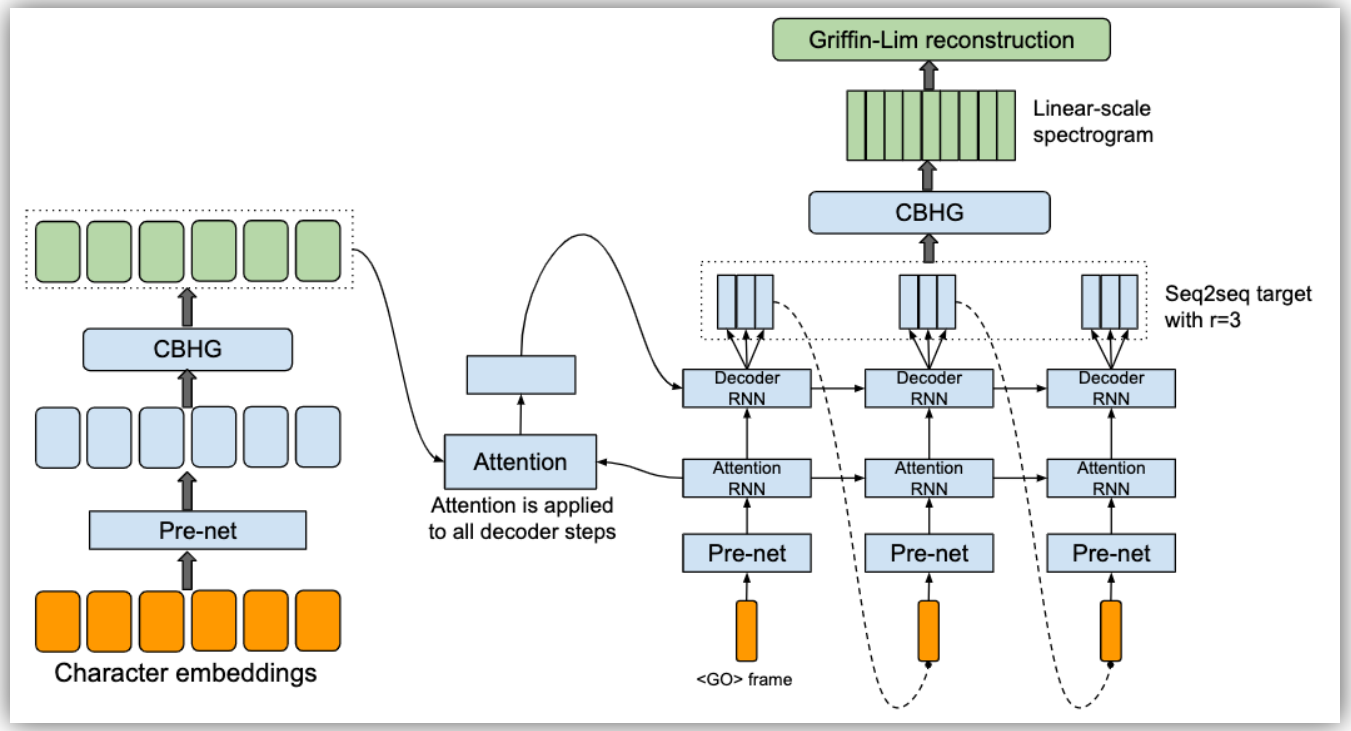


H. Ze, A. Senior and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 7962-7966, doi: 10.1109/ICASSP.2013.6639215.

深度学习语音合成（端到端语音合成）

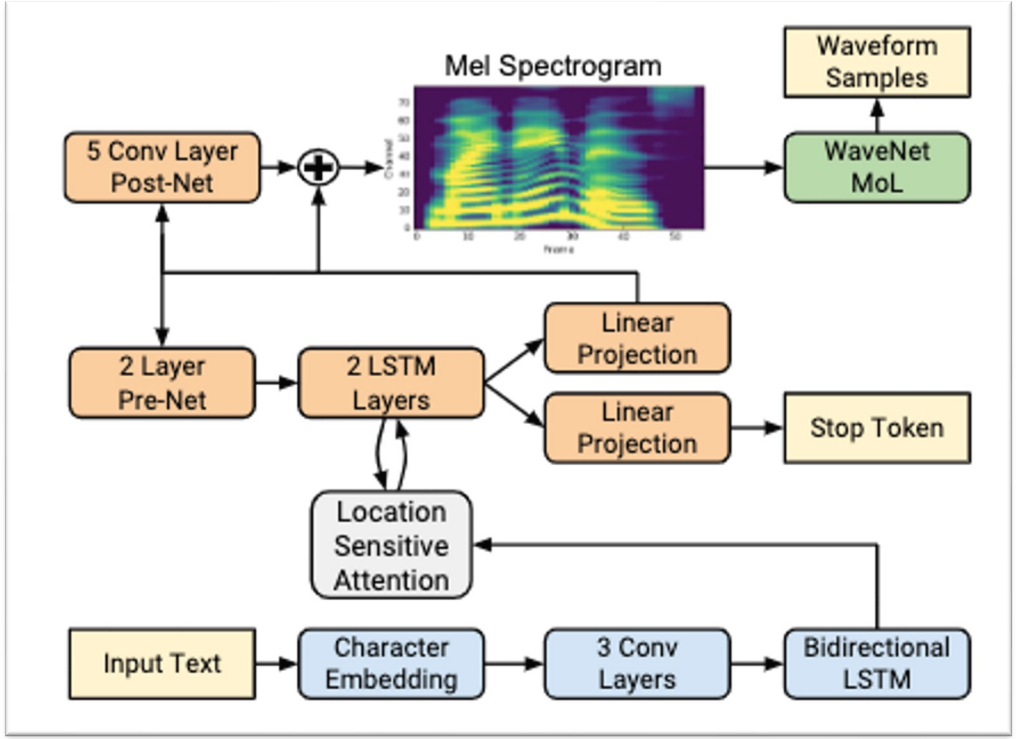


深度学习语音合成（端到端语音合成）



Wang, Yuxuan, et al. "Tacotron: Towards End-to-End Speech Synthesis." *Proc. Interspeech 2017* (2017): 4006-4010.

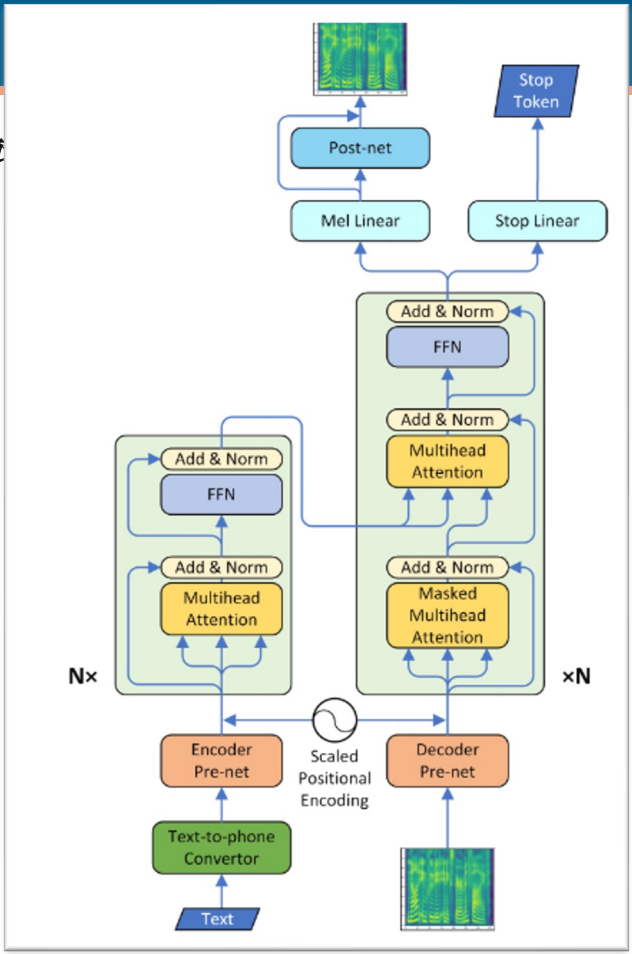
深度学习语音合成（端到端语音合成）



Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018.

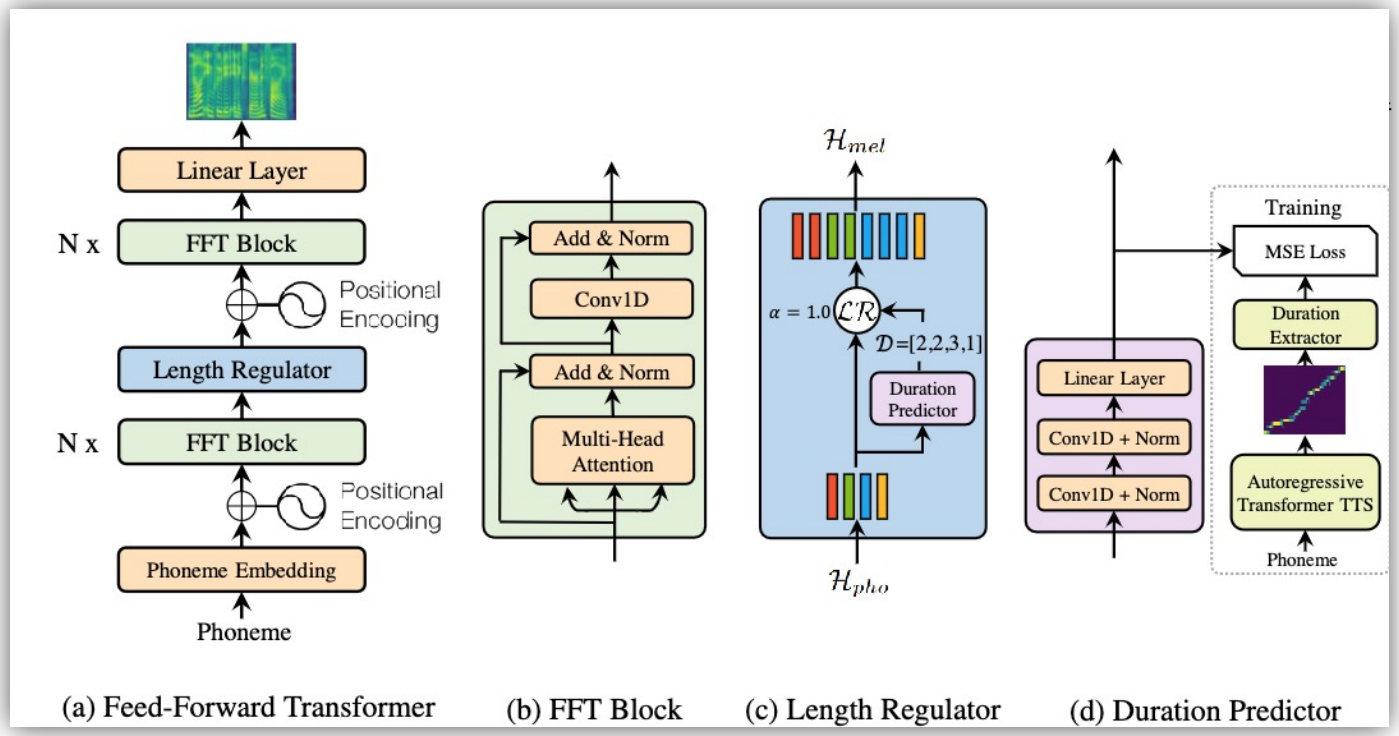
语音合成

深度学习语音合成（端到端



Li, Naihan, et al. "Neural speech synthesis with transformer network." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 2019.

深度学习语音合成（端到端语音合成）



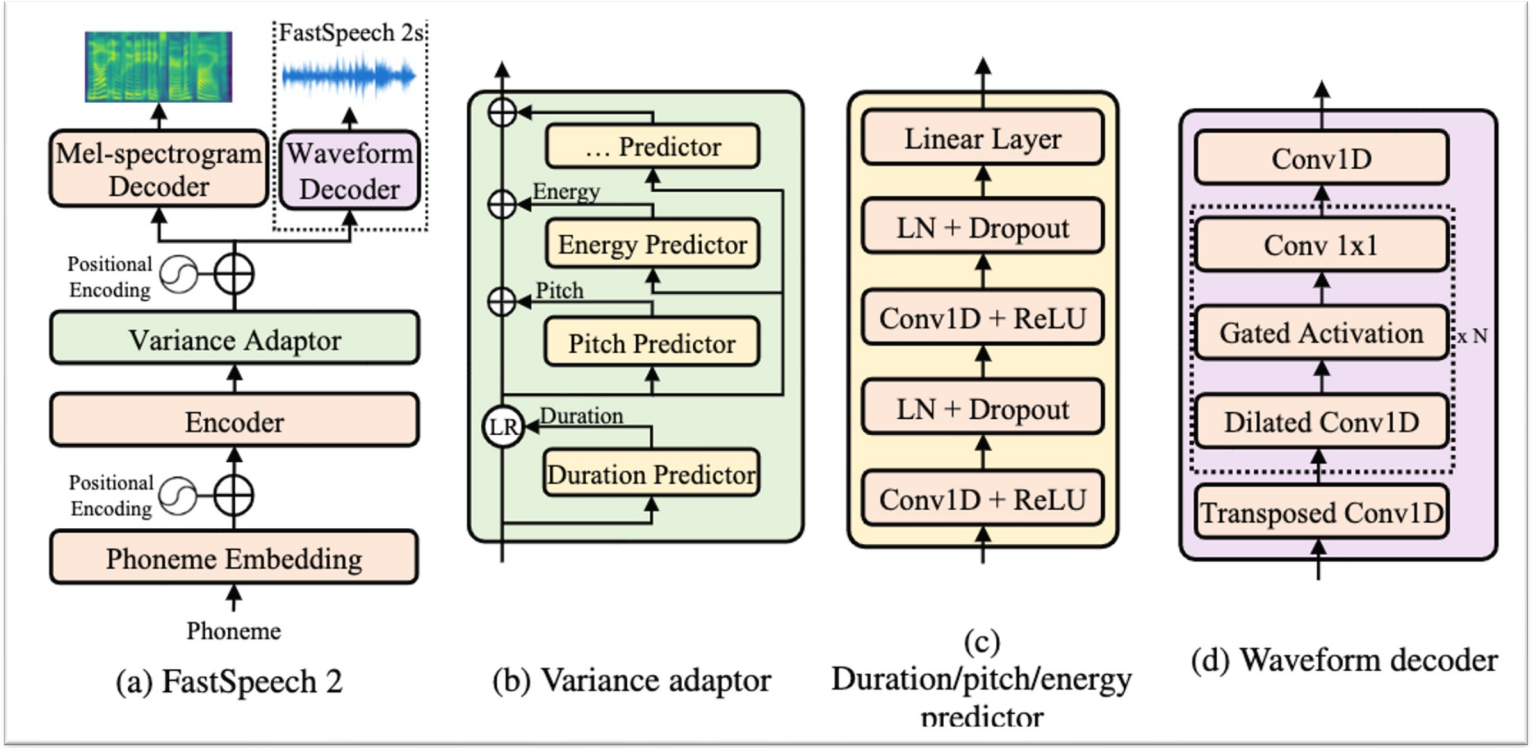
Ren, Yi, et al. "Fastspeech: Fast, robust and controllable text to speech." *Advances in Neural Information Processing Systems* 32 (2019).

深度学习语音合成（端到端语音合成）

FastSpeech2s/ViTS/NaturalSpeech/...

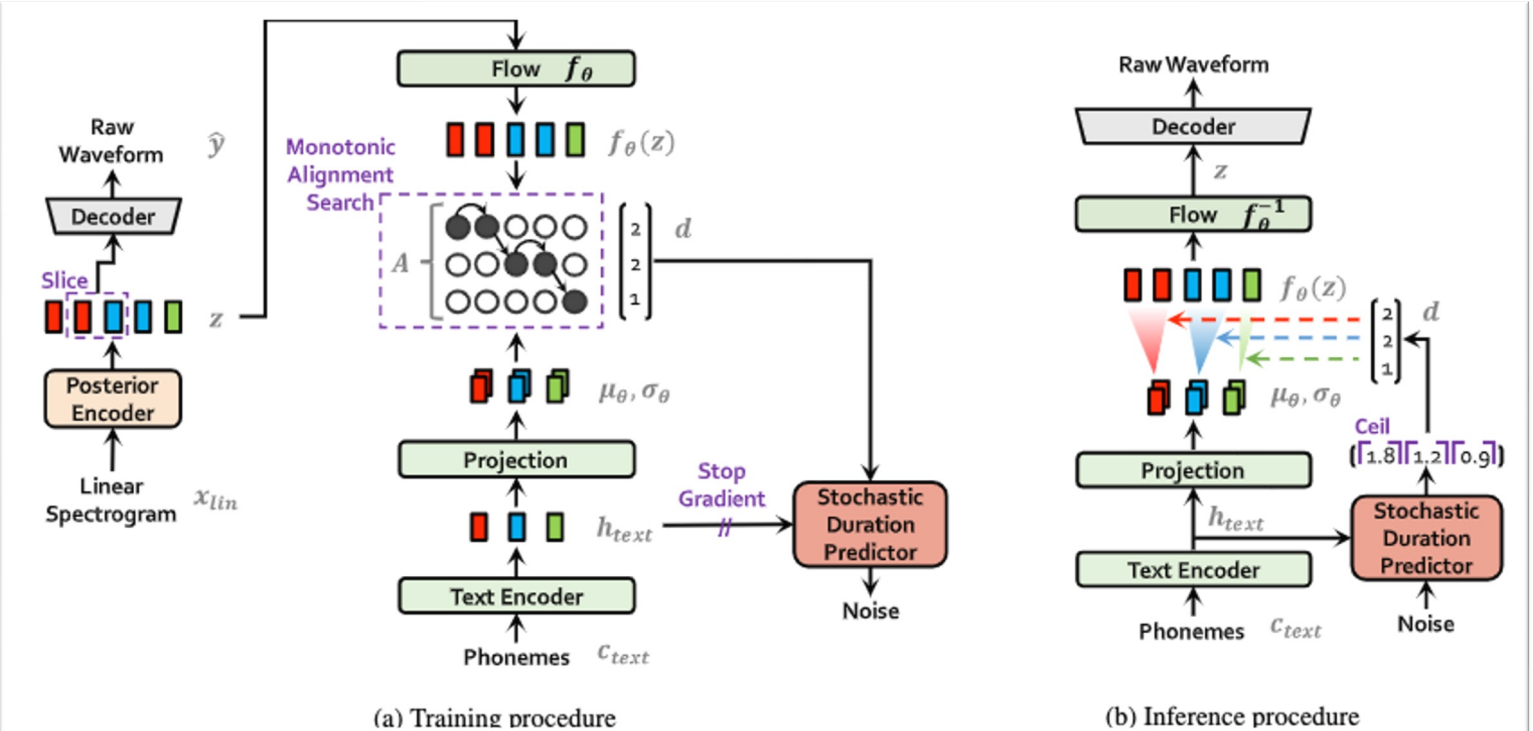
文本 → 端到端声学模型 → 合成语音

深度学习语音合成（端到端语音合成）



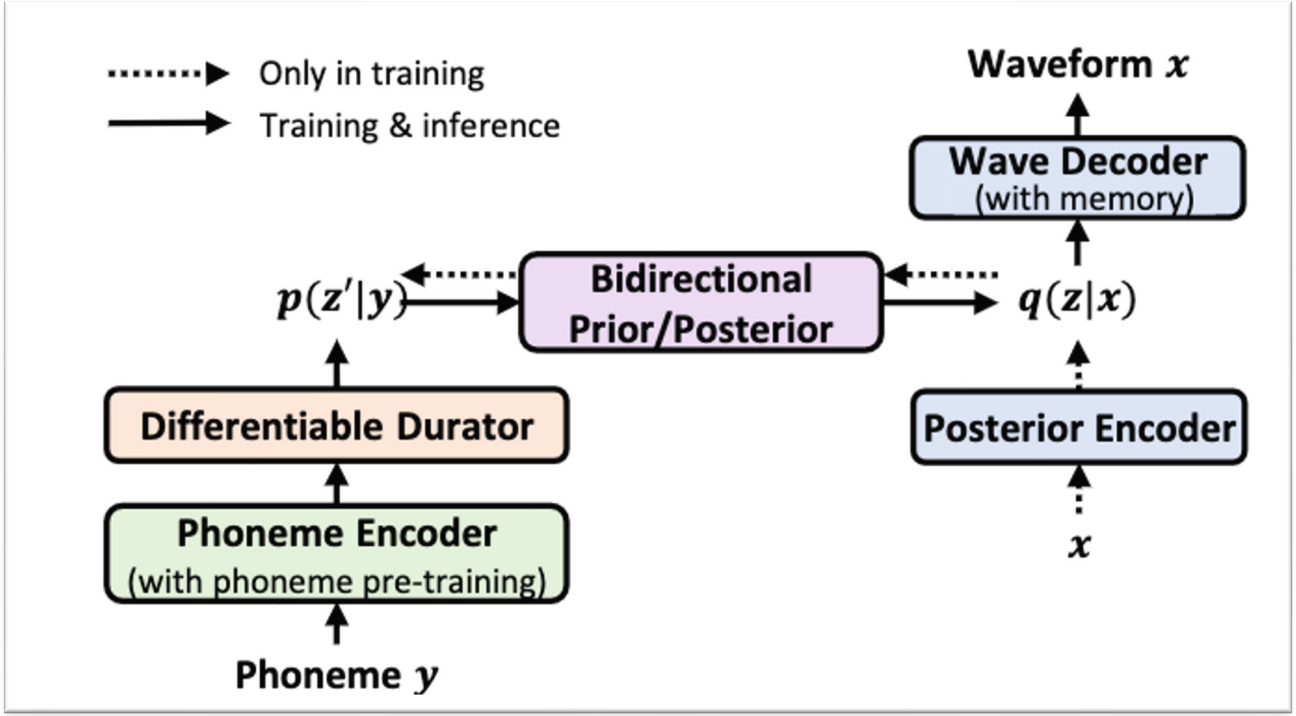
Ren, Yi, et al. "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech." *International Conference on Learning Representations*. 2020.

深度学习语音合成（端到端语音合成）



Kim, Jaehyeon, Jungil Kong, and Juhee Son. "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech." *International Conference on Machine Learning*. PMLR, 2021.

深度学习语音合成（端到端语音合成）



Tan, Xu, et al. "NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality." *arXiv preprint arXiv:2205.04421* (2022).

01 语音合成

02 **语音情感/口音表达**

03 情感可控语音合成

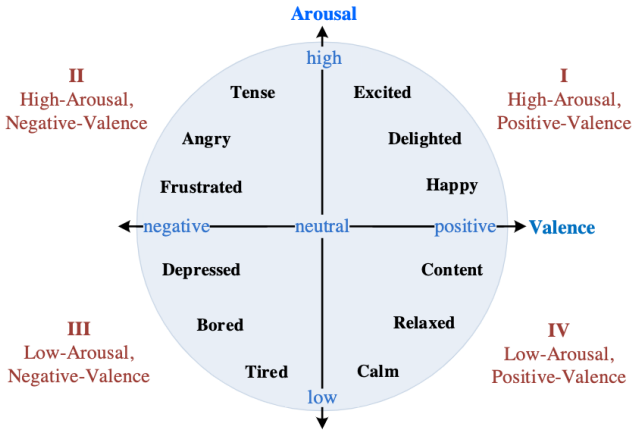
04 口音可控语音合成

语音情感/口音表达

- 离散情感模型

Ekman六元基本情感

- 连续情感模型



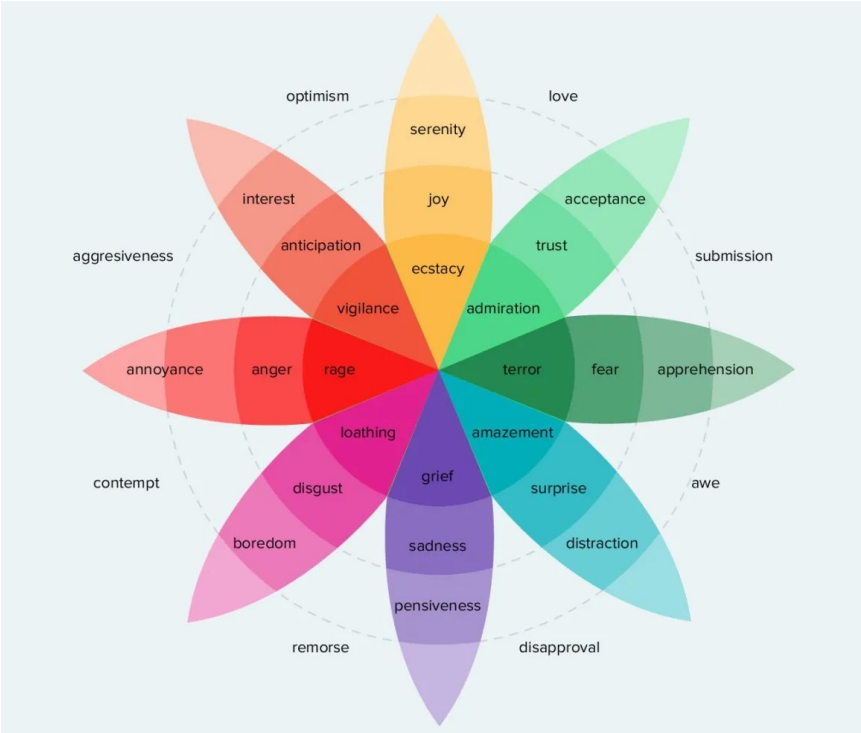
二维连续情感模型

Ekman P. An argument for basic emotions. Cognition & emotion, 1992, 6(3-4): 169-200.

Russell, James A. A circumplex model of affect. Journal of personality and social psychology 39.6 (1980): 1161.

语音情感/口音表达

- 情感强度模型



Plutchik 情感轮模型

Plutchik R. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice[J]. American scientist, 2001, 89(4): 344-350.

语音情感表达

语义信息 → 声学特征 → 情感状态

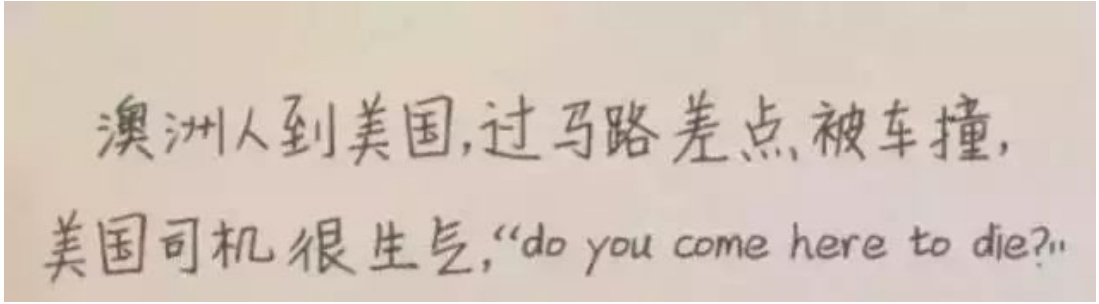
“层次化”

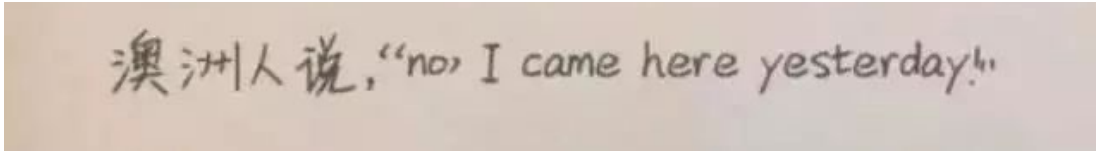
Huang, Chun-Fang, and Masato Akagi. "A three-layered model for expressive speech perception." *Speech Communication* 50.10 (2008): 810-828.

语音口音表达

160 distinct English accents throughout the world

语音口音表达


澳洲人到美国,过马路差点被车撞,
美国司机很生气,"do you come here to die?"


澳洲人说,"no, I came here yesterday!"

语音口音表达

today

e → aɪ

Phoneme substitution

音素替换现象

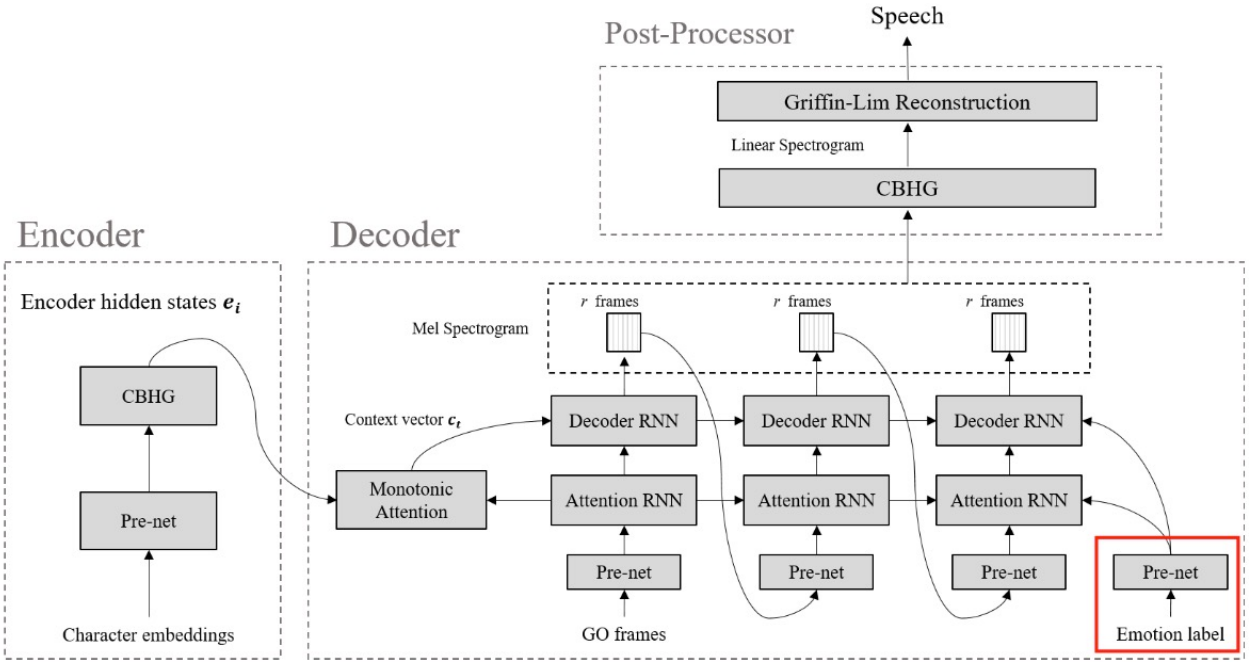
01 语音合成

02 语音情感/口音表达

03 **情感可控语音合成**

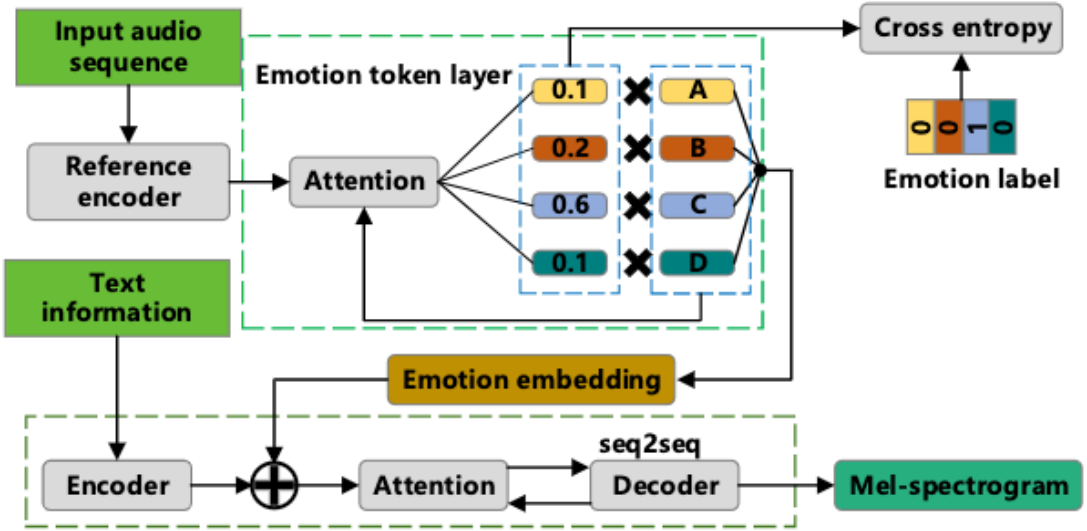
04 口音可控语音合成

情感可控语音合成



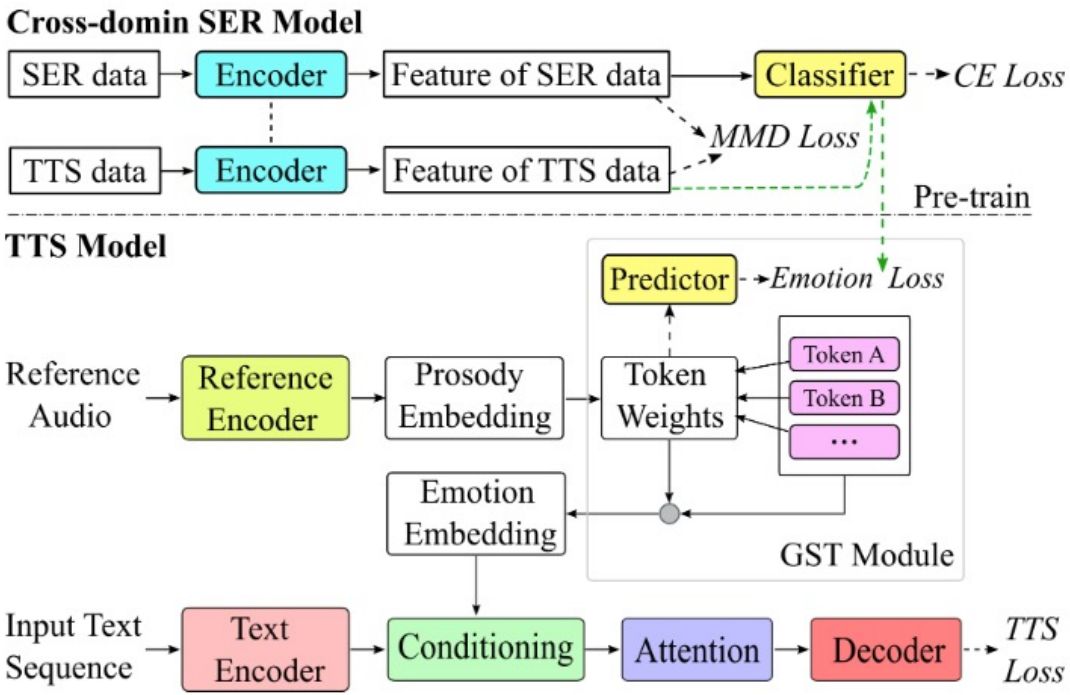
Lee, Younggun, Soo-Young Lee, and Azam Rabiee. "Emotional End-to-End Neural Speech Synthesizer." *NeurIPS2017*. Neural Information Processing Systems Foundation, 2017.

情感可控语音合成



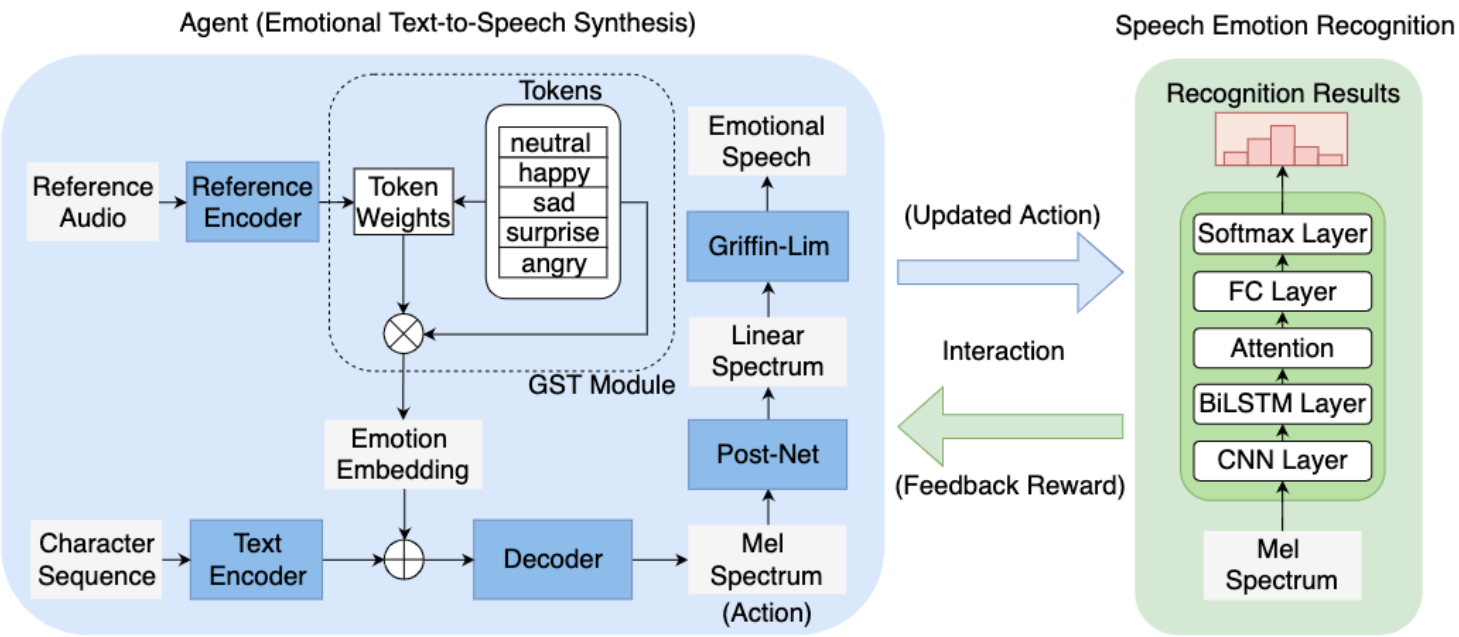
Wu, Pengfei, et al. "End-to-end emotional speech synthesis using style tokens and semi-supervised training." *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019.

情感可控语音合成



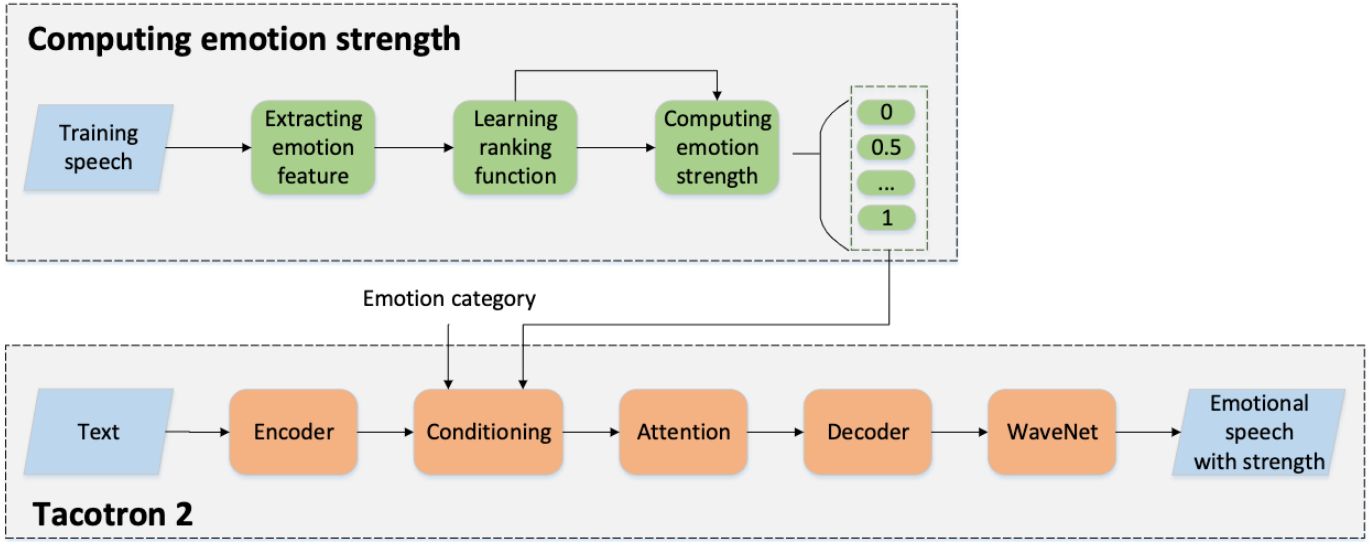
Cai, Xiong, et al. "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition." *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

情感可控语音合成



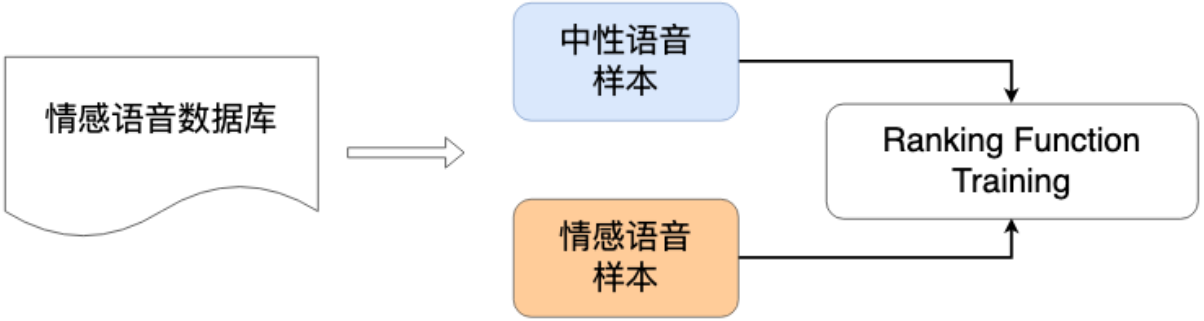
Liu, R., Sisman, B., Li, H. (2021) Reinforcement Learning for Emotional Text-to-Speech Synthesis with Improved Emotion Discriminability. Proc. Interspeech 2021, 4648-4652

情感可控语音合成



Zhu, Xiaolian, et al. "Controlling emotion strength with relative attribute for end-to-end speech synthesis." *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019.

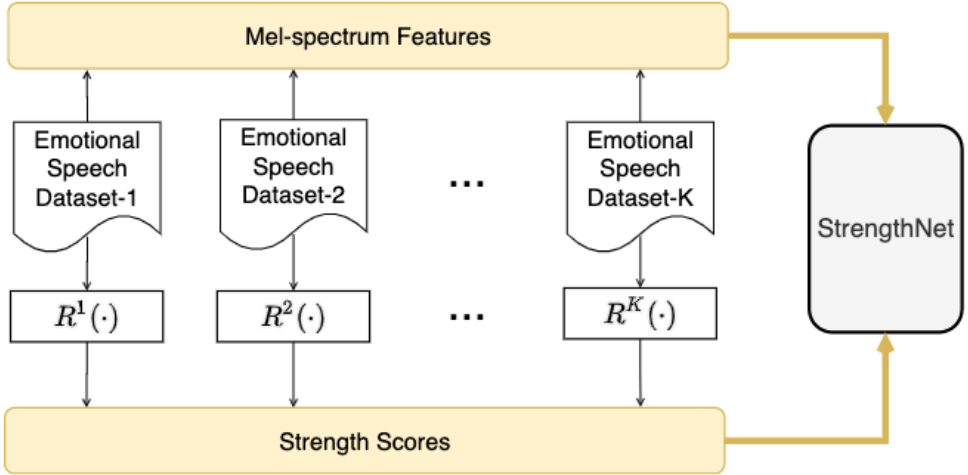
情感可控语音合成



Max-margin optimization problem -> Support Vector Machines (SVM)

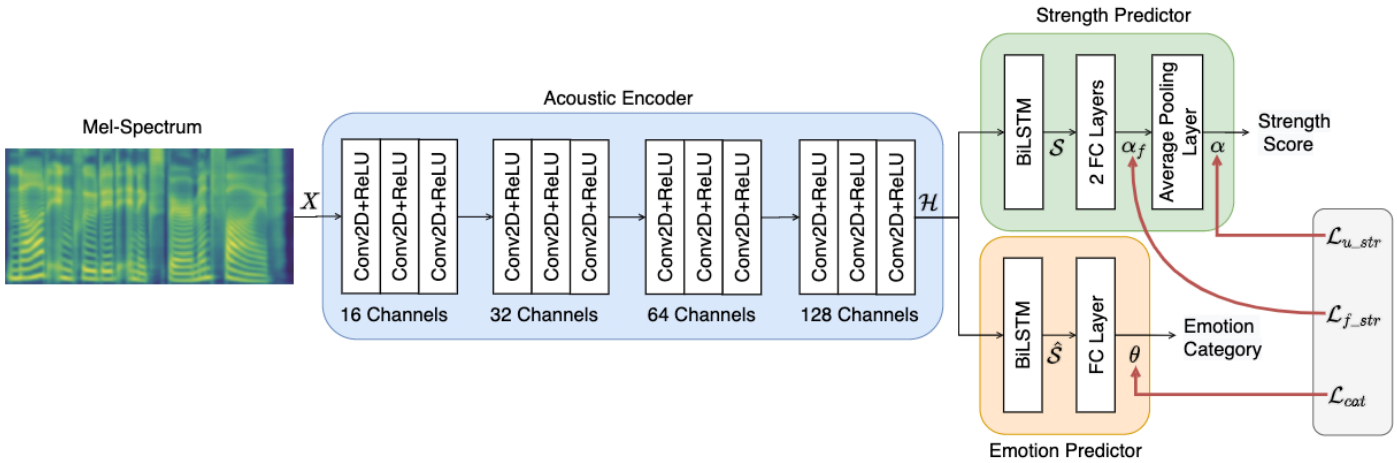
Zhu, Xiaolian, et al. "Controlling emotion strength with relative attribute for end-to-end speech synthesis." *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019.
Devi Parikh and Kristen Grauman, “Relative attributes,” in Proc. ICCV. IEEE, 2011, pp. 503–510.

情感可控语音合成



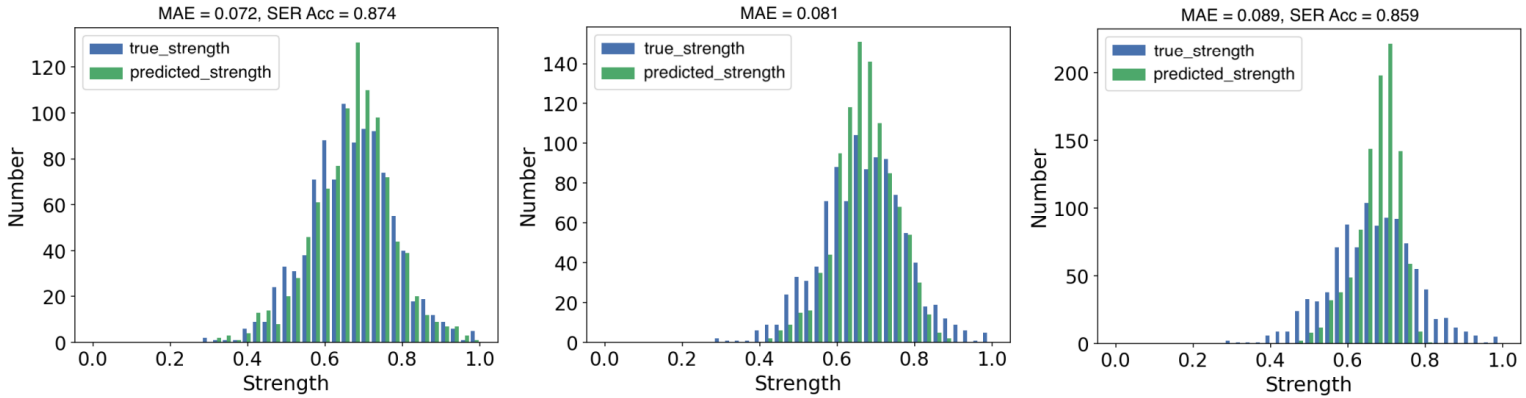
Liu, R., Sisman, B., Schuller, B., Gao, G., Li, H. (2022) Accurate Emotion Strength Assessment for Seen and Unseen Speech Based on Data-Driven Deep Learning. Proc. Interspeech 2022, 5493-5497.

情感可控语音合成



Liu, R., Sisman, B., Schuller, B., Gao, G., Li, H. (2022) Accurate Emotion Strength Assessment for Seen and Unseen Speech Based on Data-Driven Deep Learning. Proc. Interspeech 2022, 5493-5497.

Dataset： ESD



(a) StrengthNet (b) StrengthNet w/o $\mathcal{L}_{cat}$ (c) StrengthNet w/o $\mathcal{L}_{f_str}$

Figure 3: *Histogram of the utterance level strength predictions for (a) StrengthNet; (b) StrengthNet w/o $\mathcal{L}_{cat}$; (c) StrengthNet w/o $\mathcal{L}_{f_str}$. The X-axis and Y-axis of subfigures represent the strength scores and the utterance number, respectively.*

Liu, R., Sisman, B., Schuller, B., Gao, G., Li, H. (2022) Accurate Emotion Strength Assessment for Seen and Unseen Speech Based on Data-Driven Deep Learning. Proc. Interspeech 2022, 5493-5497.
Zhou, Kun, et al. "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

情感可控语音合成

Dataset： ESD， RAVDESS， SAVEE

Table 1: *MAE results for the RAVDESS and SAVEE datasets in the comparison.*

Method	**MAE**	
	RAVDESS	SAVEE
$R_{\mathrm{RAVDESS}}(\cdot)$	NA	0.304
$R_{\mathrm{SAVEE}}(\cdot)$	0.283	NA
$R_{\mathrm{ESD}}(\cdot)$	0.266	0.272
$\mathrm{StrengthNet}_{\mathrm{ESD}}$	0.238	0.243
$\mathrm{StrengthNet}_{\mathrm{ESD+RAVDESS}}$	NA	**0.173**
$\mathrm{StrengthNet}_{\mathrm{ESD+SAVEE}}$	**0.102**	NA

Liu, R., Sisman, B., Schuller, B., Gao, G., Li, H. (2022) Accurate Emotion Strength Assessment for Seen and Unseen Speech Based on Data-Driven Deep Learning. Proc. Interspeech 2022, 5493-5497.

Dataset： RAVDESS

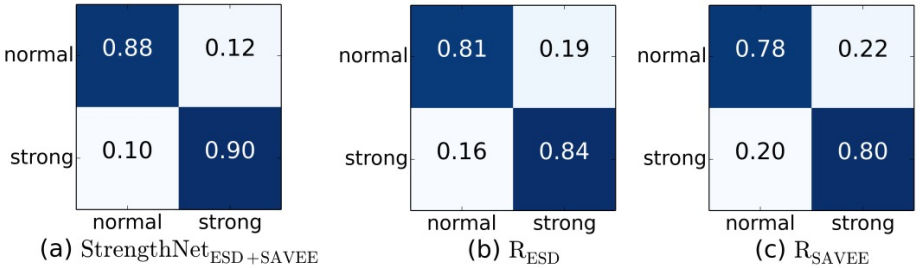
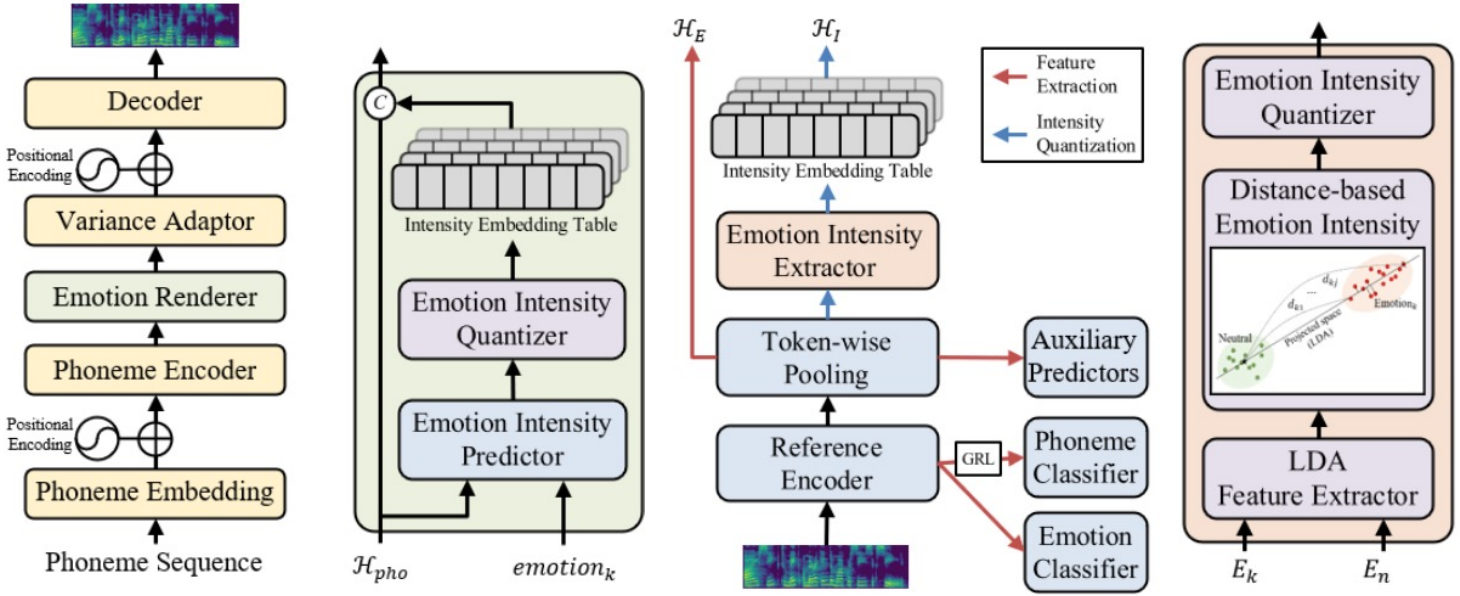


Figure 4: *Confusion matrices between predicted and perceived emotion strength categories for unseen speech. The X-axis and Y-axis of the figures represent the predicted and perceived category, namely "normal" and "strong".*

Liu, R., Sisman, B., Schuller, B., Gao, G., Li, H. (2022) Accurate Emotion Strength Assessment for Seen and Unseen Speech Based on Data-Driven Deep Learning. Proc. Interspeech 2022, 5493-5497.

Distance-based Emotion Intensity Quantization



(a) Overall structure (b) Emotion renderer (c) Emotion intensity modeling (d) Intensity extractor

Im, Chae-Bin, et al. "EMOQ-TTS: Emotion Intensity Quantization for Fine-Grained Controllable Emotional Text-to-Speech." *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

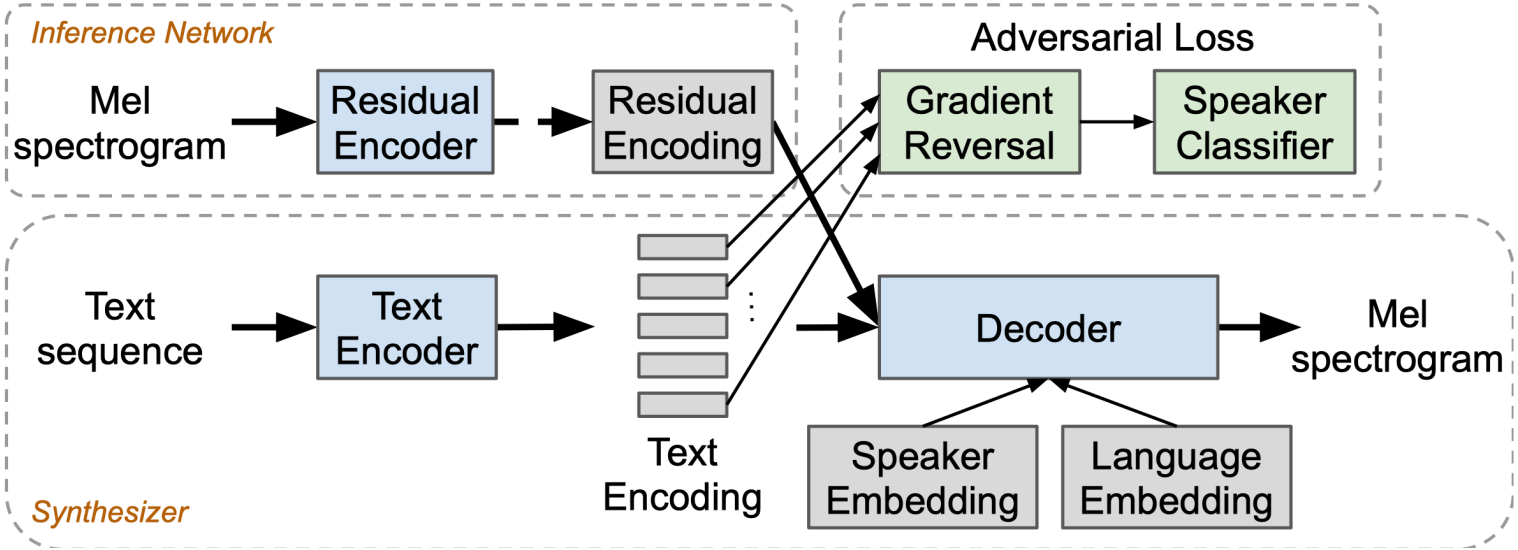
01 语音合成

02 语音情感/口音表达

03 情感可控语音合成

04 口音可控语音合成

口音可控语音合成



Zhang, Yu, et al. "Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning}}." Proc. Interspeech 2019 (2019): 2080-2084.

口音可控语音合成

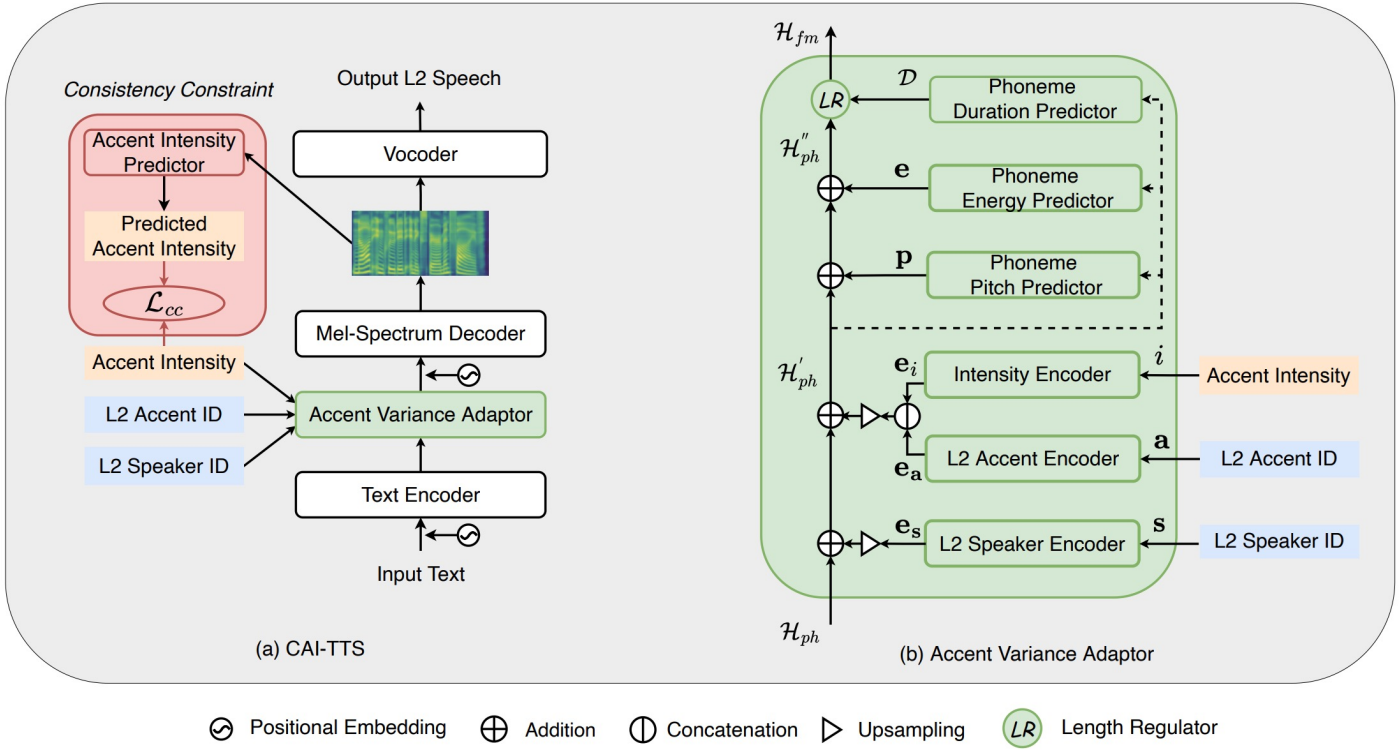
Accent Control

	Text: No somos seres aislados, formamos parte de la cadena universal de la vida.	*Text: Esta demostración es presentada por el equipo de Google.*
Reference Speaker	0:00 / 0:02	
Language id=Mandarin, adv weight=0.02	0:00 / 0:04	0:00 / 0:03
Language id=English, adv weight=0.02	0:00 / 0:04	0:00 / 0:03
Language id=English, adv weight=0.05	0:00 / 0:04	0:00 / 0:03
Language id=English, adv weight=0.1	0:00 / 0:04	0:00 / 0:03

Accent level (light to strong)	*Text: 美国人学习中文, 很难区分不同音调.*	*Text: 第一班商业航班在一九一九年来往于美国和加拿大.*
Reference Speaker	0:00 / 0:02	
Language id=Mandarin, adv weight=0.02	0:00 / 0:03	0:00 / 0:04
Language id=English, adv weight=0.02	0:00 / 0:03	0:00 / 0:04
Language id=English, adv weight=0.05	0:00 / 0:02	0:00 / 0:04
Language id=English, adv weight=0.1	0:00 / 0:03	0:00 / 0:04

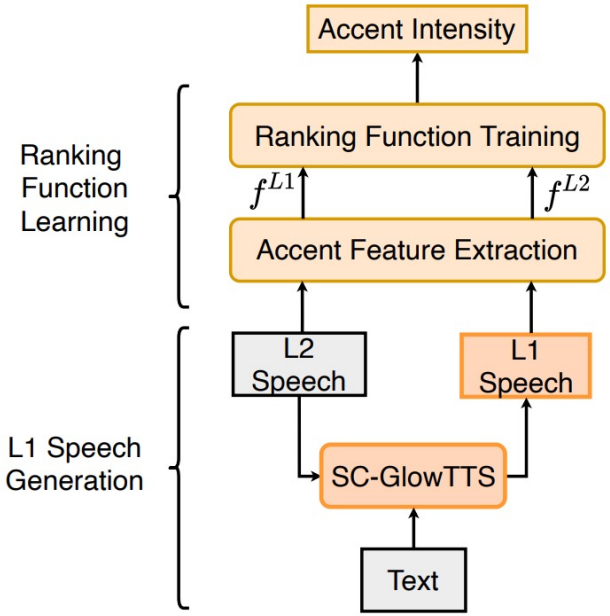
https://google.github.io/tacotron/publications/multilingual/

口音可控语音合成



Liu, Rui, Berrak Sisman, Guanglai Gao, and Haizhou Li. "Controllable Accented Text-to-Speech Synthesis." arXiv preprint arXiv:2209.10804 (2022).

口音可控语音合成



Liu, Rui, Berrak Sisman, Guanglai Gao, and Haizhou Li. "Controllable Accented Text-to-Speech Synthesis." arXiv preprint arXiv:2209.10804 (2022).
Casanova, Edresson, et al. “SC-GlowTTS: an efficient zero-shot multi-speaker text-to-speech model.” InterSpeech 2021.

Dataset: L2-ARCTIC

TABLE I: The comparison of the audio quality for different systems in terms of MCD in objective experiments, MOS and BWS in subjective experiments.

System	Audio Quality			
	MCD (dB)	*MOS*	BWS Evaluation	
			Best (%)	*Worst* (%)
GT	NA	4.62 ± 0.02	NA	NA
GT (Mel + HiFi-GAN)	NA	4.59 ± 0.03	NA	NA
Tacotron2 [9] (Mel + HiFi-GAN)	6.57	4.33 ± 0.01	6	34
Transformer TTS [10] (Mel + HiFi-GAN)	6.43	4.35 ± 0.02	8	22
FastSpeech2 [11] (Mel + HiFi-GAN)	6.31	4.38 ± 0.03	5	18
CAI-TTS *(proposed)* (Mel + HiFi-GAN)	**6.08**	**4.57 ± 0.01**	**45**	**4**
w/o phoneme pitch & energy	6.11	4.52 ± 0.02	9	8
w/o accent intensity	6.14	4.48 ± 0.03	13	6
w/o consistency constraint	6.16	4.43 ± 0.01	14	8

Liu, Rui, Berrak Sisman, Guanglai Gao, and Haizhou Li. "Controllable Accented Text-to-Speech Synthesis." arXiv preprint arXiv:2209.10804 (2022).
Zhao, Guanlong, et al. "L2-ARCTIC: A non-native English speech corpus." INTERSPEECH. 2018.

口音可控语音合成

TABLE II: The comparison of the accent variance information for different systems, including standard deviation (σ), skewness (γ), kurtosis ($\mathcal{K}$) and average dynamic time warping (DTW) distances (ϱ) for pitch, mean absolute error (*MAE*) for energy and the average of absolute boundary differences (Δ) for duration.

System	Accent Variance					
	Pitch				Energy	Duration
	σ	γ	$\mathcal{K}$	ϱ	*MAE*	Δ (ms)
GT	49.8	0.627	0.854	NA	NA	NA
GT (Mel + HiFi-GAN)	49.5	0.629	0.852	15.27	0.192	NA
Tacotron2 [9] (Mel + HiFi-GAN)	32.5	0.906	1.212	19.80	0.297	22.12
Transformer TTS [10] (Mel + HiFi-GAN)	35.8	0.887	1.025	19.32	0.288	21.98
FastSpeech2 [11] (Mel + HiFi-GAN)	43.2	0.752	0.973	18.01	0.274	20.08
CAI-TTS *(proposed)* (Mel + HiFi-GAN)	**48.9**	**0.613**	**0.882**	**16.78**	**0.235**	**18.54**
w/o phoneme pitch & energy	48.3	0.610	0.896	17.06	0.240	18.72
w/o accent intensity	47.9	0.611	0.890	16.95	0.239	18.70
w/o consistency constraint	47.8	0.594	0.889	17.13	0.245	18.85

Liu, Rui, Berrak Sisman, Guanglai Gao, and Haizhou Li. "Controllable Accented Text-to-Speech Synthesis." arXiv preprint arXiv:2209.10804 (2022).

口音可控语音合成

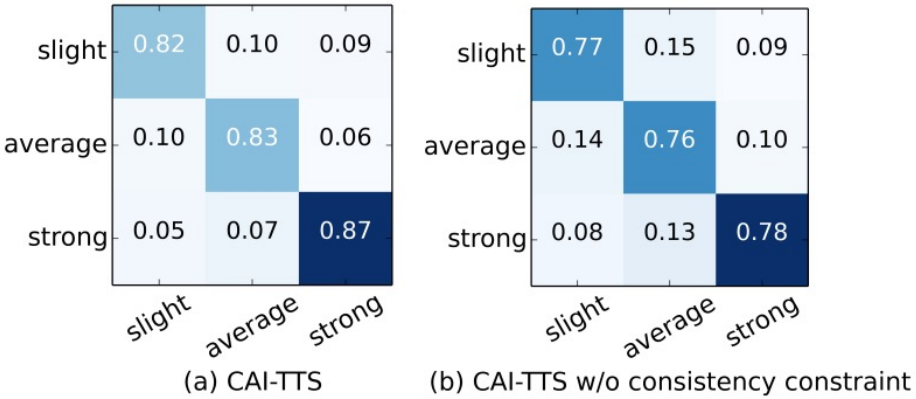
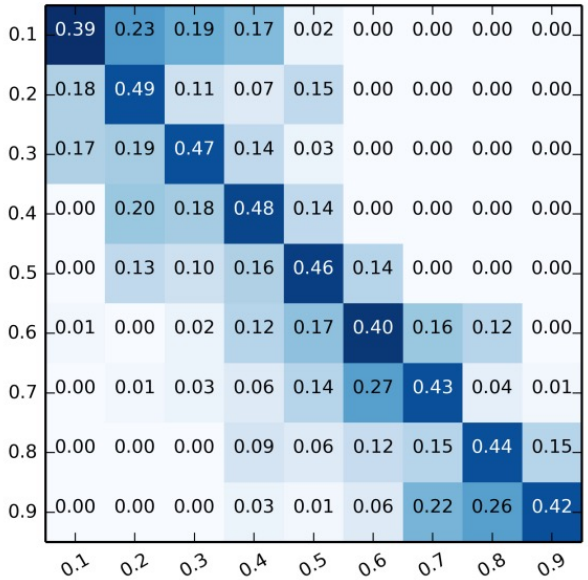


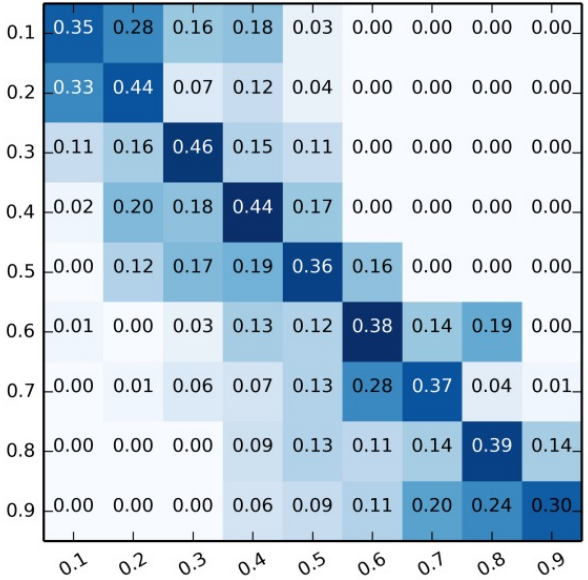
Fig. 4: Confusion matrices between perceived and intended accent intensity categories of synthesized speech. (a) CAI-TTS; b) CAI-TTS w/o consistency constraint. The X-axis and Y-axis of the figures represent the perceived and intended category, namely slight, average, and strong.

Liu, Rui, Berrak Sisman, Guanglai Gao, and Haizhou Li. "Controllable Accented Text-to-Speech Synthesis." arXiv preprint arXiv:2209.10804 (2022).

口音可控语音合成



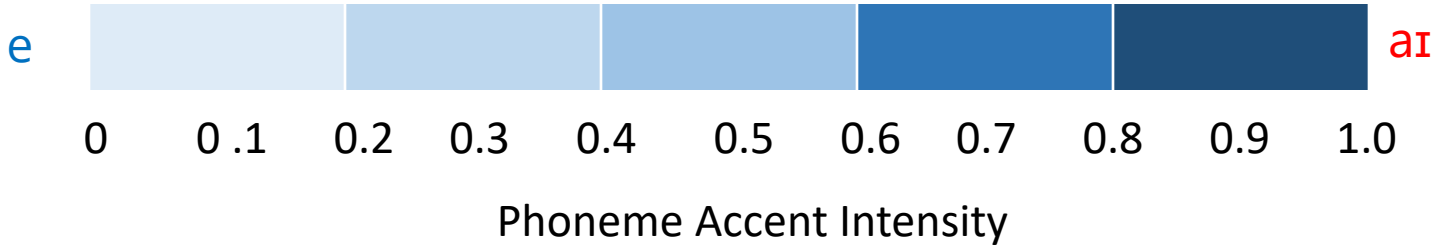
(a) CAI-TTS



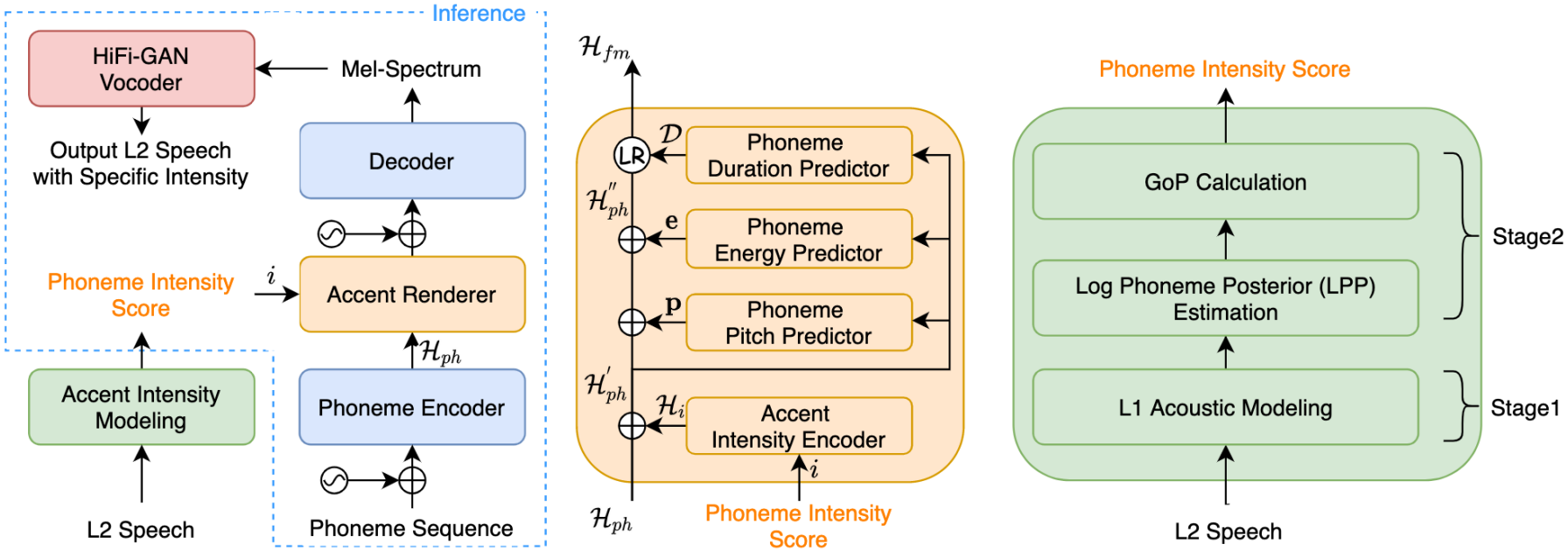
(b) CAI-TTS w/o consistency constraint

https://speechdemo.github.io/caitts/

口音可控语音合成



口音可控语音合成



(a) Ai-TTS Architecture (b) Accent Renderer (c) Accent Intensity Modeling

Liu, Rui, Haolin Zuo, De Hu, Guanglai Gao, and Haizhou Li. "Explicit Intensity Control for Accented Text-to-speech." arXiv preprint arXiv:2210.15364 (2022).

口音可控语音合成

Dataset： L2-ARCTIC （Mandarin）

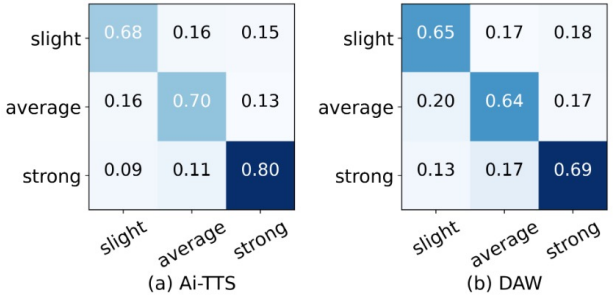


Fig. 2. Confusion matrices between perceived and intended accent intensity categories of synthesized speech. (a) Ai-TTS; b) DAW. The X-axis and Y-axis of the figures represent the perceived and intended category, namely slight, average, and strong.

Liu, Rui, Haolin Zuo, De Hu, Guanglai Gao, and Haizhou Li. "Explicit Intensity Control for Accented Text-to-speech." arXiv preprint arXiv:2210.15364 (2022).

口音可控语音合成

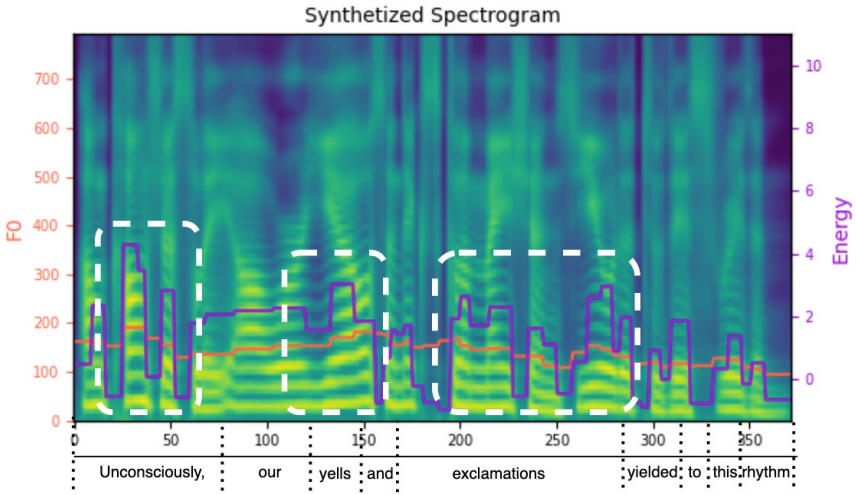


Fig. 3. Spectrogram, F0 contour and energy of an utterance with specific phoneme-level intensity score. While boxes indicates the acoustic performance with higher accent intensity. It is observed that Ai-TTS provides an effective explicit accent intensity control mechanism.

https://ttslr.github.io/Ai-TTS/

内蒙古大学
INNER MONGOLIA UNIVERSITY

谢谢！