Session:A7-P1-A
Paper #2216

INTERSPEECH 2024

UNIVERSITY of ROCHESTER

Audio Information Research Laboratory

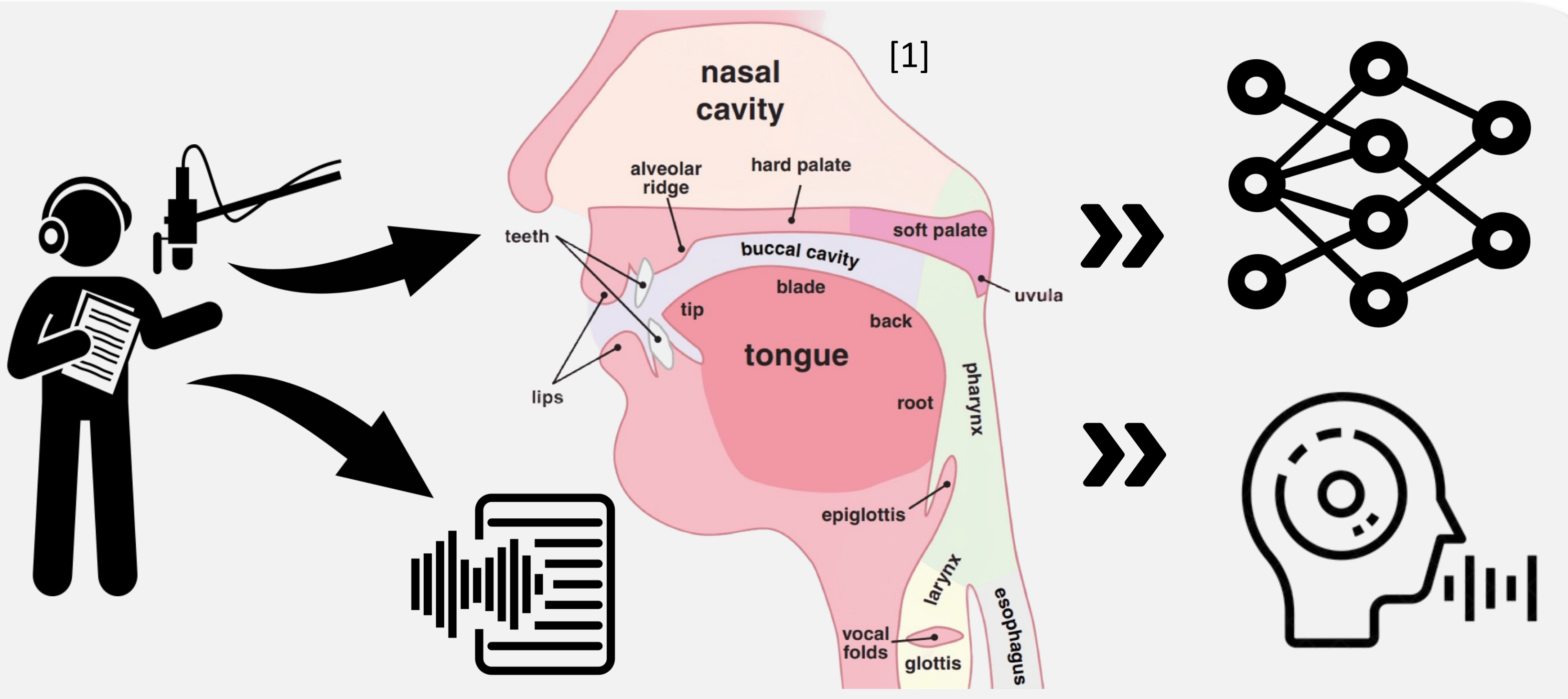
GTR-Voice

Articulatory Phonetics Informed Controllable Expressive Speech Synthesis

Zehua Kcriss Li, Meiying Melissa Chen, Yi Zhong, Pinxin Liu, Zhiyao Duan

Background and Motivation

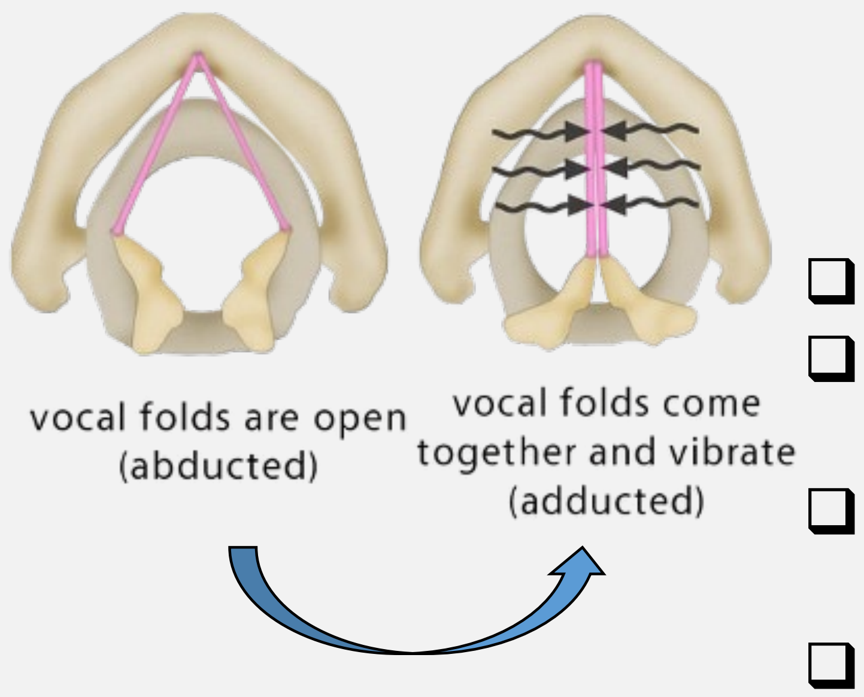
- ❑ Current speech synthesis excels in emotion but falls short in capturing **nuanced articulatory features** achieved by professional voice actors.
- ❑ This study introduces a novel **GTR framework** and dataset to improve control over expressive speech synthesis by focusing on **Glottalization**, **Tenseness**, and **Resonance**.
- ❑ Experimental results show controllability in expressive TTS, with user studies confirming GTR-based models in capturing articulatory nuances across various speech dimensions.



The GTR-Voice Dataset and GTR Controllable Speech Synthesis

Articulatory Phonetics Inspired Dimensions[2]

Glottalization[3]	Tenseness [4]	Resonance [5]
0-Whisper Voice	1-Laxest	0-Whisper Voice
1-Slack Voice	2-Slightly Lax	1-Chest Voice
2-Modal Voice	3-Moderate	2-Head Voice
3-Stiff Voice	4-Slightly Tense	3-Chest-Nasal Mix
4-Creaky Voice	5-Tensest	4-Chest-Head Mix
		5-Head-Nasal Mix
		6-Chest-Head-Nasal Mix


vocal folds are open (abducted)
vocal folds come together and vibrate (adducted)

Dataset Description

- ❑ 3.6 hours of 48Khz/24bits HQ speech audio
- ❑ 2500 clips, ~6 seconds each, representing one of the **125 unique GTR combinations**.
- ❑ All recorded by a **professional** 30-year-old male Mandarin **voice actor**
- ❑ **Fully accessible** under CC BY-NC-ND 4.0 license

Model Architecture

- ❑ **FastPitch**[6] Feedforward Transformer TTS model with pitch and duration predictors for mel spectrogram generation. We added **three embedding layers** to condition the encoder output on GTR labels.
- ❑ **StyleTTS**[7] Two-stage TTS model that captures prosody and emotion. We replaced the style encoder with a **GTR embedder,** retaining other pre-trained components.

Training

- ❑ **FastPitch** Pre-trained on AISHELL3[8] for **80 epochs**, then fine-tuned for **3000** epochs on GTR-Voice with GTR label embeddings.
- ❑ **StyleTTS** Pre-trained on Libri-TTS (460 hours) for 200 epochs. GTR embedder trained for **500 epochs** using an RTX 3090, fixing other pre-trained weights.

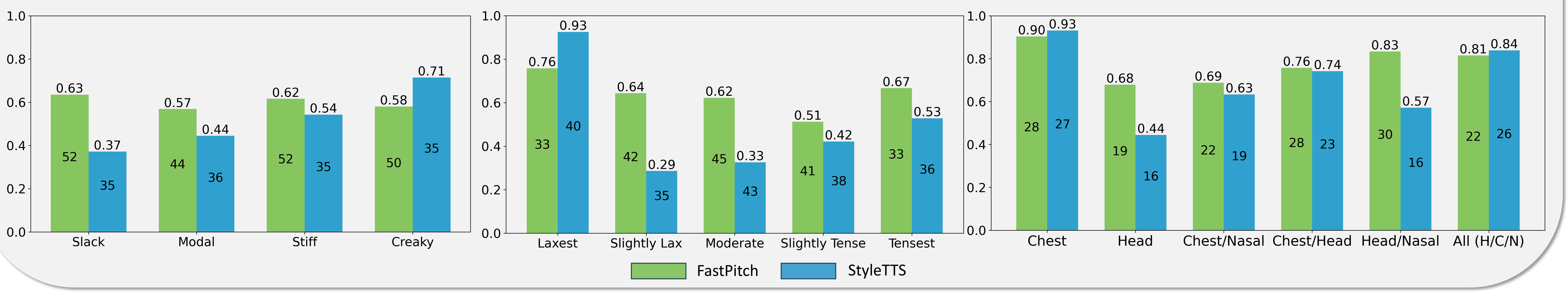
Experiments Result

- ❑ **Evaluation Setup** User study with **60 participants**, **40 webpages** (20 Chinese, 20 English). Participants compared **model-generated speech** with a **reference** and rated MOS.
- ❑ **MOS Scores** Both models scored above **3.00**, laying the foundation for controllability experiments.

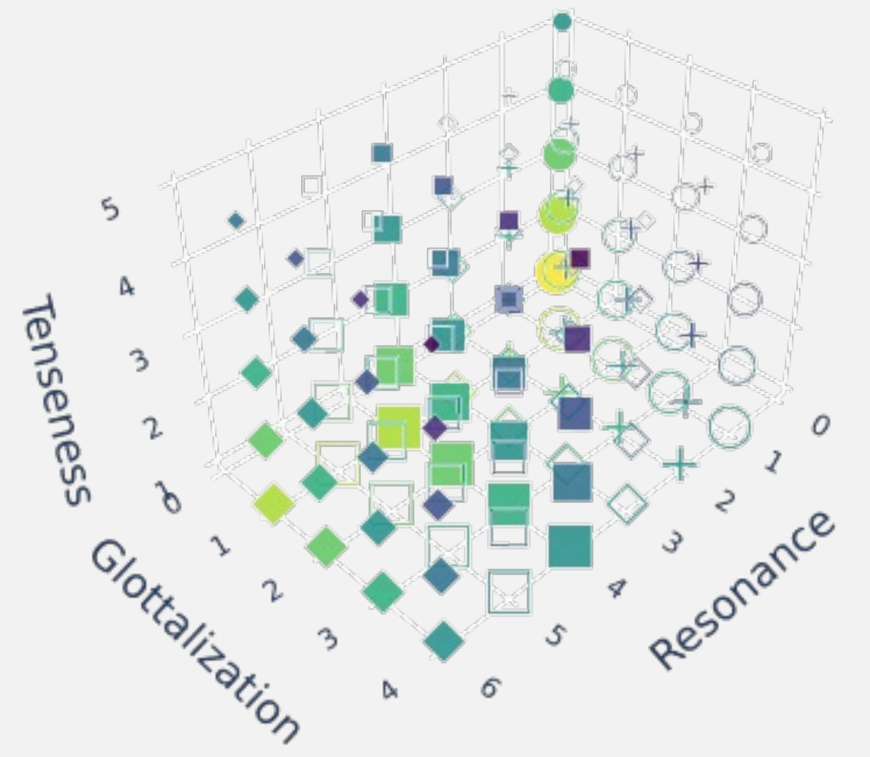
Model	Quality↑ (1-5)	Naturalness↑ (1-5)
FastPitch	3.05 ± 0.05	3.14 ± 0.11
StyleTTS	4.21 ± 0.14	4.16 ± 0.12

GTR Controllability

- ❑ **Glottalization** FastPitch: 67%, StyleTTS: 57%. Best for **Creaky Voice**, worst for **Slack Voice** (StyleTTS).
- ❑ **Tenseness** FastPitch led except for Laxest (StyleTTS: 68%). Significant **accuracy gaps** favoring FastPitch.
- ❑ **Resonance** Highest for Chest Voice (79% FastPitch, 71% StyleTTS). StyleTTS struggled with **Head Voice**.
- ❑ **Models Average** FastPitch: 67.07%, StyleTTS: 57.14%. **Best for R** dimension, **lowest for G** dimension.



Demo



3D
INTERACTIVE
VOICE
PLOT
PLEASE SCAN!

To learn more about the GTR-Voice, visit `https://GTR-Voice.com`

Reference

[1] Figure 2: IPA articulation points (left) Human vocal tract (right) IPA (vowels, consonants) articulation points.
[2] G. S. Nathan, The Sounds of the World's Languages. JSTOR, 1998.
[3] M. Garellek, "Voice quality strengthening and glottalization," Journal of Phonetics, vol. 45, pp. 106–113, 2014.
[4] J. Kuang and P. Keating, "Glottal articulations in tense vs lax phonation contrasts," The Journal of the Acoustical Society of America, vol. 134, pp. 4069–4069, 11 2013.
[5] H. Hollien, "On vocal registers," Journal of Phonetics, vol. 2, pp. 125–143, 1974.
[6] A. Łancucki, "Fastpitch: Parallel text-to-speech with pitch ´ prediction," in in Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 6588– 6592.
[7] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," in Advances in Neural Information Processing Systems, vol. 36, 2023, pp. 19 594–19 621.
[8] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A Multi-Speaker Mandarin TTS Corpus," in Proc. Interspeech, 2021, pp. 2756–2760.
[9] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in Proc. Interspeech, 2019, pp. 1526–1530.