Learning Fine-Grained Controllability on Speech Generation via Efficient Fine-Tuning

Chung-Ming Chien[1], Andros Tjandra[2], Apoorv Vyas[2], Matt Le[2], Bowen Shi[2], Wei-Ning Hsu[2]
TTI-Chicago[1], FAIR at Meta[2]

Overview

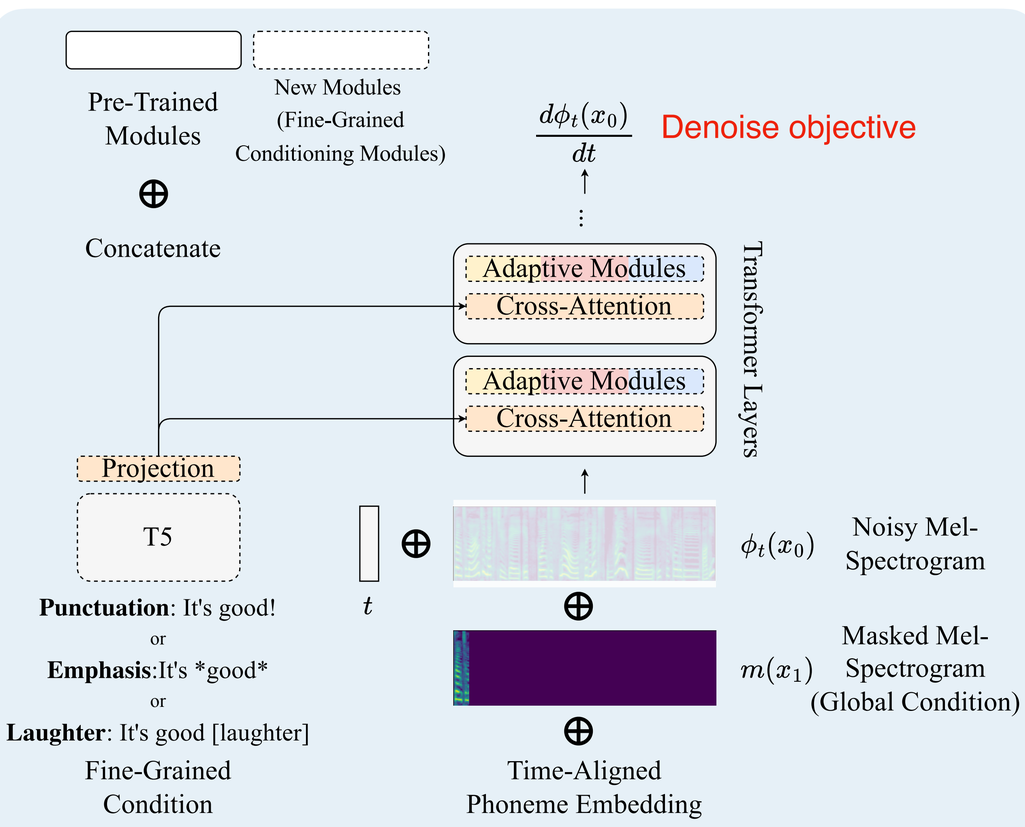
Fine-grained controllable speech generation

- Punctuation, pauses, laughter, etc.
- Previous work: task-specific design & training.

Efficient fine-tuning on fine-grained conditions

- Speech generation models **pre-trained on large corpora** already learn fine-grained vocalizations.
- We teach pre-trained models to follow user-specified **fine-grained conditions via efficient fine-tuning methods**.

Model Architecture



- $x_0 \sim p(x)$: Gaussian noise.
- Initial condition: $\phi_0(x_0) = x_0$.
- Evaluate $\frac{d\phi_t(x_0)}{dt}$ with the model.
- Solve $\phi_1(x_0)$ with an ODE solver.

Pre-Training

Global conditioning (speaker, style) from masked Mel-spectrogram

- Telephone conversation + audiobook + podcast
- 61.8k hrs, **20k speakers**
- **Diverse styles, expressiveness, quality**, etc.

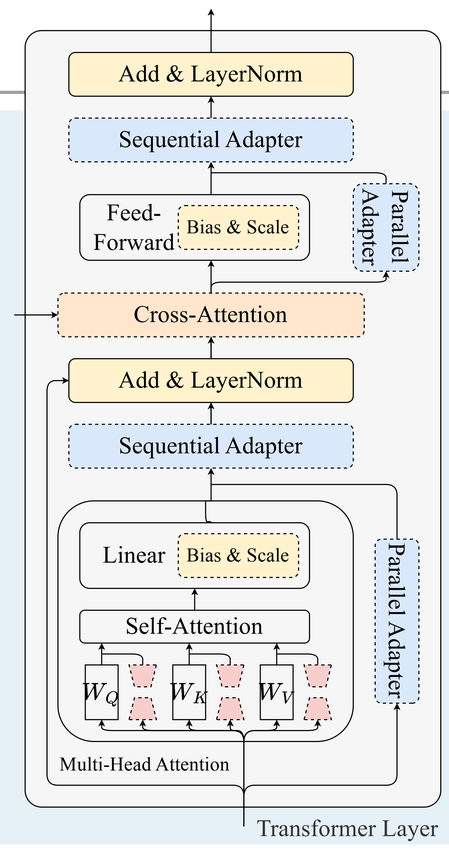
Fine-Tuning

Fine-grained conditional speech generation tasks

- **Punctuation** (550 hrs), **emphasis** (20 hrs), **laughter** (250 hrs)

Efficient fine-tuning methods

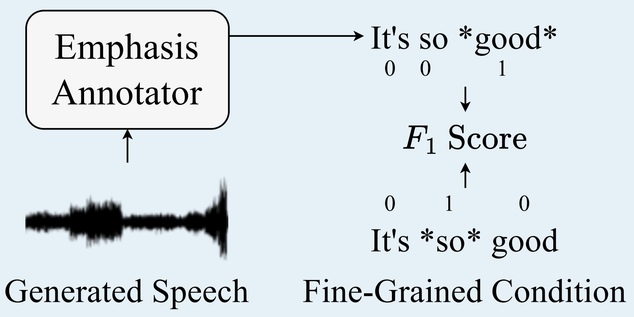
- Adapters
- LoRA
- Bias-tuning



Evaluation

Fine-grained controllability

- FC-MOS: subjective fine-grained controllability.
- F_1 score evaluated with annotation models.



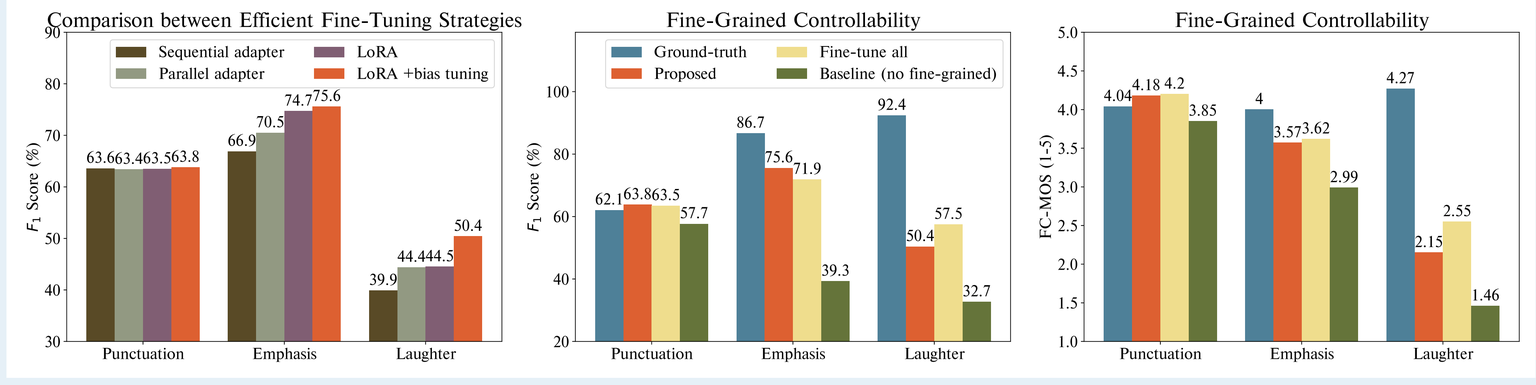
Speech quality

- Q-MOS: subjective quality.
- WER: word-error rate.

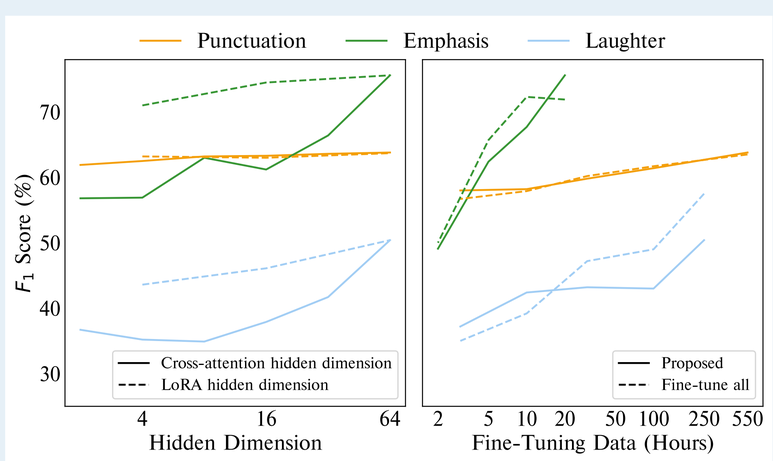
Global controllability

- SIM-o: speaker similarity between $m(x_1)$ and generated speech.

Result & Conclusion



Models	Speech Quality		Global Controllability
	WER(%)	Q-MOS	SIM-o
Ground-truth	5.3	3.84	0.718
Proposed	3.2	3.88	0.593
Fine-tune all	3.3	3.89	0.568
Baseline	3.5	3.82	0.565



Augment pre-trained speech generation model with fine-grained controllability via efficient fine-tuning.

- Comparable to full fine-tuning.
- Struggle on laughter generation.
- Does not hurt quality & global control.
- Robust across different data sizes.

INTERSPEECH 2024