Applying LoRA To TTS Models

Andrew Zhou & Huy Huynh

Agenda

Introduction

Motivation

Current Models

Mainland Mandarin

是 吃 知

shì chī zhī

Lots of retroflexes

The Gap

Taiwanese Mandarin

是 吃 知

sì cī zī

No retroflexes!

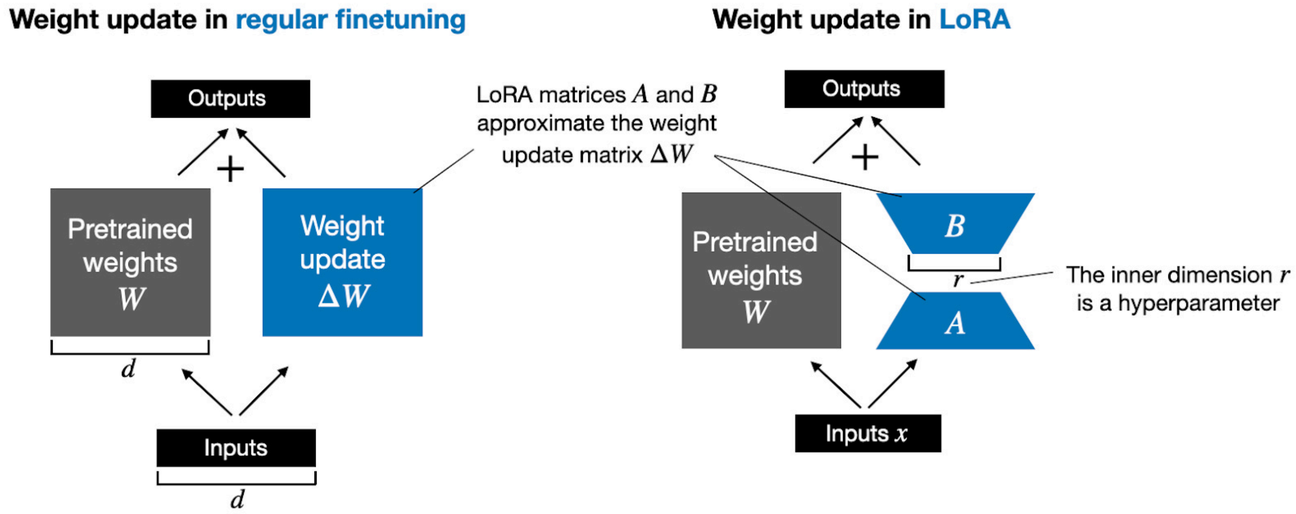
Impact

Why are we doing this?

- Accessibility
- Diversity
- More

LoRA

What is it, what are its tradeoffs, and when should I use it?



Training
Jack Of All Trades

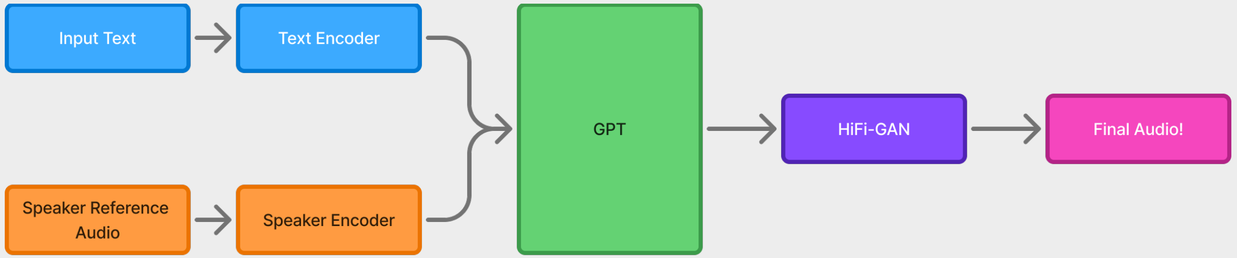
Time	High
Space	High
Breadth	Very High
Depth	Low

Reg. Fine Tuning
Domain Expert

Time	Medium
Space	High
Breadth	Medium
Depth	High

LoRA
Healthy Intermediate

Time	Low
Space	Low – Med
Breadth	Med – High
Depth	Med - High



443,721,923 Params → 2,703,360 params (99.4% reduction)

Why use LoRA for TTS?

Currently, good models require lots of parameters which presents a barrier for consumers

- → Models are expensive to train
- → Even minor tweaks = Big resource expenditure
- → LoRA can provide good performance with little downside

Goals

Personal Goals

Understanding The Problem & Complexity

- Learn modern architectures
- ML ❤ Linguistics

Broader Goals

Training & Assessing Performance

- Demonstrate effectiveness
- Connect to real applications

Experiment

Setup

Environment

Google Colab – T4 GPU & High Ram

Dataset

Mozilla Data Collective (zh-TW)

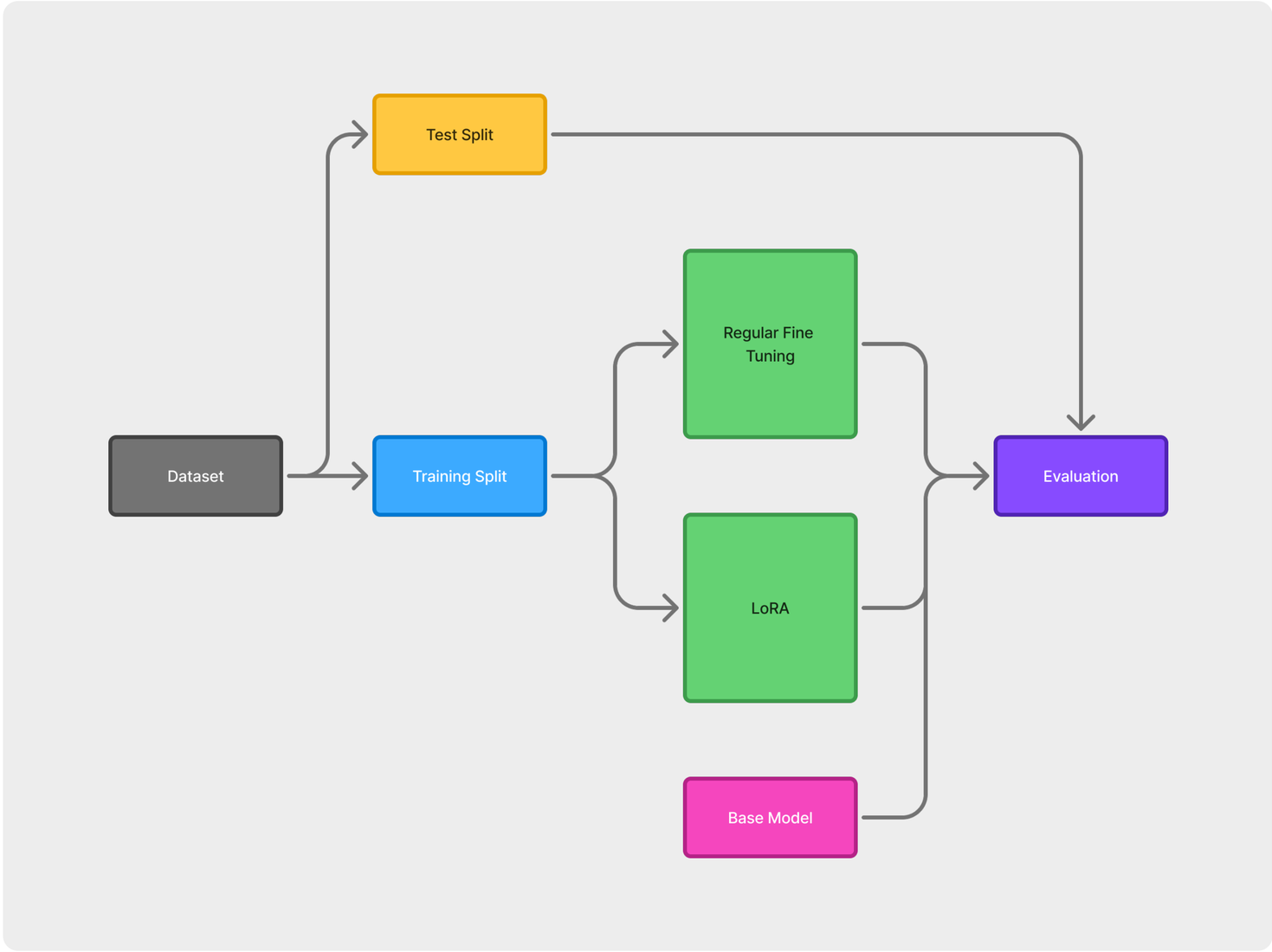
30 minutes of audio used to train

Hyperparameters

Rank 8 LoRA & 2e-4 LR

5e-6 LR fine tuning

30 epochs for both



Quantitative Results

We used the metrics defined at the right

Base Model

SECS	Medium
CER	High

Reg. Fine Tuning

SECS	High
CER	High

LoRA

SECS	Medium
CER	High

Speaker Encoder Cosine Similarity (SECS)

How close does a model's "speaker" sound to the reference speaker?

Higher is better

Character Error Rate (CER)

When transcribing a model's audio output back to speech, how many wrong characters were there?

Lower is better

Qualitative Results

Treat these assessments with a grain of salt – we're just two people

Base Model

Authenticity	Medium
Flow	High

Reg. Fine Tuning

Authenticity	High
Flow	High

LoRA

Authenticity	Medium
Flow	High

Authenticity

Did the model sound like a plausible Taiwanese speaker?

Did it learn to omit retroflexes?

Higher is better

Flow

This measures qualities like pauses, punctuation, and prosody (stresses & intonation)

Smoother is better

Resource Usage

Regular Fine Tuning

Learning rate 5e-6

1 Hour

Training Time

5.61 Gb

Average Checkpoint Size

LoRA

Learning rate 2e-4

30 Min

Training Time

2.11 Gb

Average Checkpoint Size

Extensions & Limitations

Extensions

Large TTS Models

e.g. FastSpeech2, GlowTTS

Similar Languages, Dialects, and Accents

e.g. British English, Beijing Accent

Many Linear Layer

Model Architectures

Phonemes

Similar
Coverage

Prosody

Similar
Dynamics

Limitations

Small TTS Models

e.g. Autoregressive RNNs, Diffusion vocoders

Languages, Dialects, and Accents

e.g. Cantonese, Highland English

Limited Linear Layer

Model Architectures

Phonemes

Different Coverage

Prosody

Different Dynamics

LoRA is good!

It sounds slightly more robotic, but when it works, the tradeoff of resources is well worth it (in our opinion).

Q&A

Thank You!