HW1 - TTS

Team 5
B09901116 陳守仁、B09602017 白宗民、B08901207 呂俐君

Outline

- **Text Processed into Model Input**
  - Number
  - Cleaners
  - Languages
- **Pre-processing of Original Audio Files**
- **Training**
  - Input Formats
  - Loss
  - Duration
  - Generating Speed

Basic Experiments

Text Processed into Model Input - Number

	Input	Output
Removing Commas	2,500	"2500"
Expanding Decimal Points	3.14	Three point one four
Expanding Currency Formats	£500 $20.50	five hundred pounds twenty dollars, fifty cents
Expanding Ordinals	1st / 22nd	First / twenty-second
General Number Expansion	2019 12345	two thousand nineteen twelve thousand three hundred forty-five

Text Processed into Model Input - Cleaners

Basic Cleaners	Input	Output
Uppercase → lowercase	HELLO WORLD	Hello world
Number Expansion	I have 2 dogs	I have two dogs
Abbreviation Expansion	Dr. Smith	doctor smith

Transliteration Cleaners	Input	Output
(for non-English text)	for "Hello"	ASCII transliteration
Japanese	こんにちは	konnichiwa

Text Processed into Model Input - English

Input:

"Deep learning is fun."

Phonetic Representation:

"{d}{iy}{p} {l}{iy}{r}{n}{ih}{ng} {ih}{z} {f}{ah}{n}"

Tokenized Sequence:

This phonetic representation is then converted into a sequence of numerical tokens that the model can process.

Text Processed into Model Input - Mandarin

Input: "機器學習很有趣"

Pinyin Conversion:

"jī qì xué xí hěn yǒu qù".

Phonetic Representation:

This Pinyin is further converted to a phonetic sequence, possibly using a lexicon.

Tokenized Sequence: Similar to the English preprocessing, this phonetic representation is converted into a sequence of tokens for the TTS model.

Pre-processing of Original Audio Files

取出原始音檔的

- mel spectrogram
- Energy
- pitch contour
- duration

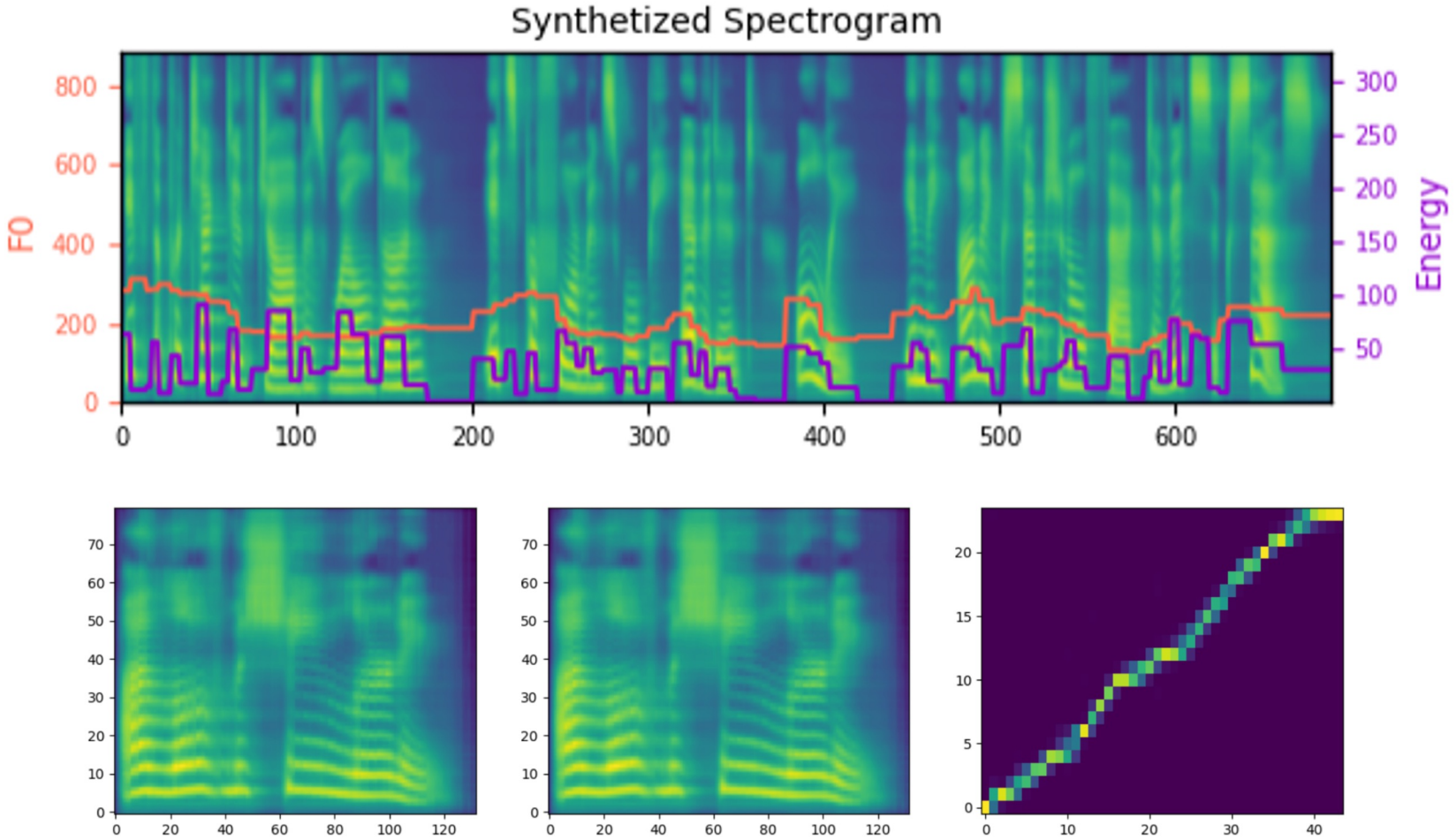
Input Formats

Tacotron2:

- Characters
- Mel Spectrogram

FastSpeech2:

- Phonemes
- Mel Spectrogram
- Energy
- Pitch contour
- Duration


Synthetized Spectrogram
F0
Energy
800
600
400
200
0
300
250
200
150
100
50
0
100
200
300
400
500
600
70
60
50
40
30
20
10
0
0
20
40
60
80
100
120
20
15
10
5
0
0
10
20
30
40

Loss of TTS Models

Tacotron2

1. Total Loss
2. Mel Loss
3. Gate Loss

FastSpeech2

1. Total Loss
2. Mel Loss
3. Mel-Postnet Loss
4. Pitch Loss
5. Energy Loss
6. Duration Loss

Training Duration

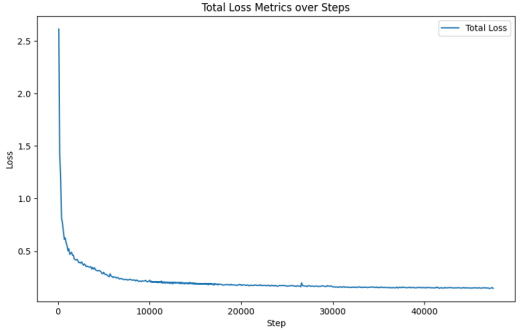
Tacotron2:

25 Hours

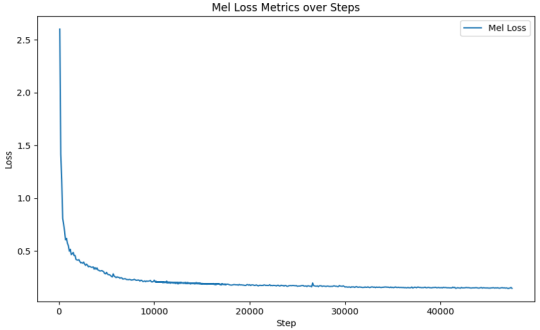
Fast Speech2:

- LJSpeech ~4hrs
- AISHELL ~4.5hrs

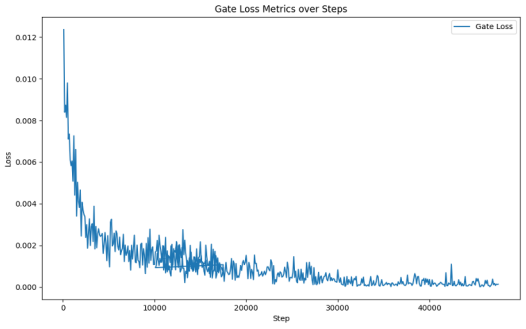
Tacotron Train on LJSpeech



Total Loss

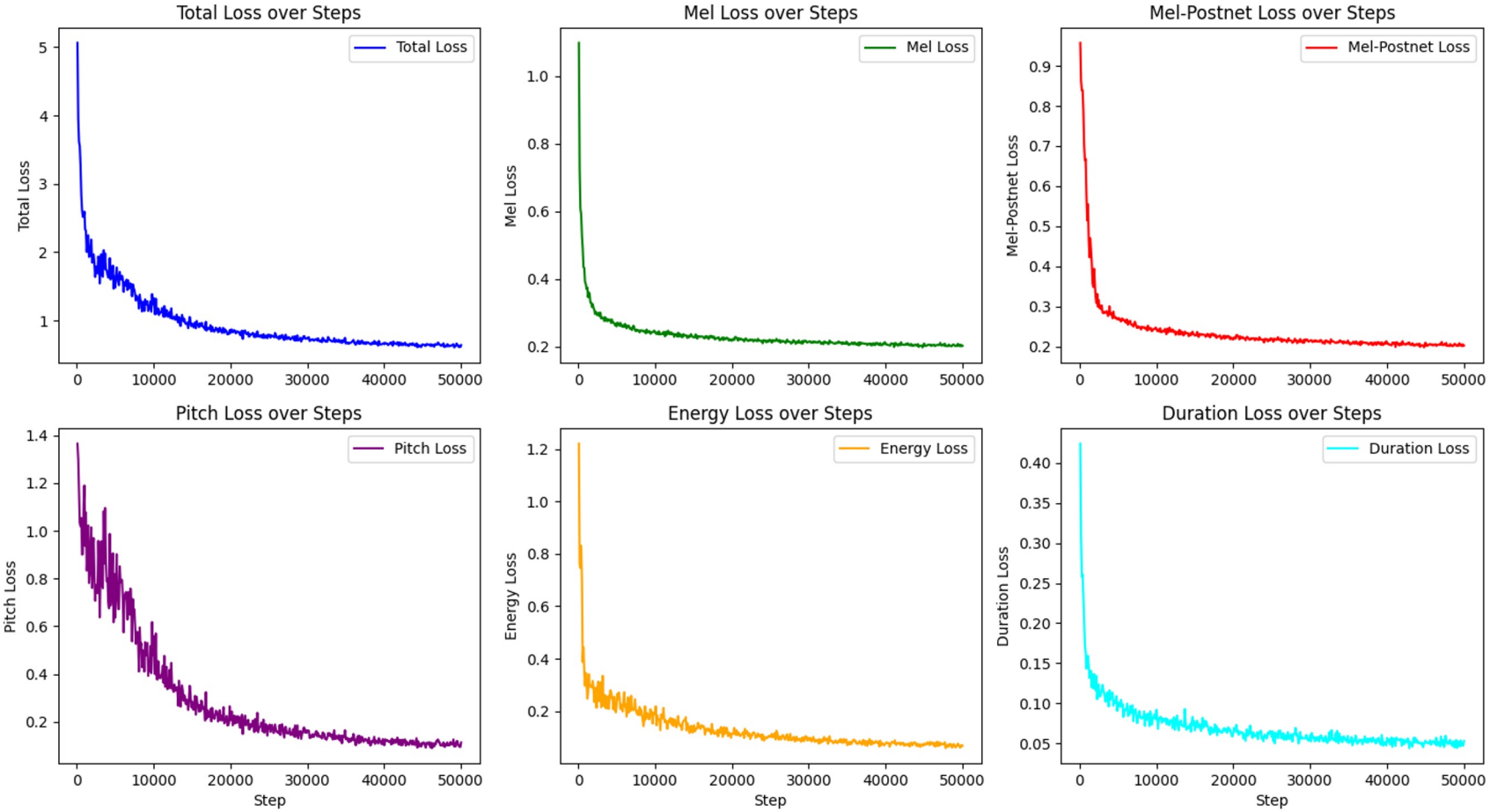


MEL Loss

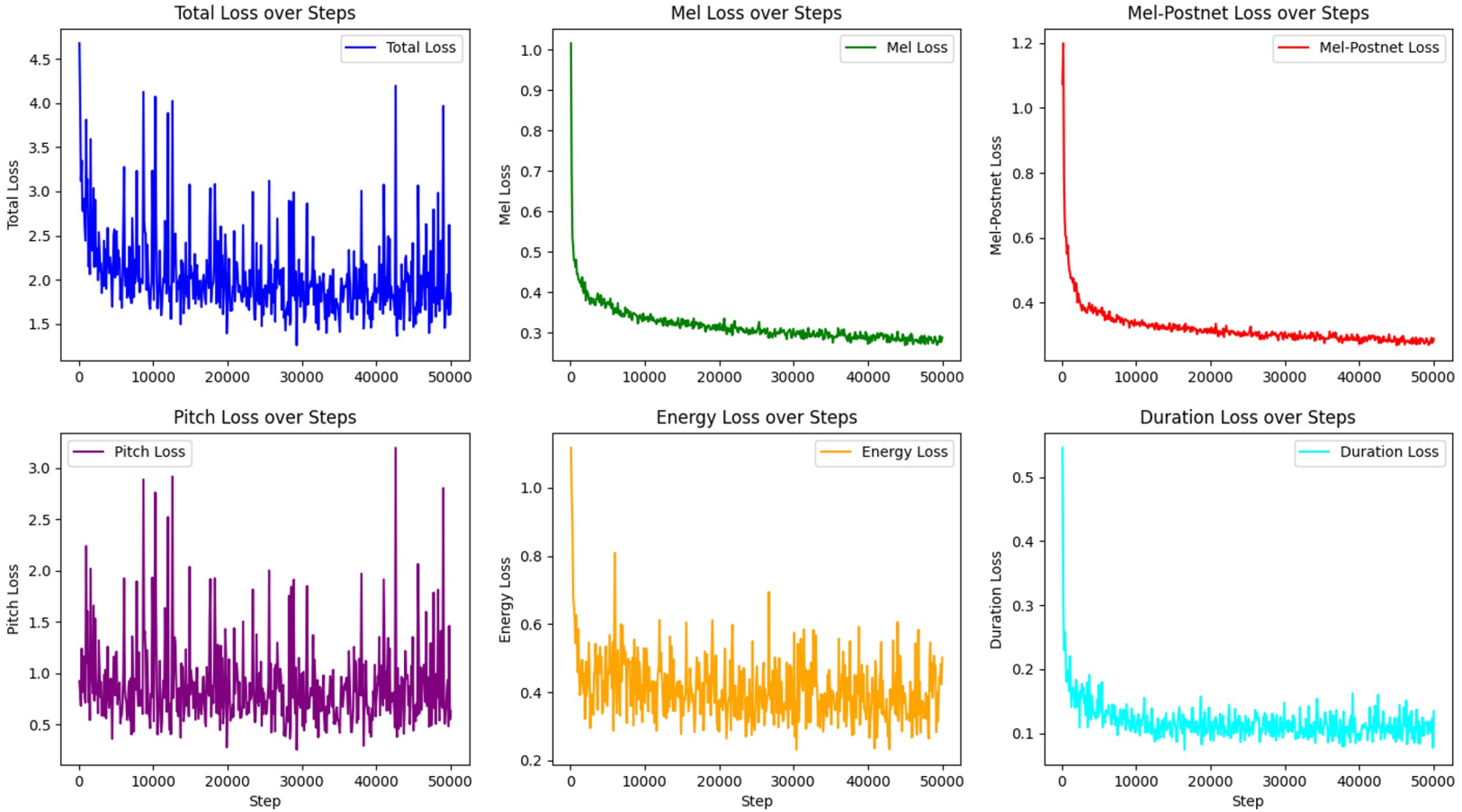


Gate Loss

FastSpeech2 Train on LJSpeech



FastSpeech2 Train on AISHELL



Generating Speed

	Model	Dataset	音檔長度(sec)	生成時間(sec)
	**FastSpeech**	**LJSpeech**	1	10.98
	**FastSpeech**	**AISHELL**	1	8.7
	**Tacotron**	**LJSpeech**	2	8.98

Tongue Twisters

Tongue Twister 1:

Tongue Twister 2:

Thanks!