QI-TTS: QUESTIONING INTONATION CONTROL FOR EMOTIONAL SPEECH SYNTHESIS

Haobin Tang[1,2,#], Xulong Zhang[1,#], Jianzong Wang[1,*] , Ning Cheng[1] , Jing Xiao[1]

[1]Ping An Technology (Shenzhen) Co., Ltd. & [2]University of Science and Technology of China

[#] Indicates equal contribution * Indicates corresponding author

Introduction

Style transfer Emotional TTS utilizes reference audio to specify the desired speech style and its intention is to generate speech that emulates the emotion of the reference audio.

Problem in previous researches:

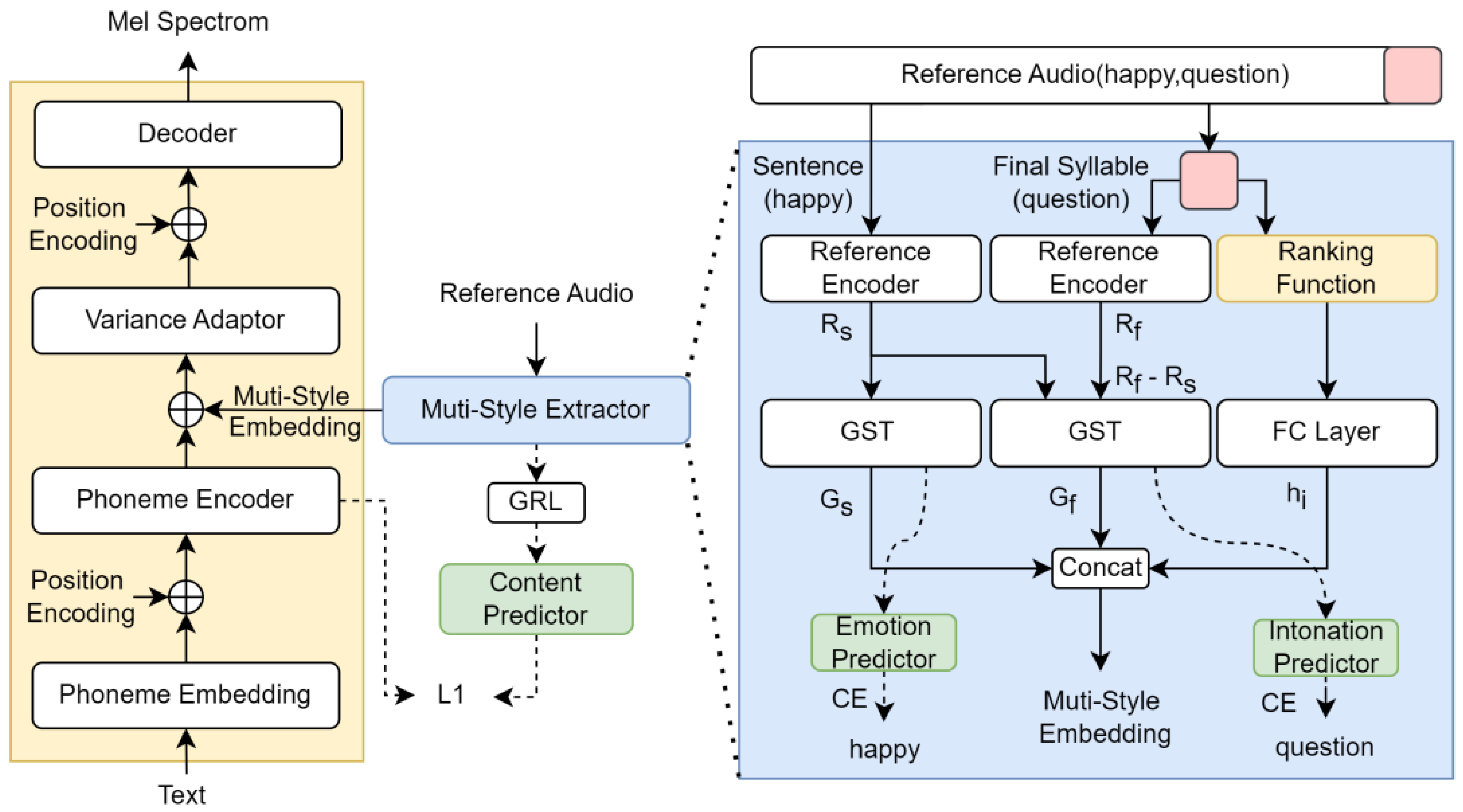
- The existing emotion modeling frameworks only consider the normal statement and lack of the ability to model questions in each emotion;
- The intonation expressions vary in intensity. Thus we desire to flexibly deliver questioning intonation with specific intensity;
- Limitation of the ability to disentangle prosody from other attributes like content, resulting in the quality degrade and expressiveness instability.

Our contributions:

- We proposed QI-TTS which jointly transfer the emotion and intonation from reference audio in an end-to-end way to further delivers the speaker's intention;
- QI-TTS can control the intonation intensity effectively using either manual instructions or reference speech without the use of explicit labels.

Proposed Method

QI-TTS can be mainly divided into two parts based on FastSpeech 2, a multi-style extractor with ranking function, and a content predictor with gradient reversal layer (GRL).



Multi-style extractor:

In addition to modeling the whole sentence scale, we model the last 0.52 seconds of the audio that contains the final syllable as intonation to capture the **duration variance** and **intonation related features.**

Intonation Intensity control:

- A ranking-based method called relative attributes is used for unsupervised intensity modelling.
- Final syllable's acoustic features extracted by the openSMILE are used for calculating intensity.
- After pre-training the ranking function the intensity can be predicted by analyzing the reference audio or assigned a value manually within the interval of [0, 1] at the run time .

Prediction tasks:

- Weighted cross entropy function is used as intonation loss because of the sparse question labels.
- The gradient from content predictor is reversed before backward propagated to the muti-style extractor to minimize the content information contained in the multi-style embedding.

Experiments

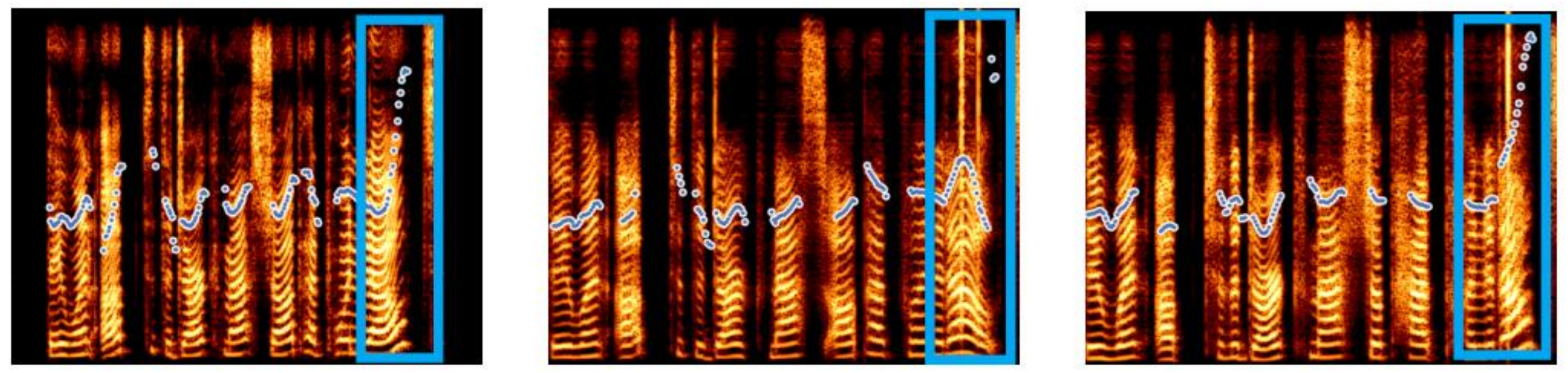
Dataset: Emotional Speech Dataset (ESD)

- English part in 5 emotions
- 10 speakers (5 male and 5 female)
- 310 questions for each speaker on average

Objective and subjective evaluation

Model	MOS ↑	SMOS ↑	Intonation ↑	MCD ↓	FFE ↓	Duration MSE ↓
GT	4.47 ± 0.08	/	/	/	/	/
GTmel + Vocoder	4.40 ± 0.09	4.47 ± 0.10	99.2%	2.40	0.07	0.031
MutiEmo FS2 [20]	3.81 ± 0.08	3.85 ± 0.08	81.6%	**3.15**	0.43	0.144
Styler [21]	3.76 ± 0.08	3.97 ± 0.08	85.9%	5.57	0.41	0.149
QI-TTS	**3.84 ± 0.10**	**4.01 ± 0.08**	**95.2%**	4.89	**0.39**	**0.141**

Ablation study



(a) Ground Truth (b) w/o final syllable (c) QI-TTS

Model	Question	Statement
QI-TTS	/	/
w/o final syllable level	-0.15	-0.09
w/o residual style	-0.08	-0.08
w/o Emotion predictor	-0.10	-0.10
w/o Intonation predictor	-0.11	-0.04
w/o GRL content predictor	-0.08	-0.09

Best-worst scaling test for intensity control

(a) Perception of questioning intonation

Configuration		Best(%)	Worst(%)
Surprise	30% Question	8	79
	60% Question	11	21
	90% Question	81	0
Angry	30% Question	8	69
	60% Question	15	31
	90% Question	77	0

(b) Perception of emotion

Configuration		Best(%)	Worst(%)
Surprise	30% Question	29	39
	60% Question	34	33
	90% Question	37	28
Angry	30% Question	39	28
	60% Question	40	27
	90% Question	21	45

Acknowledgement

Supported by the Key Research and Development Program of Guangdong Province (grant No. 2021B0101400003) and corresponding author is Jianzong Wang (jzwang@188.com).