Emotion Controllable Speech Synthesis Using Emotion-unlabeled Dataset With The Assistance Of Cross-domain Speech Emotion Recognition

Xiong Cai[1], *Dongyang Dai*[1], *Zhiyong Wu*[1,2], *Xiang Li*[1], *Jingbei Li*[1], *Helen Meng*[1,2]

[1] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Shenzhen International Graduate School, Tsinghua University

[2] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong

1. Introduction

1.1 Motivation

- Emotional Text-to-Speech (TTS) synthesis can help TTS systems generate more human-like speech
- A problem: emotion-labeled TTS datasets are usually difficult to obtain
- The field of Speech Emotion Recognition (SER) has many achievements in terms of datasets and approaches
- *Can we build an emotional TTS model on an emotion-unlabeled dataset using the achievements of SER?*

1.2 Contribution

- Propose an emotion controllable TTS with GST-based model
  - ✓ Using **emotion-unlabeled** TTS dataset
  - ✓ Can generate speech with desired emotional expressiveness
  - ✓ Keep nearly the same overall speech quality as neutral TTS systems
- Design three key components to ensure the model works well
  - ✓ Design a **MMD-based cross-domain SER model** to provide effective emotion labels for the TTS dataset
  - ✓ Design an **auxiliary emotion prediction task** to help the GST module learn more emotion-related features
  - ✓ Design a **top-K scheme** to choose a more reliable reference audio set for each emotion category

2. Methodology

2.1 Overall idea

- Step 1: train a cross-domain SER model on the SER and TTS datasets
- Step 2: predict emotion labels for TTS dataset by the trained SER model
- Step 3: train emotional TTS model using the predicted emotion labels

2.2 Overall structure

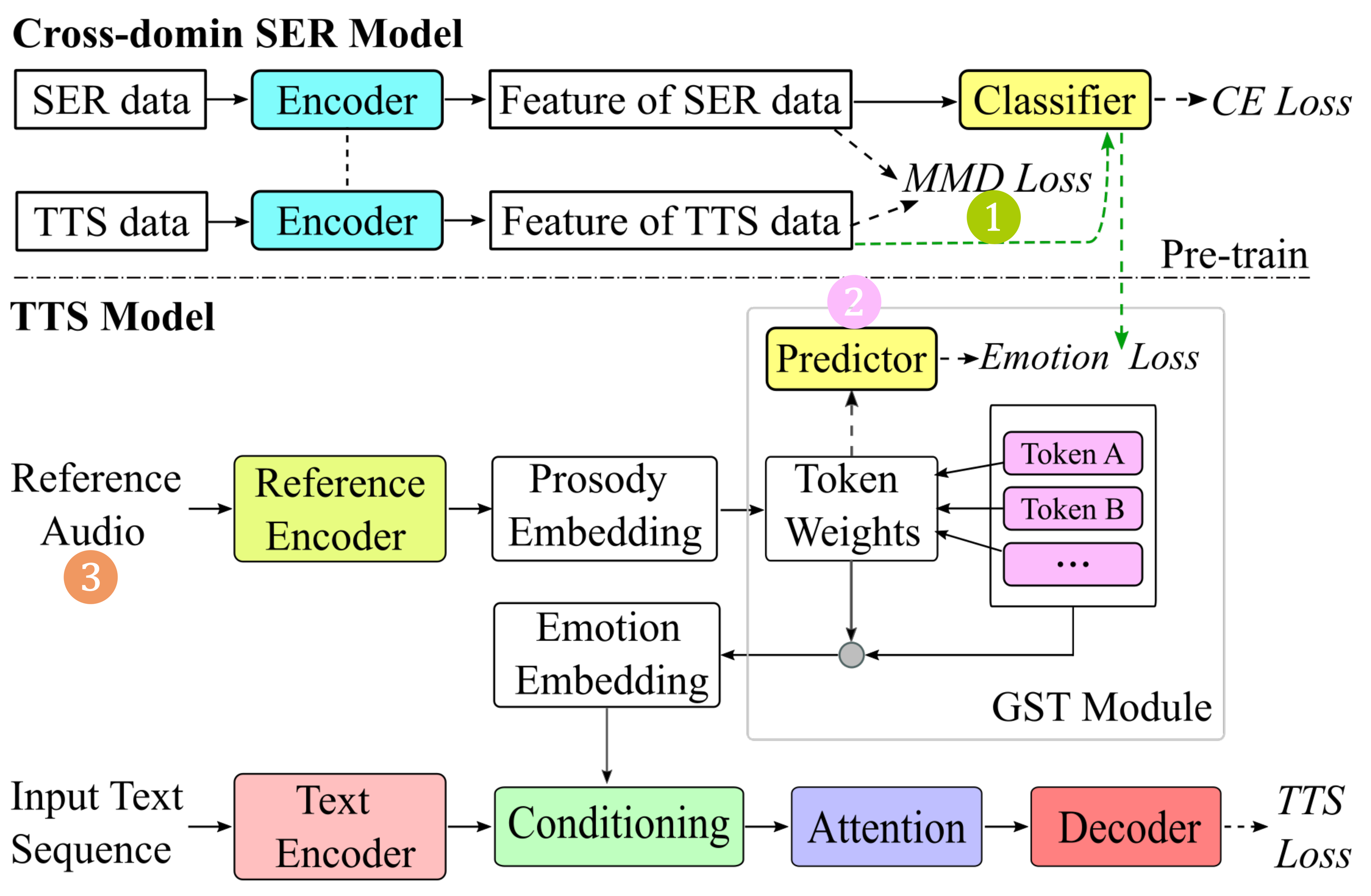


Fig.1. Overall structure of the proposed model which includes a cross-domain SER sub-model and a Tacotron2-GST TTS sub-model

2.3 Cross-domain SER sub-model

- Encoder: 4 Conv-2D layers + 1 bi-GRU layer
- Emotion Classifier: a 2-layer dense network
- Trained on an emotion-labeled SER dataset (source domain) and an emotion-unlabeled TTS dataset (target domain)
- The trained model is used to predict emotion labels for the TTS dataset

2.4 Emotional TTS sub-model

- Consists of a reference encoder, a GST module and a TTS module
- Reference encoder: 6 Conv-2D layers + 1 bi-GRU layer + 1 dense layer
- GST module: 10 style tokens && 4 heads of multi-head self attention
- TTS module: the same as Tacotron2 except we add a CBHG post-net to convert the mel spectrum to linear spectrum
- Vocoder: Griffin-Lim algorithm, to verify the validity of the approach

2.5 The proposed three key components

- 1 **MMD (Maximum Mean Discrepancy) Loss**
  - ✓ Distribution shift between features of SER and TTS data
    - Lead to performance of SER model on TTS data drops significantly
  - ✓ We add a training-stable MMD Loss to reduce this shift

$$L_{MMD} = \frac{1}{m^2}\Sigma_{i=1}^{m}\Sigma_{j=1}^{m}k(\boldsymbol{s}_i,\boldsymbol{s}_j) + \frac{1}{n^2}\Sigma_{i=1}^{n}\Sigma_{j=1}^{n}k(\boldsymbol{t}_i,\boldsymbol{t}_j) - \frac{1}{mn}\Sigma_{i=1}^{m}\Sigma_{j=1}^{n}k(\boldsymbol{s}_i,\boldsymbol{t}_j)$$

Where $\boldsymbol{s}_i$ and $\boldsymbol{t}_i$ are the features of SER and TTS data respectively, and the $k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ is the multiple RBF kernels function.

- 2 **Auxiliary emotion prediction task**
  - ✓ The original GST module
    - Learn the prosody styles from a given reference audio
    - The styles are uncertain since they are learned unsupervisedly
  - ✓ Our auxiliary emotion prediction task
    - Guides GST module to focus on the more emotion-related styles
    - Is a single dense layer that takes as input the style token weights and outputs emotion categories
    - The ground truth labels of this task are the soft labels predict by the trained cross-domain SER model
- 3 **Top-K scheme**
  - ✓ GST-based emotional TTS models
    - Usually select all audios of the same emotion category as the reference audio set to generate this kind of emotion speech
  - ✓ Our proposed model
    - Only have labels predicted by the cross-domain SER model
    - These predicted labels may contain a lot of mispredictions
  - ✓ To get a reliable reference audio set, we propose a *top-K scheme*
    - Only selects the *K* utterances with the highest posterior for each emotion category
    - Greatly reduces the impact of prediction error by the SER model

3. Experiments

3 .1 Datasets

- SER dataset: IEMOCAP
  - ✓ 10,039 utterances about 12.5h
  - ✓ 4 emotion categories: *neutral, happy, angry, sad*
  - ✓ 2 emotion dimensions: *arousal* and *valence*
- TTS dataset: Blizzard Challenge 2013 English
  - ✓ 95k utterances about 73h
  - ✓ An emotion-unlabeled audiobook dataset with rich emotional expressiveness

3.2 Compared models

- **our-4cls:** proposed model for 4 categories
- **our-2d:** proposed model for 2 dimensions
- **base-4cls:** same as **our-4cls** but no auxiliary task
- **full-4cls:** use the same model checkpoint as **our-4cls** but no *top-K scheme*

3.3 Subjective evaluation settings

- 10 text sentences
- 20 postgraduate subjects
- Test1: MOS evaluation for overall speech quality
- Test2: subjective emotion prediction evaluation for the emotion expressiveness

3.4 MOS evaluation results

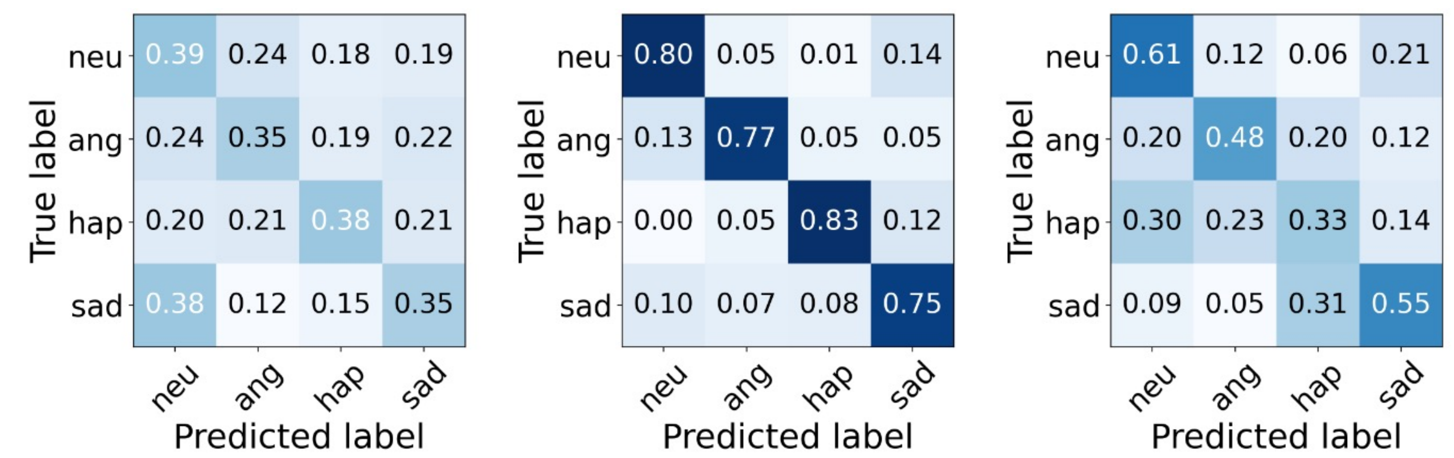
Table 2. MOS of **base-4cls** and **our-4cls** for 4 emotion categories

model	neu	ang	hap	sad	average	p-value
base-4cls	3.90	3.84	3.45	3.74	3.73	–
our-4cls	4.12	3.80	3.11	3.61	3.66	**0.20**

Table 3. MOS of **our-2d** for arousal and valence dimensions.

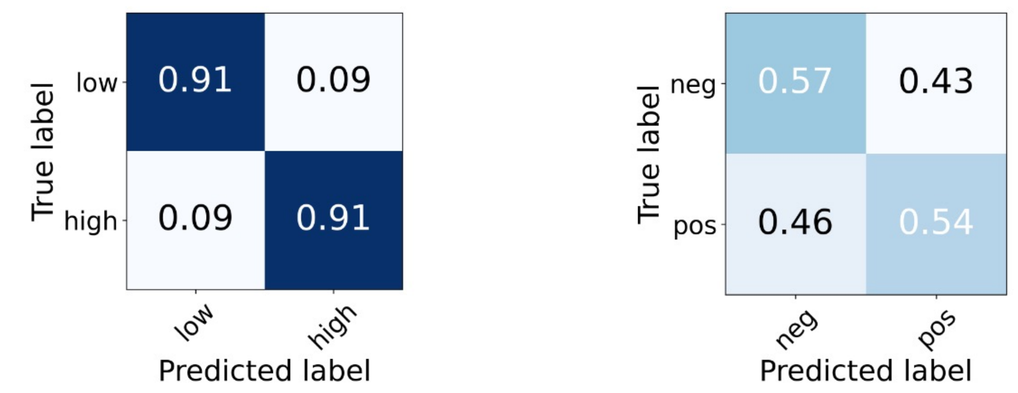
model	low	high	neg	pos	average	p-value
our-2d	3.99	3.33	3.91	3.41	3.66	**0.18**

3.5 Emotion expressiveness evaluation results



(a) **base-4cls** (b) **our-4cls** (c) **full-4cls**

Fig.2. Confusion matrices of 4 emotion categories for the three methods

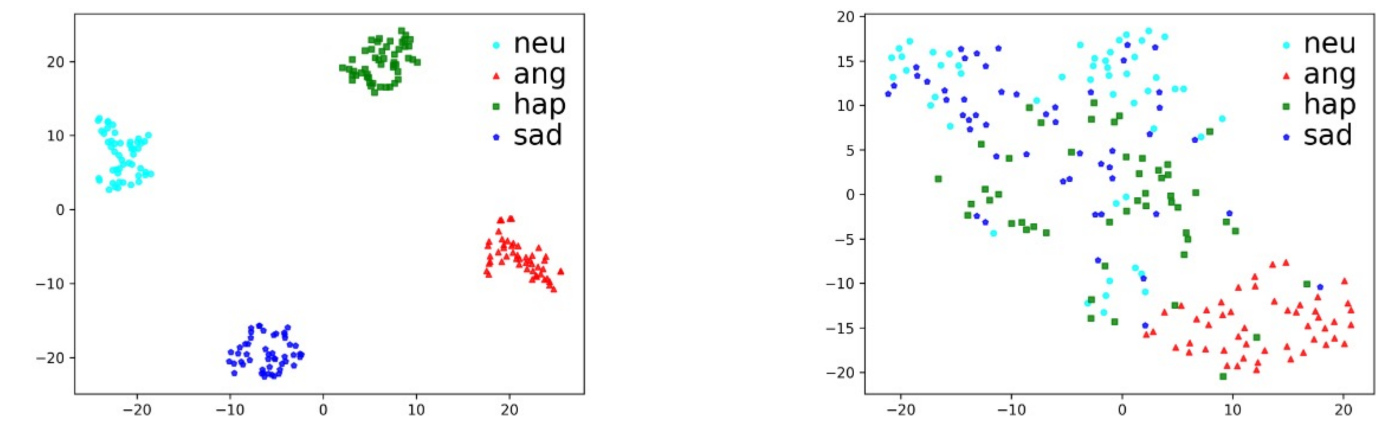


(a) arousal dimension (b) valence dimension

Fig.3. Confusion matrices of polarities of arousal and valence for **our-2d** model

Audio Demos

3.6 Visualization analysis



(a) **our-4cls** (b) **base-4cls**

Fig.4. The t-SNE 2D visualization for style token weights on the reference audio set