A. DETAILS OF METHODOLOGY

A.1. Hyperparameters

We list hyperparameters and configurations of ReActSpeech used in our experiments in Table 4. Note that "ResBlock" in the table denotes to dilated residual convolutional block, which contains a 1D-convolutional layer, ReLU activation and frequency dimension normalization.

Table 4. Hyperparameters of ReActSpeech.

Hyperparameter	ReActSpeech
Embedding Dimension	384
Speaker Embedding Dimension	384
Encoder ResBlocks	13
Encoder ResBlock Kernel	4
Encoder ResBlock Dilations	4*[1,2,4]+[1]
Decoder Residual Blocks	25
Decoder ResBlock Kernel	4
Decoder ResBlock Dilations	6*[1,2,4,8]+[1]
Encoder/Decoder Hidden	384
Encoder/Decoder Dropout	0.2
Encoder/Decoder Activation	ReLU
Encoder/Decoder Normalization	FreqNorm
Predictor Hidden	384
Predictor ResBlocks	3
Predictor ResBlock Kernels	[4,3,1]
Predictor Dropout	0.5
Redistribution Hidden	384
Redistribution Attention Heads	4
Discriminator Layers	12
Discriminator Conv1D Kernel	3
Discriminator Conv1D Filter Size	128
frequency bands	[40,40,40]
Time Window Lengths	[32,64,128]
Number of Parameters	48M

B. EXPERIMENTAL SETTINGS

B.1. Dataset

We utilize an internal multi-speaker mixlingual (Chinese and English) speech dataset, which contains 105,856 audio clips of 12 speakers and corresponding text transcripts, as base dataset to train our base model. And then finetune it on several small target dataset with about 500 audio clips of single speaker (given speaker ID 6 which is specially reserved), respectively. Specifically, we split each target dataset into three parts: 64 samples for testing, 32 samples for validation, and the rest for training. We convert the text sequence into the phoneme sequence with an internal grapheme-to-phoneme tool to alleviate the mispronunciation problem. Importantly, we utilize an internal phoneme set that is slightly different from the IPA (International Phonetic Alphabet) according to the reality of mixlingual (Chinese and English) speech.

B.2. Training and Inference

We train ReActSpeech on 1 NVIDIA V100 GPU, with batch size of 32. We use the Adam optimizer (Kingma & Ba, 2014) with $\beta_1 = 0.9, \beta_2 = 0.98, \varepsilon = 10^{-9}$ and follow the same learning rate schedule in Vaswani et al. (2017). It takes about 20k steps for training (base model) on base dataset until convergence. In the inference process, the output spectrograms of our ReActSpeech are transformed into audio samples using a pretrained HifiGAN (Kong et al., 2020).

C. MORE ANALYSES

C.1. Explanation of RBDR

We utilize a new metric called relative boundary difference rate (RBDR) to measure the alignment accuracy, which is formulated as Equation 1, and a higher value means more accurate. Importantly, RBDR reaches 1.0 only when the phoneme boundaries are strictly aligned that both begin and end boundaries of phoneme are exactly the same as the ground-truth. Moreover, it is entirely possible for RBDR to be negative, which means that the phoneme boundaries are completely misaligned with the ground-truth.

C.2. More Visualizations

We plot more marginal and joint distributions (along time and frequency) of spectrograms in Figure 7, 8, 9 (ground-truth), and Figure 10, 11, 12, 13 (generated by different methods).

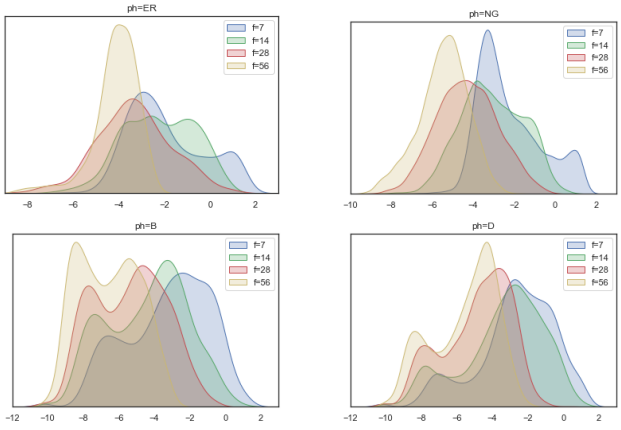


Fig. 7. More marginal distributions $P(y(t, f)|x = ph)$ for different phonemes ph.

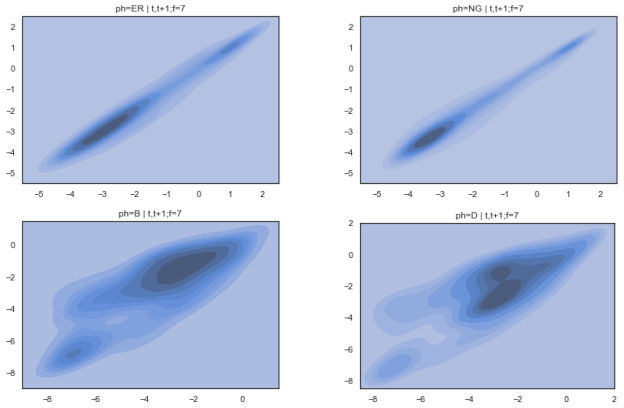


Fig. 8. More joint distributions $P(y(t, f1), y(t, f2)|x = ph)$ along time dimensions.

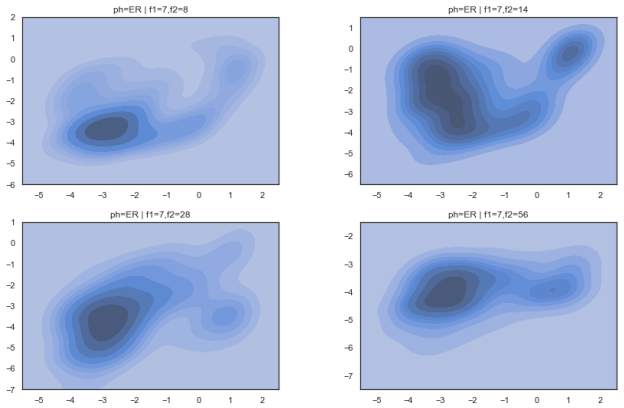


Fig. 9. More joint distributions $P(y(t, f1), y(t, f2)|x = ph)$ along frequency dimensions.

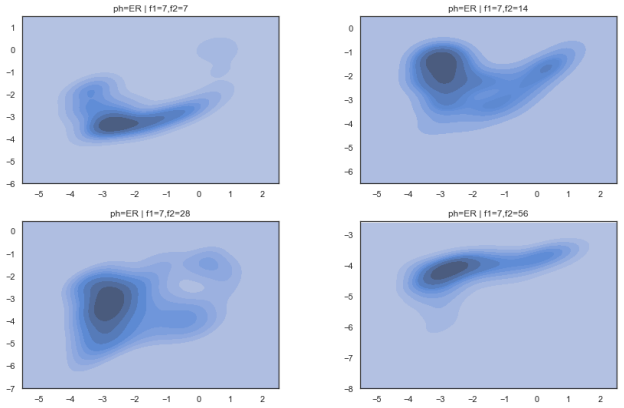


Fig. 10. The joint distributions calculated from spectrograms generated by ReActSpeech (v1). Compare with Figure 9.

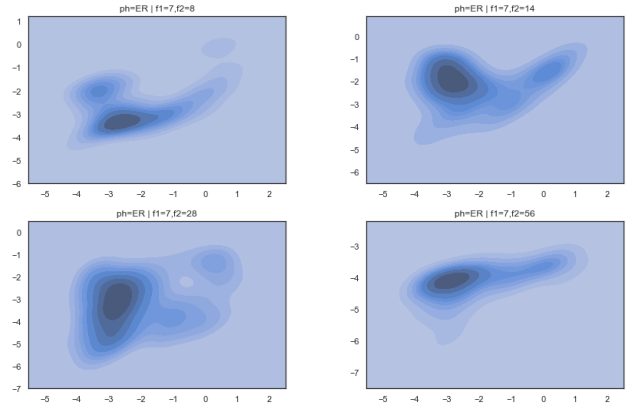


Fig. 11. The joint distributions calculated from spectrograms generated by ReActSpeech (v2). Compare with Figure 9.

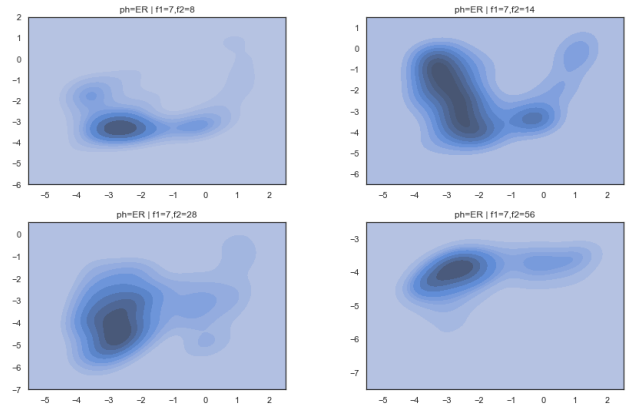


Fig. 12. The joint distributions calculated from spectrograms generated by ReActSpeech (v3). Compare with Figure 9.

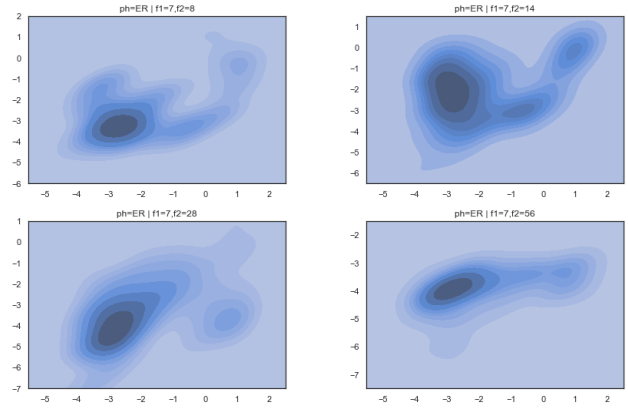
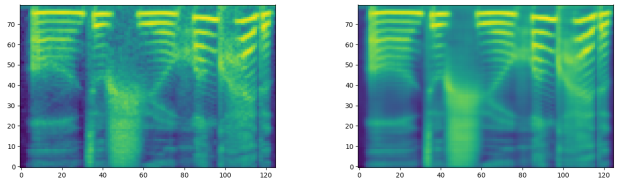


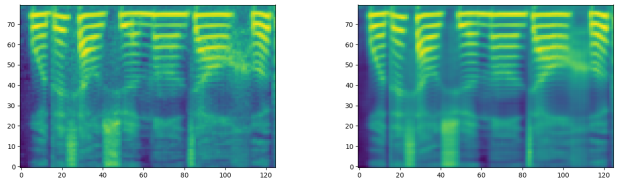
Fig. 13. The joint distributions calculated from spectrograms generated by ReActSpeech (v4). Compare with Figure 9.



(a) Ground-Truth (b) Tacotron 2

(c) FastSpeech 2 (d) ReActSpeech (v4)

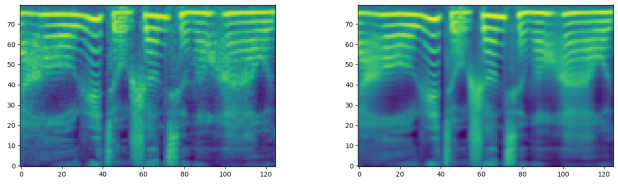
Fig. 14. Visualizations of the ground-truth and generated spectrograms by different methods.



(a) Ground-Truth (b) Tacotron 2

(c) FastSpeech 2 (d) ReActSpeech (v4)

Fig. 15. Visualizations of the ground-truth and generated spectrograms by different models.



(a) Ground-Truth (b) Tacotron 2

(c) FastSpeech 2 (d) ReActSpeech (v4)

Fig. 16. Visualizations of the ground-truth and generated spectrograms by different models.

Note that: (1) ReActSpeech (v1) denotes to raw ReActSpeech **without** redistribution module & iterative decoder (revisor) & multiple frequency bands discriminator; (2) ReActSpeech (v2) denotes to ReActSpeech (v1) + redistribution module; (3) ReActSpeech (v3) denotes to ReActSpeech (v2) + iterative decoder (revisor); (4) ReActSpeech (v4) denotes to ReActSpeech (v3) + multiple frequency bands discriminator with GAN training pipeline. And **ReActSpeech (v4)** is our final model.

In addition, we also plot several spectrograms generated by different models (while training) to compare with the corresponding ground-truth in Figure 14, 15, and 16. Obviously, the spectrograms generated by ReActSpeech (v4) have much more details than those generated by Tacotron 2 and FastSpeech 2, which further demonstrates that ReActSpeech (v4) has more powerful capability to fit complex data distributions.

C.3. More Evaluation

To further demonstrate the effectiveness of our model, we conduct several other evaluations (spectrogram).

Dynamic Time Warping Insted of MAE and MSE, we utilize the Dynamic Time Warping (DTW) (Springer et al., 2007), which can capture nonlinear correspondence of sequences, to evaluate the quality of spectrogram.

Structural Similarity Index We also utilize Structural Similarity Index (SSIM) (Wang et al., 2004), which can capture structural information and texture, to evaluate the quality of spectrogram. Notably, we provided ground-truth durations while inference to ensure consistent size of spectrograms ($*$).

Variation of the Laplacian The same with the research by Ren et al. (2022) [18], we employed the Variation of the Laplacian (VoL) (in terms of Var_L) to measure the amount of blur in a spectrogram. Higher is better.

Degree of Multimodality Moreover, we conduct Hartigan's dip test (Hartigan et al., 1985) on the marginal distributions $P(y(t,f)|x = ph)$ to measure the degree of multimodality (in terms of $\overline{\mathcal{D}}$) of each model. Lower is better.

Table 5. Results of other evaluations among different models. The best scores are in bold.

Model	DTW	$SSIM^*$	Var_L	$\overline{\mathcal{D}}$
GT (Mel)	0.000	1.000	0.580	0.049
Tacotron 2 [2]	1.672	0.413	0.336	0.061
FastSpeech 2 [5]	1.126	0.787	0.386	0.058
ReActSpeech	**0.694**	**0.838**	**0.510**	**0.054**

D. ORIGIN OF NAME

Redistribution Module and **Re**visor for **A**uto-**c**orrec**t**ing (to be **ReAct**).