1.设定好多个其他的style，经过style的叠加来合成最终，那么网络分为两部分，style部分，正常的text内容+声学参数部分，两部分需要叠加来合成最终。

2.如果每部分都训练好了，那么一个无监督style的就要枚举style，来找loss最小，但是都不确定，所以不好训练。

3.用encoder-decoder模型，先脱离掉style，然后看怎么和2关联，或者用强化学习能不能行。

4. [6] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, “Adapting

and controlling DNN-based speech synthesis using input codes,”

in Proc. ICASSP, 2017, pp. 4905–4909.

[7] S. Xue, O. Abdel-Hamid, H. Jiang, L.-R. Dai, and Q. Liu, “Fast

adaptation of deep neural network based on discriminant codes for

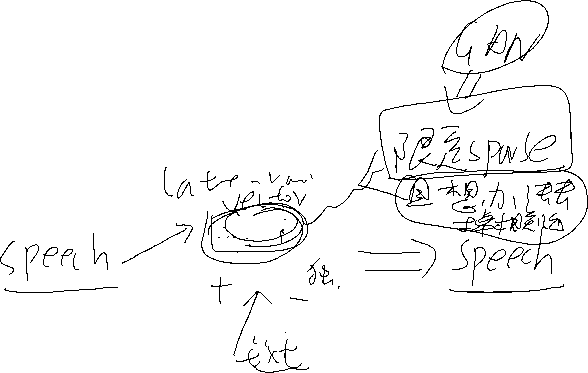
speech recognition,” IEEE/ACM T. Audio Speech, vol. 22, no. 12,

pp. 1713–1725, 2014.

[8] O. Watts, Z. Wu, and S. King, “Sentence-level control vectors

for deep neural network speech synthesis,” in Proc. Interspeech,

2015, pp. 2217–2221



5.我的掷筛子思想，然后消除。

6.情感识别对抗网络

7.style是公共的k个向量，每次依赖于阿尔法的选择，还是每个sample都有自己的，做为额外的输入。

F:/我的坚果云/研一-语音基础/follow比赛代码/follow\_geek\_tactron

8.用中文大学的那个数据集，情感丰富，数据量大

9.详细讨论下style token和阿尔法选择的学习过程。