Experiment 2

Experiment details

1. Train data : WSJ/train_si284 (add noise with probability 0.3 , 37318 utterances in total, about 80 h)
2. Resample strategy : evaluate the WER of each utterance, and resample the training data using WER in sentence level. Keep the amount of utterances to be same.
3. Ensemble strategy : assemble the model after softmax layer. The weight for each model is calculated base on the test set's WER.

basic results:

	1280*6	256*6 (1)	256*6 (2)	256*6 (3)	256*6(4)	256*6(5)
bd_tgpr_dev93	15.97	19.38	27.52	29.04		
bd_tgpr_dev93_fg	14.45	17.59	25.49	27.69		
bd_tgpr_eval92	13.20	16.89	22.24	23.57		
bd_tgpr_eval92_fg	11.89	15.59	20.26	22.08		
tg_dev93	17.40	20.03	28.11	30.58		
tg_eval92	14.05	17.92	23.25	24.61		
tgpr_dev93	18.00	21.11	28.59	31.33		
tgpr_eval92	14.78	19.01	23.91	25.64		
train set		19.83	46.87			

Some consistent conclusion :

1. When training data is selected to be harder , model tends to perform worse.
2. Model performs bad on resampled training data but performs relatively good on test set.

Ensemble result:

	256*6 (1)	**256*6 (2)**	**256*6(2)_(1)**	**Ensemble(1)_(2)** weight1: 0.75595 weight2: 0.24405	**Ensemble(1)_(2)** weight1: 0.5647 weight2: 0.4353
bd_tgpr_dev93	19.38	27.52	53.50	20.66	27.41
bd_tgpr_dev93_fg	17.59	25.49	52.87	19.04	26.12
bd_tgpr_eval92	16.89	22.24	44.16	17.77	19.71
bd_tgpr_eval92_fg	15.59	20.26	43.45	16.23	18.36
tg_dev93	20.03	28.11	55.00	21.54	28.39
tg_eval92	17.92	23.25	45.95	18.18	20.95
tgpr_dev93	21.11	28.59	54.48	22.74	29.04
tgpr_eval92	19.01	23.91	45.47	19.19	21.76

256*6(2)_(1) means model(2) with model(1)'s prior. It performs much more worse.

Some thoughts:

- The ensemble model uses the prior of model 1, maybe it has a bad influence on the result.

- model 1 performs better than model 2. When assemble them, we can get better result if model 1's weight is higher. But it's interesting that some of the result of **Ensemble(1)_(2) (weight1: 0.5647 weight2: 0.4353)** is worse than model 2 (the worse one).