Speechflow with random embedding replacement cycle loss

I designed a speechflow with random embedding replacement cycle loss. In this model, we introduce another part into training loss, which measures the loss between directly obtained embeddings and reconstructed embeddings calculated with reconstructed spectrograms. Otherwise, for this part, we randomly select an embedding from content, rhythm and pitch, the selected embedding will be replaced with which of another frame. Then, the new joint embeddings will be sent to the decoder to construct a spectrogram with characters of two frames. Finally, the obtained spectrogram will be sent to the encoders again to calculate the loss between the new embeddings and the original joint embeddings. By this method, maybe we can improve the ability of reconstruction and decrease mutual information between embeddings. The whole loss can be defined as the following equation:

$$Loss = ||S - \hat{S}|| + ||Z' - \hat{Z}'||$$
$$Z_c = E_c(A(S))$$
$$Z_r = E_r(S)$$
$$Z_f = E_f(A(P))$$
$$Z' \in \{(Z'_c, Z_r, Z_f, Z_u), (Z_c, Z'_r, Z_f, Z_u), (Z_c, Z_r, Z'_f, Z_u)\}$$

Some results are as follows:

Mutual information with 10-classes cluster:

original speechflow

	Spectrogram	Content	Rhythm	Pitch
Spectrogram		0.6547	0.5550	0.2337
Content			0.5204	0.2566
Rhythm				0.2856
Pitch				

improved speechflow

	Spectrogram	Content	Rhythm	Pitch
Spectrogram		0.5321	0.5015	0.1670
Content			0.3587	0.1243
Rhythm				0.2103
Pitch				

It seems that there are less mutual information between embeddings in the improved speechflow. For original speechflow, mutual information between rhythm and content is badly 0.5204, while for improved speechflow, it's much better 0.3587.

Results of emotion recognition

original speechflow

No.	**Factors**			**Test Sets**	
	Content	Rhythm	Pitch	IEMOCAP	SAVEE
1	-	-	-	59.08	45.00
2	√	√	√	57.50	47.08
3	×	×	×	25.00	25.00
4	√	×	×	50.77	42.08
5	×	√	×	54.12	36.67
6	×	×	√	45.73	32.71
7	√	√	×	56.45	42.71
8	√	×	√	51.91	37.08
9	×	√	√	56.22	36.04

improved speechflow

No.	**Factors**			**Test Sets**	
	Content	Rhythm	Pitch	IEMOCAP	SAVEE
1	-	-	-	59.08	45.00
2	√	√	√	58.52	48.12
3	×	×	×	25.00	25.00
4	√	×	×	47.41	44.17
5	×	√	×	61.88	33.33
6	×	×	√	45.92	28.75
7	√	√	×	59.82	39.58
8	√	×	√	50.50	43.12
9	×	√	√	57.45	35.21

For No.2, results of improved speechflow are better than original speechflow, which might prove that our model could perform better on reconstruction.

For other results, rhythm performs better on in-corpus emotion recognition, while content performs better on cross-corpus emotion recognition.