BiWEEKLY REPORT

Text-independent unsupervised speaker recognition Model

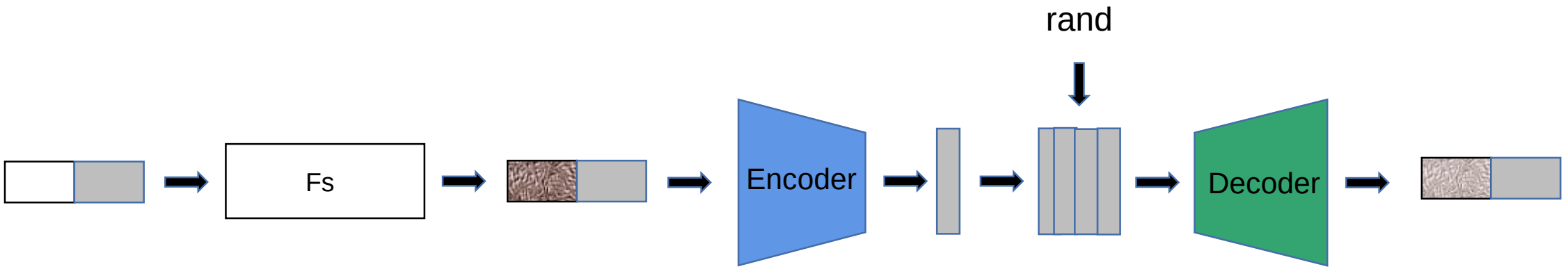
汇报人：林琬

时间日期：2022.7.22

- Motivation

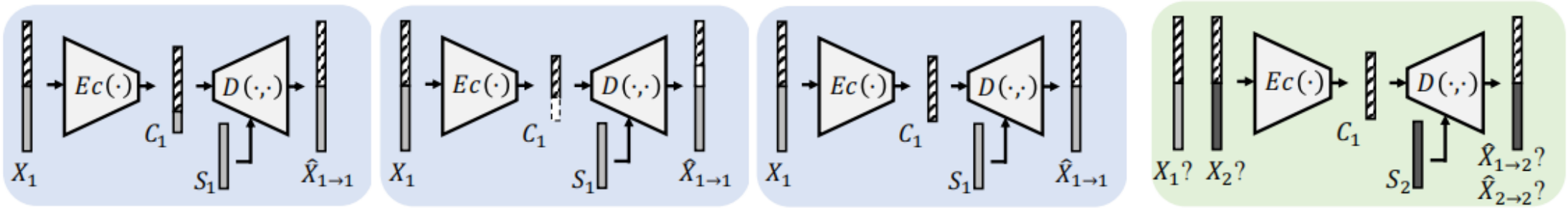
  Text-independent unsupervised speaker recognition

- Model



$$\text{loss} = \| x_{\text{in}} - x_{\text{rec}} \|^2$$

- Information bottlenecks



(a) Bottleneck too wide (b) Bottleneck too narrow (c) Bottleneck just right (d) Conversion

- Probability distribution

speaker: U frame: $p_x(\cdot|U)$

- Dataset:VCTK

  VCTK Corpus includes speech data uttered by 109 native speakers of English with various accents. Each speaker reads out about 400 sentences, most of which were selected from a newspaper plus the Rainbow Passage and an elicitation paragraph intended to identify the speaker's accent.

  train_data:p225-p345(100)　　%3!=0

  test_data:p347-p376(9)　　%3==0

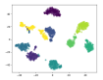
  each data duration:1~3s

validation

- Dataset:VCTK

Iteration：1000
acc(<50%):0.465
acc(<10%):0.212

……

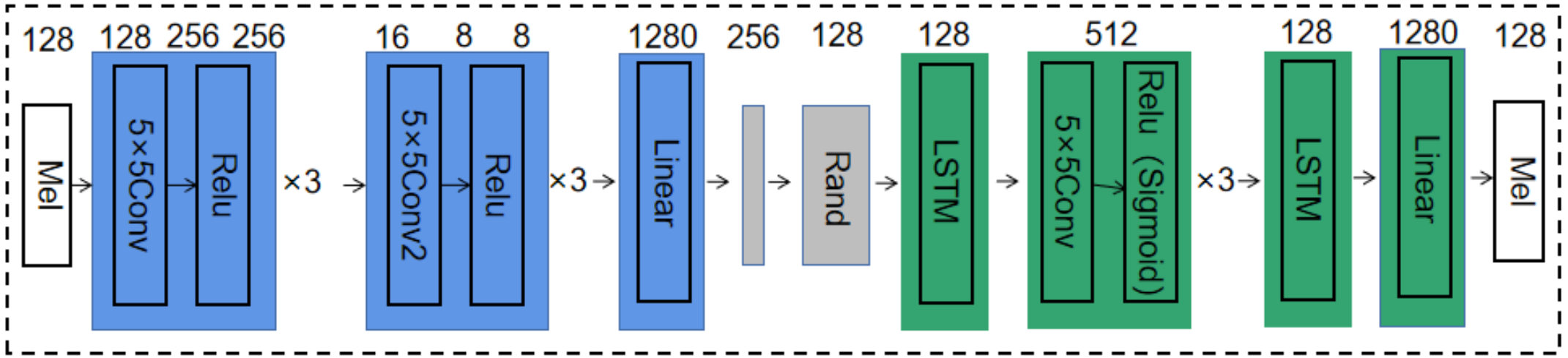


Iteration：10000
acc(<50%):0.856
acc(<10%):0.727

The main factors of model

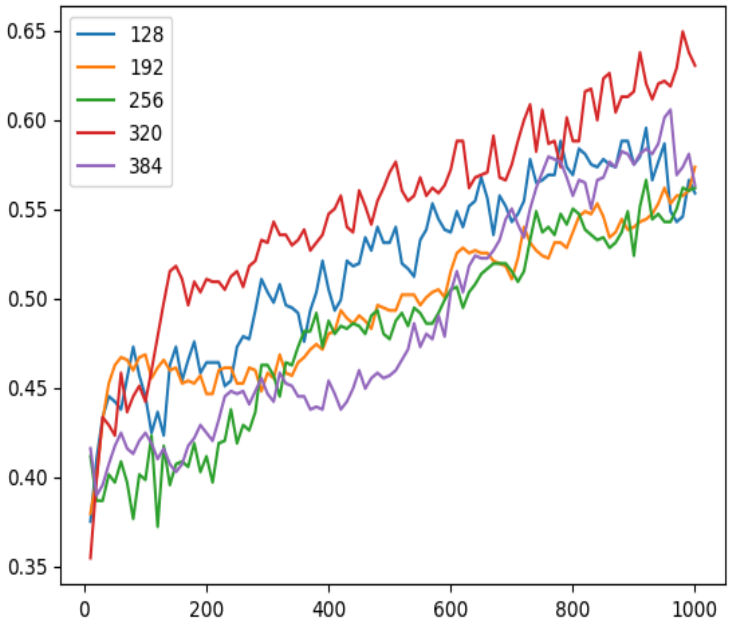
- Hidden layer architecture
- Appropriate loss function
- Encoding size
- Appropriate batch size

Hidden layer architecture



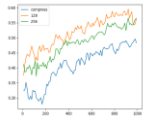
Encoding size

Only change the size of code



accuracy

Change the encoding size



accuracy

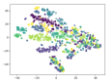
Limitations of the model

- Highly dependent on training duration
- Sensitive to the environment of the voice

Noisy_Dataset(Noisy speech database for training speech enhancement algorithms and TTS models)

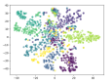
Not-Finetune

acc(<50%):0.725
acc(<10%):0.772

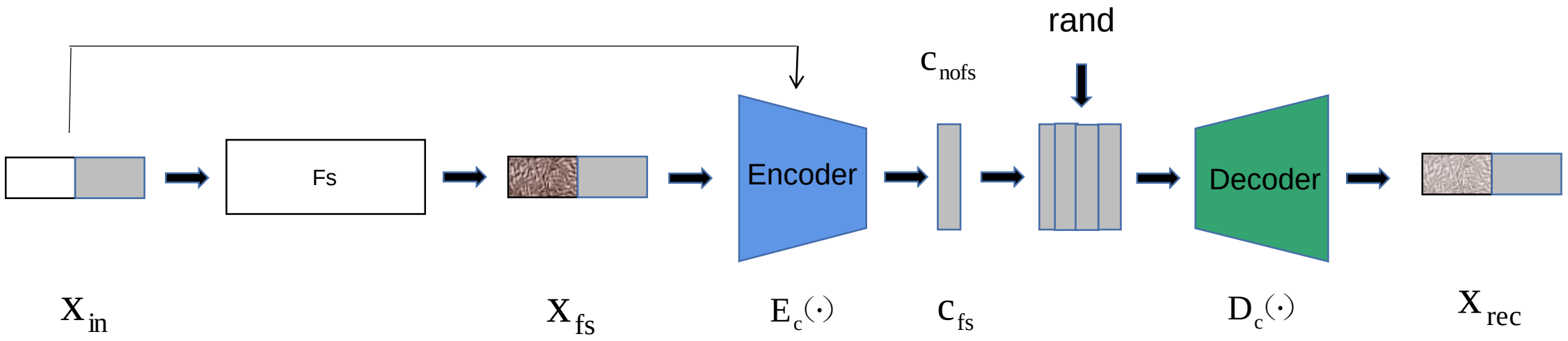


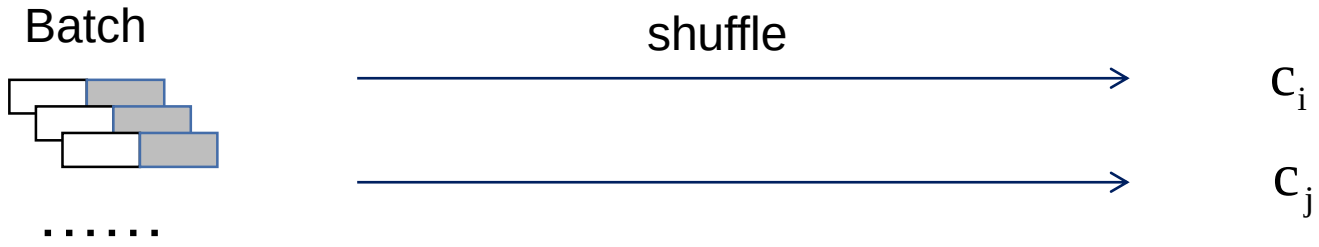
Finetune

acc(<50%):0.836
acc(<10%):0.772



Improvement of loss function



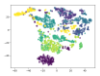


$$loss = \| x_{in} - x_{rec} \|^2 + \lambda_1(1 - \cos(c_{nofs}, c_{fs})) + \lambda_2 \cos(c_i, c_j)$$

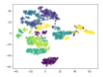
Predict results

- Dataset:VCTK

Iteration：1000
acc(<50%):0.667
acc(<10%):0.475



Iteration：2000
acc(<50%):0.769
acc(<10%):0.576



Iteration：3000
acc(<50%):0.824
acc(<10%):0.624

……

Iteration：10000
acc(<50%):0.985
acc(<10%):0.866

Results Comparison

$$loss = \| x_{in} - x_{rec} \|^2 + \lambda_1(1 - \cos(c_{nofs}, c_{fs})) + \lambda_2 \cos(c_i, c_j)$$

Iteration：1000
acc(<50%):0.667
acc(<10%):0.475

Iteration：10000
acc(<50%):0.985
acc(<10%):0.866

$$loss = \| x_{in} - x_{rec} \|^2$$

Iteration：1000
acc(<50%):0.465
acc(<10%):0.212

Iteration：10000
acc(<50%):0.856
acc(<10%):0.727

- Conclusion
  - The Autoencoders based on lstm has the ability to identify the speaker
  - A loss function closer to the target can effectively improve the performance and efficiency of the model
- Possible Trying
  - Combined with a Speech Enhancement Model
  - Modify the network structure or loss function so that the model can adapt to different noise environments