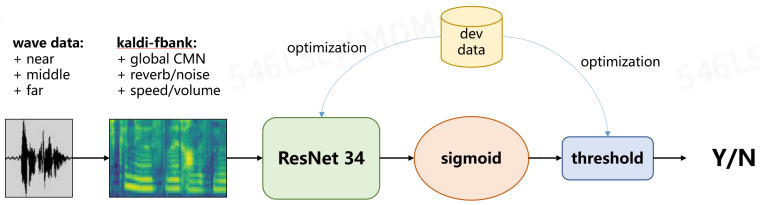
MISP2021 track1

1. Audio Scheme

a. mono-channel pretrained model

【Pipeline】



【Input】

- We collect all channel-level samples for training
- We make data augment firstly, then extract 40-dim Fbank feature by kaldi, and adopt global CMN finally.
- We fix the spectrogram into 300x40x1 shape as input.

【Method】

- We use 34-layers ResNet to learn the representation in TensorFlow.
- We choose sigmoid activation with CE target for training

【Post Processing】

- The best threshold is obtained according to the score in DEV dataset.

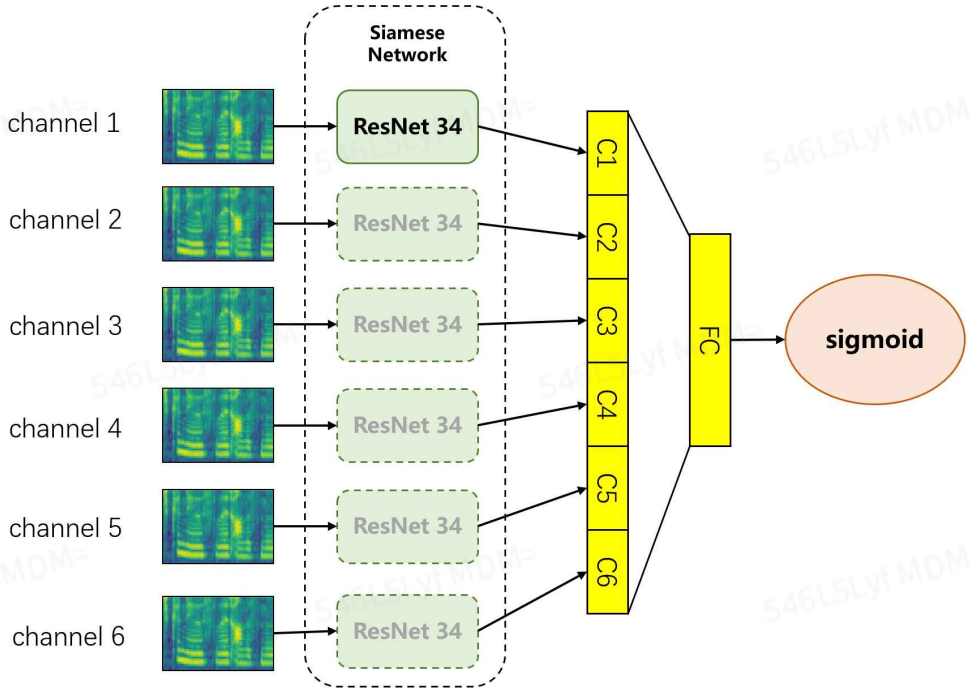
【Result】

- Here, we calculate the average mono scores of one utterance.
- dev : 0.12

b. Multi-channels model

- On the basic of mono-channel model, we further propose a multi-channel model to support microphone array.

【Pipeline】



【Method】

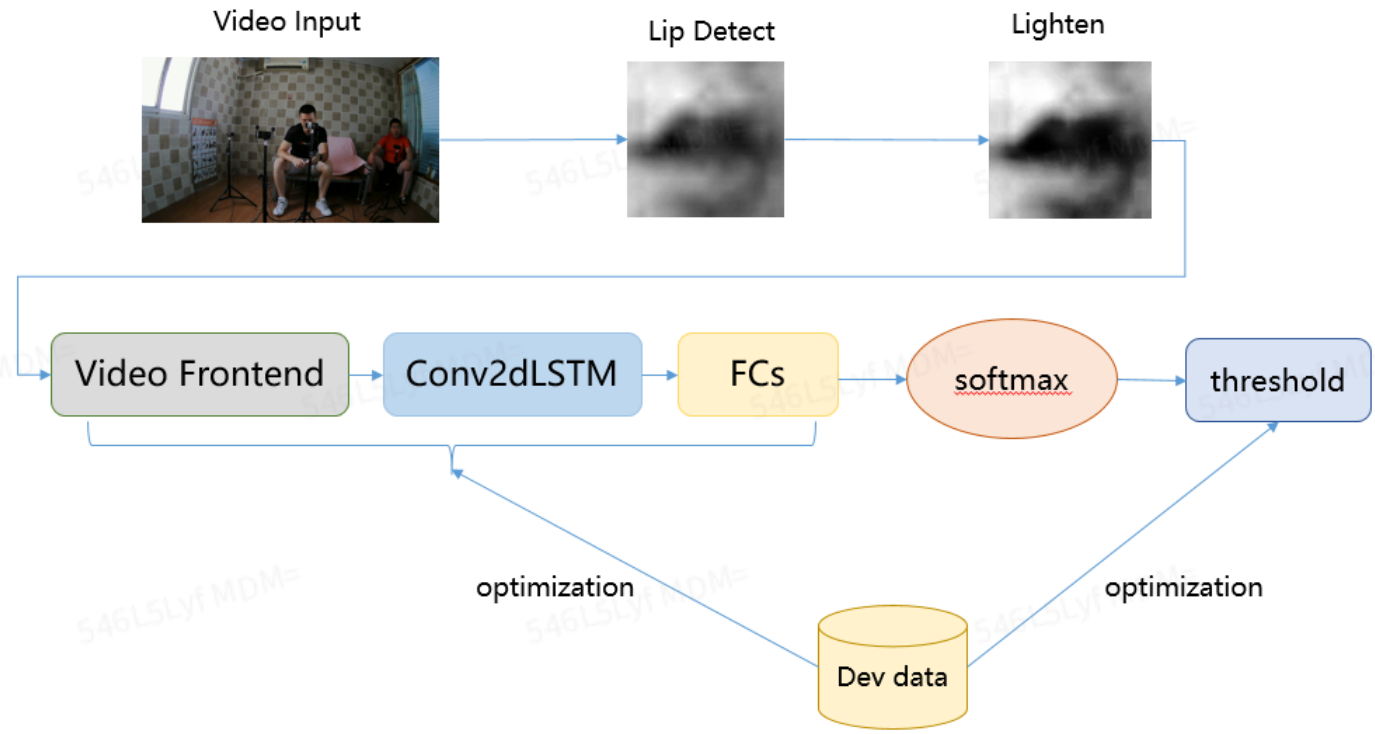
- Building 6-channels siamese network, where the weights of ResNet are initialized by the mono-channel network.
- Getting the corresponding embedding vector for each input using the shared ResNet, and concatenating into the one.
- Training and giving the utterance-level score by sigmoid.

【Result】

- dev : 0.11

2. Video Scheme

【Pipeline】



【Input】

- A series of grey-scale map about lip from the video

【Method】

- Model
  - Video Frontend:
    - extract the coarse-grained feature of the lip from the video
  - Conv2dLSTM:
    - model the spatio-temporal information of the lip movement and get salient feature for classifier
  - FCs:
    - reduce the dimension of feature and classify
- Loss function
  - Class imbalance
    - assign weights for positive and negative class, i.e. the positive weight is the ratio of the number of negative samples and positive samples
  - Label smoothing
    - apply label smoothing to avoid over-fitting and improve the ability of the model to adapt.

【Post Processing】

- Lighten: Use Histogram equalization to lighten the lip input.

- Use grid search to find the best threshold.
  - Supposed threshold is α and the model output is $[p_{out}, n_{out}]$ both of which are not activated by softmax, we apply $+\alpha$ and $-\alpha$ on positive and negative score separately, and then get the prediction based on softmax result.
  - Considering the equivariance of the argmax with softmax, α can be smaller than 0.

【Result】

- dev: 0.4136

3. Fusion Scheme

【Pipeline】

- Through the above two models,we get dev's video(V) and audio(A) scores
- Two parameters in below formula:alpha represents the weight coefficient of V relative to A;t represents different threshold settings which determines whether to be waked up
- We use grid search to find the best combination

$$\hat{\alpha}, \hat{t} = \underset{(\alpha,t)}{\operatorname{argmax}} \; Score(\frac{A+\alpha V}{1+\alpha}, t)$$

【Input】

- dev's video(V) and audio(A) scores

【Result】

In test,we use the two parameters and get results:

- dev: 0.055
- test: 0.108

Team: "XiaoMi No1"