The OPPO System for CHiME-6 Challenge

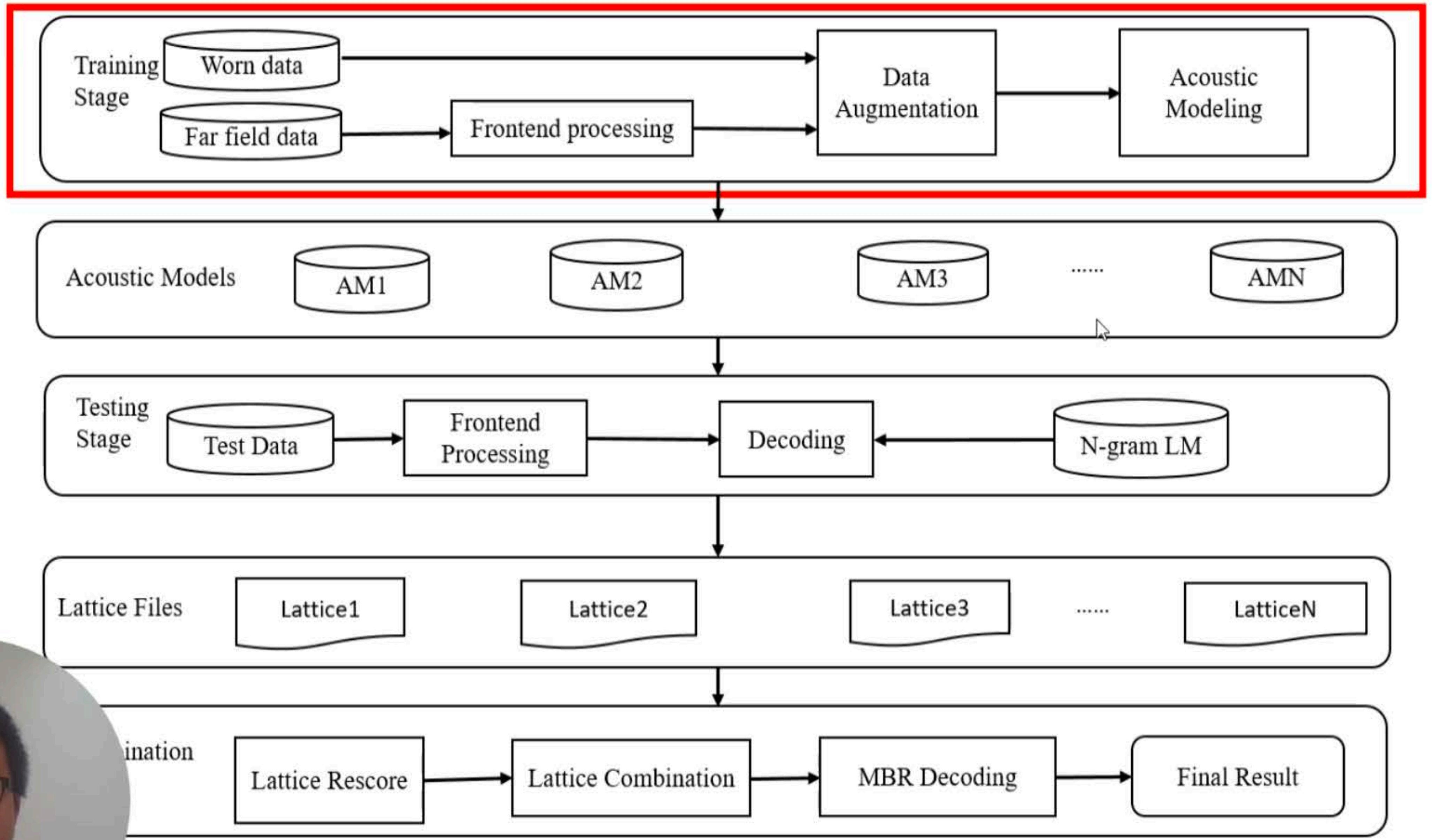
Xiaoming Ren, Huifeng Zhu, Liuwei Wei, Linju Yang, Ming Yu, Chenxing Li, Dong Wei, Jie Hao

2020-04-28

O1_3_video.webm

PowerPoint 幻灯片放映 - [CHiME6_OPPO.pptx] - PowerPoint

System Framework for Track1(ASR only)



1

recordscreen.io 正在共享窗口。 停止共享 隐藏

00:14

System Description

- Data Preparation
- Frontend processing
- Acoustic modeling
- Language modeling
- Decoding

O1_3_video.webm

PowerPoint 幻灯片放映 - [CHiME6_OPPO.pptx] - PowerPoint

System Description

- Data Preparation
  - ✓ For the worn(L+R) microphone training data, realign original utterance segmentation using ASR model
  - ✓ Clean up the training data by filtering out segments which are less than 1 second
  - ✓ Remove noises which can be recognized as words from noises used in Room Impuse Responses(RIR) [1] convolution
  - ✓ Apply only speed perturbation for the training data without the volume perturbation

  Finally, we obtain about 1400 hours of training data, which contains the following dataset
  - ➢ The realigned worn (L+R) training data
  - ➢ The far field enhanced by GSS module
  - ➢ The worn data and enhanced far field data both convolved with RIRs
  - ➢ The augmented previous dataset by speed perturbation

Ko, V. Peddinti, D. Povey, M. L. Seltzer, S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” 2017 IEEE ernational Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5220-5224, 2017

recordscreen.io 正在共享窗口。 停止共享 隐藏

幻灯片 第 6 张

02:34

System Description

- Frontend processing
  - ✓ Compared to the official baseline setup, apply the GSS[2] not only in testing stage but also in training stage.
- Acoustic modeling
  - ✓ Use two different kinds of acoustic model structures (CNN-TDNN-F and TDNN-F[3]) based on LF-MMI training
  - ✓ Train 8 acoustic models with different parameters using Kaldi[4] toolkit

...leker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in ...5, pp35-40,2018.

...G. Cheng, Y.Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in Proceedings ...nual Conference of the International Speech Communication Association (INTERSPEECH), Hyderabad, India, Sep. 2018.

...y, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi ...cognition toolkit," in Proceedings IEEE Automatic Speech Re... ...hop (ASRU), 2011.

System Description

- Acoustic modeling
  - ✓ All the acoustic models is described as follows
    - ➢ CNN-TDNN-F{1, 2, 3, 4}: GSS module with {10, 15} context, 40-dim MFCC, 6-layer CNN + 19-layer TDNN-F with bottleneck-dim = 512, NUM-PDFS = {2500, 3500}
    - ➢ CNN-TDNN-F{5, 6}: GSS module with {10, 15} context, 40-dim MFCC, 6-layer CNN + 19-layer TDNN-F with bottleneck-dim = 768, NUM-PDFS = 3500
    - ➢ CNN-TDNN-F7: GSS module with 15 context, 80-dim MFCC, 6-layer CNN + 19-layer TDNN-F with bottleneck-dim = 512, NUM-PDFS = 3500
    - ➢ TDNN-F8: GSS module with 15 context, 40-dim MFCC, 25-layer TDNN-F with bottleneck-dim = 512, NUM-PDFS = 2500

O1_3_video.webm

PowerPoint 幻灯片放映 - [CHiME6_OPPO.pptx] - PowerPoint

System Description

- Language modeling
  - ✓ Build a 2-layer LSTM-based language model.
  - ✓ Rescore the lattice using the score of LSTM-base LM and official n-gram LM with a weighting of 0.55 and 0.45

recordscreen.io 正在共享窗口。 停止共享 隐藏

06:35

O1_3_video.webm

PowerPoint 幻灯片放映 - [CHiME6_OPPO.pptx] - PowerPoint

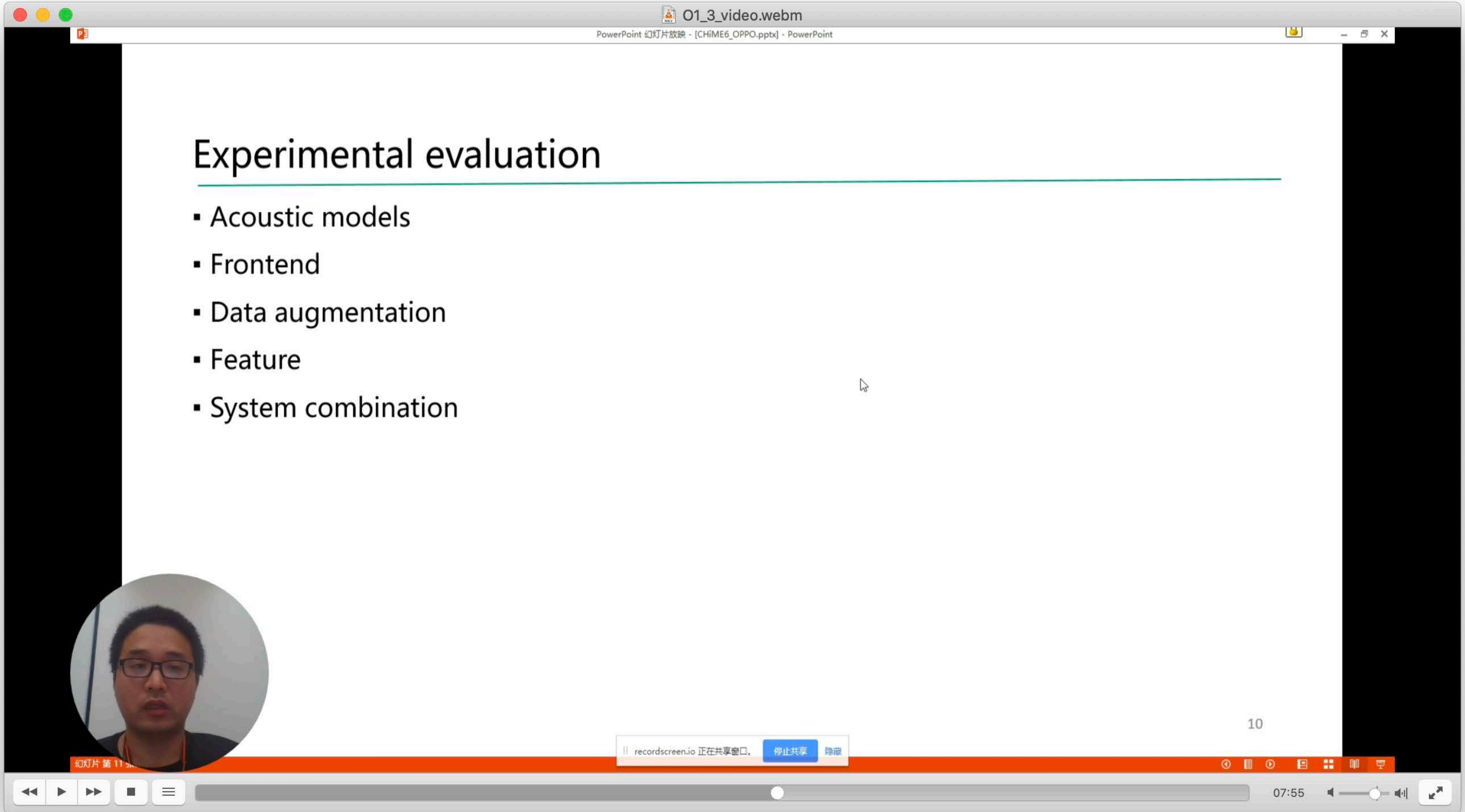
System Description

- Decoding
  - ✓ STEP1: generate lattices using eight acoustic models described in acoustic modeling section
  - ✓ STEP2: for Category B, rescore the lattice with LSTM-LM and official n-gram LM, for Category A, skip this step
  - ✓ STEP3: Combine all the lattices and apply MBR[5] decoding to get the final result

Xu, D. Povey, L. Mangu and J. Zhu, "Minimum Bayes Risk decoding and system combination based on a recursion for edit distance,"
ter Speech & Language, vol. 25, pp. 802-828, 2011.

recordscreen.io 正在共享窗口。 停止共享 隐藏

07:14


O1_3_video.webm
PowerPoint 幻灯片放映 - [CHiME6_OPPO.pptx] - PowerPoint
Experimental evaluation
▪ Acoustic models
▪ Frontend
▪ Data augmentation
▪ Feature
▪ System combination
10
recordscreen.io 正在共享窗口。
停止共享
隐藏
07:55

Experimental evaluation

- Acoustic models
  - ✓ Compared to official TDNNF model with 15 layers, deeper TDNNF model with 25 layers and CNN-TDNNF model which has 6 convolution layers and followed by 19 TDNNF layers can reduce WER obviously

AM	DEV(%)	EVAL(%)
Official TDNNF	51.76	51.29
Deeper TDNNF	50.77	50.30
CNN-TDNNF	48.53	48.15

Table 1: WER of different acoustic models on the dev and eval sets

Experimental evaluation

- Frontend
  - ✓ Based on CNN-TDNNF model, in order to match the data in testing stage, apply GSS for all multi-array data in training stage
  - ✓ Compared to the official baseline which select 400K utterance from multi-array data, apply GSS in training stage can reduce WER by 2% absolutely on the dev set

AM	Frontend	DEV(%)	EVAl(%)
CNN-TDNNF	baseline	48.53	48.15
CNN-TDNNF	apply GSS in training stage	46.54	48.02

Table 2: WER of different frontends on the dev and eval sets

Experimental evaluation

- Data augmentation
  - ✓ Apply GSS module in training stage, reduce the amount of training data
  - ✓ Replace the L channel worn data with L+R channel worn data
  - ✓ Realign L+R channel worn data and make RIR data augmentation, and it can greatly improves the performance

DATA	DEV(%)	EVAL(%)
multi-array GSS + worn(L)	46.54	48.02
multi-array GSS + worn(L+R)	45.55	47.30
Multi-array GSS + aligned worn(L+R) +RIR	45.31	45.81

Table 3: WER of different datasets on the dev and eval sets

Experimental evaluation

- Feature
  - ✓ In addition, we find that using 80-dimension MFCC can slightly improve the performance of the system

Feature	DEV(%)	EVAL(%)
40-dim MFCC	45.31	45.81
80-dim MFCC	44.99	45.28

Table 4: WER of different features on the dev and eval sets

Experimental evaluation

- System combination
  - ✓ Combine lattice produced by 8 acoustic models, WER of each model are presented in Table 5
  - ✓ For category A and B in Track1, apply MBR decoding on the combined lattice and get the best result of the system

AM	DEV(%)	EVAL(%)
CNN-TDNN-F1(10,2500)	45.20	45.76
CNN-TDNN-F2(10,3500)	45.00	45.58
CNN-TDNN-F3(15,2500)	45.61	45.74
CNN-TDNN-F4(15,3500)	45.31	45.81
CNN-TDNN-F5(10)	45.46	45.80
CNN-TDNN-F6(15)	45.50	46.17
CNN-TDNN-F7	44.99	45.28
TDNN-F8	46.66	47.14

Table 5: WER of different acoustic models on the dev and eval sets

Category	AMs	LM	DEV(%)	EVAL(%)
A	8 AMs	official n-gram LM	41.99%	42.41%
	8 AMs	+ LSTM-LM	41.18%	42.02%

Table 6: WER of system combination on the dev and eval sets

Experimental evaluation

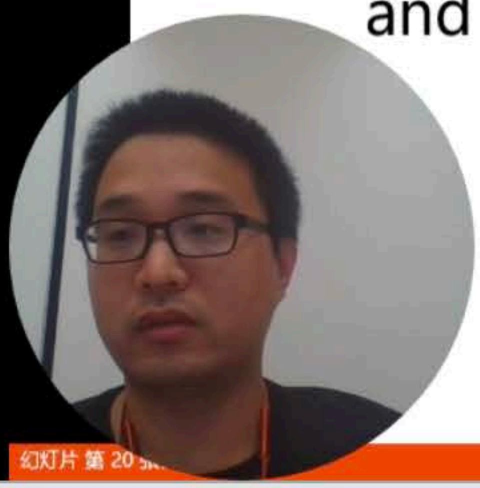
- Results summary

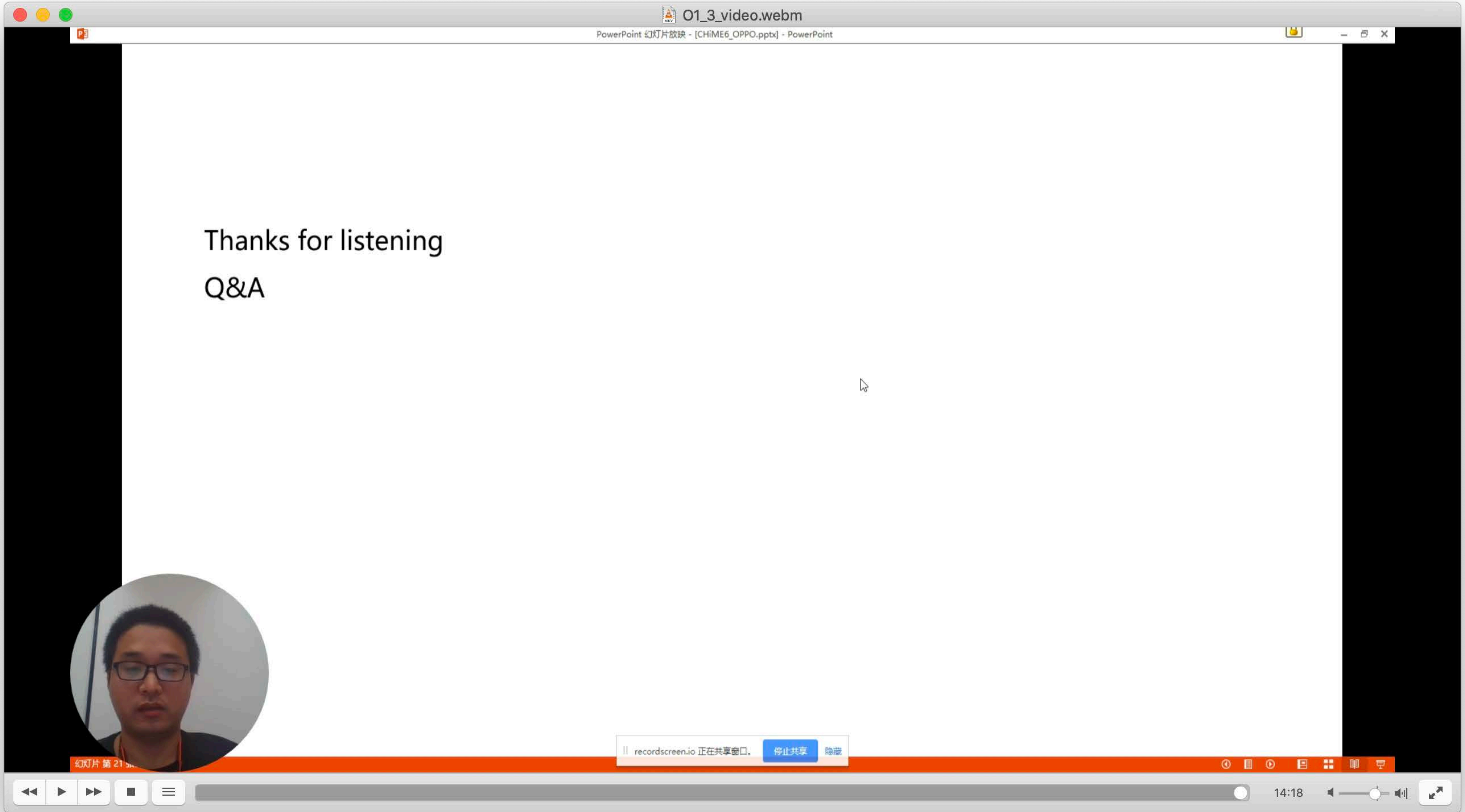
Categroy	Session		Kitchen	Dining	Living	Ave
A	Dev	S02	47.42	45.57	38.31	41.99
		S09	39.28	43.22	38.80	
	Eval	S01	58.00	35.83	47.80	42.41
		S21	51.49	35.09	34.43	
B	Dev	S02	46.66	45.00	37.47	41.18
		S09	38.48	41.99	38.04	
	Eval	S01	57.56	35.47	47.76	42.02
		S21	50.95	34.95	34.75	

Table 7: WER of the system in Track1(ASR only) for category A and category B

Conclusion

- Apply GSS module in training stage can improve the performance
- Data augmentation is helpful to boost the system performance
- Compared to TDNNF model, CNN-TDNNF can get better result
- Combine various acoustic model lattices, do MBR decoding and it can greatly reduce WER




O1_3_video.webm
PowerPoint 幻灯片放映 - [CHiME6_OPPO.pptx] - PowerPoint
Thanks for listening
Q&A
recordscreen.io 正在共享窗口。
停止共享
隐藏
14:18