The RoyalFlush System of Speech Recognition for M2MeT Challenge

Shuaishuai Ye, Peiyao Wang, Shunfei Chen, Xinhui Hu and Xinkang Xu
Hithink RoyalFlush AI Research Institute, Zhejiang, China

OUTLINE

1. System configuration
2. Data Preparation
3. Experimental Settings and Results
4. Conclusions

OUTLINE

BACKGROUND

- M2MeT: Multi-channel Multi-party Meeting Transcription Challenge
- Organizers: Alibaba Group and Aishell
- Tracks

  - Track1 : speaker diarization

  - Track2 : multi-speaker ASR

  - Organizers set up two subtrack for both above tracks:

    -- subtrack1 : the first sub-track limits the usage of data.

    -- subtrack2 : the second sub-track allows the participants to use extra constrained data.

- Dataset

  - Dataset1 : AliMeeting (~ 120 hours)

  - Dataset2 : Aishell4 (~120 hours)

THE FINAL RANK

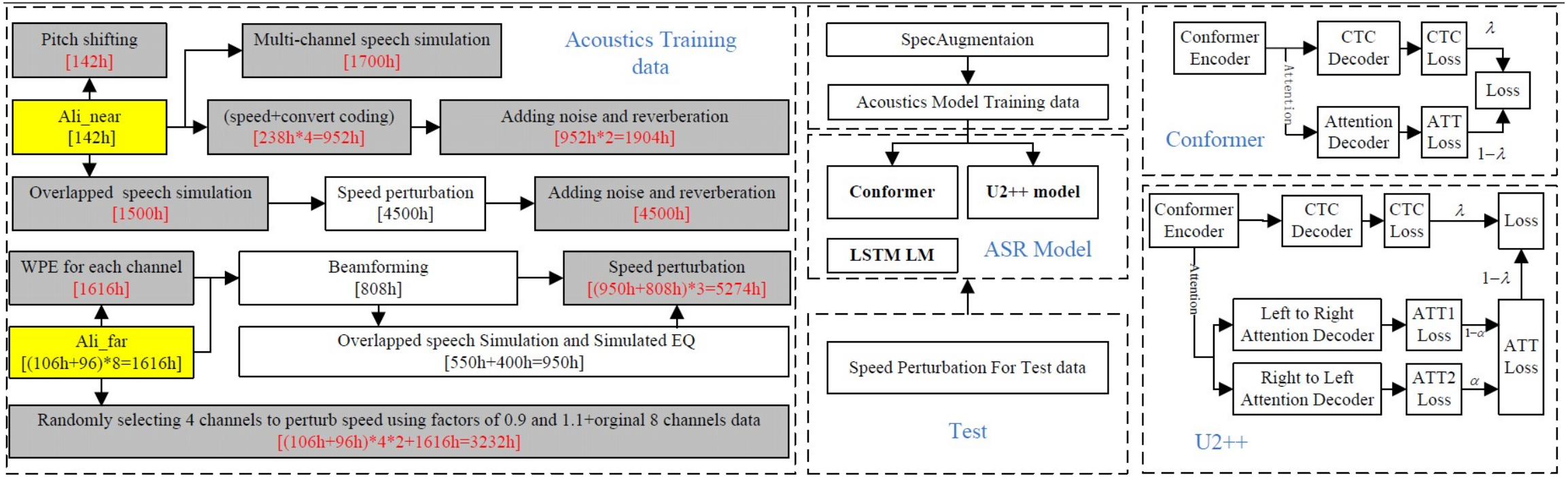
- Track1 : 5

- Track2 : 1

We only introduce the system of track2 .

System configuration

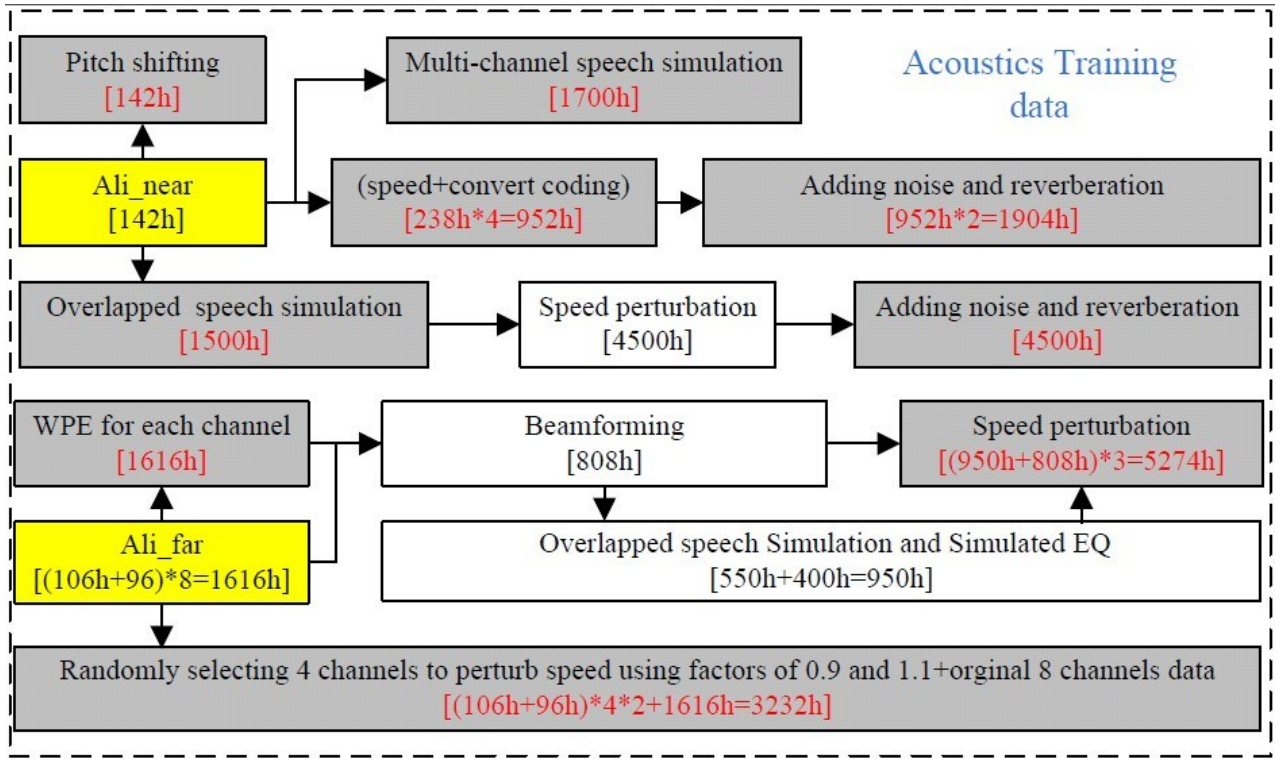
Overall architecture



- Part 1: Acoustics Training data
  1, Front-end data processing
  2, Data augmentation

- Part 2: ASR Model
  1, Conformer
  2, U2++
  3, LSTM Language model
  4, Result fusion

Acoustics Training data



Front-end Data Processing:

- WPE : long-term linear prediction method
- Beamforming : delay-and-sum

WPE + Beamforming

And we processed our train set, development data and test set using WPE + Beamforming

Data augmentation:

See the next section.

System configuration

ASR Model

TRAINING: The modeling capabilities of different acoustic models can complement each other.

- Conformer: the standard conformer based joint CTC/Attention ASR model (ESPnet)
- U2++: the conformer based U2++ ASR model (WeNet)

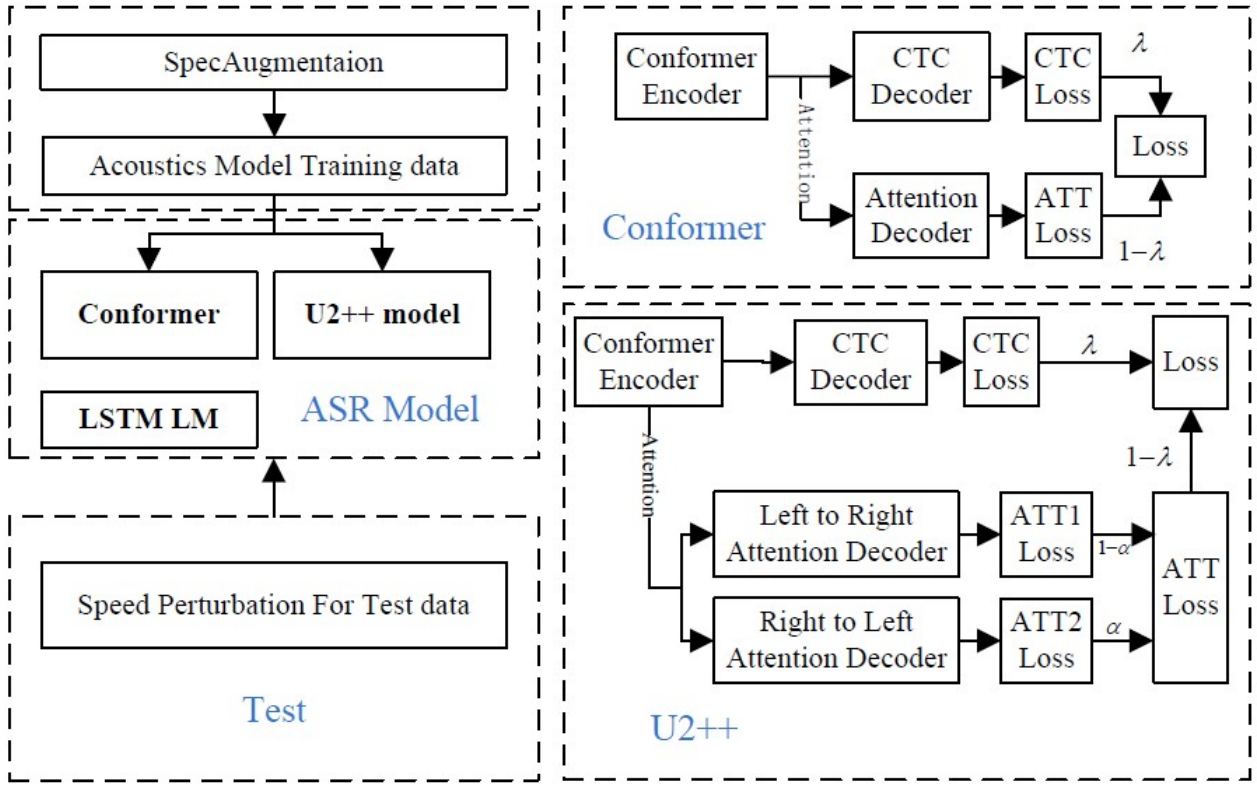
Language Model

- LSTM

TESTING: Results are fused from two dimensions: model and test set.

- Conformer -> Averaging 8 best models.
  speed 1.0 , speed 1.1, speed 0.9
- U2++ -> Averaging 10 best models.
  speed 1.0 , speed 1.1, speed 0.9
- Conformer + LSTM language

So, Fusing 7 the above results using ROVER tools to get the final result.

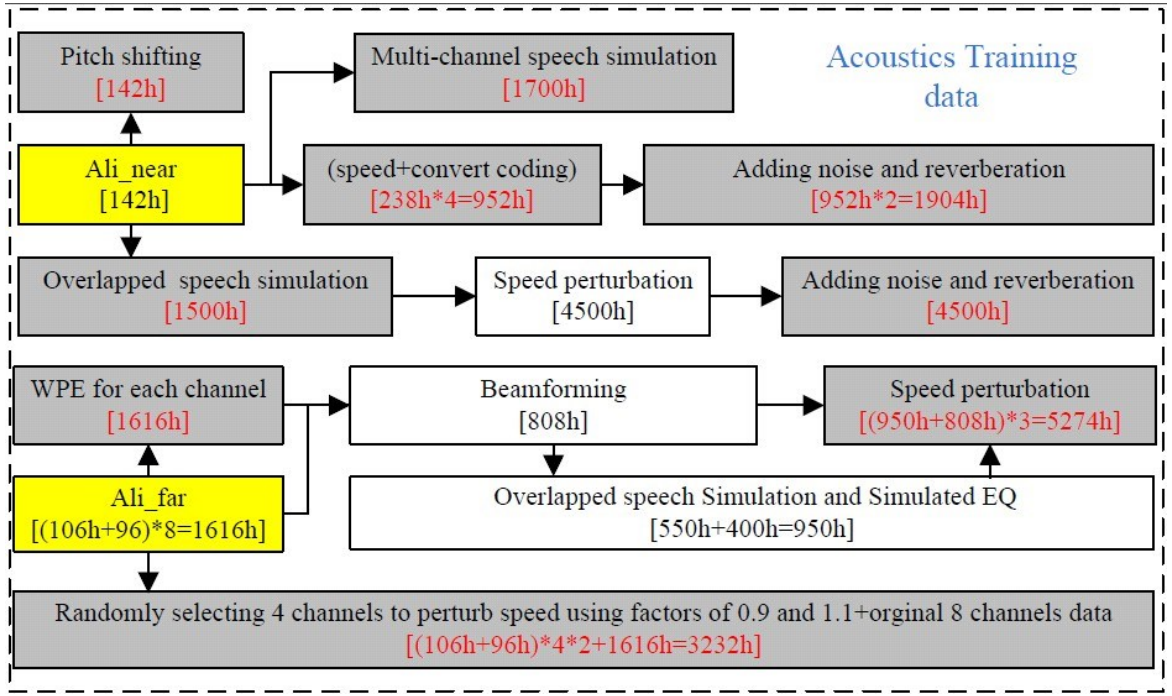


OUTLINE

1. System configuration
2. Data Preparation
3. Experimental Settings and Results
4. Conclusions

Data Preparation

Data Augmentation



Yellow parts are original training data, and grey parts are the final training data. the final duration of effective training data is about 18,000 hours.

Data Augmentation

- Speed perturbation
  - speed factor 0.9, 1.0 and 1.1
- Adding noise and reverberation
  - noise sources: MUSAN Corpus (no bable), AliMeeting noise
  - reverberation sources: RIRS_NOISES Corpus, RIRS_AliMeeting
- Pitch Shifting
- Multi-channel far-field speech simulation
  - 17,400 multi-channel RIRs are generated according to AliMeeting room size and some random room size.
  - simulated multi-channel array data using the above RIRs.
- Overlapped speech simulation
  - Source data: headset data.
  - 50% for 2 speakers, 30% for 3 speakers, 20% for 4 speakers with a random overlapped ratio factor.
- SpecAugment
- Simulated equalization
  - various filters to process data, including weighted low-pass/high-pass filter, de-emphasis filter, simulated frequency response curve filter

OUTLINE

1. System configuration
2. Data Preparation
3. Experimental Settings and Results
4. Conclusions

Experimental Setting

Acoustic features : 80-dimensional Fbank concatenated with 3-dimensional Pitch.

ASR Model settings

- Conformer:
  - ✓ a 12-layer encoders with 2048-dim feed forward, a 6-layer decoder with 2048 units and 8 heads attention with 512.dimensions.
  - ✓ averaging the best 8 models' weights based on loss value.
- U2++:
  - ✓ an 18-layer encoder, a 3-layer left-to-right decoder and a 3-layer right-to-left decoder and 4 heads attention with 256 dimensions.
- averaging the best 10 models' weight based on loss value.
- Learning rate: 0.001
- Beam_size: 20

Language Model

- LSTM
  - ✓ 2 LSTM layers with 1024 dimensions
  - ✓ Training data: original text + concatenating text

The results of different models on the AliMeeting far-field validation and test set. 'wb' refers to 'wpe-bf', and 'Eval' refers to evaluation set. 'U2++ _*channel' represents the number of channel. 'U2++_big_data' represents the final model trained by all data.

Group	System	Eval (CER%)						Test (CER%)		
		Eval-1c	Eval-wpe	Eval-bf	Eval-wb	Eval-wb-sp0.9	Eval-wb-sp1.1	Test-wb	Test-wb-sp0.9	Test-wb-sp1.1
Group1	baseline [24]	30.8	-	29.7	-	-	-	30.9		
Group2	Conformer_Ali-near	47.44	-	-	-	-	-	-	-	-
	+SOT	46.06	-	-	-	-	-	-	-	-
	Conformer_5k-hours	24.02	24.99	23.32	22.77	-	-	23.15	-	-
	Conformer_10k-hours	21.49	21.71	20.27	**19.06**	**19.74**	**19.31**	**20.14**	**20.65**	**20.68**
	+LSTM LM	21.64	21.84	20.52	**19.14**	19.90	19.59	**20.29**	-	-
Group3	U2++_1channel	28.49	28.95	26.83	26.13	-	-	26.78	-	-
	U2++_4channels	27.93	28.47	26.56	26.04	-	-	26.64	-	-
	U2++_8channels	27.45	27.86	26.15	25.50	-	-	26.20	-	-
	U2++_big_data	21.10	21.08	19.92	**18.68**	**18.82**	**19.12**	**19.99**	**20.10**	**20.46**
Group4	Fusion*				**17.48**				**18.79**	

* The numbers with bold font are fused as the final result using the *ROVER* tool.

Group	System	Eval (CER%)						Test (CER%)		
		Eval-1c	Eval-wpe	Eval-bf	Eval-wb	Eval-wb-sp0.9	Eval-wb-sp1.1	Test-wb	Test-wb-sp0.9	Test-wb-sp1.1
Group1	baseline [24]	30.8	-	29.7	-	-	-	30.9		
Group2	Conformer_Ali-near	47.44	-	-	-	-	-	-	-	-
	+SOT	46.06	-	-	-	-	-	-	-	-
	Conformer_5k-hours	24.02	24.99	23.32	22.77	-	-	23.15	-	-
	Conformer_10k-hours	21.49	21.71	20.27	**19.06**	**19.74**	**19.31**	**20.14**	**20.65**	**20.68**
	+LSTM LM	21.64	21.84	20.52	**19.14**	19.90	19.59	**20.29**	-	-
Group3	U2++_1channel	28.49	28.95	26.83	26.13	-	-	26.78	-	-
	U2++_4channels	27.93	28.47	26.56	26.04	-	-	26.64	-	-
	U2++_8channels	27.45	27.86	26.15	25.50	-	-	26.20	-	-
	U2++_big_data	21.10	21.08	19.92	**18.68**	**18.82**	**19.12**	**19.99**	**20.10**	**20.46**
Group4	Fusion*				**17.48**				**18.79**	

* The numbers with bold font are fused as the final result using the *ROVER* tool.

Our experiments were conducted on the following several aspects to investigate the influence on the ASR model:
(1) SOT scheme; (2) the amount of training data; (3) the number of channel used to train model; (4) front-end methods, namely the WPE and Beamforming approach; (5) model fusion.

OUTLINE

Conclusions

Experimental results showed :

1. WPE and Beamforming based on multi-channels can effectively decrease the CER.
2. Using a variety of data augmentations to expand the training set can greatly improve the performance and robustness.
3. The fusion of different results from two dimensions, model and test set, can achieve 1.2% absolute CER reduction on development set when compared with the best result of U2++ ASR model.
4. Comparing with the official baseline system, our system got a 12.22% absolute CER reduction on the development set, from 29.7% to 17.48% and 12.11% absolute CER reduction on the test set, from 30.9% to 18.79%.

THANKS for your attention

Contact Email: ***yeshuaishuai@myhexin.com***