Punctuation Prediction for Streaming On-Device Speech Recognition

Zhikai Zhou*, Tian Tan^, Yanmin Qian*

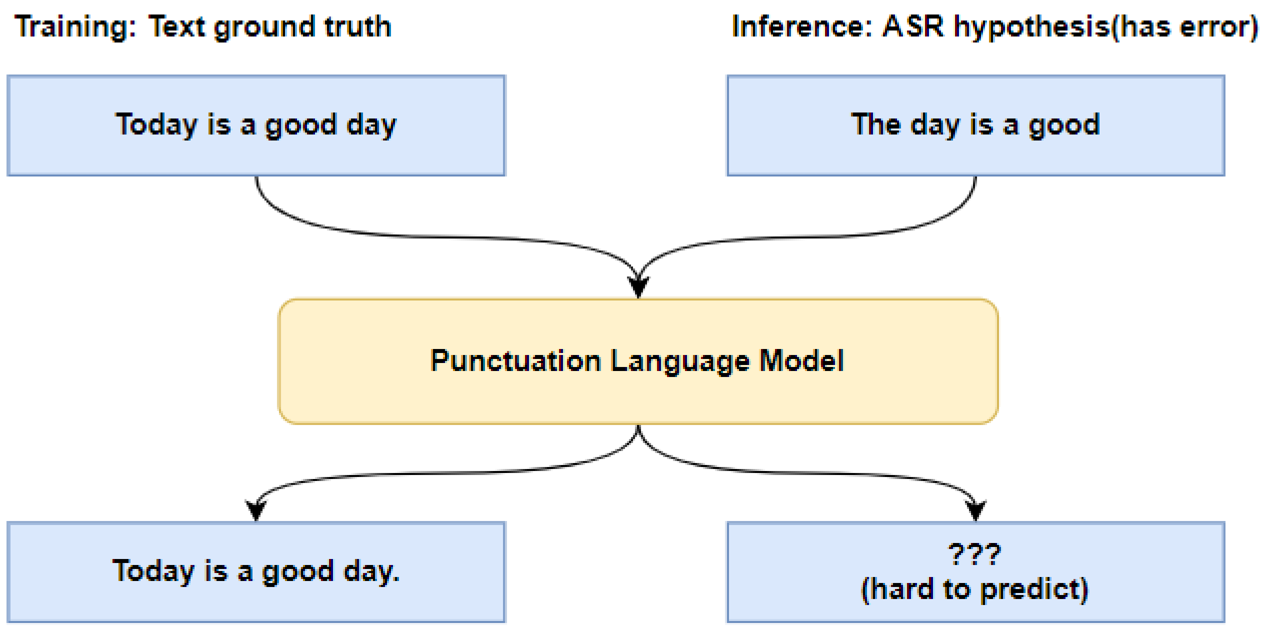
X-LANCE Lab, Shanghai Jiao Tong University, China*
AISpeech Ltd, Suzhou, China^

May 2022

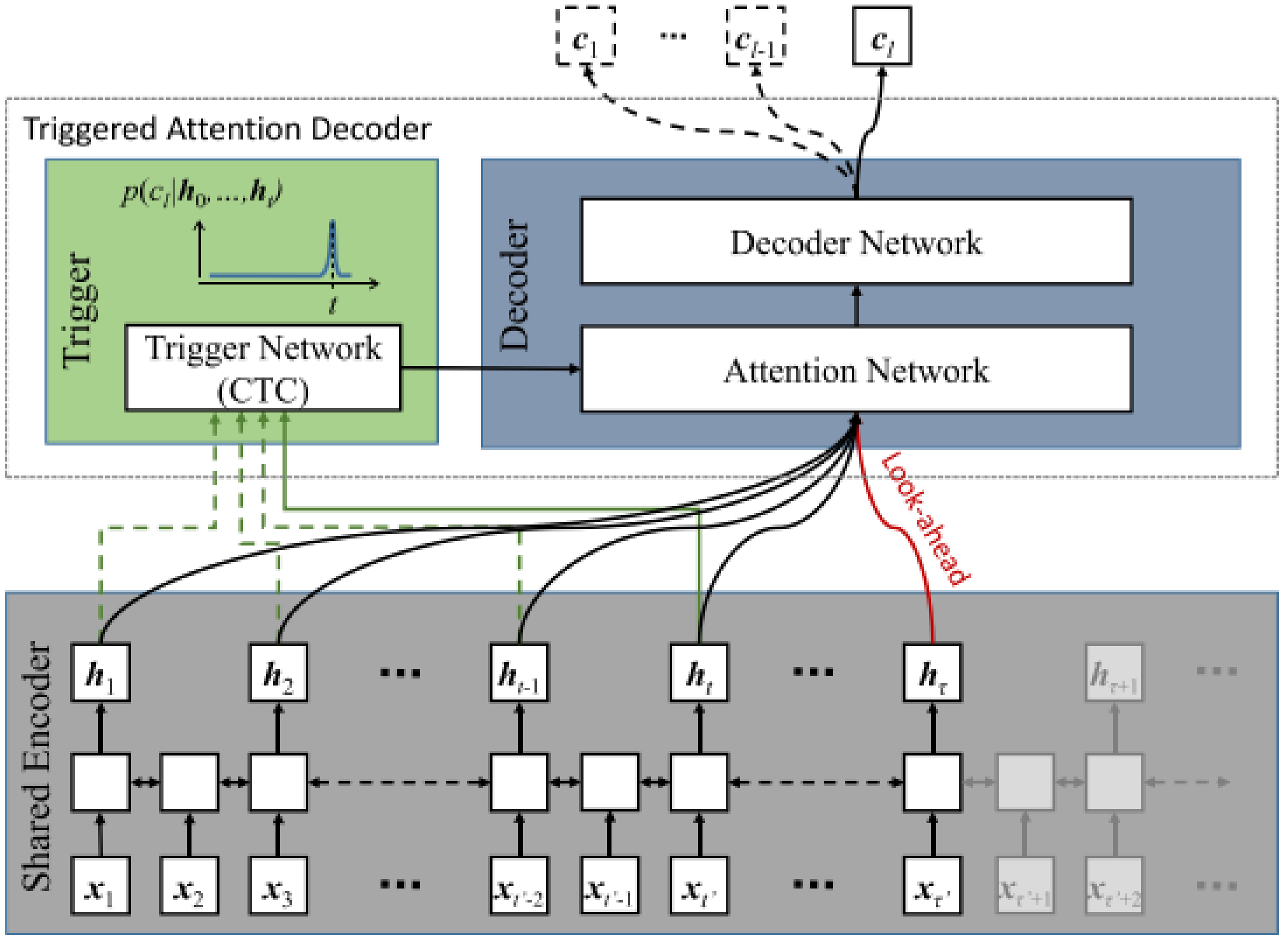
Background

Punctuation Prediction for On-device Scenarios

- Common ASR does not model punctuations
- ASR post-processing procedure for punctuations
  - An extra post-processing model is needed, while costly for on-device scenarios
  - Mismatch between text sequence(training) and ASR hypothesis(inference)



Related work – Streaming Speech Recognition

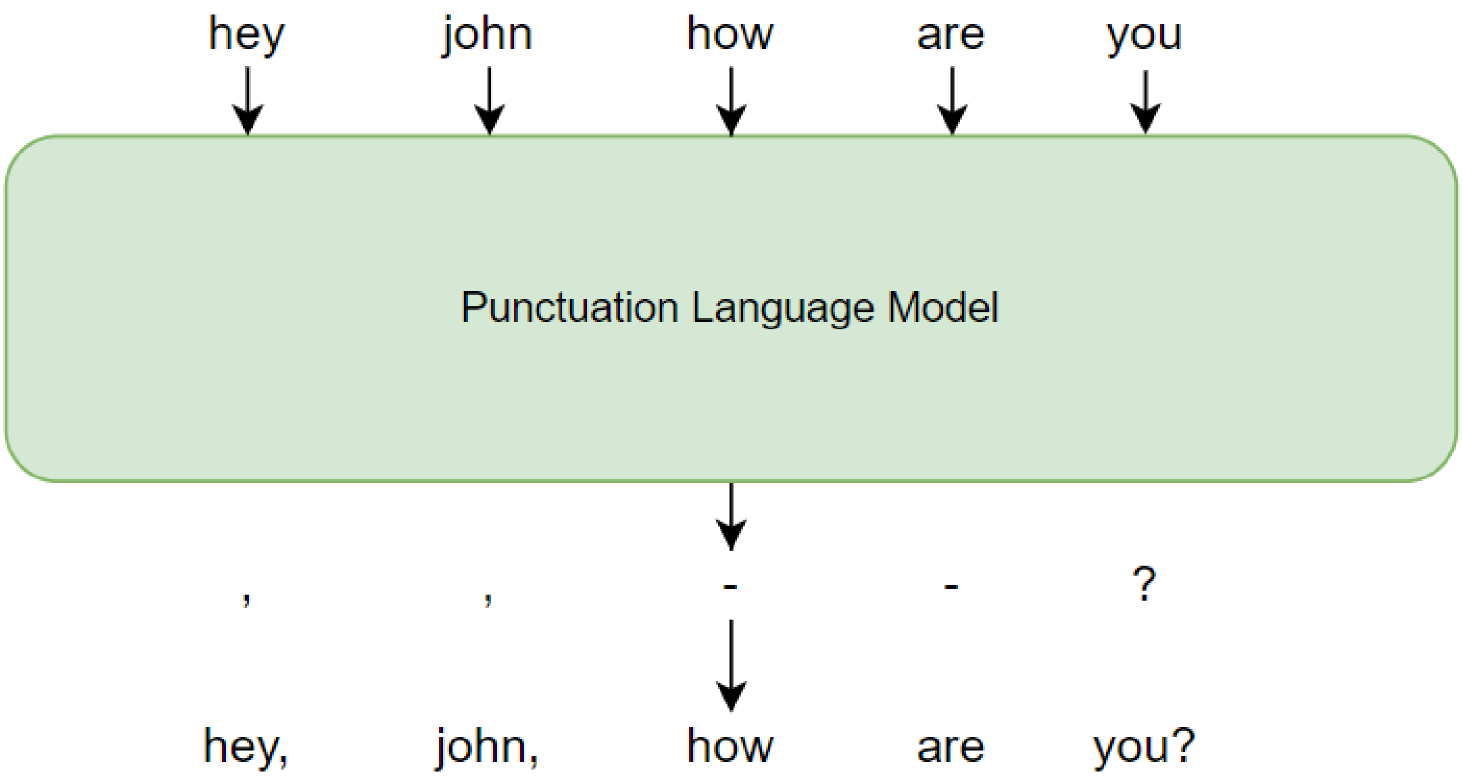


- Spikes are detected by the CTC trigger network
- Once the spike is detected, the decoder take a step.
- In practice, we count the spikes and decode chunk by chunk.

[1] Moritz N, Hori T, Le Roux J. Triggered attention for end-to-end speech recognition[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 5666-5670.

Background

Related work – Punctuation Prediction

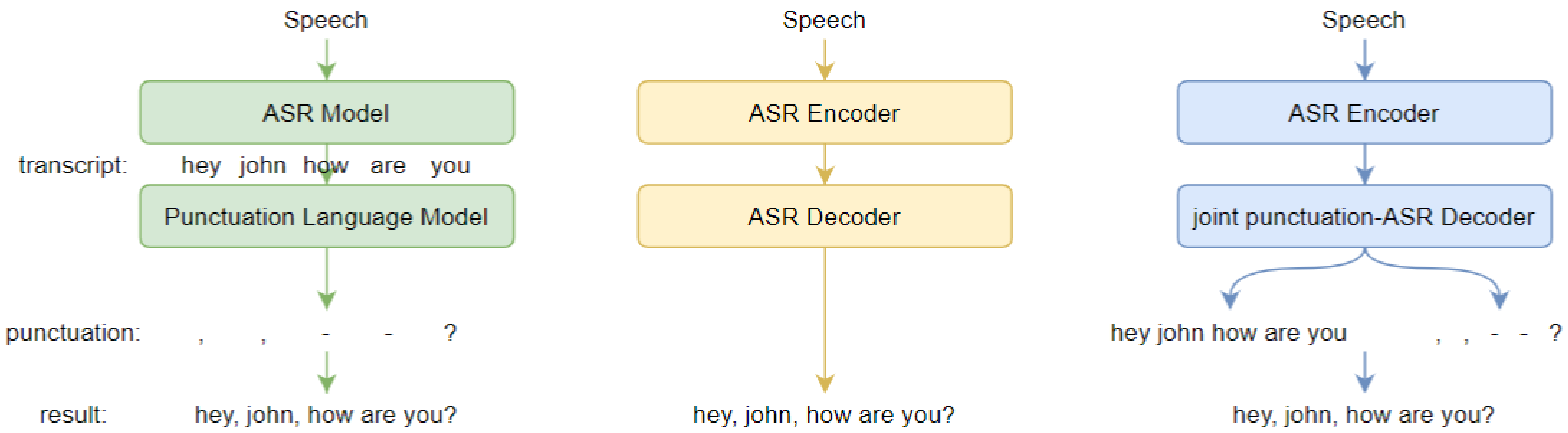


- Input is text only
- For each token, the model predict which the punctuation follows(including blank)
- Can be initialized by MLM or other pretrained models

Drawbacks of Language Model based Punctuation Prediction

- Text-only input makes model's output have no difference in different speech. E.g.
  - "Onetwothreefourfive."
  - "One,two,three,four,five."
  - "One,twothree,fourfive."
- An independent model needs many parameters to model the task

Proposed Method



(a) Two-pass pipeline of ASR post processing

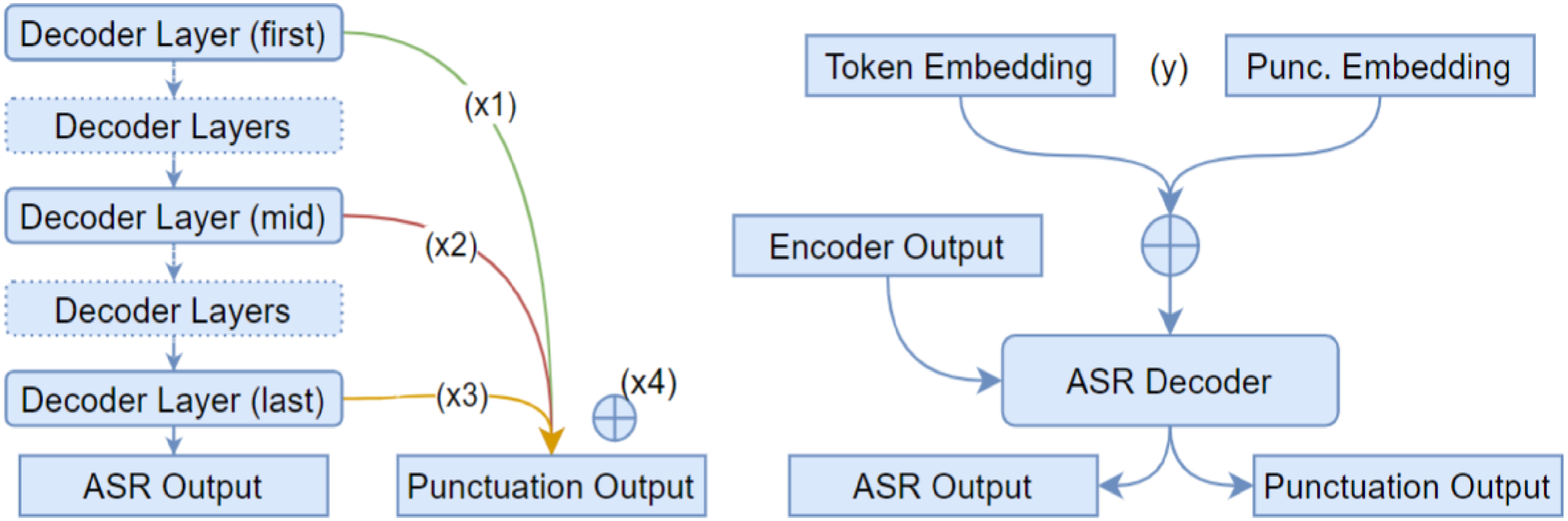
(b) Direct ASR modeling of transcripts with punctuation

(c) Joint modeling of both ASR transcription and punctuation

- Three methods are explored to model punctuations with ASR for on-devices scenarios
- The joint modeling of ASR and punctuation is proposed in this work

Proposed Method

Different methods for the joint model



- Feature from different layers of the decoder
- Auto-regressive decoding for joint punctuation and ASR

Example
Ref: Hey, John, how are you**?**
Hyp: Hey, John, how**’re** you

- Punctuation errors and ASR errors are tangled with each other
- We need to separate them
- The teacher-forcing decoding scheme is proposed to evaluate the punctuation performance as follows

$$P(s_t|x) = P(s_t|h, \hat{y}_{<t}, \hat{s}_{<t})$$

$$P(h|x) = Encoder(x)$$

Experiment

Dataset

- Training set: 3000 hrs of in-house Chinese spoken Dataset, split into 90% - 10% for development set
- Test set: 10 hrs Indoor, 5 hrs Meeting, 16hrs Mobile
- Punctuation: Comma, Period, Question Mark, Enumeration Comma, and Blank.

Model	#params	Indoor TER/F_1	Mobile TER/F_1	Meeting TER/F_1
Trans-2L	9.88M	15.93/86.80	27.94/70.69	28.89/72.64
Trans-4L	16.19M	15.88/87.28	27.84/71.48	28.76/74.09
Trans-6L	22.49M	15.72/87.59	27.73/71.79	28.64/74.15
ASR	72.6M	13.49 (CER)	24.89 (CER)	25.91 (CER)

Table 1: Performance Comparison of the Two-Pass Strategy with Punctuation Models

- TER: Whole sequence token error rate
- CER: The raw text sequence character error rate
- F1: Averaged punctuation F1-score

Experiment

Results

Model	α	#Ext par.	Indoor TER/CER/F_1	Mobile TER/CER/F_1	Meeting TER/CER/F_1	Average TER/CER/F_1
ASR + Trans-6L	-	22.49M	15.72/13.49/87.59	27.73/24.89/71.79	28.64/25.91/74.15	24.03/21.43/77.84
ASR with Punc	-	11.3K	15.49/14.45/**92.02**	31.73/28.87/71.66	31.88/29.95/78.06	26.37/24.42/80.58
Joint Model -x3	1.0	2.0K	14.62/**13.19**/91.01	24.27/21.39/72.38	**27.76/25.33**/78.82	22.22/19.97/80.74
Joint Model -x3	2.0	2.0K	**14.51**/13.20/91.45	**23.66/20.60**/71.70	28.46/26.15/78.29	**22.21/19.98**/80.48
Joint Model -x3	5.0	2.0K	14.53/13.36/92.00	24.48/21.59/**72.17**	28.77/26.53/**79.11**	22.59/20.49/**81.09**
Joint Model -x1	2.0	2.0K	39.51/17.54/50.51	57.92/35.58/35.21	46.35/37.74/53.51	47.93/30.29/46.41
Joint Model -x2	2.0	2.0K	20.10/13.68/79.84	35.44/25.99/58.57	34.68/27.77/69.74	30.07/22.48/69.38
Joint Model -x3	2.0	2.0K	**14.51/13.20**/91.45	**23.66/20.60**/71.70	28.46/26.15/78.29	**22.21/19.98**/80.48
Joint Model -x4	2.0	2.0K	14.61/13.23/91.17	25.31/22.40/70.91	**28.29/25.83**/77.72	22.74/20.49/79.93
Joint Model -y	2.0	4.0K	14.56/13.24/91.20	24.82/21.95/**72.30**	29.23/27.01/**78.46**	22.87/20.73/**80.65**

Table 2: Performance comparison of different strategies for both ASR and punctuation prediction. ASR+Trans-6L: The two-pass pipeline using punctuation language models. ASR with punc: The one-pass direct ASR modeling on transcripts with punctuation. Joint Model utilizes feature from which output of the decoder layer: x1: 1st, x2: 3rd, x3: Last, x4: Sum of all, y: Last, but feed punctuation result to the input.

- The direct modeling(ASR with Punc) has good punctuation result while worse in ASR.
- Joint Model-x3 with $\propto = 2.0$ achieves the best position, which is better than first two methods.

Experiment

Results

Model	α	#Ext par.	Indoor TER/CER/F_1	Mobile TER/CER/F_1	Meeting TER/CER/F_1	Average TER/CER/F_1
ASR + Trans-6L	-	22.49M	15.72/13.49/87.59	27.73/24.89/71.79	28.64/25.91/74.15	24.03/21.43/77.84
ASR with Punc	-	11.3K	15.49/14.45/**92.02**	31.73/28.87/71.66	31.88/29.95/78.06	26.37/24.42/80.58
Joint Model -x3	1.0	2.0K	14.62/**13.19**/91.01	24.27/21.39/72.38	**27.76/25.33**/78.82	22.22/19.97/80.74
Joint Model -x3	2.0	2.0K	**14.51**/13.20/91.45	**23.66/20.60**/71.70	28.46/26.15/78.29	**22.21/19.98**/80.48
Joint Model -x3	5.0	2.0K	14.53/13.36/92.00	24.48/21.59/**72.17**	28.77/26.53/**79.11**	22.59/20.49/**81.09**
Joint Model -x1	2.0	2.0K	39.51/17.54/50.51	57.92/35.58/35.21	46.35/37.74/53.51	47.93/30.29/46.41
Joint Model -x2	2.0	2.0K	20.10/13.68/79.84	35.44/25.99/58.57	34.68/27.77/69.74	30.07/22.48/69.38
Joint Model -x3	2.0	2.0K	**14.51/13.20**/91.45	**23.66/20.60**/71.70	28.46/26.15/78.29	**22.21/19.98**/80.48
Joint Model -x4	2.0	2.0K	14.61/13.23/91.17	25.31/22.40/70.91	**28.29/25.83**/77.72	22.74/20.49/79.93
Joint Model -y	2.0	4.0K	14.56/13.24/91.20	24.82/21.95/**72.30**	29.23/27.01/**78.46**	22.87/20.73/**80.65**

Table 2: Performance comparison of different strategies for both ASR and punctuation prediction. ASR+Trans-6L: The two-pass pipeline using punctuation language models. ASR with punc: The one-pass direct ASR modeling on transcripts with punctuation. Joint Model utilizes feature from which output of the decoder layer: x1: 1st, x2: 3rd, x3: Last, x4: Sum of all, y: Last, but feed punctuation result to the input.

- Joint Model achieves better performance on both ASR and punctuation while needs limited extra parameter.

Thanks !