SJTU SpeechLab System Description for SLT2021 CSRC Challenge

Wei Wang, Zhikai Zhou, Yizhou Lu, Yanmin Qian

SpeechLab
Shanghai Jiao Tong University

Overview

- Architecture
- Encoder Pretraining
- Data Augmentation
- Adaptation
- Decoding

Contents

1 Overview

2 Architecture

3 Augmentation and Adaptation

4 Decoding & Results

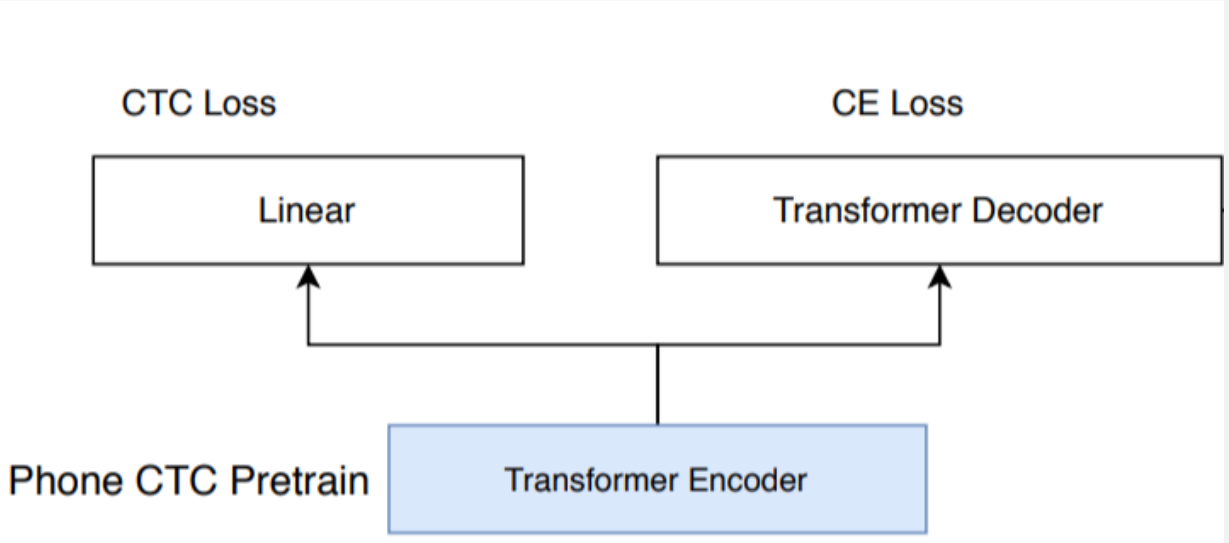
Larger Model

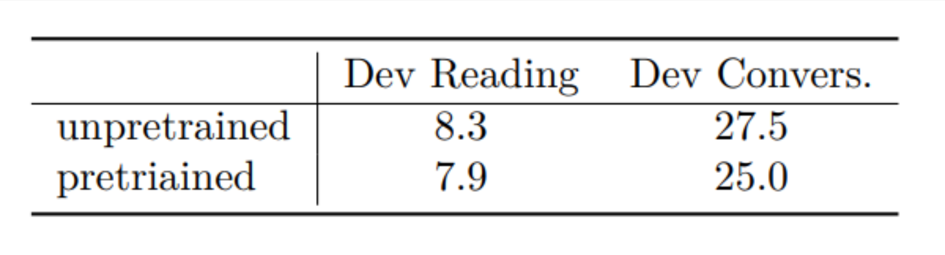
- Toolkit: ESPnet
- Data: 340h adult + 60h children
- Modeling unit: 3667 characters (#freq.>10) and 100 English subword.

Model	#attention heads	Attention dim	Parameters
Transformer Base (12 enc + 6 dec)	4	256	30M
Transformer Large (20 enc + 6 dec)	8	512	100M

- Learning rate is set to 5.0 (default 1.0)

Encoder Pretraining





	Dev Reading	Dev Convers.
unpretrained	8.3	27.5
pretriained	7.9	25.0

Since our transformer has a **deeper encoder**, model achieves better performance when phone based CTC pretraining is adopted on encoder.

M. Huang, Y. Lu, L. Wang, Y. Qian and K. Yu, "Exploring Model Units and Training Strategies for End-to-End Speech Recognition," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), SG, Singapore, 2019, pp. 524-531, doi: 10.1109/ASRU46091.2019.9003834.

Contents

1 Overview

2 Architecture

3 Augmentation and Adaptation

4 Decoding & Results

Data Augmentation

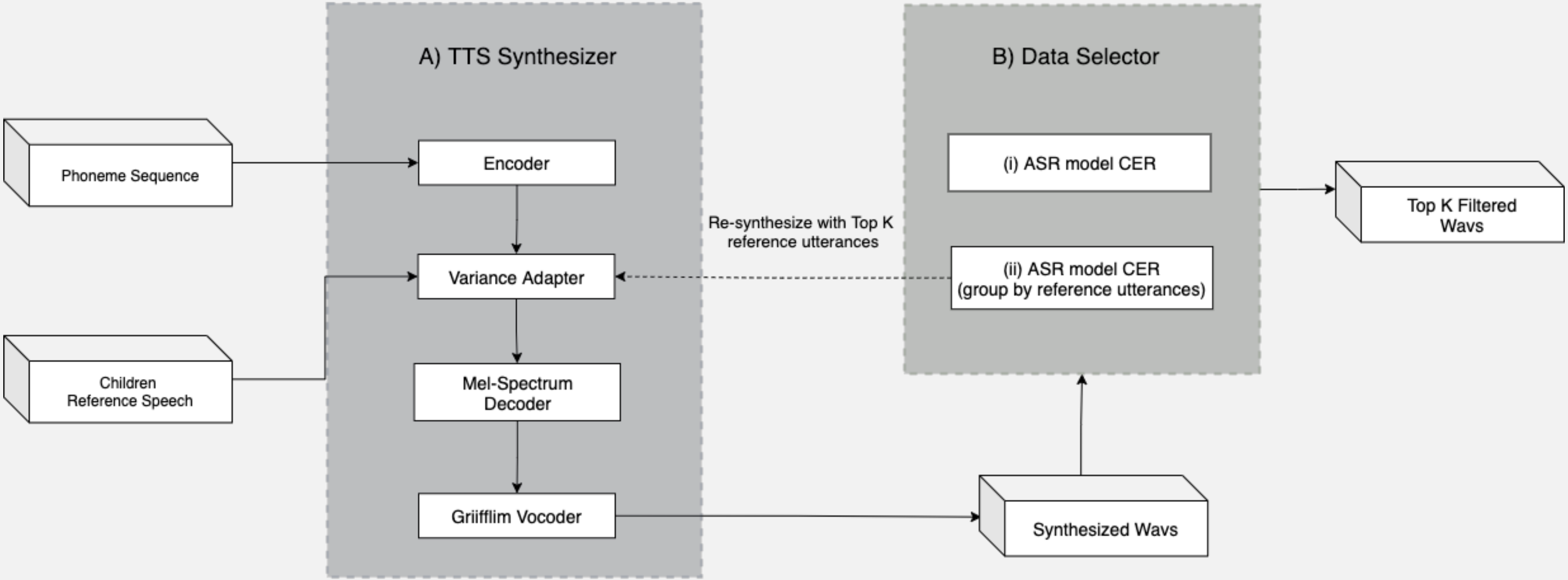
- Speed Perturbation (0.9, 1.0, 1.1)
- SpecAugment
- Prosody Modification
  - WSOLA from SoX "tempo" is conducted on adult speech signals to **simulate children speech**.
  - Pitch modification factor is set to **1.1** on adult speech

Li, C., Qian, Y. (2019) Prosody Usage Optimization for Children Speech Recognition with Zero Resource Children Speech. Proc. Interspeech 2019, 3446-3450, DOI: 10.21437/Interspeech.2019-2659.

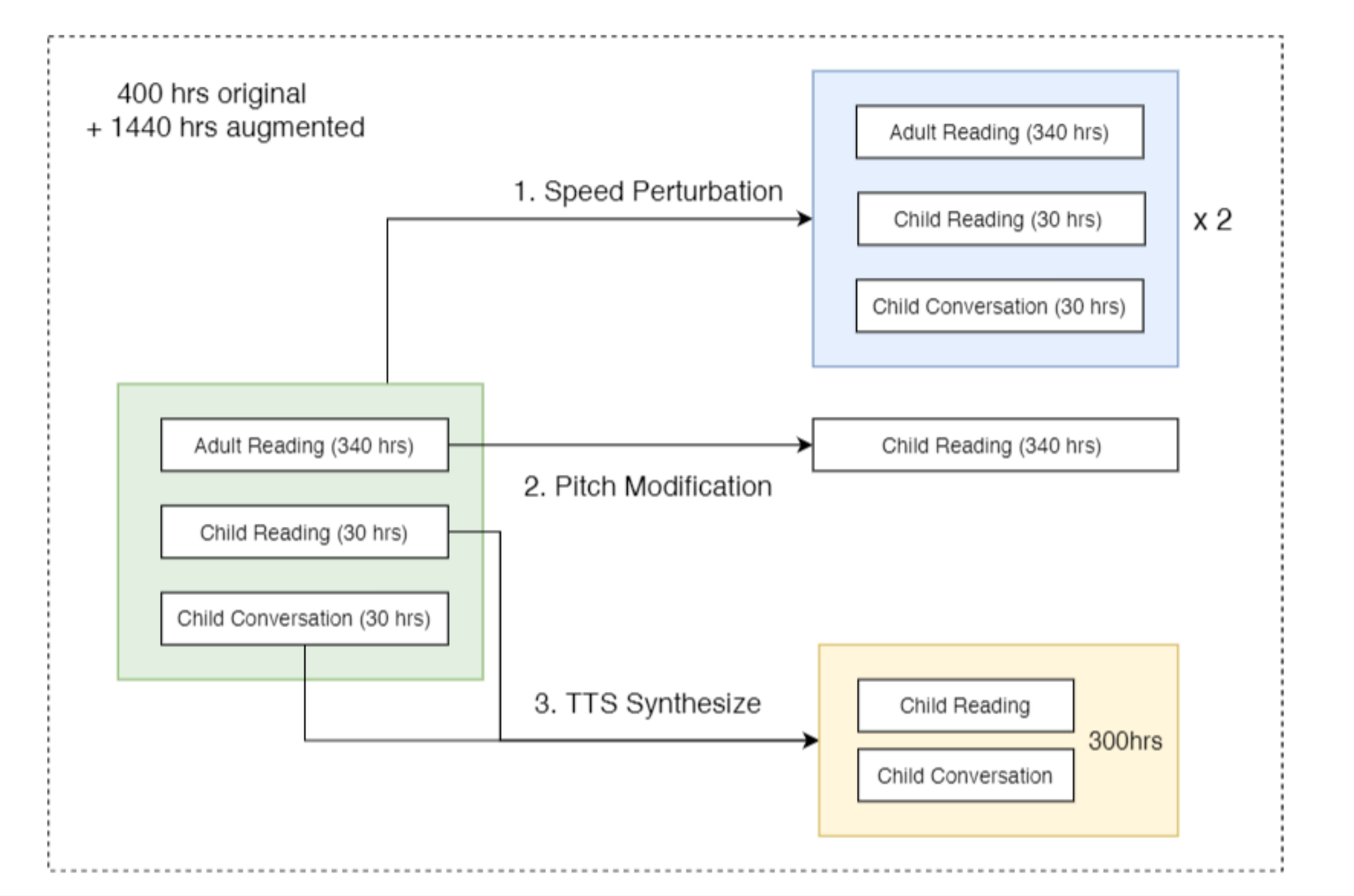
TTS Data Augmentation

- Model: Fastspeech2
- Training Data: Adults Reading+ Children Reading
  - Children Conversation is too dirty for TTS training.
  - Add adults’ speech for better performance
- Generated data: Given Transcript + Children Pattern
  - CER based filtering
  - CER based filtering + Re-synthesize

TTS Data Augmentation



Overall Data Augmentation



400 hrs original
+ 1440 hrs augmented

1. Speed Perturbation

Adult Reading (340 hrs)

Child Reading (30 hrs)

Child Conversation (30 hrs)

x 2

2. Pitch Modification

Adult Reading (340 hrs) → Child Reading (340 hrs)

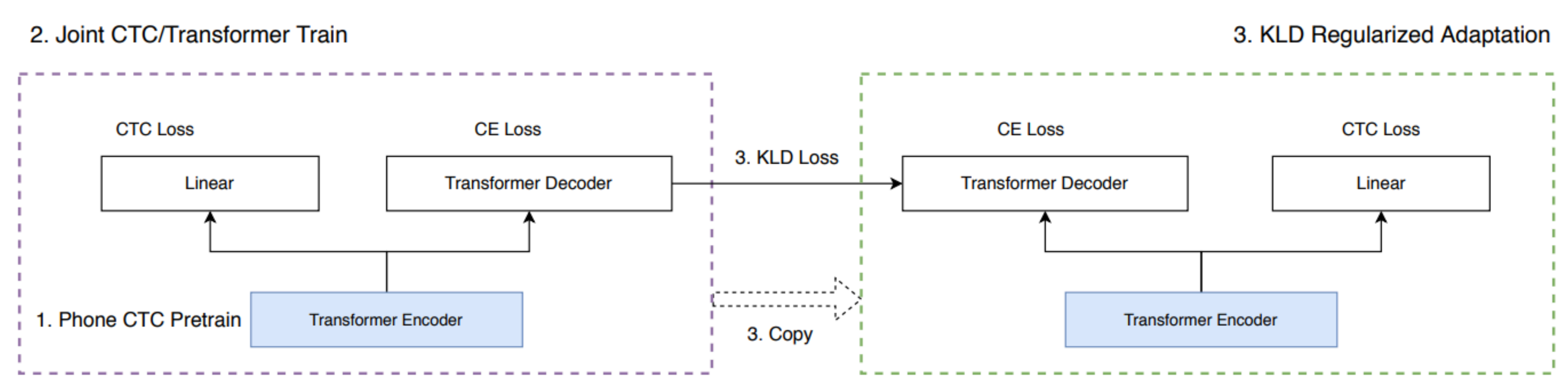
3. TTS Synthesize

Child Reading

Child Conversation

300hrs

Adaptation



After training with augmentation, model is fintuned on real children speech data(30h+30h).

Furthermore, the KLD loss with pretrained model is incorporated with ASR loss.

Yu, Dong & Yao, Kaisheng & Su, Hang & Li, Gang & Seide, Frank. (2013). KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on. 7893-7897. 10.1109/ICASSP.2013.6639201.

Contents

1 Overview

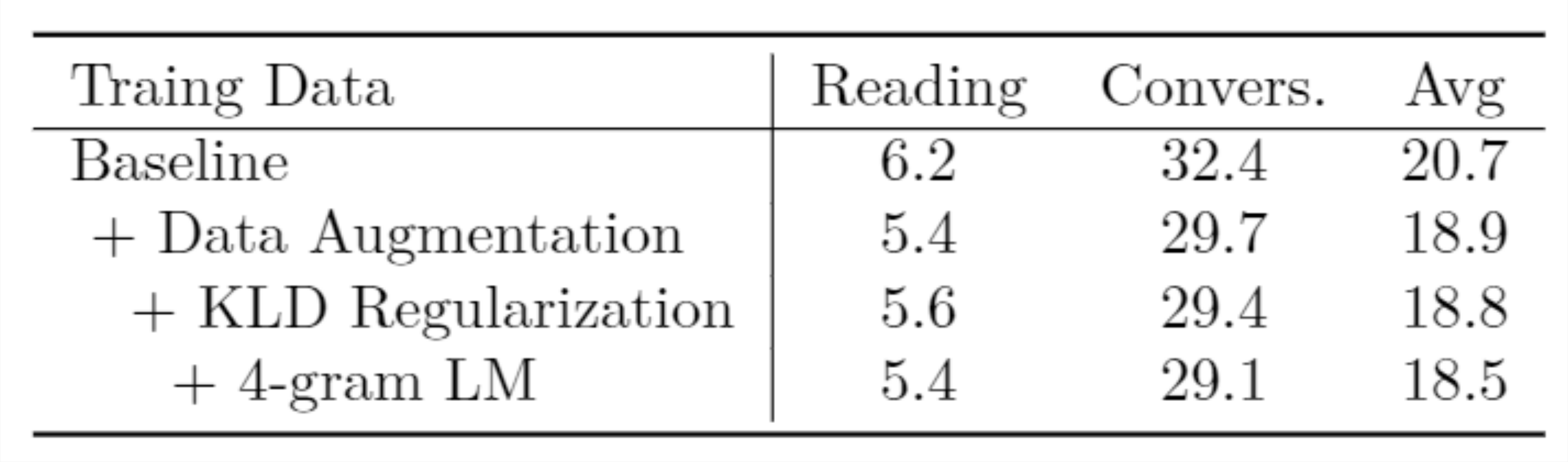
2 Architecture

3 Augmentation and Adaptation

4 Decoding & Results

Decoding and Results

- Model average on last 10 snapshots
- 4-gram character LM on all transcripts
- N-best results from joint CTC-Attention are rescored by LM of factor 0.3.
- The posteriors of English units are masked.



Traing Data	Reading	Convers.	Avg
Baseline	6.2	32.4	20.7
+ Data Augmentation	5.4	29.7	18.9
+ KLD Regularization	5.6	29.4	18.8
+ 4-gram LM	5.4	29.1	18.5

Thanks!