HAVE BEST OF BOTH WORLDS: TWO-PASS HYBRID AND E2E CASCADING FRAMEWORK FOR SPEECH RECOGNITION

Guoli Ye, Vadim Mazalov, Jinyu Li, and Yifan Gong
Microsoft Corporation, USA

1. Introduction

Hybrid and end-to-end (E2E) systems have their individual advantages, with different error patterns in the recognition results.

- E2E: jointly modeling audio and text, performs better in matched scenarios and scales well with a large amount of paired audio-text training data.
- Hybrid: modularized design, easier for customization, and better to make use of a massive amount of unpaired text data.

We proposed a **two-pass hybrid and E2E cascading (HEC)** framework to combine the hybrid and E2E model in order to take advantage of both sides, with hybrid in the first pass and attention-based encoder decoder (**AED**) model in the second pass.

2. Two-pass Hybrid and E2E Cascading (HEC) Framework

Fig. 1: Two-pass *Hybrid and E2E Cascading (HEC)* framework.

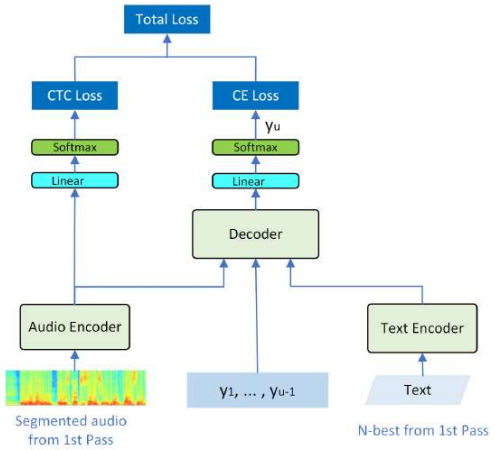

Fig. 2: Second pass AED model in two-pass HEC Framework.

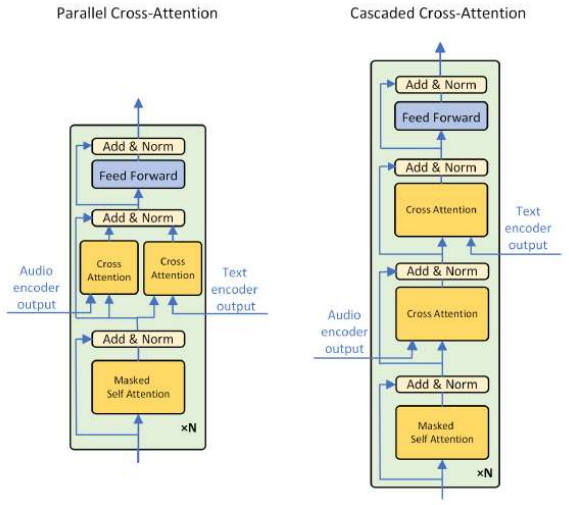

Fig. 3: Two Structures to incorporate the audio and text encoders into the second-pass AED decoder: (left) Parallel Cross Attention (PCA). (right) Cascaded Cross Attention (CCA).

3. Experiments

- Training data:
  - 65 k hours of transcribed Microsoft en-US production data from different scenarios
- Testing data:
  - **en-US General**: en-US production data, similar source as training data
  - **en-Dialect**: production data, with English speakers from 4 other countries
  - **en-Accent**: recorded in clean environment, from English speakers with accent
- Model structure
  - 80-dimension log Mel filter bank
  - **First pass hybrid:** an ensemble of two layer-trajectory bi-directional LSTM models, with 9404 senones as output
  - **Second pass E2E:** a joint CTC-attention AED model with 4k sentence piece as output

Test Set	Hybrid	AED	HEC		PCA WERR over	
			PCA	CCA	Hybrid	AED
en-US General	8.37	7.27	6.86	6.84	18.0	5.6
en-Dialect	10.97	11.48	10.53	10.47	4.0	8.3
en-Accent	11.79	11.84	10.74	10.85	8.9	9.3
Avg.	10.31	10.04	9.24	9.26	10.4	8.0

Table 1: WERs of Hybrid, AED and HEC models

Test Set	Old Hybrid	New Hybrid	HEC-PCA	
			Old Hyb	New Hyb
en-US General	8.37	8.5	6.86	6.92
en-Dialect	10.97	10.47	10.53	10.21
en-Accent	11.79	10.06	10.74	10.11
Avg.	10.31	9.58	9.24	8.94

Table 2: WERs of HEC-PAC model with old and new first pass

4. Conclusions

- We propose a two-pass HEC (hybrid and E2E cascading) framework to combine a hybrid and an E2E model, with the hope to keep the key advantages of each system.
- The proposed system shows 8~10% relative lower WER than each of the individual systems.
- The second pass model is robust with respect to the change of the first pass model.