End-to-End ASR TTS Report 1

Shigeki Karita*

2018 年 9 月 27 日

概要

2018/09/27 meeting

Overview

- INTERSPEECH2018: combining ASR and text-to-text (T2T) with inter-domain loss (KL, MMD) to train with unsupervised datasets in WSJ
- NEXT: combining ASR, text-to-speech (TTS), T2T, speech-to-speech (S2S) with MMD to improve unsupervised learning in Librispeech

I call this model ASR/TTS/T2T/S2S

- ASR/TTS task Librispeech train_clean_100
- S2S task Librispeech train_clean_360
- T2T task Librispeech train_other_500

Issues

- (Librispeech ESPnet baseline) I could not run unigram 5000 model because of its GPU memory requirement
- Instead of unigram baseline, I used char baseline here
- ASR/TTS/T2T/S2S is very slow (4-6 times slower than ASR only. about 10 days)
  - also char model
  - smaller minibatch training (current setting: 20 samples x 4 tasks)
  - difficult to find the best learning rate for each tasks (current setting: ASR 1e-3, TTS 1e-3, S2S 1e-4, T2T 1e-4)

Results

- char-based ASR baseline (Librispeech clean 100)

* karita.shigeki@lab.ntt.co.jp

- char-based ASR/TTS/T2T/S2S without MMD
- char-based ASR/TTS/T2T/S2S with MMD

We need discussion what to investigate (too many combination)

表 1 current running experiments (WIP)

name	ASR	TTS	S2S	T2T	MMD
ASR (baseline)	1	0	0	0	0
ASR/T2T (INTERPSEECH2018)	1	0	0	1	1,0
ASR/S2S	1	0	1	0	1,0
ASR/S2S/T2T	1	0	1	1	1,0
ASR/TTS	1	1	0	0	0
ASR/TTS/S2S/T2T	1	1	1	1	1,0

WIP results

name	dev_clean Acc	dev_clean CER	test_clean CER	dev_clean WER	test_clean WER
ASR (baseline)	87.5	9.4	9.1	24.3	23.6
ASR/TTS/S2S/T2T with MMD	86.0	14.7	15.4	27.0	27.7
ASR/TTS/S2S/T2T without MMD	85.7				

???

- End-to-End ASR: argmax_t p(t|s)
- End-to-End ASR with LM: argmax_t p(t|s) p(t) <- ???
- DNN-HMM hybrid ASR: argmax_t p(s|t) p(t)

p(s|t) can be probabilitic end-to-end TTS model?

- End-to-End ASR with TTS-LM: argmax_t p_asr(t|s) p_tts(s|t) p_lm(t)

Emacs 25.2.2 (Org mode 9.0.3)