The AlexHT system for FSR Challenge

Hsiao-Tsung Hung

Tainan, Taiwan

AlexHT_Hung@asus.com

Abstract

This paper describe the AlexHT system for Formosa Speech Recognition (FSR-2018) Challenge. Our concern is to investigate the applicability of end-to-end automatic speech recognition (ASR) systems to transcribe Mandarin broadcast radio. Furthermore, we attempt to explore better ESPnet receipts that compatible with Kaldi ones.

Several features were analyzed, including traditional MFCC39, high resolution MFCC and FBank features. The experimental results shows that 80 dimension FBanks was appropriate when BLSTMP encoder was chosen. We also try to apply an RNN language model, which is trained from CLMAD news corpus, but it drags down accuracy in every condition.

Index Terms: speech recognition, ESPnet

1. Introduction

There are a number of ways in which the vocabulary increases, and some of these words were pronounced in unusual way. These situation highlight the difficulty of maintaining a Mandarin pronunciation dictionary. For instance, ”藍瘦香菇”(feeling sad) is a mispronounced word by a weeping man. If we apply typical receipts to build up systems, these words may be either dropped or trained with improper phone labels. Moreover, code-switching utterances are frequently used in daily life, and each language has many different dialect regions in Taiwan.

Fortunately, the area of sequence-to-sequence modeling with attention on ASR tasks has been widely studied in recent years. The main advantage of these methods is reducing the effort of maintaining pronouncing lexicons. For above reasons, this work forces on building an end-to-end speech recognition system for FSR-2018 challenge.

2. System Overview

All systems were developed using the ESPnet[1] and Kaldi[2] toolkits. Most of settings were follow-up to the character-RNNLM based receipts in ESPnet. Our experimental script can be found in the website[1]. Figure 1 shows the flow of our receipts, and details are described as follows.

2.1. Acoustic features

Several features were compared in experiments, including

- MFCC39: MFCC13 with deltas and delta-detlas,
- MFCC40+Pitch: 40 dimension MFCC features and 3 dimension pitch features,
- FBANK80+Pitch: 80 dimension Mel-filterbank features and 3 dimension pitch features.

In order to compare above feature sets, we chose BLSTMP as encoder instead of VGG-BLSTMP. Our submitted system was trained with MFCC39, but we would not recommend it.

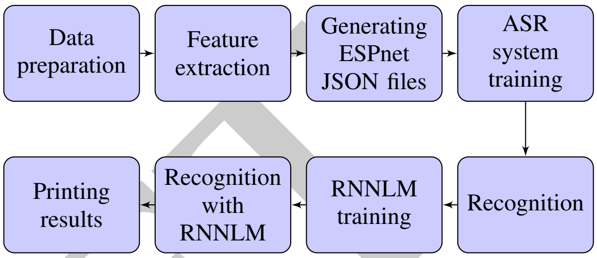

Figure 1: *flowchart of the* ***run.sh*** *script.*

Name	Shows
Test	JZ, GJ, KX, YX
FSR-final-test	ET, JK, QN

Table 1: *Data sets used for evaluation.*

2.2. Language modeling

ESPnet provides both word and character-based language modeling scripts. Since Chinese characters are logographs, we believe that character-based language model with rich character level input features are potential for further investigations.

The CLMAD[3] corpus was added to train the character-based LSTM-RNN language model. These documents were translated from simplified Chinese to traditional Chinese by *cconv*[4] tool, and word sequences were split into character sequences.

3. Experiments

All experiments were done in same encoder-decoder network architectures. The encoder architecture is a 4-layer BLSTMP, and each layer contains 320 units and a 160 dimension projection layer. The decoder network has a CTC network and an attention decoder network. Further details of architectures and algorithms are found in [5].

Only NER-Trs-Vol1 speech corpus were used in training stages. Data sets are organized into *train* and *test* sets likes baseline scripts, and the *test* is including four shows as listed in Table1. The submitted system were trained with whole *train* set, and validated in *test* set in training phase. In following experiments, 1% utterences in *train* set are split out for cross-validation purposes.

One should note that all character error rate are computed by *SCLITE* without any prepossessing, but submitted results will be normalized[2] before evaluation.

[1] https://github.com/alex-ht/formosa_espnet

	Test set		FSR-final-test	
Feature \Epoch	16	30	16	30
MFCC13+Δ+$\Delta\Delta$	23.9	**21.2**	21.2	21.1
MFCC40+Pitch(3)	23.4	21.4	21.2	19.2
FBank80+Pitch(3)	**21.9**	22.2	**18.9**	**18.8**

Table 2: *Character error rates(CERs) for* test *set in* NER-Trs-Vol1 *and* final-test *set in FSR challenge.*

Feature	wo/RNNLM	w/RNNLM
MFCC13+Δ+$\Delta\Delta$	23.9	24.3
MFCC40+Pitch(3)	23.4	24.3
FBank80+Pitch(3)	21.9	22.7

Table 3: *Character error rates(CERs) for* test *set in* NER-Trs-Vol1. w/RNNLM *and* wo/RNNLM *denote respectively the results are decoded with and without RNNLM.*

3.1. Effects of resolution on the accuracy and generality

First experiment was designed to explore the effects of resolution of acoustic features on the accuracy and generality. Three difference feature sets are included and compared under two training epoch settings. The settings of 16 training epochs can be viewed as a kind of early stopping strategy to prevent overfitting. The experimental data are presented in Table 2.

Before receiving the FSR final-test data, our system was tuned according to character error rates of *test* set in NER-Trs-Vol1. It appears that traditional MFCC39 feature set is most simpler and easy to train on a medium size corpus. However, table 2 shows consistent results across a broad range of parameter choices and implies that properties of final-test data are close to training data.

Last row of table 2 shows that early stopping method is important when high resolution acoustic features are applied. Note that ESPnet using a find-tuned parameter scheduling method for Adadelta optimizer, so there is room for further investigation of over-fitting issue.

3.2. Effects of RNNLMs

Since we don't have any popular text corpus, likes Chinese Gigaword, to train a language model, the following experiment was performed using the CLMAD[3] corpus. CLMAD is an open Chinese Language Model Adaptation Dataset, and contains 14 classes of 740,000 news. The network architecture of language model is two stacked LSTMs with 320 units in each layer. Table 3 shows the experimental results on the character-based RNNLM.

Obviously, the character-based RNNLM drags down accuracy in every condition. At first thought, the fundamental problem should be the mismatch of word usage between two corpora. Another problem is caused by the natural of end-to-end system. In this work, the training data are limited and only contains about 4,000 Chinese characters. Therefore, it is hard to improve generality of the end-to-end system without adding more training speech data.

Although several methods are examined in this work, but all of them performed worse then the baseline *chain-tdnn-1a-sp*[3] model. Our findings, however, are subject to many limitations from toolkit. ESPnet implement a Hybrid CTC and Attention based decoders, but Chain model appears to be worth trying to combined with Attention-based model.

System	CER	Ins.	Del.	Sub.
baseline	25.0	3.3	4.1	9.2
M1	23.2	2.4	**2.4**	8.9
iFlytek	**18.8**	**2.3**	5.0	**5.9**

Table 4: *FSR-2018 resluts*

3.3. Submitted system

Out submitted system was trained with MFCC39 features as described previous chapters. The main difference is that the weight of CTC is 0.3 in testing phase. Since the joint decoder is much slower then Attention-based decoder, we don't take the discussion further in this paper. Results are presented in Table 4 and our system is denoted as M1.

4. Conclusions

The aim of this work is to investigate the applicability of end-to-end ASR systems to transcribe Mandarin broadcast radio. From the present results obtained, it difficult to improve the End-to-end system without increasing training speech data. Several recommendations may be made for the further research. In acoustic modeling part, Zhou *et al.* suggest to replace characters with syllables[6]. Moreover, we could extend the research topic to multilingual speech recognition problem. Duo to the lack of complete pronunciation dictionary of dialects, it is also a suitable task for end-to-end based system.

5. References

[1] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 2207–2211.

[2] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[3] Y. Bai, J. Tao, J. Yi, Z. Wen, and C. Fan, "Clmad: A chinese language model adaptation dataset," in *The Eleventh International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2018.

[4] xiaoyjy. (2015) A iconv based simplified-traditional chinese conversion tool. [Online]. Available: https://github.com/xiaoyjy/cconv

[5] T. Hori, S. Watanabe, and J. R. Hershey, "Multi-level language modeling and decoding for open vocabulary end-to-end speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec 2017, pp. 287–293.

[6] S. Zhou, L. Dong, S. Xu, and B. Xu, "Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese," in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 791–795.

[2] https://github.com/106368024yuchenlin/asr-evaluation-for-FSR-2018-

[3] https://github.com/yfliao/kaldi/tree/master/egs/formosa