National Institute of Information and Communications Technology

同花顺
WWW.10JQKA.COM.CN

AN INVESTIGATION OF USING HYBRID MODELING UNITS FOR IMPROVING END-TO-END SPEECH RECOGNITION SYSTEM

Shunfei Chen[1], Xinhui Hu[1], Sheng Li[2] and Xinkang Xu[1]

[1]Hithink RoyalFlush AI Research Institute, Zhejiang, China

[2]National Institute of Information and Communications Technology (NICT), Kyoto, Japan

OUTLINE

1. Background
2. System description
3. Experiments and results
4. Conclusions

OUTLINE

1. Background
2. System descriptions
3. Experiments and results
4. Conclusions

Background

- The end-to-end (E2E) speech recognition has become popular and its performances can compete with that of traditional automatic speech recognition (ASR).

- The E2E ASR greatly simplifies the ASR modeling process , and lexicon and language model are not needed.

- However, modeling unit is still important and necessary for E2E ASR.

Background

- In E2E ASR systems: there are more choices than in the traditional DNN-HMM ; such as

~~CD-phone:~~ HELLO → ~~sil-HH+AH0 HH-AH0+L AH0-L+OW1 L-OW1+sil~~

which is widely used in the traditional DNN-HMM based ASRs, but is rarely used in E2E ASRs.

character: HELLO → H E L L O

word: HELLO → HELLO

subword: HELLO → _HE LLO

Disadvantage: OOV problem is easily happened.

For English, the subword is the most used unit for E2E ASR systems.

- Current situations of the modeling units for E2E Mandarin ASR:

  1) most of the studies focus on individual units [1,2]
  such as character, subword, syllable, etc.

  2) few researches pay attention to using different units' combinations. But different modeling units have their own disadvantages, such as character causes data sparseness problem and syllable difficult to distinguish homophones.

- Purpose of this study

  From the viewpoints of taking advantages of different modeling units, we propose to apply the hybrid units to a CTC/attention multi-task learning architecture.

[1] Chiu C C, Sainath T N, Wu Y, et al. State-of-the-art speech recognition with sequence-to-sequence models[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 4774-4778.
[2] Zhou S, Dong L, Xu S, et al. A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin chinese[C]//International Conference on Neural Information Processing. Springer, Cham, 2018: 210-220.

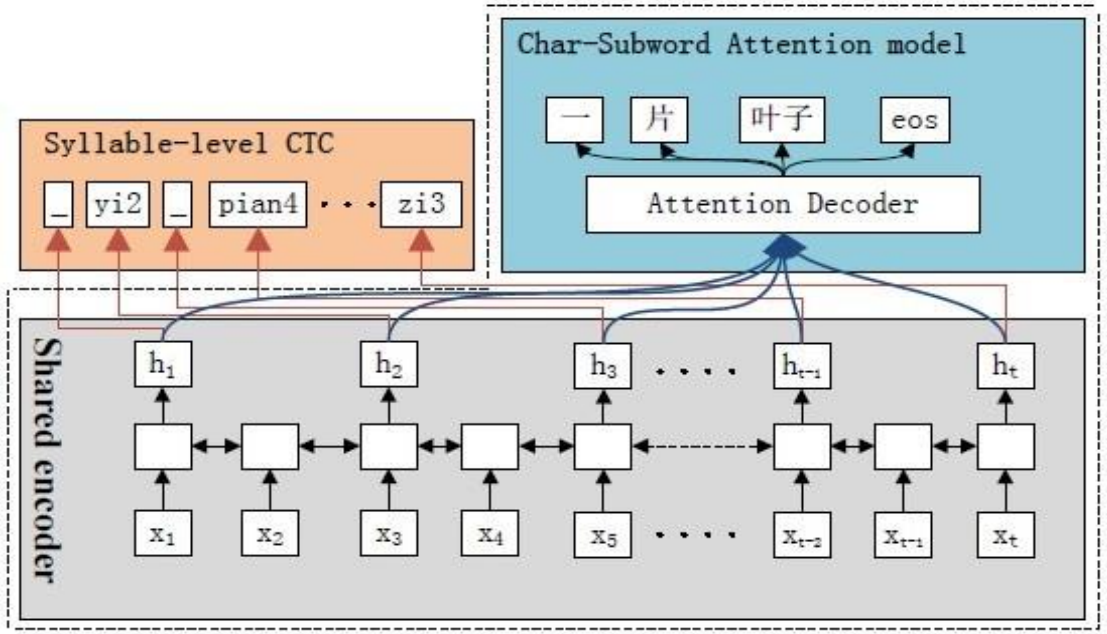
OUTLINE

1. Background
2. System descriptions
3. Experiments and results
4. Conclusions

System descriptions

Overview

- character and subword are used as grapheme modeling units for the attention decoder
- Mandarin syllable is used as an unit for the CTC module
- During training stage， the joint CTC/Attention multi-task learning is adopted.
- During inference stage，the attention decoder’s output is directly used as the recognition result.



System descriptions

The reasons to add Mandarin syllable as the modeling unit:

- Mandarin is a syllable-based language and syllable is the logical unit of pronunciation.
- Different from English where a word may map to several syllables, each Chinese character only maps to a tonal syllable
- The advantage of syllable: 1) eliminating OOV; 2) dealing with data sparseness for less common characters
- The disadvantage of syllable: 1) needing a lexicon (the same as in CD-phone,); 2) difficult to distinguish homophones

In this study,

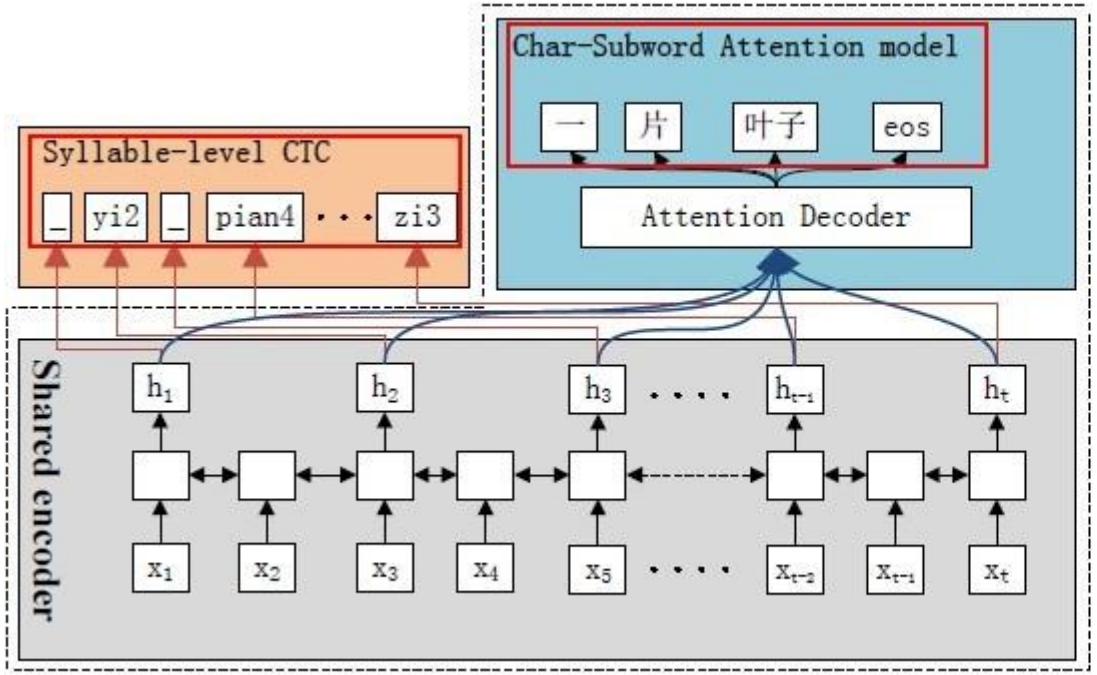
- Mandarin E2E ASR modeling units include character, subword, **syllable**

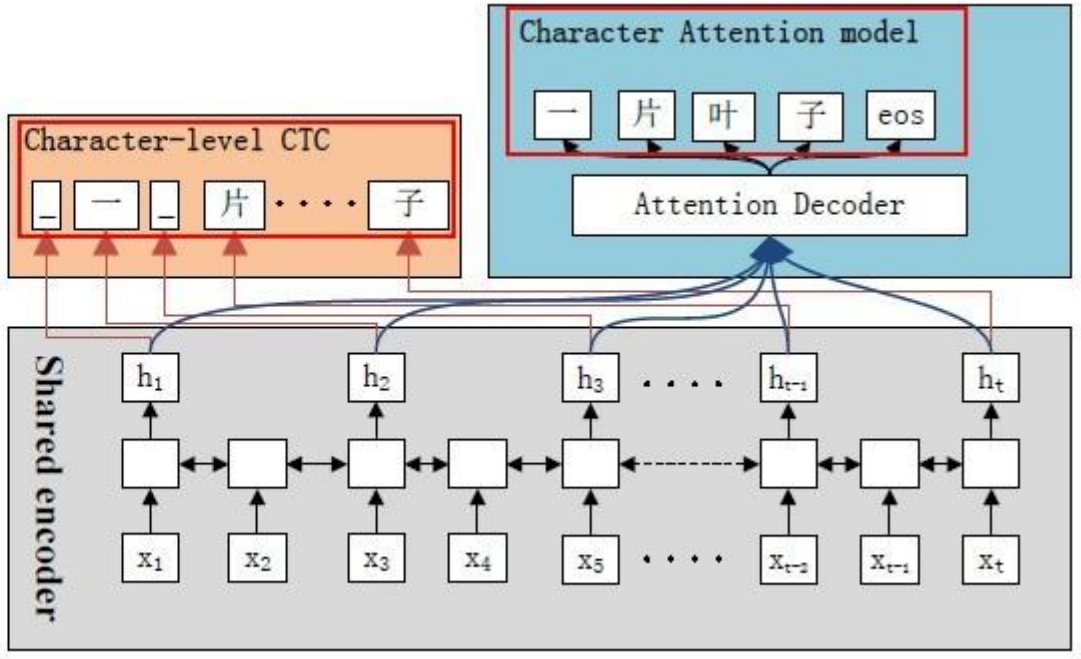
System descriptions

Innovations of our system:

1. char-subword-based attention model : instead of character only, character and subword are mixed and are used as units in attention model.

2. Syllable-level CTC : instead of character, syllable unit is used in CTC module.





System descriptions

Char-subword-based attention model

- Why subword is added? subword has more contextual information than character.

- How to build char-subword ?
  The **character unit** set A is built by collecting all characters in the training transcripts.
  The **subword unit** set B is obtained by using BPE algorithm first, then selecting high-frequency subwords, and deleting those single characters.
  The final char-subword unit is composed of A and B.

- The difference between char-subword and BPE:
  In char-subword, all Chinese characters are ensured to be used,
  but in BPE, only **high frequency** subwords (including characters）are used.

Table 1. Examples of different modeling units. The original sentence is ”一片叶子” which means ”a piece of leaf.”

Modeling Unit	Converting Results
syllable without tone	yi pian ye zi
tonal syllable	yi2 pian4 ye4 zi3
character	一 片 叶 子
subword	一片 叶子
char-subword	一 片 叶子

System descriptions

Syllable-level CTC

- Why syllable is needed? :
  1) The Mandarin syllable is the basic unit of its speech, and has a fixed number.
     There are about 400 syllables without tone and 1500 syllables with tone (include light tone).
  2) Each Chinese character corresponds to a tonal syllable. Using character as an unit will cause OOV problem
     or data sparseness problem for those low frequency characters.
  3) A syllable is generally shared by many characters.

- We conducted experiments with tonal syllable and non-tonal syllable in this study.

Table 1. Examples of different modeling units. The original sentence is "一片叶子" which means "a piece of leaf."

Modeling Unit	Converting Results
syllable without tone	yi pian ye zi
tonal syllable	yi2 pian4 ye4 zi3
character	一 片 叶 子
subword	一片 叶子
char-subword	一 片 叶子

System descriptions

How to combine different modeling units

- Joint CTC/Attention multi-task learning

- 1) Why not directly mix syllable with character + subword as modeling unit ?
  - An additional pronunciation dictionary module is necessary to convert syllables to characters in decoding stage.

  In this study, we ignore the output of syllable in CTC decoder, and use only the attention decoder for output.

  2) Why do we use syllable in CTC rather than in attention model?
  - The shared-encoder in the transformer plays the listener's role, and the decoder is a speller. Using a syllable unit to train the shared-encoder can make the shared-encoder more robust for distinguishing different syllables.
  - The alignment estimation effect of CTC is better than the attention model.

OUTLINE

1. Background

2. System descriptions

3. Experimental settings and results

4. Conclusions

Experiments

- Data:
  - Mandarin Corpus from the OpenSLR
- Experimental Setup：
  - Acoustic feature: Fbank + pitch (83-dimension)
  - Network hypermeters:
    - 12 encoder blocks and 6 decoder blocks
    - 4 heads multi-head attention with 256
    - Adam optimizer
    - Warmup = 25000
    - Dropout = 0.1

Table 1. Data sets for experiments

dataset name	training set	val. set	test set
AISHELL-1	**150h**	**18h**	**10h**
ST-CMDS	102h	2.64h	5h
Primewords	90h	2.77h	6.25h
aidatatang 200zh	140h	20h	40h
MAGICDATA	712h	14h	28h
Total	1194h	57.41h	89.25h

Experimental results

- Results on Aishell-1

Here，

- CER1 is the character error rate by using LM and CTC
- CER2 is the character error rate without using LM and CTC weight is zero.

The recognition results are shown in the Table 2. We can see that

- The char-subword is slightly better than the character and subword.
- The size of subwords will affect the recognition results

Table 2. Recognition results on AISHELL-1: Comparisons among the char-subword and separating units

Modeling Unit	CER1		CER2	
	val.	test	val.	test
Character[1]	6.00	6.70		
character	6.03	6.68	6.70	7.61
subword(BPE)	5.82	6.52	6.68	7.65
char-subword(460)	5.84	6.52	6.68	7.65
char-subword(200)	5.78	6.45	6.61	7.49

[1] Karita S, Chen N, Hayashi T, et al. A comparative study on transformer vs rnn in speech applications[C]//2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019: 449-456.

Experimental Results

- Results on Aishell-1

  - In CER1, syllable$_{tone}$-char-subword is better than char-subword or subword.

  - In CER2, either syllable$_{tone}$-char-subword or syllable$_{notone}$-char-subword has a significant reduction compared with char-subword or subword.

  - **Therefore, we regard that, in the cases of no LM is used, the model trained by syllable-char-subword is more robust than models trained by the other units.**

Table 3. Recognition results on AISHELL-1:

Modeling Unit	CER1		CER2	
	val.	test	val.	test
character[1]	6.00	6.70		
character	6.03	6.68	6.70	7.61
subword(BPE)	5.82	6.52	6.68	7.65
char-subword(460)	5.84	6.52	6.68	7.65
char-subword(200)	5.78	6.45	6.61	7.49
syllable$_{notone}$-char-subword	5.79	6.37	6.10	6.91
syllable$_{tone}$-char-subword	5.77	6.32	6.02	6.73

Experimental Results

- Results on the OpenSLR
  - The syllable-char-subword-based model outperforms the others.
  - We also find that the final performance changes with the CTC weight during the training stage, and the best weight is 0.2

Table 4. The results of different modeling units and performance changes with different CTC weight in training

Modeling Unit	CTC weight	AISHELL		ST-CMDS		Primewords		aidatang_200zh		MAGICDATA		Average	
		val.	test	val.	test	val.	test	val.	test	val.	test	val.	test
character	0.3	5.87	6.37	7.71	8.57	15.33	15.11	5.58	6.29	5.42	5.55	6.17	6.46
subword	0.3	5.72	6.36	7.58	8.62	15.03	14.77	5.65	6.30	5.49	5.57	6.14	6.45
char-subword	0.3	5.80	6.40	7.61	8.66	15.05	14.98	5.54	6.15	5.38	5.52	6.10	6.38
syllable-char-subword	0.3	5.31	6.04	7.18	8.06	13.78	13.92	**5.03**	**5.66**	5.79	**5.42**	5.75	5.96
syllable-char-subword	0.2	**5.27**	**5.94**	7.19	**7.91**	**13.65**	**13.89**	5.04	5.67	**5.38**	5.51	**5.66**	**5.96**
syllable-char-subword	0.1	5.31	6.00	**7.17**	8.17	13.85	13.96	5.13	5.78	5.66	5.45	5.76	6.03

Experimental Results

- Results on OpenSLR
  - Analysis of the errors in the results.
  - Sub1, sub2 refer to substitution errors by non-homophone characters and homophone characters.

  - In table 5, Compared with char-subword, the relative reductions of the errors corresponding to sub, sub1, and sub2 in the case of syllable-char-subword are 8.35%, 9.88%, and 5.14%, respectively.
  - **So, syllable is regarded as being able to reduce the substitution errors.**

Table 5. the detail of different error character on test set: insert, delete and substitute.

Modeling Unit	ins	del	sub	sub1	sub2
character	1643	4627	53992	35676	18316
subword	1699	4968	53520	35631	17889
char-subword	1782	5252	52537	35545	16992
syllable-char-subword	2468	5019	48148	32030	16118

OUTLINE

1. Background
2. System Introduction
3. Experiments and Results
4. Conclusions

Conclusions

From this study, we obtained following findings:

- In this work, we **proposed** a hybrid modeling unit of syllable-char-subword in a joint CTC/Attention multi-task learning framework for the Mandarin E2E ASR system.
- With the addition of syllable and subword to the modeling unit of character, the trained model becomes **more robust** than using the other modeling units
- **In particular**, the substitution errors are considerably reduced.
- **In our experiments**, using the syllable-char-subword hybrid modeling unit can achieve 6.6% relative CER reduction on our 1200-hour data compared with the conventional units of char-subword (from 6.38% to 5.96%).
- **In the future**, we plan to do some experiments about adding a module to convert syllable to character for the output of CTC.

THANKS for your attention

Contact Email: *chenshunfei@myhexin.com*, *huxinhui@myhexin.com*, *sheng.li@nict.jp*, *xuxinkang@myhexin.com*