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

Perspective

- **Background**
  - Modeling unit is still important and necessary for E2E ASR.
  - Mandarin is a syllable-based language. Using syllable unit can solve OOV and data sparseness problems.
  - Different modeling units have their own advantages and disadvantages.
- **Research Status**
  - Most of the studies on Mandarin ASR focus only on an individual unit.
  - Few attention are paid to using different units' combinations for the E2E ASR systems.
- **Objective**
  - From the viewpoints of taking advantage of different modeling units, we propose to apply the hybrid units to a CTC/attention multi-task learning architecture.

Modeling units

- **Character:** There are tens of thousand Chinese characters, in which only about 6700 are commonly used.
- **Syllable:** About 400 syllables without tone and 1500 syllables with tone . Syllable has a strong disambiguation effect in Mandarin speech.
- **Char-subword**: A new grapheme unit that we proposed in this study, and it is different from the conventional BPE.

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 description

- **Overview**
  - Character and subword are used as grapheme modeling units for the attention decoder.
  - Mandarin syllable is used as an unit for the CTC.
  - Training stage: joint CTC/Attention multi-task learning is adopted.
  - Inference stage: the attention decoder's output is directly used as the recognition result.
- **Innovations**
  - char-subword-based attention model : Instead of only using character, subword is added with it, they are mixed as the units in attention model.
  - Syllable-level CTC : Instead of character, syllable unit is used in the CTC module.
- **System construction**
  - How to build char-subword

The **character unit** set is built by collecting all characters in the training transcripts. The **subword unit** set is obtained by using BPE algorithm first, then selecting high-frequency subwords, and finally deleting those single characters.

- Syllable-level CTC

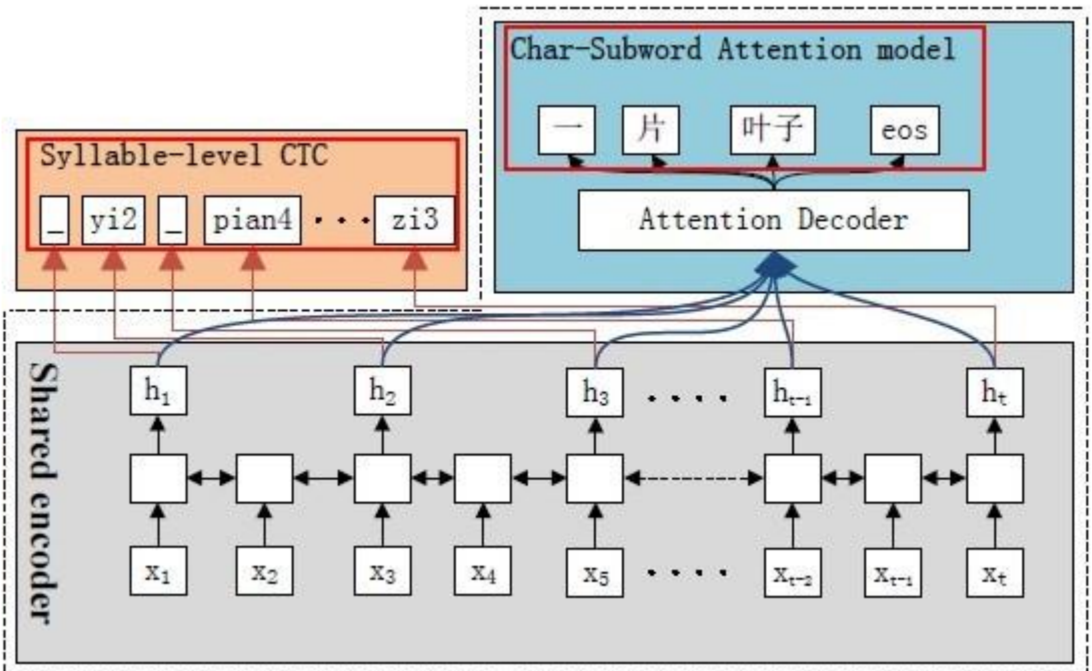
Each Chinese character corresponds to a tonal syllable. Using character as an unit will cause OOV problem or data sparseness problem for those low frequency characters.

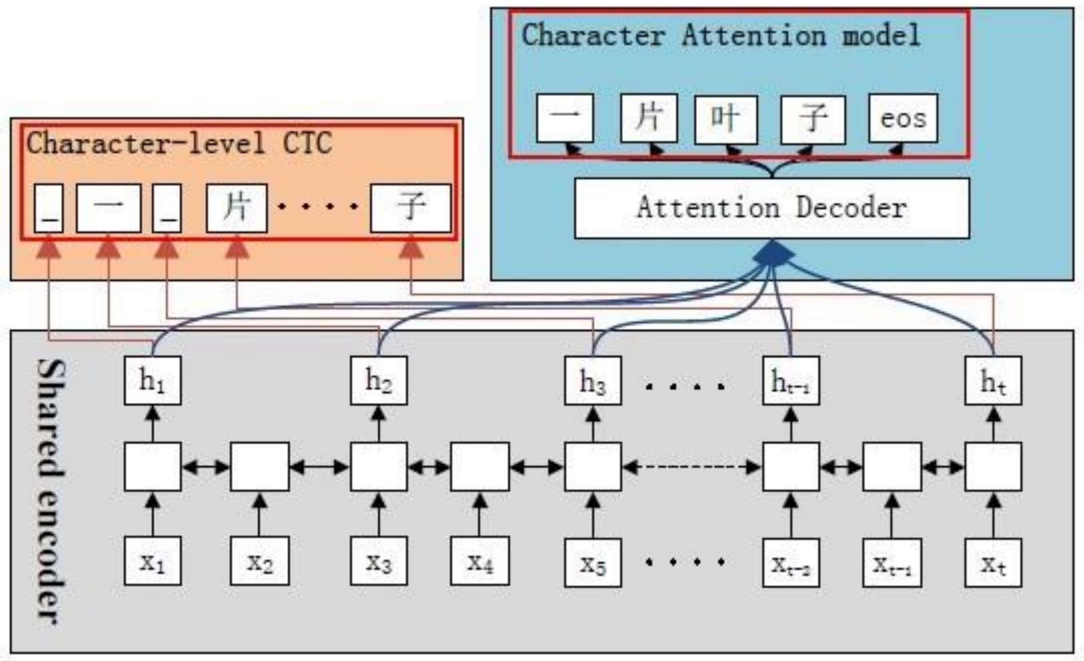
- How to combine different modeling units

□ It is realized by the joint CTC/Attention multi-task learning architecture.

- The advantages of using multi-task learning

1) If directly mixing syllable with character + subword, an additional lexicon module is necessary to convert syllables to characters.
2) Using the syllable unit to train the CTC module can make the shared-encoder more robust for distinguishing different syllables, and further benefit the attention decoder.





Experiments and results

- **DataSet**: Mandarin Corpus from OpenSLR
- **Model**: Transformer(12 ecoder blocks+6 decoder blocks)
- **Results On Aishell-1**:
  - **CER1**: character error rate (CER) by using LM and CTC; **CER2**: CER without LM;
  - char-subword is shown better than character and subword
  - 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 effective than models trained by the other units.

Table 3. the results of different modeling unit on AISHELL-1

Modeling Unit	CER1		CER2	
	val.	test	val.	test
character[3]	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**

- **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 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

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

- **Analysis of the error in the results on OpenSLR**
  - Sub1, sub2 refer to substitution errors by non-homophone characters and homophone characters, respectively.
  - 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-sub-word are 8.35%, 9.88%, and 5.14%, respectively.
  - So, syllable is regarded as being able to reduce the substitution errors effectively.

Conclusions

- With the addition of syllable and subword to the modeling unit of character, the trained model becomes **more robust** than using the individual ones or char-subword combination.
- **In particular**, the substitution errors are considerably reduced with the addition of syllable unit.
- **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 unit of char-subword **(from 6.38% to 5.96%).**
- **In the future**, we plan to do some experiments utilizing the output of CTC.