ASRU中英混杂语音识别挑战赛线下技术交流会

Towards Language-Universal Mandarin-English Speech Recognition Using DFSMN-CTC-sMBR

张仕良

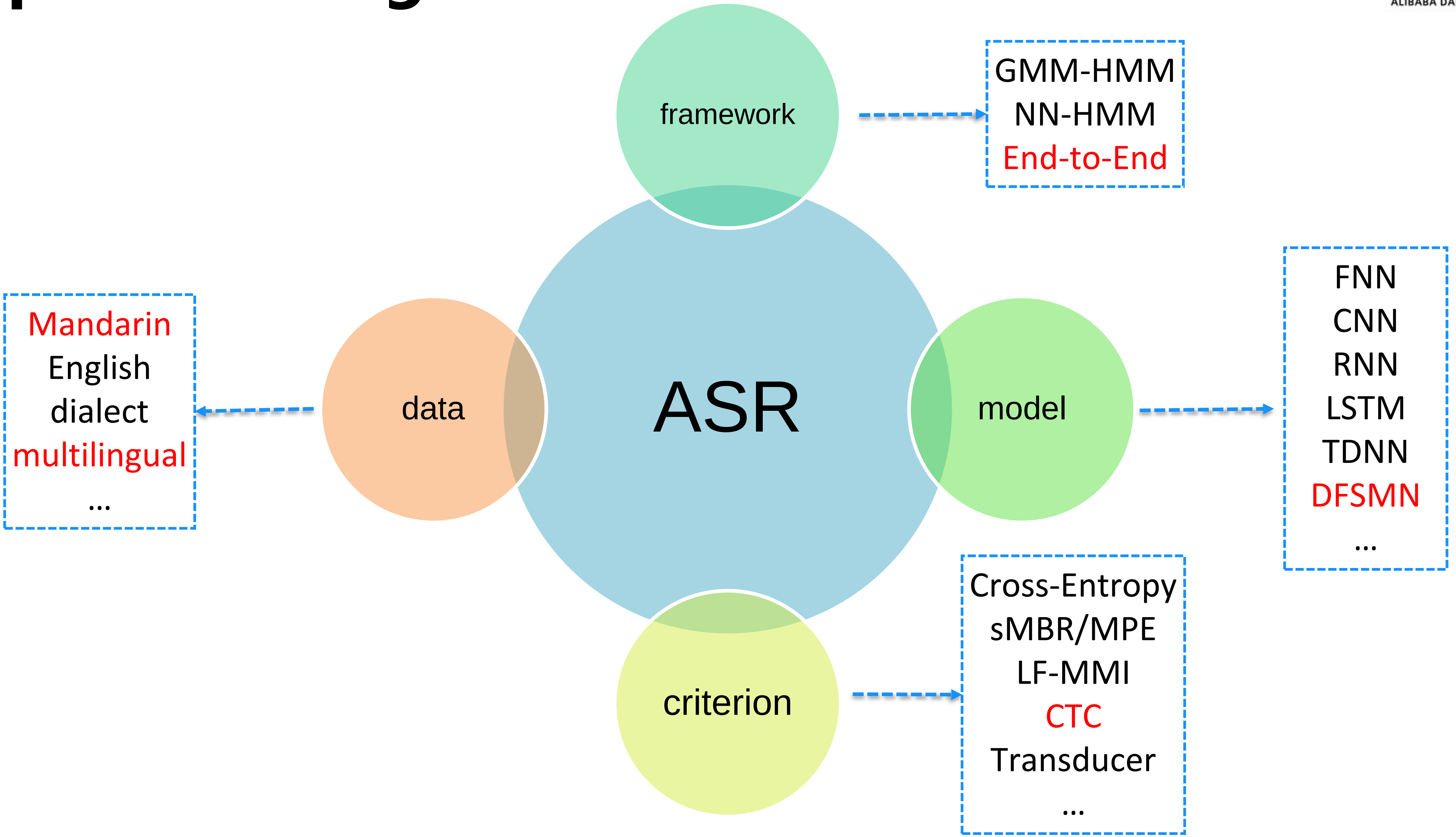
阿里巴巴-机器智能技术-智能语音交互团队

2019年11月23日

Main

- **Mandarin Speech Recognition using DFSMN-CTC-sMBR**
  - Zhang S., Lei M., Acoustic Modeling with DFSMN-CTC and Joint CTC-CE Learning. In *Interspeech* 2018 (pp. 771-775).
  - Zhang S., Lei M., Liu Y, et al. , Investigation of Modeling Units for Mandarin Speech Recognition Using DFSMN-CTC-sMBR. Proc. of ICASSP 2019: 7085-7089.
- **Language-Universal Mandarin-English Speech Recognition**
  - Zhang, S., Liu, Y., Lei, M., Ma, B., Xie, L. , Towards Language-Universal Mandarin-English Speech Recognition. Proc. of Interspeech 2019, 2170-2174.

Speech Recognition



ASR

framework

GMM-HMM
NN-HMM
End-to-End

data

Mandarin
English
dialect
multilingual
…

model

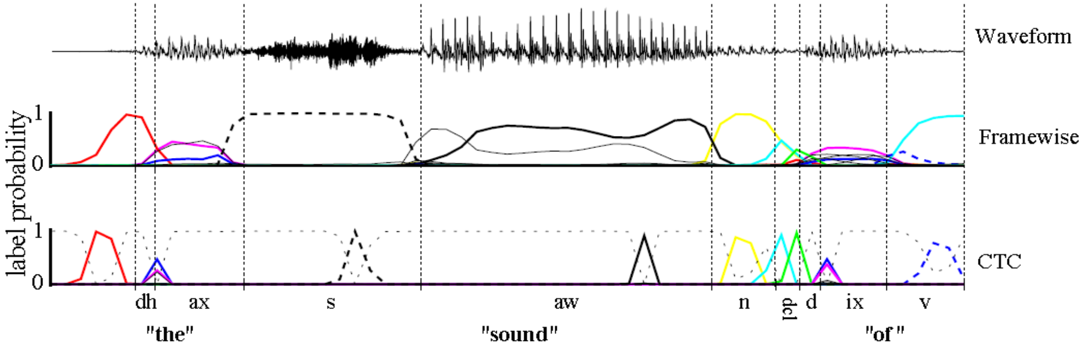
FNN
CNN
RNN
LSTM
TDNN
DFSMN
…

criterion

Cross-Entropy
sMBR/MPE
LF-MMI
CTC
Transducer
…

CTC based Acoustic Modeling

- Connectionist Temporal Classification
  - First proposed by Grave @2006 @2013
  - Successfully applied to LVCSR by Google @2016
  - LSTM/BLSTM-CTC



- CTC Vs. Cross-Entropy
  - Advantage：better performance, faster decoding speed
  - Problem： unstable, spike delay et al.
- CTC for Mandarin speech recognition
  - Deep Speech 2【2016】， Zhehuai Chen【2016】， Zhongdi Qu【2017】et al.
  - Zhang S., Lei M., Acoustic Modeling with DFSMN-CTC and Joint CTC-CE Learning. In *Interspeech* 2018 (pp. 771-775).
  - Zhang S., Lei M., Liu Y, et al. , Investigation of Modeling Units for Mandarin Speech Recognition Using DFSMN-CTC-sMBR. Proc. of ICASSP 2019: 7085-7089.

Mandarin Speech Recognition with DFSMN-CTC-sMBR

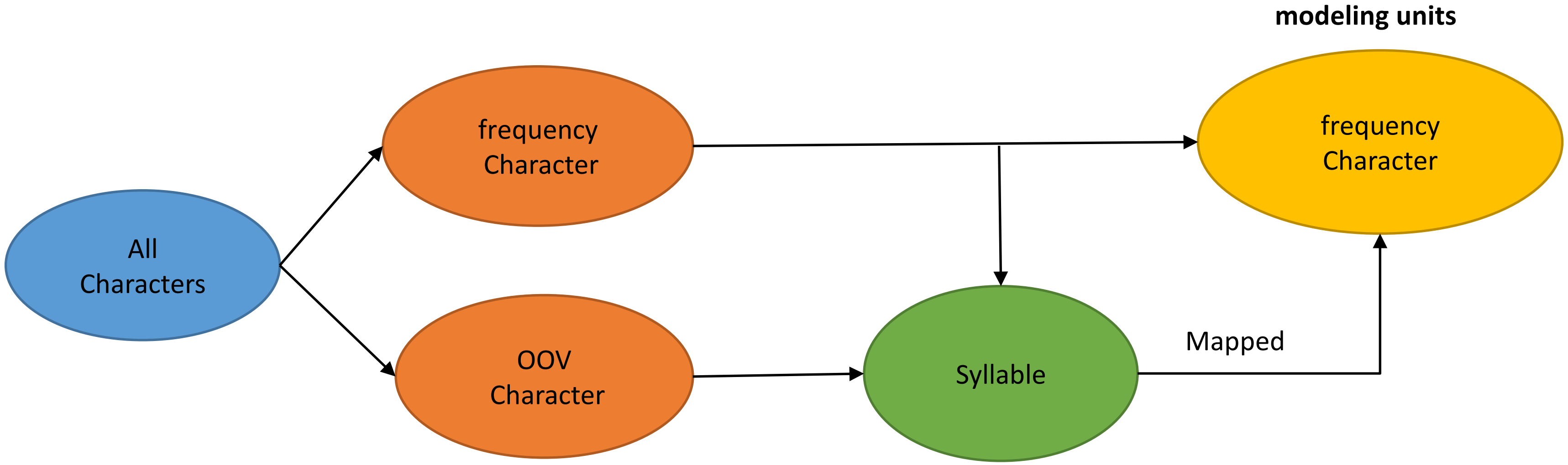
■ Acoustic Modeling units for Mandarin

序列(Sequence)	• 玩具总动员	数目无穷
词(word)	• 玩具 总动员	上百万
字(character)	• 玩 具 总 动 员	数十万，常用5千
音节(syllable)	• wan2 ju4 zong3 dong4 yuan2	一千多
绑定的声韵母(CD-IF)	• /s-w-an2 w-an2-j an2-j-u4 …	决策树, 数千
声韵母(IF)	• w an2 j u4 z ong3 d ong4 y uan2	143个

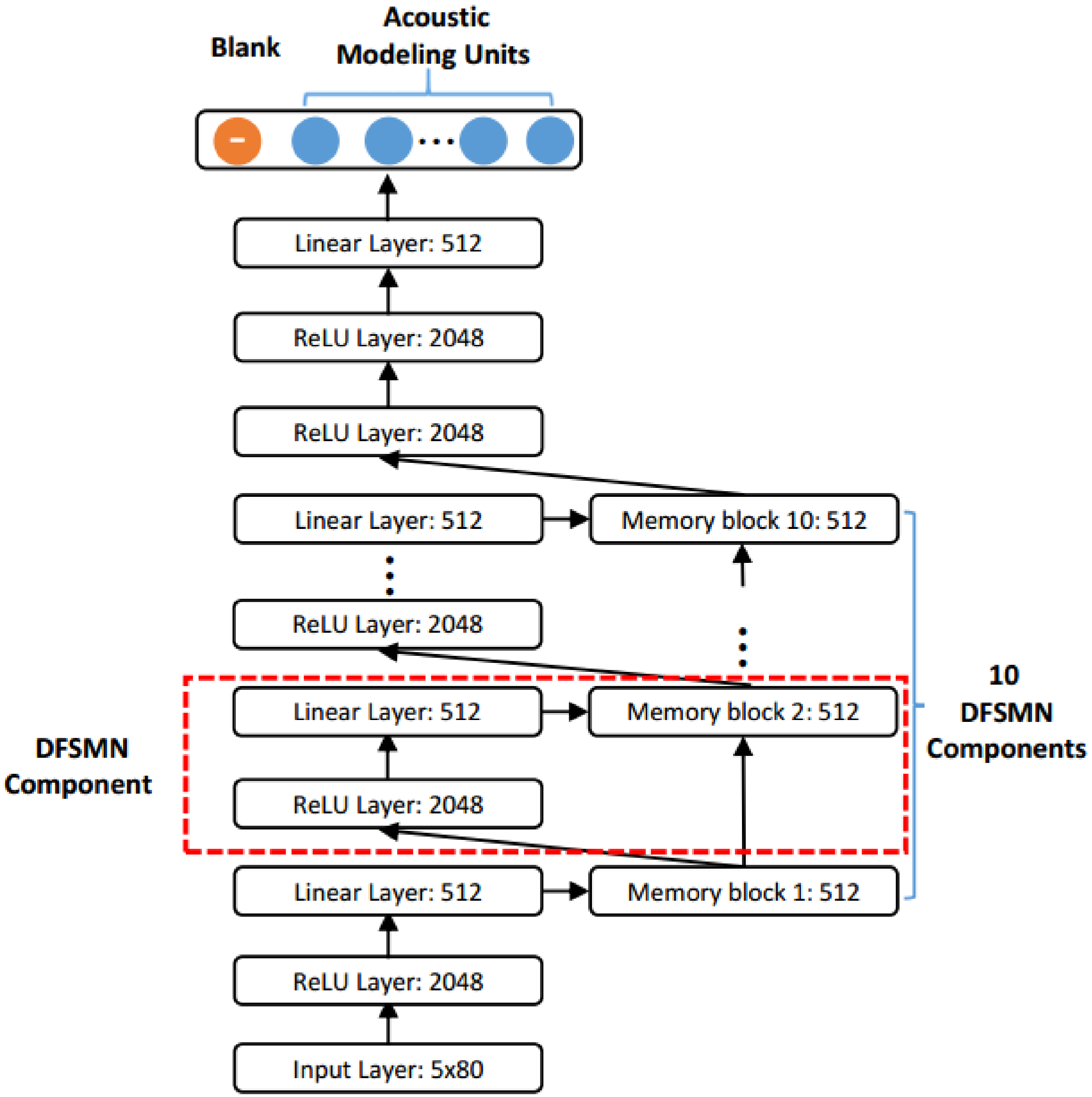
- 传统语音识别: 绑定的声韵母
- 基于CTC的识别模型：声韵母、绑定的声韵母、音节、字

Mandarin Speech Recognition with DFSMN-CTC-sMBR

- ■ Modeling units： Mixed-syllable-character
  - Method：frequency character + syllable
  - Advantage：without OOV problem； lower frame rate
- ■ Modeling units： All-Character
  - Method：frequency character+ character mapping
  - Advantage： small vocabulary size



Mandarin Speech Recognition with DFSMN-CTC-sMBR



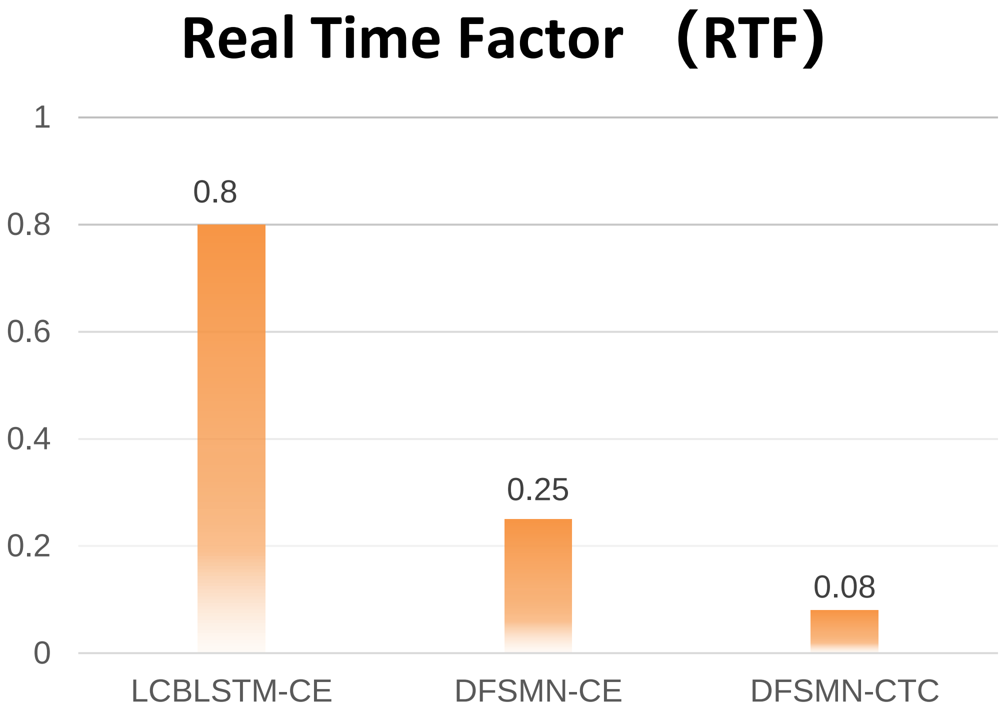
- Why DFSMN？
  - Feedforward architecture
  - Effectively model long-term dependency
- Acoustic Modeling units

Modeling units	Detailed composition
CI-IF	23 Initials + 185 tonal Finals
CD-IF	7951 context-dependent Initial/Finals
Syllable	1319 tonal syllables
Char(2k)-Syllable	2000 high frequency Chinese characters + 1319 tonal syllables
Char(3k)-Syllable	3000 high frequency Chinese characters + 1319 tonal syllables
All-Char(2k)	2000 high frequency Chinese characters

DFSMN-CTC-sMBR for Mandarin Speech Recognition

- 20000-hours-task

Exp	Model	Modeling Units	CER(%)	
			CE	+sMBR
1	LCBLSTM	CD-IF	11.32	10.59
2	DFSMN(10)	CD-IF	10.53	9.49
			CTC	+sMBR
3	DFSMN(10)	CI-IF	10.38	9.37
4	DFSMN(10)	CD-IF	9.70	-
5	DFSMN(10)	Syllable	9.03	7.94
6	DFSMN(10)	Char(2k)+Syllable	8.87	7.61
7	DFSMN(10)	Char(3k)+Syllable	8.81	7.45
8	DFSMN(12)	Char(3k)+Syllable	8.46	7.28
9	DFSMN(10)	All-char(2k)	8.08	6.98


Real Time Factor （RTF）

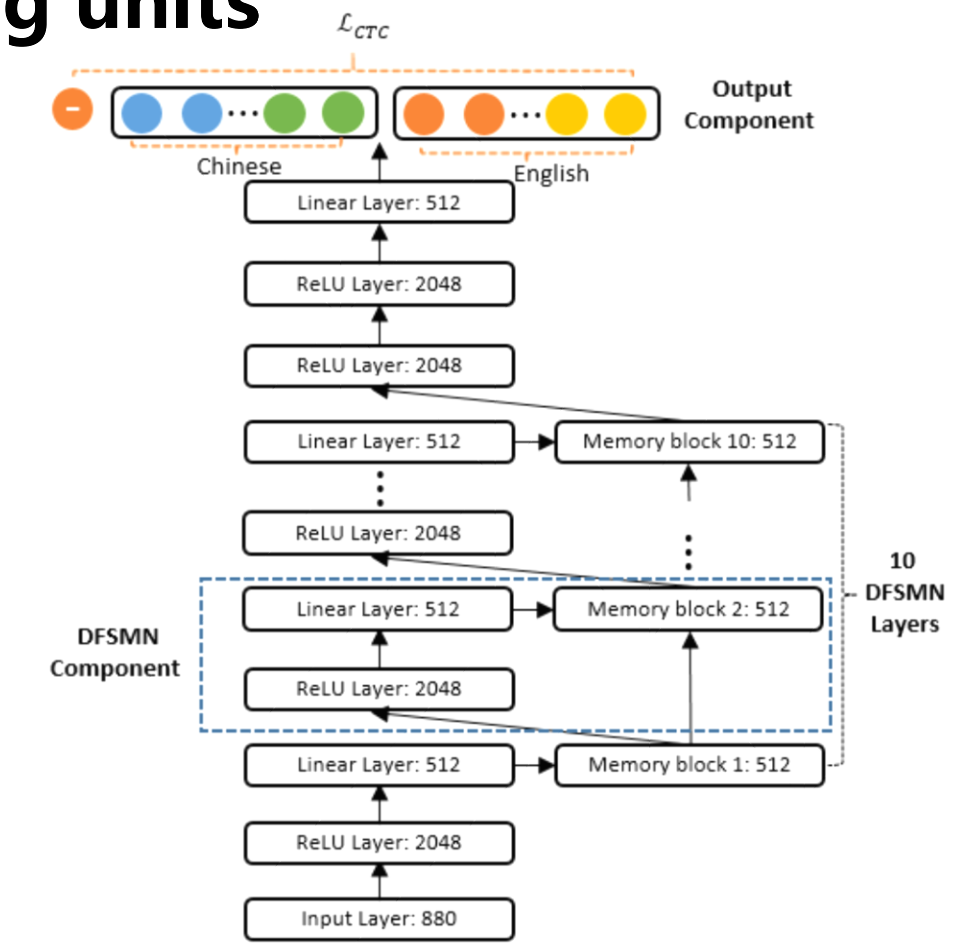
Language-Universal Mandarin-English ASR

- Background: language-specific ASR system
- Language-Universal Mandarin-English ASR System
  - ASR system can recognize Mandarin, English and code-switching speech without any language-specific information.
- Challenge
  - How to model two languages
  - Code-switching acoustic & text data
  - Discriminative acoustic score
- Our works
  - CTC-sMBR
  - Bilingual-AM
  - Word Space Mapping

Language-Universal Mandarin-English ASR

■ Idea 1、Mixed Acoustic Modeling units

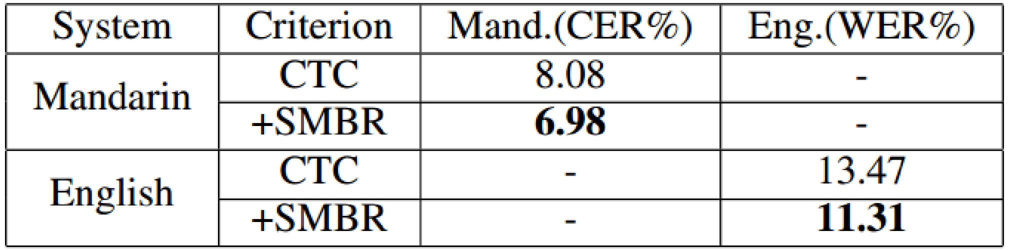
- Acoustic Model
  - DFSMN-CTC-sMBR
- Training Data
  - monolingual Mandarin + English
- Acoustic Modeling Units
  - Mandarin： All-character
  - English: wordpiece



Language-Universal Mandarin-English ASR

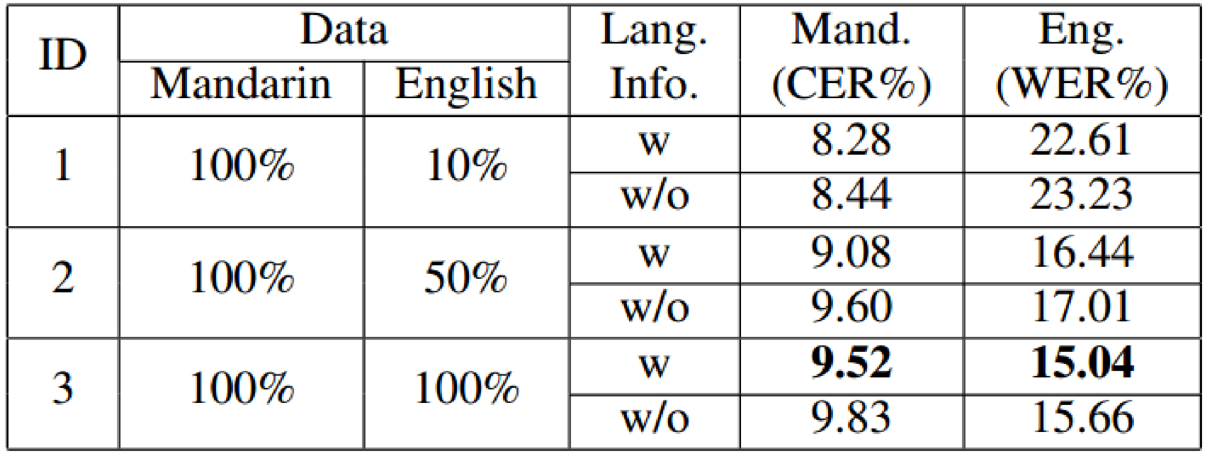
■ Idea 1、Mixed Acoustic Modeling units

- Training data
  - 20000-hours-Mandarin + 15000 hours English



System	Criterion	Mand.(CER%)	Eng.(WER%)
Mandarin	CTC	8.08	-
	+SMBR	**6.98**	-
English	CTC	-	13.47
	+SMBR	-	**11.31**

Language-specific DFSMN-CTC-sMBR system

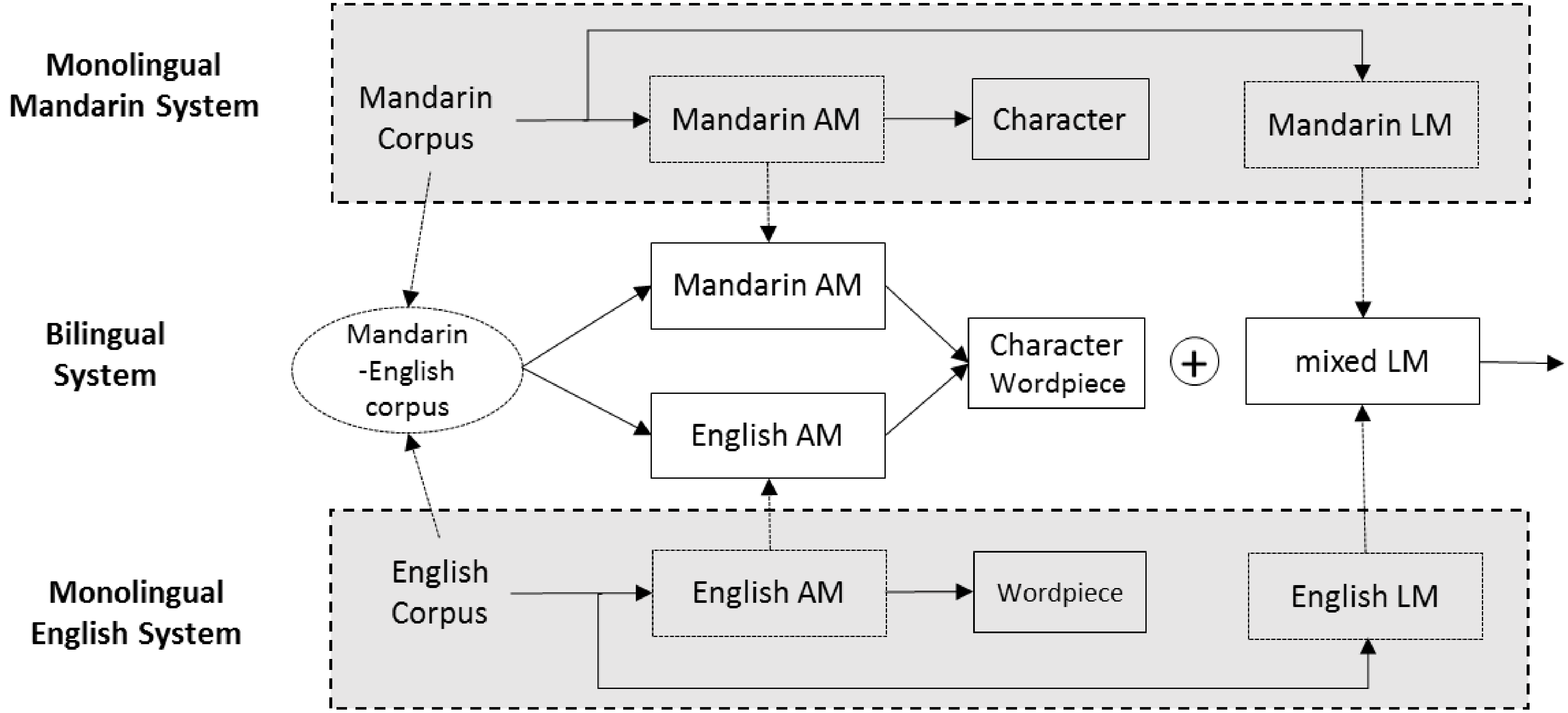


ID	Data		Lang. Info.	Mand. (CER%)	Eng. (WER%)
	Mandarin	English			
1	100%	10%	w	8.28	22.61
			w/o	8.44	23.23
2	100%	50%	w	9.08	16.44
			w/o	9.60	17.01
3	100%	100%	w	**9.52**	**15.04**
			w/o	9.83	15.66

Language-universal DFSMN-CTC system with mixed modeling units

Language-Universal Mandarin-English ASR

- Idea2、Monolingual pre-trained Bilingual-AM



Language-Universal Mandarin-English ASR

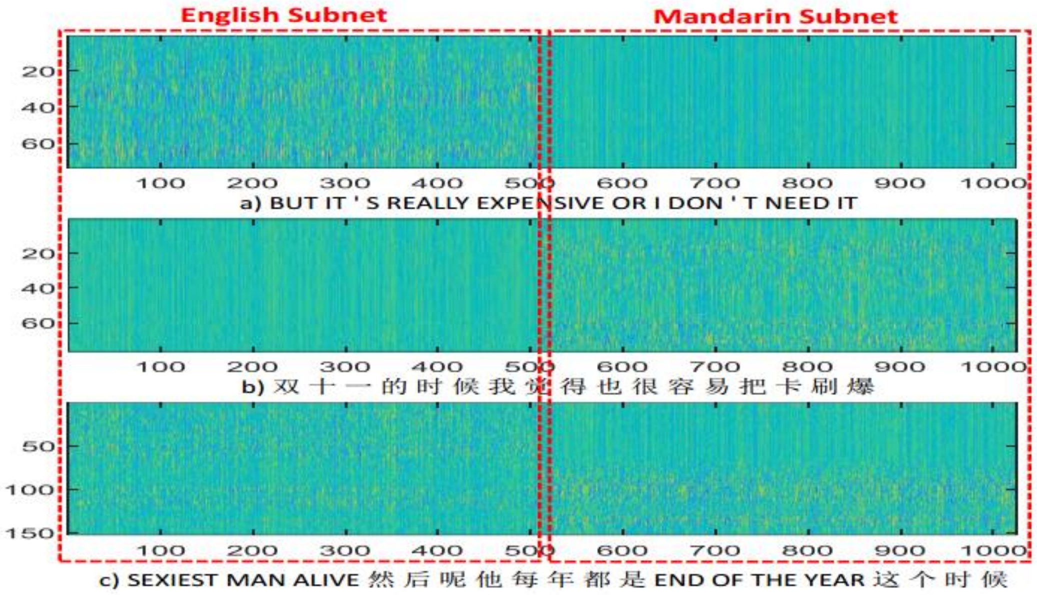
- Idea2、Monolingual pre-trained Bilingual-AM

Language-specific DFSMN-CTC-sMBR system

System	Criterion	Mand.(CER%)	Eng.(WER%)
Mandarin	CTC	8.08	-
	+SMBR	**6.98**	-
English	CTC	-	13.47
	+SMBR	-	**11.31**

Language-universal DFSMN-CTC-sMBR system

Lang. Info.	Criterion	Mand.(CER%)	Eng.(WER%)
w	CTC	8.04	12.80
	+SMBR	**6.94**	**11.33**
w/o	CTC	8.14	12.94
	+SMBR	**7.02**	**11.60**



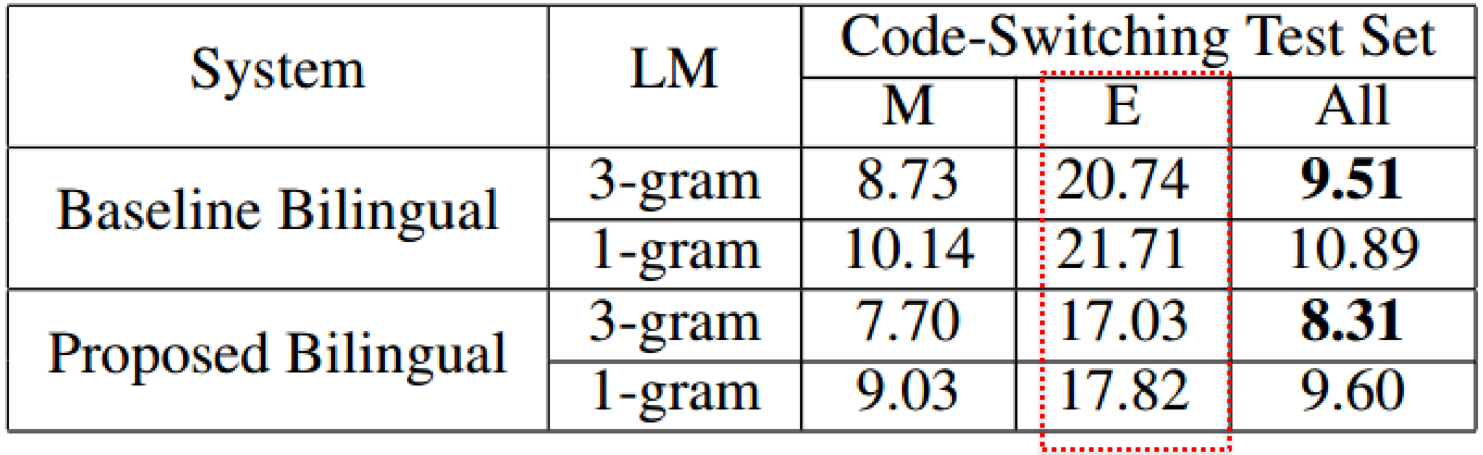
a) BUT IT ' S REALLY EXPENSIVE OR I DON ' T NEED IT

b) 双十一的时候我觉得也很容易把卡刷爆

c) SEXIEST MAN ALIVE 然后呢他每年都是 END OF THE YEAR 这个时候

Language-Universal Mandarin-English ASR

- Idea2、Monolingual pre-trained Bilingual-AM
  - Code-switching test set



System	LM	Code-Switching Test Set		
		M	E	All
Baseline Bilingual	3-gram	8.73	20.74	**9.51**
	1-gram	10.14	21.71	10.89
Proposed Bilingual	3-gram	7.70	17.03	**8.31**
	1-gram	9.03	17.82	9.60

  - Code-switching LM score is mostly back-off to 1-gram

Language-Universal Mandarin-English ASR

- Idea3、 Word Space Mapping

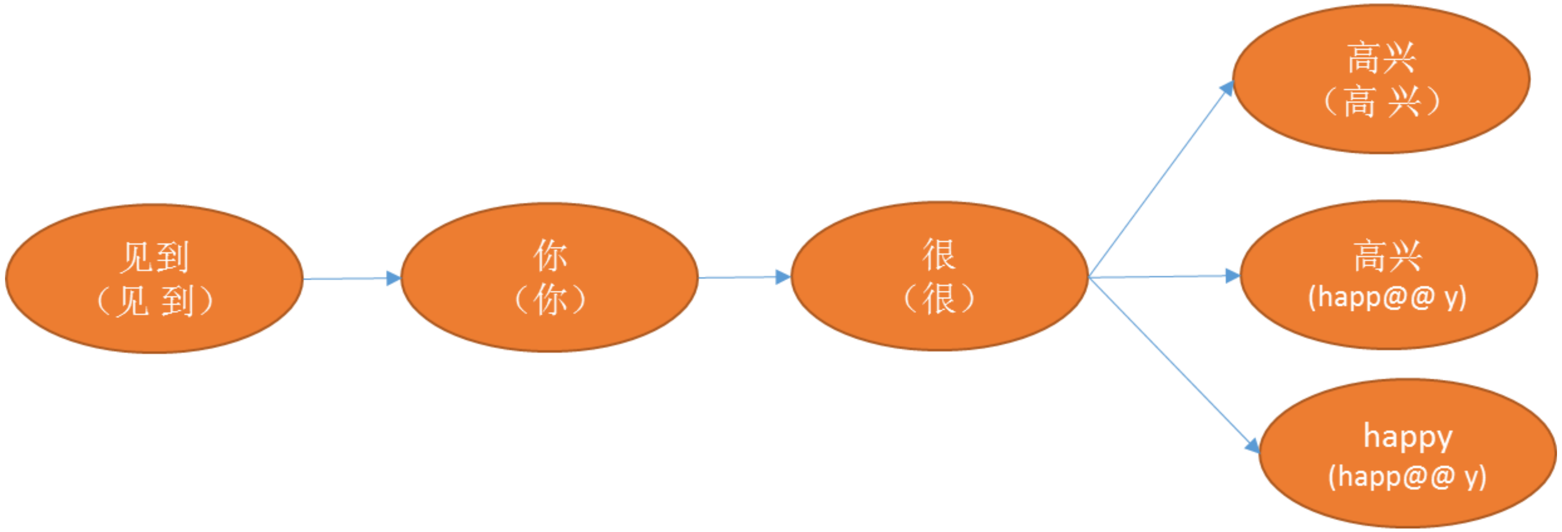
Case： 见到 你 很 happy

LM score： P(happy| 你 很) —>P(happy)

Word Mapping：f(happy) -> 高兴， P(happy| 你 很) —> P(高兴|你 很)

Lexicon：

词条	发音
高兴	高 兴
高兴	happ@@ y
happy	happ@@ y



Language-Universal Mandarin-English ASR

- Idea3、 Word Space Mapping

Method	Test1	Test2
Base	17.12(33.40)	18.54(25.15)
WSP	**16.98[32.73]**	**17.56[23.63]**

Label	处 理 data 的 部分
Base	true data 部 分
LM-Mapping	处 理 data【数据】 部 分
Label	那 这 个 machine learning 我 会 主 要 觉 得 它 是 一 个 三 块 三 块 这 个 knowledge 的 结 合 哦
Base	那 这 个 machine learning 我 会 主 要 觉 得 他 是 一 个 三 块 三 块 这 个 knowledge that 结 合
LM-Mapping	那 这 个 machine learning 我 会 主 要 觉 得 他 是 一 个 三 块 三 块 这 个 knowledge 【知识】 的 结 合

Language-Universal Mandarin-English ASR

- Demo

達摩院
ALIBABA DAMO ACADEMY

thanks