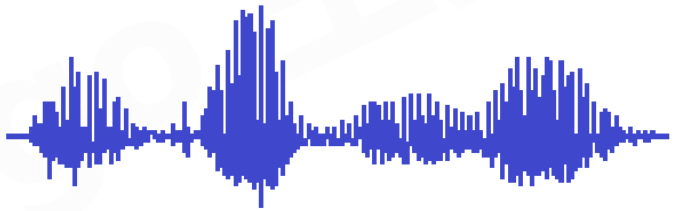
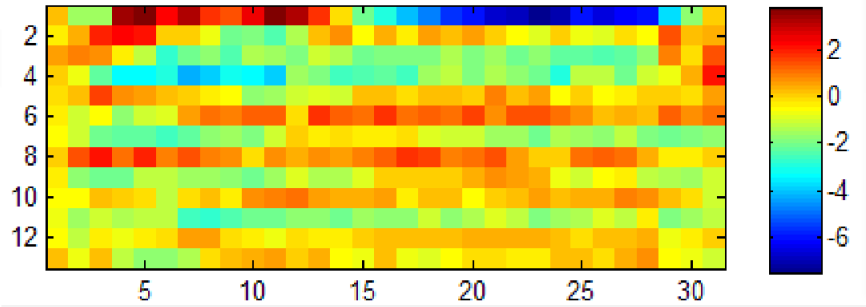
语音识别系统

吴嘉瑶

什么是语音识别?

- 输入





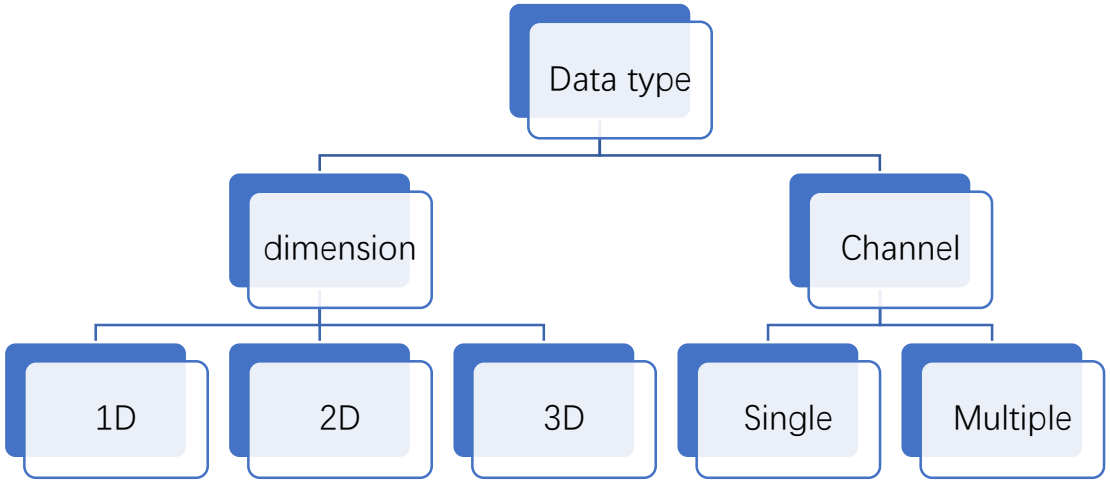
- 输出

普通话

p, u, t, o, ng, h, u, a

语音识别问题分析

- 输入数据类型
  - Audio Signal, Stock Value
  - Word embedding
  - 黑白图片
  - 彩色图片
  - 黑白Video
  - 彩色Video



- 语音识别问题类型：分类
- 生成式模型 $Y = \underset{Y^*}{argmax}\, P(X|Y)P(Y)$
- 鉴别式模型 $Y = \underset{Y^*}{argmax}\, P(Y|X)$

语音中的“空间”维度与频率分布和特征提取的数学变换相联系，它包含多重声学上的变换属性，例如来自环境的因素、说话人、口音、说话方式和速率等。环境因素包括麦克风特性、语音传输信道、环境噪声和室内混响，后几种因素则包括空间和时间维度的相关性。

——《语音识别实践》P35

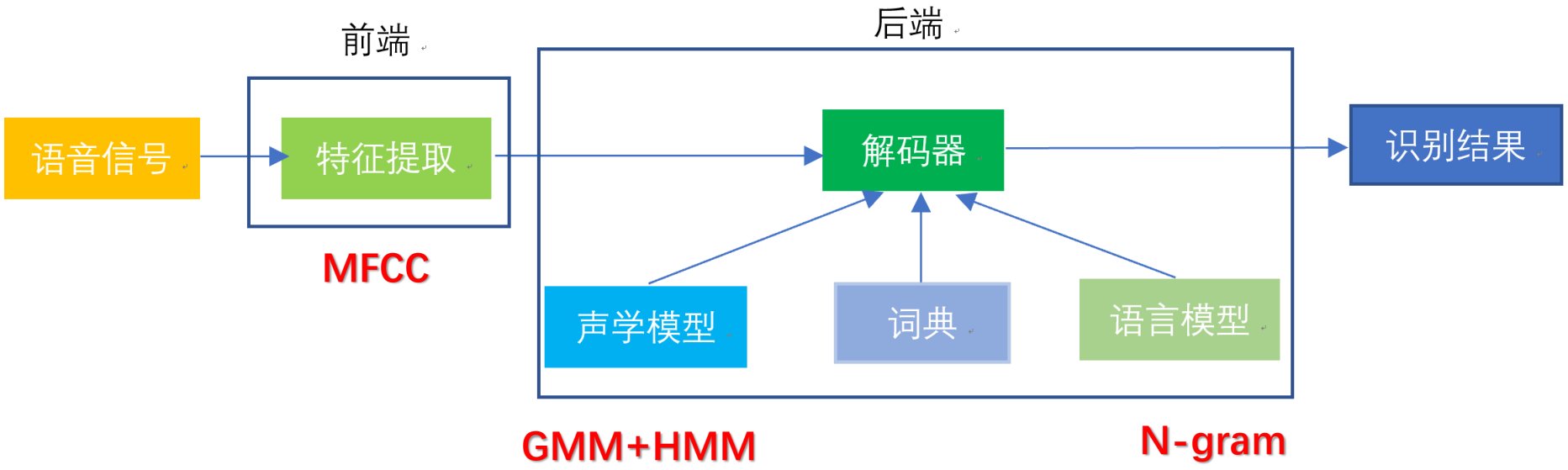
语音识别基本步骤

- 不定长
- 输入输出长度不对等

- 对齐状态
- 识别音素
- 组成单词
- 组成句子

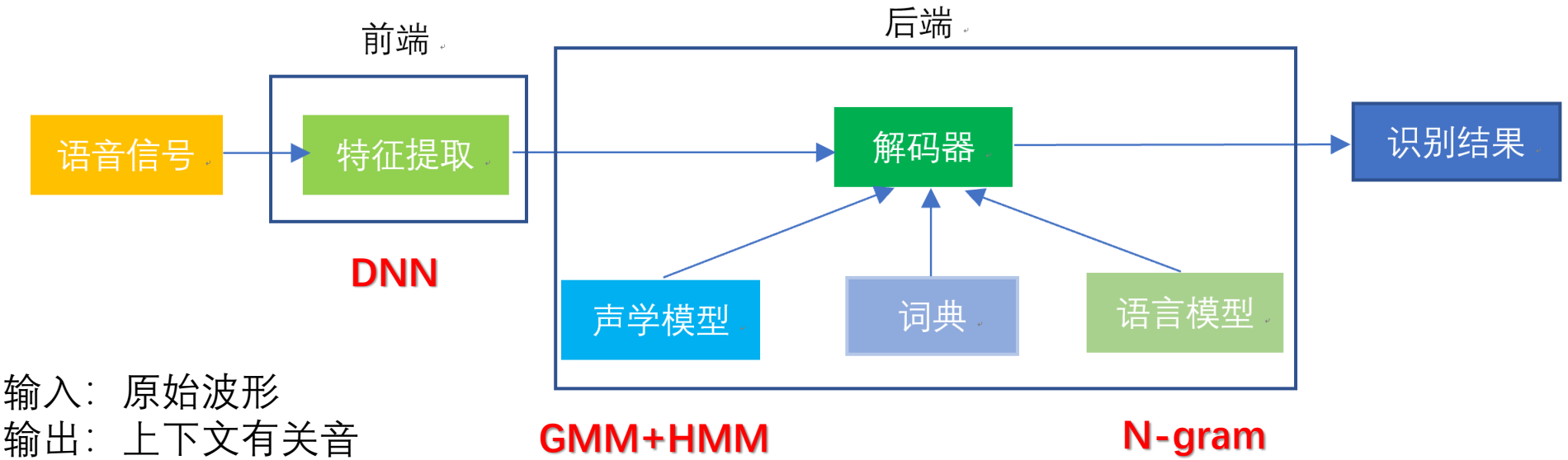
- 时序性
- 上下文相关性

语音识别基本框架的变迁



基本结构

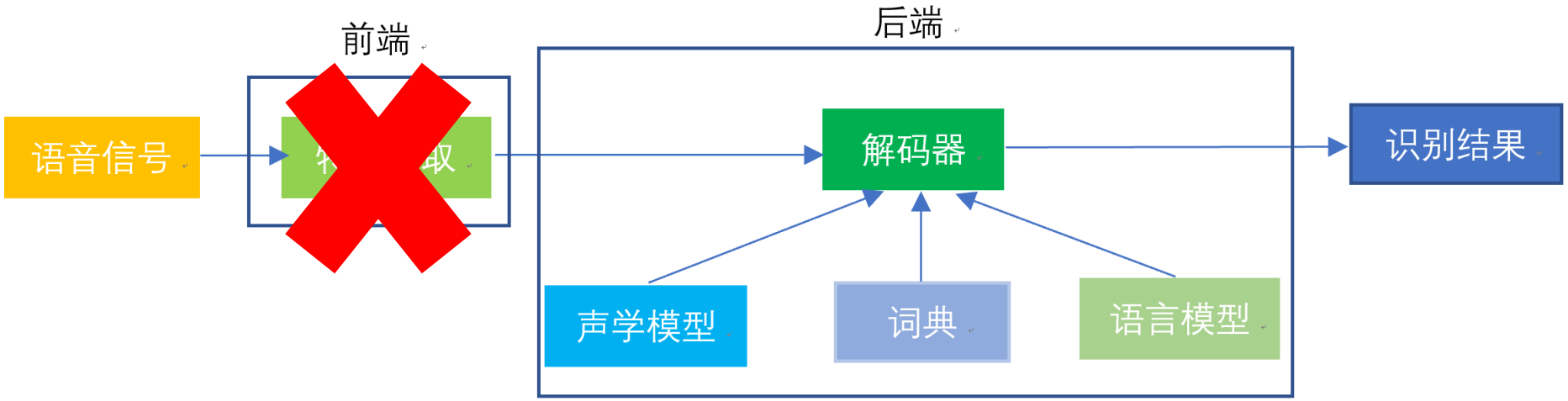
语音识别基本框架的变迁



- 输入：原始波形
- 输出：上下文有关音素分布，判别问题
- 中间层作为特征，bottleneck layer(几十维)

Tandem结构

语音识别基本框架的变迁



- 输入：语音信号
- 输出：判断每帧是哪个状态，DNN代替GMM的功能
- 训练时还是需要传统GMM+HMM提供对齐

DNN+HMM　　**trigram**

Hybrid结构

Why do we need HMM ?

- 神经网络只进行逐帧判别
- 训练时需要HMM系统提供各音素起止时间
- 解码时需要考虑状态转移概率

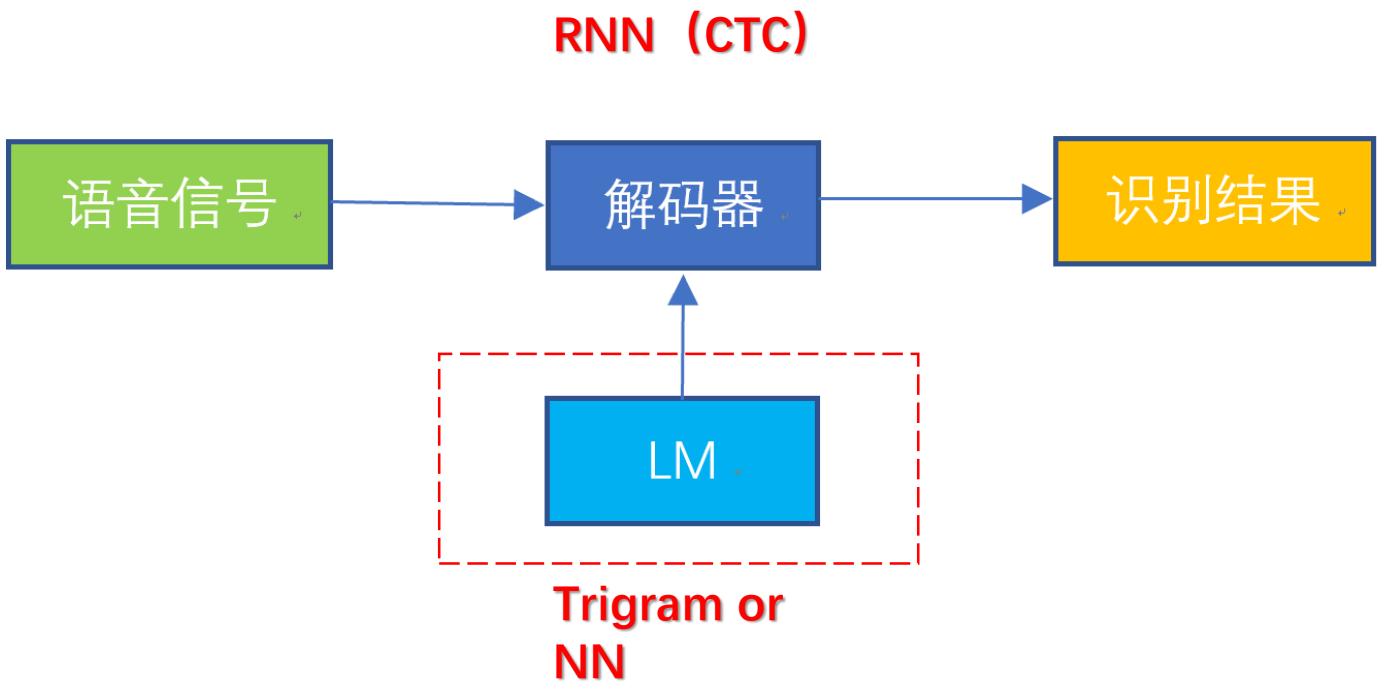
But...

- 上下文建模能力有限

Can we abandon HMM ?

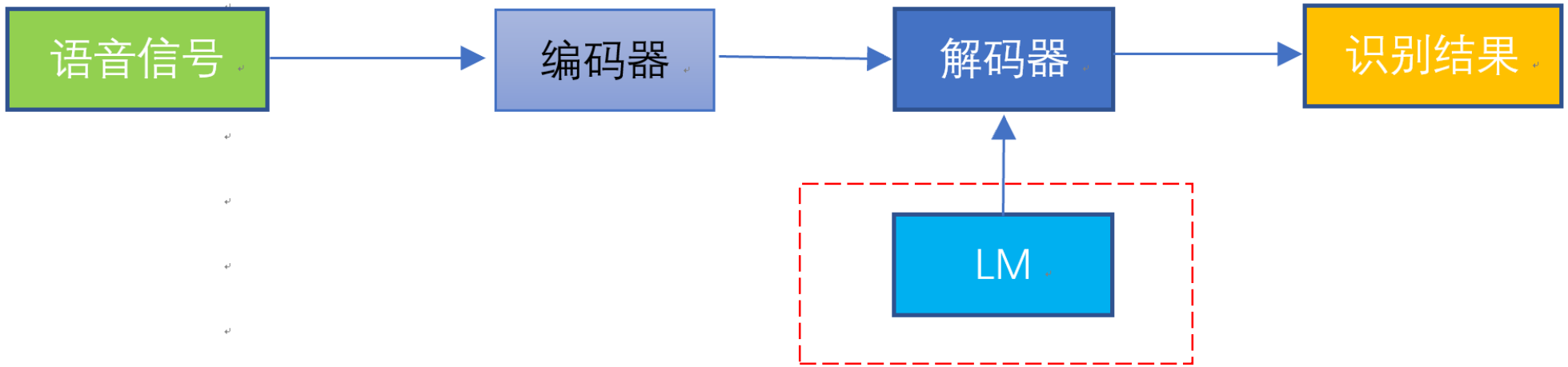
- 如果我们不进行逐帧的判别呢?

语音识别基本框架的变迁



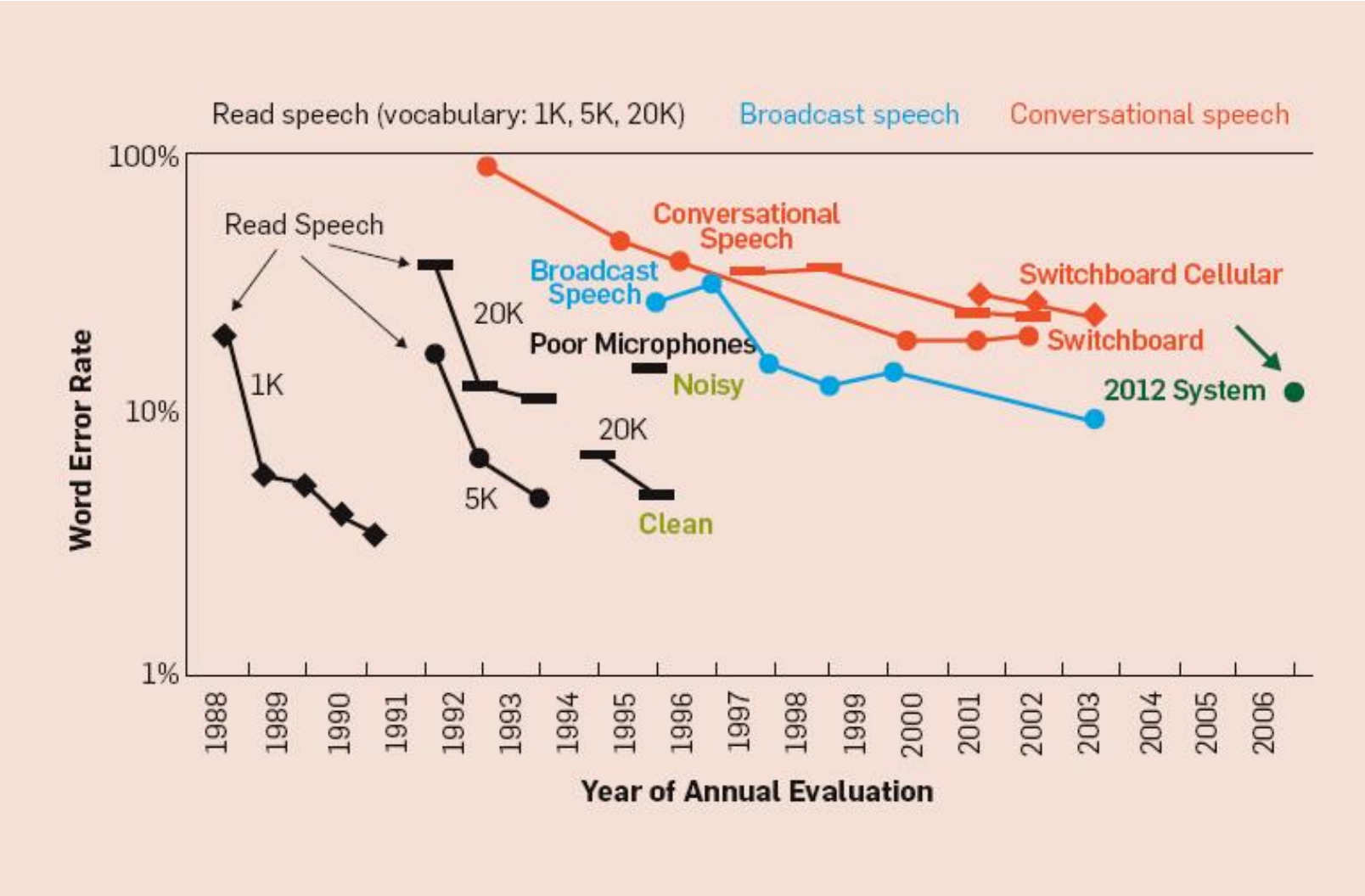
CTC结构

语音识别基本框架的变迁

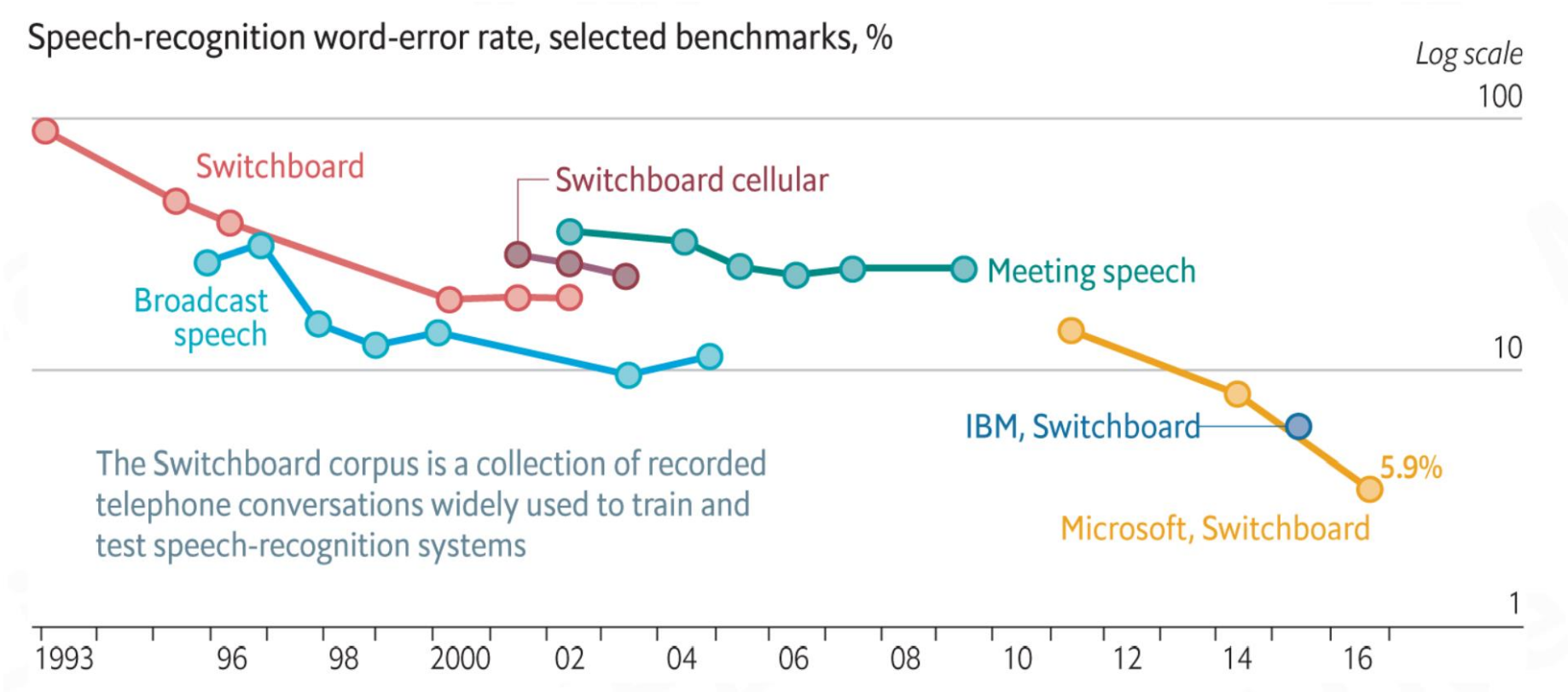


Attention结构

语音识别系统性能


Read speech (vocabulary: 1K, 5K, 20K)
Broadcast speech
Conversational speech
100%
10%
1%
Word Error Rate
Read Speech
1K
20K
5K
20K
Clean
Conversational Speech
Broadcast Speech
Poor Microphones
Noisy
Switchboard Cellular
Switchboard
2012 System
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
Year of Annual Evaluation

语音识别系统性能

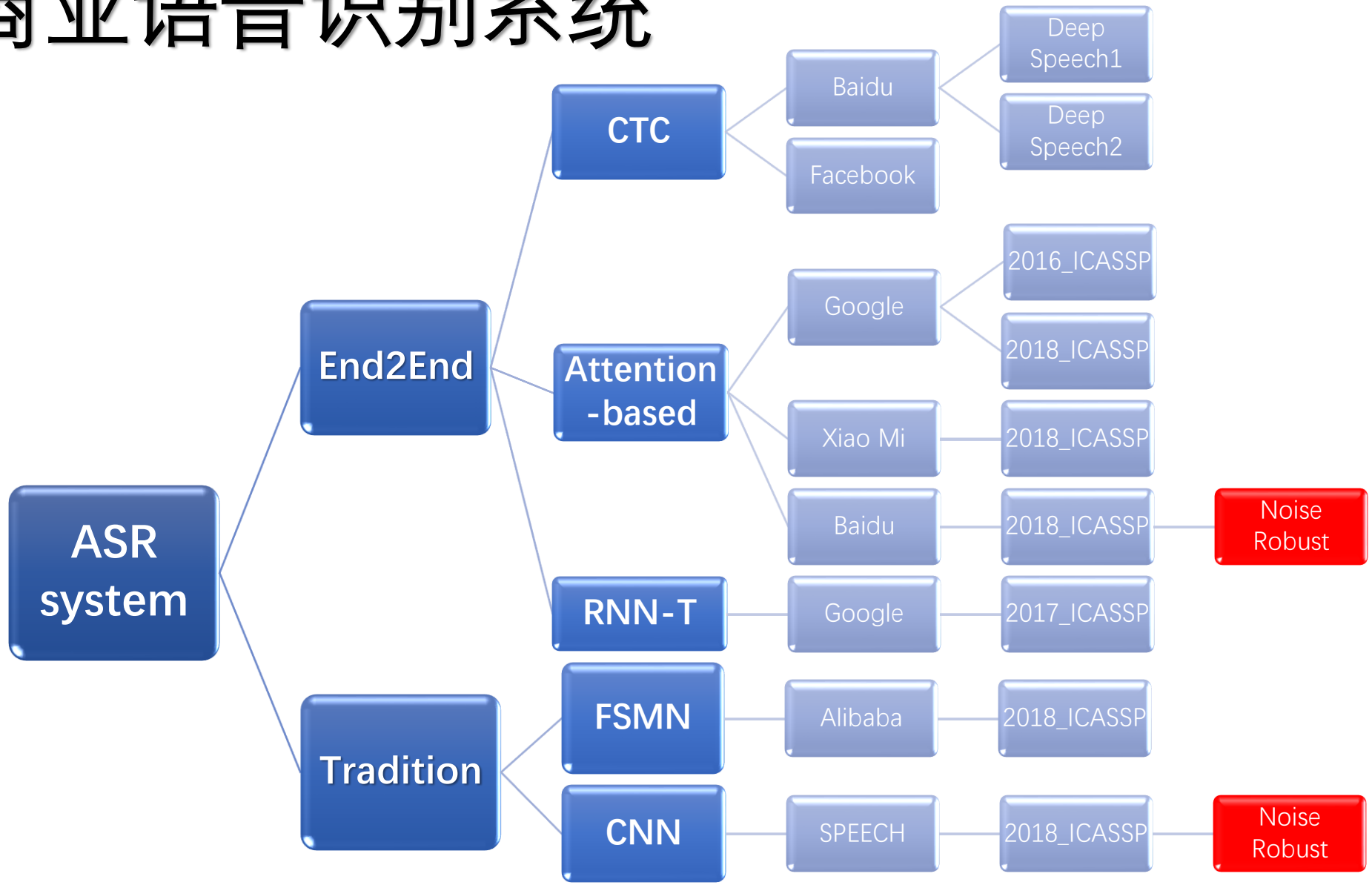

Speech-recognition word-error rate, selected benchmarks, %
Log scale
100
10
1
Switchboard
Switchboard cellular
Broadcast speech
Meeting speech
IBM, Switchboard
5.9%
Microsoft, Switchboard
The Switchboard corpus is a collection of recorded telephone conversations widely used to train and test speech-recognition systems
1993
96
98
2000
02
04
06
08
10
12
14
16

Is ASR system good enough?

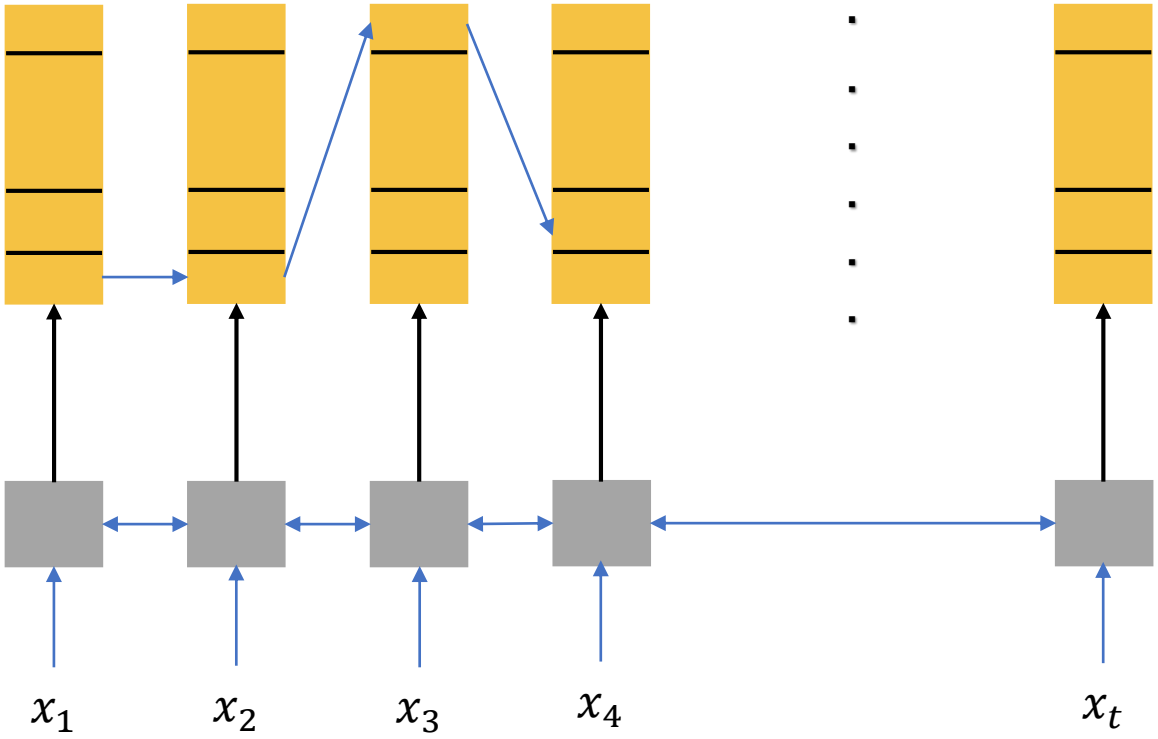
- 远场麦克风语音识别——ALEXA
- 高噪音环境下的语音识别——车载
- 带口音的语音识别——方言
- 不流利的自然语音，变速或者带有情绪的语音识别。

13581887557

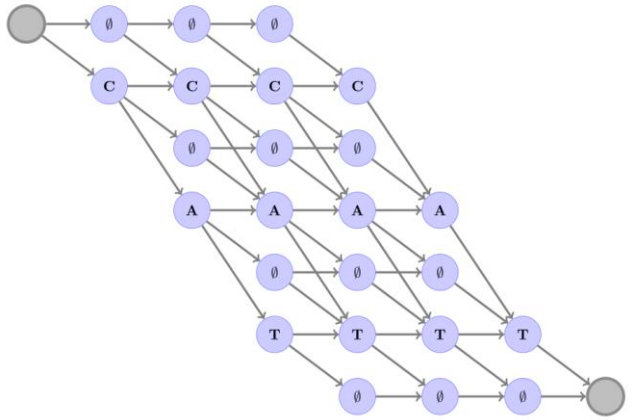
商业语音识别系统


ASR system
End2End
Tradition
CTC
Attention -based
RNN-T
FSMN
CNN
Baidu
Facebook
Google
Xiao Mi
Baidu
Google
Alibaba
SPEECH
Deep Speech1
Deep Speech2
2016_ICASSP
2018_ICASSP
2018_ICASSP
2018_ICASSP
2017_ICASSP
2018_ICASSP
2018_ICASSP
Noise Robust
Noise Robust

CTC Loss Function



K category



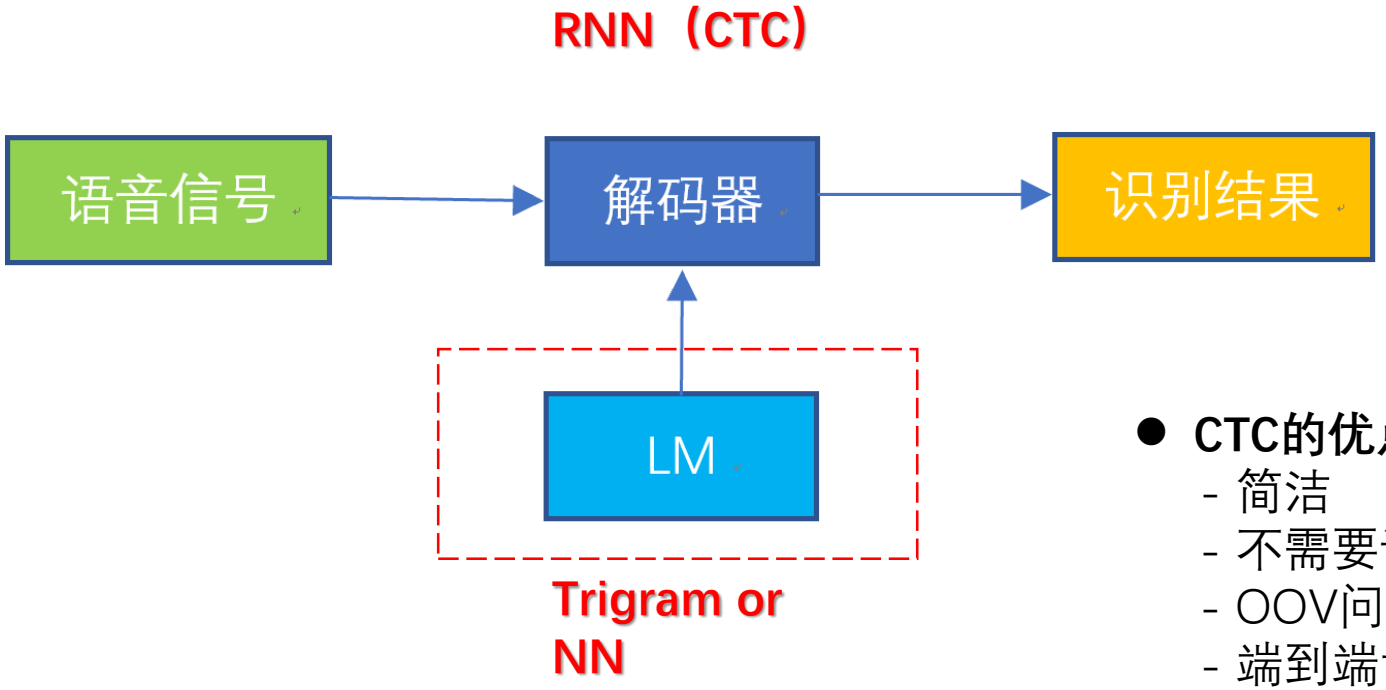
- Softmax over vocabulary
- Score(k, t)=log P(k, t| X)

- Total Score of one path
  the sum of scores at different time steps
- The probability of any transcript
  the sum of probabilities of all paths

- 不再进行逐帧判别
- 添加blank，让输出能缩成标答即可
  （实际输出位置接近真实位置）

普通声学模型：n n i i i i h h h ao ao ao ao ao
CTC　　：- n - - i - - h - - - ao - -

CTC Loss Function

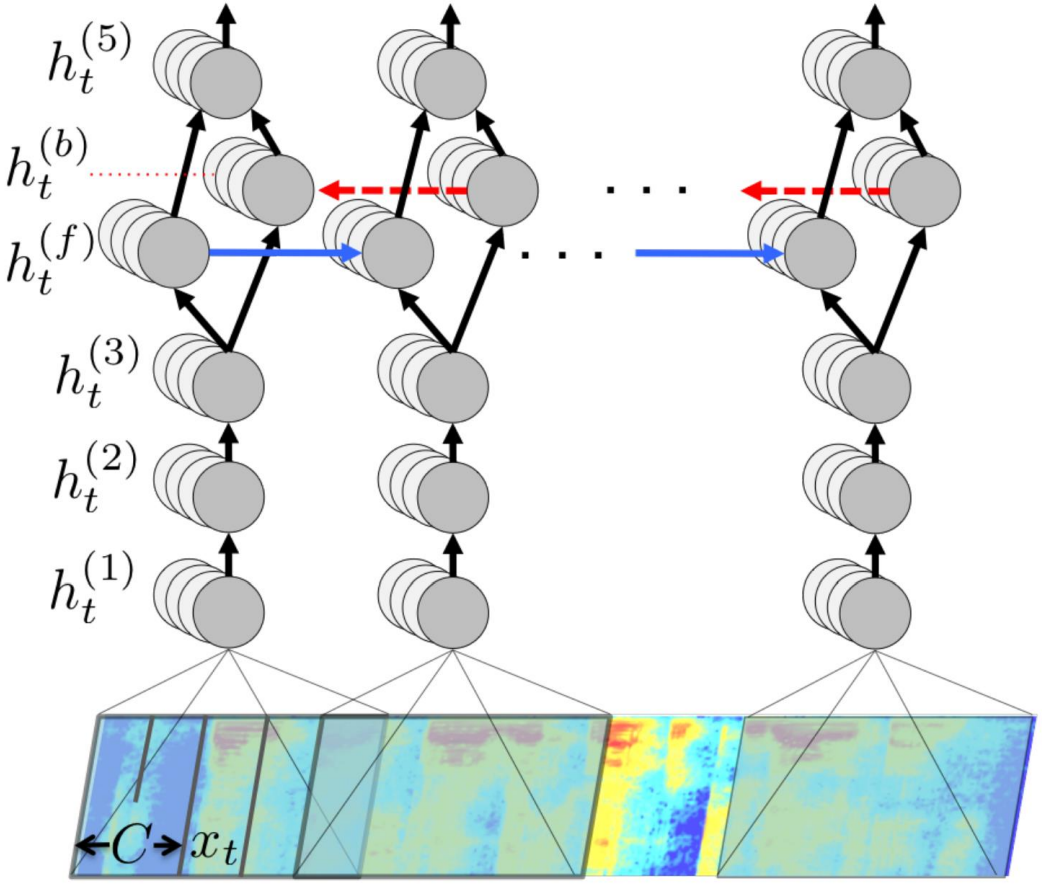


- **CTC的特点**
  - 帧独立假设
  - 假设上下文已由RNN处理
- **训练**
  - 所有能缩成标准答案的总概率
  - 动态规划算法(前向后向算法)
- **解码**
  -beam search

- **CTC的优点**
  - 简洁
  - 不需要语言知识（词典和语言模型）
  - OOV问题（词典既是铠甲，又是软肋）
  - 端到端训练（模块单独训练整个系统不一定最好）
- **CTC的缺点**
  - 大量的训练数据（身兼数职）
  - 语音数据里上下文有关信息少（外接LM）
  - 帧独立假设。/greit/可以拼成great或者grate，但CTC可能会拼成grete.

Graves, Alex, et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.

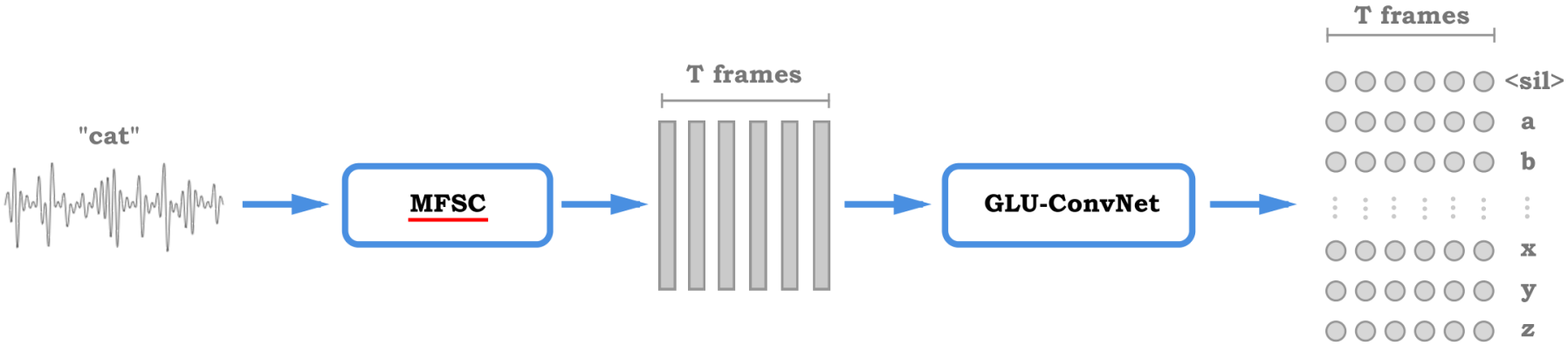
Deep Speech 1.0



- 输入：$X = (x_1, x_2, \dots, x_t)$
- 输出：$P(c_t|X)$
  $c_t \in \{a, b, c, \dots, z, space, apostrophe, blank\}$
- Loss function：CTC
- 训练：Nesterov's Accelerated gradient（NAG）二阶方法

Hannun, Awni, et al. "Deep speech: Scaling up end-to-end speech recognition." *arXiv preprint arXiv:1412.5567* (2014).

Facebook——ConvNet CTC

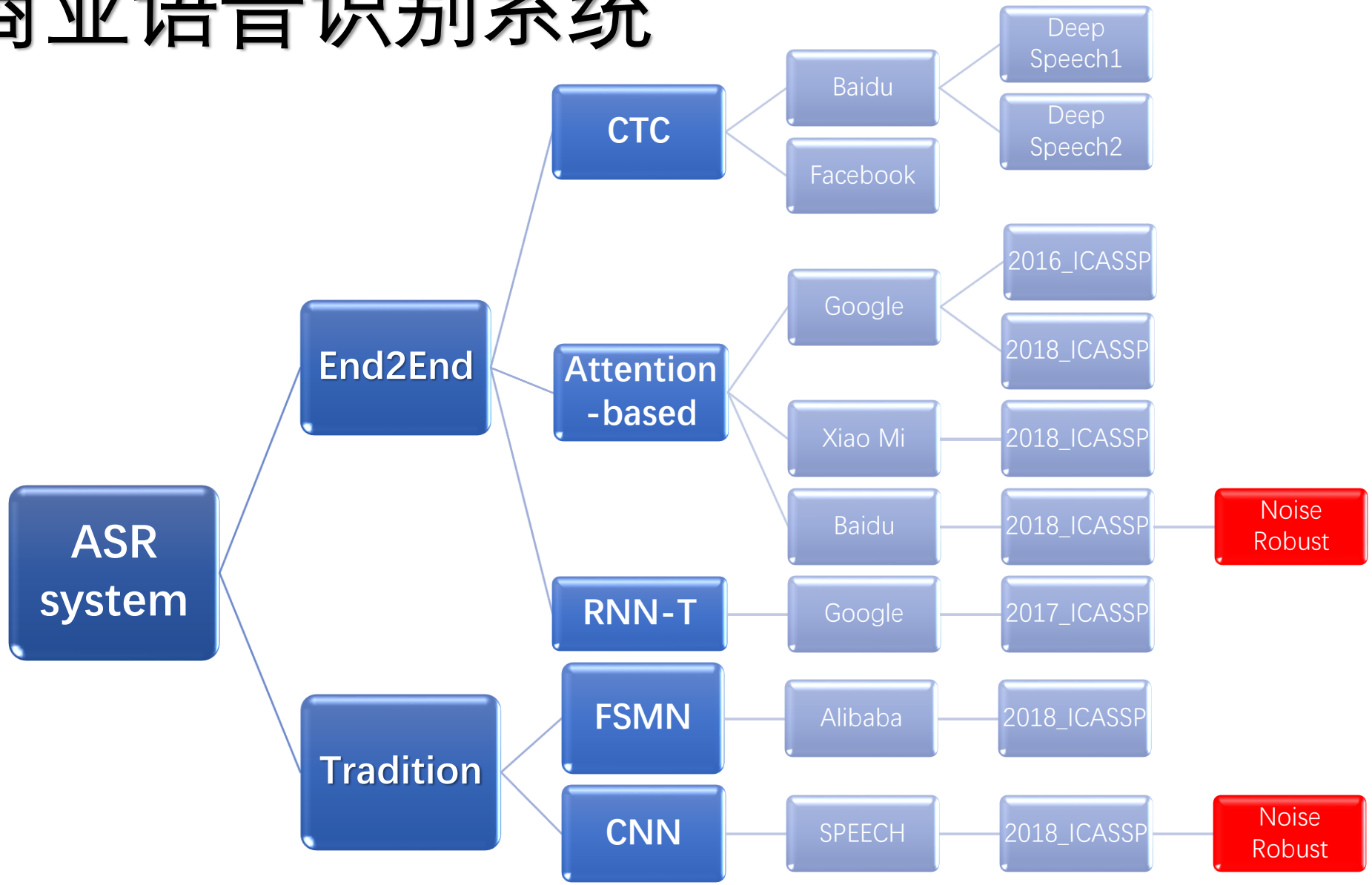


- 输入：$X = (x_1, x_2, \ldots, x_t)$
- 输出：$P(c_t|X)$ $c_t \in \{30\ graphemes\}$
- 网络结构：GLU-ConvNet $h_i(X) = (X * W_i + b_i) \otimes \sigma(X * V_i + c_i)$
  缓解梯度消失的问题
- Loss function：CTC

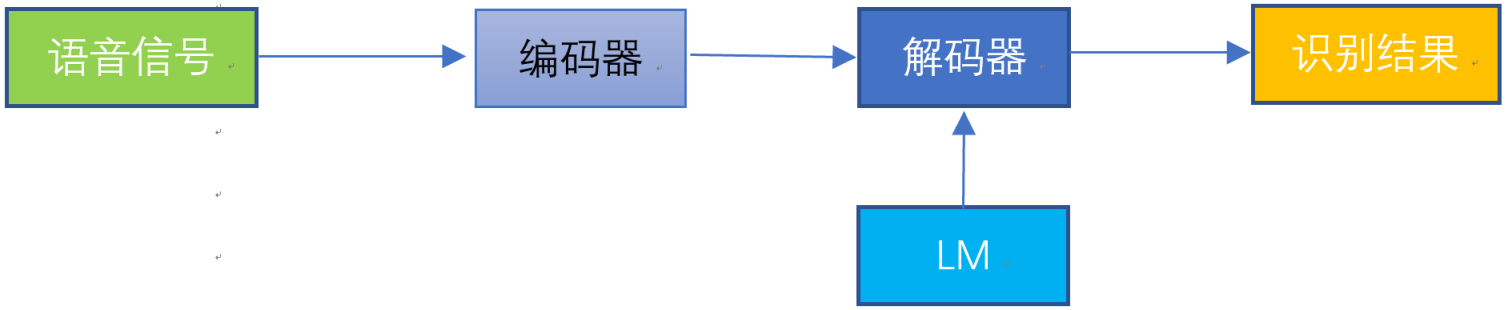
Collobert, Ronan, Christian Puhrsch, and Gabriel Synnaeve. "Wav2letter: an end-to-end convnet-based speech recognition system." *arXiv preprint arXiv:1609.03193* (2016).

Liptchinsky, V., G. Synnaeve, and R. Collobert. "Letterbased speech recognition with gated convnets." *CoRR, vol. abs/1712.09444* 1 (2017).

商业语音识别系统

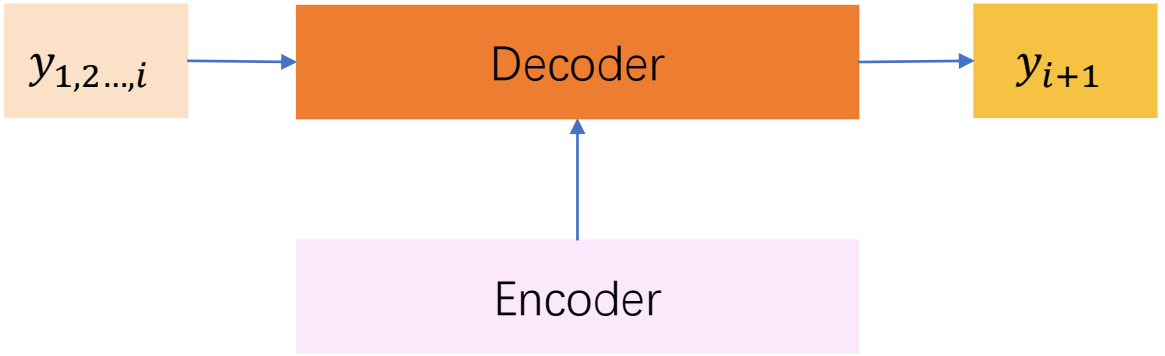

ASR system
End2End
Tradition
CTC
Attention -based
RNN-T
FSMN
CNN
Baidu
Facebook
Google
Xiao Mi
Baidu
Google
Alibaba
SPEECH
Deep Speech1
Deep Speech2
2016_ICASSP
2018_ICASSP
2018_ICASSP
2018_ICASSP
2017_ICASSP
2018_ICASSP
2018_ICASSP
Noise Robust
Noise Robust

Seq2Seq with attention



- 编码器
  -序列的特征提取和信息压缩（声学模型）
- 解码器
  -每步主动寻找输入需要那几帧（语言模型）

Seq2Seq with attention



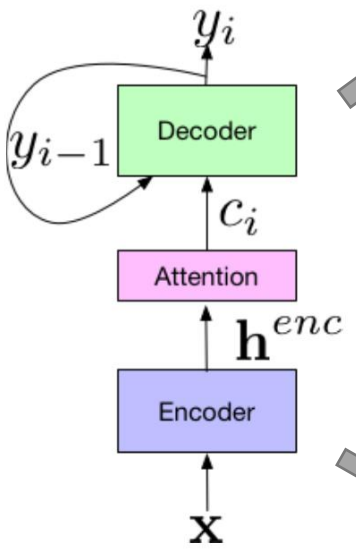
$$P(y_{i+1}|y_{1,2,\ldots,i}, X)$$

- **Seq2Seq的特点**
  - 编码器+解码器
  - 每步都输出，没有Blank
- **训练**
- **解码**
  - beam search

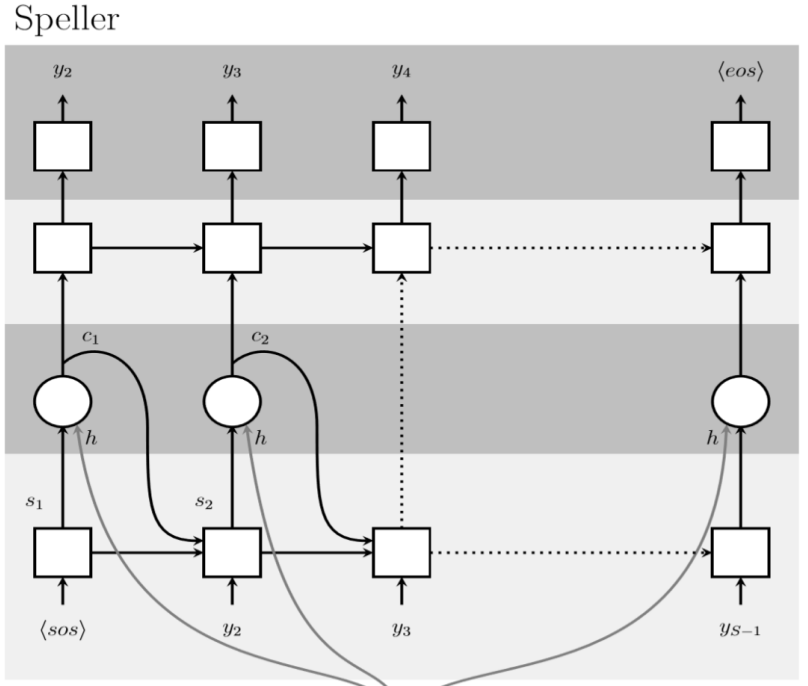
- **Seq2Seq的优点**
  - 帧非独立假设
  - 端到端训练
- **Seq2Seq的缺点**
  - 不能在线识别

Google LAS

$Y = (\langle sos \rangle, y_1, \ldots, y_s, \langle eos \rangle)$



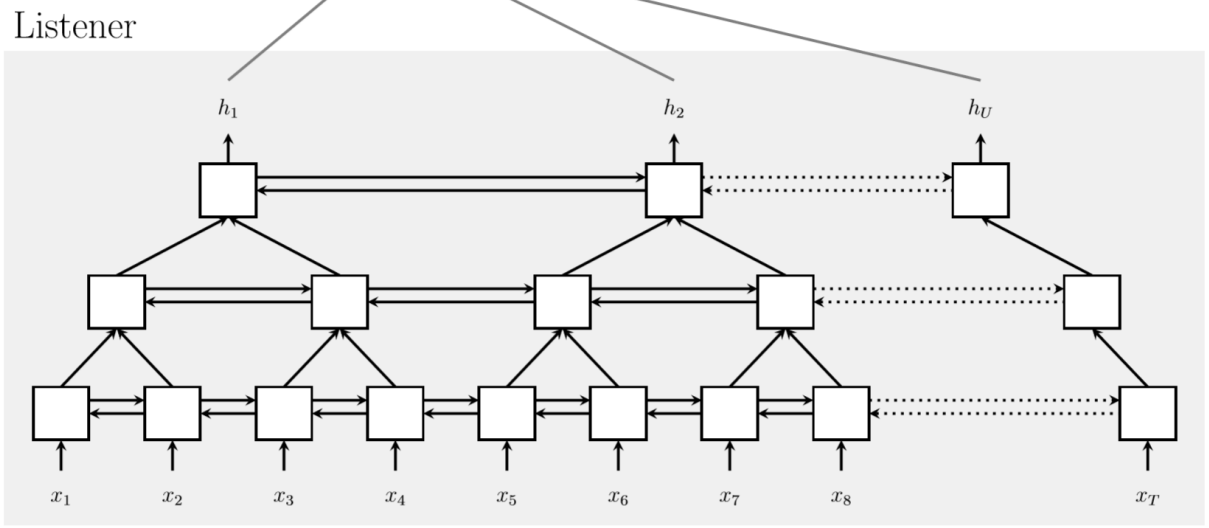
$X = (x_1, x_2 \ldots x_T)$



Grapheme characters y_i are modelled by the CharacterDistribution

AttentionContext creates context vector c_i from $\mathbf{h}$ and s_i

Long input sequence $\mathbf{x}$ is encoded with the pyramidal BLSTM Listen into shorter sequence $\mathbf{h}$



Google LAS

$P(y_i|X, y_{<i})$

RNN

$c = (c_1, c_2 \dots)$

Attention

$h = (h_1, h_2 \dots h_u)$

Pyramidal RNN

$X = (x_1, x_2 \dots x_T)$

- 类似于对齐操作
- 信息压缩
- 特征提取

- 公式:

$$c_i = Attention\ Context(s_i, h)$$
$$s_i = RNN(s_{i-1}, y_{i-1}, c_{i-1})$$
$$P(y_i|x, y_{<i}) = Character\ Distribution(s_i, c_i)$$

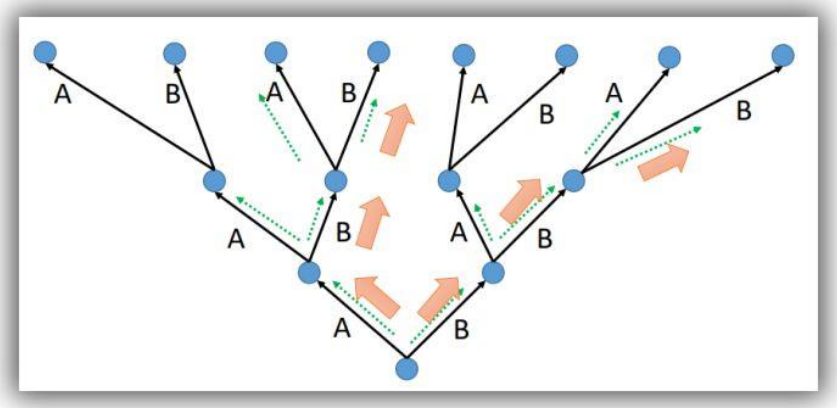
- 训练

$$\max_{\theta} \sum_{i} \log P(y_i|X, y_{i-1,\dots,1,\theta})$$

- 解码

Beam search

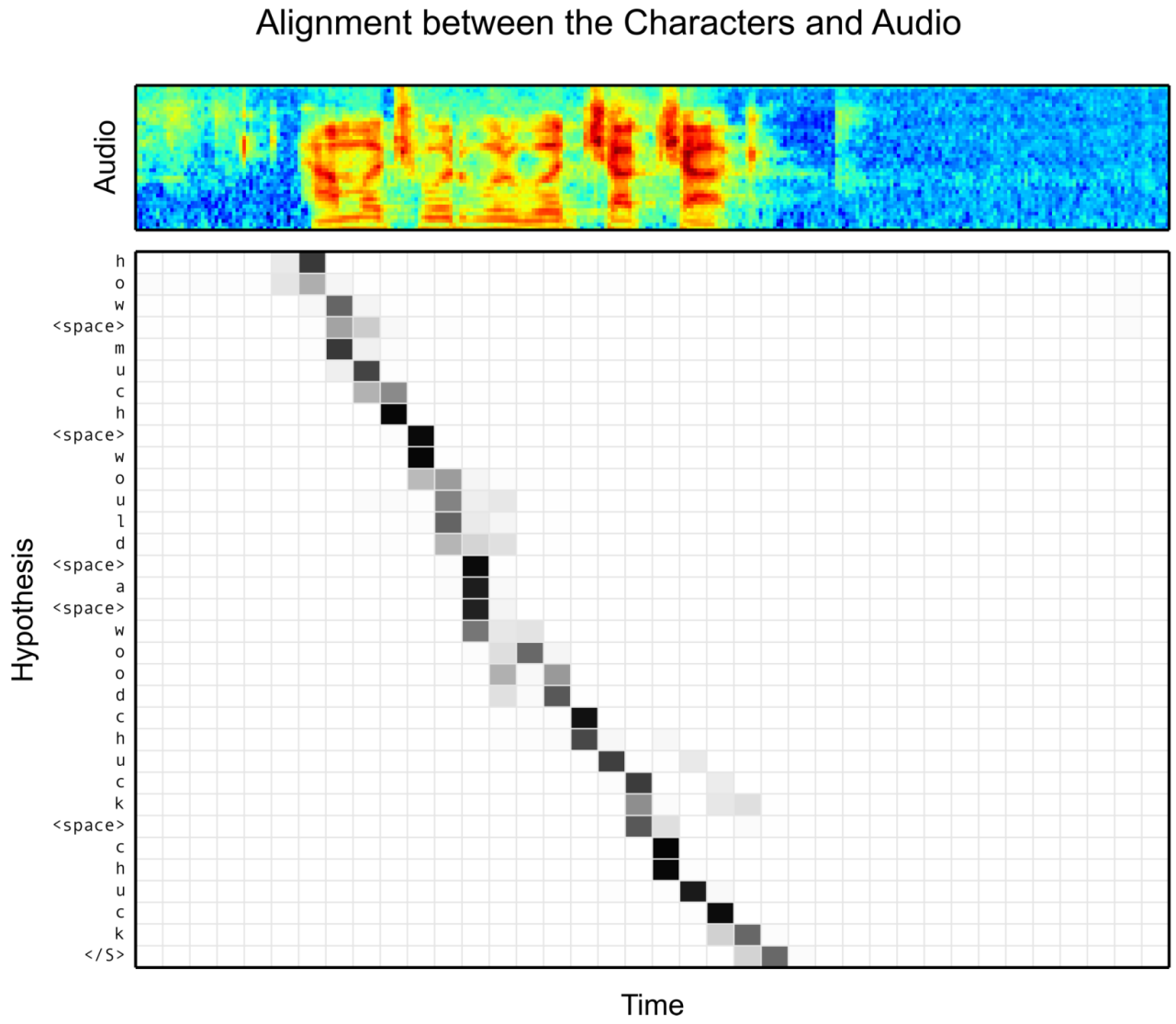
Beam size



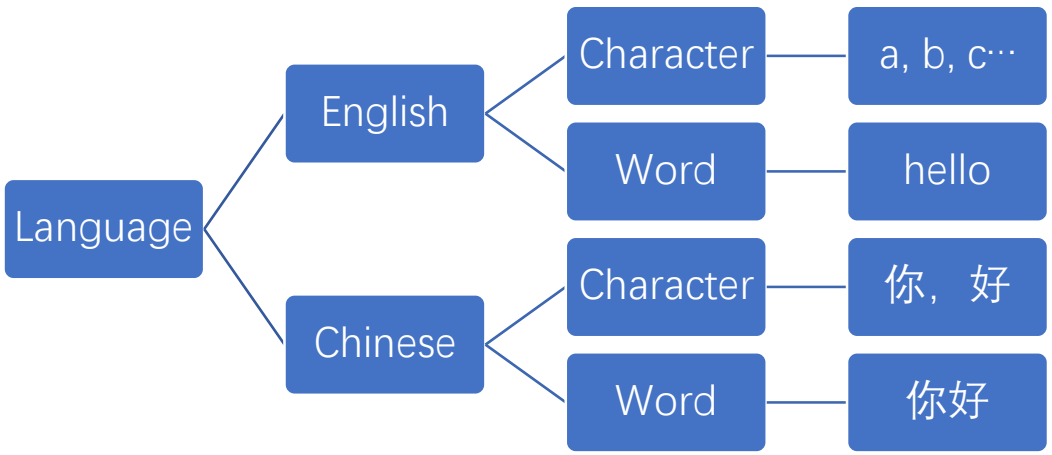
1. Chan, William, et al. "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition." *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016.

2. Chiu, Chung-Cheng, et al. "State-of-the-art speech recognition with sequence-to-sequence models." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

Google LAS


Alignment between the Characters and Audio
Audio
Hypothesis
h
o
w
<space>
m
u
c
h
<space>
w
o
u
l
d
<space>
a
<space>
w
o
o
d
c
h
u
c
k
<space>
c
h
u
c
k
</S>
Time

Xiao Mi

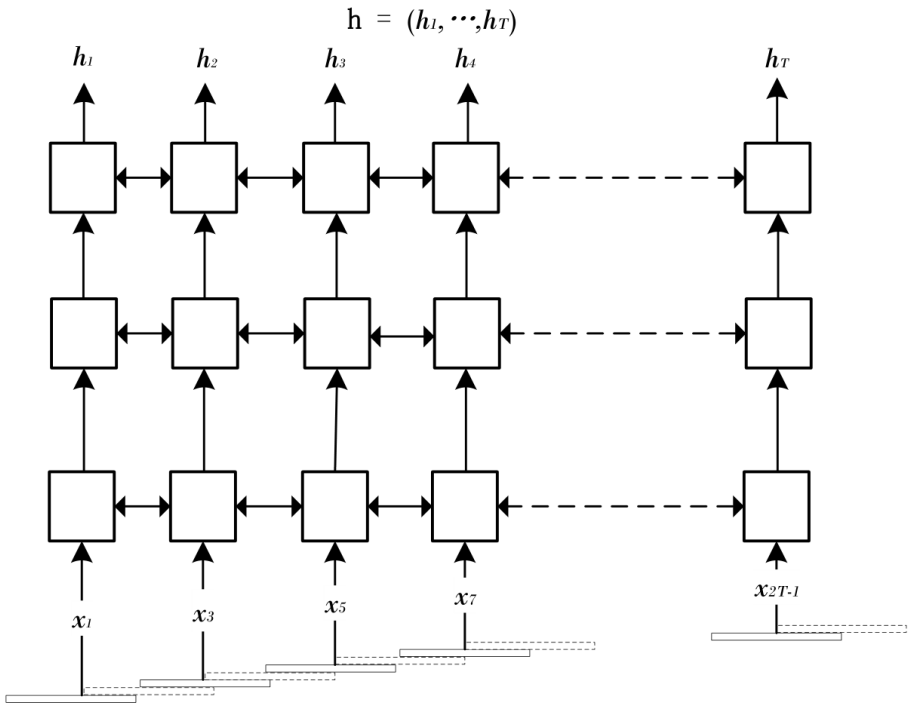


From English to Mandarin on Voice Search Task

1. Structure
   - character Embedding
2. Training
   - L2 regularization
   - Gaussian weight noise
   - Frame skipping
   - Attention smoothing

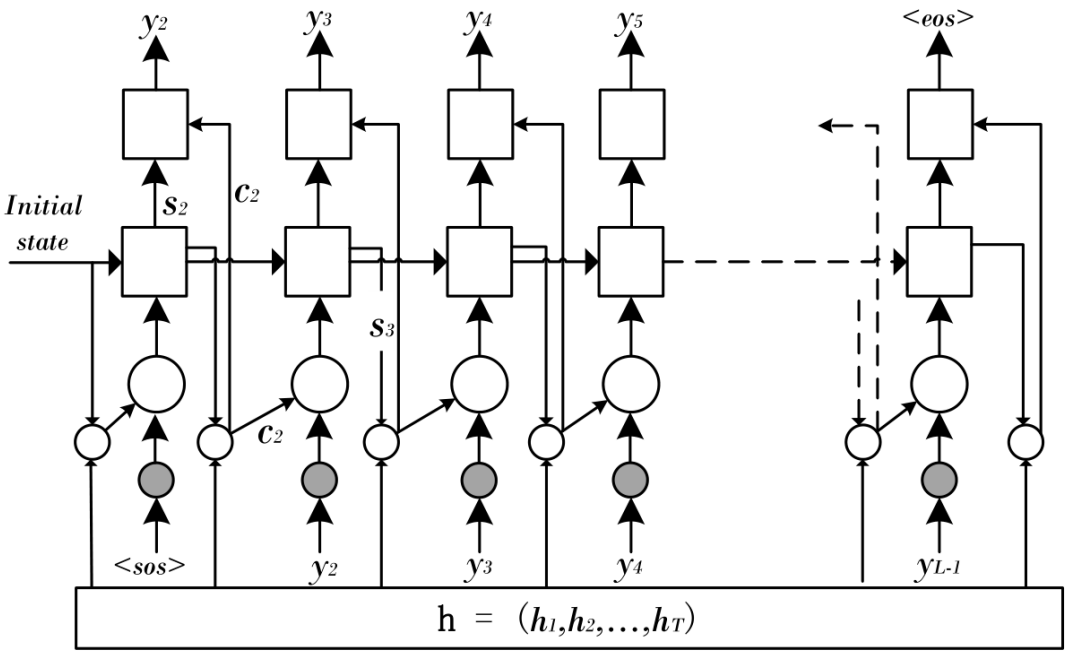
Shan, Changhao, et al. "Attention-based end-to-end speech recognition on voice search." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

Xiao Mi



encoder

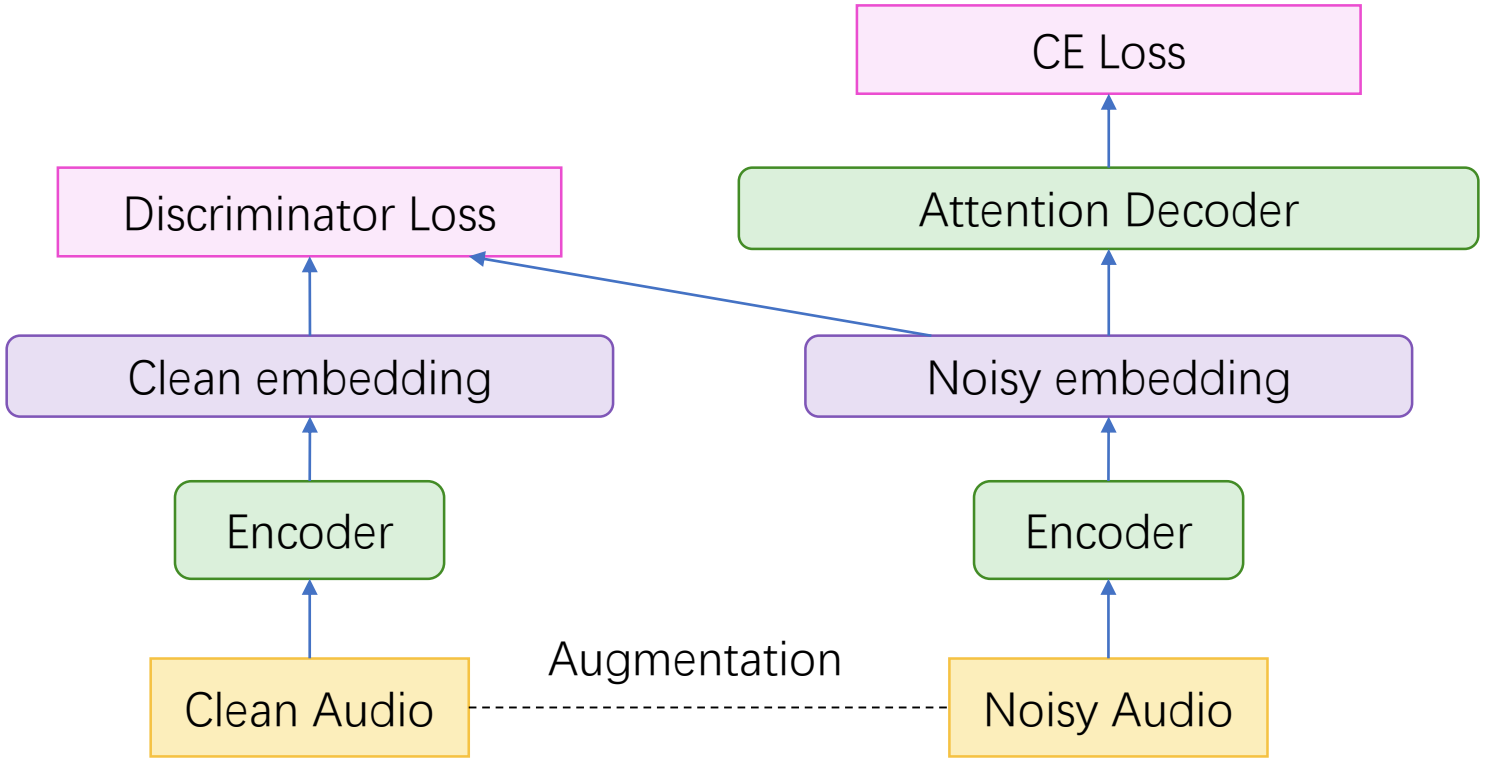
$$(h_1, h_2, \dots, h^T) = BLSTM(x_1, x_3, \dots, x^{2T-1})$$



decoder

$$P(y_i|X, y_{i-1}, \dots, y_1) = CharacterDist(s_i, c_i)$$
$$s_i = DecodeRNN([y_{i-1}, c_{i-1}], s_{i-1})$$
$$c_i = AttentionContext(s_i, H)$$

Baidu ASR GAN

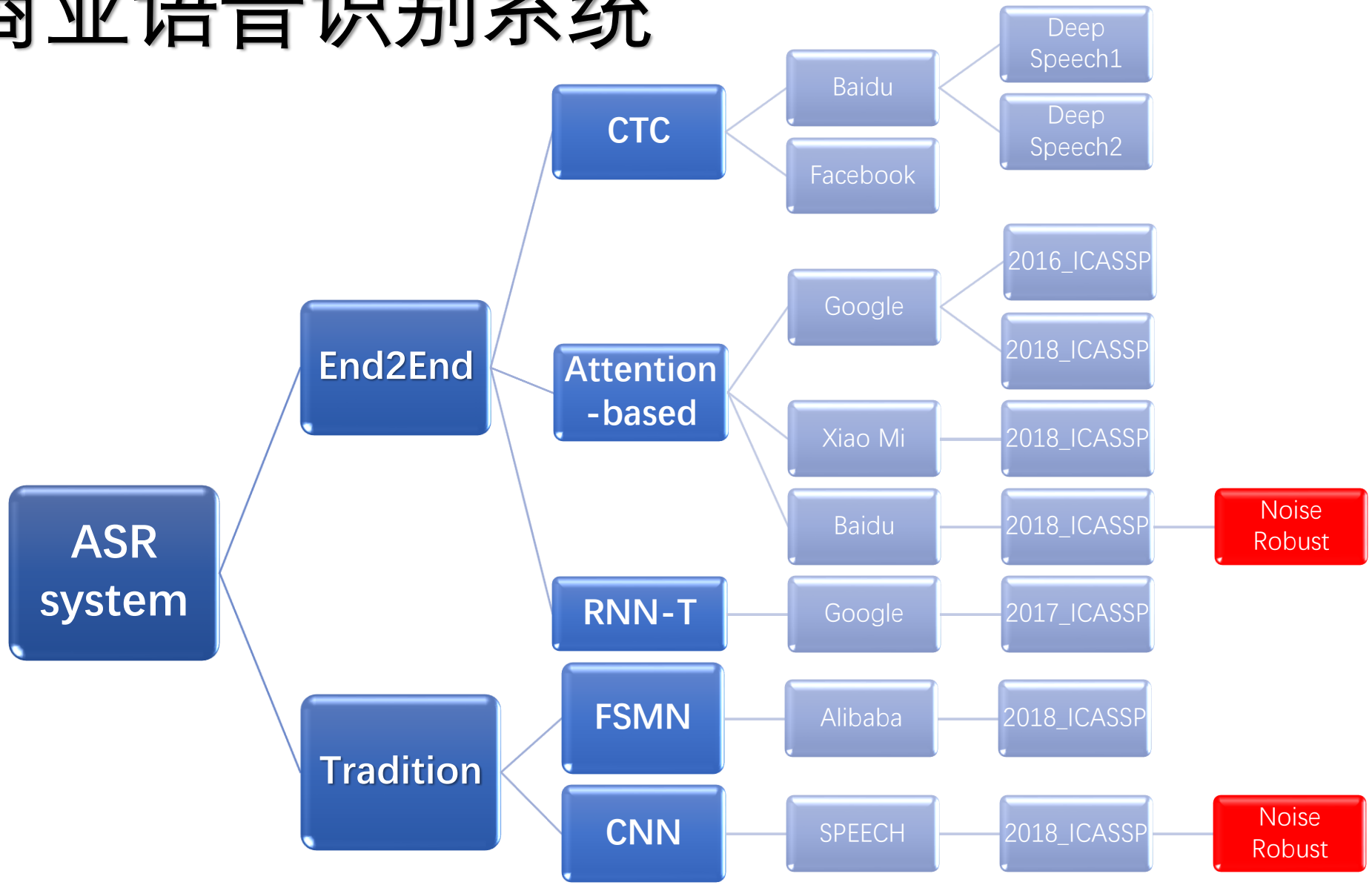


Sriram, Anuroop, et al. "Robust speech recognition using generative adversarial networks." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

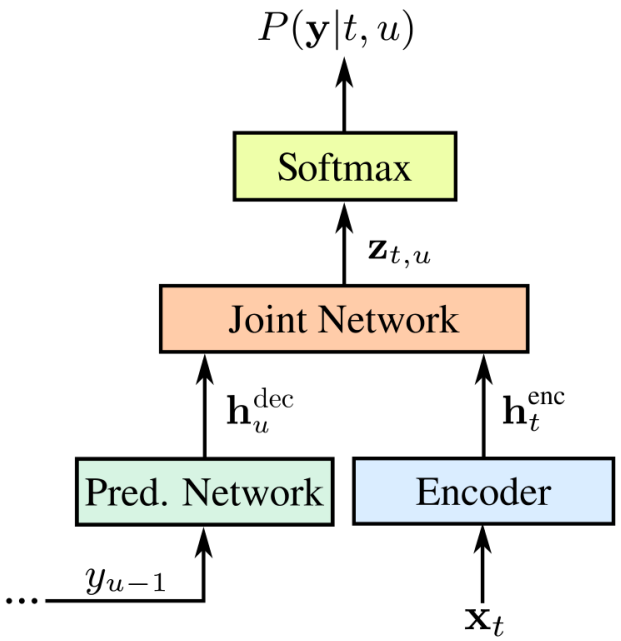
Think

- 编码结果
- Match function（两个向量的匹配度）
- Encoder和Decoder的网络

商业语音识别系统


ASR system
End2End
Tradition
CTC
Attention -based
RNN-T
FSMN
CNN
Baidu
Facebook
Google
Xiao Mi
Baidu
Google
Alibaba
SPEECH
Deep Speech1
Deep Speech2
2016_ICASSP
2018_ICASSP
2018_ICASSP
2018_ICASSP
2017_ICASSP
2018_ICASSP
2018_ICASSP
Noise Robust
Noise Robust

RNN-T



- Encoder（RNN）把x_t映射为特征向量h_t^{enc}.类似于AM特征提取
- Pre. Network将最后一个非空白标签作为输入

训练：动态规划算法
解码：beam search

Rao, Kanishka, Haşim Sak, and Rohit Prabhavalkar. "Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer." *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*. IEEE, 2017.

End2End比较

	CTC	Transducer	Attention
输出语言模型	无	有	有
对齐	单调	单调	不单调
	硬	硬	软
解码所需步数	输入长度	输入长度 +输出长度	输出长度

CTC做输出独立假设，而Transducer和Attention不独立

Graves, Alex, et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.

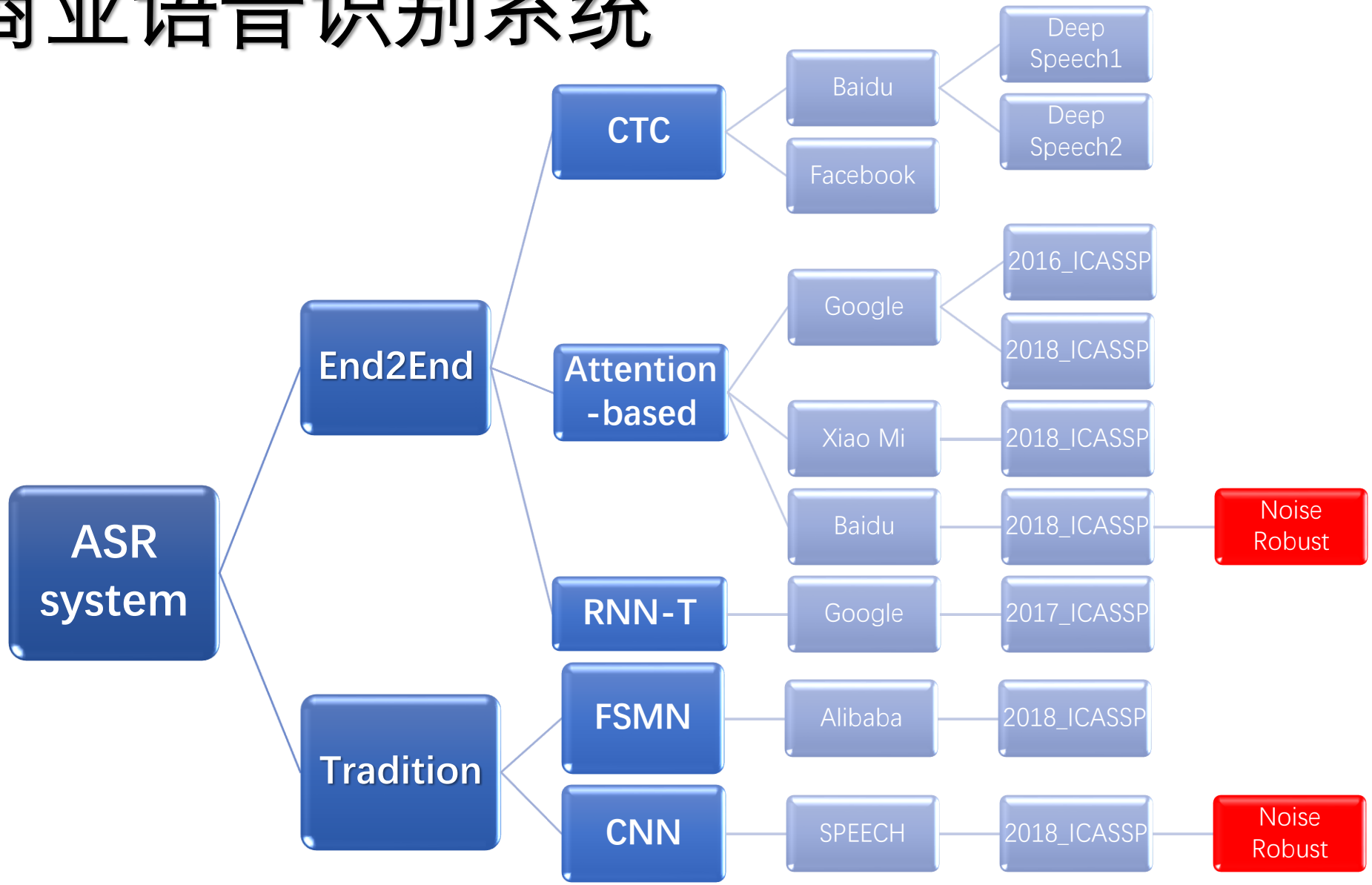
Graves, Alex. "Sequence transduction with recurrent neural networks." *arXiv preprint arXiv:1211.3711* (2012).

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

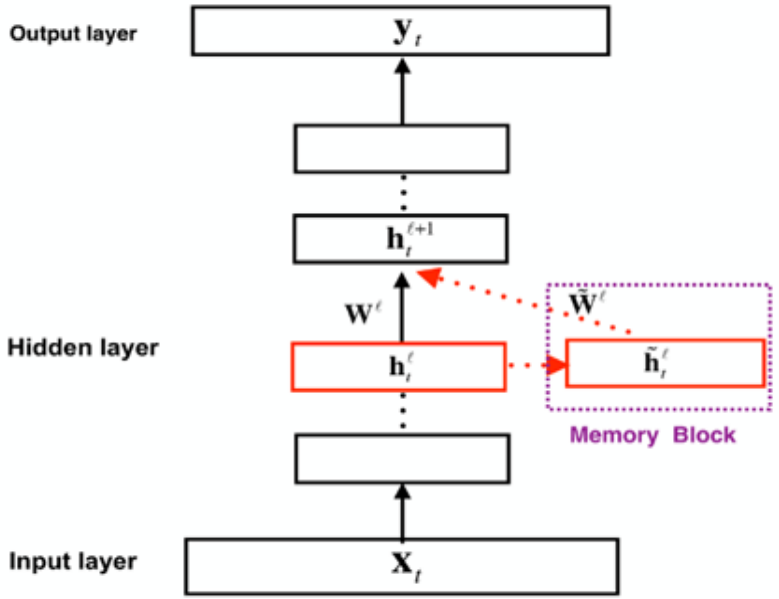
Better with LM！

- 端到端学习中包含语言模型但是较弱
- 文本数据比语音数据更好获得

商业语音识别系统


ASR system
End2End
Tradition
CTC
Attention -based
RNN-T
FSMN
CNN
Baidu
Facebook
Google
Xiao Mi
Baidu
Google
Alibaba
SPEECH
Deep Speech1
Deep Speech2
2016_ICASSP
2018_ICASSP
2018_ICASSP
2018_ICASSP
2017_ICASSP
2018_ICASSP
2018_ICASSP
Noise Robust
Noise Robust

Alibaba-FSMN



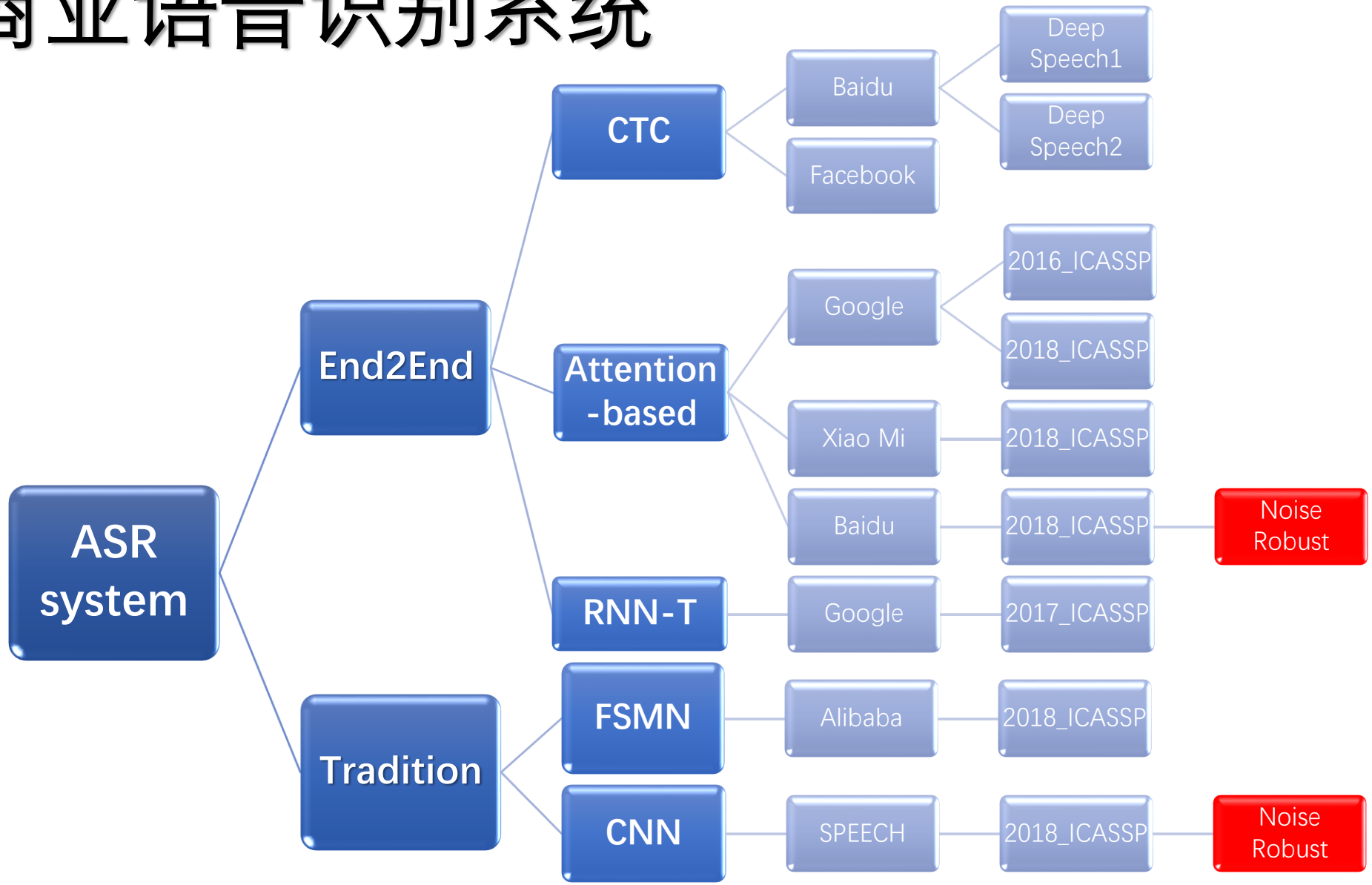
在信号处理学科中，有两种滤波器，分别叫做 IIR 和 FIR，它们和两种神经网络相对应。所提出的FSMN受到数字信号处理中的滤波器设计知识的启发，任何无限脉冲响应（IIR）滤波器都可以使用高阶有限脉冲响应（FIR）滤波器很好地近似。

Zhang, Shiliang, et al. "Feedforward sequential memory networks: A new structure to learn long-term dependency." *arXiv preprint arXiv:1512.08301* (2015).

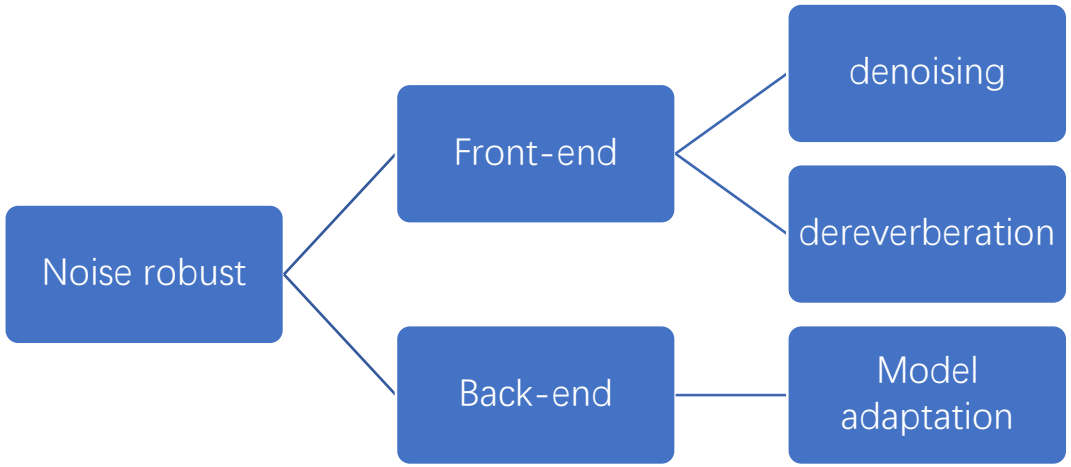
Zhang, Shiliang, et al. "Compact Feedforward Sequential Memory Networks for Large Vocabulary Continuous Speech Recognition." *INTERSPEECH*. 2016.

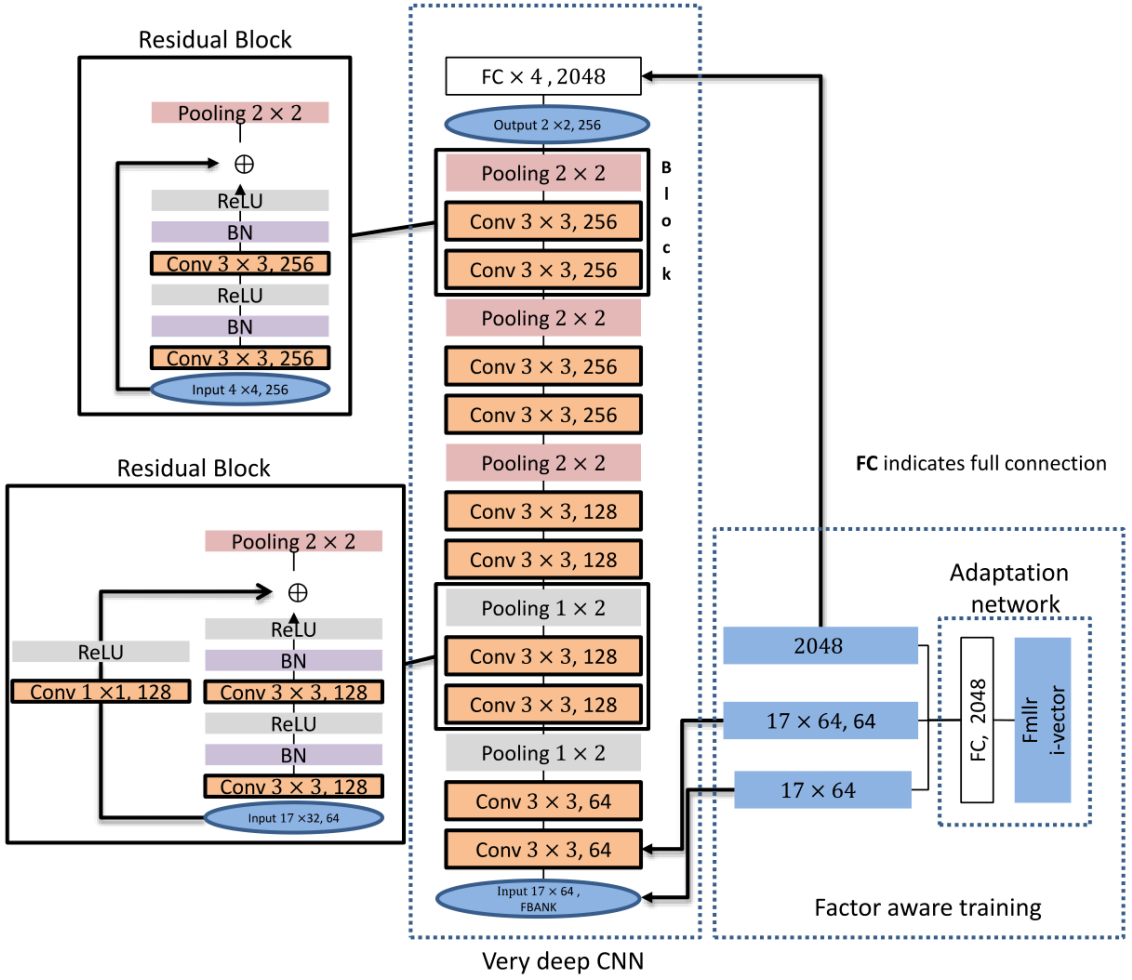
Zhang, Shiliang, et al. "Deep-FSMN for Large Vocabulary Continuous Speech Recognition." *arXiv preprint arXiv:1803.05030* (2018).

商业语音识别系统


ASR system
End2End
Tradition
CTC
Attention -based
RNN-T
FSMN
CNN
Baidu
Facebook
Google
Xiao Mi
Baidu
Google
Alibaba
SPEECH
Deep Speech1
Deep Speech2
2016_ICASSP
2018_ICASSP
2018_ICASSP
2018_ICASSP
2017_ICASSP
2018_ICASSP
2018_ICASSP
Noise Robust
Noise Robust

VDCRN





- Based VDCNN : noise robust superior than other models VDCNN + BN + residual learning
- FAT(facture aware training) & CAT(cluster adaptive training)

Tan, Tian, et al. "Adaptive very deep convolutional residual network for noise robust speech recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*26.8 (2018): 1393-1405.

谢谢！