深度神经网络与语音

陶建华

jhtao@nlpr.ia.ac.cn

中国科学院自动化研究所

模式识别国家重点实验室

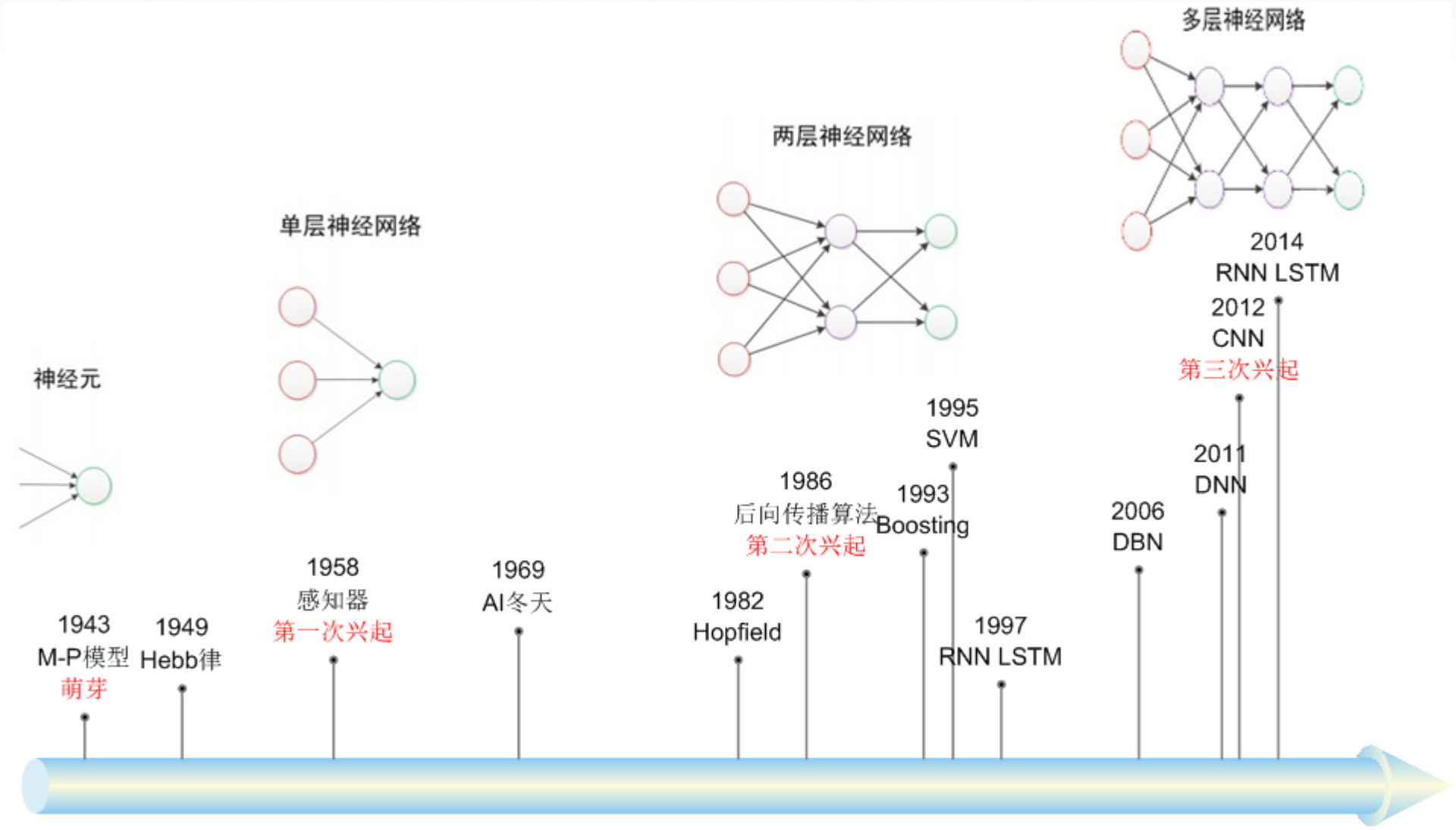
内容

- 神经网络发展历程
- 基于深度神经网络的语音识别
- 基于深度神经网络的语音合成
- 基于深度神经网络的语音增强
- 基于深度神经网络的说话人识别
- 基于深度神经网络的语种识别
- 展望

内容

- 神经网络发展历程
- 基于深度神经网络的语音识别
- 基于深度神经网络的语音合成
- 基于深度神经网络的语音增强
- 基于深度神经网络的说话人识别
- 基于深度神经网络的语种识别
- 展望

神经网络发展历程


多层神经网络
两层神经网络
单层神经网络
神经元
2014
RNN LSTM
2012
CNN
第三次兴起
1995
SVM
2011
DNN
1986
后向传播算法
第二次兴起
1993
Boosting
2006
DBN
1958
感知器
第一次兴起
1969
AI冬天
1982
Hopfield
1997
RNN LSTM
1943
M-P模型
萌芽
1949
Hebb律
1940
2020

神经网络发展历程

- 1940年代
  - 心理学家McCulloch和数学家Pitts建立了M-P模型（1943）
  - 心理学家Hebb提出神经元之间突触联系是可变(可学习)的假说——Hebb律（1949）
- 1950、1960年代
  - 提出并完善了单层感知器（Perceptron）（1958）

    代表人物：Marvin Minsky，Frank Rosenblatt，Bernard Widrow

神经网络发展历程

- 1980年代
  - J. Hopfield提出Hopfield网络（1982）
  - Hinton、Sejnowsky、Rumelhart等人提出了著名的Boltzmann机（1986）
  - Rumelhart等提出多层网络的学习算法—后向传播（BP）算法（1986）
  - Yann LeCun等人提出卷积神经网络（CNN）（1989）

神经网络发展历程

- 1990年代(神经网络的冬天)
  - Schapire等提出Boosting （1993）
  - Vapnik 等提出支持向量机SVM （1995）
  - S.Mike,K.Paliwal等提出RNN（1997）
  - H.Sepp,S.Jürgen等提出LSTM（1997）

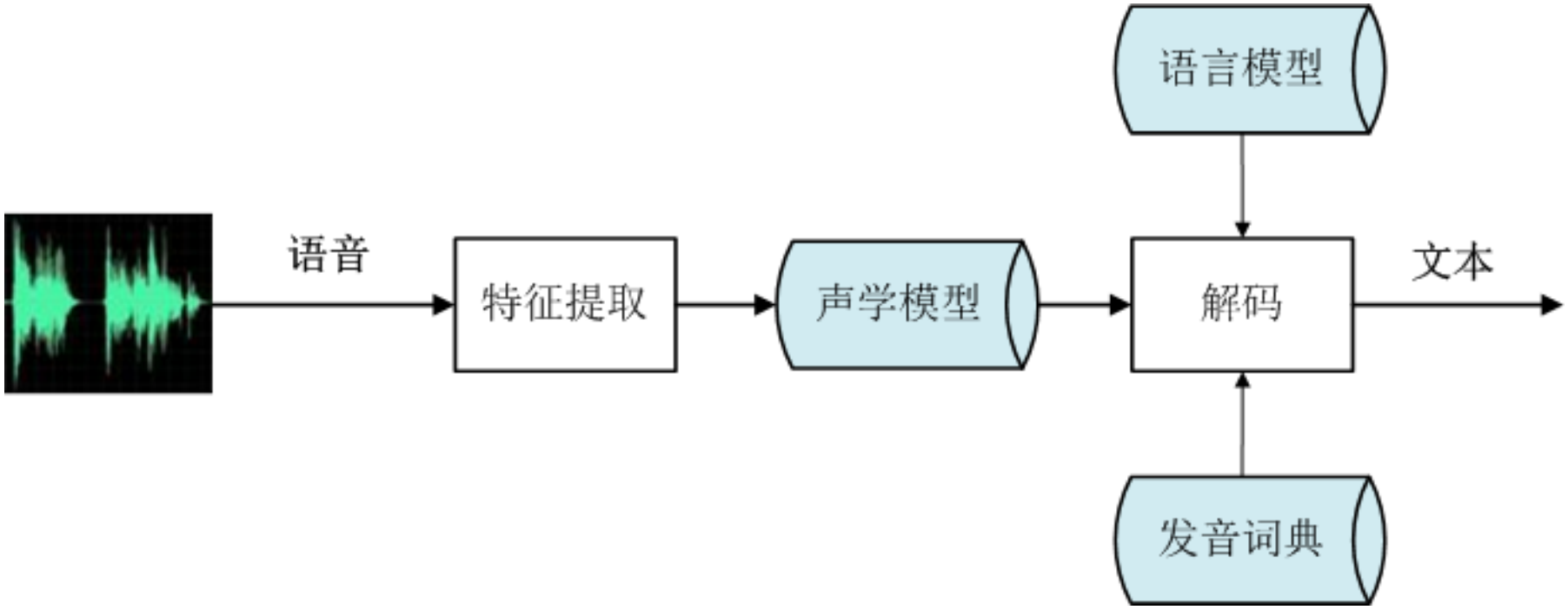
神经网络发展历程

- 2000年代
  - Hinton等提出DBN（2006）
  - 美国国防部DARPA计划首次资助深度学习项目（2010）
  - 微软、谷歌语音识别采用DNN，词错误率降低20%-30%（2011）
  - CNN技术在ImageNet上分类top5错误率由26%降低至15% （2012）
  - RNN LSTM用于语音识别性能超过DNN（2014）
  - 微软提出深度残差网络（2015）
  - 谷歌DeepMind的AlphaGo战胜人类围棋冠军李世石（2016）
  - 谷歌DeepMind发布语音合成WavNet（2016）

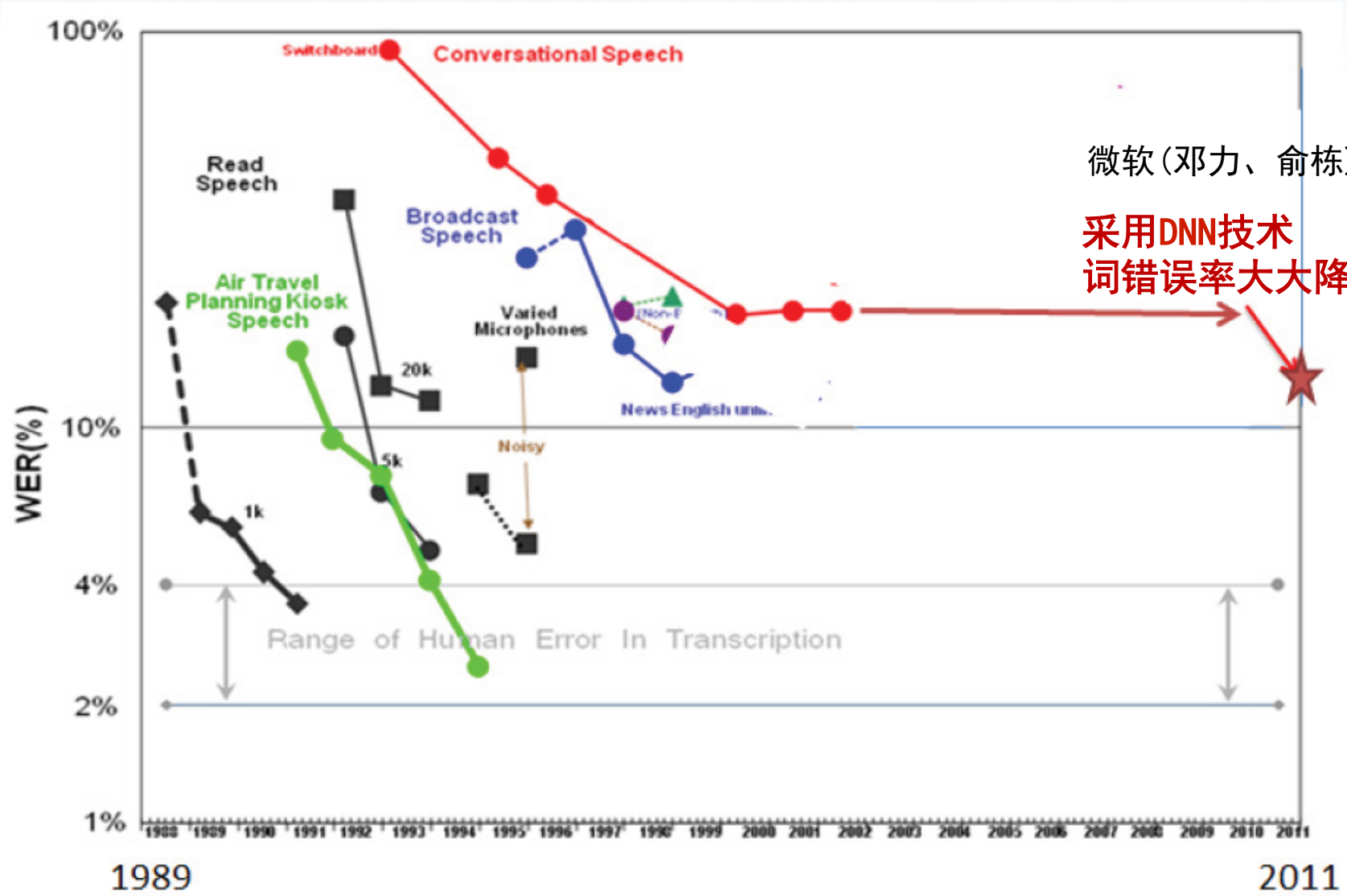
内容

- 神经网络发展历程
- 基于深度神经网络的语音识别
- 基于深度神经网络的语音合成
- 基于深度神经网络的语音增强
- 基于深度神经网络的说话人识别
- 基于深度神经网络的语种识别
- 展望

语音识别



语音识别词错误率



微软(邓力、俞栋)，201

采用DNN技术
词错误率大大降低

声学模型

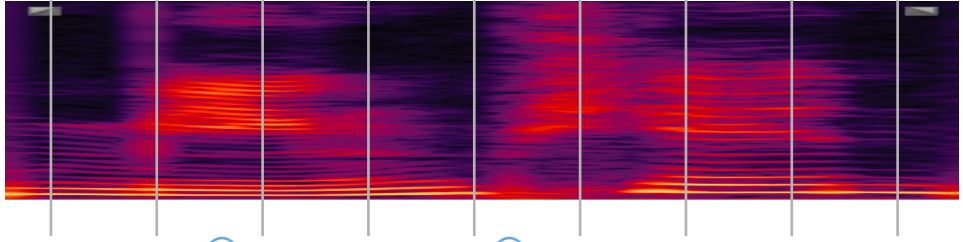
- 混合声学模型
  - 混合高斯-隐马尔科夫模型（GMM-HMM）
  - 深度神经网络-隐马尔科夫模型（DNN-HMM）
  - 深度循环神经网络-隐马尔科夫模型（RNN-HMM）
  - 深度卷积神经网络-隐马尔科夫模型（CNN-HMM）
- 端到端声学模型
  - 连接时序分类-长短时记忆模型（CTC-LSTM）
  - 注意力模型（Attention）

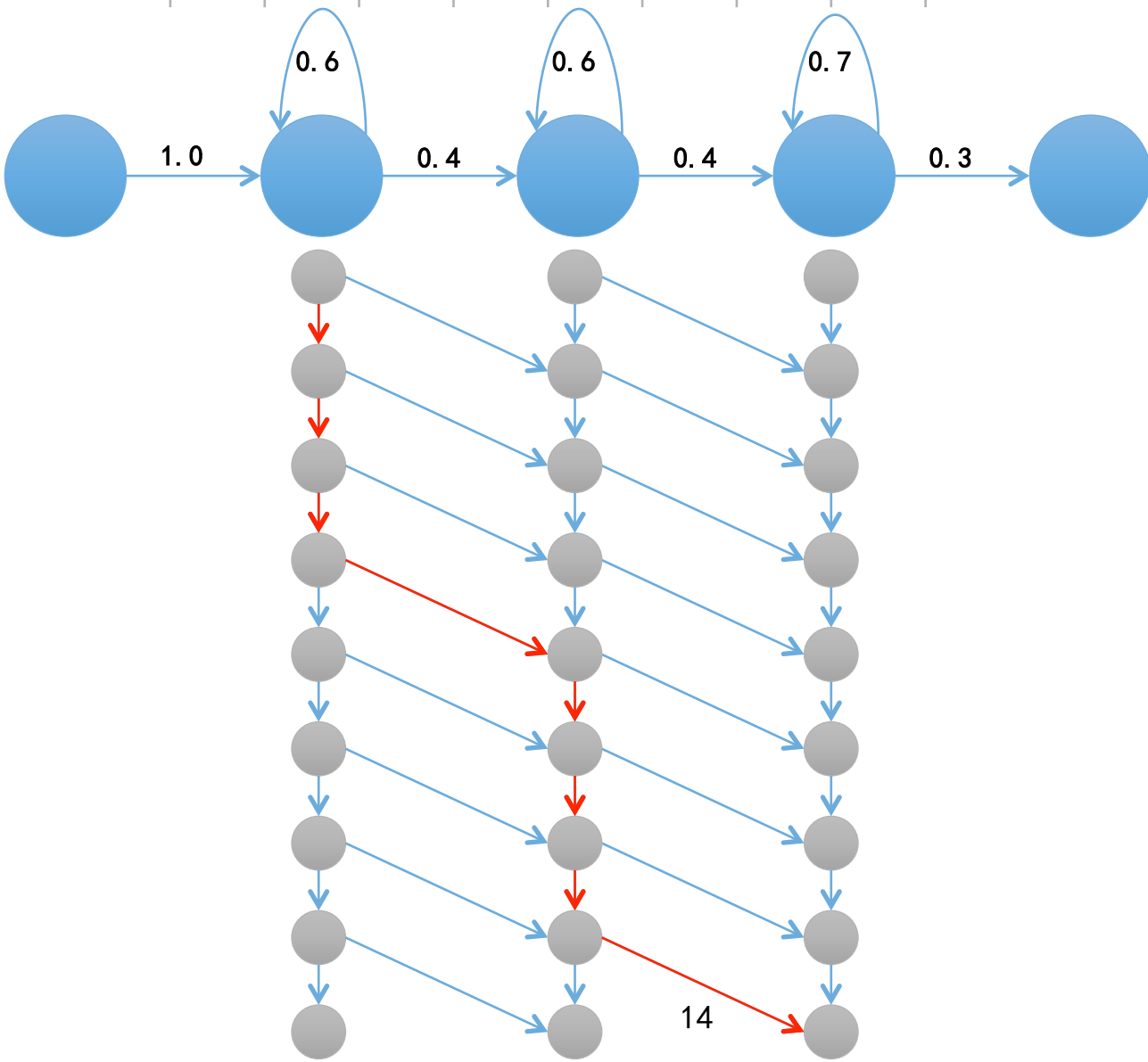
声学模型

- 混合声学模型
  - 混合高斯-隐马尔科夫模型（GMM-HMM）
  - 深度神经网络-隐马尔科夫模型（DNN-HMM）
  - 深度循环神经网络-隐马尔科夫模型（RNN-HMM）
  - 深度卷积神经网络-隐马尔科夫模型（CNN-HMM）
- 端到端声学模型
  - 连接时序分类-长短时记忆模型（CTC-LSTM）
  - 注意力模型（Attention）

基于GMM-HMM模型

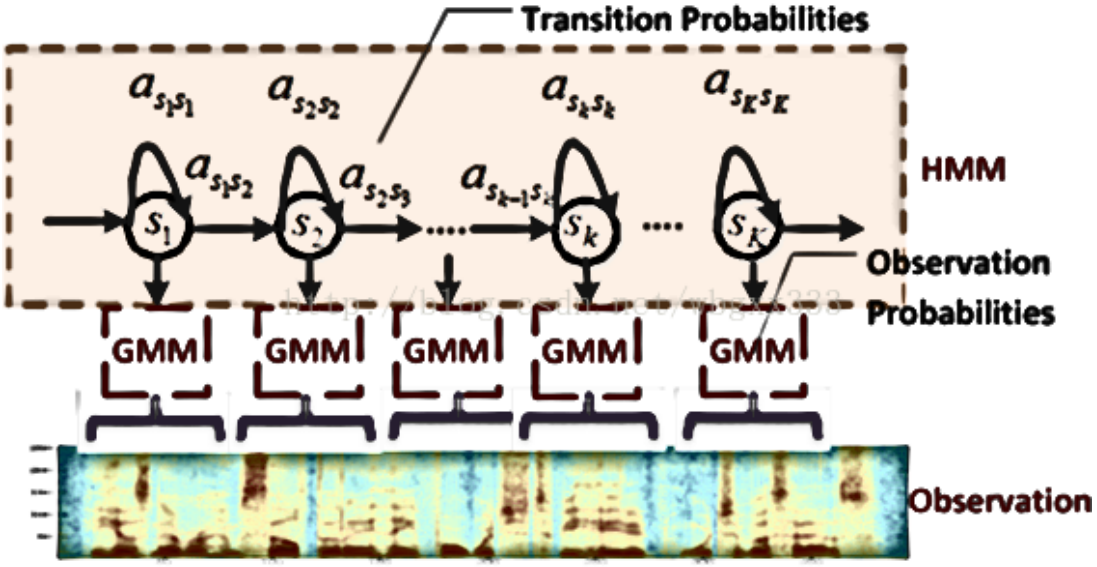
```
~o <VecSize> 39 <MFCC_0_D_A>
~h "proto"
<BeginHMM>
 <NumStates> 5
 <State> 2
    <Mean> 39
      0.0 0.0 0.0 ...
    <Variance> 39
      1.0 1.0 1.0 ...
 <State> 3
    <Mean> 39
      0.0 0.0 0.0 ...
    <Variance> 39
      1.0 1.0 1.0 ...
 <State> 4
    <Mean> 39
      0.0 0.0 0.0 ...
    <Variance> 39
      1.0 1.0 1.0 ...
 <TransP> 5
  0.0 1.0 0.0 0.0 0.0
  0.0 0.6 0.4 0.0 0.0
  0.0 0.0 0.6 0.4 0.0
  0.0 0.0 0.0 0.7 0.3
  0.0 0.0 0.0 0.0 0.0
<EndHMM>
```





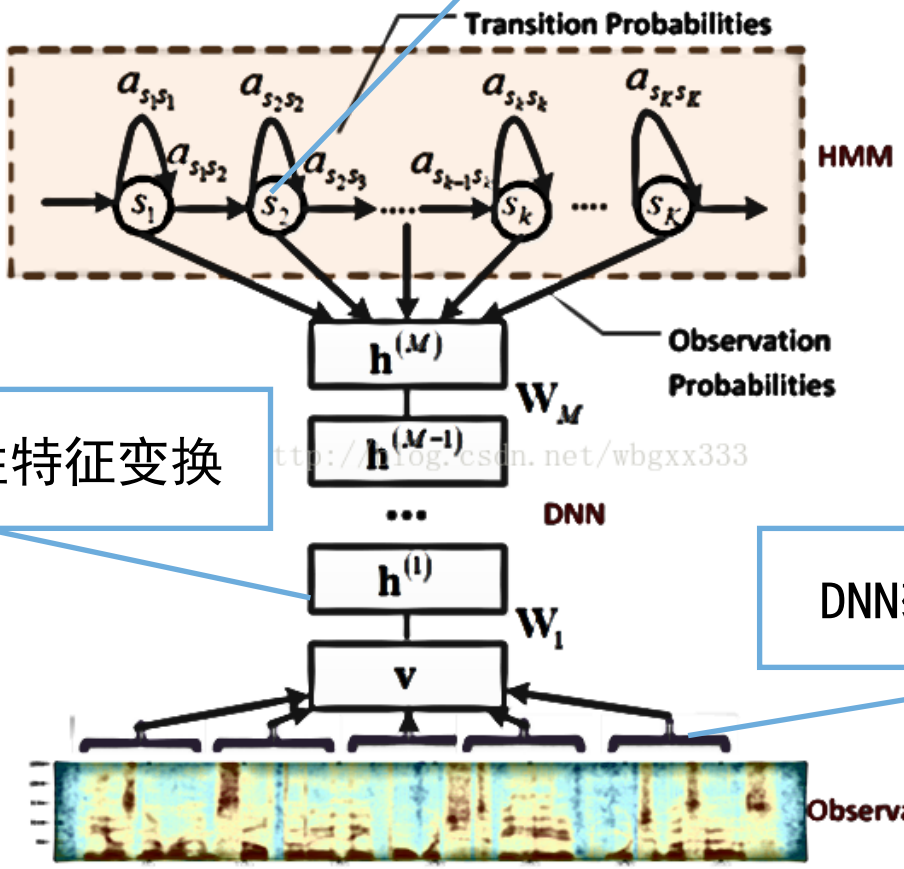
基于GMM-HMM的声学模型

- 优点
  - GMM训练速度快
  - GMM的声学模型小，容易移植到嵌入式平台
- 缺点
  - GMM没有利用帧的上下文信息
  - GMM不能学习深层非线性特征变换



基于DNN-HMM的声学模型

DNN的输出为senones（绑定的三音子）



多层非线性特征变换

DNN输入为较长的帧窗长

混合声学模型

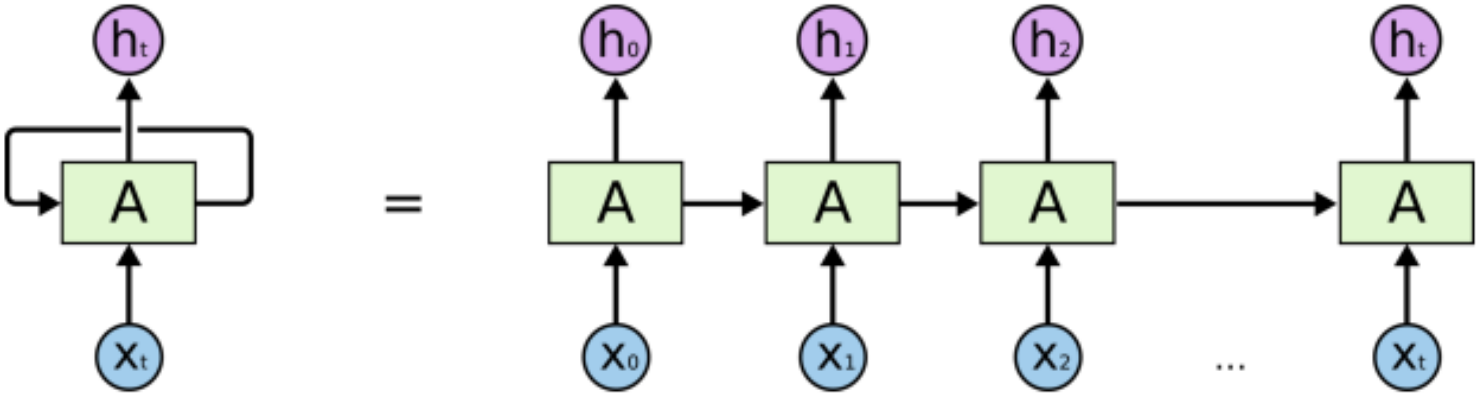
- 深度神经网络的输出标签通过隐马尔科夫模型（HMM）得到，一般为senones（绑定三音子）
- 深度神经网络的训练数据由GMM-HMM模型进行帧对齐得到，即给每帧打上标签（senones）
- 深度神经网络的训练准则：交叉熵（CE）

基于DNN-HMM的声学模型

- 优点
  - DNN能利用帧的上下文信息，比如前后各扩5帧
  - DNN能学习深层非线性特征变换，表现优于GMM
- 缺点
  - 不能利用历史信息来辅助当前的任务

基于RNN-HMM的声学模型

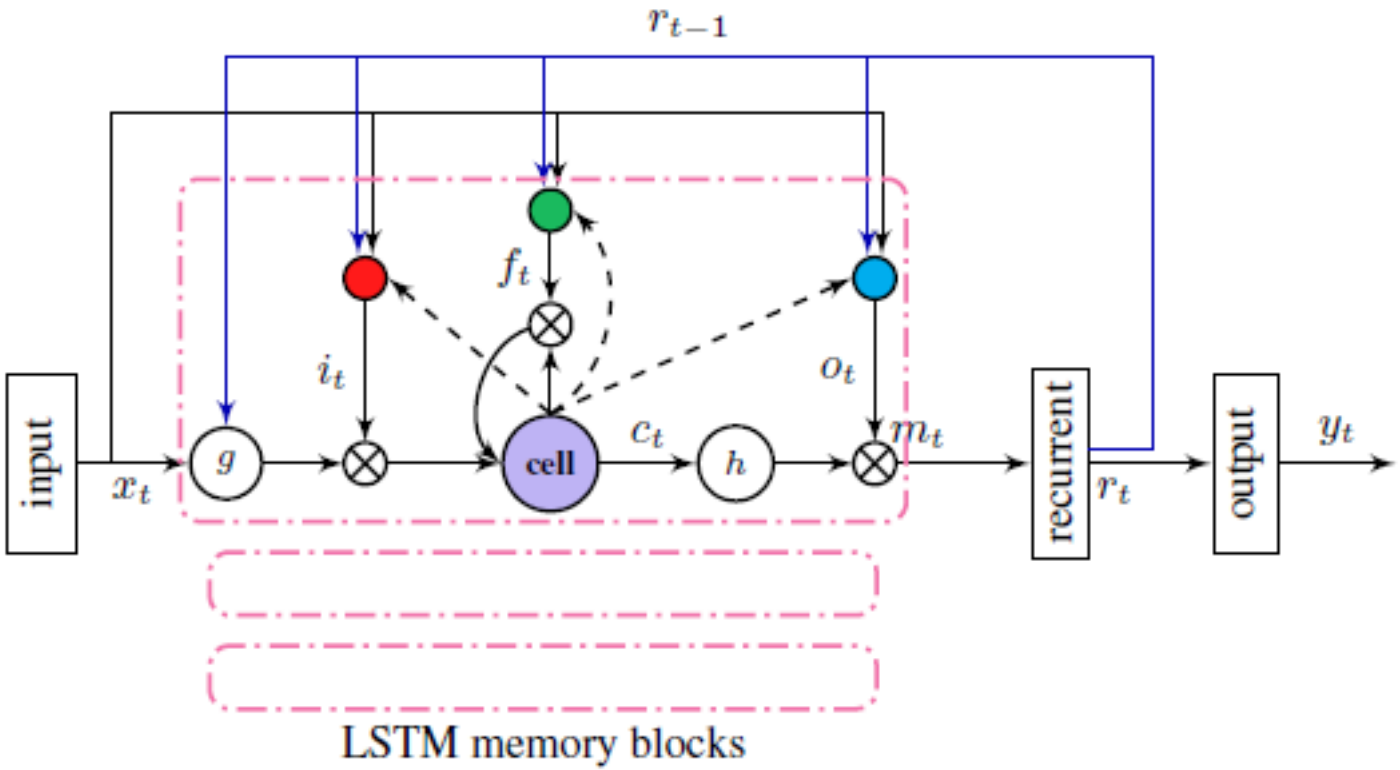
- 优点
  - RNN能有效利用历史信息，将历史信息持久化
  - 在很多任务上，RNN性能表现优于DNN
- 缺点
  - RNN随着层数的增加，会导致梯度爆炸或梯度消失



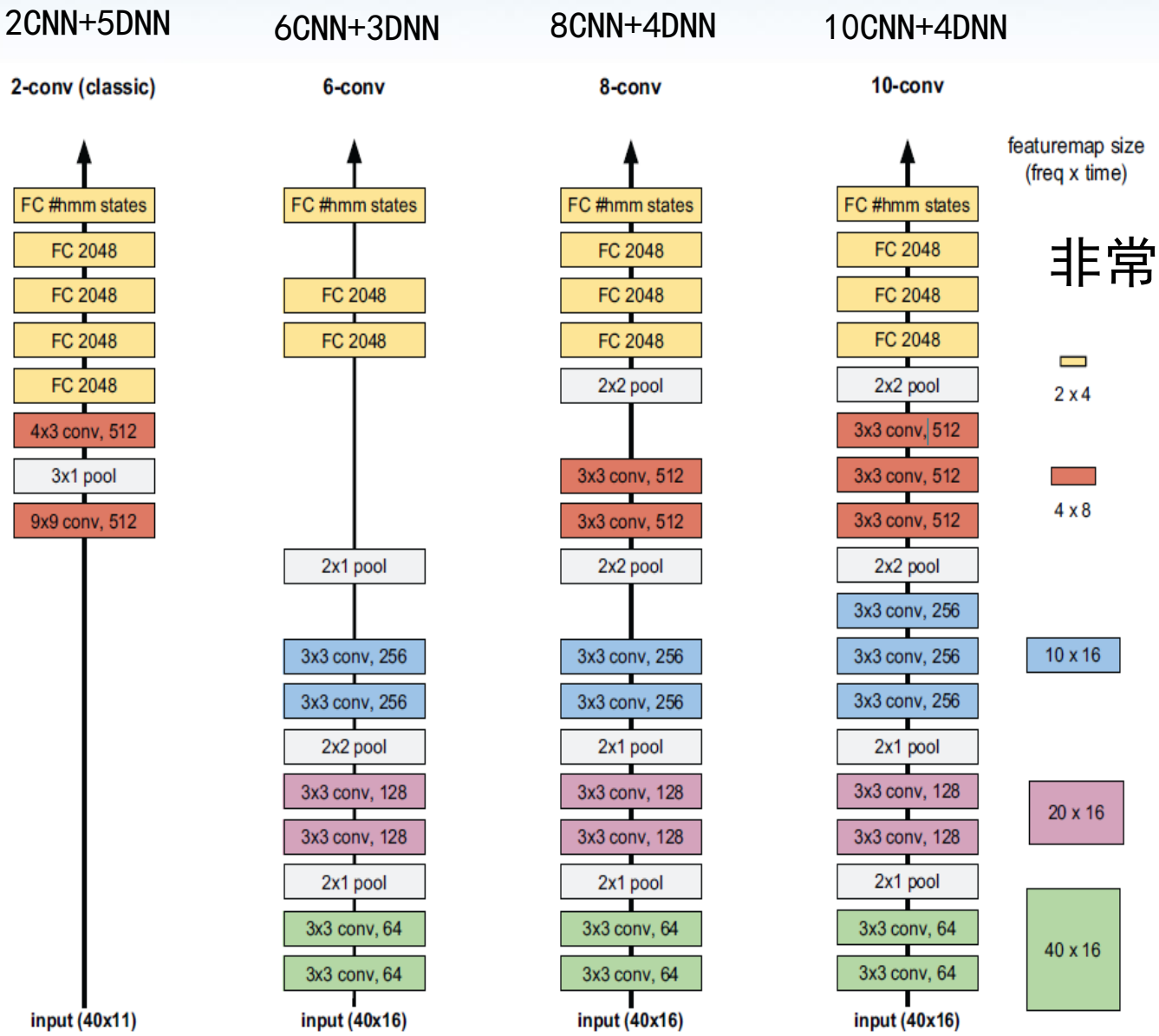
展开的 RNN

基于RNN-HMM的声学模型

- 优点
  - LSTM采用一些控制门（输入门，遗忘门和输出门）来减少梯度累积的长度，一定程度上解决了RNN训练时梯度消失和梯度扩散的问题。



基于CNN-HMM的声学模型



基于CNN-HMM的声学模型

- 优点
  - CNN对于语音信号，采用时间延迟卷积神经网络可以很好地对信号进行描述学习
  - CNN比其他深度神经网络更能捕获到特征的不变性
- 性能
  - 谷歌、微软、IBM均在2016年发表成果证明非常深的CNN声学模型已超越其他深度神经网络的声学模型

Table 2: *Word error rate on the SMD task*

Model	WER
DNN	16.1%
LSTM	14.4%
CNN-LACEA	13.0%

声学模型

- 混合声学模型
  - 混合高斯-隐马尔科夫模型（GMM-HMM）
  - 深度神经网络-隐马尔科夫模型（DNN-HMM）
  - 深度循环神经网络-隐马尔科夫模型（RNN-HMM）
  - 深度卷积神经网络-隐马尔科夫模型（CNN-HMM）
- 端到端声学模型
  - 连接时序分类-长短时记忆模型（CTC-LSTM）
  - 注意力模型（Attention）

端到端声学模型

- 深度神经网络的输出为单因素，三因素或者上下文相关的三因素
- 深度神经网络的训练数据不需要由GMM-HMM强制对齐得到
- 整个训练过程较为简洁

连接时序分类（CTC）

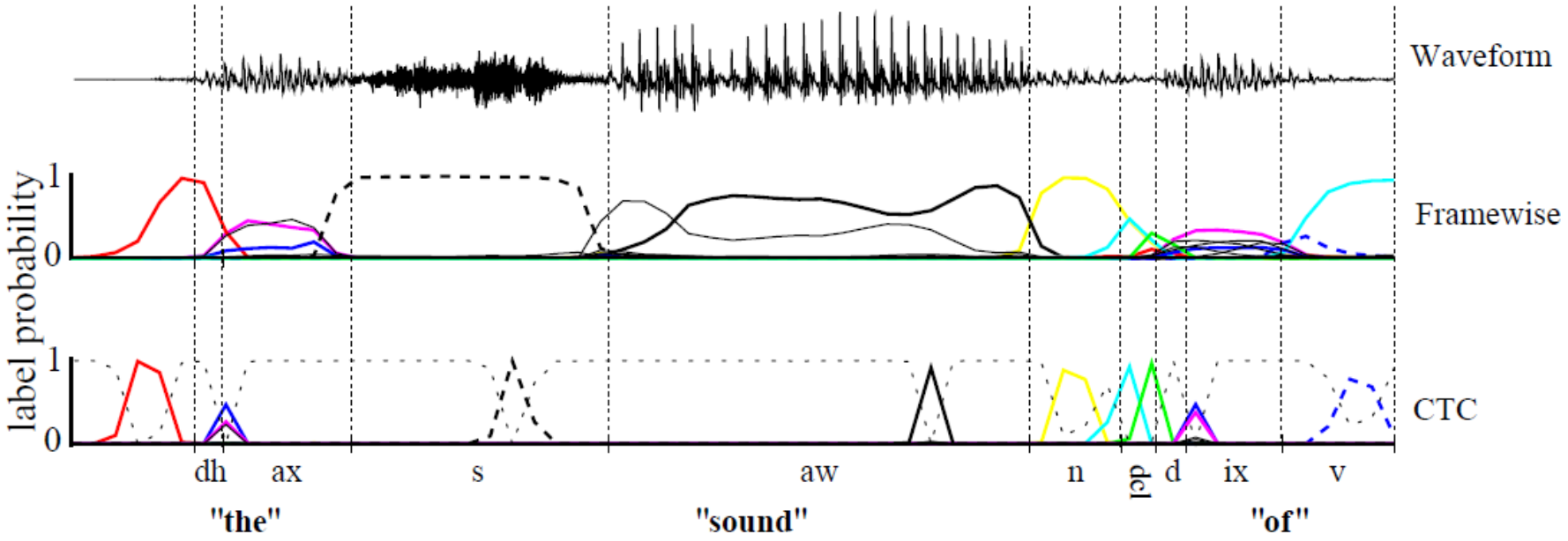
- CTC是一种训练准则，只需要输入和输出在句子级别对齐，不需要帧级别对齐，因此不需要HMM模型

$$L_{ctc} = -\sum_{(x,l)} \ln p(z^l \mid x) = \sum_{x,l} L(x, z^l)$$

$$\frac{\partial L(x, z^t)}{\partial a_l^t} = y_l^t - \frac{1}{p(z^l \mid x)} \sum_{u\{u:z_u^l\}} \alpha_{x,z}(t,u)\beta_{x,z}(t,u)$$

$$p(z^l \mid x) = \sum_{u=1}^{|z^l|} \alpha_{x,z}(t,u)\beta_{x,z}(t,u)$$

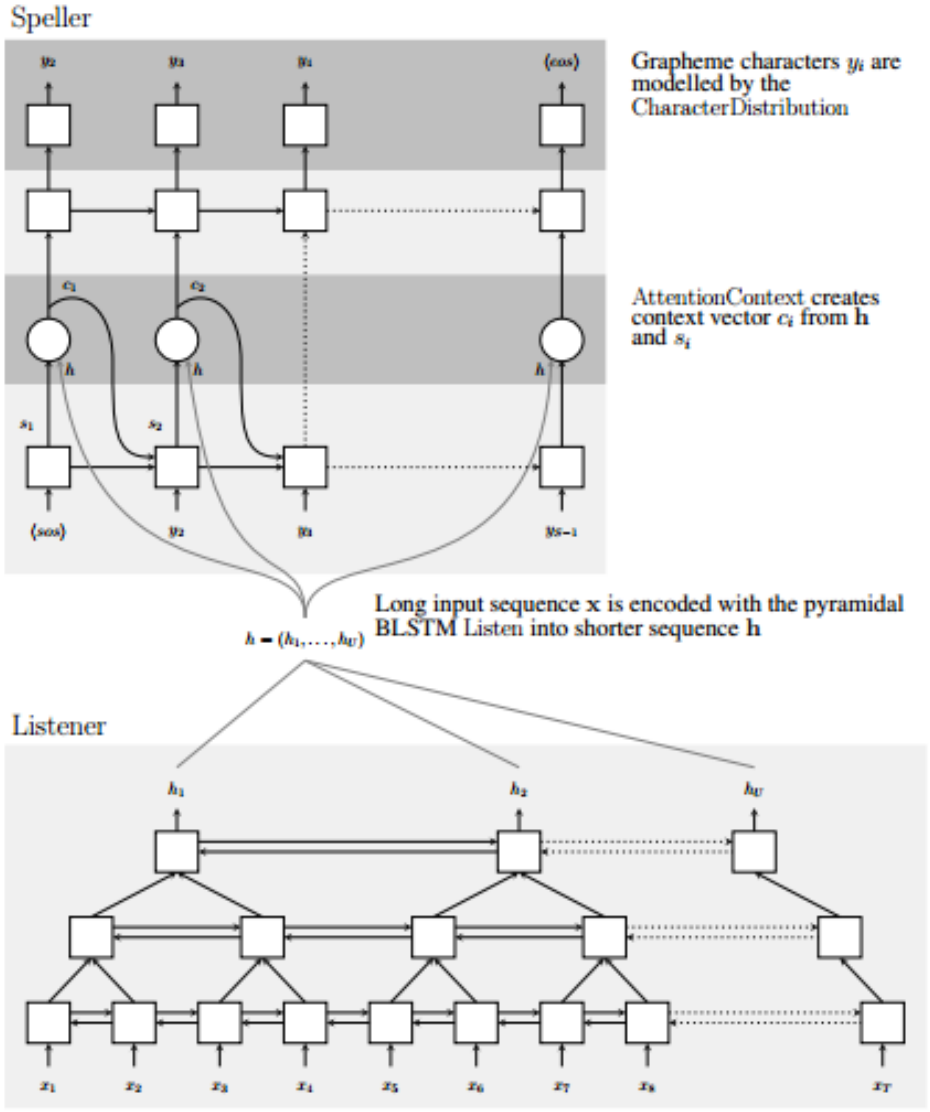
基于CTC-LSTM的声学模型



基于CTC-LSTM的声学模型

- 优点
  - 不要需要输入与输出帧级别的对齐信息，不用和HMM模型结合
  - 约90%的帧其对应的输出为空（blank），可以采取跳帧，加快解码速度
  - 因解码速度快，识别性能也较优，所以工业界大多采用这种模型

基于Attention的声学模型



基于Attention的声学模型

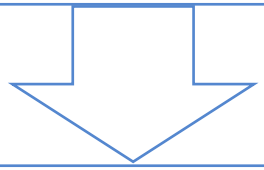
- 优点
  - 从RNN的历史信息中挑选比较重要的信息
  - 思想有趣，模型优雅，训练步骤减少
- 缺点
  - 目前识别精度较低

Model	Clean WER	Noisy WER
CLDNN-HMM [20]	8.0	8.9
LAS	16.2	19.0
LAS + LM Rescoring	12.6	14.7
LAS + Sampling	14.1	16.5
LAS + Sampling + LM Rescoring	10.3	12.0

声学模型训练速度慢

DNN模型：6隐层1026节点，2000小时的语音单GPU（K20）约65天

5000小时，1万小时，10万小时等耗时更多

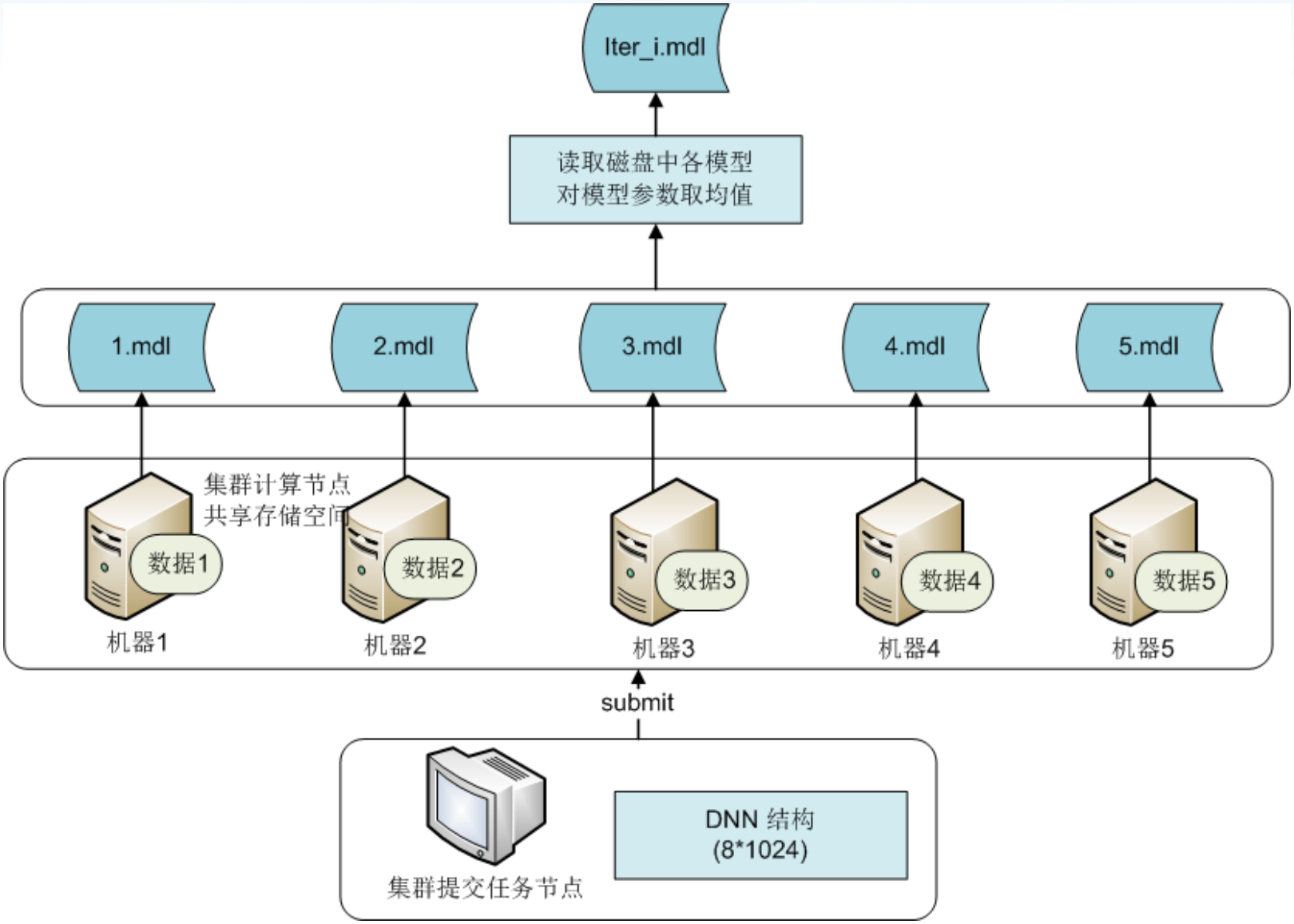


并行计算（多机多GPU）

并行计算方法

- 多机共享存储（网络文件系统，NFS）
- 同步随机梯度下降（MPI）
- 异步随机梯度下降（参数服务器）

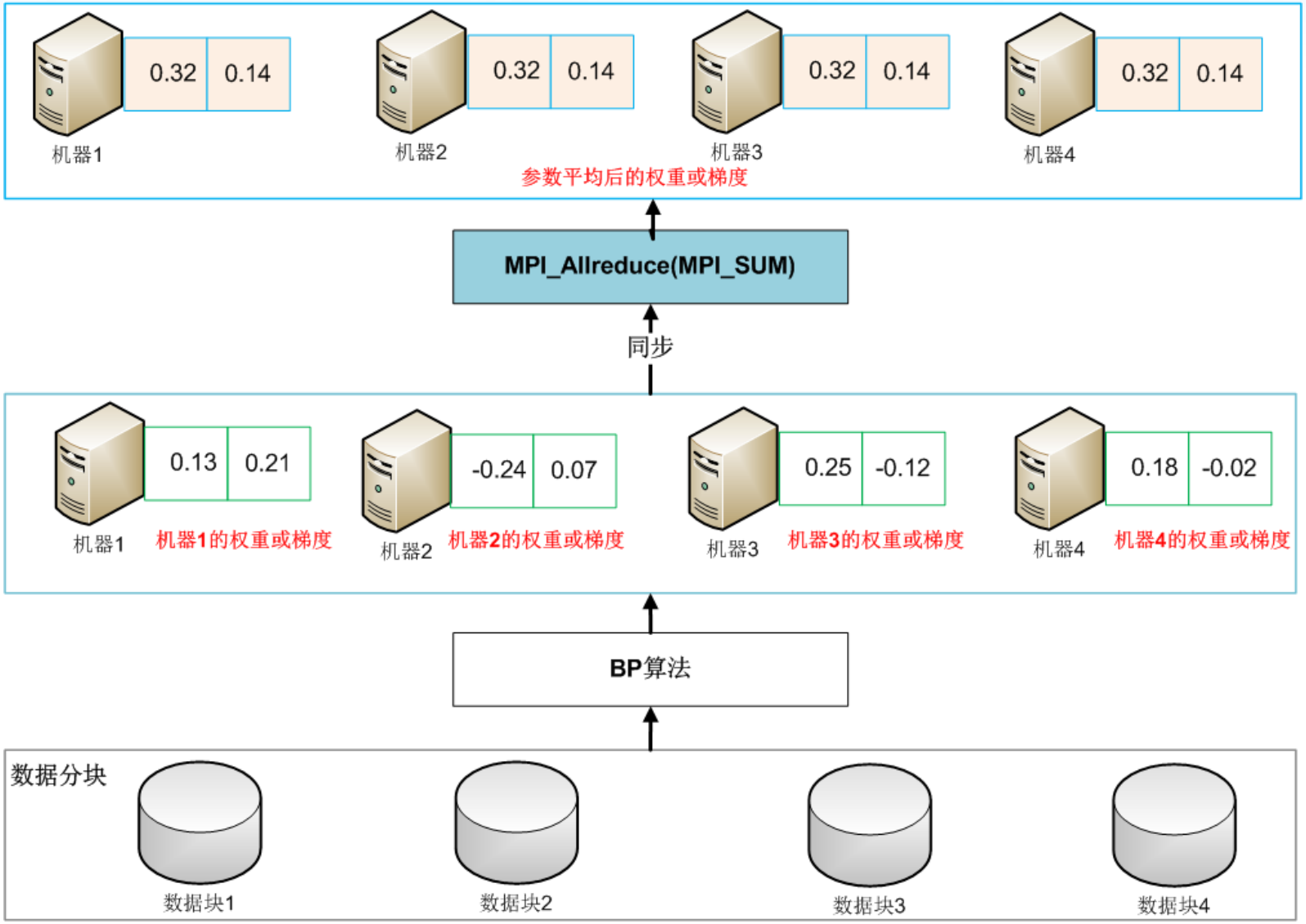
多机共享存储


Iter_i.mdl
读取磁盘中各模型
对模型参数取均值
1.mdl
2.mdl
3.mdl
4.mdl
5.mdl
集群计算节点
共享存储空间
数据1
数据2
数据3
数据4
数据5
机器1
机器2
机器3
机器4
机器5
submit
集群提交任务节点
DNN 结构
(8*1024)

多机共享存储

- 优点
  - 各个模型分开训练，互不干扰
  - 如果某个机器的模型训练失败，只用单独训练此模型
  - 并行计算实现简单
- 缺点
  - 频繁存储和读取磁盘文件，占用大量I/O带宽，训练速度慢
  - 训练过程占用大量磁盘空间

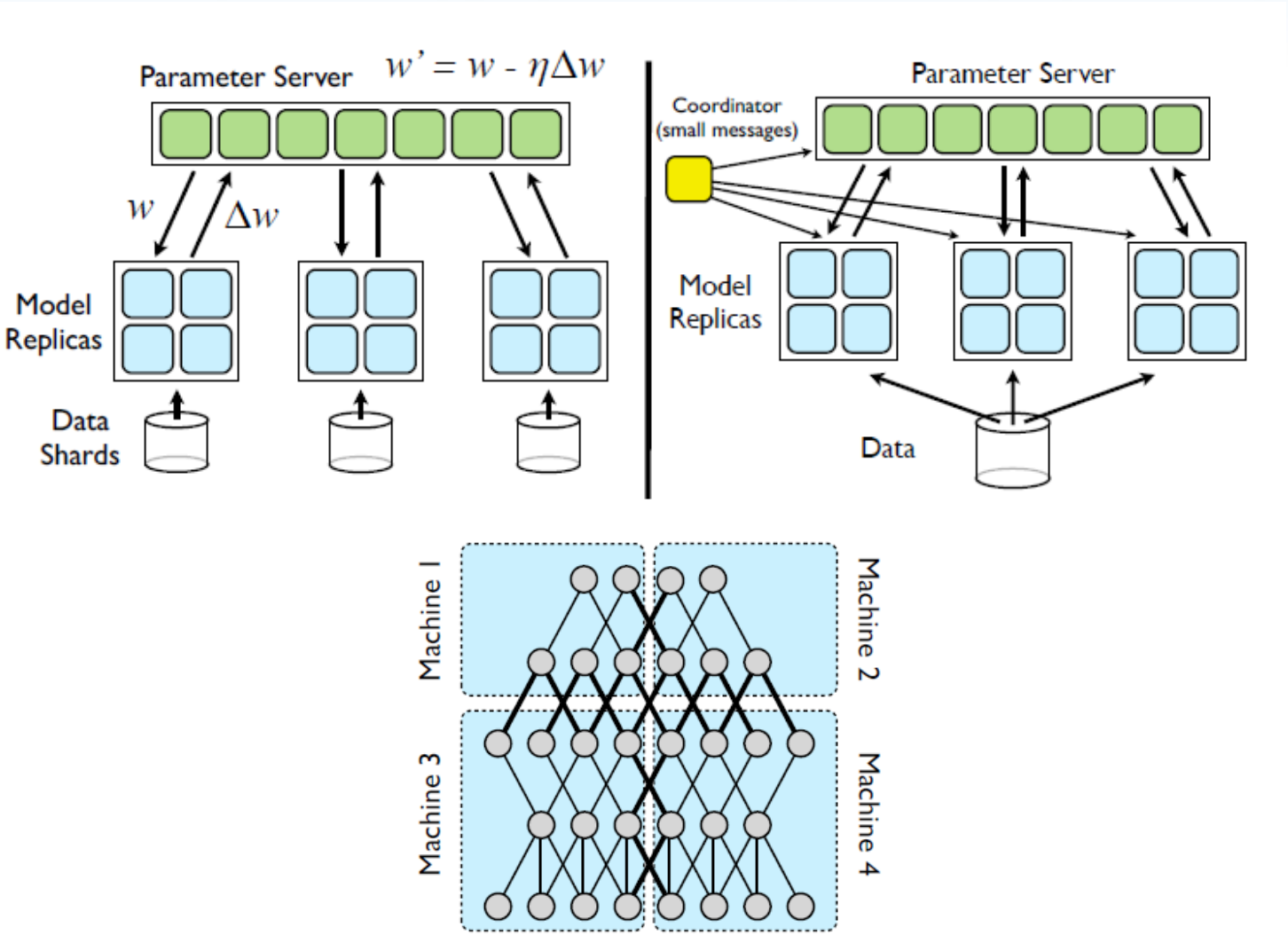
同步随机梯度下降（MPI）



同步随机梯度下降（MPI）

- 优点
  - 模型参数在内存中进行同步和操作，训练速度快
  - 精度损失较小
  - 并行计算实现难度适中（学术界，工业界）
- 缺点
  - 一旦某台机器的模型训练失败，所有机器需从头开始训练

异步随机梯度下降（参数服务器）



Google DistBelief Distributed Framework

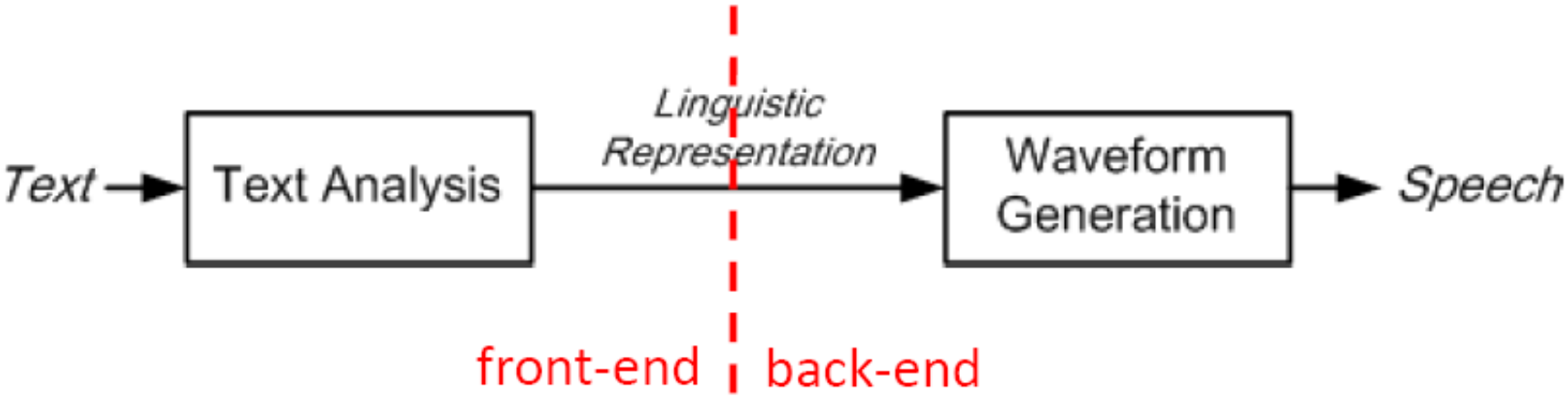
异步随机梯度下降（参数服务器）

- 优点
  - 模型可以分块更新
  - 异步更新模型参数，某台机器训练失败，不需要所有机器重新训练
  - 并行计算实现难度较大（工业界：谷歌，微软，阿里等）
- 缺点
  - 需要过硬的软件和硬件资源，一般只有大企业才有实力开发参数服务器

内容

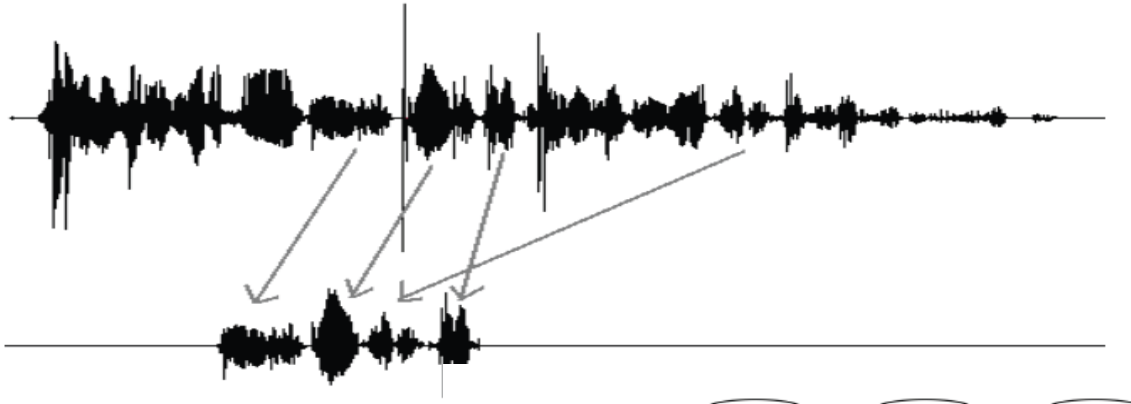
- 神经网络发展历程
- 基于深度神经网络的语音识别
- **基于深度神经网络的语音合成**
- 基于深度神经网络的语音增强
- 基于深度神经网络的说话人识别
- 基于深度神经网络的语种识别
- 展望

语音合成

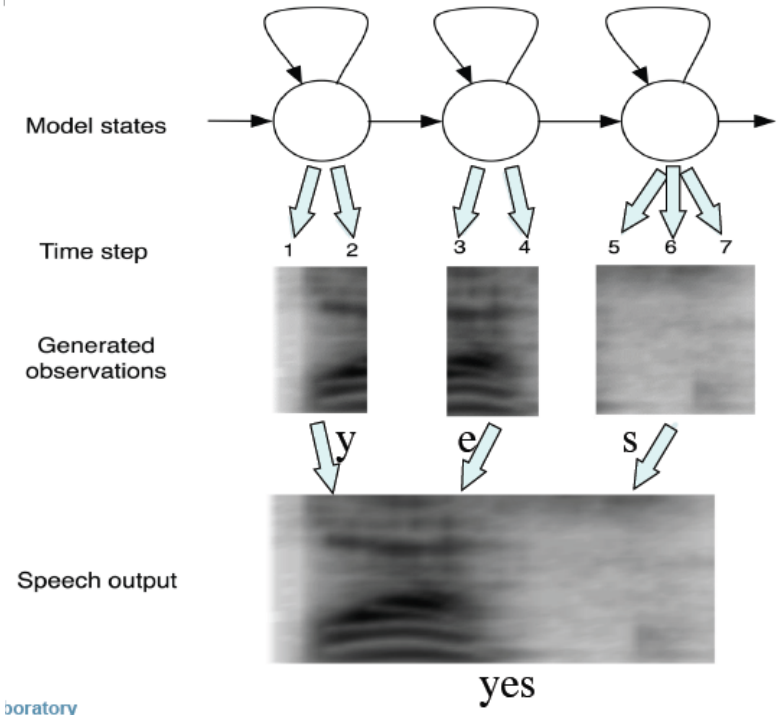

Text
Text Analysis
Linguistic Representation
Waveform Generation
Speech
front-end
back-end

语音合成方法

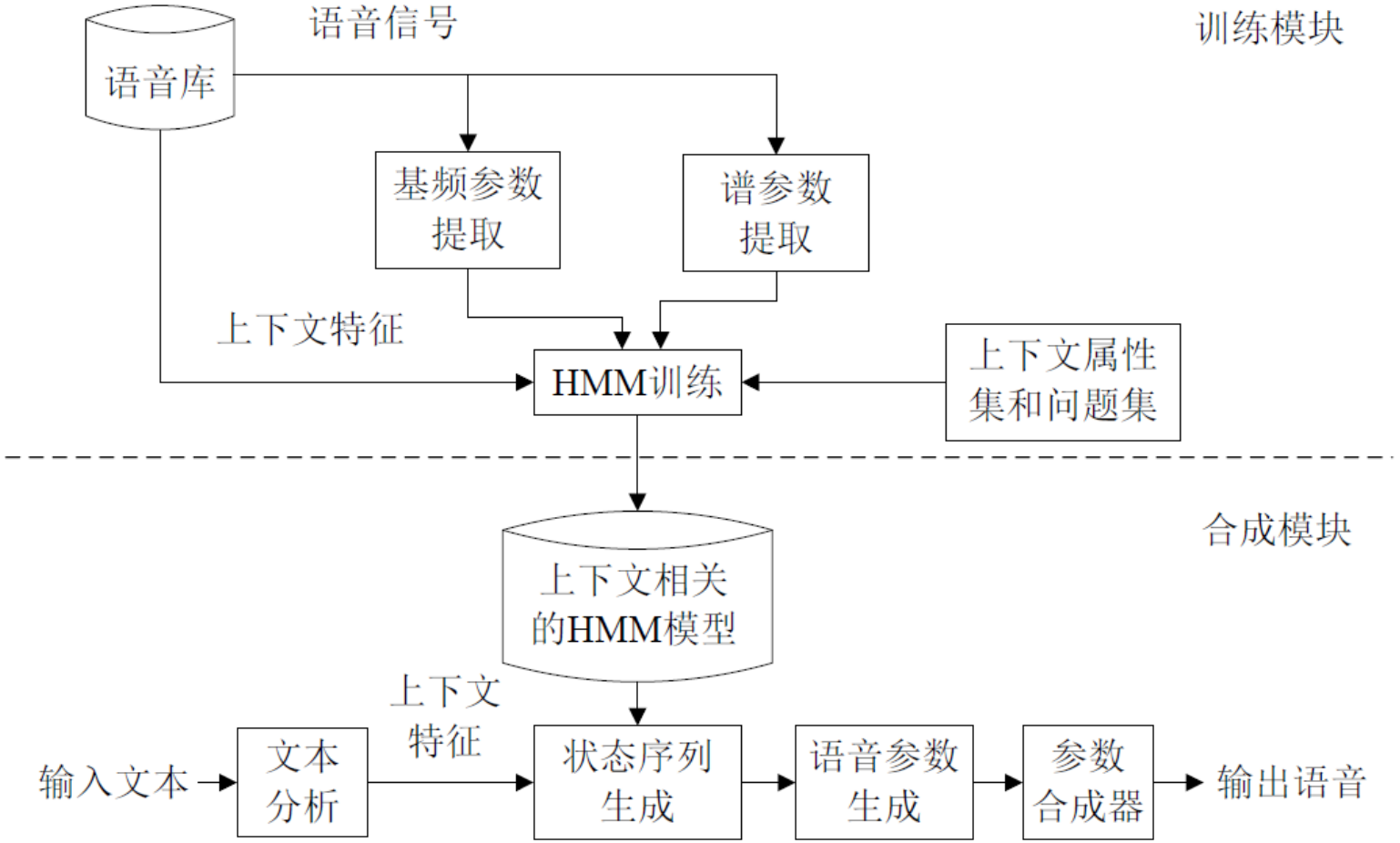
- 基于波形拼接



- 基于统计参数



基于HMM统计参数语音合成



基于HMM统计参数语音合成

- 不足
  - 受限于声码器
  - HMM建模不准确
  - 生成参数不够平滑

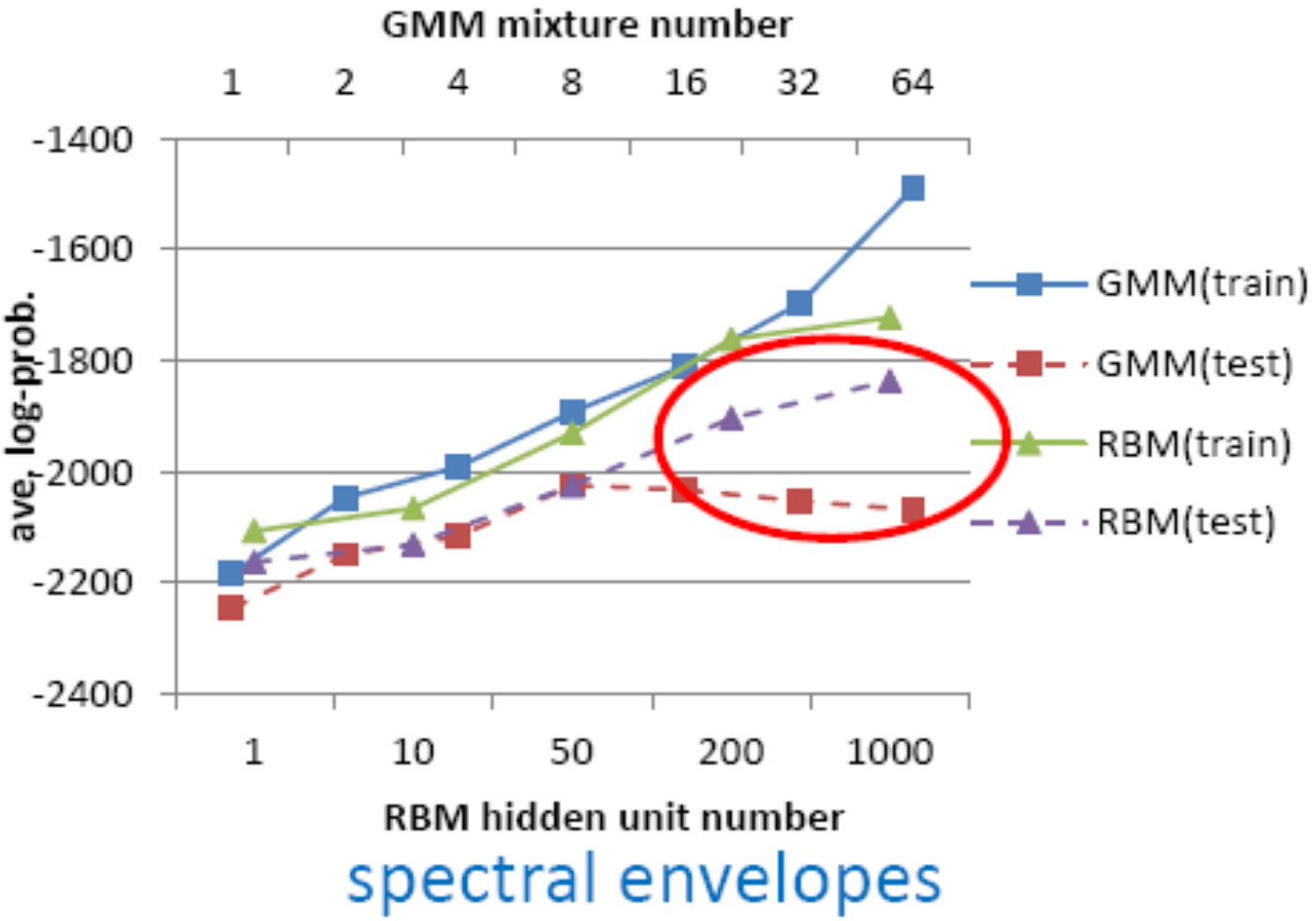
基于深度神经网络的语音合成

- 特点
  - 输入和输出参数通过训练好的HMM模型进行帧对齐处理
  - 神经网络的训练准则为最小均方误差（MSE）
- 声学建模
  - 受限玻尔兹曼机模型（RBM）
  - 深度信念网络（DBN）
  - 深度混合密度网络（DMDN）
  - 深度循环神经网络（RNN）
  - WavNet

基于RBM的语音合成

- GMM $p(v) = \sum_{i=1}^{M} m_i N(u_i, \sigma_i)$
  - 高斯混合密度模型产生的分布不会比组成它的单高斯成分更“尖锐”（过平滑）
  - 需要很多的数据来估计模型参数
- RBM $p(\mathbf{v}) \propto e^{-\frac{1}{2}(\mathbf{v}-\mathbf{b})^{\mathrm{T}}(\mathbf{v}-\mathbf{b})} \prod_{j} \left(1 + e^{c_j + \mathbf{v}^{\mathrm{T}}\mathbf{W}_{*,\mathrm{j}}}\right)$
  - 可以产生比单个成分更“尖锐”的分布
  - 只需要少量数据就可以估计模型参数

基于RBM的语音合成



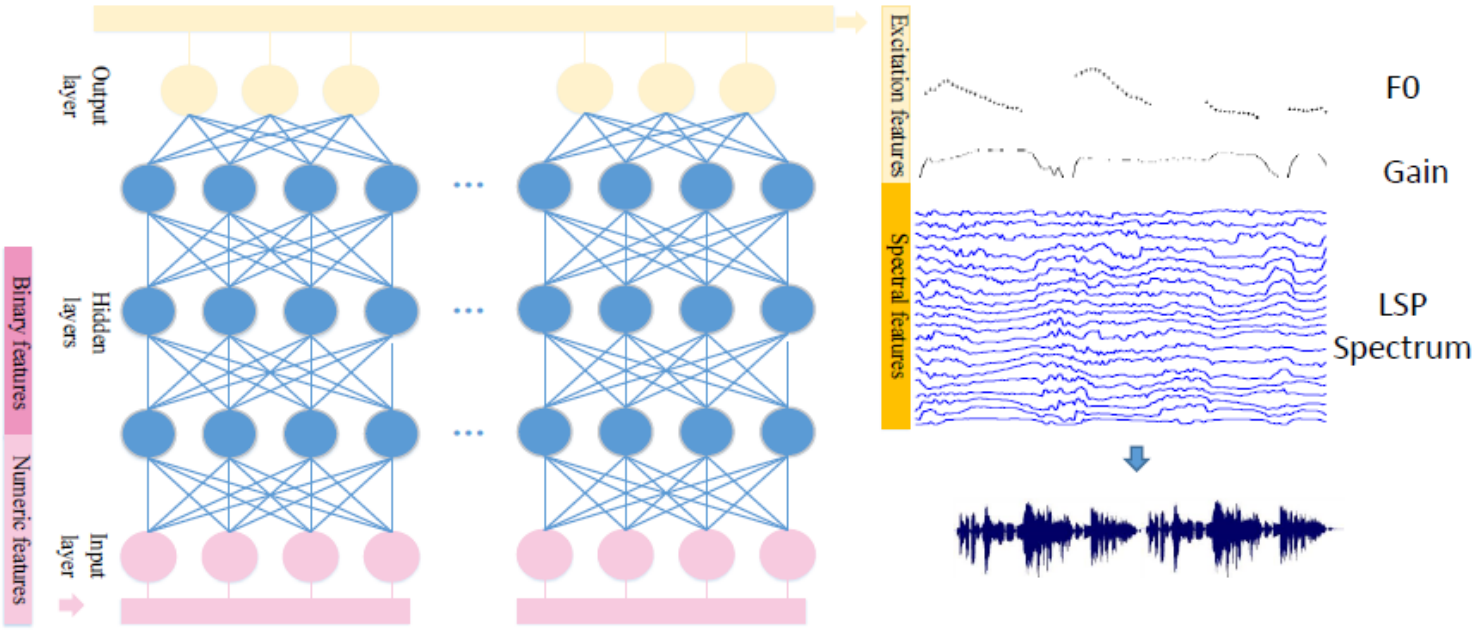
spectral envelopes

基于DBN的语音合成

利用DBN构建文本参数到语音参数之间的映射

不管以前谁
对谁错

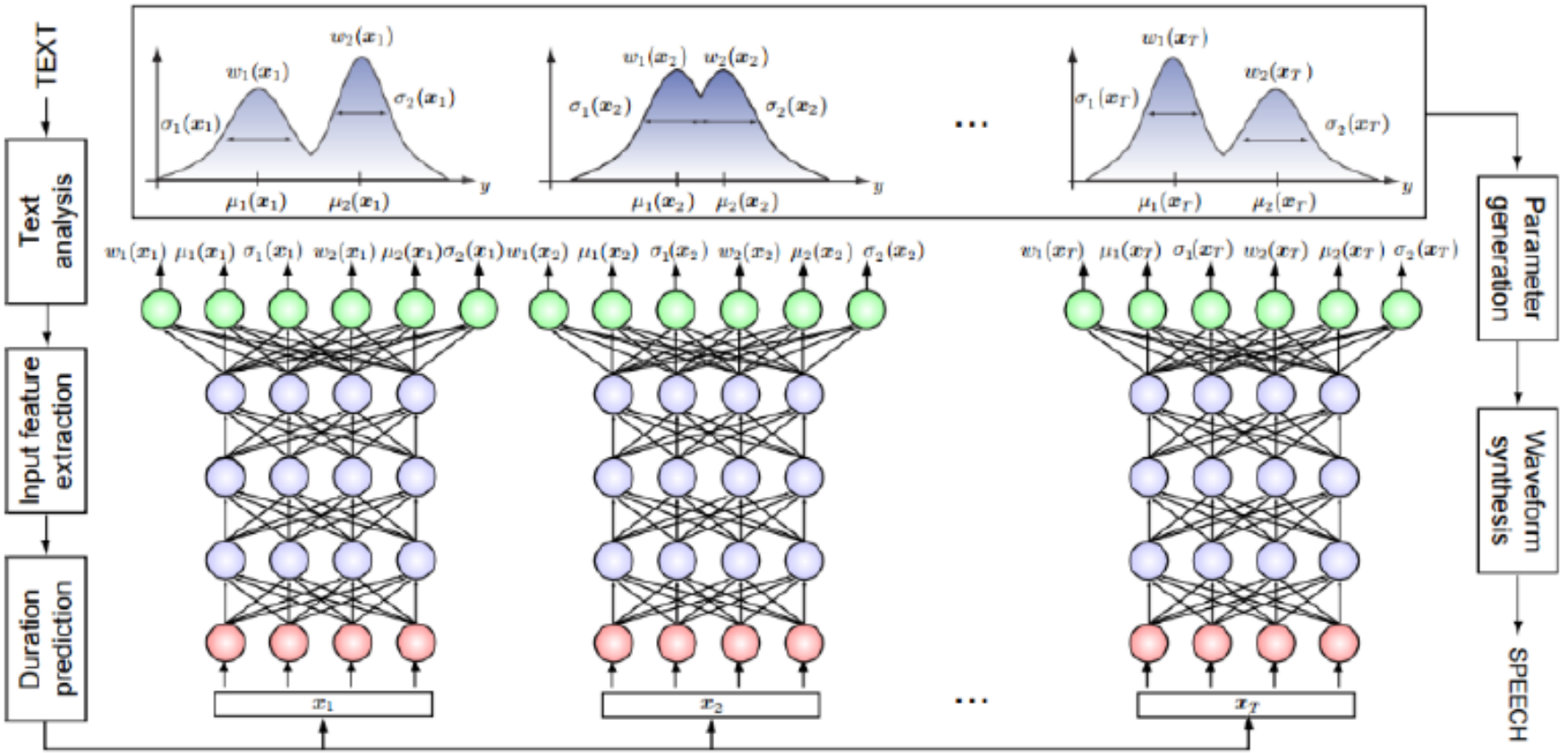
音素标签
词性标签
…
词数
时长
…



预测得到的只有静态及一二阶语音参数，而其方差是通过设定为0.0
1倍的全局方差得到的

基于DMDN的语音合成

- DMDN能够解决DNN不能预测目标参数的方差问题

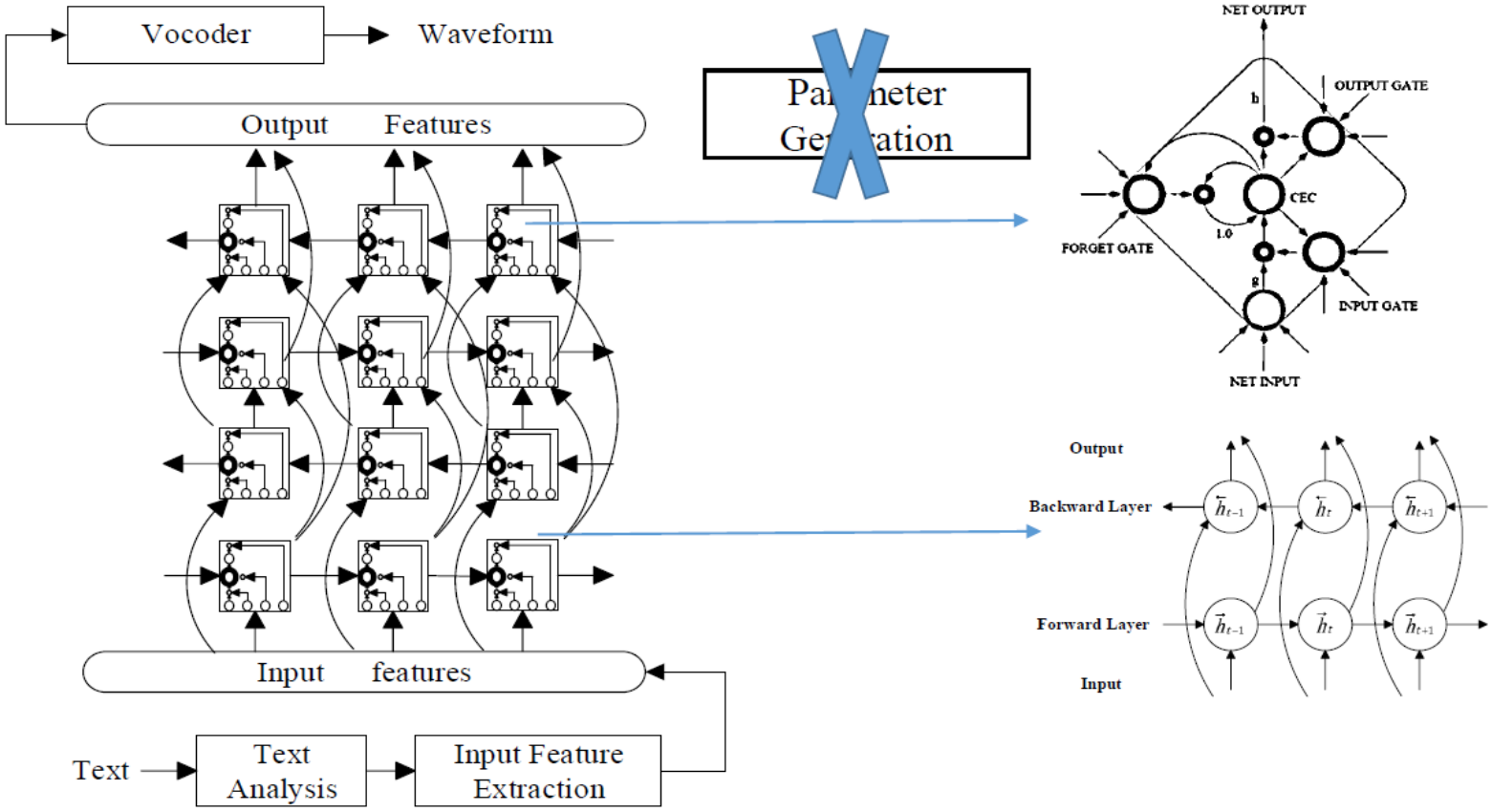


基于DBLSTM-RNN的语音合成

- 语音的生成是一个连续的动态过程
  - 考虑了语意、句法、词性等信息
  - 这些信息与其所在的上下文信息关联性很强
- DBLSTM-RNN
  - 能够综合考虑过去和未来的信息(DNN只能考虑固定的窗口长度)

基于DBLSTM-RNN的语音合成

较之前的网络跳过了参数生成算法



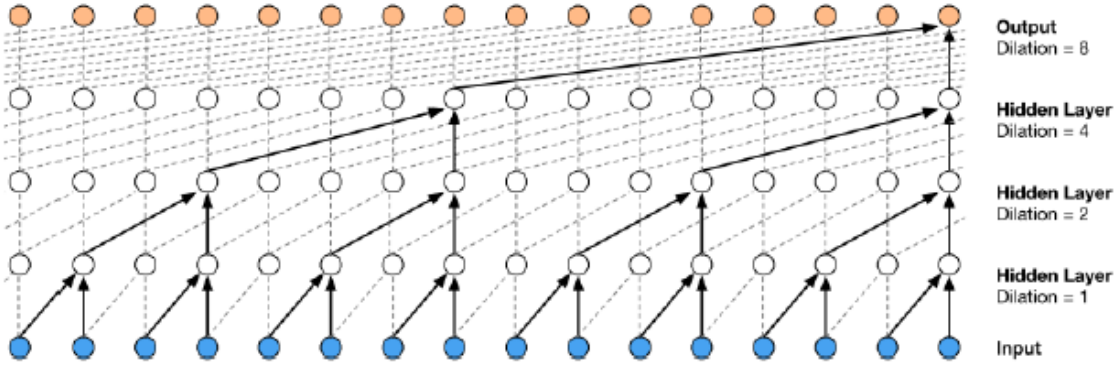
谷歌WavNet

WaveNet by DeepMind [van den Oord 2016]

- Model the joint probability of a waveform using a product of conditional PDFs

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

- The conditional PDF is modelled by a stack of convolutional layers



dilated causal convolutional layers to increase receptive field

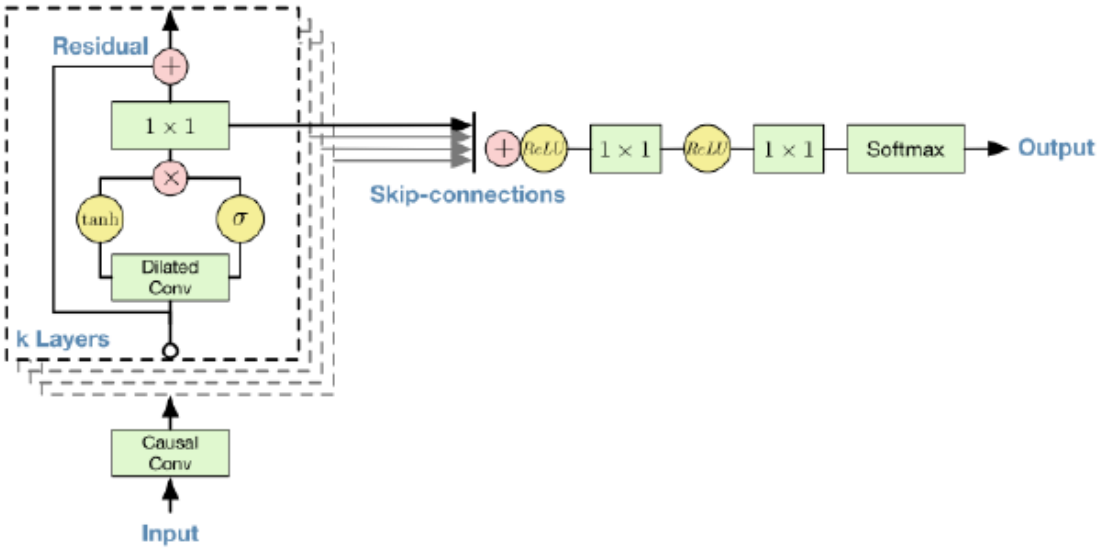
- Gated convolution: works better than ReLU

$$\mathbf{z} = \tanh(W_{f,k} * \mathbf{x}) \odot \sigma(W_{g,k} * \mathbf{x})$$

谷歌WavNet

WaveNet by DeepMind [van den Oord 2016]

- Softmax at output layer
  - μ-law companding, 16bit → 8bit, 65536 → 256
- Residual and skip connections for entire architecture



- Conditional WaveNet for integrating linguistic features for TTS

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y}\right)$$

y为文本特征

谷歌WavNet

Performance

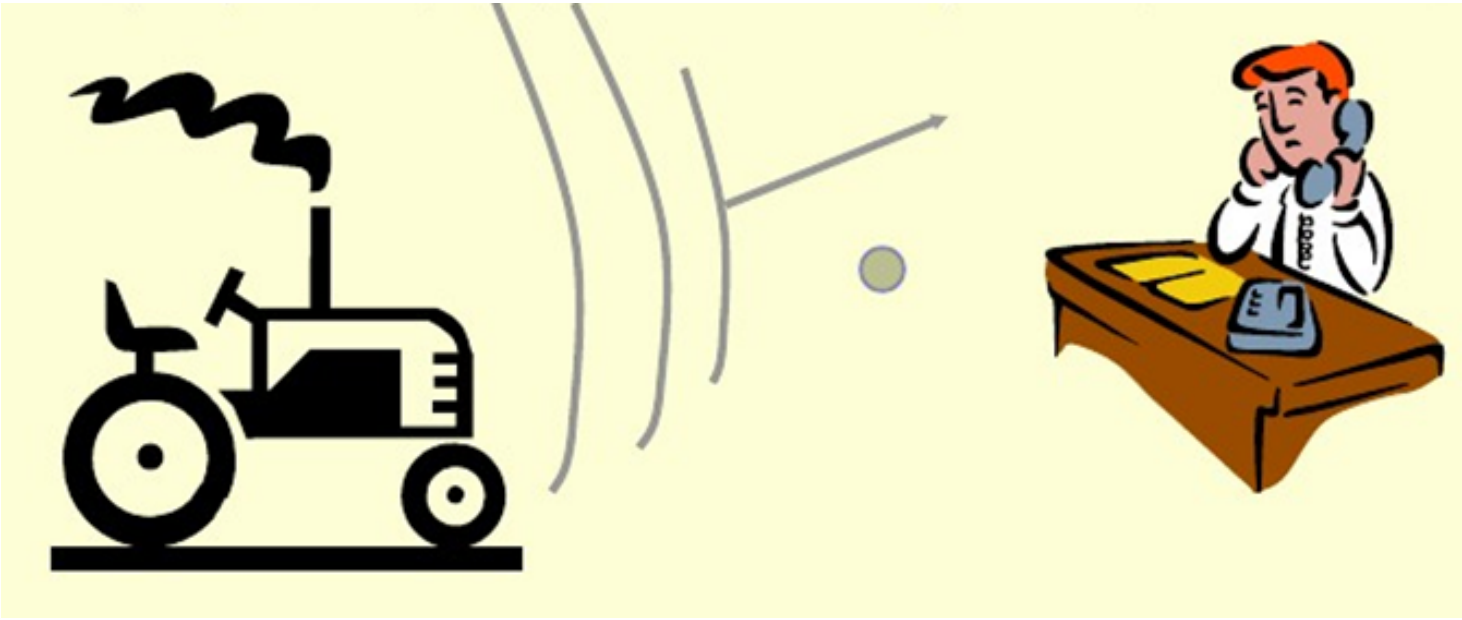
Speech samples	Subjective 5-scale MOS in naturalness	
	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	$\mathbf{4.21} \pm 0.081$	$\mathbf{4.08} \pm 0.085$
Natural (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071

内容

- 神经网络发展历程
- 基于深度神经网络的语音识别
- 基于深度神经网络的语音合成
- **基于深度神经网络的语音增强**
- 基于深度神经网络的说话人识别
- 基于深度神经网络的语种识别
- 展望

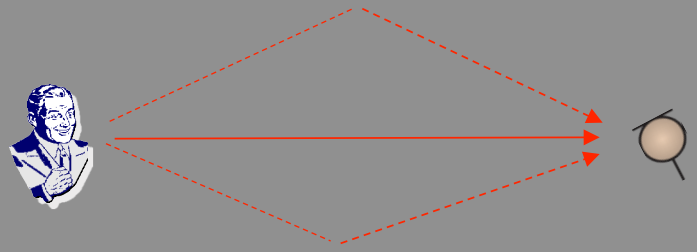
语音增强

- 语音增强是指当语音信号被不同噪声干扰、甚至淹没后，从噪声背景中提取有用的语音信号，抑制噪声干扰的技术。

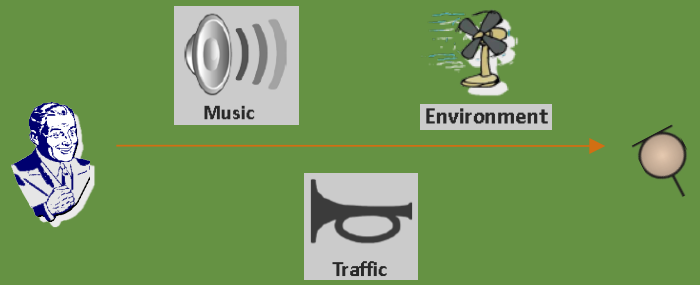


噪声类型

混响： Reverberation

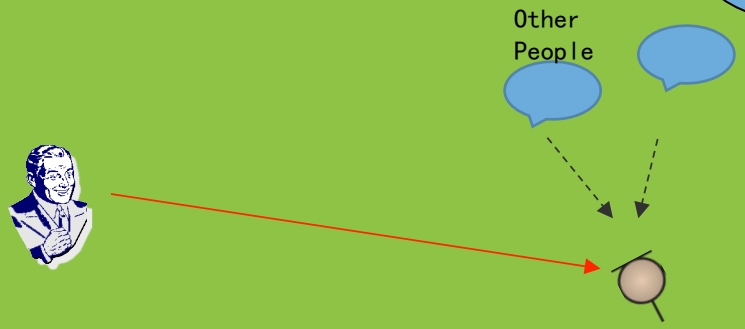


背景噪声： Background Noise

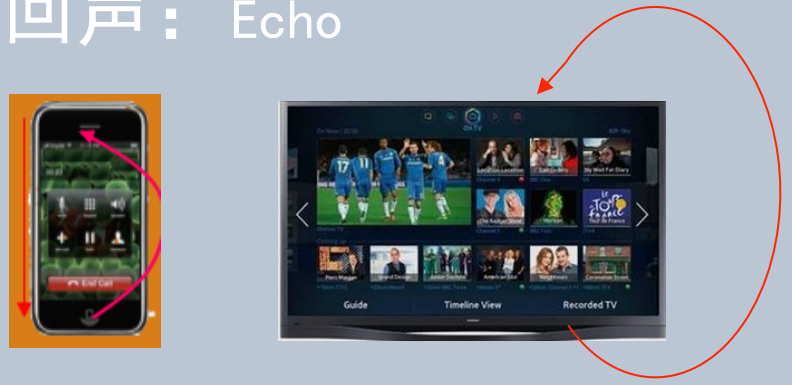


噪声

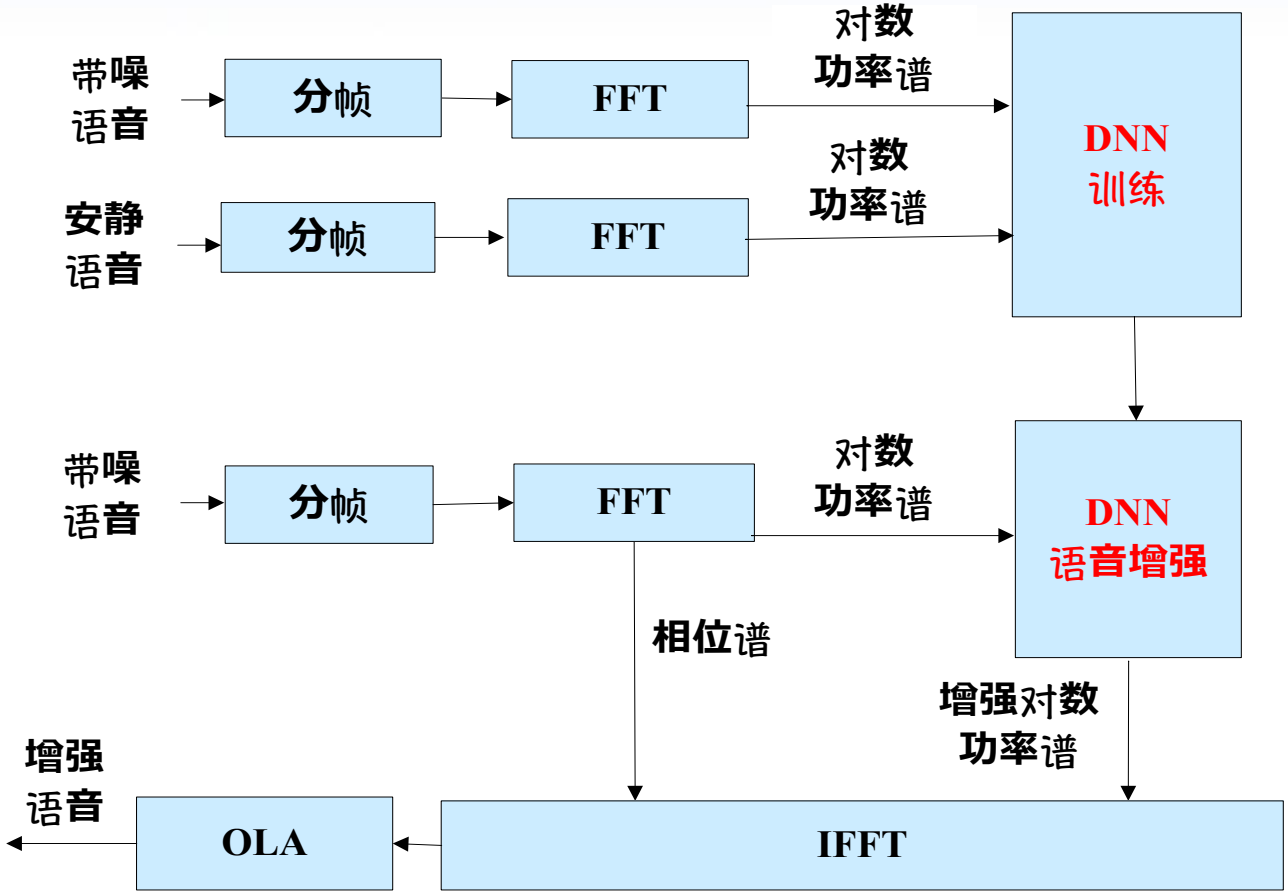
人声干扰： Interference



回声： Echo



基于DNN的语音增强



建立带噪语音与安静语音对数功率谱的映射关系

基于DNN的增强

语谱图对比分析

5dB信噪比babble噪声语谱图

左上：带噪语音

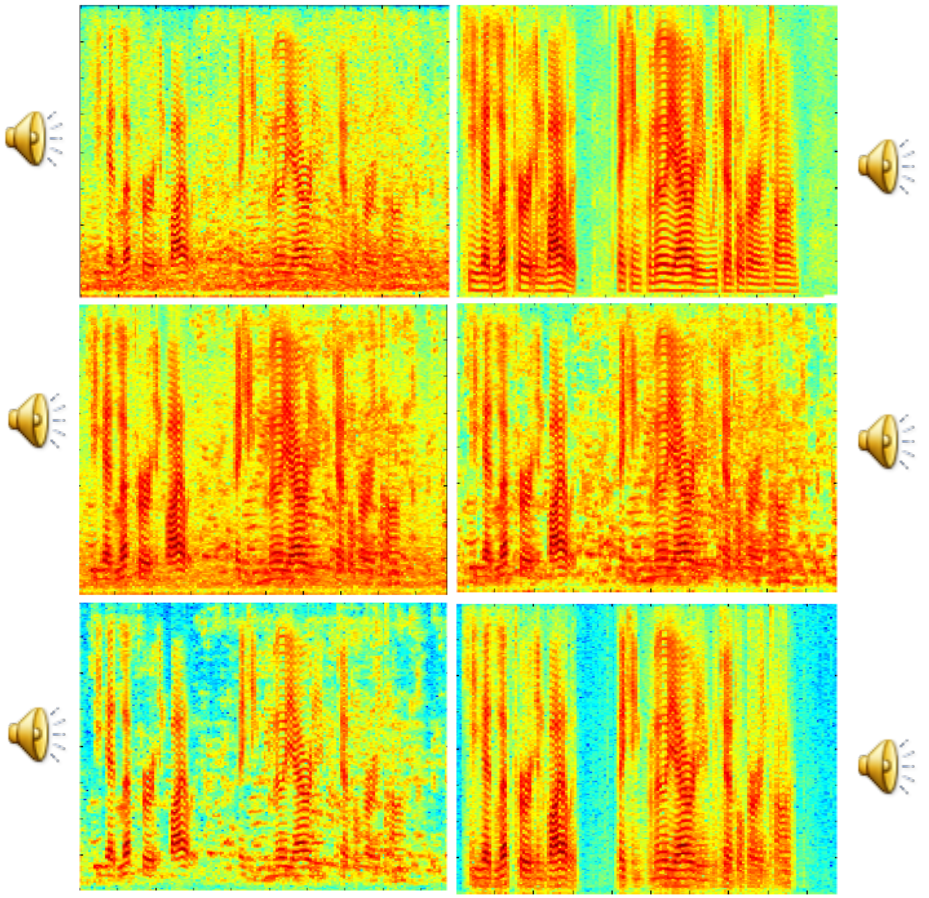
右上：安静语音

左中：子带谱减法

右中：维纳滤波法

左下：logmmse法

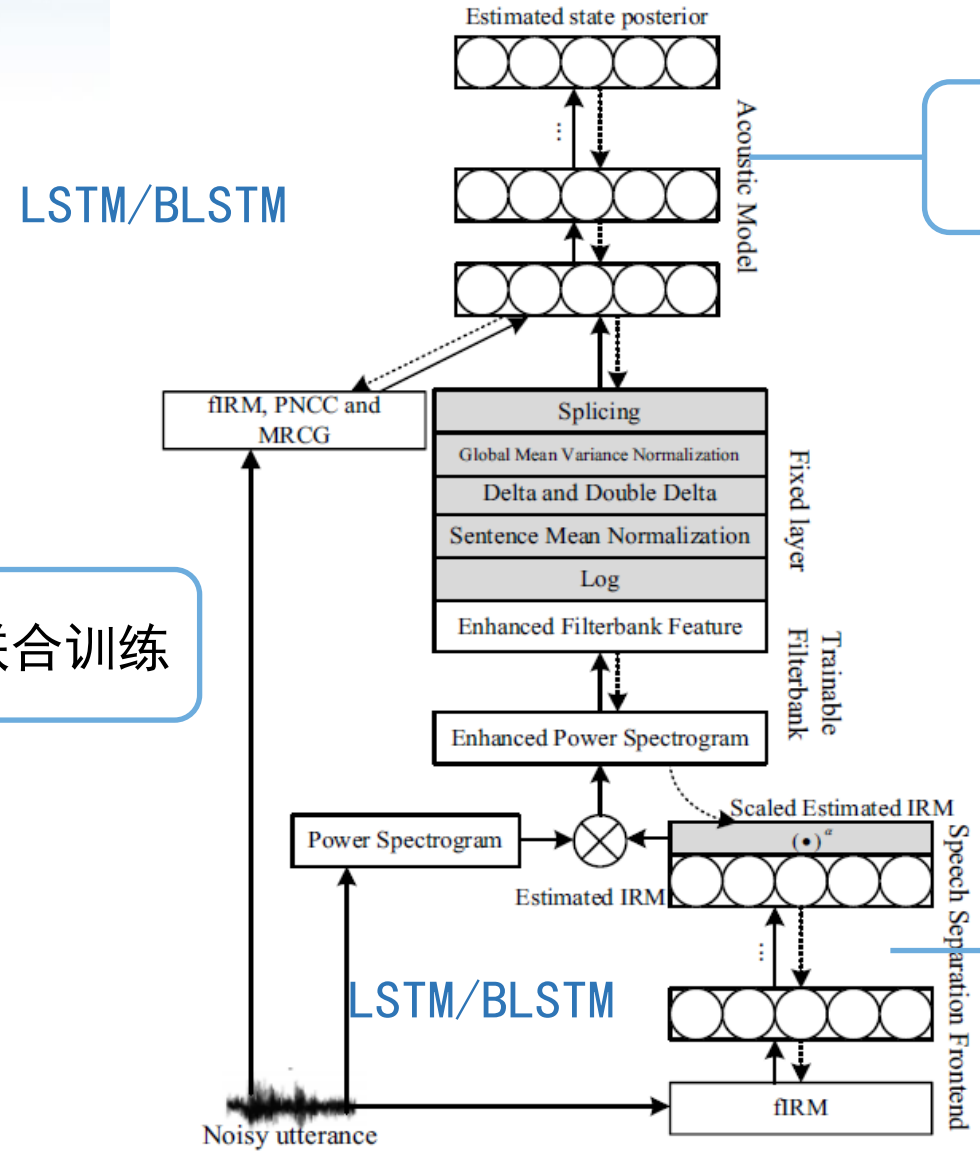
右下：深度学习方法



分析

基于DNN的语音增强方法可以有效的抑制非平稳噪声

基于RNN的语音增强



语音识别声学模型

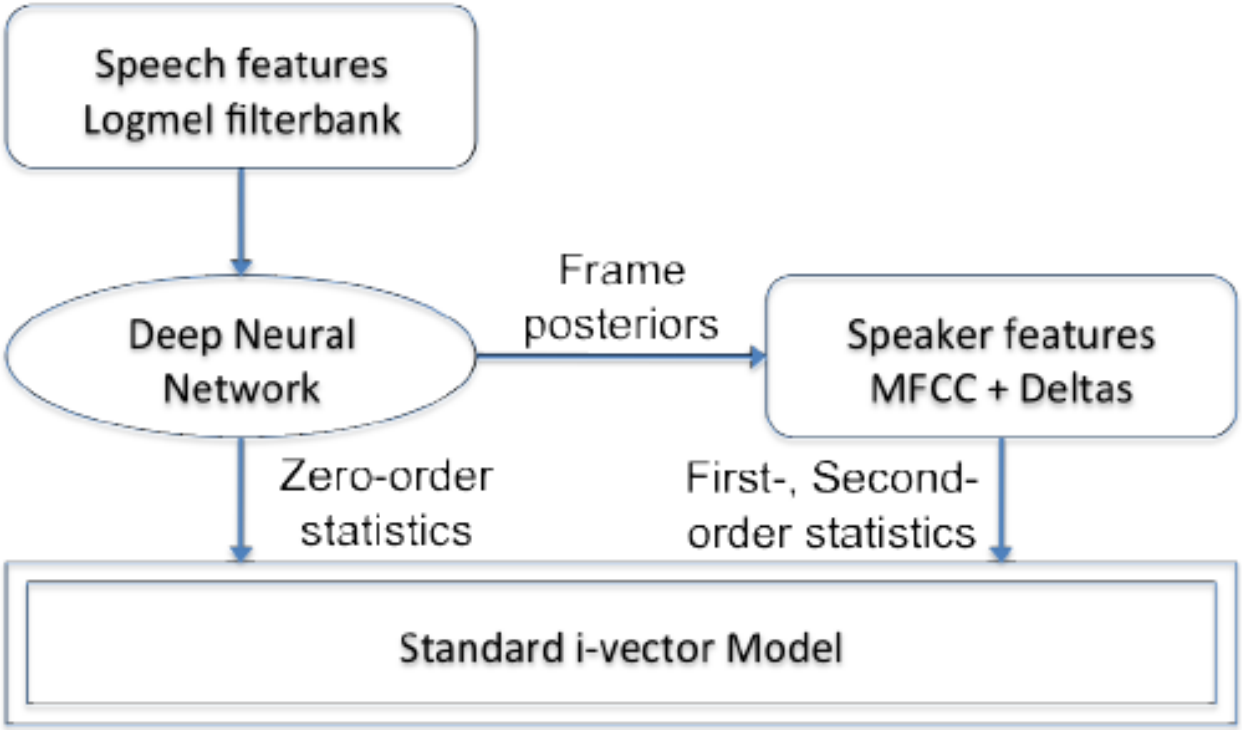
识别和增强联合训练

语音增强

内容

- 神经网络发展历程
- 基于深度神经网络的语音识别
- 基于深度神经网络的语音合成
- 基于深度神经网络的语音增强
- **基于深度神经网络的说话人识别**
- 基于深度神经网络的语种识别
- 展望

基于深度神经网络的说话人识别



b. NIST SRE'12 C5 extended condition - female

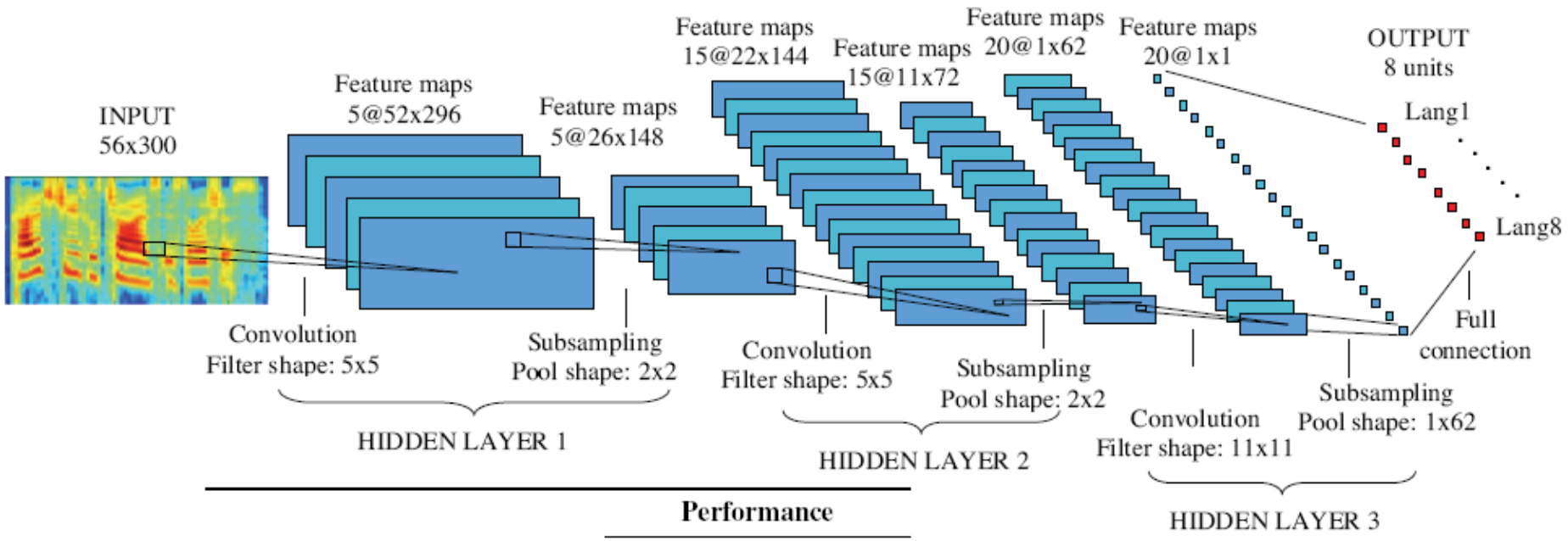
System	$P_{0.001}^{tar}$	$P_{0.01}^{tar}$	EER(%)	FA@M10
UBM-EM(2048)	0.421	0.252	2.84	0.36
UBM-EM(4096)	0.401	0.237	2.55	0.26
UBM-sup(3450)	0.451	0.272	2.94	0.44
DNN	0.291	0.177	1.92	0.10

内容

- 神经网络发展历程
- 基于深度神经网络的语音识别
- 基于深度神经网络的语音合成
- 基于深度神经网络的语音增强
- 基于深度神经网络的说话人识别
- **基于深度神经网络的语种识别**
- 展望

基于深度神经网络的语种识别

基于CNN的端到端语种识别



ID	Size	Performance	
		EER_{avg} (%)	C_{avg}
i-vector	~23M	**16.94**	**0.1535**
ConvNet 1	~198k	22.14	0.2406
ConvNet 2	~39k	25.90	0.2700
ConvNet 3	~39k	24.69	0.2616
ConvNet 4	~39k	23.48	0.2461
ConvNet 5	~78k	21.60	0.2282
ConvNet 6	~78k	21.11	0.2293
AllConvNets	-	17.93	**0.1836**
ConvNet 6+i-vector	-	**15.96**	**0.1433**
AllConvNets+i-vector	-	**15.04**	**0.1360**

内容

- 神经网络发展历程
- 基于深度神经网络的语音识别
- 基于深度神经网络的语音合成
- 基于深度神经网络的语音增强
- 基于深度神经网络的说话人识别
- 基于深度神经网络的语种识别
- 展望

展望

- 语音识别
  - 简化训练流程
  - 将大数据模型知识迁移到小数据模型上
- 语音合成
  - 小说、口语语音合成
  - 多风格，富有情感的语音合成
- 语音增强
  - 真实环境中人对语音生成、听觉感知和认知的机理
  - 实现真正的人-人、人-机无障碍交流与通信

参考文献

- G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. NSainath, et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” Signal Processing Magazine, IEEE, vol. 29, no. 6, pp. 82–97, 2012.
- Y. LeCun and Y. Bengio, “Convolutional networks for images, speech,and time series”, The handbook of brain theory and neural networks,vol. 3361, pp. 1995, 1995.
- Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard,W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwrittenzip code recognition”, Neural computation, vol. 1, pp. 541–551,1989.
- T. Robinson and F. Fallside, “A recurrent error propagation networkspeech recognition system”, Computer Speech & Language, vol. 5, pp.259–274, 1991.
- S. Hochreiter and J. Schmidhuber, “Long short-term memory”, euralcomputation, vol. 9, pp. 1735–1780, 1997.
- H. Sak, A. W. Senior, and F. Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling.”,in Interspeech, pp. 338–342, 2014.
- G. Saon, T. Sercu, S. J. Rennie, and H. J. Kuo, “The IBM 2016 English conversational telephone speech recognition system”, CoRR, vol.abs/1604.08242, 2016.

参考文献

- Zen H, Agiomyrg iannakisY, Egberts N, et al. Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices[J]. arXivpreprint arXiv:1606.06061, 2016.
- MoriseM. Cheaptrick, a spectral envelope estimator for high-quality speech synthesis[J]. Speech Communication, 2015, 67: 1-7.
- Hu Y J, Ling Z H. DBN-based Spectral Feature Representation for Statistical Parametric Speech Synthesis[J]. IEEE Signal Processing Letters, 2016, 23(3): 321-325.
- TakakiS, Yamagishi J. A DEEP AUTO-ENCODER BASED LOW-DIMENSIONAL FEATURE EXTRACTION FROM FFT SPECTRAL ENVELOPES FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.Speech enhancement theory and practice, Philipos C. Loizou，2007 .
- Microphone Arrays: Signal Processing Techniques and Applications (Digital Signal Processing) by Michael Brandstein, Darren Ward, Springer, 2001.
- Ding C, XieL, Yan J, et al. Automatic prosody prediction for Chinese speech synthesis using BLSTM-RNN and embedding features[C]//2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015: 98-102.

参考文献

- Wang W, Xu S, Xu B. First Step Towards End-to-End Parametric TTS Synthesis: Generating Spectral Parameters with Neural Attention[C]//Proceedings Interspeech. 2016: 2243-2247.
- Merritt T, Clark R A J, Wu Z, et al. Deep neural network-guided unit selection synthesis[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 5145-5149.Speech enhancement theory and practice, Philipos C. Loizou，2007 .
- Microphone Arrays: Signal Processing Techniques and Applications (Digital Signal Processing) by Michael Brandstein, Darren Ward, Springer, 2001.

谢谢！