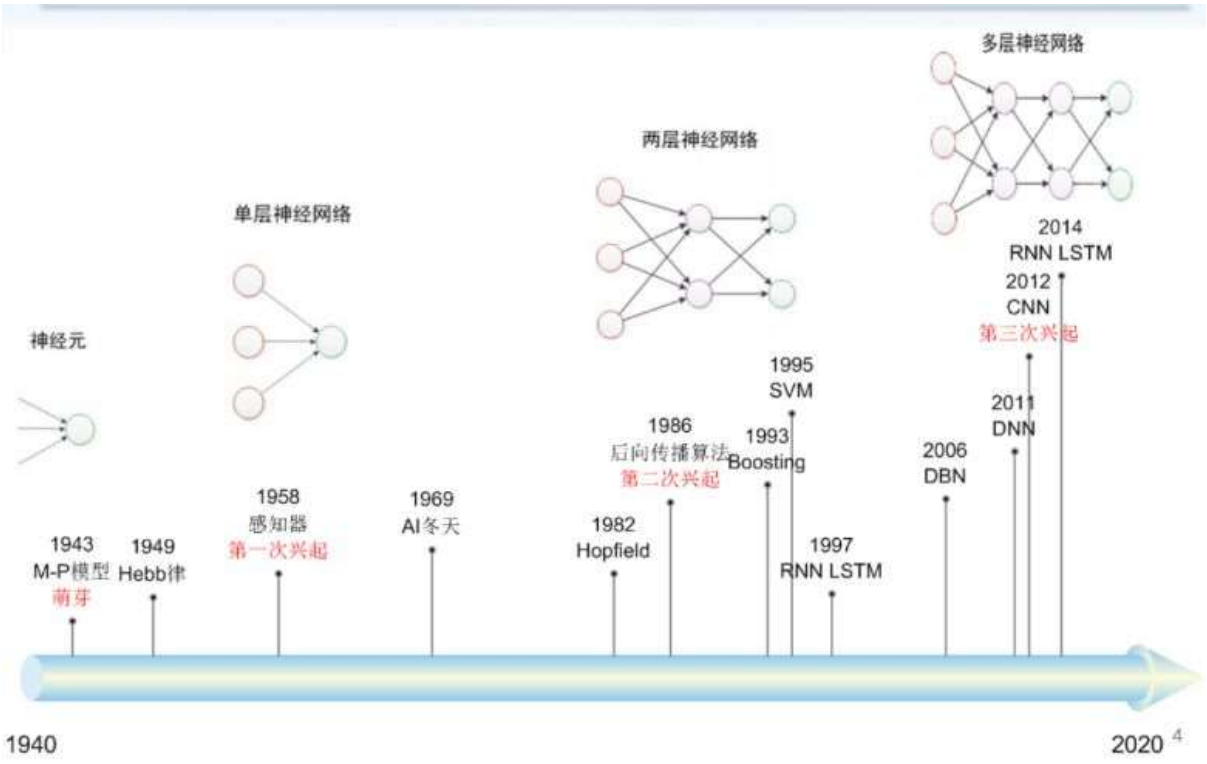
【AIDL专栏】陶建华：深度神经网络与语音

- 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

一、神经网络发展历程



深度神经网络是神经网络发展的复兴。在80-90年代，人工神经网络有很多热点，但当时学界认为，语音方面采用纯机器学习方法无法达到很好的性能，因为浅层神经网络能达到的性能非常清楚，数据规模也不乐观。97年IBM推出的第一个商用语音识别输入系统ViaVoice，训练量仅为1000个小时，可见当时能够处理的数据量和机器学习方法能达到的性能是有限的。

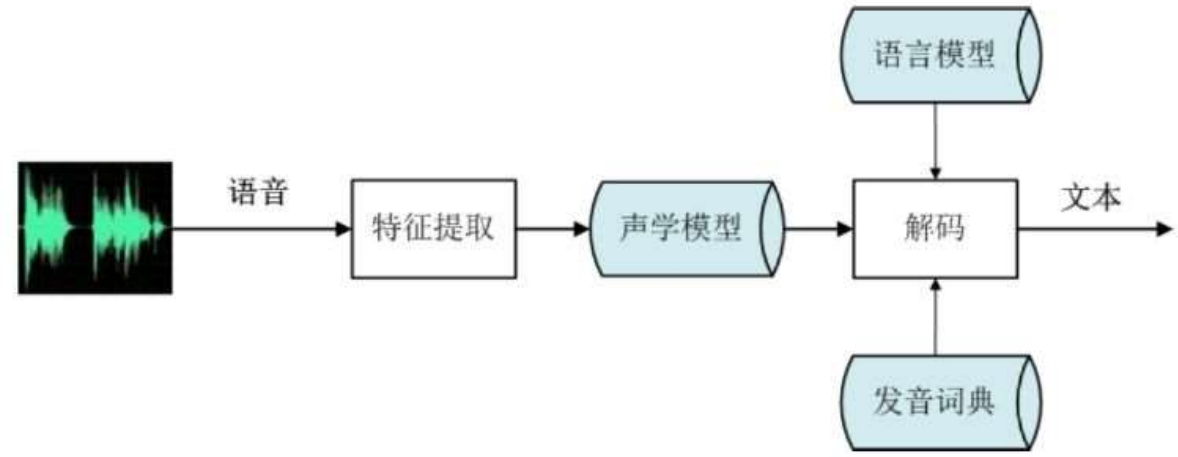
90年代到2000年初，神经网络进入冬天。Hinton在2006年提出的DBN（Deep Belief Network）成为新的发展和契机。深度神经网络在语音方面真正被重视是在2011年，微软学者俞栋实现用深度神经网络做语音识别的工作。这是深度神经网络第一个成功的应用，使用不同类型的训练集和测试集，识别词错误率均稳定相对降低了20%-30%，引起巨大轰动。谷歌随即也开展了这一工作。

此后就是卷积神经网络（CNN）。深度CNN凭借在ImageNet上图片分类任务的出色表现为人熟知，因此很多人认为2012年是深度神经网络的爆发性元年，实际上对语音来说2011年更为重要。深度机器学习方法在语音方面的应用经历了不同的发展阶段，每个阶段有不同的变体，不同的模

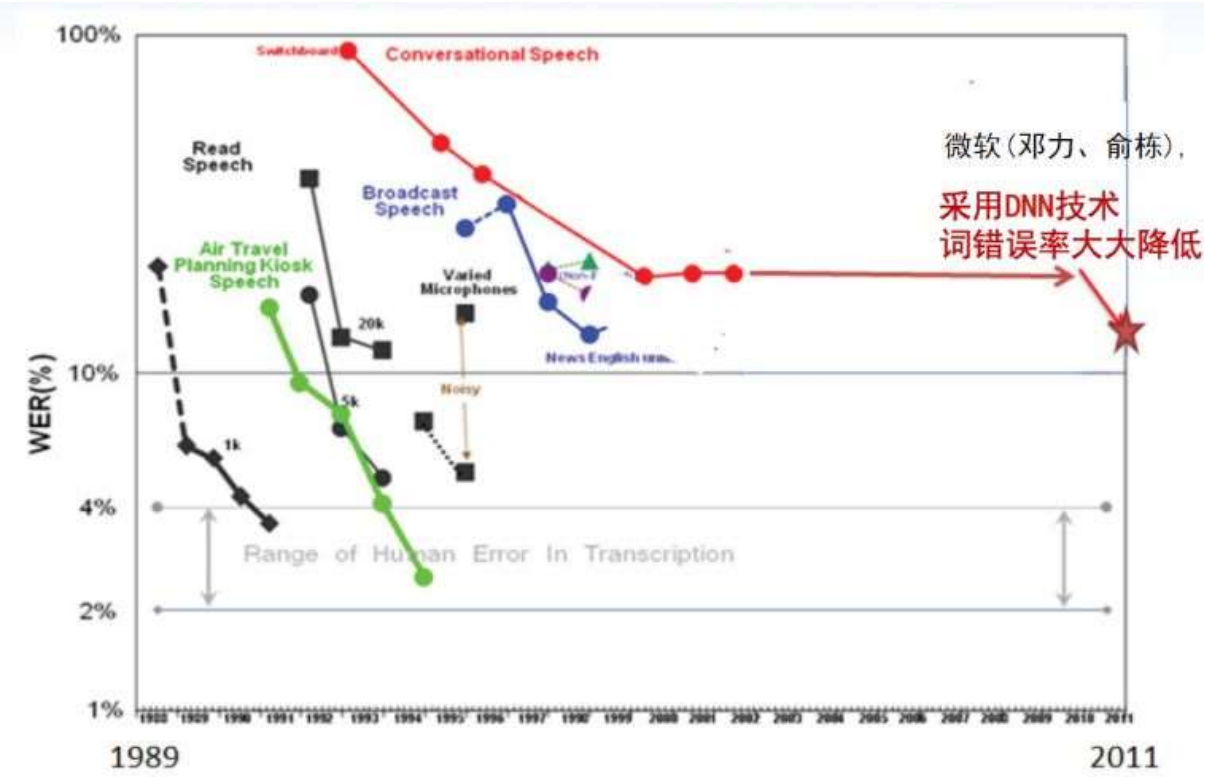
型甚至出现了组合应用，对整个语音的性能，包括识别、合成、增强等产生了很多影响。下面针对这几个不同领域稍作回顾。

二、基于深度神经网络的语音识别

2.1 语音识别回顾



语音识别首先对语音进行特征提取，一般提取频谱值。语音中频谱值通常为梅尔倒谱参数，如MFCC、频带参数、子带参数等，之后送入声学模型识别出它大体是什么音素或音节，但是由于同音不同字，故无法识别出它对应文字，所以引入语言模型，构建一个解码搜索空间。根据利用大量文本数据训练的语言模型，可以将声学模型的音素或者音节有效地转换为文字，有效提高识别正确率。这里主要讲深度神经网络在声学模型中的贡献，先看语音识别这几年的词错误率。

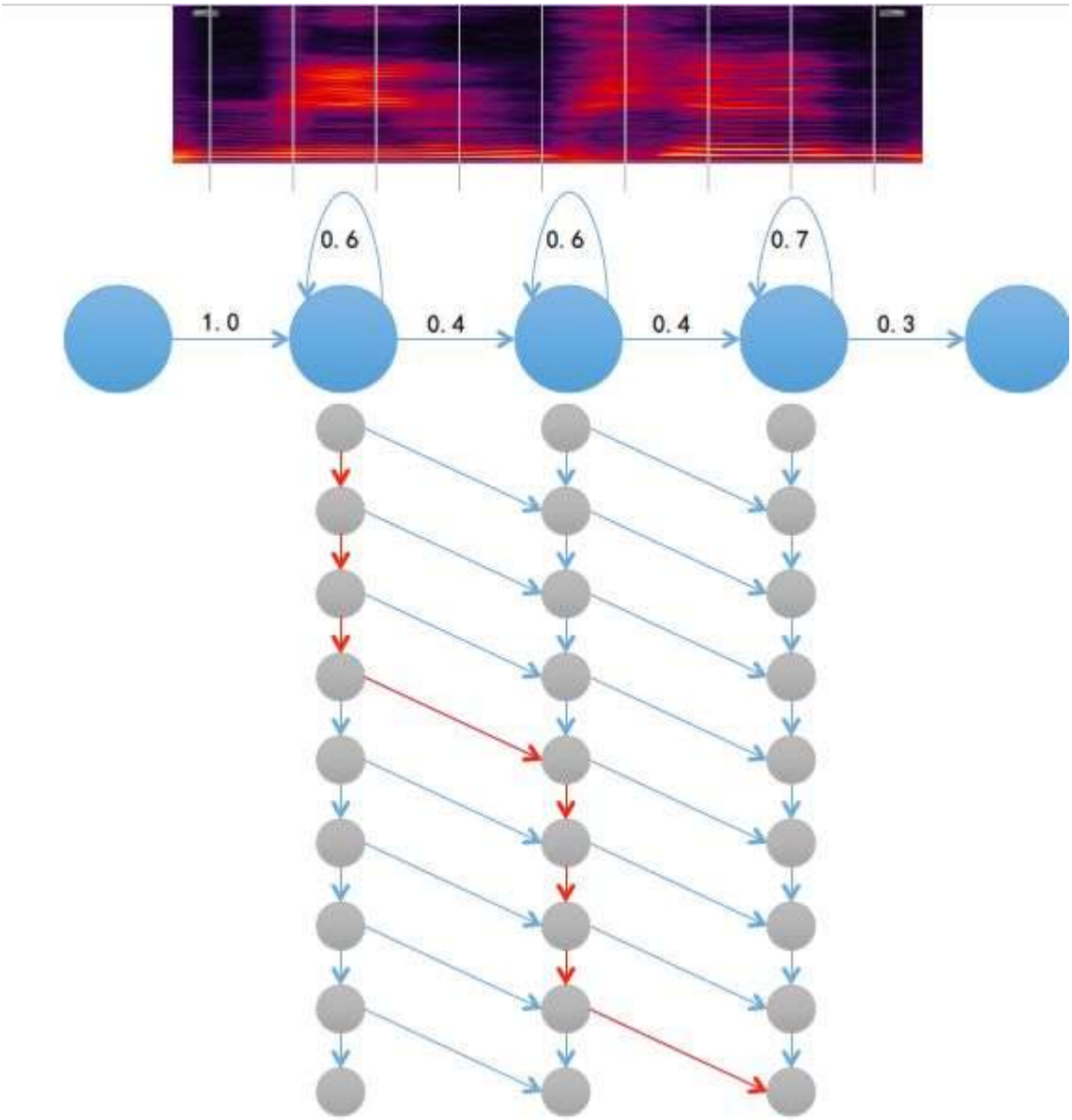


可以看到错误率每个阶段会出现一个大下降。在上世纪90年代末到2000年初，朗读语音识别错误率已有很大下降，但对话语音（Conversational Speech）识别错误率没有大变化，直到深度学习出来后大幅降低。深度学习和语音的结合不是一步到位，首先介绍传统语音识别中的混合高斯-隐马尔科夫模型（GMM-HMM）

2.2 声学模型

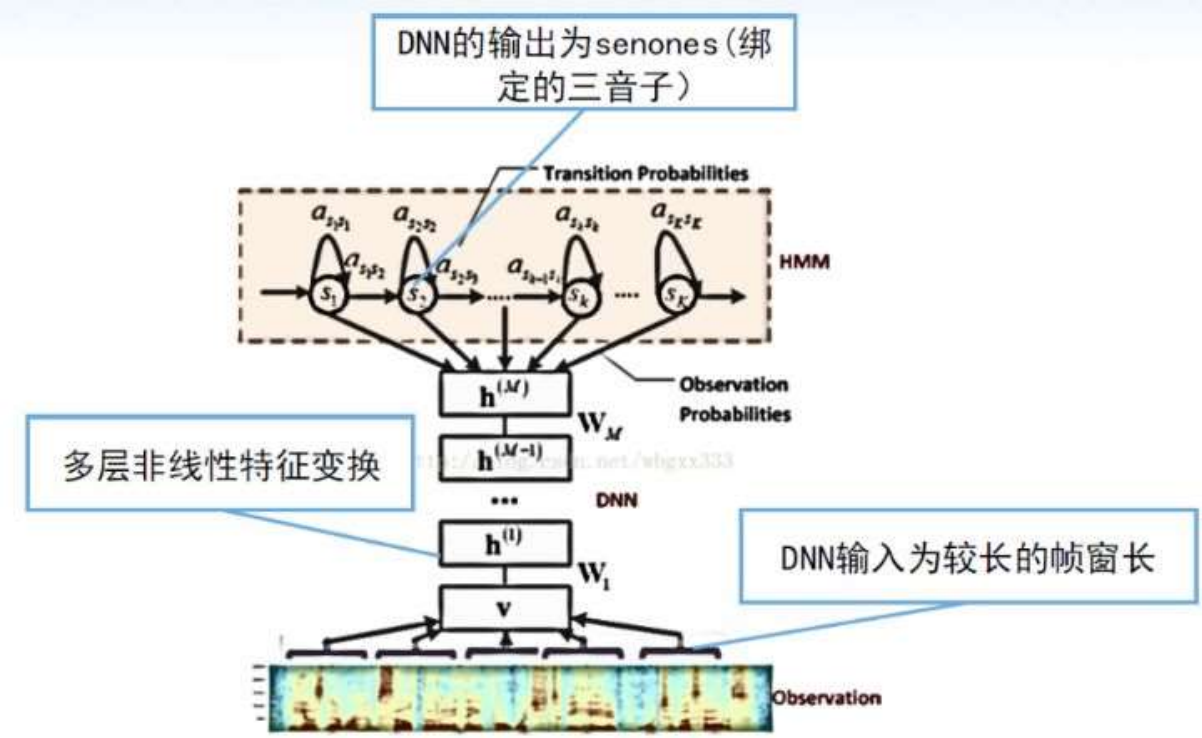
①.混合声学模型

高斯混合模型-隐马尔科夫模型（GMM-HMM）：隐马尔科夫模型（HMM）的参数主要包括状态间的转移概率以及每个状态的概率密度函数，也叫出现概率，一般用高斯混合模型（GMM）表示。



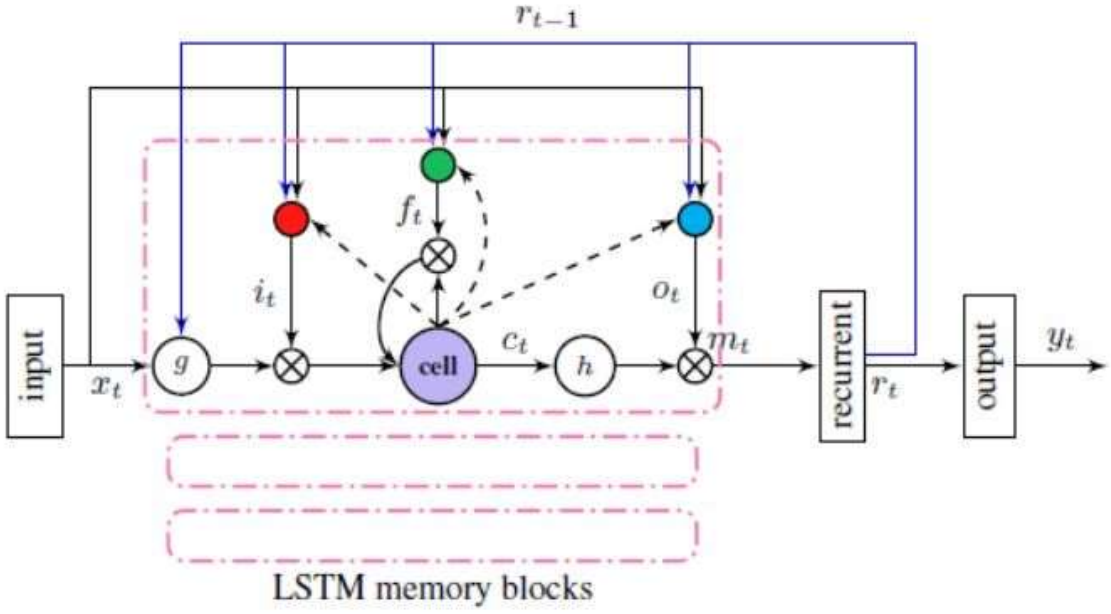
最上方为输入语音的语谱图，将语音第一帧代入一个状态进行计算，得到出现概率，同样方法计算每一帧的出现概率，图中用灰色点表示。灰色点间有转移概率，据此可计算最优路径（图中红色箭头），该路径对应的概率值总和即为输入语音经HMM得到的概率值。如果为每一个音节训练一个HMM，语音只需要代入每个音节的模型中算一遍，哪个得到的概率最高即判定为相应音节，这也是传统语音识别的方法。出现概率采用GMM，具有训练速度快、模型小、易于移植到嵌入式平台等优点，但缺点是没有利用帧的上下文信息，缺乏深层非线性特征变化的内容。GMM代表的是一种概率密度，它的局限在于不能完整模拟出或记住相同音的不同人间的音色差异变化或发音习惯的变化。

深度神经网络-隐马尔科夫模型（RNN-HMM）：俞栋将GMM替换为深层神经网络DNN，性能提升很大，这是深度神经网络在语音方面真正意义上的第一个应用。



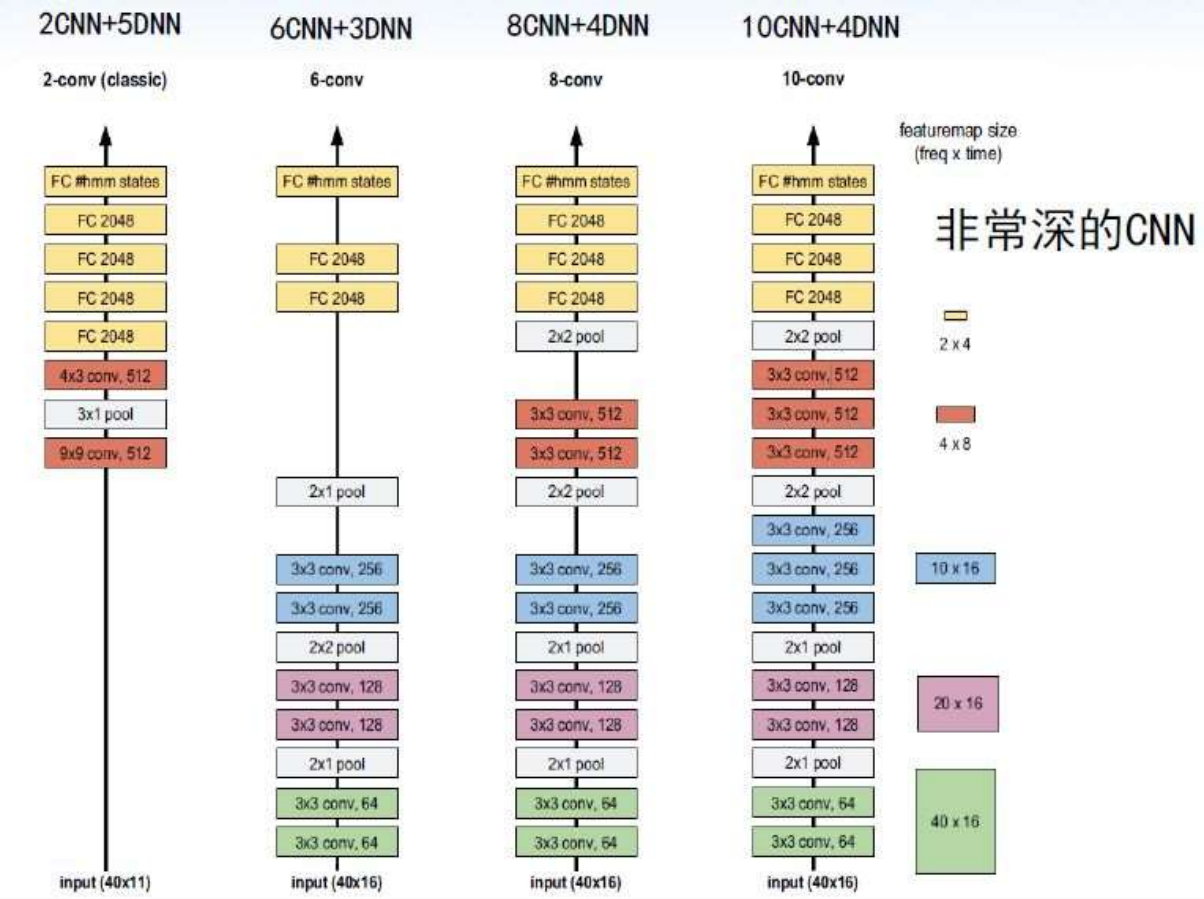
深度神经网络通常与HMM绑定，深度神经网络的输出标签通过HMM得到，一般为senones（绑定三音子），训练数据由GMM-HMM模型进行帧对齐得到，即给每帧打上标签（senones），训练准则为交叉熵（CE）准则。深度神经网络能利用帧的上下文信息，并能学习深层非线性特征变换，缺点是不能利用历史信息来辅助当前任务。

深度循环神经网络-隐马尔科夫模型（RNN-HMM）：如上所述，单纯用深层神经网络没有历史信息来辅助当前任务，所以提出RNN-HMM模型。RNN增加了时间的反馈，能够有效的利用时间的历史信息，将历史信息持久化，在很多任务上性能表现优于DNN，尤其是针对语音，因为语音具有很强的时序概念。但RNN随层数的增加会出现梯度爆炸或梯度消失，难以训练。也有人将长短时记忆模型LSTM融入RNN，LSTM采用一些控制门（输入门、遗忘门和输出门）来减少梯度累计的长度，一定程度上解决了RNN训练时的梯度消失和扩散的问题。RNN是迄今为止语音识别最好的模型之一，它的变体LSTM和BLSTM的性能都是相对最好的。



深度卷积神经网络-隐马尔科夫模型（CNN-HMM）：卷积神经网络CNN把视觉空间信息逐层抽象化，在视觉处理中起了很大作用，下面讨论CNN为何能应用到语音处理。语音的时序状态的波形

可以转化为频谱，语音和语谱图可以一一对应，即能从图像上“看懂”语音。尽管每个人之间说话具有差异性，有不同的口音，但从语谱图上能够反映相似性，所以引入CNN成为可行的方式。



CNN+LSTM+DNN达到了迄今为止已知报道中的最好识别结果。语音谱图本身隐含时间特性，是时间延迟的图像，所以应用于语音谱图的CNN也叫时间延迟的卷积神经网络，它能很好地对信号进行描述学习，也比其他深度神经网络更能捕获到特征的不变性。谷歌、微软、IBM均在2016年发表成果证明非常深的CNN声学模型已超越其它深度神经网络声学模型。

目前提到的深度神经网络依然和HMM进行了结合，仅仅把GMM替代了。它的训练过程还要依赖于传统的GMM-HMM的强制对齐信息，即利用GMM-HMM对每帧语音打一个标签，再利用这种有标签的数据训练深度神经网络，但这种方式依然是目前性能最好的方法之一。

②.端到端声学模型

连接时序分类-长短时记忆模型（CTC-LSTM）：汉语有调音节约为1300个，为每个音节训练一个深度神经网络并不困难。但训练一句话时，需要找到这句话中每个音节发音的起始和终止位置，几万小时数据需要的人工标注量巨大。因此，2014年CTC训练准则引入深度神经网络，主要针对LSTM模型。

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

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

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

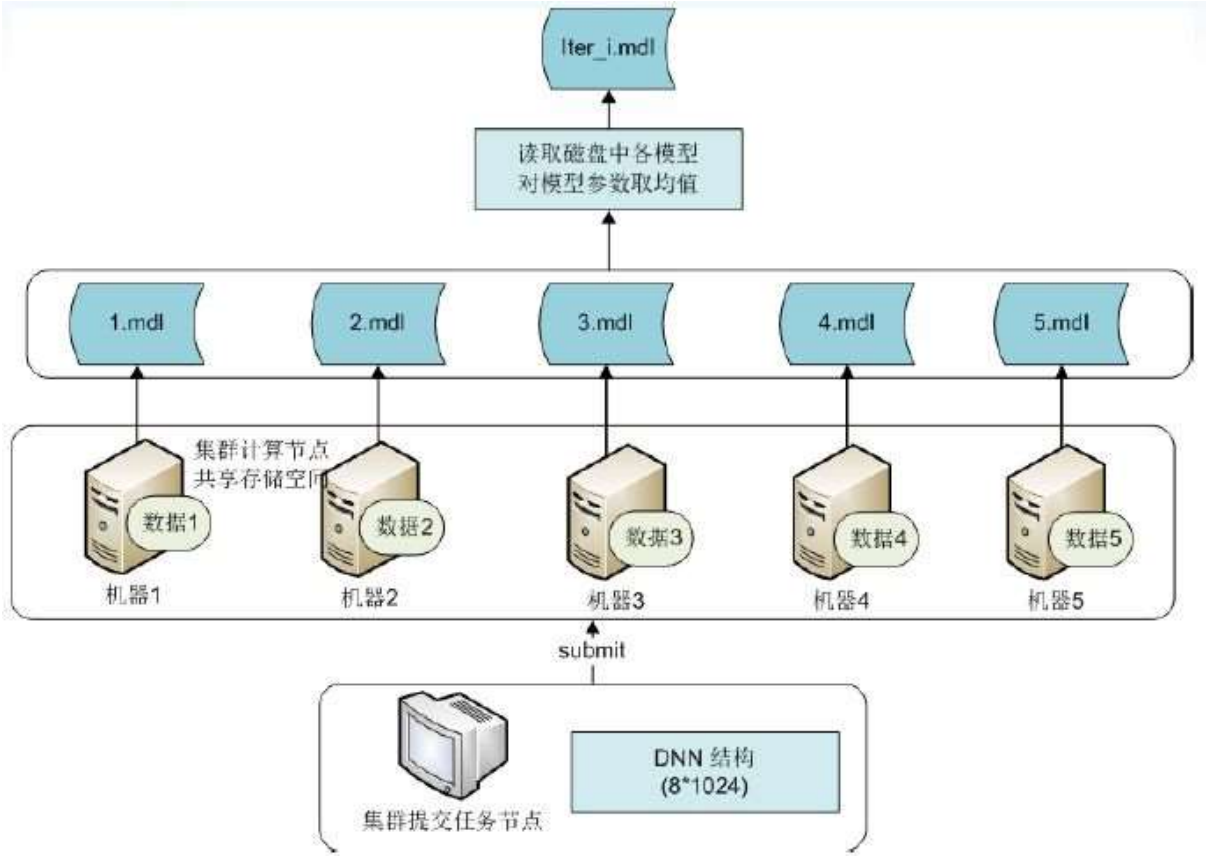
CTC准则只需要输入和输出在句子级别对齐，将句子中多个音节的神经网络串在一起，整句话直接送到这个深度神经网络组合中训练，算法能自动将每个音节与相关语音帧对齐，不需要先用GMM-HMM进行帧对齐，训练过程简洁。CTC还引入blank概念，即静音段或音节间的过渡段。约90%的帧对应输出为blank。对blank进行跳帧，可跳过绝大多数的静音段或过渡段，将有价值的声音堆积到某几帧语音上，加快解码速度。因解码速度快，识别性能较优，工业界大多采用CTC-LSTM模型。

注意力模型（Attention）：在基于RNN的Encoder-Decoder模型的基础上引入Attention机制，能从RNN的历史信息中挑选重要信息，抑制非重要信息。此方法接近人脑思维活动，思想有趣，模型优雅，训练步骤减少，但语音识别精度较低。

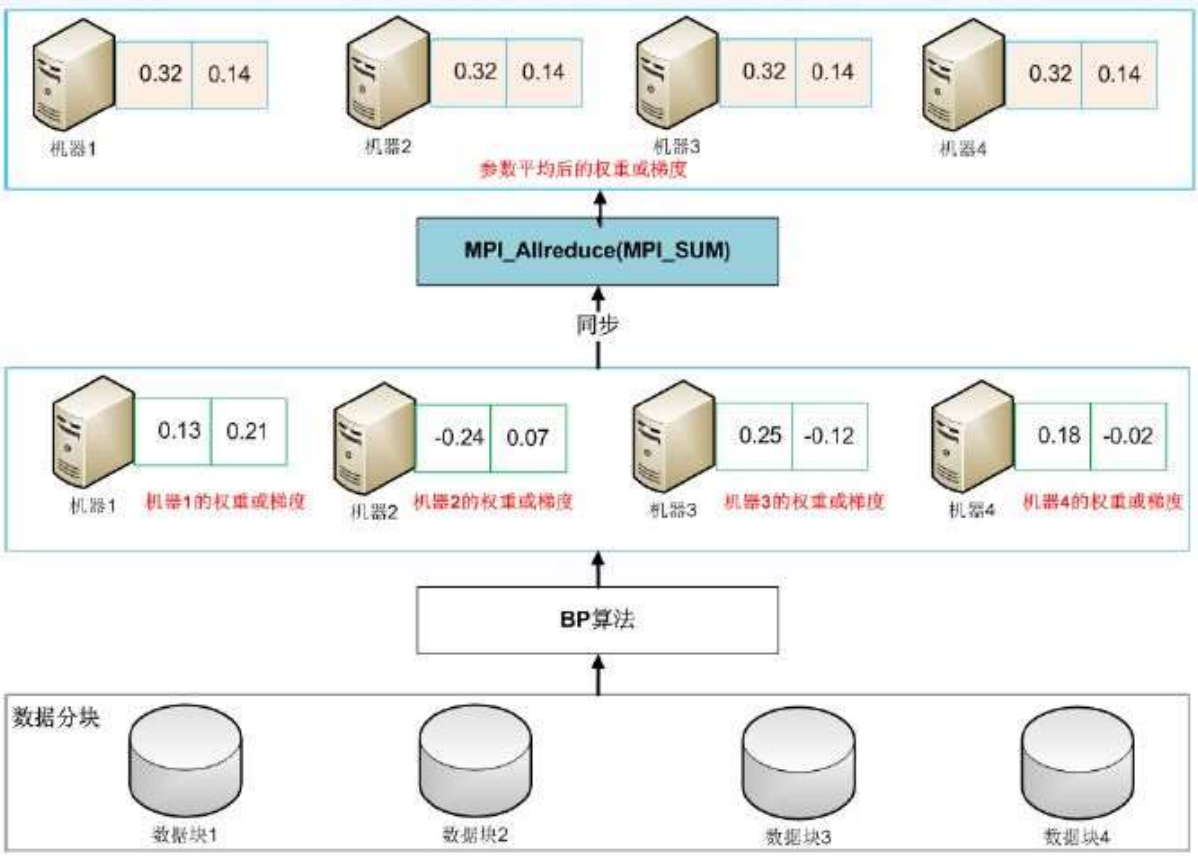
③.模型训练并行计算方法

深度神经网络训练时间长，需要并行计算加速。主要方法有多机共享存储、同步随机梯度下降和异步随机梯度下降。

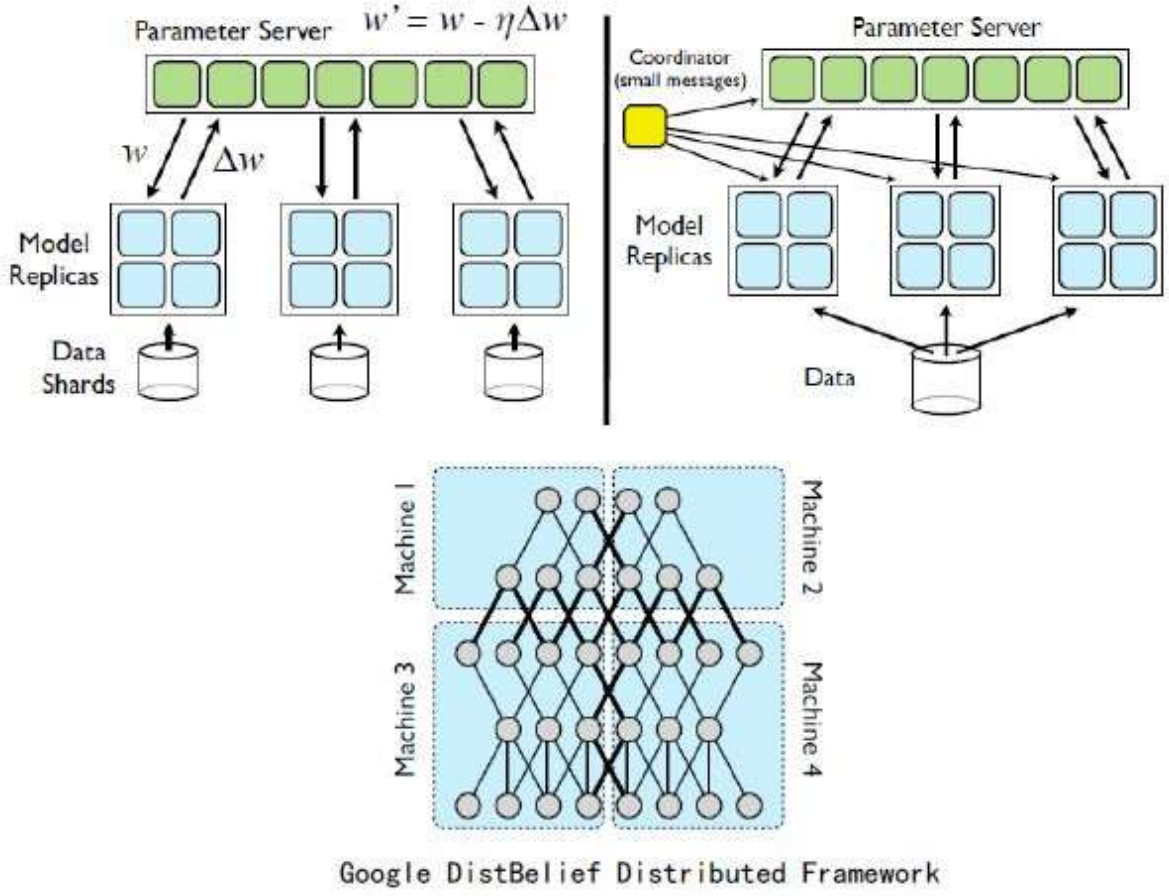
多机共享存储：每个机器单独训练模型，将训练参数取均值。优点是实现简单，各个模型互不干扰，某个机器模型训练失败只需单独重新训练，缺点是频繁存储和读取磁盘文件，占用大量I/O带宽，训练速度慢。



同步随机梯度下降（MPI）：多进程训练，每个进程训练结束后将参数在服务器上平均后对所有进程重新同步，其他机器在这个新的参数基础上训练。该方法实现难度适中，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。但一旦某台机器模型训练失败，所有机器需从头开始训练。

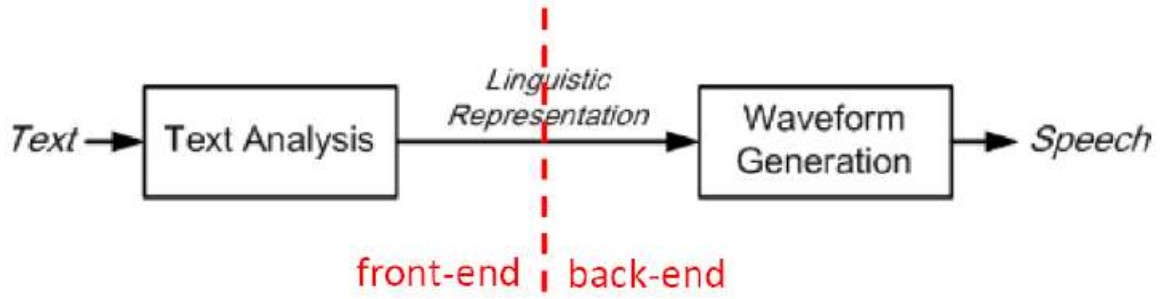


异步随机梯度下降（参数服务器）：多进程训练，模型可以分块更新。模型参数采用异步更新方式，某台机器训练失败不需要所有机器重新训练。该方法实现难度较大，需要过硬的软件和硬件资源，通常大企业才有实力开发，是现在工业界如谷歌、微软、阿里、百度等的主流训练方法。



三、基于深度神经网络的语音合成

3.1 语音合成简介



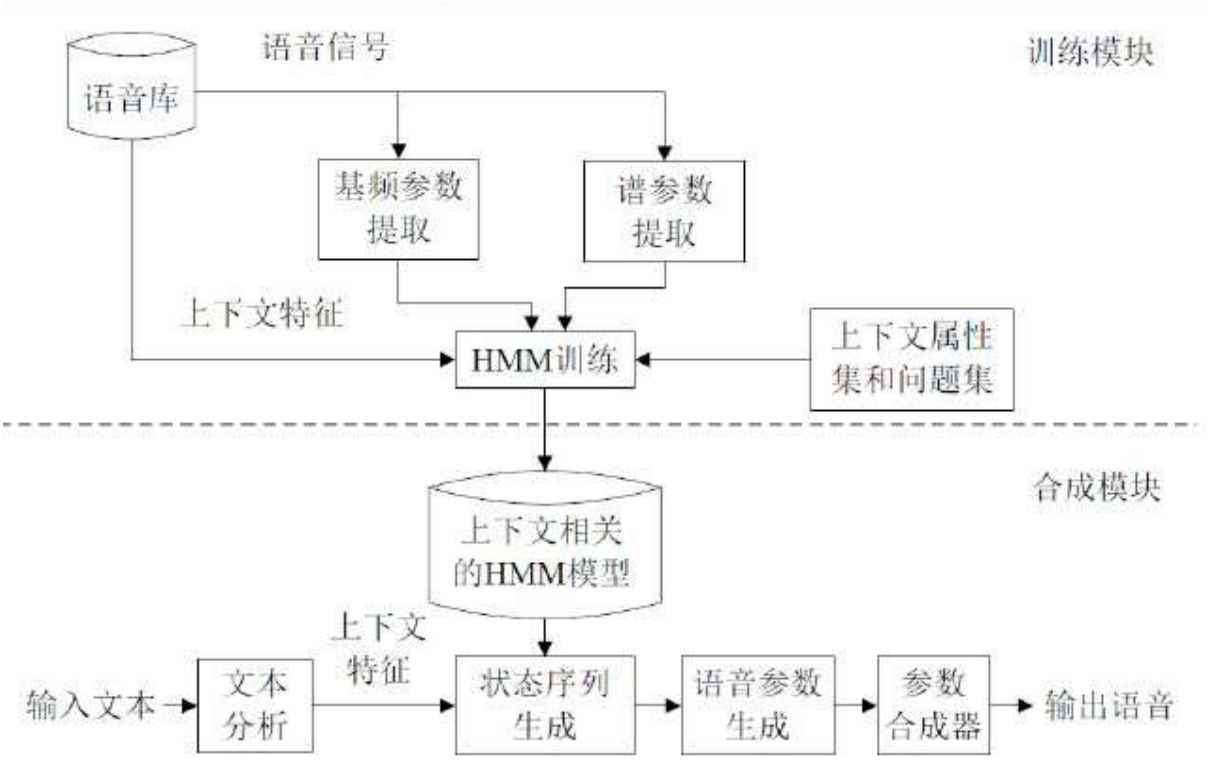
将文本送到计算机，经过一个模型得到参数，该参数经过声码器可产生声音，这就是语音合成的基本过程。

3.2 语音合成方法

①.基于波形拼接的方法

成句地录制各种声音，需要合成某句话中某个字的音时，可以根据这个字丰富的上下文信息，如在短语中的位置、词性等在录制的音库中找到相似的声音。用同样方式找到所有音节拼在一起形成整句话发音。此方法优点是声音清晰，缺点是声音之间可能不连贯，或找到的音仅是相似但不是最理想的，导致合成的声音忽高忽低。

②.基于HMM统计参数的语音合成



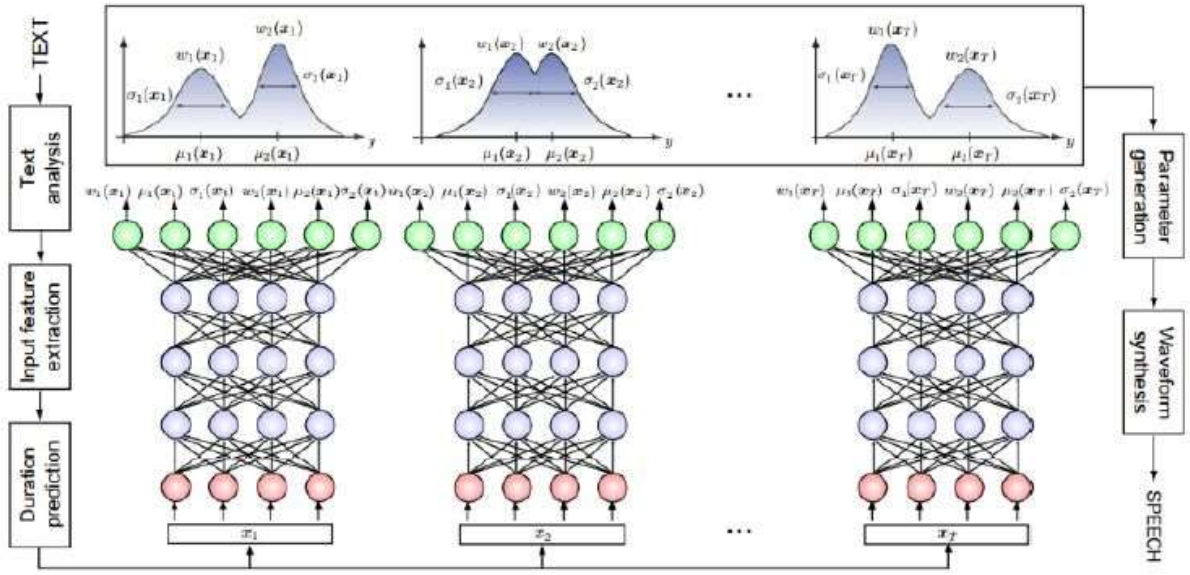
基于HMM统计参数的语音合成简单来说将HMM的高斯混合函数的均值作为语音参数来生成语音。此方法在2001年发表，但当时效果较差，经改善在2004-2005年的时候超过了拼接法。该方法优点是合成声音圆润，缺点是受限于声码器、HMM建模不准确，最重要的是生成参数不够平滑，不具有表现力，合成声音发闷。

③.基于深度神经网络的语音合成

类似语音识别，将高斯函数替换为深度神经网络，逐帧预测输出。特点是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）。下面介绍几个代表性的声学建模模型。

深度信念网络（DBN）：用DBN替换GMM，用DBN构建文本参数到语音参数之间的映射。

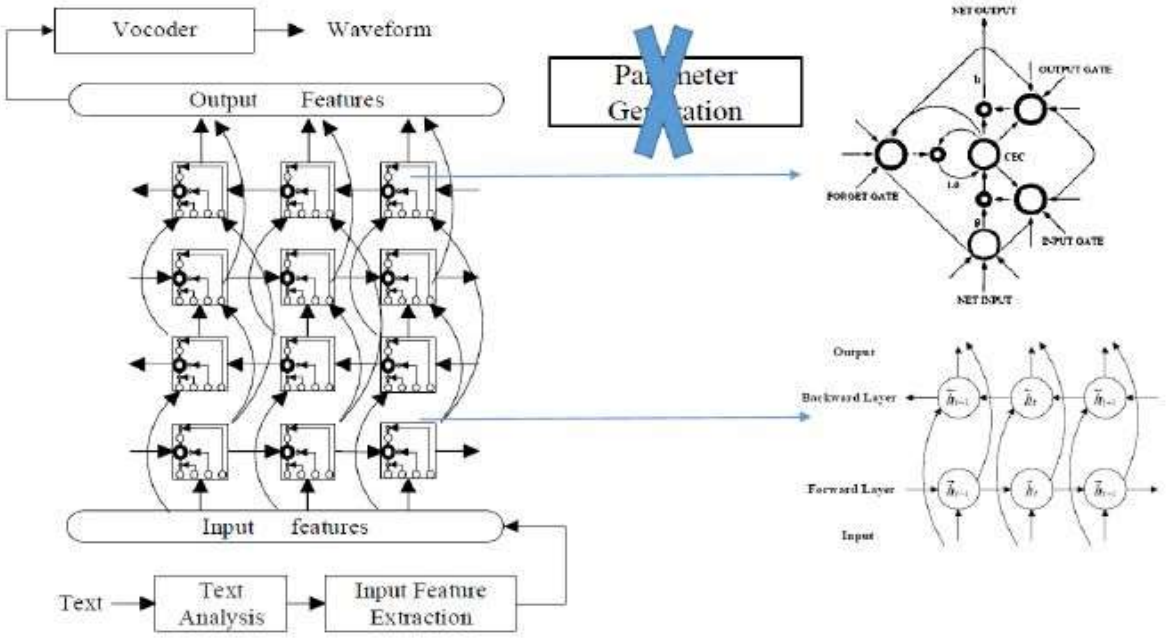
深度混合密度网络（DMDN）：如图，前面是DNN，但输出层的参数经过简单计算可以对应到高斯函数的参数（均值、方差）上，尤其是多高斯混合时每个高斯的权重都都可以通过输出层参数简单地公式映射得到，可以认为网络输出就是高斯混合函数的参数。



该方法的好处是，输入文本经过分析和预处理后送到DMDN模型中，输出不是语音生成参数而是高斯混合函数参数，在此基础上利用传统HMM思路生成语音，性能提高很多。

总体来说，语音生成是连续动态过程，需要考虑语义、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，所以语音合成时需要考虑文本的一系列上下文的内容以及声学层中历史信息的影响，所以DBLSTM-RNN相对效果更好。

深度双向长短时记忆网络（DBLSTM-RNN）：优势在于能完全做到端到端，较之前的网络跳过了参数生成算法。语音合成需要对文本做很多处理，如分析短语边界、词性、拼音等，通常使用贝叶斯决策、条件随机场、最大熵等方法，这些都可以用深度神经网络代替。声学层再做一个深度神经网络，将两个网络嫁接即可，不需要HMM，输入文本，经神经网络输出语音参数，再经声码器就可得到很好的声音。

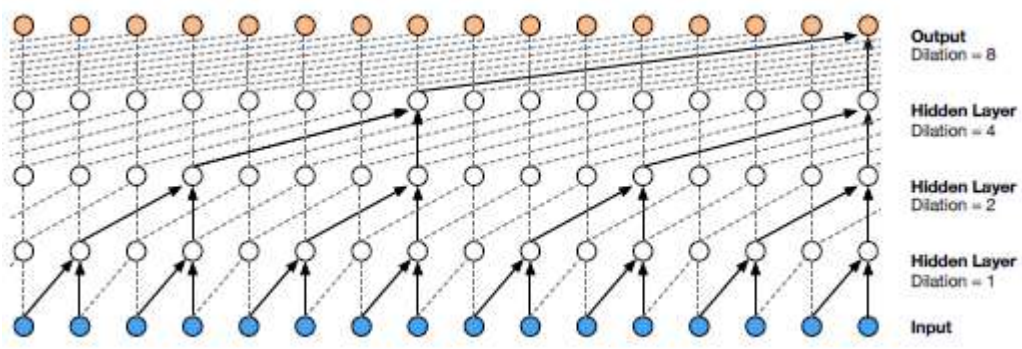


谷歌WaveNet：主要思想如下，语音的波形就是一个个采样点，每个采样点都受前面一系列采样点约束，存在条件概率密度函数，波形的联合概率可用条件概率分布的乘积来建模。

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

WaveNet将条件概率分布用多层卷积层建模，输出层不是普通意义上采样的语音波形，而是采用μ-

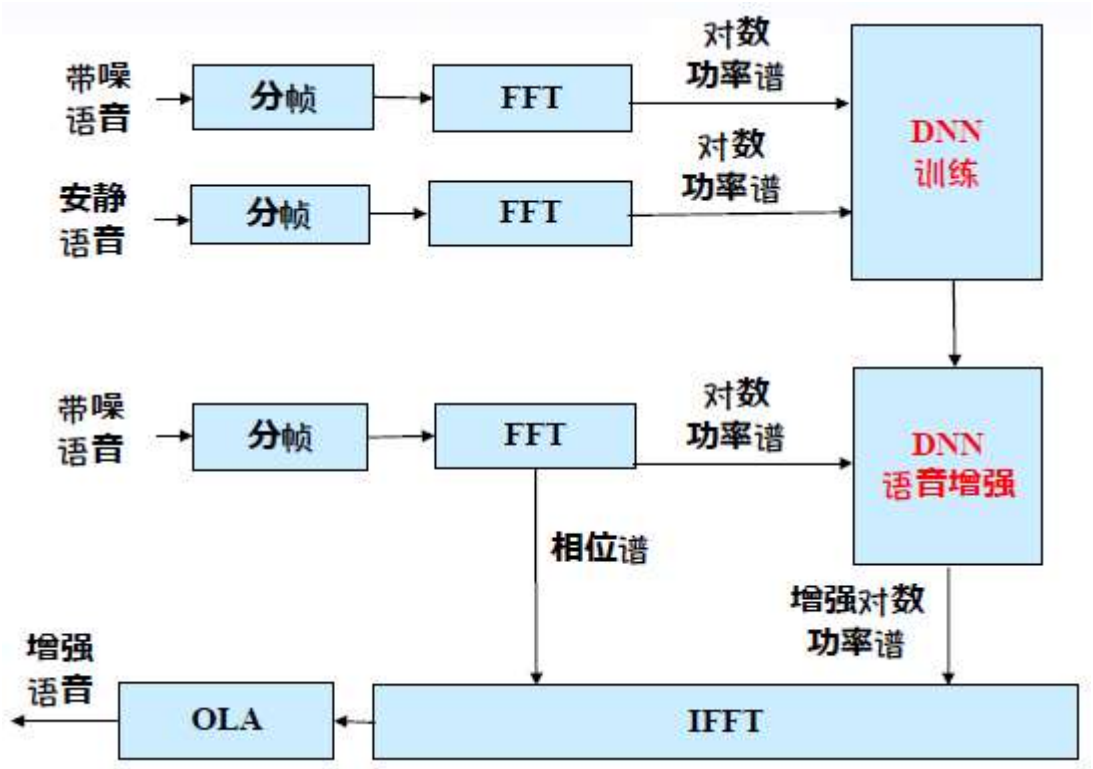
律压缩后的结果。训练的细节包括用残差反馈进行区分性训练，以及采用skip connections，跳跃某些时序特征的约束，增多训练层数，最后采用Conditional WaveNet激活函数将信息综合起来训练。WaveNet的结果超过之前所有系统。详细内容参考文献：Oord A V D, Dieleman S, Zen H, et al. WaveNet: A Generative Model for Raw Audio[J]. 2016.



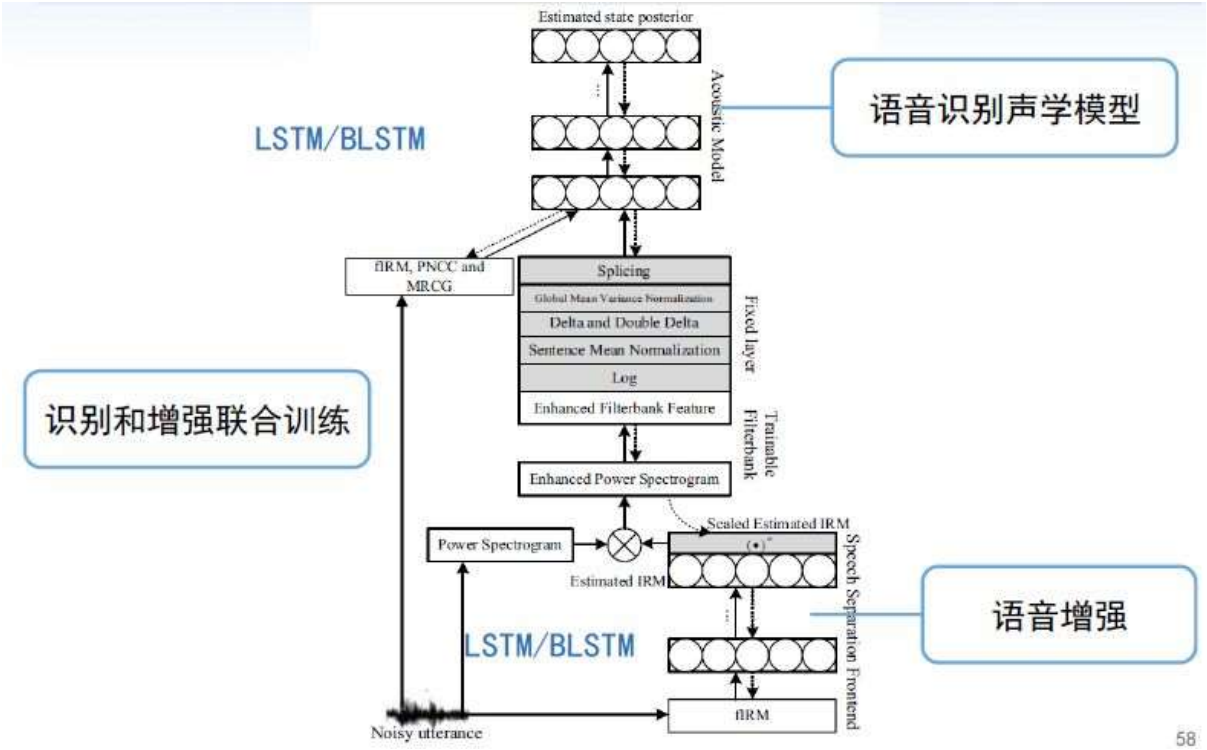
图为WaveNet采用的dilated causal convolutional layers

四、基于深度神经网络的语音增强

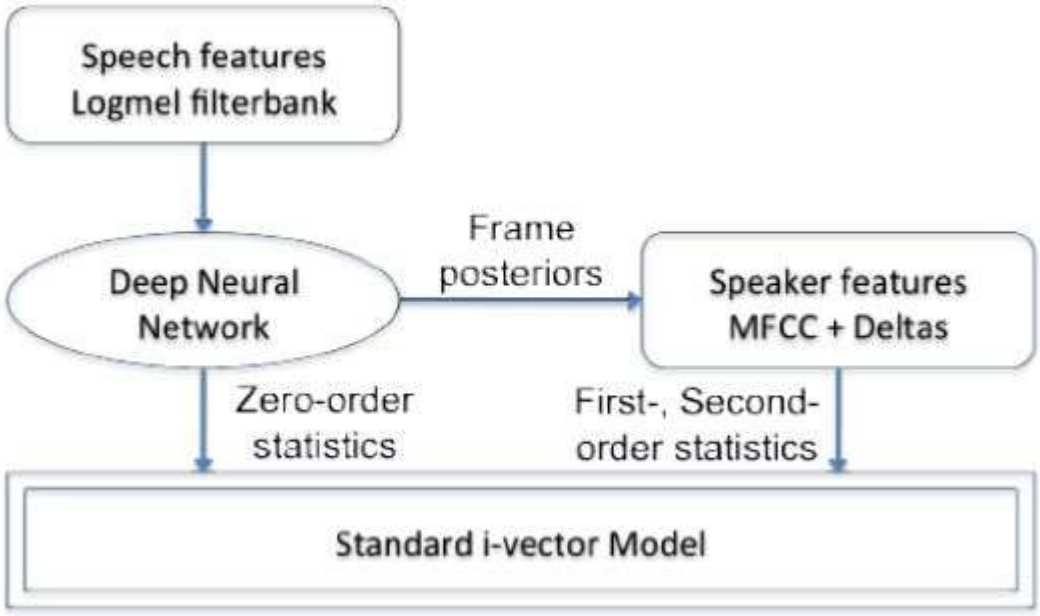
语音增强是指语音信号被不同噪声干扰甚至淹没后，从噪声背景中提取有用语音信号，抑制噪声干扰的技术。噪声类型包括混响、背景噪声、人声干扰、回声等。近两年深度神经网络被用于语音增强。将带噪语音输入，输出原干净声音，训练DNN，建立带噪语音与安静语音对数功率谱的映射关系，结果相比传统的子带谱减法、维纳滤波法、logmmse法等更能有效抑制非平稳噪声。



目前应用于语音识别的语音增强方法，一般分为两种：一是语音增强和识别单独训练，二是语音增强和识别联合训练。一般后者的性能优于前者。



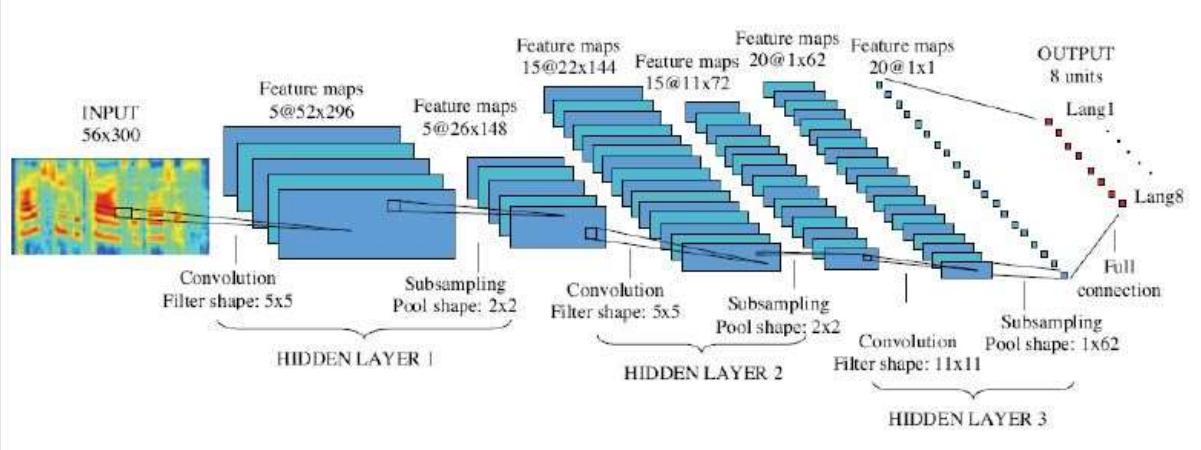
五、基于深度神经网络的说话人识别



说话人识别中的经典方法是I-Vector，I-Vector建模方式称为全局差异空间建模（Total Variability Modeling，TVM），采用该方法提取的I-Vector记为TVM-I-Vector。在基于TVM-I-Vector的声纹识别系统中，我们一般可以分为三个步骤。第一步是统计量的提取，第二步是提取I-Vector，第三步是进行信道补偿技术。统计量的提取是指将语音数据的特征序列，比如MFCC特征序列，用统计量来进行描述，提取的统计量属于高维特征，然后经过TVM建模，投影至低维空间中得到I-Vector。在TVM-I-Vector建模中，统计量的提取是以UBM为基础的，根据UBM的均值及方差进行相应统计量的计算。基于DNN的说话人识别的基本思想是取代TVM中的UBM产生帧级后验概率。即采用DNN进行帧级对齐的工作，继而计算训练数据的统计量，进行全局差异空间的训练以及I-Vector的提取。目前没有详细证据证明深度神经网络或组合i-vector的深度神经网络性能一定优于i-vector方法，可能原因是说话人识别中信道干扰较多，难以搜集足够数据训练深度神经网络。。

六、基于深度神经网络的语种识别

普遍认为将语音频谱变化为类似图像的方式的语谱图，采用CNN方法能得到最好的语种识别结果。



七、展望

语音识别方面，将大数据模型知识迁移到小数据上非常重要；语音合成方面，如何合成多风格、富有情感的语音和口语语音；语音增强方面，真实环境中人对语音生成、听觉感知和认知的机理，以及实现真正的人-人、人机无障碍交流通信，还有很多问题需要解决。我们需要在深度学习方法中持续努力，也需要探索新的方法，更多分析深层次的物理问题。

Q&A

问1：对于语音识别，深度模型相比传统模型占优势的数据拐点大约是多长时间的信号？

陶老师：语音具有一定规律性，通过“记忆”的方式能记住更多人发相同音时的频谱分布，因此深度学习方法提高语音识别性能并不奇怪。不是数据量增加才提高，而是早期在100甚至几十小时的数据上证明性能提高后才增大数据量，探索什么情况下性能提高更大。

问2：语音控制无人机在噪声环境下如何降噪？

陶老师：硬件降噪永远是最好的解决方案。软件降噪有两个思路，一种是采用报告中提到的降噪方法，另外一种是语音识别训练时用融合噪声的数据做训练，这样输入带噪声语音数据能达到相对较好性能，但不如降噪方法。

问3：深度学习在商业化中的应用发展远超过高校做研究的速度，高校研究人员如何弥补这一鸿沟？

陶老师：这是学术界公共问题。第一，高校研究人员需要考虑深度学习以外的问题，要对机器学习进行更深层次的分析，探索新的解决思路；第二，加强同企业的合作很必要。

陶建华老师的心得分享

1.深度学习有很多潜力未被发掘，但也要对深度机器学习方法以后面临的新挑战有充分的思想和预期。

2.深度学习依赖于大数据，但不能迷信大数据，要留心“大数据陷阱”。

3.深度机器学习随着开源工具的广泛使用，学习和应用门槛越来越低，这让人工智能更受重视，但也变得更廉价，如果学术界没有更多更新的创新，“冬天”就会到来。要思考深度学习之后，我们还要做什么新的内容。