语音识别技术的现状与未来

The Present and Future of Speech Recognition

（CSLT-TRP-20160034）

王东（Dong Wang）

2017/01/08

CSLT, RIIT, Tsinghua Univ.

语音识别任务及其研究意义

语音识别 (Automatic Speech Recognition, ASR)是指利用计算机实现从语音到文字自动转换的任务。在实际应用中，语音识别通常与自然语言理解、自然语言生成和语音合成等技术结合在一起，提供一个基于语音的自然流畅的人机交互方法。

早期的语音识别技术多基于信号处理和模式识别方法。随着技术的进步，机器学习方法越来越多地应用到语音识别研究中，特别是深度学习技术，它给语音识别研究带来了深刻变革。同时，语音识别通常需要集成语法和语义等高层知识来提高识别精度，因此和自然语言处理技术息息相关。另外，随着数据量的增加和机器计算能力的提高，语音识别越来越依赖数据资源和各种数据优化方法，这使得语音识别与大数据、高性能计算等新技术产生广泛结合。综上所述，语音识别是一门综合性应用技术，集成了包括信号处理、模式识别、机器学习、数值分析、自然语言处理、高性能计算等一系列基础学科的优秀成果，是一门跨领域、跨学科的应用型研究。

语音识别研究具有重要的科学价值和社会价值。语音信号是典型的局部稳态时间序列，研究这一信号的建模方法具有普遍意义。事实上，我们日常所见的大量信号都属于这种局部稳态信号，如视频、雷达信号、金融资产价格、经济数据等。这些信号的共同特点是在抽象的时间序列中包括大量不同层次的信息，因而可用相似的模型进行描述。历史上，语音信号的研究成果在若干领域起过重要的启发作用。例如，语音信号处理中的隐马尔可夫模型在金融分析、机械控制等领域都得到了广泛应用。近年来，深度神经网络在语音识别领域的巨大成功直接促进了各种深度学习模型在自然语言处理、图形图象处理、知识推理等众多应用领域的发展，取得了一个又一个令人惊叹的成果。

在实用价值方面，语音交互是未来人机交互的重要方式之一。随着移动电话、穿戴式设备、智能家电等可计算设备的普及，基于键盘、鼠标、触摸屏的传统交互方式变得越来越困难。为了解决这种困难，手势、脑波等一系统新的人机交互方式进入人们的视野。在这些五花八门的新兴交互方式中，语音交互具有自然、便捷、安全和稳定等特性，是最理想的交互方式。在语音交互技术中，语音识别是至关重要的一环：只有能“听懂”用户的输入，系统才能做出合理的反应。今天，语音识别技术已经广泛应用在移动设备、车载设备、机器人等场景，在搜索、操控、导航、休闲娱乐等众多领域发挥了越来越重要的作用。随着技术越来越成熟稳定，我们相信一个以语音作为主要交互方式的人机界面新时代将很快到来。

研究内容和关键科学问题

语音识别研究主要包括如下三方面内容：语音信号的表示，即特征抽取；语音信号和语言知识建模；基于模型的推理，即解码。语音信号的复杂性和多变性使得这三方面的研究都面临相当大的挑战。图 1 给出一个语音识别系统的典型架构。

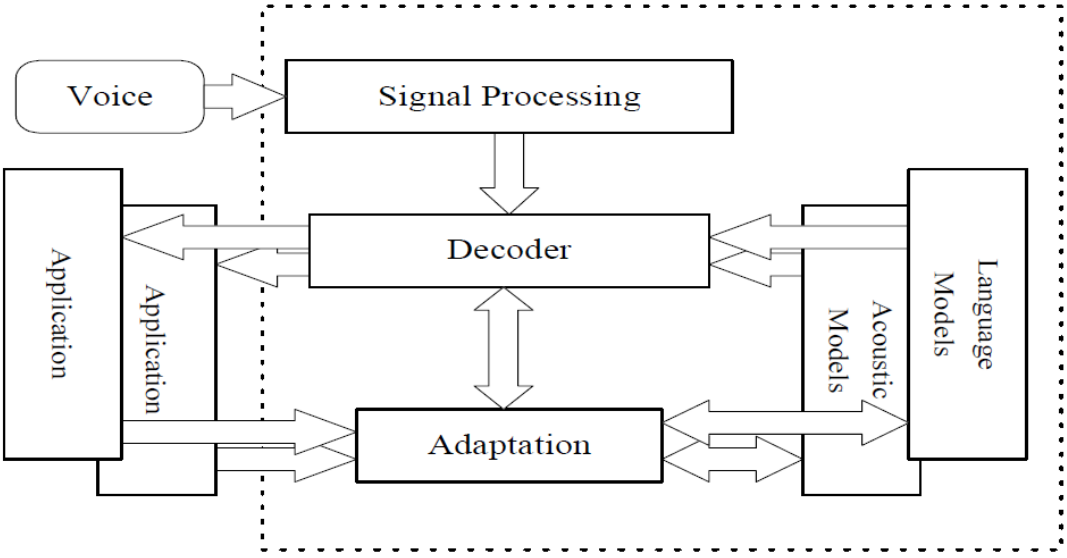

图 1. 语音识别系统结构。(Huang, X.,1996)

语音特征抽取

语音识别的一个主要困难在于语音信号的复杂性和多变性。一段看似简单的语音信号，其中包含了说话人、发音内容、信道特征、口音方言等大量信息。不仅如此，这些底层信息互相组合在一起，又表达了如情绪变化、语法语义、暗示内涵等丰富的高层信息。如此众多的信息中，仅有少量是和语音识别相关的，这些信息被淹没在大量其它信息中，因此充满了变动性。语音特征抽取即是在原始语音信号中提取出与语音识别最相关的信息，滤除其它无关信息。

语音特征抽取的原则是：尽量保留对发音内容的区分性，同时提高对其它信息变量的鲁棒性。历史上研究者通过各种物理学、生理学、心理学等模型构造出各种精巧的语音特征抽取方法，近年来的研究倾向于通过数据驱动学习适合某一应用场景的语音特征。

模型构建

语音识别中的建模包括声学建模和语言建模。声学建模是对声音信号（语音特征）的特性进行抽象化。自上世纪 70 年代中期以来，声学模型基本上以统计模型为主，特别是隐马尔可夫模型/高斯混合模型(HMM/GMM)结构。最近几年，深度神经网络(DNN)和各种异构神经网络成为声学模型的主流结构。

声学模型需要解决如下几个基本问题：（1）如何描述语音信号的短时平稳性；（2）如何描述语音信号在某一平稳瞬态的静态特性，即特征分布规律；（3）如何应用语法语义等高层信息；（4）如何对模型进行优化，即模型训练。同时，在实际应用中，还需要解决众多应用问题，如：（1）如何从一个领域快速自适应到另一个领域；（2）如何对噪音、信道等非语音内容进行补偿；（3）如何利用少量数据建模；（4）如何提高对语音内容的区分性；（5）如何利用半标注或无标注数据，等等。

语言建模是对语言中的词语搭配关系进行归纳，抽象成概率模型。这一模型在解码过程中对解码空间形成约束，不仅减小计算量，而且可以提高解码精度。传统语言模型多基于 N 元文

法 (n-gram)，近年来基于递归神经网络（RNN）的语言模型发展很快，在某些识别任务中取得了比 n-gram 模型更好的结果。

语言模型要解决的主要问题是如何对低频词进行平滑。不论是 n-gram 模型还是 RNN 模型，低频词很难积累足够的统计量，因而无法得到较好的概率估计。平滑方法借用高频词或相似词的统计量，提高对低频词概率估计的准确性。除此之外，语言建模研究还包括：（1）如何对字母、字、词、短语、主题等多层次语言单元进行多层次建模；（2）如何对应用领域进行快速自适应；（3）如何提高训练效率，特别是对神经网络模型来说，提高效率尤为重要；（4）如何有效利用大量噪声数据，等等。

解码

解码是利用语音模型和语言模型中积累的知识，对语音信号序列进行推理，从而得到相应语音内容的过程。早期的解码器一般为动态解码，即在开始解码前，将各种知识源以独立模块形式加载到内存中，动态构造解码图。现代语音识别系统多采用静态解码，即将各种知识源统一表达成有限状态转移机（FST），并将各层次的 FST 嵌套组合在一起，形成解码图。解码时，一般采用 Viterbi 算法在解码图中进行路径搜索。为加快搜索速度，一般对搜索路径进行剪枝，保留最有希望的路径，即 beam search。

对解码器的研究包括但不限于如下内容：（1）如何加快解码速度，特别是在应用神经网络语言模型进行一遍解码时；（2）如何实现静态解码图的动态更新，如加入新词；（2）如何利用高层语义信息；（3） 如何估计解码结果的信任度；（4）如何实现多语言和混合语言解码；（5）如何对多个解码器的解码结果进行融合。

技术方法和研究现状

语音识别研究可追溯到 20 世纪 50 年代，例如贝尔实验室的 AUDREY 系统，用模拟电路实现了对 10 个数字的识别。从那以后，语音识别技术经历了模式识别、统计模型、机器学习、深度学习等几个发展阶段。需要注意的是，语音识别技术包括特征提取、声学建模、语言建模、解码等几个方面，其中声学建模的发展最为显著。上述发展阶段基本上是以声学模型的发展而划分的。因而，本节主要关注声学模型技术（特征提取在深度学习方法中成为声学模型的一部分），同时简述其它几种技术的发展现状。

概率模型方法

语音识别技术发展初期以模式匹配方法为主，对不同词保留若干各自的样本，将待测试语音信号与这些标准样本进行匹配，取距离最近的样本所对应的词标注为该语音信号的发音。这一方法有两个主要问题：(1)不能有效描述语音信号在时序上的不确定性，即短时平稳属性；（2)不能有效描述语音信号在发音特征上的不确定性，即不同条件下同一发音的不确定性。为解决上述因难，Reddy、Jelinek、Baker 等研究者提出基于概率模型来描述这些不确定的发音。这一模型主要包括两个部分：在描述时序动态性上，认为一个发音单元（词或音素）包括若干状态，同一状态内部的发音特性保持相对稳定，不同状态间的转移具有随机性；在描述发音特征的不确定性上，通过概率模型描述某一发音状态内部的特征分布。应用最广泛的

概率模型是 HMM/GMM 模型（如图 2 所示），其中 HMM 用来描述短时平稳的动态性，GMM 用来描述 HMM 每一状态内部的发音特征。

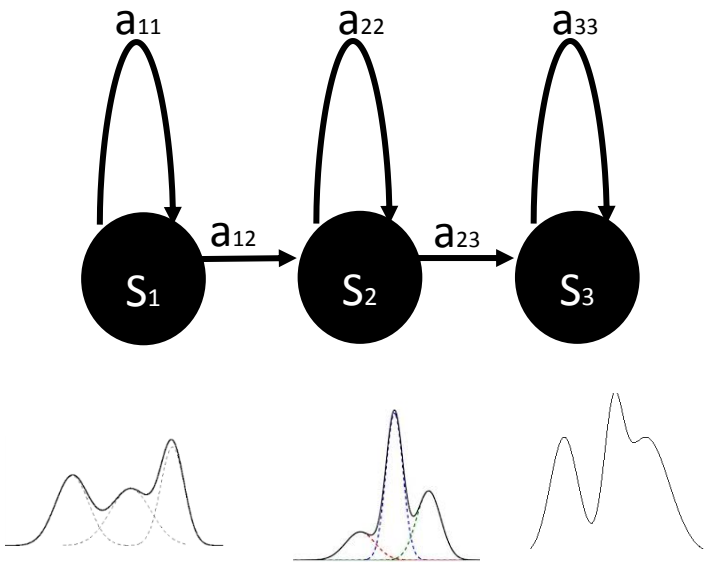

图 2 . HMM/GMM 模型

HMM/GMM 模型结构简单，有保证收敛的快速训练方法，可扩展性强，因此一直到 2011 年一直是语音识别领域的主流方法。基于 HMM/GMM 框架，研究者提出各种改进方法，如结合上下文信息的动态贝叶斯方法、区分性训练方法、自适应训练方法、HMM/NN 混合模型方法等。这些方法都对语音识别研究产生了深远影响，并为下一代语音识别技术的产生做好了准备。

深度学习方法

深度学习是“使用包含复杂结构或由多重非线性变换构成的多个处理层对数据进行高层抽象的算法”。深度学习在语音识别领域中的应用始于 2009 年，Mohamed 等在 NIPS workshop 上发表的“Deep Belief Networks for phone recognition”报告了基于 DNN 的声学模型在 TIMIT 数据集上得到了 23%的错误率，远好于其它复杂模型。之后，微软、IBM、谷歌等公司对深度学习模型进行了深入探索，尝试了各种深度学习模型在不同识别任务上的效果。今天，深度学习技术已经成为语音识别中的主流方法，基于深度模型的语音识别系统不论是识别率还是鲁棒性都远好于基于 HMM/GMM 的系统。

2013 年以前，DNN 是语音识别中应用最广泛的深度模型。DNN 是具有多个隐藏层的多层神经网络，具有强大的特征学习和分类能力。经过合理的初始化（如预训练），DNN 可通过随机梯度下降(SGD)算法进行优化。DNN 在声学建模中的应用可分为两种方式，一是混合建模方式，即用 DNN 代替 GMM 来描述状态输出概率;另一种是特征提取方式，即利用 DNN 提取抽象特征，再送入传统的 HMM/GMM 模型进行声学建模。这两种方式各有优势，其中混合建模更简单有效，是大多数商用系统采用的方式，而特征提取方式对资源的要求比较低，通常用在小语种识别等数据稀疏场景中。

随着研究的深入，研究者对 DNN 声学模型特性的理解也越来越全面。首先，人们发现 DNN 具有很强的特征提取能力，可以从频谱甚至时域信号中直接学习语音特征。这种纯数据驱动得到的特征在很多识别任务上远好于基于听觉感知特性设计的特征（如 MFCC 和 PLP）。第二，人们发现 DNN 具有强大的环境学习能力，可以对多种噪音、口音条件下的信号进行统一学习，极大提高了系统鲁棒性。第三，人们发现 DNN 非常适合多任务学习和转移学习，利用一种语言的数据训练出的 DNN，可以直接用到另外一种语言上做为特征提取模型。

DNN 的成功激励研究者探索更有效的深度模型，其中具有重要意义的是卷积神经网络（CNN）和循环神经网络（RNN）。这两种网络在深度学习提出之前已经被研究多年，在深度学习框架下取得了更好的效果。

CNN 是一种比 DNN 更有效的特征提取方法，它利用语音信号中典型模式（如音素）的重复性和局部性，将 DNN 的全连接结构变成时频空间中的局部连接结构，相当于设计了一系列具有局部关注特性的滤波器，并通过训练学习滤波器的参数。这一思路将 DNN 模型的特征学习方式进一步结构化，不仅减小了参数量，减小了模型训练难度，得到的 CNN 模型也更加符合特征提取的结构化要求。

RNN 模型是一种状态累积的时序动态模型。通过时间上的循环连接参数，RNN 可以学习更长时的历史信息，进而提高模型的预测和分类能力。历史上人们曾尝试过将 RNN 应用于语音信号建模，但由于系统的复杂性和模型训练上的困难，这些尝试大多局限在小数据任务上，效果不显著。随着深度学习中模型优化方法的进步和数据量的增多，人们发现当有足够多的训练数据时，RNN 模型确实可以有效学习信号的动态性，提高语音识别的性能。进而，研究者探索出一系列更适合语音建模的 RNN 结构，如 LSTM， GRU，双向 LSTM 等。同时，人们发现将多层 RNN 迭加起来形成深层 RNN 结构，可进一步提高识别性能。

RNN 更深远的影响是对 HMM 模型统治地位的冲击。同 HMM 一样，RNN 模型是一种时序模型，通过累积历史信息进入不同的状态，进而改变模型输出特性。和 HMM 的离散状态结构不同的是，RNN 是一种连续状态模型，因而适合描述语音信号从起始到结束的动态发展过程。因此，利用 RNN 代替 HMM，用连续状态序列代替离散状态序列，进而把语音识别的所有模块统一成神经网络模型，是件非常吸引人的事。研究者曾做过一些这方面的探索，但直到 2014 年端对端训练方法出现以后，这一思路才最终确定下来。以 CTC 准则为目标的端对端训练方法不再依赖一个初始 GMM 模型对信号和标注进行逐帧对齐，而是考虑所有可能的路径来计算损失函数，因而有望得到更精确的模型。特别重要的是，基于 RNN 结构，音素内部的状态变化不再用 HMM 来描述，而是依赖 RNN/LSTM 内部的状态累积。这意味着统治语音识别研究近 40 年的 HMM 模型至少已经变成一个可选项。

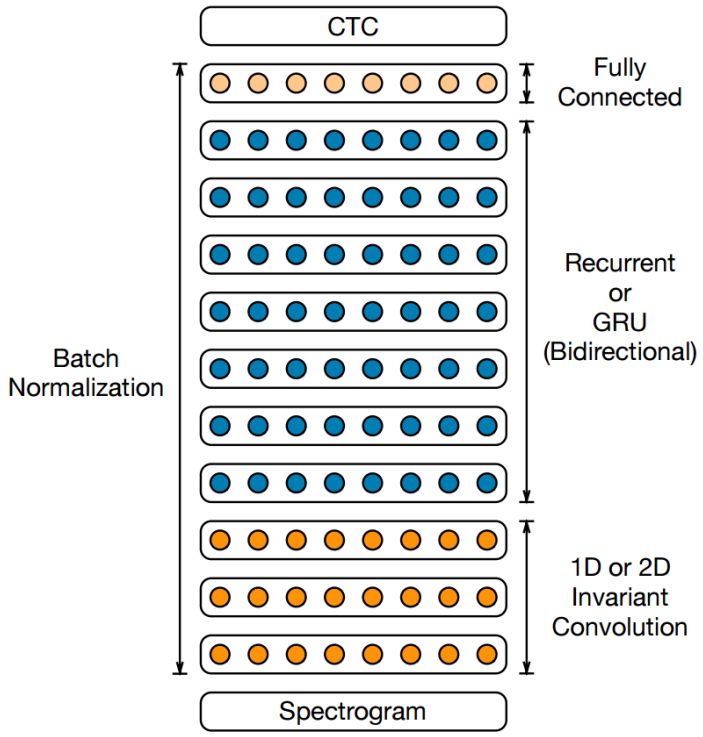

图 3. 当前语音识别系统中的声学模型结构（DeepSpeech2）

图 3 是当前语音识别系统所采用的一种典型的神经网络模型结构。该结构从频谱开始，通过 3 层 CNN 学习发音特征，通过 7 层 RNN 学习信号的静态和动态特性，最后通过 1 层全连接网络输出音素（或其它语音单元）的后验概率。RNN 层可采用 GRU 或 LSTM 结构，并可采用双向结构。训练时以 CTC 为目标，同时加入 BN 等控制梯度的方法，保证训练的收敛。

语言模型和解码器

前述内容主要是声学模型上的进展。相对而言，语言模型和解码器并没有发生太大变化。在语言模型方面，绝大部分系统依然基于传统的 n-gram 模型加上各种平滑算法。近年来，基于深度神经网络的语言模型(NNLM)取得很大进展，但 NNLM 不论训练和推理都显著慢于 n-gram，特别是应用到一次解码中，还需要较多的工程化工作。这使得 NNLM 还不能取代传统 n-gram 成为主流的语音模型结构。另一方面，随着训练速度、新词处理、应用框架等问题的解决，NNLM 应该很快会取代 n-gram 成为主流，甚至成为端到端网络的一部分。

在解码器方面，绝大多数系统依然采用基于 FST 的静态解码方法。这一方法预先将 LM、词表、决策树、HMM 模型等各层次知识统一表达成 FST，并将这些 FST 编译成一个“端到端”的解码图，其中输入为音素状态，输出为词。在 CTC（端到端）系统中，解码图仅需包括 LM 和词表，因而极大简化了构图流程。基于这一解码图，应用动态规划算法（如 Viterbi）即可实现解码。这一静态解码方法不需考虑众多各异化的资源，极大简化了解码器的结构和工作流程，而且可以对解码图进行确定化、最小化等离线优化，提高解码效率。这一方法的缺点是编译后的解码图难以进行动态更新，如增加新词等。研究者提出了嵌套子图算法和相似对

方法等技术，可以部分解决这一问题。

技术展望

语音识别技术已经逐渐走向成熟，在特定领域、特定环境下已经达到实用化程度。然而，在自由发音、高噪声、同时发音、远端声场等环境下，机器识别的性能还远远不能让人满意。本节对这一技术的未来发展做一展望，希望引起更多兴趣。

远端语音识别

当前近端语音的识别性能基本可以满足很多应用场景的需求，但远端语音识别的性能较差。这是因为远端声音包含更多背景噪音，且有回声干扰。当前远端语音识别多依赖各种麦克风阵列技术，包括各种 beamforming 技术和最近提出的基于 DNN 的信道融合技术。除了麦克风阵列，分布式麦克风技术也引起关注，但在理论和实践上还需进一步发展。相对人耳对远端声音的鲁棒性，远端语音识别性能的急剧下降可能意味着我们需要新的方法和思路，以便更深入地理解和描述声音信号的特性及其与声学模型的匹配性。

多语种、小语言、方言识别

当前基于 DNN 的语音识别对资源丰富语言（如英语、汉语）的识别性能已经可满足实用性要求，但对小语种和方言这些资源稀缺语言的识别性能还比较差。如何利用多任务学习和转移学习，实现对资源稀缺语言的“共享学习”，依然是比较困难的问题。特别是，如何实现多种语言在统一解码空间中解码，还需要一些探索。

多任务协同学习

语音信号中包括说话内容、说话人、情绪 、信道等各种信息，这些信息混杂在同一信号中，在不同任务中的重要性各有不同。例如，语音识别希望只保留说话内容而去掉说话人信息，反之说话人识别希望保留说话人信息而去掉说话内容。如果将这两个任务放在一起协同学习，让每一任务可利用其它任务的信息，则有望同时提高各个任务的性能。这一协同学习也是人类学习的典型方式。

语音-语义协同学习

语音识别的最终任务是让机器能理解人的语义，而非简单转换成文字。因此，语音识别最终要包含语义理解模块。当前语音识别和语义理解的研究还相对割裂。幸运的是，当前语义理解的主流方法同样基于 DNN/RNN 模型，这为两者的结合提供了基础。端对端训练有可能是一种有效的方法。

神经网络结构学习

当前深度学习越来越依赖计算图模型(computing graph)架构。基于这一架构，研究者可对神经网络的结构进行任意设计，计算图模型架构可自动计算梯度，从而极大节约了设计优化算法上的成本。一些大公司依赖其计算资源优势，利用这一方法寻找最合适的网络结构。未来这一方法在工程上可能会设计出极其复杂的网络，显著提高最终系统的性能。然而，这种穷举式的结构搜索方法可能会被结构学习方法所替代，即将网络结构也作为参数的一部分进行学习，从而通过数据驱动得到优化网络。

神经网络持续学习

当前网络优化多基于 SGD 算法。这一算法的一个显著缺陷是当网络学习完成以后，很难对新的数据进行学习。这显然不能满足实际应用的需要：我们希望对持续得到的新数据进行连续学习，使得模型可以持续更新，逐渐忘记以前的环境，适应到新环境。研究者提出了一些方法（如 AdaGrad）来解决这一缺陷， 但这些方法是否能实现一个持续学习的语音识别系统，需要进一步研究。