语音识别研究的发展、现状以及进展

一、 引言与语音识别的研究意义

语言，传递信息的声音，是人类最重要的交际工具，是人类相互交流最常用、有效的通信形式，语音是语言的声学表现。能够与机器进行沟通交流是我们人类一直想要做的事情，

语音识别（Speech Recognition），是一门研究实现人与计算机用语音进行有效通信的各种理论和方法的新兴科学，是融计算机科学，数学，统计学，声学，语言学于一体的交叉学科，是当前一门快速发展并已经得到一定广泛应用的研究领域。随着1952年贝尔研究所Davis等人研究成功了世界上第一个能识别10个英文数字发音的实验系统的诞生，语音识别技术经过了多年的沉淀与积累，随着当今机器学习理论，人工智能硬件和算法原理的发展而焕发出了新的活力。如今，语音识别已经在搜索引擎、智能玩具、智能家电、智能手机、智能客服等领域都被广泛地应用。

用语音识别与计算机进行通信，具有明显的实际意义和理论意义。人类可以进行“说话”与计算机进行交换，不必再用大量双手放在键盘上，可以解放双手。这将进一步发挥人类和计算机各自的优势能力。人类也可通过它进一步了解人类的语言、语音能力和智能的机制。在当今数据量急剧增长，人机交互更加频繁的数据时代，语音识别技术将具有更加重要的重要应用价值。

二、 发展脉络及研究背景

语音识别技术最早起源于上个世纪的50年代，在那个时间，由于研究处于空白时期，许多研究都要从零开始，所以起步时期，语音识别的研究主要针对元音、辅音、数字和许多孤立词的识别。

上个世纪60年代时期，语音识别的研究突破了当时的瓶颈，突飞猛进地发展，取得了一定的成果。线性预测分析和动态规划的提出较好地解决了语音信号模型的产生和语音信号不等长两个问题，并通过语音信号的线性预测编码，有效地解决了语音信号的特征提取。

上个世纪70年代左右，语音识别技术又取得了实质性地进展。基于动态规划的动态时间规整技术（Dynamic Time Warping, DTW）已经成熟起来，随后提出了矢量量化（Vector Quantization, VQ）和隐马尔可夫模型（Hidden Markov Model, HMM）理论[1]。

时间又推进到了20世纪80年代，语音识别开始从孤立词、连接词转向大词汇量、非特定人、连续语音的识别，较前三十年有了一定进步。进行识别的算法也通过传统的基于标准模板匹配的方法转变到了基于统计模型的方法。在声学方面，模型采用也由于HMM能够很好地描述语音时变性和平稳性，开始被光放应用于大词汇量连续语音识别（Large Vocabulary Continuous Speech Recognition, LVCSR）的声学建模[2-3];在语言方面，模型采用也是以n元文法为代表的统计语言模型开始广泛应用于语言识别系统[4]。

来到了20世纪90年代，语音识别在细化模型的设计、参数提取和优化、系统的自适应方面取得较大进展[5]。此时，语音识别技术正在与其他相关领域的技术进行有机地结合，通过该手段以此来提

高识别的准确率，便于语音识别技术向产品、商品的方向发展。

三、 目前研究水平、存在问题及未来发展方向

3.1 当前研究水平

进入21世纪，2006 年，Hinton 等人[5]提出逐层贪婪无监督预训练深度网络之后，微软成功地将深度学习应用到自己的语音识别系统中，比起之前的最优方法，使单词错误率降低了约30%[6]，这称得上是语音识别领域中的再一次重大突破。随后，微软的基于上下文相关的深度神经网络—隐马尔可夫模型(context-dependent DNN-HMM，CD-DNN-HMM) 对大词汇量语音识别的研究成果，彻底改变了语音识别系统的原有技术框架[7]。目前许多国内外知名研究机构，如微软、讯飞、Google、IBM 都积极开展对深度学习的研究[8]。在人们生活的应用层面上，由于移动设备对语音识别的需求与日俱增，以语音为主的移动终端应用不断融入人们的日常生活中，如国际市场上有苹果公司的 Siri、微软的 Cortana 等虚拟语音助手; 国内有百度语音、科大讯飞等。还有语音搜索(VS) 、短信听写(SMD) 等语音应用都采用了最新的语音识别技术。现在，绝大多数的SMD 系统的识别准确率都超过了 90% ，甚至有些超过了 95% ，这意味着新一轮的语音研究热潮正在不断兴起[8]。2012年，微软邓力和俞栋老师将前馈神经网络FFDNN（Feed Forward Deep Neural Network）引入到声学模型建模中，将FFDNN的输出层概率用于替换之前GMM-HMM中使用GMM计算的输出概率，引领了DNN-HMM混合系统的风潮。长短时记忆网络（LSTM，LongShort Term Memory）可以说是目前语音识别应用最广泛的一种结构，这种网络能够对语音的长时相关性进行建模，从而提高识别正确率[9]。双向LSTM网络可以获得更好的性能，但同时也存在训练复杂度高、解码时延高的问题，尤其在工业界的实时识别系统中很难应用。

3.2 当前所面临的问题

由上文我们可以知道，语音识别技术已经取得了长足的进步，各项子任务的处理速度和准确度也在不断地提升。但是这一领域仍然存在一些问题，亟待人研究人员和开发人员探索解决。

3.2.1强化学习与语音识别技术仍待结合

强化学习是指学习者（计算机）可以通过和环境的交互获知自己学习的结果，通过接收到的外界信号来指导自己的行为，从而逐渐获得预期的最好效果。强化学习在图像识别等领域已经取得了很好的效果，计算机可以通过人工标注的权重来进行自调整。然而，不容知否的是，强化学习在语音识别领域的应用仍旧较少。

但是，由于语音识别涉及的评价维度更加多种多样，人工标注反馈的成本巨大等原因，强化学习尚未在语音识别领域发挥其最大的价值。例如用户在与智能客服进行交互后的评价过程实际上就是语音识别的信号回馈过程[10]，但用户往往只能给出模糊的评分，缺乏其他维度的反馈，这使得智能客服的表现并不能通过客户的回馈得到显著的改善。这一领域仍待研究者进行探索。

3.2.2 语音识别在口声和噪声中存在的问题

语音识别中最明显的一个缺陷就是对口音和背景噪声的处理。最直接的原因是大部分的训练数据都是高信噪比、美式口音的英语。比如在交换台通话的训练和测试数据集中只有母语为英语的通话者（大多数为美国人），并且背景噪声很少[11]。

而仅凭训练数据自身是无法解决这个问题的。在许许多多的语言中又拥有着大量的方言和口音，我们不可能针对所有的情况收集到足够的加注数据。单是为美式口音英语构建一个高质量的语音识别器就需要 5000 小时以上的转录音频。

3.2.3 语音识别在模型、自适应、强健性方面的问题

从算法模型方面来说， 需要有进一步的突破。目前使用的语言模型只是一种概率模型，还没有用到以语言学为基础的文法模型，而要使计算机实实在在地理解人类的语言，就必须在这一点上取得进展[12]。

从自适应方面来讲，语音识别技术也有待进一步改进，做到不受特定人、口音或者方言的影响，这实际上也意味着对语言模型的进一步改进。

从强健性方面来说，语音识别技术需要能排除各种环境因素的影响。 目前，对语音识别效果影响最大的就是环境杂音或嗓音，个人能有意识地摒弃环境嗓音并从中获取自己所需要的特定声音，如何让语音识别技术也能达成这一点是一个艰巨的任务[13]。

3.3未来发展

多语言混合识别以及无限词汇识别方面：将来的语音和声学模型可能会做到将多种语言混合纳入，用户因此就可以不必在语种之间来回切换。 此外，对于声学模型的进一步改进，以及以语义学为基础的语言模型的改进，也能帮助用户尽可能少或不受词汇的影响，从而可实行无限词汇识别[14]。多语种交流系统的应用是将语音识别技术、机器翻译技术以及语音合成技术的完美结合，全世界说不同语言的人都可以实时地自由地交流，不存在语言障碍。可以想见，多语种自由交流系统将带给我们全新的生活空间。语音情感识别：近年来随着人工智能的发展，情感智能跟计算机技术结合产生了情感计算这一研究课题，这将大大地促进计算机技术的发展。 情感自动识别是通向情感计算的第一步[15]。 语音作为人类最重要的交流媒介，携带着丰富的情感信息。 因此，如何从语音中自动识别说话者的情感状态，近年来受到了各领域研究者的广泛关注。

四、 总结和感想

在计算机以及人工智能等技术的发展历程中，语音识别的起步相对较晚，理论基础相较于计算机视觉等较为成熟的领域仍旧较浅，但正因如此，语音识别也是一个具有广阔研究领域和应用领域的

研究方向。语音识别面临实际问题多种多样，难以用一成不变的套路应对变化万千的需求，这对自然语言处理的研究人员是巨大的挑战，对具体的实现技术也提出了很高的要求。

可以预见，在不久的将来，随着语音识别技术的不断进步，语音识别系统的研究将会更加深入，语音识别系统的应用将更加广泛。各种各样的语音识别系统产品将出现在市场上。人们也将调整自己的说话方式以适应各种各样的识别系统。在短期内还不可能造出具有和人相比拟的语音识别系统，要建成这样一个系统仍然是人类面临的一个大的挑战，我们只能朝着改进语音识别系统的方向一步步地前进

参考文献：

[1] 蘧鹏里.语音识别技术综述[J].计算机产品与流通,2018(08):105.

[2] Zixing Zhang, Jürgen Geiger, Jouni Pohjalainen. Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2018,7(9-5)

[3]Hinton G E，Osindero S，Teh Y． A fast learning algorithm for deepbelief nets［J］． Neural Computation，2006，18(3) : 1527-1554．

[4] Bengo Y，Courcille A，Vincent P． Representation learning: a reviewand new perspectives[J]．IEEE Trans on Pattern Analysis andMachine Intelligence，2013，35(8) : 1798-1828．

[5] Dahl G，Yu Dong，Deng Li，et al． Context-dependent pretraineddeep neural networks for large vocabulary speech recognition［J］．IEEE Trans on Audio，Speech，and Language Processing，2012，20(1) : 30-42．

[6] Hinton G E，Deng Li，Yu Dong，et al． Deep neural networks foracoustic modeling in speech recognition: the shared views of four re-search groups［J］． IEEE Signal Processing Magazine，2012，29(6) : 82-97．

[7] Yekutiel Avargel and Israel Cohen. 2007. System identification in the short-time fourier transform domain with crossband filtering. IEEE Trans. Audio Speech Lang. Process. 15, 4 (Mar. 2007), 1305–1319.

[8] Zhuo Chen, Shinji Watanabe, Hakan Erdoğan, and John R. Hershey. 2015. Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks. In Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'15). Dresden, Germany, 1–5.

[9] Tian Gao, Jun Du, Li-Rong Dai, and Chin-Hui Lee. 2015. Joint training of front-end and back-end deep neural networks for robust speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP'15). 4375–4379.

[10] Yedid Hoshen, Ron J. Weiss, and Kevin W. Wilson. 2015. Speech acoustic modeling from raw multichannel waveforms. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'15). 4624–4628.

[11] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. 2015. Joint optimization of masks and deep recurrent neural networks for monaural source separation. IEEE/ACM Trans. Audio Speech Lang. Process. 23, 12 (Dec.2015), 2136–2147.

[12] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël A. P. Habets, Reinhold Haeb-Umbach, Walter Kellermann,Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, and others. 2016. A summary of the REVERB challenge:state-of-the-art and remaining challenges in reverberant speech processing research. EURASIP J. Adv. Sign. Process. 2016,1 (Dec. 2016), 1–19.

[13]侯一民,周慧琼,王政一.深度学习在语音识别中的研究进展综述[J].计算机应用研究,2017,34(08):2241-2246.

[14] Kang Hyun Lee, Woo Hyun Kang, Tae Gyoon Kang, and Nam Soo Kim. 2017. Integrated DNN-based model adaptation technique for noise-robust speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17). 5245–5249.

[15] Andrew L. Maas, Quoc V. Le, Tyler M. OŃeil, Oriol Vinyals, Patrick Nguyen, and Andrew Y. Ng. 2012. Recurrent neural networks for noise reduction in robust ASR. In Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'12). 22–25.