语音合成让计算机声文并茂、有声有色

蔡莲红 王玮 陶建华 吴志勇 王志明
清华大学计算机科学与技术系
人机交互与媒体集成研究所
智能技术与系统国家重点实验室
北京 100084

1.什么是语音合成

机器说话
说什么
怎样说
文字到语音的转换

2 用户接受合成的声音

用户听懂合成的声音
用户听清合成的声音
用户需要合成的声音（即时性）

机器要象人那样说话

语音合成技术把可视的文本信息转化为可听的声音信息，其应用和经济社会效益前景非常良好。尤其对于汉语语音合成技术应用而言，面对着有十几亿人使用中文的泱泱大国，市场需求、应用前景和经济效益等等都可见一斑。

语音技术在智能电话查询系统中的应用，语音技术已逐渐在电信的声讯信息服务领域内的智能电话查询系统中展开应用，并在迅速地推广，在电话高度普及的今天，如果打电话就能查询到所需信息，无疑将给人们的日常生活带来极大方便。汉语语音合成技术应用到声讯服务领域内，对现有的电话查询系统将产生革命性的影响。

语音技术与互联网的成功结合，电话因特网关是一种用于实现电话网和因特网之间的信息互访的系统。简而言之，就是让电话用户能够轻松地通过电话访问因特网，系统的功能主要体现在两个方面。一方面，让用户通过电话、手机或传真随时随地访问因特网上各种信息，例如新闻、电子邮件等，大大扩展了因特网信息的用户群和地域范围，同时大大降低了用户参与到因特网的技术难度。另一方面，能够将电话终端上信息流或控制指令发送到因特网上，例如用户可以通过电话方便地发送电子邮件和类似的留言信息，不仅具有传统的语音邮箱的功能，还可以将用户语音以 IP 的方式廉价地发送到全球的任何一个电脑或电话终端，大大降低了信息交流的成本，语音合成技术的信息服务得到了用户的广泛接纳，给用户生活提供了极大的方便。

3 语音合成的应用

电子图书的有声输出

网络信息的即时有声发布

办公自动化

口语机器翻译受到重视

由于残疾人渴望与健康人进行交流，因此口语翻译越来越受到人们的重视。口语翻译的目的就是帮助聋哑人和正常人交流，首先，聋哑人要戴上一双特制的手套，计算机根据他打出的手语并且进行识别，然后，通过语音合成系统就可以把图像信息翻译成语言信息。同时，系统也能够完成将健康人的语言翻译成聋哑人的手语，只要将健康人先把要说的话键入计算机，经程序分析处理之后，翻译成有表情、有动作的三维图像，从而最终达到了聋哑人与正常人之间通过翻译机进行交流的目的。口语翻译的研究在其他很多方面都有重要价值，如用手势控制计算机，甚至用手势导航等等。

2. 语音技术的展望

2.2 高自然度、具有表现力的合成语音

提高合成语音的自然度仍然是高性能文语转换的当务之急。就汉语语音合成来说，目前在单字和词组一级上，合成语音的可懂度和自然度已基本解决，但是到句子乃至篇章一级时其自然度问题就比较大。未来的文语转换系统发展趋势是应该采用基于语境相关的合成思想进行设计的，能够将发音人的原始发音特征最大限度的保留下来，辅助以先进的层次化语言韵律模型，通过分散统计的模型方法来涵盖语义语音之间的内在联系，使系统能够输出具有高自然度和表现力的合成语音，但是目前合成系统中普遍存在合成输出的机器味比较浓、语境的知识层次模型研究不完善等问题。因此获得高自然度、具有表现力的合成语音也是今后语音技术的研究目标之一。

2.3 语音技术与多媒体的结合

伴随着现代语音技术的不断发展，人类对语音信号的需要已经不在仅仅停留在可懂性和正确性上，当前语音合成技术的研究方向是合成语音的美感并同时输出辅助的视频特征，实现虚拟主持人的效果，通过将视觉效果包括人的头部建模，唇形同步等技术和表情因素等视频信息的加入，可以更好地体现语音合成系统的表现力和感染力。因此完全有理由相信将语音技术和多媒体技术的有机结合将会提供给合成系统广阔应用前景。

2.4 语音技术与网络的结合

目前语音技术已经逐渐在电信的声讯信息服务领域和互联网消息收发进行应用，随着电话网与互联网的融合，联网信息项目的增多和时效性要求的逐步提高，建立适合于股票交易、航班动态查询、电话自动报税等业务成为可能，电话用户可以通过传统的语音、传真获取互联网上无穷无尽的信息，这些业务都将全面彻底解决传统数字录音回放技术所无法解决的海量信息库和动态变化信息的实时生成和存储的难题，因而将语音技术与网络进行完美的结合是有强大的生命力的。

3. 语音计算的拓展

3.1 韵律研究与感知相结合

韵律是语音信号的自身属性，它反映了一个人说话时的语调高低时间的长短信息，同时反映了说话人说这句话时的语境信息，韵律模块也是语音合成系统中的重要组成模块，韵律参数研究的成功与否直接影响合成系统的输出。感知信息体现说话人对一句话中某些部分的强调和语句重音信息，语句重音也对系统的合成输出产生很大的影响，因此要想得到较好的语音合成效果需要对韵律和感知进行深入的研究。

3.2 从语法、语义层面探索语音计算的理论和方法

语音计算中包含对语言的语法、语义的理解，语音合成系统的输出不仅仅取决于语音数据音质的好坏，同时在很大程度上受到所处理文本的语法及语义现象的制约，如果没有正确的语法描述，合理地体现语义信息，就不可能产生很好的合成效果，而获得这种相互关系只有通过对大量的语言现象进行分析总结形成规则描述，为了更加客观的进行描述可以借助于人工智能领域里的数据挖掘方法研究，因此语音计算的关键技术是挖掘语法、语义和语音之间的相互关系，将这种关系采用规则描述结合到实际合成语音系统中，提高语音合成输出的自然度。

3.3 建设海量语音数据资源

语音计算的成功与否在很大程度上取决于语音资源的积累。目前比较先进的语音处理方法中都无一例外的提到了采用基于数据的驱动方式进行，然而这种基于数据驱动的方式就首先就需要大量的语料数据，没有大语料就无从谈起数据的驱动，因此为了尽可能地覆盖各种语言现象，就需要长期积累各种语音资源，同时对于语音信号的处理也需要大量的语音处理软件，这些都是一种日积月累的过程。

4. 语音合成的新进展

4.1 神经网络用于训练韵律模型

基于人工神经网络具备良好的自学习和自适应能力，将神经网络技术应用于语音合成系统中韵律模型的研究具有很重要的意义。将神经网络模型与已有的文语转换系统有机结合，

可以改变传统的文语转换系统的韵律模型具有了更强的适应性和可训练性，使得合成语音的自然度得到了显著提高，增加了系统的灵活性和风格的多样性。

4.2 数据挖掘用于发现语音知识

数据挖掘作为一种在大量数据库中发现隐藏新知识的计算技术方法，通过语音定性模型的建立，将数据分析和挖掘结果转化为逻辑规则或用可视化的形式进行表达。因此，将数据挖掘和人机交互和接口紧密的联系在一起会对计算机语音信号处理的研究工作产生巨大的推动能力，为语音信号处理提供了一条崭新的研究途径。

4.3 文本-可视语音转换系统研制成功

文本-可视语音转换技术的出现，是多媒体技术迅速发展的产物，也迎合了社会发展的需求。它给人们的生活增添了新的色彩，使计算机更人性化，人们与计算机的交流变得更为简单。相信在不久的将来，它将会在众多的技术、商业和娱乐领域得到广泛的应用，并逐步进入我们每个人的生活。