语音合成技术最新研究进展及其应用展望

王仁华
(中国科学技术大学，安徽 合肥 230027)

摘要

文章介绍了语音合成技术研究的最新进展，展望了语音合成在网络信息服务、人机自然交互、移动信息终端及各种嵌入式设备上的应用前景。

关键词

语音合成；文语转换；研究进展；应用展望

编者按

语音合成技术研究的目标是让信息时代的各种机器像人一样能“开口说话”，该技术因其广阔的应用前景而倍受关注。随着该技术的不断成熟，语音产品正处于规模化进入市场的关键阶段，极有希望形成一个全新的产业。

中图分类号:TN912.33
文献标识码:C
文章编号:1009-6868(2003)05-0037-03

语音合成技术是计算机“开口说话”的关键，现阶段语音合成的最大进展是已经能够实时地将任意文本转换成连续可懂的自然语句输出，相应技术通常称为文语合成或文语转换(TTS)。TTS使得数据通信和语音通信在终端一级实现交融，人们将有望在获取因特网信息时，使短消息服务、电子邮件等多数以文本方式提供的信息也用语音的方式输出，极大地方便终端用户[1]。

1 语音合成技术新进展

1.1 合成语音自然度大幅度提高

所谓自然度就是指合成语音听起来是否自然，是否像人的自然语音。这是制约语音合成应用的第一要素。近年来流行的基于大语料库的合成系统采用自然语音波形直接拼接的方法，进行拼接的语音单元是从一个预先录下的自然语音数据库中挑选出来的，因此有可能最大限度地保留语音的自然度。基于听感量化理论的语音合成新方法[2]，目前对语料库设计、韵律预测、音库裁减、以高层韵律描述为输入合成单元挑选等基于大语料库语音合成关键技术的研究取得了重要突破。在最近一次对采用听感量化理论语音合成方法的语音系统的语音自然度评测中，选择日常生活中高频使用的语料内容，将合成系统输出语音和国家一级播音员、自然发音人的语音进行对比测听。参加测试人员为有正常听辨能力和表达能力、中专以上文化程度、35周岁以下人员24人组成，评分标准采用5分制。语音对比测听评分标准如表1所示。评测结果表明合成语音的自然度大幅提高，已经超过了普通自然人的发音。语音对比测听评分结果如表2所示。若以播音员的语音5.0分作为

表1 语音对比测听评分标准

评分	1	2	3	4	5
效果	极差，不能接受	较差，不愿接受	尚可，可以接受	较好，愿意接受	优秀，很自然

表2 语音对比测听评分结果

序号	系统代号	发音者	测听值
1	A	播音员	4.72
2	B	自然人	3.69
3	C	语音系统	3.98

基础，则自然人语音为4.0分，对比计算出语音系统的自然度指标为4.28分，结果令人鼓舞。

1.2 文语转换系统音库减小

目前高质量的文语转换系统一般需要几百兆字节甚至上千兆字节的存储容量，这在以PC机或工作站为硬件平台的应用中尚可接受，而对于像个人数字助理(PDA)、商务通及无线通信手机等资源有限的嵌入式设备上就没法承受。解决途径有两个：一是降低音库的冗余度，二是采用低速率的语音编码技术来存储音库。音库的裁减一般是通过聚类的方法，将听感上相似的合成基元归并，达到降低音库容量的目的。

目前采用的方法是使用最有利于表现听感距离的基频和线谱对频率(LSF)等语谱特征声学数据，计算单元之间的听感差异距离，在得到单元之间听感差异矩阵的基础上，利用数据挖掘码本设计算法(LBG)对数据进行聚类。考虑到LSF系数和基频曲线并不能唯一地确定音节单元在听感上的距离，因此在聚类的过程中还要考虑除声学层数据以外的“环境约束因素”。通过两个层面的聚类，音库的容量可以大幅度减小。

音库裁减后对合成语音的质量会有一定的影响。音库裁减的幅度和合成语音音质之间应折衷考虑。采用最新技术，当每个音节平均采用6个样本时，自然度得分可以达到3.9以上，每个音节平均采用3个样本时也可接近3.8分，再结合适当的语音编码技术，音库所需的存储量可降到1兆字节左右。

1.3 合成语音表现力提高

多语种的文语合成有着独特的应用价值。例如在自动电话翻译、有声电子邮件中都提出多语种的合成，即使是对汉语合成也有多方言文语转换的需求。近年来由于采用与语种无关的研究路线，使得中、英文混读合成系统取得了重要进展[3]。过去中国的语音合成系统，对于输入文本里的英文，往往舍弃或调用国外厂商的语音合成引擎，这不但需要付给国外语音厂商高额使用许可费用，而且合成文本中高度相关的中、英文文本合成的音色完全不一致，引起可懂度和自然度的急剧下降。最近采用统计和数据挖掘为主的方法开发语音系统，可以做到同时合成中文和英文，在中、英文混读时语音平稳、自然，并具有较好的语种扩展性，可以使用统一的技术平台实现不同语种的语音合成问题，避免了在合成混合语种文本时需要调用多个独立的语音合成引擎的问题，也为中国多方言的语音合成，打下了基础。

为了合成语音更人性化，语音合成与人脸合成结合也逐步受到重视。该项技术也被称为多模态(Multimodel)、视觉语音(Visual speech)、说话头(Talking head)等。将语音和图像两种模态结合到一起，能够对单一使用某一种模态的不足进行合理、有益的补充，使人们对信息的了解更全面。通过虚拟人脸，人机交互变得更加和谐。这一技术在个人数字助理、客户服务、电子商务、虚拟播音、游戏等方面都有相应的需求。近几年，基于大语料库的语音合成技术的发展给人脸合成研究带来了启发，研究者们已经开始把基于大语料库语音合成的思想借鉴到人脸合成的研究中来：首先制作一个合适的图像数据库，合成时按照一定的规则挑选出最好的单元，然后对这些单元进行变形插值处理，最后连接这些单元形成人脸动画视频。

2 语音合成应用展望

语音合成技术的逐步成熟，特别是在一般陈述语气下合成语音已超过普通自然人说话的水平，使得语音合成技术的应用前景无限(见图1)。

2.1 在网络信息服务中的应用

语音合成技术在网络信息服务中的应用主要表现在呼叫中心和各种计算机与电信集成系统(CTI)应用中，在中国最典型的就是中国电信的160/168声讯服务系统。2001年中国电信160/168系统的收入为25亿元，但由于现有系统采用的是传统数字录音回放技术，不能解决信息时代所迫切需要的大规模动态海量信息存储和即时生成问题，因此全国各地声讯台都需要进行语音合成技术的改造，即采用语音合成技术来代替数字录放，预计这一工作5年内完成，而且，当完成160/168系统改造以后，在该系统上还将开发更多的增殖业务，丰富多样的语音节目所蕴涵的市场价值至少是平台改造软件支出的2倍以上。一个成功的例子是2002年在韩

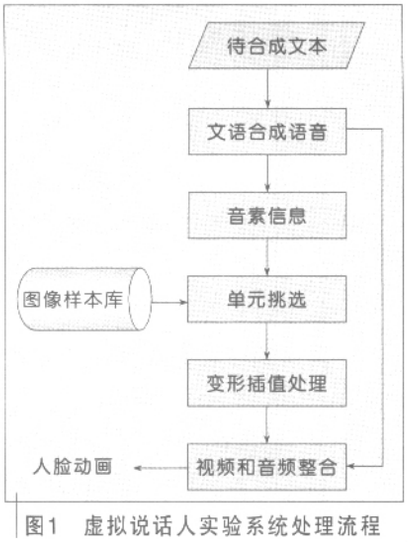

图1 虚拟说话人实验系统处理流程

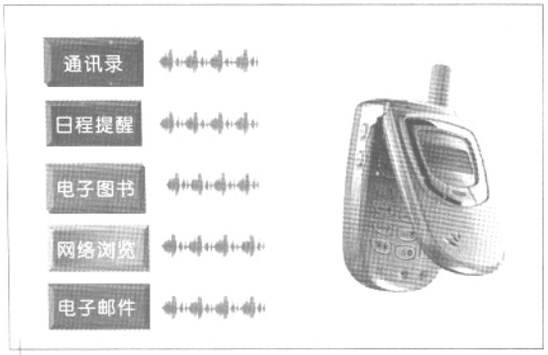

图2　语音合成技术在嵌入式设备上的应用

日足球世界杯期间建设的世界杯赛况查询系统。该系统使中国396个地方声讯台都可以通过拨打16897168同步了解到最新赛况、比赛花絮、球星风采等内容，取得了很好的经济效益和社会效益。另一个实例是国元证券呼叫中心。在国元证券的客服中心中，为了给用户提供丰富、多方位的资讯，有今日特别提示、财经信息、动态股评、公司介绍、营业部介绍、业务介绍等信息。但是由于这些信息都是随着经济活动时刻变化的，并且信息量巨大，根本无法用预先录音的方式来实现。在线语音合成系统就可以很好地解决这个问题，自动语音应答系统可以全天24小时给用户提供海量、实时的资讯，从而既承担了很大一部分客户服务工作，又节省了管理、维护成本。在电信领域，语音合成技术除了在160/168系统改造中应用以外，还有其他应用：如邮政185、电信1000号客服等。

2.2 在网络终端上的应用

利用语音和语言处理技术能增加电脑使用的趣味性并降低使用门槛。例如：语音日程提醒、时间播报等更人性化的语音秘书功能，语音听网、听书，朗读各种来源的新闻及小说，对各种编辑软件实现有声语音校对等。结合语音识别技术还可以实现语音听写、语音排版、声控上网、人机对话等。

语音和语言处理技术的应用等于在现有的Windows、Unix、Linux等操作系统之上又构建了一个更加适合中文电脑用户使用习惯的语音平台，从而满足信息社会中人们更高的需求：如在电脑前工作了一天，可以听听网上新闻休息休息，同时还可以做些别的事情。由于知识层次和年龄层次的差异，部分网民无法通过键盘、鼠标操作电脑，检索自己需要的信息，语音网络导航将使这部分人也能在网上尽情冲浪。

现在已经有产品能利用语音合成技术，将任意文本框以及网上浏览的内容用清晰自然的声音朗读出来，受到广大用户的欢迎。

2.3 在移动信息终端及各种嵌入式设备上的应用

在嵌入式设备，如在PDA、手机、智能玩具、信息家电和车载GPS上，利用语音合成技术在后PC时代有着越来越广泛的应用。

随着移动通信的发展，手机日益普及，需求量成倍增长。手机作为移动通信终端正朝着小型化、多功能化、个性化方向发展。带有语音合成功能的手机，可以用语音播报来电号码，概述电子邮件内容，给予日程提醒，收听网络信息等(见图2)。从手机的发展趋势来看，语音技术在手机上的全面运用已成为不可扭转的趋势。

采用全球卫星定位系统(GPS)来提供道路状况和定位信息是运输交通行业的一大趋势，几乎已成为未来车辆的必备设备。在车载GPS上加入语音技术，可使得驾驶员在眼与手忙的情况下，通过语音实时接受动态路况信息及通知、公告，及时获取感兴趣的车主个性化定制的信息，将平面显示导航上升到立体语音导航。

在消费类产品中结合MP3播放器，使MP3播放器不仅可以听音乐，还可以听小说！

此外，嵌入式语音技术还能在电子图书、智能语音玩具、“会说话的书”、测量仪器等众多领域得到广泛应用。

归纳起来，语音合成技术近5年来的发展突飞猛进，合成语音的可懂度、自然度已达到用户可以接受的程度，语音合成已基本进入大规模产业化的应用阶段。随着信息网络时代人们对信息获取多样性需求的不断增加，以及后PC时代各种嵌入式终端和移动通信设备在移动状况下和小屏幕终端上信息交互需求的不断提高，语音合成技术迎来了巨大的应用契机。语音合成的应用前景，既取决于技术的进步，又取决于市场开拓的力度。从技术上讲进一步研究发展的方向还有合成情感语音、对话语音等。相信随着语音合成技术的不断成熟与完善，语音合成技术必将会走进每个人的生活。

3 参考文献

[1] 王仁华. 智能通信终端 [J]. 中兴通讯技术, 2001,7(5):44-48.
[2] 刘庆峰. 基于听感量化的语音合成研究 [Z]. 中国科技大学博士学位论文, 2002.
[3] 吴晓如. 多语种语音合成中的关键技术研究 [Z]. 中国科技大学博士学位论文, 2002.

收稿日期：2003-07-30

作者简介：
王仁华，中国科学技术大学电子工程与信息科学系教授、博导。长期从事人机语音通信、数字信号处理、多媒体通信方面的科研和教学工作。自1990年以来共主持国家“863”项目10项、国家自然科学基金项目4项、中国科学院重大工程项目1项、国家“九五”攻关项目1项，研究成果多次获得中国科学院和省部级奖励，其中“KD系列汉语文语转换系统”获2002年国家科技进步二等奖。已出版专著3本，发表论文150余篇。