文章编号：1003-0077(2011)06-0137-05

言语信息处理的进展

蔡莲红[1]，贾 珈[1]，郑 方[2]

（1. 清华大学 计算机科学与技术系，普适计算教育部重点实验室，清华信息科学与技术国家实验室（筹），北京 100084；
2. 清华大学 信息技术研究院语音和语言技术中心，北京 100084）

摘 要：该文介绍了言语信息处理的进展，特别提到汉语言语处理的现状。言语信息处理涉及到言语识别、说话人识别、言语合成、言语知觉计算等。带口音和随意发音的言语识别有力的支持了语言学习与口语水平测评等应用；跨信道、环境噪音、多说话人、短语音、时变语音等因素存在的情况下提高识别正确率，是说话人识别的研究热点；言语合成主要关注多语言合成、情感言语合成、可视言语合成等；言语知觉计算开展了言语测听、噪声抑制算法、助听器频响补偿方法、语音信号增强算法等研究。将言语处理技术与语言、网络有效结合，促进了更加和谐的人机言语交互。

关键词：言语识别；说话人识别；言语合成；言语知觉计算

中图分类号：TP391　　**文献标识码**：A

The Research Progress of Speech Information Processing

CAI Lianhong[1], JIA Jia[1], ZHENG Fang[2]

(1. Key Laboratory of Pervasive Computing, Ministry of Education,
Tsinghua National Laboratory for Information Science and Technology (TNList),
Departmetn of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
2. Center for Speech and Language Technologies,
Research Institute of Information Technology, Tsinghua University, Beijing 100084, China)

Abstract: This paper introduces the progress of speech information processing, especially the researches on Chinese speech processing. Speech information processing includes speech recognition, speaker recognition, speech synthesis and computational speech perception. Researches on speech recognition with accent and personal style support the systems of language learning and evaluation, while speaker recognition focuses on how to improve the performance in different conditions. Researches on speech synthesis pay more attention on cross-language, emotional and audio-visual speech synthesis. Fomputational speech perception focuses on the implementation on speech testing and rehabilitation, denoising, and speech enhancement. Through these researches, especially the combination of speech information processing, linguistics and web technology, we can build more harmonious human-computer speech interaction system.

Key words: speech recognition; speaker recognition; speech synthesis; computational speech perception

1 概述

语言是人类最基本的信息交流方式之一。言语（语音，Speech）指人类按给定自然语言模式发出的声音；有时也指人类说话的声音信号。汉语言语处理的研究始于 20 世纪 50 年代。目前汉语言语处理的研究工作基本跟上了国际的步伐，并有所创新。

收稿日期：2011-10-24　定稿日期：2011-10-25

基金项目：国家自然科学基金资助项目(61003094,60928005,60805008)

作者简介：蔡莲红(1945—)，女，教授、博士生导师，主要研究方向为言语处理与合成；贾珈(1981—)，女，讲师，主要研究方向为信号处理、言语感知计算；郑方(1967—)，男，教授、博士生导师，主要研究方向为言语识别、说话人识别、自然语言处理。

言语处理涉及到言语识别、言语合成、说话人识别、言语知觉计算等，其成果正在向实用化方向迈进，已有一些商品问世。言语处理的研究也不断深入，如集成多种技术的计算机辅助学习，可视言语合成，语音翻译，语音检索等。本文介绍了言语信息处理的进展，特别提到汉语言语处理的现状。

2 言语识别与语言学习

2.1 言语识别

言语识别是指利用计算机识别出语音信号所表达的内容，理解其所蕴含的含义。隐马尔科夫模型在言语识别与建模中的应用[1]是近年来言语识别领域最突出的成果，使识别系统性能得到了显著的提高。目前，基于隐马尔科夫模型的言语识别框架仍然是言语识别研究的一个主要方向。言语识别可将声音转换成文字，辨认说话人身份，识别说话人的情感。随着语音识别和互联网技术的进步，基于言语识别技术拓展了研究领域和应用范围，例如，多语种言语识别、语种识别、言语情感识别、声音转换、机器翻译、口语水平自动测评、语音搜索、哼唱搜索、声乐演唱水平评价等。

言语识别性能的提高受到多种因素的影响[2]，主要体现在：(1)协同发音造成的影响；(2)不同说话人或说话方式不同造成的影响，如朗读式发音和随意发音会对识别模型的结果造成影响；(3)应用环境、采集设备和传输信道等不同造成的影响。《中国语音识别系统通用技术规范》(标准 GB/T21023—2007)的发布有力的支撑了中文言语识别技术的发展。

言语识别系统面对异常复杂的语音输入，其鲁棒性是言语识别技术实用化的关键问题[3]。因此，消除说话人相关因素对语音识别造成的影响和环境相关因素对语音识别造成的影响，提高言语识别系统的鲁棒性，才能解决以上因素引发识别率的退化问题。目前，言语识别的主要研究热点是带口音的语音识别和随意发音的语音识别。嵌入式言语识别、语音搜索、情感识别、基于云模式的系统构建等方面应用前景潜力巨大。

2.2 语言学习与口语水平测评

语言学习(特别是第二语言)的学习需要花费大量的时间和财力。据估计，中国约有 2 亿人正在学习外语；全世界另有 3 000 万人正在学习中文。据教育部估计，到 2010 年全世界有 1 亿人学习中文。如何发现语言学习中的问题，提高学习效率是人们关注的问题。

言语识别技术可辅助学习者进行发声训练，对学习者的错误发音进行检测和诊断，不断训练学习者的发音以达到增强其控制自身发音器官运动的能力；另一方面，利用可视言语合成技术，通过发音模型，从语音和视觉两个模态对学习者的错误发音与正确发音之间的区别进行矫正性的认知反馈，从而让学习者从两者的对比中不断地增强其准确区分不同发音的能力，并进而鼓励学习者在随后的发声过程中减少相应的错误发音。然而，现有的计算机辅助语言学习的研究工作，在针对发音错误提供适当而有效的诊断并反馈信息方面，还处于研究的起步阶段[4]。

相关研究包括英语口语学习评测、计算机辅助普通话水平测试评分、在线语言学习的交互平台、语言水平考试系统等。这些研究将言语处理技术与语言、网络有效结合，促进了言语处理技术的深化和拓展。研究中建立了具有矫正性认知反馈功能的基于网络的交互式在线语言学习平台，针对学习者在练习发音时的认知过程，为其提供了一个无所不在的计算机辅助的语言学习和训练环境[5]。

3 话者识别与身份验证

说话人识别属于生物特征识别技术的一种，是一项根据语音波形中反映说话人生理和行为特征的语音参数，自动识别说话人身份的技术。与言语识别不同的是，说话人识别利用的是语音信号中的说话人信息，而不必关注语音中的字词信息，它强调说话人的个人特性；而言语识别的目的是识别出语音信号中的言语内容，并不考虑说话人是谁，它强调共性。与其他生物特征的识别技术相比，声纹在应用方面有获取方便、使用简单、适于远程身份确认、算法复杂度低等特殊的优势。

说话人识别根据应用的范畴，可分为两类：(1)说话人辨认：用以判断某段语音是若干人中的哪一个所说的，是“多选一”问题；(2)说话人确认：用以确认某段语音是否是指定的某个人所说的，是“一对一判别”问题。根据进行识别的内容，又可分为三类：(1)文本相关：要求用户按照规定的内容发音，每个人的声纹模型逐个被精确地建立，而识别

时也必须按规定的内容发音；(2)文本无关：不规定说话人的发音内容，模型建立相对困难，但用户使用方便，可应用范围较宽；(3)指定文本：为防止通过预先盗取录制说话人语音等方式非法闯入系统，在进行识别时，系统会随机地指定说话人说出某段或某些段文本，只有说话人说出的语音与指定的文本一致且说话人识别结果为接受时才可以能被系统接受；或系统随机提问说话人某个或某些预先设定的问题，只有说话人回答的内容与预先设定的答案文本一致且说话人识别结果为接受时才可以被系统接受。此外，说话人识别还可分为语言无关、语言相关；按说话人不同可以分为：单说话人、多说话人等。

目前说话人识别的研究主要集中在各种识别参数的提取、选择和实验上。此外，分析各种声学参数的线性或非线性处理以及新的模式匹配方法，如动态时间规整、主成分分析、隐马尔可夫模型、神经网络和多特征组合等技术，也是目前语音信息处理热门研究方向之一[6]。在相关研究成果的支持下，说话人识别技术已逐渐走入实际应用。目前说话人识别技术的识别率，T-NETIx 公司的 SPeakEZ 达到 94%～95%；日本在此基础上研制的同类产品据称其识别率已达到 99.8%；而国内的声纹识别技术，说话人辨认的正确率不低于 99%，说话人确认的错误识别率和错误拒绝率均低于 1%，并首次在电话银行中用于身份认证。我国政府和科研单位、商业机构高度重视自主知识产权的说话人识别技术研发、标准制订和应用推广工作。在信息产业部科技司批准成立了中文语音交互技术标准工作组，并设立了“声纹识别特定领域技术标准”专题组。《中国声纹识别系统通用技术规范》(标准 SJ/T11380—2008)经信息产业部批准后正式成为国家标准，是我国说话人识别技术发展的重要标志。

说话人识别目前的研究热点主要集中在跨信道、环境噪音、多说话人、短语音、时变语音等因素存在的情况下，如何提高说话人识别的正确率。说话人识别在电话银行、安全监听、个性化应用等方面有着广泛的应用前景。

4　言语合成与语言表达

言语合成就是让机器像人那样的说话，实现自然的人机交互。言语合成的研究已有 200 多年的历史，经历了机械式、模拟、数字的不同阶段；从合成算法到系统集成，进而实现了从实验室研究到应用的进步。TTS 涉及到文本分析、合成算法、韵律控制等技术。研究的目标是合成语音自然且具有表现力。

4.1　言语合成技术与系统

当前典型的言语合成系统是 TTS(Text To Speech)，它涉及到文本分析、合成算法、韵律控制等技术。

(1) 文本分析。文本分析是 TTS 系统的前端。文本分析的主要任务是将以文字形式表示的文本转换为计算机可识别的读音表示，并对文本进行韵律结构预测。文本分析涉及语言学、语音学等多个学科的知识。文本转换的正确率直接影响了 TTS 系统输出语音的正确性，也影响语音的自然度。文本分析主要包括文本正则化、语法分析、韵律结构预测、字音转换等部分。文本分析主要涉及自然语言处理的技术，目前多使用基于规则或机器学习方法。

(2) 合成算法。合成算法的作用是将文本分析生成的读音表示生成语音波形。这是言语合成系统中的核心部分。在现在的言语合成系统中，主要的合成方法有拼接合成与参数合成两种[7]。拼接合成利用一个大语料库，从中选出合成语句中的每一个音节，再将它们拼接在一起生成合成语音。参数合成将语料库中的语音分解为声源与滤波器参数，利用隐马尔可夫模型(HMM)对这些参数进行建模；在合成时先进行参数的生成，再由参数生成语音。

拼接合成直接使用从自然语音中得到的语音片断进行拼接，因而可以获得较高的可懂度和清晰度。但主要问题是提高合成语音的自然度。这就需要对韵律模型、选音算法与语音修改及平滑算法、大规模语料库建设进行研究。从 20 世纪 80 年代开始，基于拼接合成的 TTS 逐渐成熟，并获得应用。

近些年，兴起基于 HMM 的参数化合成。在训练过程中，从语音中提取出基频和谱参数进行 HMM 参数化建模，并通过语境参数对模型进行聚类，得到一个语境相关的参数化模型。合成过程中，文本分析对输入文本进行语言学上处理得到每个训练单元的语境信息，之后从模型中选取合适的 HMM 序列，进行状态时长估计和声学参数生成，得到基频和谱参数序列，经过参数化合成器生成合成语音。

在参数合成的方法中，可以直接对生成的语音参数进行修改，从而可以更加方便地生成情感、个性

化的语音。参数合成的这一优势，使其成为了目前言语合成研究的重要方向之一，而且参数合成占用资源少，适宜用于嵌入式平台、手机系统中。

(3) 韵律控制。生成自成的韵律是提高合成语音自然度的关键。语言中的韵律信息还包含了如情感、态度个性化等信息。韵律研究是一个复杂的系统工程，涉及到语言学、语音学、心理学、语用学等学科的综合知识。目前采用的韵律参数主要是音高、音长、音强等参数以及它们的分布规律，研究目标是要生成自然语流的重音和语调。在基于隐马尔可夫模型的言语合成系统中，韵律模型与声学模型一起训练，使用参数化模型表示，可以较好地实现韵律参数的预测。

4.2 言语合成的展望

言语合成的发展方向及研究热点主要在(同音色的)多语言合成、情感言语合成、可视言语合成等方面，从而实现更加和谐的人机语音交互。

多语言合成主要面对现在的国际化环境，主要面临的挑战是在没有多语言的发音人的情况下，如何生成同一音色的不同语言的合成语音。现在主要解决方式是跨语言的自适应[8]。情感言语合成主要面临解决的问题是情感的分类定义、情感韵律的生成等。由于很难得到情感状态下的大语料库，情感的言语生成多使用参数修改的方法[9]。可视言语合成可以在生成语音的同时提供说话的图像，这可以进一步提高人机交互的自然度。可视言语合成主要研究内容包括说话人唇形、脸部动作与表情、头部动作生成等[10]。

5 言语知觉计算

听觉是人类交流的“言语链”中的重要一环，在人类的生活中起着重大的作用。声波通过介质传到内耳，刺激耳蜗内的纤毛细胞而产生神经冲动。神经冲动沿着听神经传到大脑皮层的听觉中枢，形成听觉。听觉的研究与心理语言学、认知神经科学、听觉生理机制相关。言语加工的认知机制、听神经计算模型、言语知觉计算模型、言语声学特征分析、汉语知觉特性建模等是近年来言语知觉计算的研究热点。这些研究成果也有助于抗噪语音识别、言语合成、音频编码等。

在言语知觉计算的相关应用中，如何评价听觉系统功能，即听力评估，是重要的研究方向之一。纯音测听和言语测听是听力评估的重要手段。相对于采用单一频率刺激声的纯音测听，言语测听(Speech Audiometry)采用行为方法测量听觉系统对言语信号的察觉、分辨、识别和理解能力，更能反映日常生活交流中言语信息获得的障碍，在临床实践中更有实际意义和诊断价值[11]，因此受到言语声学、听力学、临床医学的广泛关注，在评价听觉中枢、语言中枢的功能，选择干预方案和评价康复效果等方面发挥着不可替代的作用。据世界卫生组织(WHO)估计，1995年全球听力残疾人数为1.2亿，到2000年超过3亿。在我国，听障碍的残疾人约2 700万，居残疾人总数的1/3。因此，为听力残疾人提供有效的言语听障评估，对于及时准确的诊断听力损失，评价残疾程度、社会交往能力、治疗或康复效果，进一步提高患者的生活质量具有十分重要的意义。

尽管言语测听在国外早已成为临床常规使用工具，但汉语言语测听在我国尚未得到普及，将汉语言语测听应用于信息化系统更是处于研究的起步阶段[12]。目前针对汉语言语测听的相关研究，主要集中在测试材料的设计、录制和等价性评估，以及言语识别率、言语识别阈的测试方法。与英语相比，汉语言语测听在理论和技术上存在着两个难以逾越的难题：1)如何分析研究汉语言语声学特征对听觉感知的影响？目前通过对言语测试表的等价性、音位平衡研究较多，而较少关注言语声学特征对言语测听的影响；2)声调是汉语区别于其它语言最显著的特点，如何描述声调特性对汉语辨义作用的影响？汉语是声调语言，因此人耳对响度的感知不仅受到能量的影响，也会受到音节的调类和调值影响[13]。在这两个技术难题的制约下，目前的汉语言语测听主要采用阶梯式降低言语声级的方法进行识别率测试和识别阈测试，将受试者听力障碍的病理和临床表现的差异进行了模糊，而差异性对评估和诊断却是至关重要。因此，在正确评价言语分辨能力和听敏度的同时，能够区分性的鉴别出受试者听觉言语功能的具体缺陷或者残障程度，提高汉语言语测听的信度(可靠性)，改善言语测听的效度(残障程度的评价、听觉言语功能缺陷的具体鉴别)，是汉语言语测听相关研究未来的发展方向。同时，利用计算机辅助技术开展言语测听信息化系统的研发[12]，将大大推动言语测听在临床的实际使用，促进言语测听的推广，并拓展在其他相关领域的应用。

在推广汉语言语测听的同时，进一步完善测试词表的设计、噪声下的言语测听与评价、针对性言语

测听与听力障碍评估等，将有望减轻医务测试员的工作量，为患者提供更有效的听力障碍评估，减轻患者的痛苦。另外基于言语信号处理技术，研究噪声抑制算法、助听器频响补偿方法、人工耳蜗编码策略[14]、语音信号增强算法等，将有益于听力障碍者的言语交流。

6 结语

作为中文信息处理领域活跃的研究方向，汉语言语处理在言语识别、言语合成、说话人识别、言语知觉计算等方面取得了一定的研究进展。未来言语信息处理将继续向集成化、实用化的方向迈进，在计算机辅助语言学习、高表现力可视言语合成、基于互联网的语音翻译与检索、汉语听障评估等方面会有更丰硕的研究成果。

参考文献

[1] Rabiner L, Juang B-H. Fundamentals of Speech Recognition[M]. Prentice Hall, 1993.

[2] Huang X D, Acero A, Hon H W. Spoken language processing: A guide to theory, algorithm and system development[M]. Prentice Hall. 2001.

[3] Liu L, Zheng F, Wu W. State-dependent phoneme-based model merging for dialectal Chinese speech recognition[J]. Speech Communication, 2008, 50(7): 605-615.

[4] Harrison A, Meng H, Lee P. Automated Feedback in Commercial Computer-Training Systems[R]. Dept. of SEEM, CUHK, 2009.

[5] Meng H, Lo W-K, Harrison A M, et al. Development of Automatic Speech Recognition and Synthesis Technologies to Support Chinese Learners of English: The CUHK Experience[C] //APSIPA 2010, Biopolis, Singapore: 2010.

[6] Wu W, Zheng F, Xu M, et al. A Channel Robust Speaker Verification Algorithm Using Cohort-based Speaker Model Synthesis[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(6): 1893-1903.

[7] Zen H, Nose T, Yamagishi J, et al. The HMM-based Speech Synthesis System (HTS) Version 2.0[C]// Sixth ISCA Workshop on Speech Synthesis. Bonn, Germany: 2007: 294-299.

[8] Qian Y, Xu J, Soong F K. A frame mapping based HMM approach to cross-lingual voice transformation [C]//2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011: 5120-5123.

[9] Chung-Hsien Wu, Chi-Chun Hsia, Chung-Han Lee, et al. Hierarchical Prosody Conversion Using Regression-Based Clustering for Emotional Speech Synthesis [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(6): 1394-1405.

[10] Jia Jia, Shen Zhang, Fanbo Meng, et al. Emotional Audio-Visual Speech Synthesis Based on PAD[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(3): 570-582.

[11] Hall J W, Mueller H G. Speech Audiometry[G]//J. W. Hall, H. G. Mueller. Audiologists' Desk Reference. Singular, 1996

[12] 黄高扬，贾珈，蔡莲红，等. 计算机辅助汉语言语测听软件的研究与实现[C]//第十八届全国多媒体学术会议. 2009.

[13] Ciocca V, Francis A L, Aisha R, et al. The perception of Cantonese lexical tones by early-deafened cochlear implantees[J]. The Journal of the Acoustical Society of America, 2002, 111(5): 2250-2256.

[14] 吴玺宏，李量，迟惠生. 汉语、英语听感知差异及适合汉语的人工耳蜗编码策略[J]. 中国听力语言康复科学杂志，2007，5：17-20.