# 声学的基础研究促进了通信技术的发展

李昌立†

(中国科学院声学研究所　北京　100190)

**摘　要**　在通信发展过程中,声学和通信始终紧密相关.近年来,在这个交叉学科的前沿领域,提出了很多新问题,其中最重要的,就是通信系统要融入更多的知识和智能,也就是要研究和破解人对声音感知和理解的物理、生理、心理过程,并应用到通信和信息系统中.中国的通信和信息产品要赶超世界先进水平,就要在这方面投入更多的科研力量.

**关键词**　听觉虚拟环境,人工智能,多媒体,耳蜗,鸡尾酒会效应

# Fundamental research in acoustics advances the development of communication technology

LI Chang-Li

(*Institute of Acoustics, Chinese Academy of Sciences, Beijing* 100190, *China*)

**Abstract**　During the development of communication technology, there has been much research on acoustics. In recent years, many new problems have appeared in this interdisciplinary science, including the very important question of how to embed artificial intelligence and a knowledge base into the communication system. Accordingly, the physical, biological and psychological processes of acoustical signal perception and understanding by human beings have to be studied. To realize state-of-the-art communication and information technology, more effort should be devoted to this field in China.

**Keywords**　auditory virtual environment, artificial intelligence, multi-media, cochlea, cocktail-party effect

## 1　引言

2005年,德国科学家简斯·布劳尔特(Jens Blauert)联合欧洲和美国的一些著名声学家,共同编写了一本书《通信声学》,阐述了声学的基础研究对现代通信和信息科学技术的推动和促进.本文支持这些观点,并结合我国情况,阐述作者的见解.

听觉是人类通信最重要的感觉模态.因此,在声学和通信发展的过程中,两者始终是紧密相关的.有3个重要的里程碑标志着它们的发展和相互促进:一是电子管的发明,使微弱声信号的放大成为可能,由此出现了无线电广播、有声电影以及扬声器扩声系统和遍及全世界的电话网络等;二是计算机在声学研究中的普及和声音信号的数字化处理,如语音处理技术、数字音响工程、听觉虚拟环境等;三是人工智能的融入,即进一步研究和破解人对声音感知和理解的物理、生理、心理过程,并应用到通信和信息系统中.这一趋势有可能成为通信领域未来几年研究的热点.因为现代的信息和通信系统,将会包含越来越多的内置智能和知识,如在语音技术中,可以把先进的语音识别和对话系统作为例子.新的算法将要沿着这条路线发展,而且将出现更多巧妙和新奇的应用.

## 2　从"鸡尾酒会处理器"和"虚拟环境发生器"的应用,来讨论听觉场景的分析与合成

术语"鸡尾酒会效应"表示这样的事实:具有健康

2010－03－24收到

† Email:li_chang_li_cn@hotmail.com

听觉能力的收听者，能够在一群同时的发音者中，把注意力集中到某一个发音者，并分辨出该发音者的语音，因为双耳听觉能够在一定程度上抑制噪声、混响和声音染色. 也就是说，使用适当的滤波算法处理这些信息，能够增强想要听的发音人的信号，并抑制不想听的发音人的信号. 更进一步，融入了专家系统并包含内置智能和知识的场景分析器，能够完成通常所说的“内容滤波器”的功能. 这些滤波器很适用于音像节目材料的自动存档和检索任务，它和 ISO/IEC（国际标准化组织）建议的 MPEG-7 有许多相同之处.

当一个礼堂或音乐厅还在设计阶段，人们就可以根据它的几何形状、建筑结构、材料的吸声性能和人在厅中的位置，使用电声器件和计算机来模拟它的音响效果. 如果不能满足人们的要求，可以修改设计，而不会造成浪费. 这就是“虚拟环境模拟”. 人们可以在虚拟环境下，获得和真实环境一样的感觉. 例如，维也纳的金色大厅并不是每一个人都有机会去欣赏，但是使用“虚拟环境发生器”，你就可以在世界上任何地方，获得亲临其境的真实感觉.

要达到上面的这些目的，就需要使用听觉场景分析（ASA）和听觉虚拟环境（AVE）. 听觉场景的分析与合成类同于声音传输，其目的是从空间上和（或）时间上的某一点，把声音传输到另一点，以至于在两种情况下的听觉感知完全一致. 要达到这种真实的重建，一个可能的方法是由双耳技术给出，它试图在收听者两耳的入口，真实地重建声音信号.

双耳技术如图 1 所示，它包括两部分：下面部分是双耳信号处理器的基本结构，上面部分是人工智能部分，也就是给系统安装了“大脑”. 双耳系统有两个前置端口，它从人（或假人头）的左右耳获取信号作为输入，经过中耳部分的一些不规则的带通滤波器之后，将两路耳信号送到耳蜗模块. 在这里首先需要将信号分解为适合于耳的频谱分量（即临界频带），然后再作压缩处理. 这种处理，左右两只耳朵都要分别执行. 这个模块在每一个临界频带内，分析左耳和右耳信号之间到达的时间差别和声级差别，并表示为每一频谱区域内左右耳之间的互相关函数，从这些互相关函数峰值位置和形式所给出的信息，可以识别不同的声源和它们在空间的位置. 这种输出信号是一组具有可变速率的神经脉冲串，最终得到“双耳活性图”（见图 1 中的双耳激活显示）. 目前，通过分析真实的听觉场景，已经能够开发出用参数表示的各种算法，人的某些分析和识别能力已经能够模拟乃至超越. 但要达到这些目的，需要使系统变得越来越智能化和知识化，要使系统中的部件能够访问它们的“大脑”，这不是自下而上以信号激励作为基础，而是由上向下在假设的驱动下工作. 在图 1 中，“双耳活性图”又作为预分段和特征提取的输入，它和“大脑”中的假设作比较，把形成的符号放进“黑板”中，不同的专家模块能够对它进行检查. “大脑”中的假设是要建立在双耳活性图的合理解释上，各种假设一步一步地评价，然后被修正，最终才得到听觉场景的合理参数. 也就是说，系统要和具有智能的专家系统结合起来，才能够完成听觉场景分析中的大量任务.

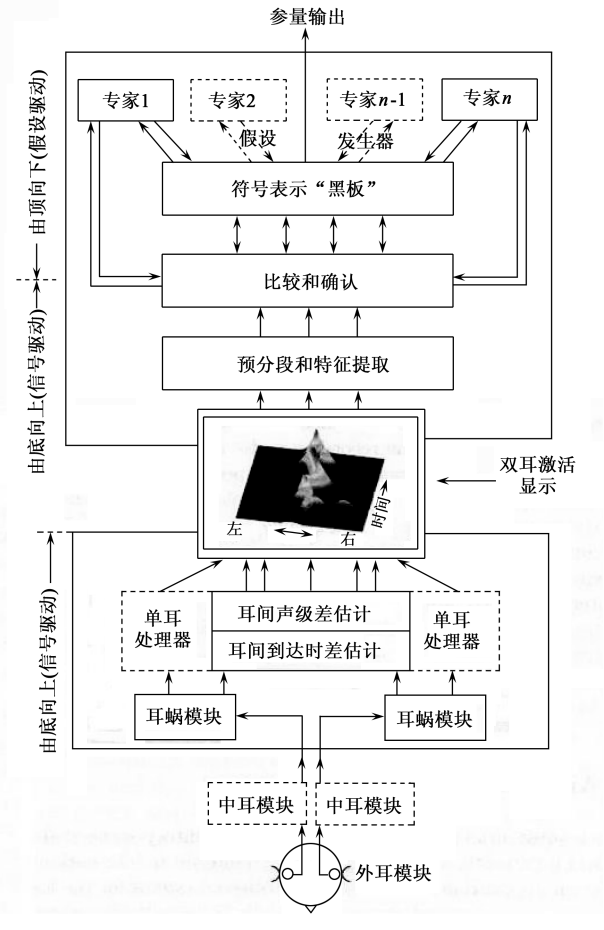


图 1　双耳信号处理模型的基本结构（增加了人工智能部分）

听觉场景的参量合成比它们的计算分析有更高的技术实用性，尤其是当收听者和合成的场景相互作用能产生效果的时候. 这些相互作用的听觉场景通常称为“听觉虚拟环境”. 由于听觉虚拟环境与通常的虚拟环境一样是人工的，即由计算机产生的，这些场景的参量表示可以包括空间，甚至包括和内容有关的方面，可使实际上驻留在不同位置的用户，在感觉上转换到一个公用的虚拟房间，例如远程会议，或参加执行远程操作. 进一步，人们可以进入一个虚拟的环境，检查它或访问环境中的目标，如虚拟

博物馆或虚拟旅游.由于虚拟空间的入口能够通过互联网提供,各种各样的应用是可以设想的,例如各种音乐演奏和演讲,远程会议,飞行员和消防队员的训练,特别是它允许一些复杂的科学研究实验方案得以灵活经济地实现.图 2 为听觉－触觉的虚拟环境发生器示意图.

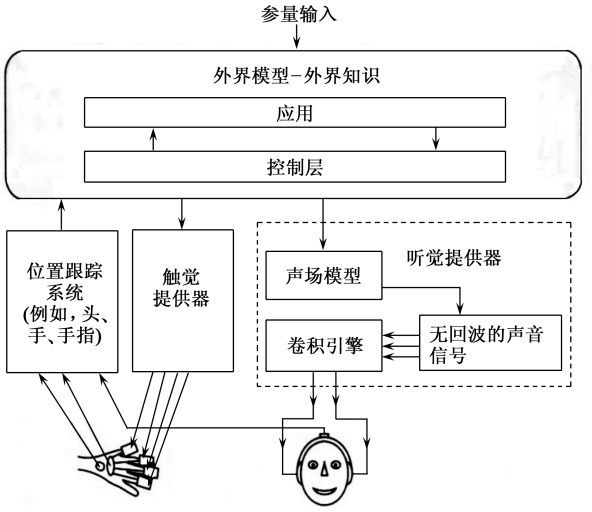


图 2　听觉－触觉的虚拟环境发生器示意图

## 3　研究动物特别是脊椎动物和人的听觉结构和机理,对通信体制和器件的发展会有深刻影响

对多种动物而言,声音通信起了重要作用.每种动物都能发展出一组声信号,这些信号并不限于反映动物的内部动机,而且也能够用它来做环境中的参考目标.为了通信,动物和人的听觉系统必须能够实现对各种信号的检测和分类,并且能够确定它们的来源.

例如,在马蹄形蝙蝠的回声定位中,其鸣叫频率表现出尖锐的调谐滤波器功能.它所用的鸣叫持续时间为 50ms 到 60ms,频率大约为 83kHz 的纯音,从开始到结束具有很短的频率扫描时间.在鸣叫频率上,它使用尖锐的调谐滤波器,分析回波的幅度调制和频移,能够对颤动的昆虫进行检测和分类,而这种回波是由运动昆虫的翅膀和多普勒效应产生的.在这种处理中,马蹄形蝙蝠甚至能根据自己的飞行速度,改变发出的鸣叫频率,使得回波总是落在蝙蝠听觉系统能够控制的频率范围内,从而能够检测出很小的频率变化,这种效应称为多普勒频移补偿.这种能力是现代雷达和声纳都难以相比的.

又如声音定位的精度是由耳间时间差和幅度差确定的.在那些比人头小得多的动物中,其定位精度预计不会比人好,而恰恰相反,谷仓猫头鹰尽管有如此小的头,对于 4—8kHz 的纯音,其定位精度在水平方向是 5°左右,垂直方向是 10°左右,它的定位精度比人要好.因此谷仓猫头鹰能够在完全黑暗中,只依靠它的听觉就能捕获鼠类动物.

人们还发现,某些树蛙在鸣叫时,会根据周围的声学环境调节能量消耗的方法.如果灰色树蛙中的雄蛙只是自己鸣叫,而没有任何竞争者时,它们会产生比有另外雄蛙竞争时消耗更少能量的信号,否则,会产生持续时间和速率都增加的信号.通过测量它们消耗的氧气,发现其差别为两倍以上.生长在婆罗州的树穴蛙,它在空心树干形成的空穴中鸣叫时,可以调节各自的鸣叫频率,使之与空腔的共振频率一致,它们 44%以上的振动能量都能够转换为辐射声能,比人的发声效率要高.

对人的听觉器官的研究,虽然比动物要深入,但目前仍然有很多问题没有解决.人的听觉器官是由耳和中枢神经共同构成的.人耳包括外耳、中耳和内耳.外耳和中耳的主要功能相当于传感器,有效地实现从空气到内耳液体中的声能传输.内耳包含了耳蜗,它的主要任务是对输入信号做频谱分析,并把它分解为很多并联的输出信号,每一种输出信号表示不同的频谱分量.这些不同的振动分量由很多内毛细胞(IHC)来感觉.这些内毛细胞被认为是耳蜗的输出口,这里的听觉信息是尖峰形状的电脉冲,它通过神经纤维传输.这个传输网络包含了一些被认为是中转站的核心,在中转站里,由不同神经纤维送来的信息借助专门的细胞参加处理;而最高级别是大脑的听觉皮层,它靠近脑的表面.

人们对听觉器官的结构和机理还不是了解得很透彻.通常认为听觉器官有惊人的信号处理能力,由于中枢神经结构的高度复杂性,所以目前还必须使用心理声学模型.当然,对听觉器官的结构和机理认识越透彻,所提出的模型也就越精确,越有实用价值.有人把听觉器官看成一台计算机,但听觉器官的功能远远超过计算机,并不是计算机的速度比不上人脑,而是目前还提不出完善的听觉器官模型和算法.众所周知,人的行为不能由声音信号直接指挥,而是由这些信号所传达的“意思”来指挥.因此,人是如何解释声音信号的?也就是说,对于人们,该信号的真正“意思”是什么?这是一个很复杂的问题,人需要知识和智能,并要求获取更多的信息,他们才能

在不同的背景情况和背景知识的状况下,各自做出判断.现代的通信设备也和人脑一样,需要智能和知识,才能发挥最大的效能.

例如早期的语音识别机器,只对外耳的声学特性和听觉系统中信号处理之类的问题建模,所用算法通常是严格地由信号驱动.但是这种策略不足以完成更复杂的任务.所以现代的语音识别机器除要求由信号驱动外,还要求由假设驱动,即由底而上和由顶而下的处理程序,因此要引入多种知识,如语义学网络、语言模型、文字模型、文法模型、句法模型、说话策略和语音学模型等.又如语音压缩编码,标准的波形编码速率是 64kbit/s;而使用参量编码,其速率最多也只能降低到 4.8kbit/s 或 2.4kbit/s.按照信息理论,汉语有 40 多个音素(符号),在正常情况下,谈话速率是每秒 10 个音素,再使用音素出现的概率表,可计算出人脑的语音信息传输速率是 50bit/s.目前各种语音压缩算法均不能达到这个目标,但人脑处理语音的速度,即听觉器官和中枢神经把语音波形转换为符号的速度,能够达到这个目标.

## 4 多媒体应用背景下的视听交互作用,能够改进通信设备的质量和效率

在自然环境中,我们通过多种感觉模态同时接收信息.这些刺激的特征是按照物理规律耦合,以致同一事件产生的听觉和视觉刺激,到达观察者时,具有特定的时间、空间和前后关系.例如,在语音情况下,可以看见嘴唇的运动和听见的发音会紧密地同步出现.现在,人们已进入到对多种感官的理解和研究.它对改进通信设备的质量和效率会有很大促进作用.

在人对语音的处理中,通常声音信号提供了足够的信息,可以得到很高的语音可懂度.但是在不同的声学条件下,像在鸡尾酒会上,附加的视觉信息(例如从发音人嘴巴运动所看到的结果)也能够对语音的可懂度有相当大的贡献.在背景噪声比较强的情况下,能测量到 40%—80%的改进,它取决于词汇表的规模.除此以外,语音刺激与嘴唇运动是否同步也会对感觉质量有很大影响,这对多媒体通信、动画片制作、电影配音都有重要作用.

在多媒体应用中,探讨音频和视频(AV)重建系统的感觉质量是很重要的.对于这样的系统仅仅知道声音信号和图像信号的感觉质量是不够充分的,人们还必须知道这两者相互之间是如何影响的,它们对 AV 整体质量是如何做出贡献的.有很多研究机构在高清电视框架内对此做了深入研究,如欧洲共同体的 MOSAIC 计划,艾恩德霍芬感觉研究所(IPO)和飞利浦研究实验室两者执行的合作研究计划,国际电信联盟 ITU“第 12 研究组”的研究计划等.这项研究早期工作是由贝尔实验室、瑞士通信实验室和英国电信实验室完成的.总之,听觉和视觉在感觉上能够相互作用的观察表明,在多媒体应用中,必须注意这种相互作用.随着多媒体应用进一步广泛,对于听觉和视觉相互作用的研究一定会更加深入.

## 5 通信系统的质量评价是一个值得深入研究的问题

最近几年,被传输语音的质量已变成了人们关注的新焦点.以前,电话语音的质量是紧密地和 300—3400Hz 带宽的模拟和数字传输通道标准相联系,两端是以传统形式的电话手机终止.由于通道特性有相对低的变化性,在这些网络上传送的语音质量比较稳定.当移动网络和基于 IP 协议的互联网络大规模建立的时候,这种情况就完全改变了.新的损伤形式是由低比特率编码技术和引起很大延迟的互联网和信号处理算法造成的,它们会产生回波和引起新的失真,使通信质量变坏.而公用接口的声音传输特性和传统的电话手机有很大的不同,如桌面或汽车驾驶环境的免提电话或计算机耳机.因此话音质量的评价就成为关键性问题.

在电信网络上,人对人的交互(HHI)和人对机器的交互(HMI),其质量不但和传输通道的特性有关,而且还和用户有关,即和人的主观感觉以及当时的语言环境有关.

在该种背景情况下,质量是感受或体验,也就是常说的“服务质量”.服务质量要从两个不同的观点探讨:第一个是服务提供者的观点,更精确地说,服务质量包含 4 个因素,也就是服务支持、服务操作能力、服务能力和服务的安全性;第二个观点是用户的观点,用户感觉到的这些表现,并把这些感觉和某些内部参考作比较,按照他们是否满意来判断它们.在研究服务质量的时候,重要的是要考虑上述两个观点.为此,必须建立两方面之间的关系:一方面是用

户的期望和要求以及他们的感觉;而另一方面是为了引起这些感觉,网络可能的响应特性.为了促进网络和服务的设计,以便于优化质量,需要建立用户和环境两者间的关系.感觉质量只能够用听觉事件来评价,被观测的项目,通常有可懂度、自然度、收听有效度、感觉噪声、语音质量或声音质量.而系统的参量和信号能够使用仪器测量,如在传输通道不同点的语音信号和噪声信号、传输路径的频率响应、延迟时间、噪声级和响度比等.

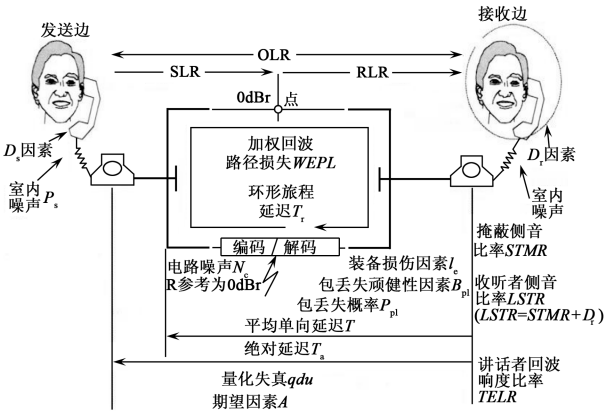


图 3 在人－人交互中,电话连接的参考模型,要详细了解见参考文献[3](图中 OLR 为发送边与接收边之间的距离)

图 3 表示在人－人交互中,电话连接的参考模型,这样的原理图已经得到国际电信联盟(I-TUT)的同意.这里使用 2/4 线模拟连接或数字连接,两边都用电话听筒作为终端.在通过网络的主要传输路径上,还会有电的侧音路径,即语音从发方和收方又返回到发音者和收听者,从而影响通信的质量.此外,电路噪声、线路延迟、AD/DA 转换和 PCM 所产生的量化噪声、丢帧丢包、设备损伤等都会影响语音的质量,图中都作了描述,可以定量计算.详细情况见文献[1] 第 135 页.

总之语音通信系统的评价是很复杂的问题,涉及到很多网络测量技术和心理声学问题,涉及到质量状况分类和预测模型等问题,目前还没有很成熟的一套方法.然而对它的深入研究,无疑会对语音通信系统的发展起推动和促进作用.

## 6 结论和展望

由于声音是人与人之间的通信中最有效和方便的方式,因此,在通信发展过程中,声学和通信始终紧密相关.尤其是最近 20 多年来,声音和图像信号的数字化和数字信号处理理论和技术的发展,给通信技术带来了革命性的变化.近年来,有一种发展趋势可能成为未来通信领域研究的热点,那就是通信系统要融入更多的知识和智能.我国实行改革开放以来,科学技术迅速发展,我国通信和信息产品的设计和制造水平已经进入世界先进行列,但主要体现在技术方面,在通信和声学交叉的前沿领域所形成的新思想、新概念和新体制仍注意不够,如果希望在通信领域有更多新的创造发明,就必须重视这些交叉学科并投入更多的科研力量.

### 参考文献[1)]

[1] [德]布劳尔特主编,李昌立,李双田译校.通信声学.北京:科学出版社,2009 年 4 月

[2] 李昌立,吴善培编著.数字语音:语音编码实用教程.北京:人民邮电出版社,2004 年 11 月

[3] Jens Blauert (Editor). Communication Acoustic. Berlin Heidelberg: Springer Verlag, 2005

1) 本文引用的参考文献太多(其中第 2 节有 71 篇,第 3 节有 107 篇,第 4 节有 91 篇,第 5 节有 48 篇),未能一一列出,只列出文献[1—3],其他的文献均可从文献[1] 和[3] 中找到——作者注