

语音一人与机器交流的又一桥梁

作者：王士元、彭刚

单位：香港城市大学电子工程系语言工程实验室

通讯地址：香港九龙达之路 83 号香港城市大学电子工程系

电子邮件：ewsyw@cityu.edu.hk（王士元） gpeng@ee.cityu.edu.hk（彭刚）

电话：（+852） 2194 2632

传真：（+852） 2788 7791

联系人：彭刚

送交：《科学技术回顾与展望》

2003 年 7 月 21 日

语音—人与机器交流的又一桥梁*

王士元、彭刚

香港城市大学电子工程系语言工程实验室

电脑业钜子比尔·盖兹曾说：“语音科技不但是 Windows 的未来，更是整个电脑界的未来！”。而语音科技最主要的两个方面即是语音合成和语音识别。语音科技成功运用的一个典范便是由美国 Computer Motion 公司开发的 ZEUS 外科医生系统。



图 1. ZEUS 外科医生系统 (摘自 <http://www.computermotion.com>)

卓越的语音控制能力使得外科医生能够用简明的语音命令来精确地控制内窥镜的移动，从而使外科医生的双手从操纵机械外科仪器的操纵杆中解放了出来。在这里我顺便提一下，在 ZEUS 中运用语音命令的思想便是我和该公司的创始人王友仑博士（本文作者王士元的次子）谈话时共同提出来的。

早在 1939 年，在美国纽约国际博览会上展示了由贝尔实验室 H. Dudley 制作 Voder，它是利用共振峰原理研制的早期语音合成器。当时由一个训练有素的操作员，就可以用 Voder 发出悦耳的“How are you?”来，在博览会引起了极大的轰动。其后，有关合成技术的研究得到了长足的发展。

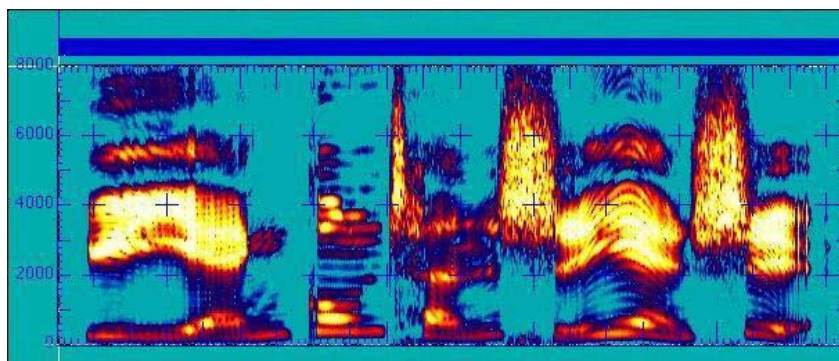


图 2. “语音工程实验室”的频谱图

从图 2 可见，在自然的连续语流中，音节之间的过度是相当平滑的。我在 1958 年便提出了如何用波形编辑的方法来合成清晰并且自然流畅的语音的这一构想[1]。其后，Holmes 于 1973 年提出了并联共振峰合成器和 Klatt 于 1980 提出的串/并联共振峰合成器。只要精心调整参数，这两个合成器都能合成出非常自然的语音。自八十年代末期至今，语音合成技术又有了新的进展，特别是基音同步波形叠加（PSOLA）技术的提出（1990），使基于时域波形拼接方法合成的语音的音色和自然度大大提高。在上世纪末，由于计算机存储能力的大大提高，又出现了一种基于语音数据库的语音合成方法，该技术的

*在此，我们衷心感谢香港城市大学及香港研究资助局 (RGC) 对我们研究工作的大力支持。

基本思想是根据合成的需要从大型语音数据库选取合适的音节或音节流。这种方法原则上只要语音数据库足够完备，就能合成出符合各种要求的语音。

相对来说，语音识别是比语音合成更富于挑战性的一个问题。我们在语音科技领域的研究也主要集中在这个方面。语音识别技术的早期研究对象主要是欧洲的语言。早在 1952 年，Bell 实验室的 K. H. Davis 等人就成功地设计了一个取名叫 Audrey 的英语数字语音识别系统。而 Audrey 这个名字便是由自动数字识别器(Automatic Digit Recognizer)美化而来。相对来说，汉语语音识别起步较晚一些。

在电脑界曾有过突出贡献的王安博士（王安电脑的创始人）和我在八十年代就已经对汉语语音识别有了一些讨论。我们当时从汉语的语言结构出发，得出汉语语音识别比西欧语言的语音识别更具优势。为此我们相当兴奋。但可惜的是王安博士于 1990 年因癌症抢救无效，带着遗憾走完了人生最后的历程。我们的这一构想一搁便是十余载春秋。直到 1995 年我来到香港城市大学，我发现香港的语言资源相当丰富。我就开始打算招收一些研究语音的博士生。张波是我在香港的第一个博士生，他系统地研究了抗噪语音识别，他的博士论文题目是：“抗噪聲語音識別中的一種基於可靠頻段的相似性度量”。本文的第二作者彭刚于二零零二年初在语言工程实验室完成了题为“語音識別中基於可靠性指標的韵律模型”的博士论文。毕业后，彭刚一直留在本实验室继续他的研究工作。

汉语是一种典型的声调语言。声调是构词必不可少的一部分。同一个音节如配上不同的声调，那么它的意思可以相差很大，甚至是完全不一样。普通话有五个声调（包括轻声）而广东话多到有九个声调（如把三个入声调并入到响应的非入声调，则只有六个声调）。

虽然一个语言或某种方言往往只有几个声调，但声调的生理实现却因发声人的不同而千差万别。为此，我们提出了自适应声调规格化方法。该方法的基本思想是将发音人的基频值（声调的生理对应）保存下来，当保存的基频值达到一定数量时，用这些基频值就可以比较客观地估计出该发音人的音域范围，进而可以非常有效地对他的声调进行规格化。而且我们也设计了一些算法来更新那些保存下来的基频值，这样就能更好地捕捉发音人音域范围的动态变化。我们在广东话不定人连续语音声调识别中采用了这种方法，实验结果表明以上方法显著优于其它方法[2]。

在我早先的一篇文章里，我从语言学的角度系统地阐述了基频信息的种种用途[3]。在语音识别，特别是声调语言的语音识别中，已经有很多方法描述如何在语音识别中运用声调信息，但它们皆有一个共同的致命弱点，那就是过于依赖声调识别的准确率。连续语流的声调识别本身就是一个高难度的问题，因而不易获得非常高的准确度。目前，普通话和广东话连续语流的声调识别率都大约在百分之七十的水平。为此，我们提出了一种把声调信息最为有效地运用到语音识别的并联声调协同（PTSA）体系[4]。这种方法的关键在于不过早地对声调识别的结果作出判定，而是把声调识别的中间结果和发声识别的中间结果并联起来，最后让语言模型来做最终选择。当我们把这种方法用于广东话不定人大词汇连续语音识别时，辨识错误率减少了近百分之二十五。而且这种方法也同样适应其它声调语言，甚至可以推广到非声调语言语音识别中重读音节的判定问题。

众所周知，现在是一个信息时代，信息的交流与获取显得尤其重要。语言是我们人类交流思想和信息的最重要的手段，如果我们能够在机器与人类之间成功地架起一座语言“桥梁”，那将引发一场新“工业革命”，最终引导我们人类进入一个全新的人性化的时代。

1. Wang W. S-Y. and Peterson G. E. 1958. Segment inventory for speech synthesis. *Journal of the Acoustical Society of America*, 30(8):743-746.
2. Peng G. and Wang W. S-Y. 2003. Tone recognition of continuous Cantonese speech based on supported vector machines. *Speech communication*. Submitted for publication.
3. Wang W. S-Y. 1972. The many uses of F_0 . In Valdman A. ed. *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*. Mouton. The Hague. 487-503.
4. Peng G. and Wang W. S-Y. 2003. Parallel tone score association for tone language speech recognition. *IEEE Transactions on Speech and Audio Processing*. Submitted for publication.