言语的感知、计算和可视

蔡莲红 陶建华 王志明 王玮

清华大学计算机科学与技术系 100084

摘 要

传统的中国音韵学和现代语音学紧密配合，给语音学及其应用带来勃勃生机。言语科学与计算技术的有机结合，推动了计算机智能的进步，但是目前还有诸多问题尚待解决，多学科的融合尚待加深。本文介绍了当前言语感知、言语计算和言语可视等方面的一些研究进展。

1. 引言

人类在用声音进行交流。言语的基本属性归感知范畴。传统语音学被称为“口耳之学”。声波刺激人的听觉系统，产生轻重缓急、抑扬顿挫的主观感觉。古代的学者无法揭示言语产生的科学原理，但是他们从知觉出发，归纳总结出一系列语音学知识。如对四声的描述：“平声哀而安，上声厉而举，去声清而远，入声急而促”。对声调与乐律的关系，指出“若以文章之音韵，同弦管之声曲，则美恶妍媸，不得顿相乖反”。现代语音学具有仪器设备、计算机，专家们指出言语四要素是音长、音强、音高和音色。将知觉进行计算，抽取声波的声学参数、计算声波的特征、归纳言语的规则、实现言语识别和言语合成。总之言语的知觉是基本、言语的计算是手段、视觉言语是辅助。我们期望将上述三方面紧密结合，探索言语的奥妙，为人类服务。

2. 言语知觉

人对声音的知觉，与众多因素相关，如听力、语境、知识背景等。这里仅介绍人的听觉特性对声音的知觉影响。

本课题得到国家自然科学基金项目（69875008）资助

2.1. 听觉特性

人的听觉系统能感知动态范围很宽的声音，也能感知到声音地细微差异：

- 人能感知声音的绝对声压大小、绝对音高、绝对时长。人能感知声音的声压级在30-100dB 范围内。对声音响度的感觉与声音的强度和频率有关。人听到同样响度的声音时，其声压级与频率的关系可以用等响曲线表示。感知声音的频率范围在 20Hz-20KHz 之间。人耳对时间分辨可短到 2ms，且与声音的频率和强度无关。
- 人也能感知到声压的差别、音高的差别、时长的差别。通常用将这种分辨能力用恰能分辨的差别（JND）或查阈（DL）来表征。查阈可以是绝对值，也可以是相对值。学者们研究了纯音和白噪声的查阈。人常遇到声音强度在 30-100dB 范围内，JND 约为 4-8%。人耳对频率的辨别与频率的绝对值、声音强度、声音时长有关。

然而言语是个复杂的特征系统。自然语流是随机信号，人对声音的知觉还受到以下因素的影响

2.2. 听觉的掩蔽效应（marking）

当两个响度不同的声音作用于人耳时，则响度较高的频率成分会影响对响度较低频率成分的感受。这种现象称为人耳的掩蔽效应。从频率角度来看，低频成分容易掩蔽高频成分。掩蔽效应使被掩蔽频率成分的听阈上升。听觉的掩蔽效应将影响声音的正确感知。另一方面，人们利用这种效应，设计了心理听觉模型，以此改善压缩编码的效果。

2.3. 听觉的残余效应（residual）：

声音知觉需要一定的时间。实验表明，听觉下降、听觉升起的时间约为 140ms。实验也表明，声音的延续时间在 0.2 秒以上，主观感觉的响度就只和它的强度有关；但如果声音再短，则响度不但与强度有关，也和长短有关；

对极短的脉冲声，响度与声音的总能量有关。这些结果说明听觉有一个积分作用。例如，有些音在连续语流中听起来的响度要比单独听时的响度大。这就提示我们在合成系统中如何根据语境选择选择适当的基元。

2.4. 听觉的溶抹效应（smearing）

让我们聆听一段连续语流，我们能够区分一个个音节，但在语音的波形图上，就很难判断音节或音素的边界；如果观察语谱图，就会发现，相邻音节结合得很紧密，甚至在边界处两个音重合在一起。如果两个共振峰差别较大的声音，将其拼接在一起。听起来好像有回声，或感觉第二个音抢在第一个音还没有完就发出来了。

3. 言语计算

计算机言语处理离不开计算。言语计算的任务是建立一种计算模型，让计算机象人们那样来感知和生成自然语言。当然要让机器真正理解自然语言，是一个非常困难的事情。它涉及到人类智能的本质的研究，因为人类的语言活动是人类极重要的制定和组织。对这其中的问题的探索需要多学科协作攻关。除了计算机科学和语言学之外，吸取认知科学、心理学、脑科学、哲学等学科的成果也是必不可少的。它是语音学、人工智能、计算机科学、语言学、心理学等互相渗透的综合性学科。

数据是明显的信息表示，而知识则是信息的含蓄表示。知识是这样一种信息，它是可以描述的，但是又不可能完全描述清楚。对于知识，可以分为两种类型：认知的（cognition）知识是一种形式化的知识，能用语言来描述。而感知的(perception) 知识很难用语言来描述。比如你可以体会多个讲话人声音的差别，但你很难清晰地描述之。

随着人机交互技术的进步，计算机将能看、听、说、学，并能用自然语言与人类进行交流。为此，语音和语言的研究日益受到计算机界重视。人们利用计算机分析语音的特征（时域特征、频域特征），进而进行韵律计算、感知计算；至于具体方法，有的采用规则的方法，也有的采用统计的方法；从特征计算，发展为进行特征预测。下面列举言语计算的新进展：

3.1. 基于神经网络的韵律模型

对汉语神经网络韵律模型研究较多的有台湾交通大学的陈信宏，中国科学院自动化所黄燕、计算所朱廷劭和声学所许洁萍等人。他们分别从不同的侧面对神经网络在汉语韵律中的应用进行了一定的研究。汉语的韵律特征表现为多层次性，且与语境密切相关。在分析和归纳汉语韵律受语境参数影响的情况下，我们研究并提出了一种适用于汉语 TTS 系统的神经网络韵律模型。针对汉语韵律表现出层次性的特点，提出了更适合汉语韵律处理的神经网络拓扑结构，并提出对其输出进行优化的方法。本文在神经网络内部引入了特殊的加权因子，从而使神经网络在汉语韵律的训练中，无论其收敛速度，还是效率都得到了较大的提高。另外，本文还利用了高斯参数分解方法，对神经网络的输出参数进行优化，一定程度上增强了网络的容错性。同时对汉语韵律特征受语境信息的影响，进行了一定的归纳和总结。提出了一种对汉语的音节基频曲线进行规格化处理的方法，该方法较为简洁，不仅适合于用大语料对神经网络进行训练，也非常适合汉语语音的基频控制。

3.2. 利用数据挖掘发现言语知识

由于语音数据具有自身复杂性、随机性、时序性和海量存储性等特点，一直以来语音数据的处理就是一个难点问题，现在伴随着语音数据存储容量的不断增长，也给语音学家总结相应的语音规则增加了难度，他们很难给出适用于全面的规则指导信息。如何在海量的数据中获得相应的有效信息，需要我们进行深入的研究。人工智能领域中的数据挖掘技术则可以在大量的数据中发现一些新颖的潜在知识信息，因此采用数据挖掘技术进行语音信号分析成为一条崭新的途径。

语音数据包含很多信息，如基频信息、时长信息、幅度信息、位置信息以及重音信息等，简单来说就是同一个音节在不同的语境中会表现出不同的信息特征，即不同的语境会使音节自身的属性信息发生变化。语音数据是一种时序数据，既一句话中音节的排列是有先后顺序的，音节所处的不同位置对于语句的分析起着至关重要的作用。同时相邻语音音节之间也存在很强的音联关系，所有这些信息特征对整个合成系统输出的可懂度以及自然度会产生很大影响。因此将数据挖掘技术应用于语音信号处理可以解决部分现阶段较难解决的语音技

术难题，同时尽可能减少人为经验因素对语音处理的影响，完成对语音处理从定性到定量的转变。

3.2.1. 时序关联规则模型获得汉语韵律参数之间的关联关系

韵律特征主要是指音节的时长、基频的包络变化、能量的变化及适当的停顿等众多参数属性，在这些属性中对合成系统的自然度影响最显著的是音节的基频变化和音长的变化。目前合成系统中的基频变化规律大多是根据语言学的研究得出的一些定性的描述，这些定性规则能够为合成提供一些参考，但是无法在合成过程中直接使用这些规则，而且这些规则也很难覆盖所有的基频变化现象，同时对这些规则的维护和完善也很困难，在具体应用中仍存在较大的不足。

数据挖掘技术中关联规则模型可以很好地发现数据项之间的存在的相互关系，同时有大量的挖掘算法可供选择，基于不同的挖掘对象可以采用不同的挖掘算法，因此基于关联规则的模型可以有效从大规模语音库中提取更为全面和准确的语音韵律相互关系。

通过对“熟语料”库中语句音节的基频数据和时长数据进行预处理，离散化成相应的属性值，获得前后音节的基频信息和时长信息之间的关联关系，从而加以指导合成系统的选音，满足在不同语境下音节参数的变化的需求。

3.2.2. 数据挖掘技术获得汉语韵律变化规律

基于语音合成中的基音同步叠加技术，可以利用数据挖掘技术进行韵律变化规律的学习，可以采用数据挖掘技术中的神经网络方法、数据项聚类以及粗糙集理论的有机结合进行综合评判，利用神经网络具有的自组织和自学习特性将经过聚类处理的语音基频数据和时长数据分别转化成神经网络的输入和输出节点，经过网络的学习来获得一些典型的基频曲线和时长的映射关系，由于神经网络自身理论存在不够完善的地方，可以辅助粗糙集理论进行适当的修正以获得期望的模式，在这些映射的基础上可以通过简单的变换即可获得典型模式，利用这些典型的模式就可以对基频的变化规律在定量的基础上从较高层次上进行韵律规则的研究。

3.3. 韵律的预测

韵律模型的建立，可以基于言语的声学参数直接建立模型，如基频数值或模式、音长模型、停顿模型等。在连续语流中，声学参数是韵律的表现，而韵律特征不仅与语法、词法、语用有关，而且与语句的长短、结构有关，甚至还与语义、情感意向有关，因此韵律建模是个相当困难的课题。

我们以自然语音为基础，曾研究了结构助词在韵律短语界定中的作用。助词是粘着于词或短语的一类虚词，除个别助词（如“所”）外，基本上都是后置的。结构助词“的”具有将前导成分和后继的“名词短语”结合成韵律短语的功能。韵律短语结合的紧密度与短语的长短有关。结构助词“地”具有将前导成分和后继的“动词短语”结合成韵律短语的功能。韵律短语结合的紧密度与“动词短语”是否含有宾语有关。

对于种种韵律现象，如不同语调中基频模式、轻重音中基频模式和音长模式等。应该利用已知文本，从语法入手，结合语义、语用信息，建立韵律预测机制。虽然现代文本分析技术还很难从深层次中获取语句或篇章的完整信息，但是基于语法结构建立声学参数的韵律模型的思路也是可行的。由于语法结构的变化毕竟非常复杂，声学参数的任一细微变化，可能相应的组合的可能将非常多。因而，采取分层韵律建模的思路，将在一定程度上降低韵律建模的复杂度。其思路为，首先利用文本分析得到的分词、注音和词性等一系列结果，建立语法结构到韵律节奏的模型，包括韵律短语预测、重音预测等。继而，进一步通过重音和韵律短语信息，并结合成统一的语境信息实现韵律声学参数的预测和进行选音的步骤。

例如，韵律短语预测的主要方法有：统计模型，既利用韵律边界相对语法结构的统计结果，来进行韵律短语边界的预测；另一种方法是采用分类决策树技术的规则学习算法，该方法利用语料中手工标注的韵律边界信息建立语法信息到韵律短语边界的预测模型。

其中，规则学习法，如决策树方法，该方法简捷，且可以融入一定的人工经验，因此，采用决策树方法对韵律重音和短语界定实用且有效。这种方法基于大量的已标注了分词、拼音和词性的语料的统计，通过实现根据经验提出的大量问题集而自动训练决策树，并使用训练后的决策树对给定的文本进行节奏的预测，其结果比传统方法有较大的改善。在汉语韵律短语的预测中，决策树节点的特征向量包括了

多个数值型或分类型的变量，从这些变量中提出了诸多问题。这些问题包括了两种类型，即枚举型和数值型。无论何种问题，其答案都是二值化的。这些问题的合理选择能够使得构造的决策树生长平衡。

可见，在上述统计方法中，语料的标注信息主要是语法信息，没有包括语义信息，另一方面，也没有包括情感意向对韵律参数的影响，因此，有必要进一步完善的特征矢量，改进算法。

4. 言语可视

4.1. 言语波形

言语波形是言语信号的直接显示。它表示言语信号的幅度随时间变化的波形形状。但是相同的声音可能是不同的波形；形似的波形可能是不同的声音。因此我们很难根据波形分辨声音。

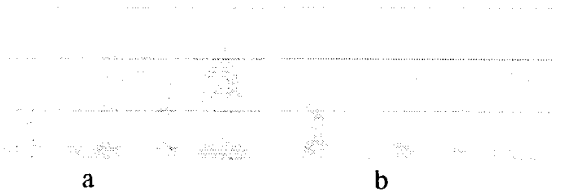
图 1：一段语音波形示意图

4.2. 言语频谱

频谱是言语信号的一种图形表示。它把声音信号转换成可视图谱。言语信号由许多频率成分组成。用幅度和相位表示每种频率成分，称为幅度谱和相位谱。幅度谱是常用的。

语谱图又可分为窄带语图和宽带语图。窄带语图显示信号的动态谐波结构，及显示各次谐波随时间的变化，宽带语图是用较宽的滤波器计算出来的。用它可以观察元音的共振峰和辅音的强频区。

人们可以看语谱，辨认声音，分析共振峰的变化，也可以利用语谱参数实现言语识别。



a　　b

图 2：a、窄带语图，b、宽带语图

4.3. 言语视位

很久以来，人们就注意到我们对语言的理解是多模态的。在许多场合中，我们不只用耳朵去听声音，而且用眼睛去观察说话人的面部运动表情。人们说话时复杂多变的面部表情不仅可以传达丰富的感情，而且可以增强对语言的理解。随着人们研究的不断深入和许多实际应用的驱动，新的国标准 MPEG-4 提出了视位（Viseme）的概念，它由英文的 Visual 和 Phoneme 两词拼接而成。MPEG-4 对视位的定义是：Viseme is the physical (visual) configuration of the mouth, tongue and jaw that is visually correlated with the speech sound corresponding to a phoneme. 即视位是指人们发某一音位时相对应的嘴、舌头、下腭等可视发音器官的物理形状。近几年来，对视觉语音的应用研究越来越受到人们的重视，已成为一个多媒体和人机交互技术研究领域相当活跃的研究方向。

言语视位包括以下几个方面的内容：

静态视位： 与某一音位对应的口形，即 MPEG-4 所定义的 Viseme，它代表了人们发某一音位时最基本的可视特征，可以用一幅图片来表示。因为许多发音动作、口形是相似的，静态视位与音位存在一对多的关系，如汉语声母/b/、/p/、/m/可以用同一个静态视位来表示；另一方面，同一个音位处在不同的语境中其口形也是不同的，如汉语声母/d/处在拼音组合/du/和/da/中口形有很大的差别，因此也存在由音位到口形的一对多的映射关系。现在的 MPEG 标准仅定义了静态视位（Static Viseme），但同时也指出不排除将来定义其它类型的视位。

动态视位： 反应人们说话过程中口形的变化规律，它以静态为基础，表达了连续语流中口形如何从一种口形过渡到另一种口形，以及协同发音对口形的影响；它可以由一个图像序列来表示。了解连续语流中口形变化规律对虚拟人脸合成和唇读都是很重要的，人们曾经试过用各种各样的函数模拟口形的变化过程，包括线性函数、三角函数、多项式函数以及指数函数。

表情视位： 是人们内在情感在面部的表现，如眉毛、眼睛、面部肌肉的动作。人们说话时的情感不仅表现在语调、语速上，也表现在面部表情上。MPEG-4 定义了六种基本的表情：高兴（Joy）、悲伤（Sadness）、生气（Anger）、害怕（Fear）、厌恶（Disgust）、吃惊（Surprise）。实事上，人的面部表情是千变万化的，各种表情也往往是掺合在一起的，如何统一的描述这些表情仍是一件有待深入研究的问题。

MPEG-4 定义一组人脸定义参数 FDP（Facial Definition Parameter），用来定义一个

人脸模型；一组人脸动画参数 FAP（Facial Animation Parameter），用来描述人脸各特征点的运动，以上这些视位都可以通过这些标准的FAP 参数来表示。

言语视位研究有着广泛的应用，主要包括以下几个领域：

识别：研究表明，不只是聋哑人，普通人在感知语音时也在一定成度上利用了视觉信息。有的声音在听觉是很容易混淆(如/bi/和/di/)，但因为它们在发音时口形有较大的差别，如果观察说话者的口形就很容易把它们区分开来。在环境噪声环境下利用视觉信息可以提高语音识别率,有研究表明，在有噪声的环境中，看到说话者的脸相当于语音的信噪比提高了约8-12dB。

合成：在人机交互的过程中，如果在给出声音信息的同时能给出一个“讲话的头”（Talking Head），表现出说话者面部表情和嘴部、眼部等变化情况，则会大大改善人们对声音的理解，方便人们和计算机的交流。在环境噪声较大的情况下，如在工厂车间、高速运行的交通工具上或战场前线进行人机交互时，合成人脸能提高人们对语音的理解。有调查表明，在许多商业和公共场合人机交互中，如果人们面对的不是单纯的文本，而是一个会说话的人物形象，则使人觉得计算机界面更为友善，带来一定经济效益。

编码：多媒体信息的巨大数据量要求人们寻找更高效的数据压缩算法，在可视电话、电视会议等应用环境中，图像的主要变化正是集中在说话者的唇部，了解口形变化的规律及口形与语音的内在联系将可能使我们更好的预测口形的变化，从而对人脸图像进行高效的参数化编码，大大提高图像数据的压缩率。

另外，视位的研究也在许多其他领域得到了广泛的应用，如音视频联合身份识别、帮助聋哑人学习等等。对言语视位的深入研究，将有利于我们更好的了解人类是如何理解言语的，也有利我们研制出更友善、更方面的新一代人机交互界面。

5. 参考文献：

[1] 蔡莲红等. 汉语文－语转换中的语言学处理. 中文信息学报. Vol9,No.1

[2] 应宏，蔡莲红，基于助词驱动的韵律短语界定的研究，中文信息学报，1999

[3] 陶建华、蔡莲红等，“汉语 TTS 系统中可训练韵律模型的研究”，声学学报,第 26 卷,P67-72

[4] Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. 1996. A stochastic finite-sate word-segmentation algorithm for Chinese. Computational Linguistics, 22(3).

[5] 王志明，蔡莲红，汉语文本-可视语音转换的研究，小型微型计算机系统（待发表）

[6] 王玮，蔡莲红，关联规则在汉语词属性中的应用，计算机工程与应用，Vol.37 No.5，P17-18,2001.3

[7] 王玮 蔡莲红，基于粗集理论的神经网络计算机工程，Vol.25 No.5，P65-67,2001.5

[8] Sin-Horng Chen et al., “An RNN-Based Prosodic information Synthesizer for Mandarin Text-to-Speech”, IEEE Transcations on Speech and Audio Processing, VOL.6, NO.3, 1998,5.

[9] International standard, Information technology-Coding of audio-visual objects-Part 2: Visual; Admendment 1: Visual extensions, ISO/IEC 14496-2: 1999/Amd. 1:2000(E).

[10] Bothe, H.H. and Wieden, E.A., A neurofuzzy approach for modeling lips movements, IEEE World Congress on Fuzzy Systems, 1994. Proceedings of the Third IEEE Conference on Computational Intelligence, Vol1, P234 –237, 1994.

[11] 吴宗济等，实验语音学，高等教育出版社

[12] 马大猷等，声学手册，

[13] 吴宗济，“中国音韵学和语言学在现代言语处理中的应用”