汉语疑问语气的声学特征研究*

100084
jdn00@mails.tsinghu.edu.cn, clh-dcs@tsinghua.edu.cn

摘要

Fisher

1. 引言

语音信号中不但包含与文字相对应的表意信息，还包含着丰富的表情信息。由于表达时语气、情感等方面的差别，文本相同的语音可能具有不同的含义和语用功能。因此，表情信息在语音交流中具有重要意义。然而，当前的计算机语音技术在处理语音中所包含的表意信息的同时，忽略和丢弃了大量的表情信息。例如：语音识别系统将语音转化为文字，但损失了与说话人说话方式、态度倾向、情感状态等相关的信息；大多数的语音合成系统只能产生朗读语气的合成语音，远远不能合成自然的、具有不同语气、表达多种情感的语音。这些问题大大降低了人机语音交互的效果和效率，严重限制了人机交互的自然性。

解决上述问题的基础是分析在语音的声学信号中，哪些声学特征与表情信息相关，并建立两者之间的映射模型。考虑到在语音合成和识别两方面应用的需要，应该分别从感知的角度和统计分类的角度建立最优的映射模型。显然，这两个模型之间可能存在着差别。

本文着重分析汉语中，与疑问语气相关的声学特征。在此方面，已有的研究工作大多从语调研究的角度，定性地分析语气不同所引起的语调上的变化，如 Garding (1987: 2)、沈炯（1983：5；1994：6）以及 Shen (1989: 4)等人的工作。在这些工作中，研究者通常精心设计少量典型的、具有相似韵律结构和声调组合的分析语句，提取其声学参数，在不同语气的语调之间加以观察、比较，最终得出结论。这些研究工作已经得到了一些重要结论，为探寻汉语疑问语气在声学上的表现做出了可贵的贡献。然而，仅仅对少量语句的声学参数加以测量、比较，无法反映其在语气感知和语气分类方面的作用，因此也无法获得声学特征和语气之间最优的映射模型。

由此，本文的工作采用了新的研究方法。一方面，为研究声学特征在语气感知方面的作用，利用语音合成方法，有选择性地将疑问语句的某种声学特征或声学特征组合复制到相应

* 本文的研究工作受国家 863 高技术项目（2001AA114072）和自然科学基金（60275014）资助

的陈述语句中，同时保持该陈述语句的其他声学特征不变。随后将得到的合成语音作为听觉感知实验的刺激材料，通过分析感知实验的结果研究该声学特征在疑问语气感知方面的作用。另一方面，从统计分类的角度，在较大规模语料上提取声学特征，在进行了针对说话人的归一化处理之后，由 Fisher 线性判别准则度量各声学特征在不同语气之间的区分特性，从而得到其在语气分类方面作用的相关结论。

本文的具体安排如下：第 2 节介绍研究中所采用的语料，第 3、4 节分别阐述感知实验以及统计分类研究中采用的具体方法，实验过程和得到的实验结果、结论等。

2. 分析语料

本文所研究的语句均为单句，具有自然的语义和声调组合，并且在其文本中不包含表达疑问的特殊结构，能够同时以陈述语气和疑问语气自然表达。在录制语料时，要求录音人分别以两种语气表达相同的文本，并且保证以两种语气表达的语句之间具有相同的韵律结构和重音分布。这样，陈述语句和与其相对应的疑问语句之间仅存在语气的差别，而在声调、重音、韵律结构等方面的特性是一致的。实验中采用的文本共有 135 句，录音人为 2 人（1 男，1 女），则共得到 540 句分析语料。

录制的语料被保存为 16khz 采样，单声道，量化比特数为 16bits 的 wave 文件，随后通过语音分析软件 Speech 对其进行音节边界的标注。Speech 能够自动地提取音节边界，同时又提供了可视界面，允许实验者对自动标注的结果进行手工修改。在标注音节边界之后，利用基频分析算法 Yin [Cheveign (2002 :1)] 估计每个音节的基频曲线，同时还对每个音节的能量曲线以及时长参数进行了估计。

3. 感知实验

3.1. 听觉刺激材料

本文在感知实验中采用的文本共有四句，其句末音节分别具有不同的声调：

语句 1：路边停着一辆汽车。

语句 2：前面是一条河。

语句 3：大家选我当代表。

语句 4：这种事发生过。

本文将上述四句所对应的疑问语句的不同声学特征（或声学特征组合），通过 TD-PSOLA 方法复制到相应的陈述语句中，同时保持其他声学特征不变。将所得到的合成语句作为感知实验的听觉刺激材料，则可通过分析感知实验的结果研究特定的声学特征及声学特征组合在语气感知中的作用。本文所研究的声学特征包括基频、能量、时长的全句平均特征以及参数曲线的包络特征。在感知实验中，听觉刺激材料分为两组。第一组刺激材料为分别单独修改上述六种声学特征之后的合成语句，第二组为同时修改多种声学特征（即声学特征组合）之后的合成语句。

3.2. 被试

被试为 9 名听觉正常的研究生，其专业均不是语音学及语言学专业。

3.3. 实验任务和过程

实验任务是要求被试分辨每个语句所表达的语气，按照与疑问语气的相似程度评分为 1~4 分。分数越高，则说明该语句越接近疑问语气。1 分表示陈述语气，4 分表示疑问语气。同一语句允许被试反复听多遍，被试的反应时间没有限制。

3.4. 实验结果

表 1 列出了分别修改六种声学特征后，被试对各语句所包含语气的平均评分。表 1 显示，当修改基频曲线包络之后，合成语句的平均评分最高，为 2.83 分；其次为平均基频，但其平均评分相较基频曲线包络有明显下降，为 1.94。能量、时长的相关特征评分均小于 1.5，较为接近 1。可见，基频曲线包络在疑问语气感知中起着最为重要的作用，而其他特征本身对于疑问语气的感知作用较小。

特征	句 1	句 2	句 3	句 4	平均
平均基频	1.78	2.00	2.33	1.67	1.94
基频曲线	3.11	3.67	2.11	2.44	2.83
平均能量	1.11	1.67	1.22	1.33	1.33
能量曲线	1.22	1.00	1.78	1.56	1.39
平均时长	1.00	1.00	1.44	1.11	1.14
时长曲线	1.11	1.11	1.33	1.33	1.22

表 1：分别修改六种声学特征之后，被试对合成语句所包含语气的平均评分。

表 2 显示了在修改基频曲线包络之后，同时修改平均基频，以及能量、时长特征之后的平均评分。需要说明的是，由于本组听觉刺激材料所修改的特征较多，因此与表 1 所列出的结果相比，对单独修改基频曲线包络的合成语音的评分有所下降，但各语句评分之间的相对关系不变，与表 1 相比均下降了大约 0.5 分左右，这说明表 1、表 2 的结果是具有一致性的。

表 2 的结果表明，若在修改基频曲线包络的同时修改平均基频，各语句的评分均有所提高；若在此基础上修改能量和时长特征，则评分能够进一步提高。这说明平均基频和能量、时长特征对于疑问语气感知具有一定的辅助作用。

特征	句 1	句 2	句 3	句 4	平均
基频曲线	2.67	3.00	1.67	1.89	2.31
基频曲线+平均基频	3.00	3.45	2.11	2.11	2.67
基频+能量+时长	3.11	3.78	2.56	2.33	2.94

表 2：同时修改多种声学特征之后，被试对合成语句所包含语气的平均评分。

表 1、表 2 的结果同时表明，声学特征对语气感知的作用与句末音节的声调相关。无论是单独修改基频曲线包络，还是在此基础上同时修改平均基频和时长、能量特征，句 3、句 4 的评分都要大大小于句 1、句 2 的评分，而在上述情况下，句 2 的评分在各语句中总是最高的。这可能是因为当末音节声调为三声、四声时，声调的基频曲线与疑问语气中上扬的基频曲线相抵触，因此基频曲线包络的作用不如末音节声调为一声、二声时明显。而当末音节为二声时，其本身上升的基频曲线帮助了疑问语气的表达。

3.5. 实验结论

本文通过感知实验得到的主要结论有：

（1）、基频曲线包络是疑问语气感知中最为重要的声学特征。

（2）、在修改基频曲线包络的前提下，同时修改平均基频和能量、时长特征能够帮助疑问语气的感知。

（3）、声学特征在疑问语气感知中的作用与语句末音节的声调相关。

4. 统计分类研究

4.1. Fisher 线性判别

本文采用 Fisher 线性判别函数度量声学特征在两类语气间的区分特性，并以此估计两者之间的线性分界面。Fisher 线性判别方法所要解决的基本问题是找到一个方向的直线，使得特征在这个方向上的投影分开的最好。Fisher 准则函数定义为：

$$J(w)=\frac{(\tilde{m}_1-\tilde{m}_2)}{\tilde{S}_1^2+\tilde{S}_2^2}$$

其中，$\tilde{m}_1$、$\tilde{m}_2$ 分别为两类特征在 w 方向上投影后的均值，$\tilde{S}_1^2$、$\tilde{S}_2^2$ 为相应的协方差矩阵。则 Fisher 准则函数的分子表示投影后两类之间的离散度，分母表示两类内部的离散度。显然，应该寻找使得分子最大，分母最小，即 $J(w)$ 取得极大值的方向 w^*。由 Lagrange 法可求得：

$$w^*=S_w^{-1}(m_1-m_2)$$

其中，m_1、m_2 分别为两类特征的中心，S_w 为总类内离散度矩阵，即两类特征的协方差矩阵之和。

在将原始特征投影到 w^* 方向之后，可采用贝叶斯决策规则，以高斯模型等概率模型估计类分布概率密度函数，从而获得一种在一维空间的“最优”分类器。也可简单地设置一个阈值 y_0，通过判断投影特征与 y_0 的相对大小关系进行分类。

4.2. 声学特征

与感知实验中所研究的声学特征相对应，本文在统计分类中同样研究基频、能量、时长三种声学参数的平均特征和曲线包络。由于分析语料中包含不同说话人的语音，因此在提取声学特征之后，还需要根据说话人进行归一化，即：

$$x'=(x-\bar{x})/S$$

其中，x 为某声学特征，x' 为归一化之后的该特征，$\bar{x}$、S 分别表示相应说话人所有语料中该特征的平均值与标准差。

在提取反映参数曲线包络的特征时，由于基频和能量参数曲线受到清音段的影响而存在间断，因此首先通过三次样条插值的方法产生连续的参数曲线，将其在全句范围内归一化，并由三次样条平滑消除曲线中小的抖动，得到平滑的参数曲线 I。在此基础上，将 I 通过截止频率为 0.5hz 的高通椭圆滤波器得到曲线 I_1，I 与 I_1 的差为曲线 I_2，如图 1 所示。则 I_1 曲线更多地反映了基频曲线在局部的特性，而 I_2 曲线更多地反映了基频曲线在较大范围内的、缓慢的变化特性 [Mixdorff (2000:3)]。

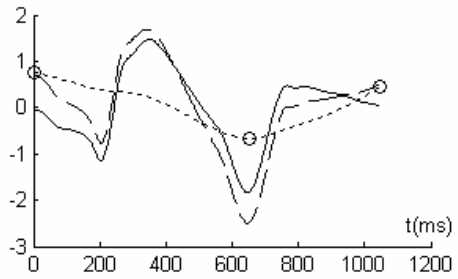

图 1：基频参数的 I、I_1 和 I_2 曲线。其中实线为 I 曲线，虚线为 I_1 曲线，点线为 I_2 曲线，I_2 曲线上的小圆圈表示极点。该语句的文本为“你不喜欢”，语气为陈述语气。

随后，本文在相应的 I_2 曲线上提取反映基频和能量参数曲线包络的特征，包括整条曲线的线性拟合斜率 $F0k$ 、 Ek ，以及最末一个极点之后的曲线斜率 $F0Finalk$ 、 $EFinalk$ 。对于时长参数曲线，提取其线性拟合斜率 Dk 。

4.3. 实验结果

表 3 列出了各种声学特征在两类语气间的区分特性。其中， $J(w^*)$ 为特征在最佳分类方向上的 Fisher 准则函数值， Ar 为训练集上的平均分类正确率。表 3 显示，平均基频 $\bar{F}0$ 在两类语气间具有最好的区分特性，其 $J(w^*)$ 和 Ar 的数值远远超过了其他特征。 $\bar{E}$ 和 $EFinalk$ 特征也具有较好的区分特性。然而，反映基频曲线包络的特征 $F0k$ 和 $F0Finalk$ 的区分特性却较差，这可能是由于 $F0k$ 和 $F0Finalk$ 更多地受到句子的声调、重音分布等因素的影响而降低了其在两类语气之间的区分特性。

声学特征	$J(w^*)$	Ar
$\bar{F}0$	3.0246	91.08%
$\bar{E}$	0.7874	79.74%
$\bar{D}$	0.4830	70.82%
$F0k$	0.4629	73.42%
$F0Finalk$	0.1312	58.36%
Ek	0.1223	62.83%
$EFinalk$	0.7962	78.44%
Dk	0.1826	62.83%

表 3：各声学特征在最佳分类方向上的 Fisher 准则函数值 $J(w^*)$ 和训练集上的分类正确率 Ar 。

表 4 列出了基频、能量、时长特征及其特征组合的区分特性。可见，基频在语气分类中的作用最为重要，能量和时长特征也具有较好的区分特性。同时使用三类特征能够改善两类语气间的区分特性。

特征	$J(w^*)$	Ar
基频	4.2302	93.49%
能量	1.4662	83.27%
时长	0.8011	75.09%
基频+能量+时长	5.5266	94.98%

表 4：基频、能量、时长特征及其特征组合在最佳分类方向上的 Fisher 准则函数值 $J(w^*)$ 和训练集上的分类正确率 Ar 。

本文已经通过感知实验发现，声学特征对语气感知的作用与语句末音节的声调相关。为研究声学特征在语气分类中的作用是否也与此因素相关，图 2（见下页）显示了当语句末音节为不同声调时的分类情况。可见，当末音节声调为一声时，与其他三个声调相比，基频特征对于语气分类的作用最大，而能量、时长特征对于语气分类的作用较小；当末音节声调为三声、四声时，能量、时长特征在语气分类中的作用相对较大。这说明声学特征在语气分类中的作用也与语句末音节的声调相关。

4.4. 实验结论

本文通过统计分类研究得到的主要结论有：

（1）、在所有的声学特征中，平均基频特征是在两类语气间区分特性最好的特征。

（2）、在基频、能量、时长三类特征中，基频特征在语气分类中的作用最大，但能量和时长特征本身在两类语气间也具有一定的区分特性。同时使用三类特征能够提高分类正确率。

（3）、声学特征在语气分类中的作用与语句末音节的声调相关。

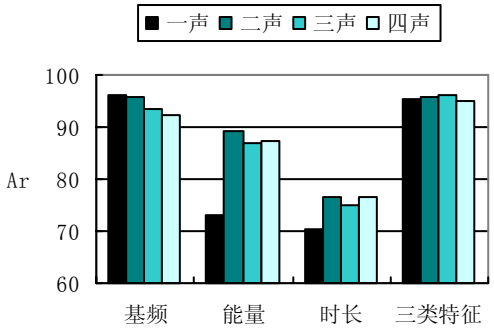

图 2：当语句末音节为不同声调时，使用基频、能量、时长及其组合特征的分类正确率。

5. 结束语

本文采用感知实验和统计分类的方法，分别研究了声学特征在汉语疑问语气感知和语气分类方面的作用。得到的主要结论有：

（1）、在语气感知中，基频参数曲线的包络起着最为重要的作用。而平均基频和能量、时长特征对语气感知起着一定的辅助作用。

（2）、在语气分类中，平均基频特征在两类语气间具有最好的区分特性。能量和时长特征本身也具有一定的区分特性。同时使用三类特征能够提高分类效果。

（3）、无论在语气感知还是在语气分类中，声学特征和语气之间的映射关系均与语句末音节的声调相关。

本文的研究还存在着一些局限之处。例如：对基频曲线包络的研究还比较粗略，在感知实验中，没有进一步研究其在句中不同位置、不同重音条件下的情况；在统计分类中，提取的反映基频曲线包络的特征还比较简单，并且由于受到其他因素的影响而导致其在两类之间的区分特性较差。在今后的工作中，作者将继续深入地研究上述问题。

参考文献：

[1] Alain de Cheveign, Hideki Kawahara, 2002, Yin, A Fundamental Frequency Estimator for Speech and Music. Accepted by *J. Acoust. Soc. Am.*

[2] Eva Garding, 1987, Speech Act and Tonal Pattern in Standard Chinese: Constancy and Variation. *Phonetic* 44. 13-29.

[3] Hansjorg Mixdorff, 2000, A Novel Approach to the Fully Automatic Extraction of FUJISAKI Model Parameters. International Conference on Spoken Language Processing (ICSLP), Beijing, China.

[4] X. Shen, 1989, *The Prosody of Mandarin Chinese*. University of California Press.

[5] 沈炯，1983，北京话声调的音域和语调。见《北京语音实验录》，林焘、王理嘉等编，北京：北京大学出版社。

[6] 沈炯，1994，汉语语调构造和语调类型。《方言》第 3 期，221-228 页。