汉语声母感知清晰度计算方法的探讨

贾珈，王永鑫，张逸嘉，田业，蔡莲红

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），
清华大学计算机科学与技术系，北京 100084

摘要 人耳特别是患耳对于声母的感知能力往往弱于对韵母的感知。研究汉语声母的感知特性，对临床听力评估具有重要的现实意义。本文将人耳对不同声母感知的难易程度定义为该声母的感知清晰度，并探讨了汉语声母感知清晰度的可计算问题。具体为，从言语声学特征的角度，挖掘汉语不同声母的关键区分特征；通过言语测听实验，分析关键区分特征与声母感知清晰度之间的相关性；进而采用线性回归策略，提出了一种基于关键区分特征的声母感知清晰度的计算方法。实验结果表明，该方法可以为规范听障言语测听的语料提供必要的技术支持。

引言

听觉是人类交流的“言语链”中的重要一环。人耳对声母、韵母和声调的感知共同构成了汉语音节的辨识与辨义。杨玉芳先生通过心理学实验，说明了音高和时长是影响音节感知的主要特征[1]。由于声母时长较短、能量较低，人耳特别是患耳对于声母的感知能力往往弱于对韵母的感知。因此，研究汉语声母的感知特性，对临床听力评估具有重要的现实意义。

传统音韵学通过发音部位和发音方法对声母进行分类，并通过统计方法和心理学实验进一步研究声母的感知特性。张家騄、吕士楠先生等利用多维分析方法对汉语辅音的知觉混淆数据进行处理，得到了辅音知觉结构，提出了影响辅音感知的主要因素依次是：清音-浊音、送气-不送气、摩擦-非摩擦，并揭示了人耳对于不同声母感知的差异性[2-4]。在汉语言语测听（Speech Audiometry）研究领域，基于上述研究成果对言语测试表的等价性、音位平衡研究较多[5,6]，而较少关注言语声学特征对听觉感知的影响。如何将语音感知与声学参数相结合，从而量化评价人耳对不同声母的感知能力，是本文探讨的问题。

我们将人耳对不同声母感知的难易程度定义为该声母的感知清晰度，并探讨利用声学特征表征声母感知清晰度的计算方法。首先，提取汉语声母多维时域和频域特征，挖掘不同声母的关键区分特征；进而通过言语测听实验，分析关键区分特征与声母感知清晰度之间的相关性；最后采用线性回归策略，提出了一种基于关键区分特征的声母感知清晰度的计算方法。通过本文提出的声母感知清晰度的计算方法，有助于规范临床汉语言语测听的语料录制与选取，提高汉语言语测听的测试信度，同时可为人机交互中语音信息的感知和理解提供必要的技术支持。

1 声母声学特征的提取

1.1 语料库及标注

本文研究使用的语料库是解放军总医院用于临床言语测听的标准语料库。其中发音人为男性的专业播音员。语音的采集频率为 44100Hz，其中有汉语单音节字 551 个。剔除掉零声母后，其余的 470 个音节基本涵盖了汉语中的全部 21 种声母（b，c，ch，d，f，g，h，j，k，l，m，n，p，q，r，s，sh，t，x，z，zh），及声韵母组合。

语料库通过清华大学计算系人机语音交互实验室自主研发的 VisualSpeech 软件进行声母起止边界的人工标注[7]。标注过程综合了波形变化、频谱变化和人工听测，具有很好的切分合理性。

1.2 时域特征提取

本文提取声母的时长、短时平均过零

率、短时平均能量、最大短时平均能量与能量均值的比值，最小短时能量与能量均值的比值。对于每个声母，得到一组 5 维时域参数。

1.3 频域特征提取

由于后接韵母会对声母频域特征提取的准确性造成影响，本文通过语言信号重采样，在声母信号尾部加零值，一次移动一个采样点取得分析窗，提高频率分辨率，提取了声母的频谱。

1.3.1 Mel 倒谱系数

Mel 频率倒谱系数（Mel Frequency Cepstral Coefficient），又简称为 MFCC。人耳的听觉系统是一个非线性的系统，人耳对于不同频率的信号响应程度不尽相同，呈现对数关系。MFCC 能够较好的反映人耳对于语音的实际感知情况。本文选取 12 个滤波器组，为每个声母提取 12 维 MFCC 系数。

1.3.2 Bark 频带能量比率

Bark 是依据人的生理特性描述人耳听感的前 24 个关键频带。由于人耳对频率的感知集中在 0-8000Hz 的范围，本文使用 Hertz 描述的 24 个频段中的前 21 个频带[8]。

声母的语谱图中蕴含大量的信息，在提取频带特征时，需要将语谱图进行压缩，进而提取其中与听感最为相关的声学特征。Bark 频带能量比率描述的是某一个频带占全部频谱能量中的比例，是将整个语谱图以频率坐标轴按 Bark 频带来切分，然后计算切分出的频带能量占整体能量的比率。

本文采用的 Bark 频带能量比率提取步骤为：

步骤 1：将语音信号分帧后求 FFT 能量谱；

步骤 3：令 $x_1, x_2, ..., x_{21}$ 分别为声母 FFT 能量谱中，每一个 Bark 频带（共 21 个）的累计分量和；

步骤 4：计算频带 *i* 占全部频带能量的比例：

$$y_i = x_i \Big/ \sum_{j=1}^{21} x_j \qquad (1)$$

步骤 5：得到 $(y_1, y_2, ..., y_{21})$ 作为声母的 21 维 Bark 频带能量比率参数；

这个方法将原来的 0-8000Hz 的语谱图压缩到 21 维参数。由于切分频率轴的方式是依据人的听感特性，得到参数即为与听感相关的特征参数。

1.4 特征归一化

在进行进一步的实验之前，需要把提取的时域和频域特征进行归一化，将声学特征值归一化到[0，1]区间，这样声学特征具有相同的数量级和量纲。处理的公式如下：

$$f_{new} = \frac{f_{source} - \min(F_{source})}{\max(F_{source}) - \min(F_{source})} \qquad (2)$$

其中，f_{new} 是归一化处理后的特征参数，范围在[0,1]之间。f_{source} 是处理之前的特征参数，F_{source} 为原始特征参数集合，即同一维度特征参数的未处理的特征参数集合，$\min(F_{source})$ 为原始特征参数集合中的最小值，而 $\max(F_{source})$ 为原始特征参数集合中的最大值。

经过特征提取，语料库中的每一个声母对应于一个 38 维的特征向量，包括 5 维时域参数、12 维 MFCC 系数和 21 维 Bark 参数。相同的声母具有不同的韵母和声调组合，所以语料库中一个声母包括多个样本。

2 关键区分特征的选取

2.1 声母层次聚类

首先定义一种距离，使得不同的声母样本之间的距离度量能够反映不同声母在人耳听感上的差异。欧氏距离在数学上可以直观描述人类感知物体之间的差异。因此，本文采用欧氏距离度量特征向量之间的距离。而对于不同的样本集合之间的距离，采用类中心距离来衡量。

为考察不同声母的关键区分特征，对语料库中的声母样本进行层次聚类，层次聚类的剩余类数设定为 1。层次聚类的流程如下：

步骤 1：初始设定每一个样本点为一类，即：$S_i = \{X_i\}$，其中 S_i 为第 *i* 个分类，X_i 为第 *i* 个样本的特征参数；

步骤 2：将所有的分类中距离最近的两个类归并为同一类；

步骤 3：重复步骤 2，直到只剩下一个类别。

通过层次聚类构造一棵叶子节点是 21 个声母的聚类树。从树的根节点向下观测是层次分类，从叶子节点向上观测则是层次聚类。

利用每个声母提取的 38 维特征向量，通过层次聚类方法，得到结果如图 1：

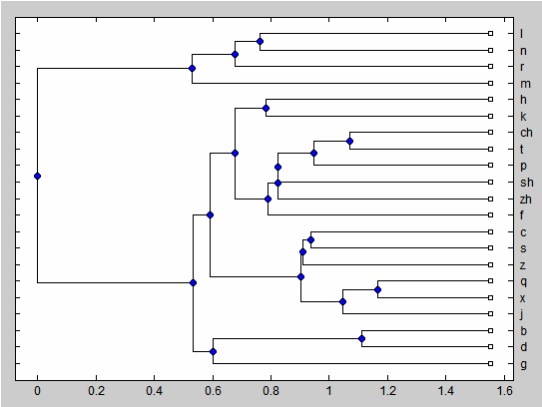

图 1 38 维特征分类树

从根节点的方向向下以分类的方式看得到的分类图，第一次分裂将清浊声母区分开。易混淆的声母{b，d，g}、舌尖前音{z，c，s}、舌面前音{j，q，x}、舌根音{h，k}进而分别被分裂到了同一个类别中。该结果显示，本文提取的 38 维参数对不同声母具有很好的区分性。但是分类结果中出现的{ch，t，p，sh，zh，f}类别缺乏合理解释，说明 38 维参数中存在冗余，需要去除与声母区分性无关的特征。

2.2 关键区分特征筛选

进一步分析不同类别之间的关键区分特征，去除冗余特征。对于任意两个类别（例如清声母与浊声母），通过依次考查 38 个维度中任意一维特征，并计算和该特征相关的以下 4 个指标，选取区分贡献度较大的声学特征，从而得到这 2 个类别的关键性区分特征。

指标 1：类间距离变化。类间距离是层次聚类中用于分类的唯一特征，该指标是关键区分特征的主要判断依据，删除某一特征后对于类间距离的影响较大，则可以认为该特征实际在分类中起到了较大贡献。

指标 2：类内距离变化。该指标衡量某一类别内部的发散程度，如果去掉某一特征后类内距离增大较多，该维度特征为关键区分特征。

指标 3：参数均值。比较两个类别某一维度特征的取值的差别，相差较大则表明特征有一定的合理性

指标 4：参数方差。该指标衡量某一维度特征在一个类别中的方差，即同一类别中该特征的发散程度。

通过计算类间距离变化、类内距离变化、参数均值、参数方差，得到各类的关键性区分特征如下（以下用 M 代替 MFCC 参数、B 代替 Bark 参数）：

关键特征组 1：平均过零率，M1、M3、M4、M5、M8，B19、B20、B21

该组特征可以较为有效的区分清声母和浊声母。清声母的短时平均过零率明显高于浊声母，而且清声母高频部分的能量分量所占能量比率明显多于浊声母高频部分所占比率，所以 B19、B20、B21 对于清浊区分有较好效果。

关键特征组 2：平均过零率，M1、M2、M3、M6、M7，B16、B17、B19、B20、B21

该组特征能较好的将舌尖前音{z，c，s}、舌面前音{j，q，x}与其他清声母区分开。声母 z,j,t,h 的语谱图如图 2：

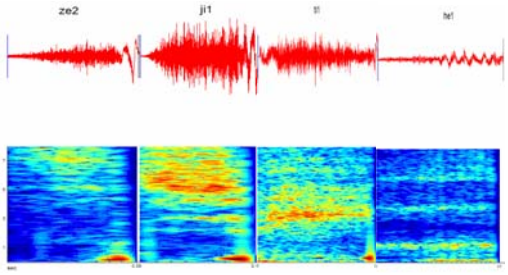

图 2 声母 z,j,t,h 语谱图

从图中可以看出，舌尖前音 z 和舌面前音 j 有一定的相似之处，但是舌尖前音和舌面前音与其他清声母的高频区域有较大区别，所以 B16、B17、B19、B20、B21 对于将舌尖前音和舌面前音与其他声母区分开有较好的效果。

关键特征组 3：M1、M2、M3、M4、M5、M6、M8，B14、B16、B18、B19、B21

该组特征能较好的将舌尖前音{z，c，

s}与舌面前音{j，q，x}区分开，舌尖前音与舌面前音虽然语谱图较为相似，但是高频区域有部分频带分量不尽相同，所以B16、B19、B21有较好的区分效果。

综合以上3组关键特征组，得到17维关键区分特征，包括：

时域参数：平均过零率；

频域参数：M1,M3,M4,M5,M8,M11,B9,B10,B12,B16,B13,B19,B20,B21,B14,B18；

采用2.1节的层次聚类方法，用17维关键区分特征向量代替38维原始特征向量，得到新的层次聚类结果如图3：

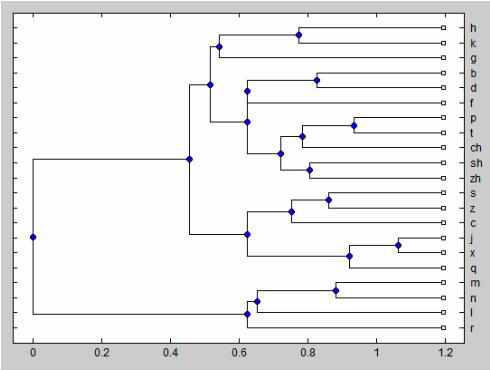

图 3 17维特征分类树

该聚类树显示，21个声母被分为7类，分别是浊声母{l,m,n,r}、舌根音{h,k,g}、舌尖前音{z,c,s}、舌面前音{j,q,x}、{b,d,f}、{p,t,ch}、{sh,zh}。

3 关键区分特征与听觉感知的相关性分析

为分析关键区分特征对声母的区分性是否与人耳对声母听感的区分性一致，本文采用临床言语测听实验进行验证。

实验共邀请20名被试，在天津听力障碍专科医院专用隔音室中进行。实验使用清华大学计算机系人机语音交互实验室与解放军总医院共同研发的《汉语普通话言语测听系统》[9]对被试分别进行纯音测听和言语测听。纯音测听结果表明20名被试均为听力正常人。言语测听按照标准的临床言语测听流程进行。将本文使用的语料库中单音节编排为25个音节长度的测试表，每名被试测试4张测试表。测听中记录被试的反馈，如果被试将声母A错听成声母B，则记录A错听成B。除此以外的情况（例如某个声母被试反馈没有听到）不予以记录。测试声强设定为被试单音节言语识别率测试所得到的PI曲线得分约50%时的声强，约为15dB。该实验得到被试声母听感混淆矩阵如表1所示，矩阵中每一列表示该声母错听成其他声母的总次数。

将混淆矩阵中的21个声母按图3叶节点所示的分类方法分为7个易混淆类。通过计算每一类的类内平均混淆度，和各类之间的类间平均混淆度，考查采用17维关键区分特征得到的声母分类是否与实际听感测试的结果一致。

指标1：类内平均混淆度。同一类别中的某个声母与该类别中其他声母的听感混淆概率的平均值，公式如下：

$$P_{within}=\frac{1}{|A|}\frac{\sum_{i,j\in A,i\neq j}x_{ij}}{\sum_{i\in A,j\in U,j\neq i}x_{ij}} \qquad (3)$$

指标2：类间平均混淆度。同一类别中的所有声母与不属于这个类别的其他所有声母的听感混淆概率的平均值，公式如下：

$$P_{external}=\frac{1}{|A^c|}\frac{\sum_{i\in A,j\in A^c}x_{ij}}{\sum_{i\in A,j\in U,j\neq i}x_{ij}} \qquad (4)$$

其中U为全体声母的集合，集合A表示某个易混淆类，集合A^c表示A在全体声母集合U中的补集，x_{ij}表示表1中声母i被错听成声母j的频次，|·|表示集合中元素的个数。计算结果如表2所示。

表2结果表明，对于所有声母的分类，其类内平均混淆概率都远远大于类间平均混淆概率。例如{j,q,x}类别，其类内混淆概率约是类间概率的12倍。

该结果说明，通过17维声母关键区分特征得到的声母分类结果，与言语测听得到的声母听感混淆的结果具有较高的符合性。因此17维声母关键区分特征与人耳对声母的听感有较为密切的相关性，可以用于量化描述人耳对不同声母的感知能力。

本文将人耳对不同声母感知的难易程度定义为该声母的感知清晰度。基于17维声母关键区分特征，本文进一步探讨度量不同声母的感知清晰度的方法，进而揭示声学特征参数与声母感知清晰度之间的关系。

	l	m	n	r	h	k	g	z	C	s	j	q	X	b	d	f	p	t	Ch	sh	zh
l		23	5	24	2	1	2				4			25	37	8	3	1	1	7	9
m	4		5	1	1		1		1					65	5	16	6				
n	6	38			1		3			1				7	41			1			1
r	13		5				9	3	3						7					1	30
h	2		1			68	101	7	2	2				7	3	2	1	30			4
k				1	14		18	1							1			9	1		
g				2	12	3		4	1					8	49		1	5		4	17
z	5		9	10	4		9			12				1	49	3		1			3
c					9	1	9	14		19										1	3
s				1	1	1	5	19	4						3			5			5
j	7			1								9	31	24	25		2	12		1	4
q	1									1	56		75					10		1	4
x				1							112	14		2	4		1	5	2	37	2
b	23	36	2			1	15								17	134	5	2	1		
d	4	12				1	4							140					3	17	18
f														51	4	1		20			
p					35	37	13	2	5		3		4	7	22	9	40		4		
t					2	1	1		5					8		1		21	2	1	1
ch			1	2			5	11				1	1		8					36	12
sh																			22		47
zh														1						14	

表 1 声母听感混淆矩阵

	类内 平均混淆度	类间 平均混淆度
b, d, f	0.1520	0.0302
l, m, n, r	0.0726	0.0417
h, k, g	0.1890	0.0241
j, q, x	0.2230	0.0184
z, c, s	0.1100	0.0372
p, t, ch	0.0976	0.0393
zh, ch	0.1631	0.0355

表 2 声母听感类间及类内平均混淆概率

4 基于关键区分特征的声母感知清晰度计算

基于 17 维声母关键区分特征，本文提出一种量化描述汉语声母感知清晰度的计算方法，其数学描述为相关声学参数的线性组合，如公式 5 所示。其中，*Index* 表示为感知清晰度，*C* 为常数项，*CrossZero* 为过零率，*a* 为过零率的系数，Bark 参数和 MFCC 参数分别取加权和。若相关维度 Bark 系数或者 MFCC 参数无效，即不在 17 维关键特征范围内，则将系数取为零。

$$Index = C + a * CrossZero + \sum_{i=1}^{21} b_i * Bark_i + \sum_{j=1}^{12} c_j * MFCC_j \quad (5)$$

基于该公式，本文尝试根据实验语料各声母的实际声学特征取值与言语测听得到的听感混淆数据，通过线性拟合算法，给出该公式各系数取值的参考值。具体为步骤为：1）提取实验语料库中各声母的 17 维关键区分声学特征；2）计算 21 种声母的每一种声母其每一维度关键区分特征的平均值，即对于每一种声母，得到一组 17 维的关键区分特征的平均值向量；3）根据实验语料库中各声母的平均混淆数据，通过线性拟合算法，给出到该公式系数的参考值列表为：

变量	系数参考值
CrossZero	-0.0571
$Bark_9$	0.0678
$Bark_{10}$	0.0374

$Bark_{12}$	0.0896
$Bark_{13}$	-0.0462
$Bark_{14}$	-0.0924
$Bark_{16}$	-0.170
$Bark_{18}$	-0.00738
$Bark_{19}$	-0.0721
$Bark_{20}$	-0.114
$Bark_{21}$	-0.105
$MFCC_1$	-0.261
$MFCC_3$	0.0117
$MFCC_4$	0.0434
$MFCC_5$	-0.131
$MFCC_8$	0.0531
$MFCC_{11}$	-0.0348
C	0.898

表 3 回归系数参考值列表

拟合结果中各维关键区分特征的系数的正负取值揭示了该维声学特征参数与声母感知清晰度之间的关系。例如，过零率越大，感知清晰度指数越小。这一点可以比较浊声母{l，m，n，r}和{c，s，z}可以看出，浊声母的过零率较小，而{c，s，z}的过零率较大。相比较而言，{l，m，n，r}类别比{c，s，z}类别更不容易类内混淆。而实际测听结果中，{l，m，n，r}的类内平均混淆度为 0.07，{c，s，z}的类内平均混淆度为 0.11，与拟合后的结果相符。再例如，Bark9，Bark10，Bark12 与感知清晰度指数是正向关系，这 3 个参数对应的频谱范围是[920,1080]Hz，[1080,1270]Hz，和[1480,1720]Hz。该结果表明，声母在这 3 个频带上的能量越大，人耳就对这个声母更加容易分辨清楚。同理，Bark16、Bark19、Bark20、Bark21 与感知清晰度指数有明显的负相关关系，这 4 个参数对应的频谱范围是 [2700, 3150]Hz，[4400,8000]Hz。由于人耳对于高频的感应能力相对较低，如果一个声母在高频区域聚集的能量越大，相对于低频区域聚集能量较大的声母，人耳更不容易分辨，就更加容易与其他声母混淆。以上结果也与汉语声母感知的定性和定量研究结果相符合[10]。

5 结论

本文探讨了人耳对声母感知的可计算性问题。从言语声学特征角度，提取声母语音信号的时域和频域参数，筛选关键区分特征；进而结合实际测听得到的听感混淆数据验证了所选取的关键区分特征与人耳感知的密切关系；最后通过线性拟合，设计了一种计算声母感知清晰度的方法。基于本文实验语料库，给出了系数参考值，并对系数参考值作出了解释和分析。

本文探讨声母感知清晰度的计算方法，其意义在于，目前汉语言语测听的相关研究对言语测试表的等价性、音位平衡关注较多。本文进一步提供了从声学特征的角度量化规范言语测听语料的录制与编排的技术支持。例如，采用该方法计算不同声母音频信号的感知清晰度，选取感知清晰度较高的文件编排言语测听语料，将会有助于提高汉语言临床语测听的测试信度。

参考文献

[1] 杨玉芳．言语知觉研究．应用声学，1997(3)：1-53.

[2] 张家騄．汉语人机语言通信基础．上海科学技术出版社，2010 年.

[3] 张家騄，齐士铃，吕士楠．汉语辅音知觉结构初探．心理学报，1981(1)：76-85.

[4] 吕士楠，张家騄，齐士铃．汉语辅音知觉混淆研究中的多维标度方法．声学学报，1981(6)：363-370.

[5] 张华，王硕，王靓等，普通话言语测听材料的数字化录制与等价性分析，临床耳鼻咽喉头颈外科杂志，2006.20(22): 1011- 1015.

[6] 郗昕，言语测听工具的效度、信度与敏感度，中华耳科学杂志. 2008.6(1): 1-6.

[7] 蔡莲红，黄德智，蔡锐．现代语音技术基础与应用．清华大学出版社，2003 年.

[8] 付强，易克初．语音信号的 Bark 子波变换及其在语音识别中的应用．电子学报，2000(10)：102-105.

[9] 贾珈，郗昕，黄高杨等，计算机辅助汉语普通话言语测听系统，2009SRBJ6657.

[10] 黄高扬，贾珈，蔡莲红. 基于声学特征分析的汉语韵母感知度量的研究.第九届中国语音学学术会议(PCC 2010).

本文研究受国家自然科学基金（61003094）的资助。