语音学知识在语音识别中的应用：案例分析

曹剑芬[1]，李爱军[1]，胡 方[1]，张利刚[1.2]

1.中国社会科学院 语言研究所；2.天津大学 计算机科学与技术学院

文 摘：计算机自动语音识别中的热门话题之一就是怎样利用语音学的知识来提高识别的正确率。在早期的数字语音识别中，2 与 8 经常容易混淆，曾经被视为一个难题。本文试图通过对这个具体案例的分析，探讨语音学特征知识在识别中的应用问题。本文采用声学和生理实验以及感知实验相结合的方法，探讨了 2 与 8 的区别性语音学特征及其在二者识别中的作用。结果表明，2 与 8 的最大差异是声调；在缺乏声调信息的情况下，第三共振峰的差异是决定性的区别特征；而它们的第一和第二共振峰非常近似，在识别上没有太大作用 。早期的自动识别恰恰忽略了声调这个最最显著的区别特征；而在自然语流中、尤其是非正式的语体中，有些 2 发音时舌尖运动不够到位，因而导致它与 8 的第三共振峰差异不十分明显，这些是识别中二者常常混淆的主要原因。由此可见，在自动语音识别中，加强对语音学特征知识的了解是个迫在眉睫的任务，在系统中充分地综合利用这些区别性特征信息，是提高识别率的有效途径。

关键词：计算机；自动语音识别；语音学知识；感知；声调；基频；声谱图；共振峰

中图分类号：TP 3；H017

在自动语音识别（Automatic Speech Recognition, ASR）的框架中利用语音学知识增加识别的效率一直是语音识别领域的重要问题之一。Ken N. Stevens 教授一直研究发音人之间的语音的不变量问题，以及各种音段的区别特征束[1, 2]。Lee Chin-hui 教授也一直提倡改进语音识别的框架模型，利用语音学的知识提高识别率[3]。Victor Zue 教授在他的对话系统中也充分重视语音学的特征[4]。此外，Chen, K. 和 Hasegawa-Johnson 建立了一个依赖于韵律的词和音素的隐马尔科夫(HMM)识别模型，结果取得了出众的、超过不带韵律的基准系统词识别的正确度[5]。Sarah Borys 采用随韵律而定的音位变体模型的言语识别，说明韵律因素在音素建模及其在言语识别应用方面的重要性[6]。Taehong Cho 等考察了韵律域起首增强的声学结果在口语词识别中的作用，证明跟音段发音增强相关的语音细节特征在词的识别方面具有重要的解歧作用[7]。王作英和肖熙等在汉语的语音识别中，利用了音段时长信息，大大提高了识别率[8]。但是，总的说来，怎样在汉语自动语音识别的框架中利用语音学知识仍然是个新的课题。

近年来，ASR 中面临的一些瓶颈问题，如口音问题，发音方式的问题等，使人们充分认识到统计模型和语音学知识结合的重要性。

在早期的汉语数字的语音识别中，2 与 8 经常容易混淆，成为一个难题。本研究试图通过解剖这个“麻雀”，探索语音学的特征知识在语音识别中的应用前景。首先，分析比较 2 和 8 的语音声谱特征，寻找 2 与 8 易混的原因；然后，通过人对 2 与 8 的语谱特征的辨识试验和各种听辨实验，分别考察声调、第一第二共振峰和第三共振峰等不同的语音学特征在 2 与 8 识别中的作用。希望通过这个典型分析，说明语音的变化虽然错综复杂，但是仍然有特征可依、有规律可循。在自动语音识别中，如果充分考虑和合理利用语音学的特征和规律，必将有益于识别率的提高。

基金项目：中国社会科学院语音与自然话语处理重点学科项目

1 2 和 8 声谱的区别性特征

从普通话语音识别数据库中选取了 43 个发音人的单音节“2” /er4/和“8”/ba1/进行声谱分析和测量，发现两者的声谱图（spectrogram）有以下几个区别性特征。

1）“8”的声调是第一声，“2”的声调是第四声，声学上表现为基频（Fundamental frequency F0）的高低和升降随时间而变化的模式不同。

2）“8”是双唇塞音声母/b/开头的，在韵母元音共振峰（Formant）的起始段，通常都有从下向上走的过渡音征，F2（2nd Formant）的尤其显著，F3(3rd Formant)多半也有，虽不十分明显；而“2” 没有声母，元音的起始段没有那种向上走的过渡音征。

3）“8” 的共振峰大部分走势比较平稳，F3 末尾一般较平或微微上扬；而“2”的最显著特点就是 F3 的逐渐下降以及 F2 末尾的微微上扬，两者趋于靠拢，只是 F3 下降的具体方式因人而异：有的一开始就逐渐下降，而有的是一开始比较平，到后半部分才逐渐下降，有的甚至直到最末尾才微微下降。总的来说，F3 的下降和 F2 末尾的微微上扬都比较一致。以上区别如图 1 所示。不过，除了声调区别以外，其余两个区别的出现情况因人

而异，但是至少能找到其中的某一种区别。

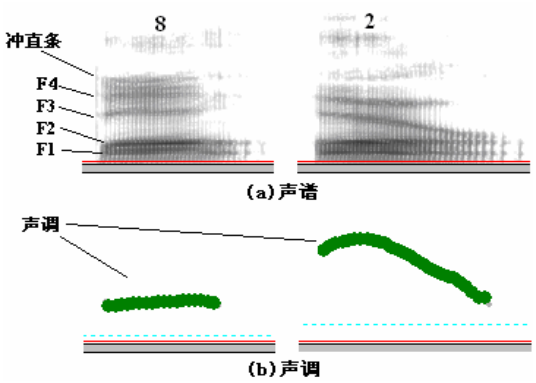

图 1 “2”和“8”声谱图和声调示例

2 共振峰模式统计分析

为了加深对“2” /er4/和“8”/ba1/声谱的认识，进一步分析了它们的共振峰分布的细节特征。

图 2、3 分别是根据 43 个发音人的数据所作的/er4/与/ba1/的/a1/共振峰模式图。

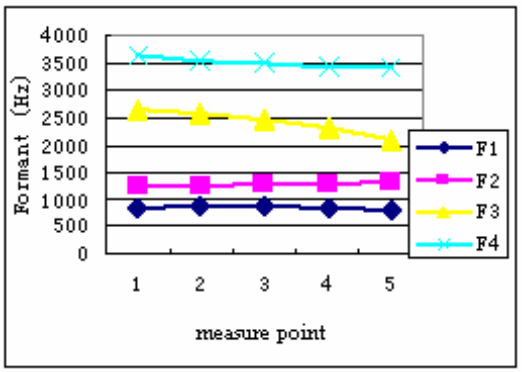

图 2 2（/er4/）的共振峰模式

根据图 2 和图 3，总体说来，/er4/与/ba1/两者的 F1 和 F2 差异很小，F3 和 F4 却有很大的差异。/er /的 F2 与 F3 的尾部有互相靠拢的趋势，F3 下降迅速，F2 尾部有上翘的趋势，F4 也有下降的趋势；/a/的 F3 与 F4 却是上翘的趋势。这些都跟图 1 的原始声谱实例所显示的区别一致。

具体地说，/er/与/a/的 F1 在第一个测量点存在显著差异，其余 4 个点都没有显著差异；/er/与/a/的 F2 在所有 5 个测量点上都没有显著差异；而/er/与/a/的 F3 则在每个测量点上都存在显著差异（后 3 个测量点特别显著）；/er/与/a/的 F4 在第一个测量点没有显著差异，在其余的点的位置有显著性差异。

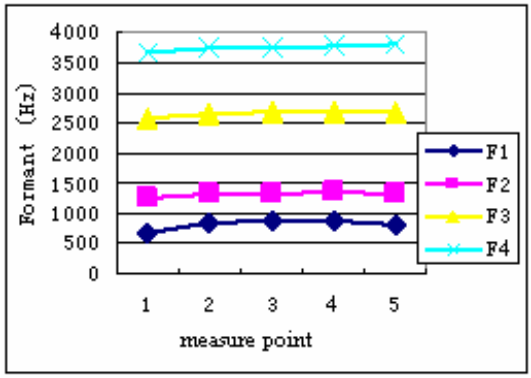

图 3 8（/ba1/）的共振峰模式

此外，就/er/本身而言，F3 从头至尾每两个点之间差异显著性水平指标 P 值（probability）随着位置的后移不断地减小，说明差异越来越显著，F3 在不断下降；而/a/的相应 P 值则随着位置的后移不断地增大，说明它的 F3 各点之间差异越来越小，基本呈平稳走势。

上述 43 个发音人共振峰模式的统计特性说明，/er/与/ba/的共振峰分布、尤其是两者的 F1 和 F2 的分布区域比较接近，所以识别上易混；但是，各自的共振峰频率随时间而滑移变化的模式却各不相同，尤其是 F3 的走势明显不同。所以，只要掌握它们的统计特性，多数情况下二者是可以识别的。

3 “2”和“8”韵母的发音特点

“2”的韵母/er/是个 r 音化的元音，一般写作带附加符号的央元音 schwa[ɚ]，但事实上如上文所述，其头两个共振峰跟/a/的很接近。本节分析/er/与/a/发音时的舌位信息，探究其发音上的特点。发音数据来自一位女性发音人，用三维电磁发音仪采集（Carstens 公司，AG500 系统）。图 4 是根据这里提取的中矢平面上的三个舌面采样点的数据（5 次重复的均值）所作的普通话元音舌位图。图中用虚线连接的国际音标表示每个元音在音段中点位置的舌位均值；用实线连接的三个实心圆点表示/er/在音段起始位置的舌位均值，而用实线连接的三个空心圆点则表示/er/在音段中点位置的舌位均值。

图 4 清晰显示出，在音段起始位置，/er/的舌位与元音/a/的基本上是一样的，这就解释了二者在共振峰模式上的相近之处。但是，由于/er/是个 r 音化的元音，发音上主要涉及舌尖的动态运动，如图中所示，到中点位置时，其舌面上的前两个采样点的数据已显著升高。正是这种动态运动形成了/er/在频谱上的动态特征，如：F3 显著下降、F4 下降、F2 尾部略升以致跟下降的 F3 呈靠拢之势。

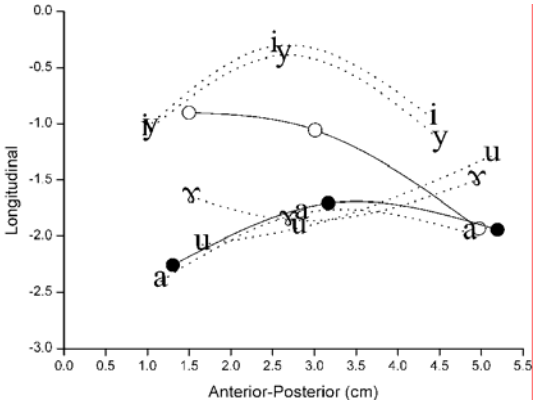

图 4　普通话元音舌位图（实心圆点：/er/的起首；空心圆点：/er/的中点）；发音人面朝左面

4 “2”与“8”声谱不易识别的原因分析

如上所述，从发音上看，发“2”/er/时，其起始段舌位的基本构造非常接近元音/a/的状态，只不过其舌尖部分有一个因 r 音化而引起的动态滑移。但是，有的人发“2”时习惯于从/a/的舌位再滑移到/er/的舌位，使得它的起始段和中段跟“8”里/a/的共振峰结构更加接近。所以，从不带声调信息的声谱上看，两者就有可能混淆。在人的自然言语交际中，由于有声调的因素及语境条件，二者不易混淆，但这却是二者在机器语音识别中产生混淆的主要因素。

同时，虽然“8”是塞音声母开头的，声谱上一开始一般有冲直条出现，如图 1 所示，但是，往往因人而异，有的人“8”的发音并不出现冲直条，只是后接元音起始如刀切般地整齐，如图 5 左图所示。“2”虽然没有辅音声母，但不少发音人发这类音往往带有一个喉塞开头，它在声谱上的表现跟一般“8”的开头很相像，会出现冲直条，如图 5 右图所示。这便加剧了二者的易混性。

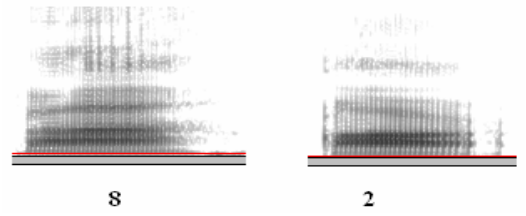

图 5　“8”与“2”起始声谱相混举例

5 人对“2”与“8”声谱辨识与感知实验

语音识别中“2”与“8”非常容易混淆，是否人在辨识它们声谱的时候也会发生混淆呢？另外，人听辨这两个音时一般是不可能混淆的。那么，在人的听辨过程中，究竟是哪些声谱特征的作用更大些。为此，本文设计了两个实验，一是人通过语图（即语音声谱图）辨识/er/和/ba/；一是感知实验，听辨人为非语音专业学生，但有一定的语音知识。

5.1 通过语图辨识/er/和/ba/

这个实验的目的是考察人通过语图辨识/er/和/ba/的正确率。辨识之前，通过讲解，让辨识人得知辨识的要点，即/er/和/ba/ 语图的特点。然后，把不带声调信息的/er/和/ba/各 43 个样本的语图（如图 6 所示）随机排列出示，请辨识人作出判断。

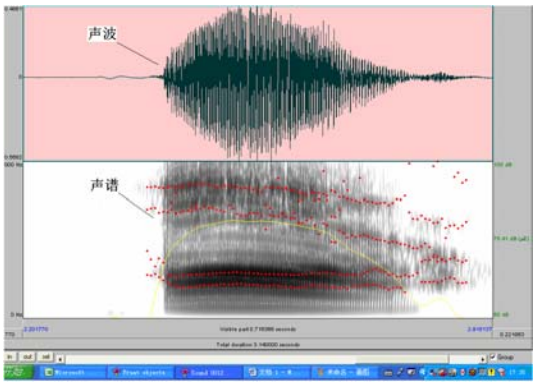

图 6　不含声调信息的语音声谱图示例

实验分两步，在两个不同的人群中、采用两种不同的方法中进行。第一步共有 7 个人参加，要求他们必须对每个样本断定是 2 还是 8。辨识结果显示，7 个人中辨识正确率最高的是 98%，最低的是 87%[1]，这种差异主要跟个人对声谱识读的经验有关系。第二步共有 8 个人参加，对每个样本的辨识可以在 2、8 或不确定三者之间选择，辨识结果如表 1（见次页）所示。

根据表 1 的数据可知，/ba/和/er/的平均正确辨识率分别为 98.25%和 81.69%，/er/和/ba/的总体辨识率为 89.82%。这说明，人辨识这种声谱有时也会发生混淆，只是错误率较低，而且，主要是 2 的声谱不易辨识。

事实上，有的声谱显示单独看起来似乎有些混淆现象，如图 7 的 64 号“8”的声谱 F3 有些往下走，所以不少人把它识别成了/er/；不过，跟这个发音人所发的/er/（59 号）比较，其区别还是十分显著的。因此，假如分别辨识各个发音人的/er/和/ba/ 的声谱，其误识率应该会低得多。而这里的声谱辨识实际上还涉及不同话者发音的个体特性差异。

[1] 第一步实验由语言研究所的华武协助实施，特此致谢。

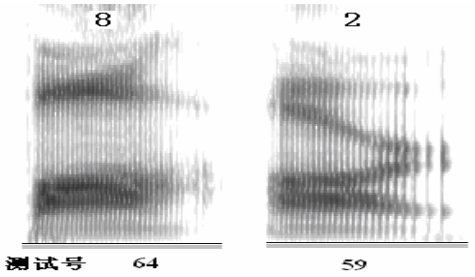

图 7　相同发音人的“8”和“2”语图比较

表 1　人对 2 和 8 语谱（语音声谱）的辨识结果统计

辨识人	1	2	3	4	5	6	7	8	均值
/ba/错识为/er/个数	2	1	0	0	0	1	2	0	0.75
/er/错识为/ba/个数	4	10	9	13	7	8	2	11	8.
/ba/错识率 %	4.65	2.33	0	0	0	2.33	4.65	0	1.75
/er/错识率 %	9.30	23.25	20.93	27.9	16.28	18.6	4.65	25.58	18.31
总错识率 %	6.97	12.79	10.46	15.12	8.14	10.5	4.65	12.79	10.18

5.2　听辨实验

为测试人在听辨 2 的过程中，究竟是哪些声谱特征的作用更大些，整个听辨实验一共对原始数据的 5 个方面作了改动，然后对修改数据后合成的音进行听辨测试。数据修改分为单因素和复合因素两个部分。单因素是指对原始数据的改动仅涉及一个单一的因素，包括改变声调（er1），只保留第一和第二共振峰（er4（F1+F2）），只保留第三及以上的共振峰（er4（≥F3））；复合因素是指修改涉及两个或两个以上的因素，包括改变声调且只保留第一第二共振峰（er1（F1+F2）），改变声调且只保留第三及以上的共振峰（er1（≥F3））。具体测试结果见表 3 至表 7。

1）按照对修改数据后合成音的听辨结果进行升序排序，为：er1(F1+F2) < er1(≥F3) < er4（F1+F2）< er1 < er4(≥F3)。他们的听辨结果正确率分别为：63.83%，65.11%，79.1%%，88.7%，95.35%。对原始数据（er4）的听辨结果为 97.67%，详见表 2。

2）统计结果显示，er4 和 er4(≥F3)的听辨结果之间没有显著的差异（P>0.01）；er4 和 er4（F1+F2）的听辨结果之间有显著的差异（P<0.01）。说明在原有声调的基础上，F3 及以上的共振峰对/er4/识别的贡献要大于 F1，F2 的贡献，即 F3 及以上的共振峰是影响/er4/识别的主要因素。

3）er1 和 er1(≥F3)的听辨结果之间有显著的差异（P<0.01）；er1 和 er1(F1+F2)的听辨结果之间也有显著的差异(P<0.01)，但 er1(≥F3)和 er1(F1+F2)的听辨结果之间没有显著差异（P>0.01）。这一点和保留原声调的情况不同，说明在无声调区别(均为阴平)的基础上，F3 及其以上的共振峰和 F1，F2 对识别的影响的差距在减小，以至于两者对影响识别的作用基本上相同（没有显著差异）。（但具体是 F3 及以上的共振峰的作用减少了，还是 F1，F2 的作用
增加了，目前还不清楚）。

4）er4（F1+F2）和 er1（F1+F2）的听辨结果有显著差异（P<0.01），er4(≥F3)和 er1(≥F3)的听辨结果也存在显著差异（P<0.01），即声调是影响识别的主要因素。这一点和保留原调的结果显然不一致：（er4（F1+F2）和 er4(≥F3)的听辨结果之间有显著差异（P<0.01）。

表 2 是对/er4/（原始 2）的听辨结果，平均正确率在 97%以上；表 3 是对/er1/（排除声调的 2）的听辨结果，听为/a/的平均百分比为 7.31%，不确定的平均百分比为 3.99%，总体错误率为 11.30%，跟表 2 情况相比，平均正确率（88.7%）大大降低，这充分说明声调对识别的重要作用；表 4 是滤除 F3 及以上高次共振峰后的/er1/的听辨结果，听为/a/的平均百分比为 31.94%，不确定的平均百分比为 4.23%，总体错误率为 36.17%；表 5 是原始 2（/er4/）滤除 F1,F2 以后的听辩结果，平均正确率高达 95.35%，非常接近对原始语音的听辨水平，说明 F1,F2 在区分 2 和 8 的听辨上基本上没有什么作用；表 6 是/er1/（排除 2 的声调信息）滤除 F1,F2 后的听辩结果，平均正确率只有 65.11%，显然说明，排除 2 的声调信息以后会极大影响它与 8 的听辨区别；表 7 是/er4/（原始 2）只保留 F1,F2 的听辨结果，误听为/a/的百分率为 12.4 %，不确定的百分比为 8.5%，总体错误率高达 20.9 %，可见滤除 F3 及以上高次共振峰以后，会大大降低听辨上区分 2 与 8 的正确率。

从以上听辨的结果比较可见，对/er1/的听辨情况跟人对语图识别的情况相仿，都排除了声调的作用，因此，总体错误率 11.30%与人对语图辨识误识率 10.18%十分接近；而滤除 F3 以上高次共振峰以后，平均有 31.94 %的都听辨为/a/，4.23%不能确定，说明就 F1 与 F2 来说，

/er/与/a/的混识率很高，F3 以上的高次共振峰对/er/与/a/ 的区分起关键作用。

表 2 /er4/（原始 2）的听辩结果

听音人	1	2	3	4	5	6	7	均值
误听为/a/的个数	1	1	0	0	0	4	0	0.86
不确定的个数	0	0	1	0	0	0	0	0.14
正确率%	97.67	97.67	97.67	100.00	100.00	90.70	100.00	97.67

表 3 /er1/（排除原始 2 的声调）的听辨结果

听音人	1	2	3	4	5	6	7	均值
误听为/a/的个数	2	2	3	1	1	4	9	3.14
误听为/a/的百分比 %	4.65	4.65	6.98	2.33	2.33	9.30	20.93	7.31
不确定的个数	0	0	0	0	1	3	8	1.71
不确定的百分比 %	0	0	0	0	2.33	6.98	18.60	3.99
总体错误率 %	4.65	4.65	6.98	2.33	4.65	16.28	39.53	11.30
正确率 %	95.35	95.35	93.02	97.67	95.35	83.72	60.47	88.7

表 4 只保留 F1-F2 的 /er1/（排除 2 的声调信息）的听辨结果

听音人	1	2	3	4	5	6	7	均值
误听为/a/的个数	25	7	18	4	11	28	3.14	13.73
误听为/a/百分率 %	58.14	16.28	41.86	9.30	25.58	65.12	7.31	31.94
不确定的个数	1	2	2	0	5	1	1.71	1.82
不确定的百分比 %	2.33	4.65	4.65	0.00	11.63	2.33	3.99	4.23
总体错误率 %	60.47	20.93	46.51	9.30	37.21	67.44	11.30	36.17

表 5 /er4/（原始 2）滤除 F1,F2 以后的听辩结果

听音人	1	2	3	4	5	6	7	均值
误听为/a/的个数	1	0	0	0	0	12	0	1.86
不确定个数	0	0	1	0	0	0	0	0.14
正确率 %	97.67	100.00	97.67	100.00	100.00	72.09	100.00	95.35

表 6 /er1/（排除 2 的声调信息）滤除 F1,F2 后的听辩结果

听音人	1	2	3	4	5	6	7	均值
误听为/a/的个数	8	0	2	31	4	34	3	11.71
不确定个数	2	1	4	6	0	1	9	3.29
正确率 %	76.74	97.67	86.05	13.95	90.70	18.60	72.09	65.11

表 7 /er4/（原始 2）只保留 F1,F2 的听辨结果

听音人	1	2	3	4	5	6	7	均值
误听为/a/的个数	N/A	1	2	7	4	18	0	5.33
误听为/a/的百分率 %	N/A	2.3	4.7	16.3	9.3	41.9	0	12.4
不确定的个数	N/A	1	10	1	7	0	3	3.67
不确定的百分比 %	N/A	2.3	23.3	2.3	16.3	0	7	8.5
总体错误率 %	N/A	4.7	27.9	18.6	25.6	41.9	7	20.9

6 结论

声学和生理实验结果表明，“2”/er4/和“8”/ba1/之间最突出的差异是声调，前者是第四声，后者是第一声；声谱的统计模式也不同，尤其是第三共振峰（F3）的分布，两者在每个测量点上的细节特征都具有显著区别，只是存在着个体差异；而它们的第一共振峰（F1）和第二共振峰（F2）都很接近，从声谱上看不易区分。感知实验结果也证明，在“2”与“8”的区分上，声调和 F3 是决定性的要素，而 F1 和 F2 的作用非常有限。

综上所述，声调区别是“2”和“8”识别中可资利用的最显著的区别特征，而早期的数字语音识别恰恰忽略了声调这个最最重要的因素。在缺乏声调信息的情况下，第三共振峰本来是个关键的区别要素；可是，在连续语音（例如数字串）中，尤其是非正式的语体中，有些“2”在发音时舌尖运动常常不够到位，因而在声谱上跟“8”的第三共振峰差异有时不十分明显，这就使得两者的识别失去了另一个最重要的声学依据，最终导致识别中“2”与“8”的常常混淆。

由此可见，在语音识别中，必须充分地认识语音学知识及其作用，才能有效地利用它们；必须抓住最根本的区别性语音特征信息，并在系统中加以综合利用，才能取得最好的效果。

注：本文原载《第九届全国人机语音通讯学术会议（NCMMSC2007）论文集》，标题是“从 2 与 8 的语音辨识看语音学知识在识别中的应用前景”；后入选《清华大学学报》（自然科学版），标题应要求而压缩修改，将于 2008 年刊出。

参 考 文 献

[1] Stevens, K.，Li，Z., Lee C-Y.，et al, "A Note on Mandarin Fricatives and Enhancement". In From Traditional Phonology to Modern Speech Processing, [D] eds., G. Fant, H. Fujisaki, J. Cao, and Y. Xu, (Beijing: Foreign Language Teaching and Research Press, 393-403, 2004).

[2] http://www.rle.mit.edu/speech/research.html

[3] Lee, Chin-Hui (2004): "From decoding-driven to detection-based paradigms for automatic speech recognition"[A]. In INTERSPEECH-2004, [C]. paper P2.

[4] Zue, Victor, Speech recognition: Acoustic-phonetic knowledge acquisition and representation. Fiscal Report, 1 Oct. 1987 - 30 Sep. 1988 Massachusetts Inst. of Tech., Cambridge. Research Lab. of Electronics.

[5] Chen, K., Hasegawa-Johnson, M., and Cole, J. (2003). Prosody dependent speech recognition with explicit duration modeling at intonational phrase boundaries. [A] Proceedings Eurospeech 2003, [C] Geneva.

[6] Sarah Borys: The importance of prosodic factors in phoneme modeling with applications to speech recognition. [A] Proceedings of HLT-NAACL, [C] Edmonton, 2003.

[7] Cho, T., McQueen, J., Cox, E: Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. Journal of Phonetics, [J] Vol.35,2007.

[8] 王作英，肖熙. 基于段长分布的 HMM 语音识别模型[J],《电子学报》，2004，32（1）：46-49。
WANG Zuoying and XIAO Xi, Duration Distribution Based on HMM Speech Recognition Models[J]. Acta Electronica Sinica, 2004，32(1):46-49. (in Chinese).

Application of Phonetic knowledge in Automatic Speech Recognition ——Case Analysis

CAO Jianfen[1], **LI Aijun**[1], **HU Fang**[1], **ZHANG Ligang**[1, 2]

1. Institute of Linguistics, Chinese Academy of Social Sciences, Beijing 100732, China
2. Institute of Computer Science and Technology, Tianjin University, Tianjin 300072, China

Abstract: One of the pop topics in Automatic Speech Recognition (ASR) is how to enhance the validity by utilizing phonetic knowledge. In early stage of numbers' speech recognition, a difficult problem was failure in discriminating 2 and 8. This paper discusses the application of phonetic knowledge in ASR through analysis to this specific case. Methods employed in

this study include acoustical and physiological experiments, and combined with a set of perception tests. The attention was mainly paid to investigate the distinctive phonetic features and their role in distinguishing of 2 and 8. The results reveal that tonal information is the most prominent distinctive feature between 2 and 8, and that of 3rd formant (F3) is the key discriminative factor in the case of absence of tonal information, but the 1st formant (F1) and 2nd formant (F2) of 2 and 8 are similar, hence have less role in their distinction. However, the tonal distinction was ignored unluckily in the early stage of numbers' speech recognition, while in continuous speech, especially in the case of informal style speech, the difference of F3 between 2(/er4/) and 8(/ba1/) is not often prominent enough due to articulatory undershoot of tongue tip movement in 2's articulation. These are the main reasons that cause the confusion between 2 and 8 in ASR. Consequently, in order to enhance the veracity in ASR, realize more phonetic knowledge and their roles in ASR is an urgent task, adequately and compositively apply these knowledge into system is an effective approach.

Key words: Computer, Automatic Speech Recognition(ASR), phonetic knowledge, perception, tone, fundamental frequency(F0), spectrogram, formant