NCMMSC — 92

第二届全国人机语音通讯学术会议

论 文 集

中国自动化学会 模式识别与机器智能专业委员会
中国计算机学会 人工智能与模式识别专业委员会
中国电子学会 信号处理学会语音图象通信专业委员会
中国声学学会 语言听觉和音乐声学分科学会
中国中文信息学会 基础理论专业委员会
中国科学院 自动化研究所

桂 林
一九九二年九月

前　言

1990年6月，由中国计算机学会人工智能与模式识别专业委员会和中国软件行业协会主持，第一届全国语言识别学术报告与展示会在北京清华大学召开，这是国内举行的第一个全国性的以语音识别为主题的学术会议，因此受到了这一领域的科技工作者的欢迎和重视，有67位专业人员与有关部门的代表出席了会议，发表了44篇论文。通过三天的学术交流和科技成果展示，与会科技工作者对这次会议给予了充分的肯定，认为有利于促进国内语音识别和合成的研究工作的发展，会议决定今后每两年举行一次这样的会议，并建议联络更多的学术团体联合主持，以扩大影响，吸引更多的同行参加会议。会议并决定由中国科学院自动化研究所负责筹办於92年召开的第二届会议。

经过筹办单位的联络，1991年1月由中国自动化学会模式识别与机器智能专业委员会、中国计算机学会人工智能与模式识别专业委员会、中国中文信息学会基础理论专业委员会、中国声学学会语言听觉和音乐声学分科学会、中国电子学会信号处理学会语音图象通信专业委员会等五个学术团体的代表举行了会议，一致同意由这五个学术团体联合主持召开这项学术会议。会议还根据国内语音处理研究的发展情况，决定从本届会议开始，将学术交流内容扩大为人机语音通讯领域，并将会议更名为“全国人机语音通讯学术会议”。同时决定第二届会议今年9月在桂林召开。

本届会议得到了国内同行的热情支持和关心，踊跃向会议投交论文。共收到学术论文一百多篇，经程序委员会评审，共录取了95篇论文，录取的论文不仅在数量上大大超过第一届，而且覆盖的研究领域面也超过了上一届会议，这批论文充分反映了近年来特别是90年以来我国人机语音通讯领域所取得的进展。

人机语音通讯涉及多种学科的研究，又有广阔的应用前景。国内在这一研究领域已经有了相当好的工作基础，不仅在汉语合成与识别的方法上取得了重要的进展，而且在实用化产品开发方面也取得了初步成效。我们相信，本次会议的召开，将进一步推动人机语音通讯的研究和向实用化发展的进程。我们衷心地祝愿人机语音通讯界同行的又一次盛会取得成功。

黄泰翼

第二届全国人机语音通讯学术会议

会议主席：黄泰翼（中国自动化学会　模式识别与机器智能专业委员会）
（中国科学院　自动化研究所）

副 主 席：方棣棠（中国计算机学会　人工智能与模式识别专业委员会）
（中国中文信息学会　基础理论专业委员会）
张家骝（中国声学学会　语言听觉和音乐声学分科学会）
李昌立（中国电子学会　信号处理学会语音图象通信专业委员会）
方惠均（桂林电子工程学院）

程序委员会主任：方棣棠(清华大学)

副主任：黄泰翼(中科院自动化所)

委　员：

王作英(清华大学)　王仁华(中国科技大学)
刘增寿(国防科技大学)　吴文虎(清华大学)
陈永彬(东南大学)　陈道文(中科院自动化所)
李昌立(中科院声学所)　杨家源(四川大学)
杨顺安(社科院语言所)　余崇智(南京大学)
迟惠生(北京大学)　严普强(清华大学)
林茂灿(社科院语言所)　张家骝(中科院声学所)
柯有安(北京理工大学)　俞铁城(中科院声学所)
赵国田(哈尔滨工业大学)　袁保宗(北方交通大学)
徐秉铮(华南理工大学)　徐近霈(哈尔滨工业大学)
柴佩琪(同济大学)

论文集责任编辑：周如玉

论文集目录

A　听觉模型与特征提取

B1　语音识别方法与系统

B2 语音识别方法与系统

C　非特定人语音识别

D　连续语音识别与语言模型

E 说话人识别

F 神经网络应用

G 语音合成

H　语音库与系统评价方法

J　语音分析与基音提取

K 语音编码与增强

L 语音技术应用

关于听觉模型研究的现状与动态

方棣棠　　刘惠华
(清华大学)

[摘要] 本文综述了听觉模型研究的意义、任务、发展历史和现状，简要介绍了作为听觉模型研究基础的听觉系统神经生理学方面的进展，并通过几个具体例子说明了听觉模型的构成原理及其在语音识别方面的应用.

一、引言

现有的计算机语音识别方法均是利用声波的物理特性和人的发声过程特性来提取语音识别的参数. 经过几十年来的飞速发展，这类方法在取得了辉煌成果的同时，也表现了其本身固有的弱点和限度，例如在对非特定人的语音识别，对外界噪声的抗噪声能力等方面，性能难于进一步改进. 自然地，这一领域的科学家们注意到人类自身所具有的优异的语音识别能力并希望能在计算机语音识别中实现这一能力. 关于听觉模型的研究就是这种努力的一部分，人们希望在听觉模型的研究中有所突破，为计算机语音识别开辟新的道路.

二、听觉模型研究的任务

虽然人们常将语音信号处理和图象信号处理相提并论，但二者有很大不同. 图象信号是一种物象 (object) 信号，是空间信号，而语音信号是一种事件 (event) 信号，是时间信号. 语音信号随时间变化剧烈、转瞬即逝. 人的视觉感觉细胞 (锥状细胞、杆状细胞等) 的数目大约是听觉感觉细胞 (内毛细胞、外毛细胞) 的数目的几十倍，而视觉接受的信息只是听觉接受的信息的几倍，因此每个听觉感觉细胞比视觉感觉细胞所承担的信息量约多十倍，这就说明人的听觉系统是一个很复杂的编码系统.

人的听觉系统是外界语音信息进入大脑的唯一通路，由下而上依次为外耳、内耳、耳蜗神经、耳蜗核、上橄榄核、下丘、内侧膝状体、大脑皮层听觉区. 在听觉通路的各个阶段上都对语音信号进行处理，例如内耳的基底膜和耳蜗神经进行声音的频谱分析，但是这种频谱分析与用物理方法所做的频谱分析很不相同，这种频谱经过听觉通路的各个阶段的多次处理后刺激大脑皮层的特定部分，形成某个音的感觉. 听觉神经生理学家研究的是在听觉通路的各个阶段中各神经纤维在声音刺激下的放电模式. 而对他们的大量的研究成果和发现进行综合分析，找出规律性的东西，弄清听觉系统是怎样对语音信号进行编码的，建立具有人类听觉系统特性的语音信号分析处理系统，使之适用于计算机语音识别，是听觉模型研究的任务. 神经生理学家探索和发现听觉系统的奥秘，而听觉模型的研究者要做的是破译听觉密码用于计算机语音识别.

语言识别与听觉和语言两个领域有关，人的听觉系统把外界语音信号进行处理后送到大脑听觉区进行音的辨识. 至于语言理解那是大脑皮层语言区的职能，而大脑皮层听觉区

只负责音的辨识，所以听觉模型的任务也是进行音的辨识而不包括语言理解。

我们希望借助于听觉模型研究解决的问题是：

(1) 找到更简洁有效的参数。在这方面语音信息在听觉通路的各个阶段特别是在大脑皮层的表示模式可能给我们以启发。

(2) 更好地利用过渡音的信息来提高识别率。可以从听觉系统信息处理的非线性和动态特性中得到启发。

(3) 找到比较统一的参数，即去掉个人信息而只保留音韵信息的参数，来解决非特定说话人问题。

(4) 提高抗噪声能力。人类能在嘈杂的场合（例如食堂）互相交谈，而目前的语音识别装置在这种场合就完全丧失了识别能力。

三、历史与现状

自从一九六一年 Békésy 由于揭示了内耳基底膜运动机制获得诺贝尔奖金以来，随着听觉生理和心理科学的发展，关于听觉模型的研究出现过几个高潮。大体上说，一是六十年代的生物物理模型，即对外耳、中耳和内耳基底膜的物理特性的模型化，例如对耳蜗管这种一端封闭短管的声学特性进行模型化等。二是七十年代的神经生理模型，即对内毛细胞将声波振动转化为电脉冲发放的机能和特性的模型化及对听觉神经纤维电脉冲发放模式的研究和模型化。三是八十年代的表示模型，即对语音信号在听觉系统中表示（representation）形式的研究和模型化，主要研究听觉系统从语音信号中提取了什么样的频谱送到大脑。九十年代以来，尚未形成新的高潮，仍然是围绕着上述几方面的模型进行工作。

迄今为止，听觉模型的研究主要局限于听觉末梢系的范围，这是由神经生理学的进展状况所决定的。虽然几十年来，听觉神经生理有了很多奇迹般的成果，但总的看进展是艰难而缓慢的。人的听觉系统构造太精密太复杂，象火柴棍那么粗的听觉神经纤维束就有几万根神经纤维有秩序的排列，并各有其不同的特性。在内耳耳蜗中 35mm 长的基底膜上均匀排列着一列内毛细胞和三列外毛细胞，每一列有 3500 个左右，所以平均每 1mm 长基底膜上，约有 100 个内毛细胞和 300 个外毛细胞，而每个毛细胞的顶部又有 100 多根纤毛。沿耳蜗顶部到耳蜗基部纤毛挠度（stiffness）逐渐增加，并且三列外毛细胞从内到外纤毛的挠度减小。这些事实对于分析作为听觉感觉细胞的内毛细胞的作用规律是有用的。音波是一种物理振动，进入内耳后引起纤毛的运动，纤毛的运动刺激内毛细胞体发放电脉冲送到听觉神经纤维，每个内毛细胞连接约 20 根神经纤维，每根纤维的频率特性和起始发放电脉冲的阈值又不相同。人的听觉系统是这样精密，有研究说人的听觉能感受到的振动只比分子的布朗运动的平均自由程大几倍，也就是说如果人的听觉再灵敏一些的话，就会被气体分子的自由运动所引起的噪声吵得不得安宁。

虽然如上所述，内耳即听觉末梢系统的机制已被基本了解清楚，但即使是对于末梢系统，仍有未知的部分，例如对占毛细胞总数的四分之三的外毛细胞，目前只知道它们连接从中枢到末梢的下行神经通路，具有运动神经的控制作用，但对控制的具体过程还不清楚。对于中枢听觉系统，目前所知道的一些主要事实包括：(1) 整个听觉通路的各个阶段始终存在着有秩序的特征频率分布。(2) 从内耳发出的神经纤维互相绝缘，在到达第一个

信息处理中间站耳蜗核时，百分之九十以上的纤维互相连接。(3) 耳蜗核中存在着强烈的侧抑制作用，并且听觉通路的其它阶段上也存在着这种侧抑制作用。(4) 左右耳神经纤维在橄榄核处交叉。(5) 大脑听觉皮层的神经细胞分成几类，有的只对信号的起始部分有反应，称为 ON 型，有的只对信号的结束部分起反应，称为 OFF 型，有的只对信号频率的变化起反应，称为 FM 型，有的只对振幅的变化起反应，称为 AM 型，还有的对所有各种类型的刺激都有反应，称为持续型。以上这些事实对中枢听觉系统的机制来说，还只是很小一部分个别事实，对于语音信号在中枢听觉系统中的处理过程，我们还完全没有掌握。作为系统，人类的中枢听觉系统仍是一个未知世界，是一个有待探索的奥秘，因此听觉模型的研究目前也是处在一条漫长道路的开始阶段。

四、神经生理学家所提出的几种听觉中的语音表示

"听觉中语音表示"的意思是"在听觉系统中语音变成了什么"即语音在听觉系统中所表现的形态，也就是在语音的刺激下听觉神经纤维电脉冲发放的模式和规律，这是听觉神经生理学家所一直孜孜寻求的，也是听觉模型研究的基础。下面是几种听觉中语音表示的例子。

(1) 神经谱图表示（即神经电脉冲发放直方图，简称神经谱图）(C.D.Geishler 美国威斯康星大学)

图 1 所示是对[tu]和[pu]两个发音的神经谱图和富里叶谱图的比较。(a) 是[tu]的神经谱图，(b) 是[tu]的富里叶谱图，(c) 是[pu]的神经谱图，(d) 是[pu]的富里叶谱图。神经谱图是从听觉末梢神经纤维中测出的，浓淡度表示脉冲发放率（单位时间内脉冲发放的次数）。

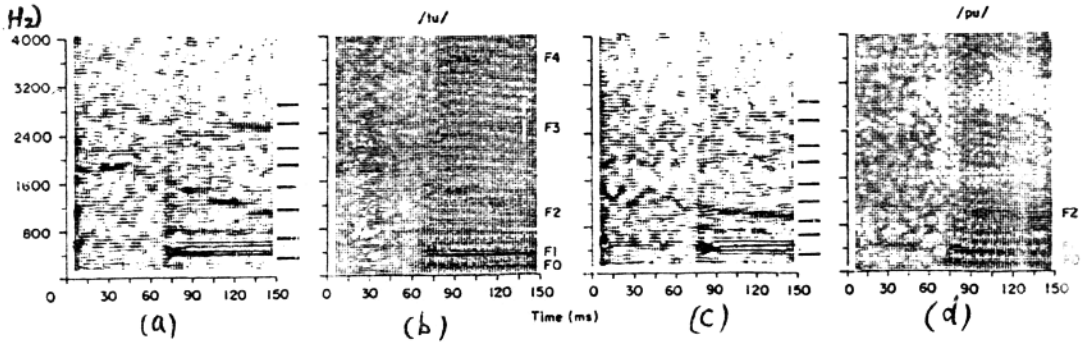

图 1 神经谱图与富里叶谱图的比较

由图 1 可见，神经谱图与富里叶谱图比起来，辅音部分的频谱更为清晰，共振峰更为突出，音韵变化部分更为明显（例如元音的开始部分和辅音的开始部分）。

(2) ALSR 表示　　(M.B.Sachs 美国马里兰大学)

ALSR 是 Average Localized Synchronized Ratio（即平均局部同步率）的缩写。这种表示的原理是：听觉末梢系统不但有空间频率分解作用，并且也传递由时间波形所表达的频率信息。图 2 所示的元音[ε]的 ALSR 图是这样做出来的：测出在某特征频率周围 0.25oct 范围内的所有神经纤维的电脉冲发放时间波形并进行富里叶变换，求出各纤维输出波形中该频率的分量并进行叠加，就是该特征频率的振幅，单位是 SPIKE／S.（每秒电脉冲发放次数）

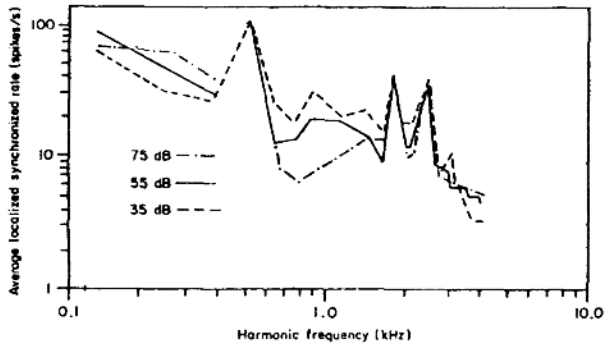

图 2　元音[ε]的 ALSR 表示

由图 2 可见，这种表示法对共振峰的表示非常清晰突出。

从以上两个例子，我们看到：语音在听觉系统中被表示为与我们平时用物理方法求出的频谱相当不同的听觉谱，其特点是强调了共振峰、音强变化部分和频谱变化部分等在知觉上有重要意义的特征。听觉模型研究要解决的是如何用数理方法把听觉系统的这些特性表现出来，形成适于计算机语音识别的性能优越的参数。

五、几种典型的听觉模型

(1) 伊藤、福岛模型　　(日本 NHK)

这是一种神经生理模型，模型构成如图 3 所示，由七层组成，各层间的输入输出函数由下式表示

$$W_l(n,t)=\varphi\left[\frac{\alpha+\int b_e(\tau)\sum_{k_e=N_1}^{N_2}a_e(k_e)W_{l-1}(n+k_e,\ t-\tau)d\tau}{\alpha+\int b_h(\tau)\sum_{k_h=M_1}^{M_2}a_h(k_h)W_{l-1}(n+k_h,\ t-\tau)d\tau}-1\right] \tag{1}$$

其中：$\varphi(x)=\begin{cases}x & (x\geqslant 0)\\ 0 & (x<0)\end{cases}$;

$b_e(\tau)=(1/T_e)e^{-\tau/T_e}$；　　$b_h(\tau)=(1/T_h)e^{-\tau/T_h}$；

i是层的次序；n是所计算的神经纤维的号数；α是神经纤维放电阈值；

T_e是兴奋性输入的时间常数；T_h是抑制性输入的时间常数；

$a_e(k_e)$是k_e号纤维的兴奋性输出加权系数；$a_h(k_h)$是k_h号纤维的抑制性输出加权系数。

当改变时间常数T_e、T_h和加权系数a_e，a_h时，上式就能表示不同特性的神经细胞，例如当选定$k_e=k_h=k$，令加权系数按高斯分布$a_e(k)=a_h(k)=\beta e^{\frac{-k^2}{2\sigma^2}}$，其中$\beta$是满足$\sum_{k=N_1}^{N_2}\beta e^{-\frac{k^2}{2\sigma^2}}=1$的常数，则(1)式就可表示出$ON$型神经细胞的特性。

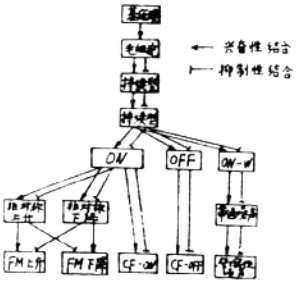

图3　伊藤、福岛模型的构成

图4　伊藤、福岛模型对语音[ichi]的输出

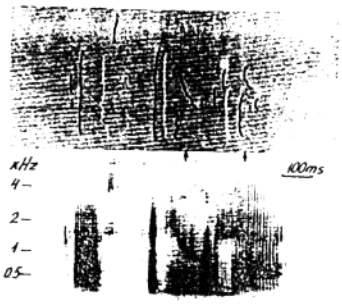

图5　Kozhevnikov模型对语音"Kastriula"的输出

图4给出了模型的输出，输入语音是"ichi"．由图可见，①持续型细胞模型的输出表现了基本频谱；②宽带ON型细胞模型只在各共振峰的始端有输出；③摩擦音型细胞模型只对输入语音中的辅音部分"ch"有输出．

(2) Kozhevnikov模型　　(前苏联科学院)

(原论文中没有说明模型的结构)

模型的输出如图5所示．输入语音"Kastriula"，下面是该语音的语谱图，上面是该模型的输出，黑线表示ON响应，白线表示OFF响应．

从图中我们看到该模型的输出相当精确地标出了元音的开始部位和结束部位及辅音的开始部位．原作者认为：这种分割界限标志真实性的假设看来是中枢听觉处理理论的关键．

(3) Li Deng 模型　　(加拿大滑铁卢大学)

模型构成如图 6 所示。

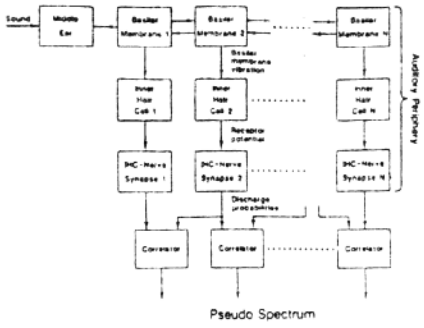

图 6　Li Deng 模型构成

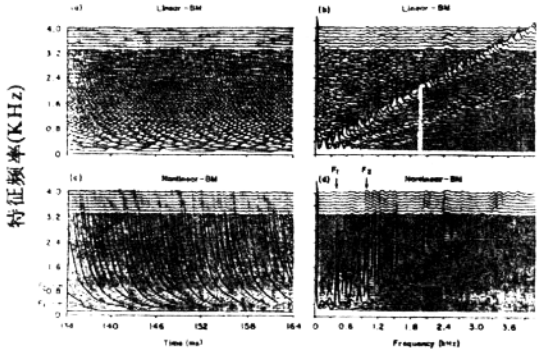

图 7　S / N = 0dB 时 Li Deng 模型对[mu]音的输出

这个模型的特点是在基底膜的行波方程中加入了非线性环节。基底膜的行波方程为:

$$\frac{\partial^2}{\partial x^2}\left(m\frac{\partial^2 u}{\partial t^2}+r(x,t)\frac{\partial u}{\partial t}+S(x)u-k(x)\frac{\partial^2 u}{\partial x^2}\right)-\frac{2QB}{A}\frac{\partial^2 u}{\partial t^2}=0 \qquad (2)$$

$$(0<x<L,\quad t>0)$$

线性时: $r_o(x)=r(CF)=b+a\times e^{-CF}$

非线性时: $r(x,t)=r_o(x)\{1+10^{[20\log_{10}(|u(x,t)|/0.5\text{Å})-20]/15}\}$

式中: u: 基底膜位移; m: 质量; r: 阻尼; s: 基底膜挠度; k: 相邻区域耦合系数; QB/A: 生理常数; L: 耳蜗管长度。

图 7 示出当输入为混有噪声且信噪比为 0dB 的[mu]时模型的输出。噪声为宽带噪声。上面是线性模型的输出，下面是非线性模型的输出，左边是时间函数，右边是频率函数。

由图可见，由于加了非线性环节，即使对信噪比为 0dB 的信号，该模型也能给出清晰的共振峰表示。

(4) Seneff 模型　　(美国 MIT)

如图 8 所示，Seneff 模型用临界滤波器群模拟基底膜振动，用半波整流模拟内毛细胞的单向放电特性，用微分电路来模拟短时适应和掩蔽效应，用快速 AGC 来模拟 ON 型细胞的机能。最后，信号经过两路并列处理输出，一路是包络检出，另一路是同步检出，其中的同步检出部分是 Seneff 本人的创作，其原理如图 9 所示，经过这种处理，对那些输出波形的频率与其本身特征频率相同的通道，输出将得到大大加强，否则输出将大大削弱，加强了频率变化的对比，表现了听觉特性。

Seneff 模型的输出如图 10 所示。输入语音是"hasitate"。

由图可见，经过同步检出处理过的频谱、共振峰和音韵变化部分都得到了加强。

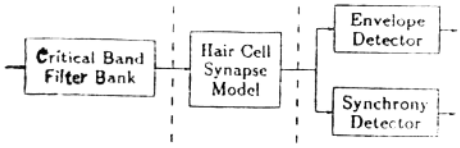

图 8 Seneff 模型的构成

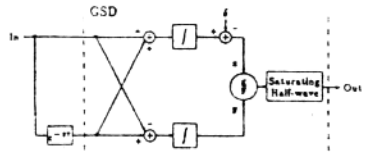

图 9 Seneff 模型中同步检出部分原理图

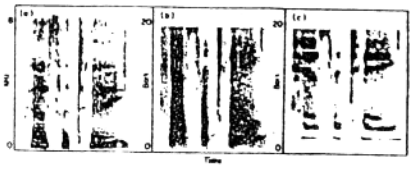

图 10 Seneff 模型对语音"hasitate"的输出
(a)富里叶变换频谱 (b)包络检出频谱
(c)同步检出频谱

(5) Ghitza 模型 (美国 Bell 实验室)

在现有的听觉模型中，Ghitza 模型很有特点．Ghitza 所提出的 EIH 表示不但表现了强调音韵特征等听觉特性，而且有量的规定．Ghitza 用 EIH 表示代替 FFT 频谱，经过 LPC 提取参数后，进行了计算机语音识别的实验．EIH 是 Ensemble Interval Histogram 的缩写，即"综合间隔直方图"．当有语音信号输入 Ghitza 模型，其输出就是该语音的 EIH 表示．

Ghitza 模型的构成如图 11 所示，由三个阶段组成．第一阶段是模拟基底膜振动的临界滤波器群，由 85 路滤波器沿对数座标的频率轴均匀分布于 200Hz 到 3200Hz．第二阶段是模拟单个毛细胞输出信号强弱的水平检出器，如图 12 所示，在这一阶段首先利用计算相邻脉冲间隔时间的方法利用$f_i = \frac{1}{T_i}$的关系确定频率 f_i，然后利用所设置的 7 个阈值来检测信号强弱，信号每超过一个阈值就在该频率 f_i 的振幅 $A_j(f_i)$上加 4dB，注意 f_i 不一定是所在滤波通道的中心频率．第三阶段是综合输出，即求出$A(f_i) = \sum_j A_j(f_i)$这样就综合了各通道对 f_i 的贡献，模拟了基底膜有时并不谐振于特征频率而是谐振于邻近更强频率分量的特性，利用时间信息 T_i 求出频率特性，这样求出的 $A(f_i)$即是 EIH 参数．

图 13 示出了元音[u]的富里叶频谱和 EIH 表示的比较．左面是富里叶频谱，右面是 EIH 表示，上面是在没有噪声的条件下求出的，中间是在信噪比为 0dB 的条件下求出的，下面是经 LPC 平滑后的包络．由图可见，在信噪比为 0dB 的条件下，富里叶表示已经面目全非，而 EIH 表示却几乎没有受到影响，显示了很强的鲁棒性．

EIH 表示用于计算机语音识别的实验结果如图 14 所示，图 14a 是比较 EIH 表示与富里叶频谱用于语音识别的比较系统框图．把求得的 EIH 表示进行 DFT 逆变换以求得自相关系数，由此求得前 19 个线性预测系数，而对富里叶频谱求出前 9 个预测系数．语音资料是 26 个英文字母，10 个英文数字和 3 个控制符号"stop"、"error"、"repeat"（2 名男