第 25 卷第 6 期　　声 学 学 报　　Vol. 25, No. 6
2000 年 11 月　　ACTA ACUSTICA　　Nov., 2000

基于决策树的汉语三音子模型

高 升　徐 波　黄泰翼
(中国科学院自动化研究所　模式识别国家重点实验室　北京　100080)

1999 年 6 月 18 日收到

摘要　基于决策树理论的上下文相关声学模型在英语语音识别中已经得到了比较深入的研究和应用，但在汉语语音识别中的应用则研究的比较少。本文基于决策树理论建立了汉语语境相关模型－三音子模型，讨论了决策树建模所要解决的几个重要问题：(1) 基本建模单元集的选择，(2) 音子类别集的设计，(3) 评估函数的选择，(4) 停止准则的选择，(5) 决策树的建立和三音子模型的生成，本文着重分析了两种不同建模单元的性能：对音子类别集的设计提出了一些一般性的准则，并对我们设计的类别集进行了统计分析；分析了三音子模型在语音库的覆盖程度。实验结果表明，基于决策树的三音子声学模型建立的识别系统与双音子声学模型系统比较，误识率下降了 24.7%。

PACS 数：　43.70

Ttiphone models for mandarin speech recognition based on decision tree

GAO Sheng　XU Bo　HUANG Taiyi
(*National Laboratory of Pattern Recognition, Institute of Automation* Beijing 100080)

Received Jun. 18, 1999

Abstract　Context–dependent acoustic model based on decision tree has been deeply investigated and applied in western language speech recognition. But in Mandarin speech recognition, diphone model was more popular and little attention was paid to triphone in the past. In this paper triphone model based on decision tree was proposed. and some key problems are discussed which must be solved when build triphones, such as how to choose the basic model unit set, how to design the question set, how to choose evaluation function, the choice of stop criterion, and how to build the decision tree and triphone model. The experiments showed that error rate was reduced by 24.7% in continuous speech recognition system based on triphone compared with one based on diphone model.

引言

在连续语言中，语流内的一连串音紧密连接，发音部位和发音方法不断改变，相互之间难免发生影响，由此产生各种各样的音变现象，例如同化作用(不相同的音在语流中相互影响而变得发音相同或相似)，异化作用 (相同或相似的音在语流中接近时变得发音不相同或不相似)，增音 (语流中两个音之间增添一个音)，减音 (语流中某些应该有的音没有发出声音来) 和连读变调等现象。语言学家很早就注意到连续语音中的这种语流音变现象，但在语音识别中这种现象也只是近年来受到越来越多的关注和研究。

音变现象在连续语流中是普遍存在的，当同一音素出现在不同的上下文语境时，它的声源的激励方式和声道的调制方法是不一样的。因此，在选择模型基元时就不能忽视它的影响。当采用基于上下文语境的基元建模时，所获得的声学模型会更加精细，模型的数目也比相应的上下文无关的模型数目要多。精细的声学模型将会更加精确地描述语音信号的特征，在识别阶段可以根据模型基元所出现的上下文语境的不同，选择不同的上下文相关的声学模型；而在采用上下文无关的声学模型时，不管上下文语境如何，都采用同一种模型。两者相比，显然前者比后者要优很多，实验证明也确实如此 [4,6]。

对于连续语流中的某个音素，可能同时受到来自它前面 n 个音素或后面 m 个音素的影响。如果音素集合中包括 N 个音素，则总共可能有 N^{n+m+1} 个上下文相关的音素，对如此多的声学模型建模是不可

能实现的。当然，有些组合是不可能存在的，但即使这样也必须考虑一些方法以减少模型数目。其中，一种直接的方法是我们仅考虑来自前面一个或后一个音素的影响，同时根据一些准则对上下文相关的音素进行聚类。聚类的方法有很多种，一种是经验确定法，即根据一些语音学知识来确定哪些音素的影响是可以合并的，从而把某些上下文相关的音素合并成一类建模。例如，我们可以依据四呼规则把韵母分成开口呼、齐齿呼、合口呼和撮口呼四类，每一类对前面声母的影响都可以认为是相同的，这样每一个声母的后续上下文语境就分成了四类，相应地把它的所有上下文有关的声母聚类成四个上下文相关的声母。当然，根据声韵母之间的连接约束关系，其中某些声韵连接在音节表中不存在。另一种是数据驱动和专家知识引导的聚类，如自下而上的合并法 (Bottom-up) 和自上而下 (Top-down) 基于决策树的分裂法。

决策树方法是基于数据驱动和专家知识引导的分类法。其基本思想是首先根据专家知识精心设计音子类别集，选择相似度度量方法及决策树停止分裂准则；在决策树每次分裂时，都要从音子类别集中选择一个最优的问题 (即根据该问题分裂时所获得的相似度的增加值最大)，并按照该问题进行分裂。基于决策树理论的分类方法具有如下优点：(1) 它是数据驱动的，分类的数目可以根据数据的多少以及预设的门限设定；因此，声学模型的数目可以根据训练数据库的大小进行调整，以使估算出来的模型参数比较鲁棒。(2) 专家知识，如语音学和语言学知识可以很方便地与模型的分类过程集成在一起，从而可以保证声学模型更加准确。IBM 的 Bahl 等人在文献 1 中提出基于决策树的建模方法，文献 3 采用基于决策树的方法建立上下文相关模型，文献 2 采用决策树建立了三音子 (Triphone) 模型。它们的实验证明基于决策树的上下文相关声学模型具有很好的性能。

在英语语音识别中，基于决策树建模的方法已经研究了很多，相对而言，这种方法在汉语语音识别中的应用研究的还比较少。在汉语语音识别系统中，目前广泛采用的还是双音子模型 (Diphone)，基于决策树的三音子模型 (Triphone) 的研究工作还比较少采用。三音子模型所要解决的主要问题在于：(1) 三音子模型的数目大大多于双音子模型，解码以及搜索都比较困难。(2) 如何设计一个好的音子类别集，以体现汉语语言学和语音学的一些规律。

本文的主要工作是建立基于决策树的三音子模型，以便提高声学模型的准确性，提高语音识别系统的性能。重点放在汉语音子集的选取，音子类别集的设计，评估函数的选择，决策树的建立以及三音子模型的建立。我们在汉语大词汇量连续语音识别中的实验表明：基于决策树的三音子模型识别器与双音子模型识别器相比，性能有很大提高，误识率平均下降 24.7% 。

1　决策树的建立

在我们的研究中，选用二叉树做决策树。在建立决策树的时候，必须考虑如下几个问题：汉语音子集的选取，音子类别集的设计，评估函数的选择，二叉树停止分裂的准则。通过优化上述参数，可以获得更好的性能。

1.1　汉语音子集的选取

汉语是一种单音节结构的语言，由 22 个声母 (含零声母) 和 37 个韵母组成。如果以声韵母作为建模单元，就产生了第一个汉语音子集，记为 P' 。它包含 24 个声母模型 (零声母分成三类)，37 个韵母模型和一个静音模型。另外，汉语的 37 个韵母可以根据四呼规则分成四类。这样，22 个声母可以根据其后所跟的韵母类别的不同分成四类声母，每一类我们称之为细化声母，它由声母及其后所跟韵母的韵头组成。因为细化声母已经考虑了韵母的韵头部分，所以可以把一些韵母合并。例如，/an/ 和 /ian/ 可以并成韵母 /an/，这样，音节 /xian/ 可以表示成 /xi/ 和 /an/，而不象采用建模单元集 P^1 时表示成 /x/ 和 /ian/ 。由此产生了第二个建模单元集，记为 P^2 。它包括 53 个细化声母模型，23 个韵母模型和一个静音模型。以上两个建模单元集的组成见表 1 。

基于这两种不同的汉语音子集，可以分别建立音子类别集、决策树以及三音子模型，具体的算法将在下面介绍。为了比较基于这两种不同汉语音子集所产生的三音子模型的性能，我们进行了初步的实验对比。实验表明：这两种方法建造的识别器的正识率基本上差不多，但是，采用 P^2 时插入和删除引起的错误要多于 P^1 汉语音子集。原因可能由于 P^2 的韵母进行了合并，相对于未合并的 P^1 来说，降低了由此产生的韵母三音子模型之间的区分能力。不过，P^2 汉语音子集的优点也在于它的韵母单元数目少，所以当考虑声调特征时，有声调模型的数目少，可以减少模型的复杂度。因为在下面的研究中，并没有考虑声调特征，所以下面的论述和实验结果都是基于 P^1 建模单元集。

表 1 建模单元集的组成

建模单元集	前半音节	后半音节	静音
P^1	null,y,w,b,d,g,p,t, k,f,s,sh, x,h,z,zh,j,c,ch,q, r,m,n,l	a,o,e,er,ai, ei,ao,ou,an,en, ang,eng,i,i1,i2, ia,ie,iao,iou,ian, in,iang,ing,u,ua,uo,uai,uei, u,an,uen,uang,ong,v,ve, van,vn,iong	silence
P^2	null,y,yv,w,b,bi,bu,c, cu,ch,chu,d,di,du,f,g,gu,h, hu,ji,jv,k,ku,l,li,lu,lv,m,mi,mu,n,ni,nu,nv,p,pi,qi,qv,r,ru,s,su,sh,shu,t,ti,tu,xi,x,v,z,zu,zh,zhu	a,ai,an,ang,ao,e,ei, en,eng,er,i,i1,i2,ie,iu, in,ing,o,ong,ou,u,vn,v	silence

注: (1) /y/, /w/ 表示以半元音开头的零声母音节的头部, /null/ 表示其他零声母音节的头部。

(2) /i1/, /i2/ 分别表示 /i/ 前面的声母是卷舌和非卷舌音的两种情况, 其他情况下的 /i/ 仍记为 /i/ 。

1.2 音子类别集

音子类别集的设计是建立决策树时必须重点考虑的问题之一, 花费些时间和精力都是值得的, 同时音子类别集的设计也是一个反复实验、分析和修正的过程。下面介绍一下我们设计音子类别集时的一些基本原则。

设计音子类别集时的一个考虑是音素之间的相似性, 这种相似性是指发音方式或发音部位的相似, 我们认为相似的音素在上下文中对其他音素会产生相似的影响。在汉语语音学中, 声母一般根据发音方式或发音部位分类。根据发音方式的不同, 可以把声母分成塞音、塞擦音、擦音、鼻音、边音和通音; 而根据发音部位的不同, 可以把声母分为双唇音、唇齿音、舌尖音、卷舌音、舌面音和舌根音。韵母的分类一般仍旧根据韵头的不同分成开口呼、齐齿呼、合口呼和撮口呼四类, 在开口呼中, 又可根据 /a/ 、 /o/ 和 /e/ 音位的不同分成更小的类; 此外, 韵母对下一个音节声母的影响主要是由韵母的最后一个音素引起的, 因此, 韵母还可依据韵母最后一个音素的不同进行分类。例如最后一个音素是 /a/ 的可以可以规成一类, 是鼻音 /n/ 和 /ng/ 的可以化成一类, 还可以把韵头和韵腹相同的韵母归成一类。在每一类的内部, 又可根据音位的细小差别做进一步的分类。总之, 考虑了各种可能的分类, 各类之间的音素相互交叉和包容。这些分类方法充分利用了语音学已有的研究结论, 语音的声学特征和发音的声学特征基本一致的。

在本文中, 我们仅考虑左边或右边的基本建模单元对中间建模单元的影响。具体来讲, 我们采用声韵母作为基本建模单元, 根据汉语拼音表确定的声韵母之间的连接关系以及前面所确定的相似性原则设计音子类别集。音子类别集中的每一个分类都是一类具有相似性的声母或韵母的集合, 表明该分类中的每一个声母或韵母对其他韵母或声母的在上下文中的影响是相同的。因此, 音子类别集的设计将直接影响决策树的分类结果。所有的分类可以根据其组成分成两大类, 即由声母构成的分类和由韵母构成的分类。如果中间建模单元是声母, 那么因声母的左边只可能出现静音或韵母, 而声母的右边只能出现韵母, 所以对于声母来讲, 它的左边只能是有关静音或韵母的分类, 而右边只能是有关韵母的分类。如果中间建模单元是韵母, 那么它的左边只能是有关声母的分类, 而右边是有关静音或韵母的分类。此外, 音子类别集中有关语境的分类, 对声母构成的分类是对称的, 对韵母构成的分类基本上是对称的。这样, 根据汉语语音学知识和一些初步的实验比较, 我们最终设计了一个有 78 个分类的音子类别集, 其中有关左边上下文的分类有 37 个, 右边上下文的分类有 41 个。下面我们给出一些分类的实例:

根据声母发音方式的不同, 有如下分类:

例 1: 塞音 /b/, /d/, /g/, /p/, /t/, / k/ 。

根据声母的发音部位, 如:

例 2: 卷舌音 /zh/, /ch/, /sh/, /r/ 。

对于由韵母构成的分类, 如果它属于左边的分类, 我们主要考虑韵母的最后一个音素对后面声母的影响; 如果它属于右边的分类, 我们主要考虑韵母的第一个音素对前面声母的影响。例如:

例 3: 第一个音素是 /a/ /a/, /ai/, /ao/, /an/, /ang/- 右边分类。

例 4: 最后一个音素是 /a/ /a/, /ia/, /ua/ - 左边分类。

一个分类包容另一个分类，如下面例 5 包容例 6, 即例 6 中的所有声母都是例 5 的元素。

例 5: 塞擦音 /z/, /zh/, /j/, /c/, /ch/, /q/ 。

例 6: 不送气的塞擦音 /z/, /zh/, /j/ 。

1.3 决策树停止分裂的准则

决策树停止分裂的准则是为了确保树叶节点有足够的训练数据来重估模型参数。在本文中，我们根据经验设定一个最低门限，如果某个树节点的样本数据低于这个门限，就把该树节点标记为叶节点。

1.4 评估函数

评估函数是为了度量样本之间的相似度, 它可以是任何一种距离测度, 如均方根距离函数和似然比函数等。当决策树的某一个节点 (称之为父节点) 分裂成两个节点 (称之为子节点) 时, 似然度会增加。假设 L_{parent} 表示父节点的相似度, $X=\{X_1,X_2,\cdots,X_N\}$ 表示父节点有 N 个样本。假设 L_{child1} 和 L_{child2} 分别表示两个子节点的相似度, $X^1=\{X_1^1,X_2^1,\cdots,X_{N1}^1\}$ 和 $X^2=\{X_1^2,X_2^2,\cdots,X_{N2}^2\}$ 分别表示两个子节点的样本, 并且 X 、 X^1 和 X^2 之间有如下关系 $X=X^1\cup X^2$, $X^1\cap X^2=\Phi$ 。在本文中，每个样本就是一个 M 维的特征矢量。假设相似度的增加以 Δ 表示，则:

$$\Delta = L_{\text{child1}} + L_{\text{child2}} - L_{\text{parent}}. \tag{1}$$

1.5 决策树的生成

在本文中，我们采用输出分布共享的隐马尔可夫 (Hidden Markov Model, 简称 HMM) 模型，基本建模单元声学模型的每一个输出分布都对应着一棵决策树 (二叉树) 。为了获得每一个输出分布所包含的训练样本, 我们基于基本建模单元 P^1 采用 Baum-Welch 算法训练出 62 个 HMM 模型，然后采用维特比算法对数据进行切割，从而得到每一个输出分布的训练样本，基于这些样本以及我们前面所设计的音子类别集、评估函数和停止准则就可以采用如下的算法生成一棵决策树。

决策树生成算法:

步骤 1: 选择某一个模型的某个输出分布，记样本集为 X 。定义开始节点为根节点，包含样本集中的所有样本，标记根节点为 “没有处理过” 。

步骤 2: 从所有节点中选择一个 “没有处理过” 的节点，如果当前节点所包含的样本数小于停止门限，则记该节点为叶节点。否则，计算该节点的相似度, 然后对音子类别集中的每一个分类都计算该节点如果按照这个分类分裂为两个子节点时 (一个子节点对该分类回答 “Yes”, 一个子节点回答 “No”) 两个子节点的相似度，然后根据式 (1) 计算相似度的增加值。选择使得相似度增加最大的分类对该节点进行分裂，并记录该分类，标记该节点 “已经处理过”, 同时标记新产生的节点为 “没有处理过”，计算新产生节点的相关参数。

步骤 3: 如果所有的节点都已经处理过，则决策树已经形成；否则，转入步骤 2 。

图 1 给出了一个二元决策树的示意图。

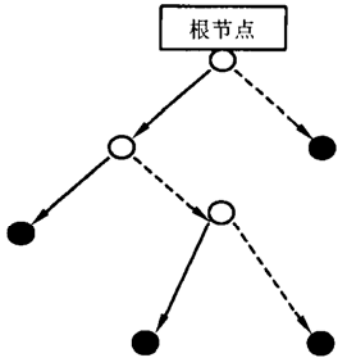

图 1　决策树例子

注:　(a) 实心圆表示叶节点，空心圆表示非叶节点

(b) 虚线表示对分类的回答是 “No”，

实线表示对分类的回答是 “Yes”

2　三音子模型

在决策树生成后，我们就可以生成和训练三音子 HMM 模型了。假设 $\lambda_p(A,B)$ 表示基本建模单元 p 的声学模型， $\boldsymbol{A}$ 表示转移概率矩阵， $\boldsymbol{B}$ 表示具有 N_λ 个输出分布的矩阵。因为每一个输出分布对应一棵决策树，则 $\lambda_p(A,B)$ 对应 N_λ 棵决策树。假设 $l_i^p=\{\text{leaf}|\text{leaf}\in \text{第 } i \text{ 棵决策树}\}$, $i=1,2,\cdots\cdots,N_\lambda$ 表示 $\lambda_p(A,B)$ 模型第 i 棵决策树的叶节点集合，$\lambda(p,p_L,p_R)$ 表示基本建模单元 p 的一个三音子模型，其左边上下文是基本建模单元 p_L，右边上下文是基本建模单元 p_R 。由决策树产生三音子模型 $\lambda(p,p_L,p_R)$ 的算法如下:

步骤 1: 从 $\lambda_p(A,B)$ 对应的 N_λ 棵决策树中选择一棵决策树，并从树根节点开始搜索。

步骤 2: 对于当前节点，如果它是叶节点，则记录该节点并终止搜索，同时，该叶节点概率密度分布

函数作为 $\lambda(p,p_L,p_R)$ 的一个输出分布。检查该节点分裂成两个子节点时所依据的分类，即有关上下文的一些信息。如果该三音子模型左边上下文 p_L 或右边上下文 p_R 与该节点分裂时依据的上下文信息一致，则检查该节点的下一个 Yes 节点；否则，检查该节点的下一个 No 节点。

步骤 3: 循环步骤 2, 直到搜索到一个叶节点。

步骤 4: 如果 N_λ 棵决策树搜索完毕, 终止搜索; 否则，返回步骤 1 。

如果 $\lambda(p,p_L,p_R)$ 模型的 N_λ 个输出分布都已经得到, 则中间 p, 左右上下文分别是 p_L 和 p_R 的一个三音子模型就产生了 (转移概率矩阵可以继承 $\lambda_p(A,B)$ 模型的) 。可以把这些三音子模型作为初始化模型进一步训练. 不过，这样得到的三音子模型数量庞大，在本文选择的汉语音子集的情况下，共有 26340 个，全部进行训练根本不可能。由于采用了决策树，许多三音子模型具有相同的输出分布，所以可以把具有相同输出分布的三音子模型合并，形成新的三音子模型，三音子模型的数目可以通过控制决策树叶节点的数目来实现。通过合并，可以大幅度减少三音子模型的数目，提高模型的鲁棒性。

3 实验结果及分析

采用前面介绍的方法，我们产生和训练了三音子模型，并与双音子模型进行了实验对比。所有的实验都是在我们实验室的大词汇量汉语非特定人连续语音听写机系统上进行的。语音识别器的特征参数由 12 维 MEL 倒谱参数和一维归一化的能量参数以及它们的一阶和二阶差分特征构成的 39 维特征参数。

训练语音库和测试语音库均由 “863” 提供，训练语音库由 116 名男生和 116 名女生录制的连续句子和孤立词；测试语音库采用 1998 年 4 月 “863” 听写机评测时录制的 6 男 6 女共 240 个连续句子。测试分男女进行，分别测试男声声学模型和女声声学模型，解码中没有应用语言模型，仅仅是声学层次的识别结果，实验结果如表 2 所示。

表 2 三音子模型识别结果

性别	模型数目		输出分布数	音节正识率 /%
男	三音子模型	4485	3670	83.58
	双音子模型	138	549	79.12
女	三音子模型	4485	2331	86.26
	双音子模型	138	549	80.88

3.1 音子类别集的统计分析

如第一部分所述，决策树是按照音子类别集构造的。因此，音子类别集的设计直接影响到决策树。但到目前为止, 怎样设计音子类别集以及如何评价什么样的音子类别集是最优的还没有一个理论上的指导。因此，我们只能根据目前所掌握的语音学、语言学以及其他一些经验知识设计音子类别集。为了进一步优化音子类别集，我们观察了音子类别集中不同分类的重要程度，并对它们的出现程度进行了分析，这些工作将会对音子类别集的优化提供帮助。下面给出一些分析结果。

(1) 对声母和韵母基本建模单元来讲，左边和右边是否静音出现的频度都很大。这说明，上下文是静音的情况与上下文是语音的情况在声学特征和语言学上有很大的不同。

(2) 对于声母基本单元，其左边分类和右边分类的出现频度基本相同, 分别为 0.55 和 0.45 这说明, 声母受到左边韵母和右边韵母的影响差不多。此外，以鼻音结尾的韵母对其右边声母的影响要大。而有些分类在决策树中根本没有提问过，这当然可能是因为其样本在训练库中很少出现，但同时说明它对声母的影响很小，如撮口呼韵母。

(3) 对于韵母基本单元，其右边分类的出现频度是左边分类出现频度的二倍，分解 0.67 和 0.33 。这说明，在连续语流中，韵母受到右边声母或静音的影响要远远大于左边声母的影响。此外，右边声母按照是否塞音、摩擦音、塞擦音、舌尖音和卷舌音设计的分类出现频度仅次于右边是否静音这个分类。这说明，对某些声母，按照发音部位分类可能比按照发音方式分类要更好一些。同样，有些分类在决策树中根本没有提问过。

3.2 三音子模型识别结果分析

表 2 实验结果表明：三音子模型要优于双音子模型。采用三音子模型后，音节误识率平均下降了 24.7% 。但同时也发现，声学模型数目及输出分布数目也都成倍增长。因此在现有训练条件下，每个三音子模型在训练语音库中的覆盖次数将大大减少，从而降低模型参数估算的准确性，可能在一定程度会降低三音子模型的性能。我们对三音子模型在训练语音库和测试语音库的出现频度进行了分析，以男声三音子模型 (3782 个三音子模型) 为例，在训练语音库中三音子模型出现 10 次以上的有 579 个，大约占全部模型数目的 15.3% 。因此，训练语料库的覆盖面是很低的，大部分三音子模型没能得到充分训

练，这说明目前采用的训练语音库的设计还存在问题，不能适合建立三音子模型的需要，应该重新设计以增加三音子模型的覆盖面。

4　结论

本文基于决策树的理论，建立了汉语三音子模型，在汉语大词汇量连续语音识别系统的实验表明：三音子模型比双音子模型更加准确，以此建立的识别器具有更好的性能，在我们的实验中误识率平均下降了 24.7%。对于音子类别集的设计，我们给出了一些设计原则并且基于所使用的训练语音库给出了一些统计分析结果，这会对音子类别集的优化设计提供一些帮助。当然，三音子和双音子相比较，其主要缺点在于模型数目太庞大，在加上语言模型后，解码复杂而且速度慢，但这可以通过研究快速搜索算法来解决。此外，根据我们的一些初步实验，三音子模型数目只要保持一定的数量就可以，也就是说，三音子模型数目超过这个数目，识别率并没有明显提高。例如男声三音子模型从 6368 个减少到 4000 个左右时，识别率降低在 1% 左右。因此，模型数目是有很大优化余地的。

参　考　文　献

1　Bahl L R, Souza P V de, Gopalakrishnan P S *et al.* Decision tree for phonological rules in continuous speech. ICASSP 89, Glasgow, 1989: 185—188

2　Reichl W, chou W. Decision tree state tying based on segmental clustering for acoustic modeling. ICASSP 98, 1998: 801—804

3　Mei-Yuh Hwang, Xuedong Huang, Alleva F A. Predicting unseen triphones with senones. *IEEE Transactions on Speech and Audio Processing*, 1998; **4**(6): 412—419

4　Bin Ma, Taiyi Huang, Bo Xu *et al.* Context-dependent acoustic models in Chinese speech language. ICASSP'96, USA, 1996

5　林　焘，王理嘉. 语音学教程. 北京：北京大学出版社

6　徐　波，张　亮，黄泰翼. 基于决策树方法的语境有关 HMM 建模. 第八届全国声学学术会议，1998: 421—424