第六届全国人机语音通讯学术会议，267-271 页，2001 年 11 月 20-22 日，深圳

汉语连续语音识别中声学模型基元比较：音节、音素、声韵母

李净，徐明星，张继勇，郑方，吴文虎，方棣棠

语音技术中心，智能技术与系统国家重点实验室，
清华大学计算机科学与技术系, 北京, 100084
[lijing, xumx, zjy, fzheng, wuwh]@sp.cs.tsinghua.edu.cn,
fangdt@tsinghua.edu.cn
http://sp.cs.tsinghua.edu.cn

摘要

本文研究的是汉语连续语音识别中声学模型基元的选择问题。根据汉语语音的特点，本文分别采用音节、音素和声韵母等三种语音识别基元进行声学建模。为了描述连续语音中的协同发音现象，本文针对音素和声韵基元，设计了相应的问题集，利用基于决策树的状态共享策略建立了上下文相关音素模型（Triphone）和上下文相关声韵模型（TriIF），并对几种声学基元进行了对比。实验结果表明，对于上下文无关模型，音素和声韵模型都要劣于音节模型，而对于上下文相关模型，Triphone 和 TriIF 模型与音节模型相比，识别性能有了很大提高，其音节误识率分别降低了 8.5%和 23.6%。

1. 引言

声学建模是连续语音识别中声学层面处理的关键步骤。声学模型用来描述识别基元对应的特征矢量序列的产生过程。通过声学建模，可以估计待识别特征矢量序列所对应的语音识别基元，从而完成特征矢量序列到语音识别基元的识别转换。

基元的选择是声学建模中一个基本而重要的问题。在汉语连续语音识别中，可以选择的基元包括：词（Word）、音节（Syllable）、半音节（Semi-Syllable）、声韵母（Initial/Final）、音素（Phone）等。识别基元的选择一般是基于语音学知识的，但是，基元也可以通过数据驱动的方式来产生，使用这种方式确定的基元可能在语音学上没有什么明确的意义，但也可以达到很好的性能。

对于词，在小词表语音识别系统中，或者命令与控制（Command & Control）系统中，使用词作为识别基元是适当的。但是，在连续语音识别中将词作为识别基元是不合适的。首先，在连续语音识别系统中，词条的数目比较多，一般都要使用几千或者几万条词条，所以声学模型的规模必然很大。这不但会增加存储的开销，还会极大地增加搜索的复杂度。其次，当词表以外的词条，即 OOV（Out Of Vocabulary）问题出现时，声学模型处理起来比较困难。第三，要对这么多基元进行训练，必然需要一个很大的数据库，并且要尽量覆盖词表中的词条，这一点是很难达到的。所以，在汉语连续语音识别系统中，采用类似于词这样较长的语音段作为识别基元是不合适的。

对于音节，在汉语中，无调音节约有 400 个，如果考虑音调，有 1300 多个有调音节[1]。在进行上下文无关的声学建模时，使用有调或者无调音节是可以的，而且还可以取得相当好的性能，因为音节作为识别基元时，它很好地刻划了音节内部的变化。但是，在连续语音识别中，音节间的协同发音现象是比较严重的，因此，必须采用适当的方式来描述这种现象。一般地，上下文相关信息应在声学建模中加以考虑，这样，识别基元就会变成上下文相关的基元。如果采用音节作为识别基元，当考虑上下文信息时，基元数目会变得非常庞大，这将会使声学模型的规模变得无法接受。同时，由于基元数目过大，也会引起训练数据稀疏的问题，从而难以对模型参数给出较为准确的估计。所以，在进行上下文相关建模时，不适宜采用音节模型。

音素在汉语中有三十多个（本文中定义的音素数目为 35 个）。音素基元在英语连续语音识别系统中得到了广泛的应用，并取得了很好的识别性能[2][3]。由此可见，音素也是一个很好的选择。但音素并没有反映出汉语语音的特点，而且，相对于声韵母，音素显得更加不稳定，这一方面给手工标注带来了困难，同时，也给声学描述带来困难。

对于半音节和声韵母，它们在形式和数量上十分接近。半音节就是将音节分为两部分，而声韵母的划分更依赖于汉语语音学的知识。可以说，声韵母基元是适合汉语特点的一种识别基元，使用这种基元，还可以有很多语言学知识可以利用，从而进一步提高声学模型的性能。声韵母作为识别基元具有以下优点：

- 汉语中的汉字是单音节的，而汉语中的音节是声韵结构的，这种独特而规则的结构，使对音节、以及词条的表示变得比较规则和统一；
- 使用声韵母作为识别基元，上下文相关信息也变得比较确定。比如，与声母相接的只能是韵母或者静音，而与韵母相接的也只能是声母或静音，而且，韵母左边相接的声母只能是与其搭配起来能够成汉语音节的那些声母。所以，上下文相关的声韵母基元的数目并不是基元数目的立方，而是远远小于这个数值的。

- 声韵母结构是汉语音节独特的一种结构，有很多关于声韵母的语音学方面的知识和研究成果可以被我们采用，以优化上下文相关声学模型。
- 选择声韵母作为基元，它的语音段长度，以及基元数目都是比较适当的。如果不考虑上下文信息，本文中定义的声韵母共有 59 个，其中声母 21 个，韵母 38 个。

同时，在连续语音中，协同发音现象是十分严重的，因此，要得到性能较高的声学模型，需要利用好上下文相关信息，即进行上下文相关建模。基于决策树的状态共享策略已经广泛的应用于连续语音识别中。本文也采用这种策略来进行上下文相关建模。

根据上面的分析，本文对适合汉语特点的三种基元：音节、音素、声韵母进行了对比实验，并利用基于决策树的状态共享策略进行了上下文相关建模，给出了对比结果。

2. 识别基元定义

2.1. 音节基元

选择音节作为识别基元是符合汉语语音特点的。本文定义了 418 个无调音节作为连续语音识别中的音节基元定义。如“a”，“ai”，“ang”，……，“zuo”等。

2.2. 音素基元集合

本节给出音素基元的定义[4]，其中辅音基元 22 个，元音基元 13 个，音素基元总数为 35，如表 1 所示。对于元音基元的定义，表 2 给出了较为详细的说明。

表 1: 音素基元定义

辅音基元 (22)	元音基元 (13)
b ,c, ch, d, f, g, h, j, k, l, m, n, ng, p, q, r, s, sh, t, x, z, zh	*aI, a, Ie, eI, eN, e, Ci, Chi, Bi, oU, o, u, v*

表 2: 元音音素基元定义

元音音素	定义
/aI/	在韵母“ai”,“an”中的音素“a”
/a/	在其它条件中的音素“a”
/Ie/	在韵母“ie”中的音素“e”
/eI/	在韵母“ei”中的音素“e”
/eN/	在韵母“en”中的音素“e”
/e/	在其它条件中的音素“e”
/Ci/	在音节“ci”，“si”，“zi”中的音素“i”
/CHi/	在音节“chi”，“shi”，“zhi”中的音素“i”
/Bi/	在其它条件中的音素“i”
/oU/	在韵母“ou”中的的音素“o”
/o/	在其它条件下的音素“o”
/u/	元音音素“u”
/v/	音节 “yu” 中的元音音素

每个音节由若干音素基元串组成。比如音节“ai”用音素“aI”和“Bi”构成，音节“bin”由音素“b”，“Bi”和“n”构成。

2.3. 标准声韵母基元集合（IF）

表 3: 标准声韵母基元定义

声母基元 (21)	韵母基元 (38)
b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, z, c, s, r,	*a, ai, an, ang, ao, e, ei, en, eng, er, o, ong, ou, i, i1, i2, ia, ian, iang, iao, ie, in, ing, iong, iou, u, ua, uai, uan, uang, uei, uen, ueng, uo, v, van, ve, vn*

表 3 给出了标准的声韵母基元定义，一共有 21 个声母和 38 个韵母。在这种定义下，有些带有零声母的音节只对应一个韵母基元，而没有对应的声母基元。

2.4. 扩展的声韵母集合（XIF）

可以认为每个音节都是由两部分组成的，分别对应其声母部分和韵母部分。根据汉语音节的这种特点，本文定义了六个零声母{_a, _o, _e, _I, _u, _v }，这样就得到了扩展的声韵母基元集合。扩展的声韵母定义见表 4。

表 4: 扩展的声韵母基元定义

声母基元 (27)	韵母基元 (38)
b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, z, c, s, r, _a, _o, _e, _I, _u, _v	*a, ai, an, ang, ao, e, ei, en, eng, er, o, ong, ou, i, i1, i2, ia, ian, iang, iao, ie, in, ing, iong, iou, u, ua, uai, uan, uang, uei, uen, ueng, uo, v, van, ve, vn*

当使用标准的声韵母基元集合时，有一些音节只有韵母部分，而没有声母部分。所以，当考虑上下文相关信息时，这些韵母既可以搭配声母，又可以搭配韵母，因此，上下文相关声韵母基元数目会很大。而使用扩展的声韵母基元集合时，韵母的上下文只能是声母或静音，声母的上下文只能是韵母或静音，所以，上下文相关基元数目会明显减少。在本文的实验中，如果采用扩展的声韵母基元，上下文相关基元数目约有 3 万个，而使用标准的声韵母集合，上下文相关基元数目则超过 10 万。

另一方面，通过实验也可以看出，如果没有引入零声母，那些带有零声母的音节将会和其它音节的韵母部分共享模型参数，从而在识别中增加了许多插入错误。

从后面的结果中可以看出，扩展的声韵母基元要优于标准的声韵母基元，所以，在进行上下文相关建

模时，本文采用扩展的声韵母基元集合。后文中的TriIF 指的是由扩展的声韵母产生的上下文相关基元。

3. 基于决策树的状态共享策略

如前所述，在连续语音中，协同发音现象是十分严重的，因此，建立上下文相关模型来描述这种现象是很有必要的。对于音节模型，由于基元数目过多，当考虑上下文信息时，基元数目会变得非常庞大，很难进行上下文相关建模。而对于音素和声韵模型，基元数目适当，则可以进行上下文相关建模。

对于音素和声韵母基元，在进行上下文相关建模时，由于基元数目庞大，训练数据就变得稀疏，一般会有一半左右的基元没有训练数据。因此，必须进行模型或参数的共享来解决这个问题。本文中使用基于决策树的状态共享策略。

基于决策树的状态共享策略已经广泛地应用于改善大词表连续语音识别系统的声学模型性能[5][6][7]。这种方法与数据驱动方法相比最大的优点就是对训练数据稀少的基元和没有训练样本的基元能够给出适当的参数估计。其次，决策树方法具有的另一个优点是可以调整分类数目，以适应声学模型的鲁棒性和训练数据稀疏的矛盾。本节介绍此技术中的两个关键问题，一个是问题集的设计，另一个是决策树的构造。

3.1. 问题集的设计

决策树是一个二叉树，每个结点都绑定着一个“Yes/No”问题，所有允许进入根结点的状态要回答结点上绑定的问题，根据回答的结果选择进入左枝还是右枝。最后，每个进入跟结点的状态都会根据对一系列结点问题的回答进入适当的（也是唯一的）一个叶子结点。进入同一个叶子结点的状态会被认为是相似的而共享起来。而问题集就是供决策树构造使用的，结点分裂时选中的那个问题，就与此结点绑定，从而决定哪些基元的哪些状态被共享起来。问题集的好坏会影响到上下文相关模型的性能。

本文中使用的问题集是基于语音学知识的[8][9][10][11]。根据这些先验知识，中心基元的上下文（即中心基元左右两边相邻的基元）被划分为若干类，每一类作为一个问题。本文针对音素和声韵基元，设计了各自的问题集。

以我们提出的 TriIF 基元为例，作为问题的声母基元类有：

响音（Sonorant）	{m, n, l}
塞音（Stop）	{b, d, g, p, t, k}
唇音（Labial）	{b, p, m, f}
塞擦音（Affricate）	{z, zh, j, c, ch, q}
……	

作为问题的韵母基元类有：

前高（HighFront）	{i, u, v}
开口 n（Open_n）	{an, en}
开口 ng（Open_ng）	{ang, eng}
……	

在每个结点分裂时，适当的问题会被提出，如“左边的基元是响音吗？”或者“右边的基元是唇音吗？”，如果是，则会分配到“Yes” 结点，反之被分配到“No”结点。所有针对此基元设计的问题都会在这里被问到，而“最佳”的问题将会被选中，作为此结点对应的问题。所谓“最佳”是指符合分裂准则的那个问题。本文中使用的分裂准则是，选择分裂后似然分增加最大的那个问题。

3.2. 决策树的构造

一般地，首先将所有可能共享的状态放入一个状态共享池（Pool of States）中，然后根据一定的分裂准则（Split Criterion）进行逐级分裂，当满足一定的条件时，即满足停止分裂准则（Stopping Criteria）时，分裂过程停止。

本文中使用的决策树是基元相关且状态相关的，即，只有同一个中心基元的同一个状态才被放到同一个共享池中。不同基元，或者同一基元的不同状态不会被共享。

在构造决策树时采用的分裂准则是，选择分裂后似然分增加最大的问题作为本结点绑定的问题。决策树的停止分裂准则采用的是阈值的方法。即当分裂后的结点中训练样本数目少于一定数量时，或者，当本结点分裂后对数似然分数的增加小于一定的阈值时，停止分裂。当所有的结点停止分裂后，决策树生成，此时，所有叶子结点对应的状态参数被重估出来，作为落到本叶子结点的所有状态的共享参数。

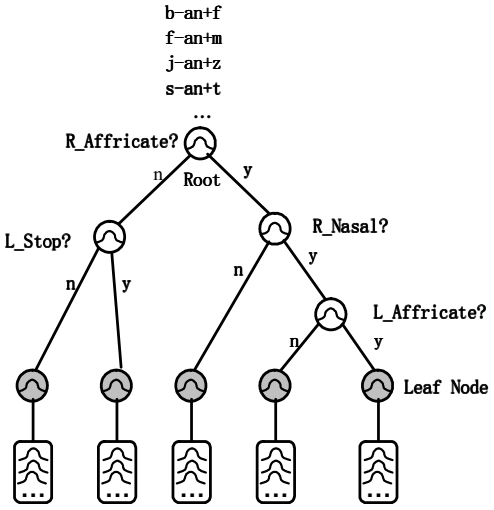

图 1. 决策树结构图

基于决策树的状态共享策略是基于知识和数据驱动方法的结合，它比较有效地解决了训练数据的稀疏问题。

4. 模型训练

本文使用隐含马尔可夫模型（HMM, Hidden Markov Model）来描述声学模型。对于音节模型，每个基元使用 6 个串联的状态来描述，每个状态只能驻留或跳转到相邻的下一个状态。对于音素和声韵基元，使用 3 个串联的状态来描述。

由于每一次结点分裂都要重新计算参数，所以，如果采用多混合的初始模型，计算量就会大得无法忍受。因此，在利用决策树进行状态共享时，使用的是单混合的初始模型，这样就可以通过参数本身直接重估出新的参数，而不用直接访问原始训练数据。这

样，构造出的状态共享的 Triphone 和 TriIF 模型是单混合的，即每个共享状态使用单个高斯混合来描述。而单混合的模型描述能力是有限的，因此，可以采用混合分裂的方式来增加混合数目。本文实验中，使用 8 混合的模型作为最终模型。

5. 实验结果

5.1. 实验条件

本文中使用的数据库是“863 数据库”中的男声数据库[12]。数据库中的句子是用略带口音的普通话读出的。数据库中共有 1560 个不同的句子，被划分为三组，分别称为 A，B 和 C 组。库中共有 80 个人的语音数据。全部语音数据的文本信息以及音节一级的标注信息都是已知的。标注信息的获得是利用手工标注和机器切分相结合的方法得到的。

本文中从数据库中选取 70 人的数据作为训练集合，剩余 10 人数据作为训练集合。所以，测试集合中的说话人都不在训练集合中。

实验中使用 42 维的 MFCC（Mel-Frequency Cepstrum Coefficients）作为特征参数，包含能量参数，以及一阶差分和二阶差分参数。

本文使用 HTK v2.2 工具进行模型训练[13]。测试结果用连续识别的音节正确率（Accuracy%）来进行评价。

5.2. 扩展声韵集合与标准声韵集合

表 5: 扩展声韵母模型与标准声韵母模型比较

模　型	音节正确率（Acc%）			
	1 混合	2 混合	4 混合	8 混合
音素模型（Phone）	29.30	37.71	42.94	48.27
标准声韵模型（IF）	38.31	44.86	50.05	54.79
扩展声韵模型（XIF）	43.02	50.85	56.12	60.28
音节模型（Syllable）	59.75	64.28	69.18	73.14

从结果中可以看出，扩展的声韵母基元集合性能优于标准的声韵母基元集合以及音素基元集合。同时，音节模型的识别率远高于音素和声韵基元模型，这是因为音节模型的基元数目远远多于前两者，它使用了更多的参数来描述模型，同时，采用音节基元时，音节内部的相关性已经得到了很好的描述，因此，音节模型的识别率较高。

如前所述，扩展的声韵母基元可以具有如下优点：（1）使发音字典变得规整，每个音节由一个声母和一个韵母组成，上下文关系比较简单和确定，使上下文相关基元数目极大地减少了，有利于建立上下文相关模型；（2）由于零声母地引入，减少了连续语音识别中带有零声母音节的插入错误。

5.3. 上下文相关模型性能

表 6: 上下文相关模型与音节模型性能比较

模　型	音节正确率（Acc%）			
	1 混合	2 混合	4 混合	8 混合
音节模型（Syllable）	59.75	64.28	69.18	73.14
音素模型（Triphone）	67.99	70.01	72.90	75.41
声韵模型（TriIF）	72.03	75.73	77.97	79.48

表 7: 上下文相关模型与音节模型规模比较

模　型	高斯混合数目
音节模型（Syllable）	20064
音素模型（Triphone）	84608
声韵模型（TriIF）	75936

从表 6 和表 7 中可以看出，基于决策树的状态共享策略应用于音素、声韵模型都可以取得很好的效果，这两种上下文相关模型的性能都要高于音节模型。对于 8 混合的模型，其音节误识率分别降低了 8.5%和 23.6%。相对于音节模型，声韵模型的性能改善尤为突出。从模型规模来看，上下文相关模型的参数数目大约是音节模型的 4 倍。但如果不使用状态共享策略，参数数目要远远大于此值，几乎不可能给出很好的估计。

6. 总结

本文根据汉语语音的特点，选择音节、音素、声韵母来进行声学模型训练，给出了音素、声韵母（包括扩展的声韵母）基元定义，并针对音素和声韵基元设计了适当的问题集，利用基于决策树的状态共享策略训练上下文相关模型。然后对使用三种基元训练的声学模型的识别性能和规模进行了对比。从中可以得到如下结论：

- 连续语音中协同发音现象十分严重，因此，进行相关性建模是很有必要的。从结果中可以看出，对于音素和声韵模型，进行上下文相关建模后，模型性能有了极大的提高。同时，与音节模型相比，上下文相关的音素、声韵模型的性能有了很大的提高，这与上下文信息的引入直接相关。
- 实验中采用的基于决策树的状态共享策略充分的利用了汉语语音学知识，并与数据驱动的方式结合，使上下文相关建模取得了很好的性能，且模型规模适当。
- 上下文相关声韵母基元（TriIF）是几种基元中最佳的选择，与音节模型相比，音节误识率降低了 23.6%，而 Triphone 模型与音节模型相比，误识率降低为 8.5%。这说明，声韵母是适合汉语语音特点的基元。

7. 参考文献

[1] 郑方，牟晓隆，徐明星，武健，宋战江，“汉语语音听写机技术的研究与实现”，软件学报，10（4）：436－444，1999

[2] Lee C., Rabiner L., Pieraccini R. and Wilpon J., “Acoustic modeling for large vocabulary speech recognition,” Computer, Speech and Language, 4, 127-165, 1990

[3] S.J.Young and P.C.Woodland, “Tree-based state tying for high accuracy acoustic modeling,” Proc. Human Language Technology Workshop, pp.307-312, March 1994.

[4] Bin MA and Qiang HUO. “Benchmark results of triphone-based acoustic modeling on HKU96 and HKU99 putonghua corpora,” International Symposium on Chinese Spoken Language Processing (ISCSLP’00), pp. 359-362, Oct. 13-15, 2000

[5] Bahl, L. R., de Souza, P.V., Copalakrishnan, P. S., Nahamoo, D. and Picheny, M. A., “Decision trees for phonological rules in continuous speech”, in *Proc. Int. Conf. Acoustics, Speech, Signal Processing’91*, Toronto, ON, Canada, May 1991, pp.185-188.

[6] Reichl, W. and Chou, W., “Decision trees state tying based on segmental clustering for acoustic modeling”, in *Proc. Int. Conf. Acoustics, Speech, Signal Processing’98*, pp.801-804.

[7] Reichl, W. and Chou, W., “Robust decision tree state tying for continuous speech recognition”, *IEEE Trans. Speech and Audio Proc*., 8(5): 555-566, 2000.

[8] 吴宗济，林茂灿，等，实验语音学概要。 北京：高等教育出版社。1989

[9] 曹剑芬，现代语音基础知识。 北京：人民教育出版社。 1990

[10] 吴宗济 (1997). 试论人－机对话中的汉语语音学. 世界汉语教学, 1997, 42(4): 3-20

[11] 罗安源 (2000). 田野语音学. 北京: 中央民族大学出版社. 2000

[12] Zheng, F., Song, Z.-J., and Xu, M.-X., “EASYTALK: A large-vocabulary speaker-independent Chinese dictation machine”, EuroSpeech‘99, Vol.2, pp.819-822, Budapest, Hungary (1999)

[13] Yong, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. and Woodland, P., “The HTK Book (for HTK Version 2.2)”, Cambridge University (1999)