文章编号:1004-9037(2013)02-0178-06

基于音位属性和边界信息的音素识别

许友亮 张连海 牛 铜

(解放军信息工程大学信息工程学院,郑州,450002)

摘要:在检测出音位属性的基础上,提出了一种基于音位属性后验概率的音素边界检测算法,并将音位属性与边界信息应用于基于条件随机场的音素识别。该方法首先计算得出相邻帧音位属性后验概率向量间的夹角,然后将夹角的极大值点所在的帧选为侯选边界,最后通过约束条件去除极值点中的错误边界。本文将音素边界与音位属性信息进行组合,作为基于条件随机场模型的识别系统的观测特征,实验结果表明,增加边界信息后,音素正确识别率有了显著提升。

关键词:音位属性;音素边界检测;自动语音识别;条件随机场

中图分类号:TP391 **文献标志码**:A

Phone Recognition Method Based on Phonological Attributes and Phone Boundaries

Xu Youliang, *Zhang Lianhai*, *Niu Tong*

(Institute of Information Engineering, The PLA Information Engineering University, Zhengzhou, 450002, China)

Abstract: A phone boundary detection method is proposed based on the phonological attributes posterior probability, taking these features and boundary information to analyze conditional-random-field-based phone recognition system. Firstly, the angles between posterior probability vectors of adjacent frames are calculated, and then the frames with the maximum angle are marked as the boundary candidates. Secondly, the false phone boundaries are removed through several restrictions in the boundary candidates. Finally, the combination of phonological attributes and phone boundaries is presented as the observation vectors of conditional random fields. Experimental results show that the accuracy rate of phoneme recognition is superior to the base system which only uses phonological attribute features.

Key words: phonological attributes; phone boundary detection; automatic speech recognition; conditional random field

引 言

语言学家通过分析机器识别和人类对语音感知(Human speech recognition, HSR)方面的差异,提出了基于事件检测的(Event detection-based)识别系统[1]。该方法通过模拟人耳的听觉感知过程,在检测各种声学事件的基础上,将各种语音学、语言学等知识引入识别系统中,从而提高语音识别系统的性能。在和各种语音学相关的知识中,音位属性特征(Phonological attributes features, PAF)的应用最广泛,相关研究主要集中于PAF的检测及建模方面。

在基于语音事件检测的系统中,音素边界的检测与应用成为研究的热点问题[2]。现有的音素边界检测方法主要分有监督和无监督两种,有监督的方法通过模型训练进行边界的检测,增加了系统的负担;而无监督的检测方法仅需要在了解先验知识的基础上,通过相应的规则进行判定,应用简单、灵活,得到了广泛应用[3]。文献[4]等在无监督的边界检测中,提出了3个优化目标函数,在边界误差为20 ms的指标上,取得了76.7%的正确检测率。

基金项目:国家自然科学基金(61175017)资助项目。
收稿日期:2011-11-10;**修订日期**:2012-01-04

文献[5]等在分析相邻帧间美尔频率倒谱系数(Mel frequency cepstral coefficient, MFCC)变化的基础上，定量地给出了音素边界与频谱幅度变化之间的关系，使边界误差在 20 ms 以内的检测率达到 75.7%。目前音素的边界信息已经被应用于语音识别中，文献[6]等提取出反映音素边界的声学特征参数，并与传统特征参数进行组合，在音素识别率上提高了 3.5%。

上述基于无监督的边界检测算法都是在声学层特征空间上进行的，而声学特征空间的冗余信息较多，对边界检测形成较大干扰。MFCC 可以看作底层的物理特征参数，而 PAF 是描述发音方法和发音状态的参数，可看做是高层次的语音描述参数，且这些参数相对稳定，基于 PAF 的边界检测是可行的。本文提出了一种基于 PAF 后验概率的边界检测方法，该方法首先得到 PAF 的后验概率，通过分析和比较相邻帧向量间的相关性，在后验概率空间上确定音素的边界。为了检验边界信息对于识别性能的影响，本文将边界信息与 PAF 进行组合，应用到基于条件随机场(Conditional random field, CRF)的识别系统中。

1　音位属性检测

本文使用的两种 PAF 分别为英语发音方式(Sound pattern of English, SPE)和支配音韵特征(Government phonology, GP)。SPE 着重从发声的角度描述语音的产生过程；GP 通过声学频谱分析而抽出音素中 11 种互补分量[7]，而这些分量的组合可以描述并区分所有音素。由于后验概率可作为传统声学层特征的补充，而神经网络在声学层特征与后验概率之间起着非线性变换的作用，将声学层特征映射为识别单元后验概率的同时，能够在一定程度上抑制特征中的冗余信息及噪声，从而将与识别单元相关的区分性信息保留在后验概率中[8]。本文首先通过神经网络得到 PAF 的后验概率，再采用 HMM 对其进行建模和识别。

本文采用基于复现时间延迟(Recurrency and time-delays neural network, RTDNN)思想的 NICO[9]神经网络进行 SPE，GP 的检测，其输出结果为对应 PAF 的后验概率。图 1 为 RTDNN 的系统结构框图，其中 μ_i 和 μ_j 为输入层节点，为隐含层权重，λ 为输出节点，F 为隐含层激励函数。RTDNN 的最大特点是在 MLP 的隐含层引入了

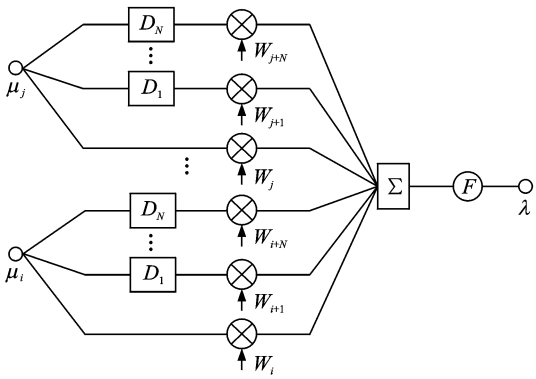

图 1　RTDNN 系统结构

时间延迟因子 D_i，即当前时刻的输入经过若干时刻延迟后，对后续的判决产生影响，相对于常规的 MLP，可以在 PAF 的检测中引入长时段的信息，提高 PAF 的检测水平。

在基于 RTDNN 的音素识别系统中，RTDNN 的每个输出为对应音素的后验概率。考虑到音素与 PAF 间的对应关系，即任一个音素 a 可由 PAF 集中 n 个并行的属性表示，即：$a=(e_1,e_2,\cdots,e_n)$，其中，e_j 为二进制值的 PAF，n 为 PAF 的个数，若将 RTDNN 的输出设定为 PAF，即每个输出对应一个 PAF 的后验概率，从而完成 PAF 的检测。

2　音素边界检测

基于 PAF 后验概率的音素边界检测方法分两个阶段：后验概率向量间夹角的计算及音素边界的筛选。首先，依次计算出相邻帧向量间夹角，然后根据筛选规则，将夹角的极大值点所在的帧选为候选边界，通过比较候选边界与其相邻帧夹角间的差异而确定边界。

2.1　后验概率向量间的夹角

音素类别与后验概率特征间具有一一对应关系，同一类别所对应的后验概率向量间差异较小，即相关性较强；而不同类别对应的后验概率间趋向于正交，相关性较弱[10]。因此，通过度量相邻帧后验概率间相关性的大小，可以准确地判别其是否属于同一类，从而找出准确的音素边界位置。本文通过度量相邻帧后验概率向量间的夹角，以度量相邻帧间的差异。假设给定两帧 k 后验概率向量 $\boldsymbol{X}_m$ 和 $\boldsymbol{X}_n$，其中 $\boldsymbol{X}_m=(x_{m,1},x_{m,2},\cdots,x_{m,K})^{\mathrm{T}}$，$x_{m,k}$ 为第 m 帧的第 k 维后验概率，和某个特定的识别单元相对应，如某个 PAF。则定义这两个 k 维非零向量

间的夹角为

$$\langle \boldsymbol{X}_n, \boldsymbol{X}_m \rangle = \arccos \frac{(\boldsymbol{X}_n, \boldsymbol{X}_m)}{|\boldsymbol{X}_n||\boldsymbol{X}_m|}$$

$$\text{且 } 0 \leqslant \langle \boldsymbol{X}_n, \boldsymbol{X}_m \rangle \leqslant \pi \tag{1}$$

若得到的夹角值越小，则这两个后验概率向量间的相关性越强；反之则相关性越弱。

本文将某一帧前后相邻帧间的夹角作为该帧的夹角，为了减小插入错误的影响，在该帧前后距离为 D 的向量夹角上进行平滑。对于第 m 帧定义其夹角为

$$A(m) = \left(\sum_{k=1}^{D} (\langle \boldsymbol{X}_n, \boldsymbol{X}_m \rangle) * k \right) \Big/ D \tag{2}$$

式中，D 为平滑长度，该值的选取对实验结果的影响较大。对于塞音如/b/，/d/和/t/等，其发音时间较短，持续时间约为 30 ms，若平滑长度过长，则“平滑”作用会使得边界点变得平坦而难以检测；而元音的持续时间较长，如/aa/，/ae/等可以达到300 ms以上，较短的平滑长度则会造成较多的插入错误。

2.2 音素边界的筛选

第 2.1 节已得到每帧的前后帧向量间的夹角，该值大小与音素边界间有紧密联系。为了从夹角信息中提取出准确的边界点，本文采用以下 3 个规则进行边界点的检测，其中规则(2)，(3)为借鉴文献[5]中的方法，与其不同的是，本文在筛选之前首先要进行静音段的剔除。

(1)排除静音段

静音帧中 PAF 的后验概率值均较小，而在计算帧间夹角时，后验概率微小的改变却对应夹角值剧烈的变化，对判决形成极大干扰；另外，静音段中不可能含有音素边界且检测率较高。因此，本文先判断出静音段，并将静音帧间的夹角值预设为较小的值。具体做法如下，若某帧的 PAF 后验概率均小于 0.2，且持续 3 帧以上，则判定该部分帧为静音段，并将这些帧所对应的夹角设为 5°以内的随机数。如图 2 所示，点 A 和点 B 为通过计算向量间角度后得到的错误突变点，若在计算前进行静音判决，将能够有效去除这些插入错误。

(2)选取候选边界

通过实验发现，从音素边界位置所得出的角度中大于 30°的帧占 94%；另外，角度的极大值点处帧间的相关性最弱，为边界的可能性最大，因此，本文以 30°为阈值，将夹角大于该阈值的极大值点选作候选边界，如图 2 中 C，D，E，F，G 和 H 点等，其目的是尽可能降低漏检的概率，但选中的点中存在很多虚假的边界点。

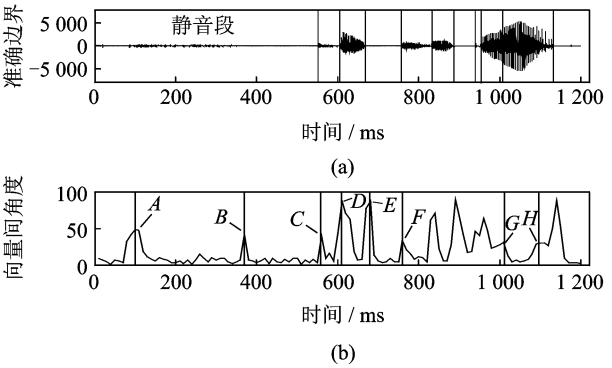

图 2 音素边界点的筛选

(3)去除虚假边界

在步骤(2)选取的候选边界中，只有峰值较尖锐的帧才有可能是边界，从实验中发现：在所有音素边界中，约 93.5%的角度值比其前、后帧大 2%以上，约 92%的角度值比其前、后 5 帧以内的最小值大 10%以上。因此，本文提出了以下两个约束条件进行边界筛选：(a)极值点的夹角值比前、后两帧的夹角均大 2%以上；(b)极值点的夹角值比前、后各 5 帧以内的最低值大 10%以上。这两个条件能够对该极值点附近的尖锐程度进行判定，如图 1 中 G 和 H 点不满足条件(1)，据此可判断其为虚假边界。

综上所述，该方法能够在降低漏检率的同时极大地减少插入错误的存在，使检测结果更加可靠。

2.3 检测结果评测

TIMIT 语料库中的边界点标注在样点上，而本文的检测结果是在帧一级，因此，在进行边界检测之前，需先将边界点转化到距该点最近的帧上，这种转换会引入一定的误差，若帧移间隔为 T ms，则最小误差为 0 ms，最大误差为 $T/2$ ms，平均误差为 $T/4$ ms。

通过比较检测结果和人工标记的音素边界之间的差异，来衡量检测方法的性能。假设本文选择的容错范围为 t ms，即若检测边界与标准边界间距在 t ms 内，则认为检测正确；若该误差范围内检测出 n 个边界，则认为其中 $n-1$ 个为插入错误；若距标准边界 t ms 内没有检测任何边界，则视为删除错误。图 3 为容错误差在 30 ms 时的检测结果，其中图(a)为语音信号的波形图，其中竖线为根据语料库音素标注文件得到的边界点；图(b)为本文方法所检测出的边界；图(c)为检测结果中的插入错误，如点 A，B 所示；图(d)为误丢弃的边界点位置，如点 C，D 所示。从图中发现：检测错误常集中于塞音/b/，/d/，/kcl/等的边界处。

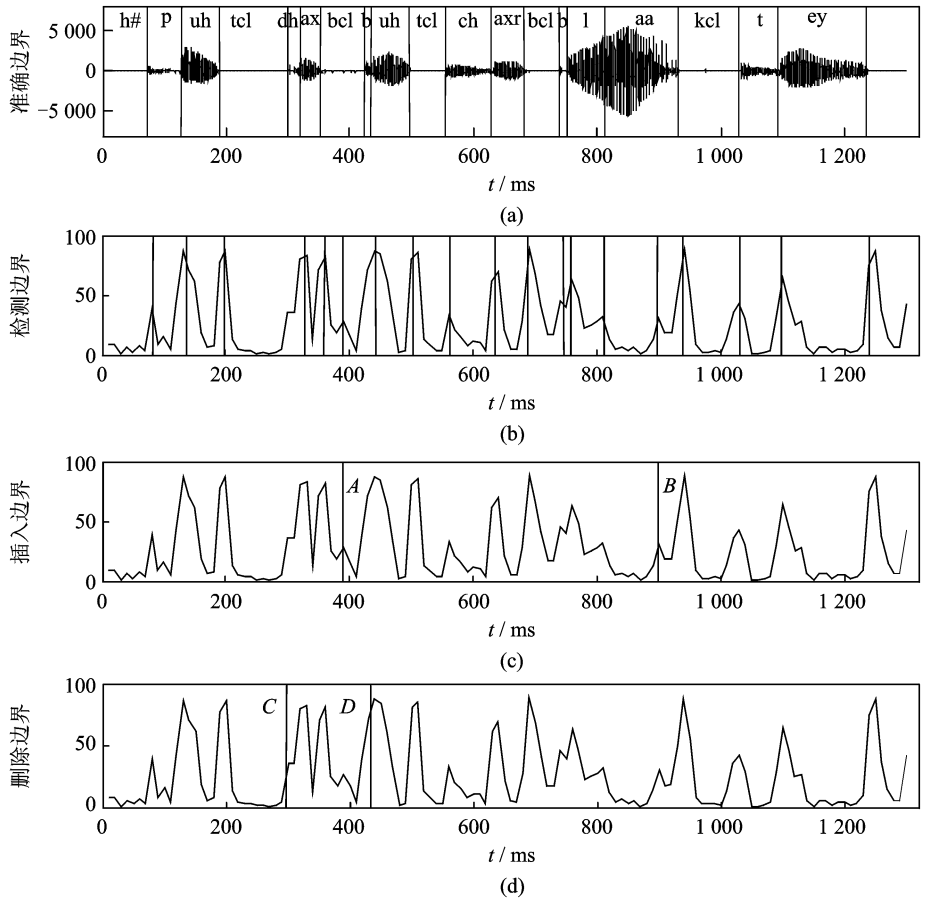

图 3　音素边界检测结果

3　实验配置及结果

3.1　实验配置

本文实验在 TIMIT 语料库上进行，选择其中 3 296 个语句作为训练集，选择 1 344 和 400 个语句作为测试集和开发集，这些集合间没有重合部分，可以认为与说话人无关。另外，排除 SA1 和 SA2 中的语句。实验中，所有的模型训练数据均来自训练集。按照 IPA 划分的标准，将 TIMIT 中 61 个音素转化为 PAF 后共 47 个，在性能评价阶段，根据 CMU/MIT 建议，音素集被映射到 39 个并进行测试[11]。

在进行 PAF 的检测时，采用的是 3 层的 TDNN，即输入层、隐含层和输出层。其中，预处理阶段所采用的帧长、帧移间隔分别为 25 ms 和 10 ms，声学特征选用 12 维 MFCC、1 维能量特征以及它们的一阶与二阶差分（对应 HTK 中的配置为 MDCC_E_D_A），共 39 维，为了考虑长时性特征的影响，本文 TDNN 的输入端为连续 9 帧数据；TDNN 的隐含层含有 300 个神经元；输出端个数与 PAF 的个数相同。为了衡量 PAF 的检测效果，将检测结果以 0.5 为阈值进行 0，1 离散化，并与从测试集语料库转化而来的结果进行对比，其识别结果如表 1 和表 2 所示。

表 1　SPE 属性的检测结果　%

音位属性	准确率	音位属性	准确率
Anterior	90.1	Nasal	97.3
Back	87.9	Round	93.9
Consonantal	90.0	Silence	97.7
Continuant	92.8	Strident	96.8
Coronal	89.5	Tense	90.2
High	88.2	Vocalic	87.4
Low	92.8	Voice	92.2

表 2　GP 属性的检测结果　%

音位属性	准确率	音位属性	准确率
A	85.5	H	92.5
I	89.7	N	97.2
U	86.4	a	96.1
E	86.1	i	94.3
S	90.2	u	95.5
h	94.6		

3.2 音素边界检测结果

在计算后验概率向量间夹角时，式(2)中的平滑长度 D 取值为 2。性能评测在测试集 1 344 个语句上进行，其中共含有 48 993 个边界点，容错误差分别为 20，30，40 ms，即分别对应的帧数为 2，3，4 帧。检测结果如表 3 所示。

表 3 基于 PAF 后验概率向量的边界检测结果

容错误差/ms	检测率/%	删除错误/%	插入错误/%
20	77.8	22.2	12.5
30	88.4	11.6	21.6
40	93.5	6.5	28.2

从表 1 中可以看出，随着容错误差的增大，虽然正确检测率不断提高，删除错误不断减少，但是插入错误也越来越严重。综合上述因素，当容错误差为 40 ms 时相对较好，此时不同误差范围内检测结果所占比重不同，最小误差约为 0 帧，即检测边界与准确边界位置一致；最大误差为 40 ms，即检测边界与准确边界相差 4 帧。

在相关边界检测试验中，通常选择容错范围为 20 ms，此处将本文和其他文献的检测结果进行比较，如表 4 所示。

表 4 本文和其他文献的检测结果

检测方法	容错误差/ms	检测率/%
文献[4]方法	20	77.5
本文方法	20	77.8

由于 PAF 后验概率是经过 MLP 得到的，在分类过程中，模型已经去除了声学特征中的冗余信息，并将与识别单元相关的区分性信息保留在其后验概率中。另外，本文在计算向量间夹角时，在当前帧的前、后分别选择 2 帧进行平滑，也在一定程度上降低插入错误的影响。因此，后验概率信息是更加精确的区分性信息，基于后验概率的检测结果更加可靠。

3.3 基于 CRF 的音素识别

条件随机场是一种新的概率无向图模型，它具有融合元素间的长距离依赖性、重叠性和非独立特征的能力，不但能够充分地利用上下文信息作为特征，还可以任意地添加其他外部特征，使模型能获取的信息、知识非常丰富[12]。为了验证边界信息对于音素识别的重要性，本文在基于 PAF 和 CRF 的系统中加入边界信息。由于本文使用的 CRF 仅二进制特征，首先将 SPE，GP 属性进行 0，1 离散化，判决阈值为 0.5；另外对于 2.3 节中的边界检测结果，将边界对应的帧标为 1；其他帧标为 0。将边界信息排列在每一帧 PAF 的最后一列，将组合后的特征作为基于 CRF 的音素识别系统的观测特征，识别系统如图 4 所示。

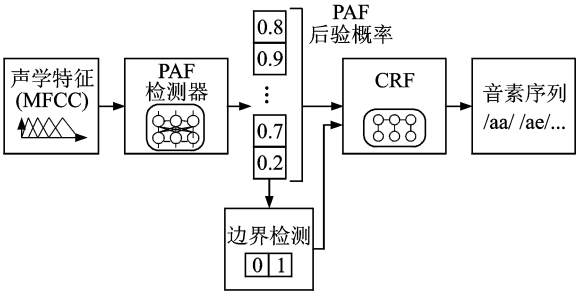

图 4 基于条件随机场模型的音素识别系统

本文在音素识别中设定了以下 4 组实验配置：基系统为在 HTK 平台上搭建的 HMM_GMM 识别系统，其高斯混元数为 32，观测特征为 39 维 MFCC 参数，与神经网络的输入特征一致；第 2 组为仅使用 SPE，GP 属性特征组合的识别结果；第 3 组在 SPE，GP 属性特征后加入了准确的边界信息，即这些边界从语料库标注文件中转化而来，该配置的目的是为了验证加入边界信息后识别性能的上界；第 4 组将属性特征与实际边界检测结果相组合，其中的边界信息不可避免存在错误。实验结果如表 5 所示。

表 5 基于 CRF 的音素识别结果 %

识别方法		正确率	准确率
HMM_GMM		69.73	59.62
CRF SPE_GP	未使用边界	66.2	61.8
	使用准确边界	69.4	64.7
	使用检测边界	68.7	64.1

从识别结果中可以发现，基于 CRF 的音素识别结果优于 HMM，体现了区分性模型相对于生成模型的优势。在基于 CRF 的系统中，由于第 3，4 组实验中加入了边界信息，其音素识别率均得到提升，说明边界信息对于提升音素识别率有着重要贡献。由于第 3 组实验中使用的边界信息是准确的，其结果可认为是边界信息对于提升音素识别率的上界；而第 4 组实验中用到的边界信息是本文算法的检测结果，其中存在一定的检测错误，因此其识别性能低于第 3 组，但音素识别率仍然提升了约

2.3%，从而也间接验证了本文边界检测算法的性能。

4 结束语

本文提出了一种基于 PAF 及边界信息的音素识别方法，该方法通过计算得出相邻帧后验概率向量间的夹角，并选定极大值点为候选的边界，然后通过筛选算法，去除极大值点中的错误边界，从而将正确的边界点保留下来，提升了音素边界的检测率。本文将音素的边界与 PAF 信息进行组合，作为基于 CRF 模型的音素识别系统的观测特征，结果表明：音素的边界信息有助于提升系统的识别性能。

参考文献：

[1] Dusan S, Rabiner L R. On integrating insights from human speech perception into automatic speech recognition[C]//Conference on the International Speech Communication Association (InterSpeech). Lisbon: Interspeech Press, 2005:1233-1236.

[2] Morris J, Fosler Lussier E. Combining phonetic attributes using conditional random fields[C]/Proc Annu Conf Int Speech Commun Assoc, INTERSPEECH. UK: Dummy Pubid, 2006:597-600.

[3] Scharenborg O, Wan V, Mirjam E. Unsupervised speech segmentation: an analysis of the hypothesized phone boundaries[J]. Journal of the Acoustical Society of America, 2010,127(2):1084-1095.

[4] Yu Qiao, Shimomura N, Minematsu N. Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, USA: [s. n.], 2008:3989-3992.

[5] Dusan S, Rabiner L. On the relation between maximum spectral transition position and phone boundaries[C]//International Speech Communication Association. Pittsburgh, USA:[s. n.], 2006:1317-1320.

[6] Omar M, Hasegawa-Johnson M, Levinson S. Gaussian mixture models of phonetic boundaries for speech recognition[J]. IEEE Workshop on Automatic Speech Recognition and Understanding, 2002:33-36.

[7] Chen I F, Wang Hsin-Min. Articulatory feature asynchrony analysis and compensation in detection-based ASR[C]//International Speech Communication Association. Brighton, United Kingdom: [s. n.], 2009:3059-3062.

[8] Zoltán Tüske, Christian Plahl. A study on speaker normalized MLP features in LVCSR[C]//Conference of the International Speech Communication Association. Florence, Italy: [s. n.], 2011:1089-1092.

[9] Strom N. The NICO artificial neural network toolkit [EB/OL]. (2011-02-10). http://nico.nikkostrom.com.

[10] Viet-Bac Le, Lori Lamel, Jean-Luc Gauvain. Multi-style MLP features for BN transcription[C]//IEEE International Conference on Acoustics Speech and Signal Processing. Dallas TX: [s. n.], 2010:4866-4869.

[11] Lee K F, Hon H W. Speaker-independent phone recognition using hidden Markov models[J]. IEEE Trans on Acoustics, Speech, and Signal Processing, 1989,37(11):1641-1648.

[12] McCallum A. Efficiently inducing features of conditional random fields[C]//UAI'03 Proceedings of the 9th Conference on Uncertainty in Articifical Intelligence. San Francisco, USA: Morgan Kaufmann Publishers Inc., 2003:403-410.

作者简介：许友亮(1985-)，男，硕士研究生，研究方向：连续语音识别，E-mail: xyl0709@yahoo.com；张连海(1971-)，男，副教授，研究方向：语音识别、语音信号处理；牛铜(1983-)，男，博士研究生，研究方向：语音信号处理。