文章编号：2095-4980(2013)06-0936-06

基于语谱能量的音素边界检测

李立永，张连海，冯志远

(信息工程大学 信息系统工程学院，河南 郑州 450002)

摘　要：根据音素发音时语谱结构的变化提出了一种基于语谱能量的音素边界检测方法。该方法首先根据语谱结构变化特点将信号频域划分为高频、中频、低频 3 个区域，并以语音帧间语谱能量向量的欧氏距离为判别依据分别对 3 个区域进行音素边界检测，然后对 3 个区域检测的边界分别进行二次筛选，最后将 3 个区域的边界信息融合，得到音素边界检测结果，相对于基于音素属性的边界检测方法，计算复杂性大大降低，边界检测率提高了 3.95%。

关键词：音素边界检测；语谱能量；语音事件

中图分类号：TN912.34　　**文献标识码：**A　　**doi：**10.11805/TKYDA201306.0936

A phone classification method based on spectrum

LI Li-yong，ZHANG Lian-hai，FENG Zhi-yuan

(Institute of Information System Engineering，Information Engineering University，Zhengzhou Henan 450002，China)

Abstract：This article proposes a phone boundary detection method based on the spectrum. The spectrum is divided into three frequency regions according to the structure of the spectrum. The potential boundaries of every region are detected through the Euclidean distance between the neighbors of two vector frames of spectrums. The false phone boundaries are removed by a second examination. The phone boundaries are obtained via the fusion of the boundaries of three regions. Comparing with the method using phonological attributes，the computational complexity is significantly reduced，and a higher score of 3.95% is achieved.

Key words：phone boundary detection；spectrum；speech event

音素边界检测技术在语音信号处理中的许多领域都有广泛应用，其可以将语音从时间维度分割为不同的单元，如词、音节和音素等。边界信息的人工标注，工作量大且具有人的主观性。因此如何快速、准确地检测出音素边界或类音素边界，对语音信号的处理具有重要作用。目前自动边界检测的方法主要有基于模型(Model-based)和基于尺度(Metric-based)2 种。基于模型主要是采用动态时间规整技术(Dynamic Time Warping，DTW)[1]和隐马尔科夫模型(Hidden Markov Model，HMM)，以一定的训练标准对数据进行训练而得到音素或音节的边界。基于尺度的方法主要是利用数据特征参数的变化，并采用合适的判定准则对数据的变化率进行描述。常用的参数有过零率、短时能量和倒谱系数等。相比于基于模型的边界检测方法，基于尺度的边界检测方法需要的数据量小，且鲁棒性好。语音信号的变化可以通过语音事件来描述。语音事件包括语音产生时各种发音特征的变化，如声带、声道和舌唇等发音器官的位置的变化，元音、辅音、摩擦音和爆破音等发音方式的变化，及其他信息的变化。2011 年 You-Yu Wang 等[2]根据语音信号的发声特征，以语音信号的包络和谱熵等为参数，提出了基于样点的音素边界检测方法。张宝奇等[3]于 2010 年提出了基于听觉事件的汉语声韵切分方法，通过检测听觉事件，聚类确认后实现对边界的检测。2012 年许友亮[4]在检测出音位属性的基础上，提出了一种基于音位属性的边界检测算法，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，提高了边界检测率和音素识别率。由于不同音素的发音特征所对应的语音事件在频谱上具有不同的结构，如音素/i/和/u/在低频区域具有相似的结构，而在高频区域则具有不同的结构。因此可以根据语音信号的这种频谱区域结构特性，将语音信号分为若干频带，对比各个频带信号帧之间的差异性，从而实现对音素边界的检测。本文根据该思想提出了一种基于语谱能量的音素边界检测方法，该方法首先计算语音信号的语谱能量，然后根据语音信号的发音特征划分频域，分析比较相邻帧语谱能量向量之

收稿日期：2012-10-10；**修回日期：**2012-11-14

基金项目：国家自然科学基金资助项目(61175017)

间的相关性，并通过判定准则和二次筛选检测，确定音素的边界候选，最后通过边界融合得到最终的检测结果。

1　语谱能量

不同的语言有不同的音素集合，对于英语而言，可以根据音素发音时声道受阻程度分为元音和辅音等，元音又可以分为单元音、半元音和双元音等，辅音又可以分为摩擦音、塞擦音、爆破音和鼻音等，其描述方式主要根据声带和声道等发音特征的变化，如声带是否处于振动，舌的位置、高度和状态，软腭的状态等。在产生语音的过程中，发音特征的变化最终导致语音波形及其声学的时域和频域特征的变化[5]。这些变化在语谱中都有相应的特点。元音发声时，声道的形状受到舌、颌、唇、软腭的影响，但肺部的气流通过声道时并不会受到声道的阻碍，具有连续性，在宽带语谱图上呈现垂直的条纹，这些条纹的间隔即为基音周期。鼻腔有较大的容积，因此鼻音的频谱成分主要为低频谐振。清摩擦音有“类噪声”频谱，而浊摩擦音经常既有噪声又有谐波结构。爆破音的语谱图中可以看到一段静默区域后的突然爆破，然后为送气噪声[6]。因此，可以根据不同音素的发音特征集合所对应的语谱结构，实现对音素的边界进行检测。

本文音素边界检测流程见图 1。首先计算语音信号的语谱能量，为了使语音信号的处理具有较好的时域分辨率，本文选取 6 ms 的汉宁窗进行滤波处理，帧移为 5 ms，然后将每帧语音进行 512 点的离散傅里叶变换。为了使得到的语谱能量数据更加准确可靠，本文对语谱能量采用了去噪和平滑的处理方式。首先，将语音信号的前 40 帧数据作为背景噪声的参考，取其中能量最大的一帧数据作为背景噪声的标准，并在后续的每帧数据处理时减去背景噪声的能量。再次，为平滑语谱能量数据，本文将当前帧、前后各 5 帧共 11 帧能量均值作为当前帧能量瞬时值。

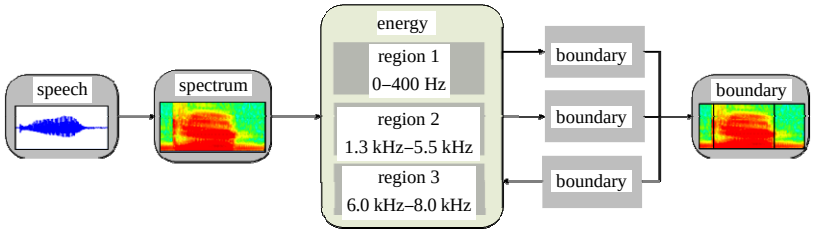

Fig.1 Flow chart of boundary detection
图 1　音素边界检测流程图

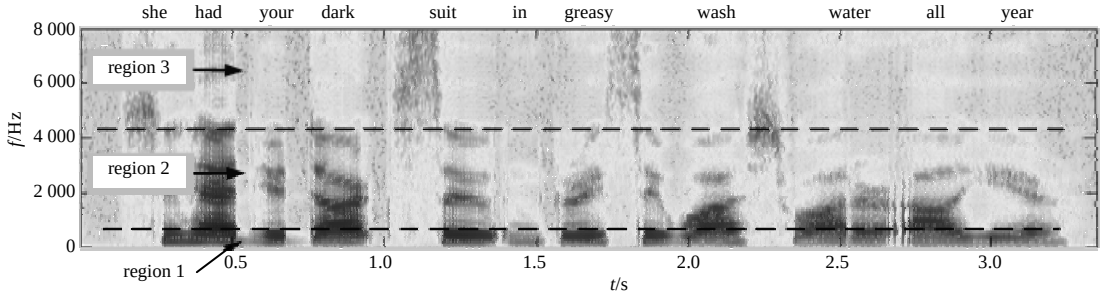

Fig.2 Spectrogram for “She had your dark suit in greasy wash water all year”
图 2　“She had your dark suit in greasy wash water all year”的语谱图

从图 2 中可以看出语谱能量在音素的开始或结束时变化十分明显，且具有清晰的区域结构特点。因此，本文根据语谱的区域结构特点，在频域 0~8 kHz 范围内将其分为 3 个区域，见图 2。根据音素的发音特征可以知道，在区域 1 中，如果能量很大(颜色深)，则表示音素发声时，声带是振动的，反之，则表示音素发声时，声带不会振动，因此可利用该区域的语谱能量变化来检测音素的声带振动信息。区域 2 是语谱频域上包含信息量最多的区域，英语的强辅音、元音等音素的发音特征变化都集中体现在该区域，因此可利用该区域的语谱结构变化检测元音和强辅音等音素的边界信息。区域 3 则是在爆破音和摩擦音等音素发声时，才会产生明显的变化。本文利用 3 个区域语谱能量向量的变化特点，分别检测 3 个区域的音素边界，最后将 3 个区域的边界信息融合得到边界的检测结果，保证了音素边界检测的准确性和完整性。

本文根据语音信号的低频、中频、高频结构特点将语音信号的频谱结构划分为 3 个区域：区域 1、区域 2 和区域 3，区域 1 为 0~400 Hz，区域 2 为 1.3 kHz~5.5 kHz，区域 3 为 6.0 kHz~8.0 kHz，并将区域 1 划分为 0~100 Hz，100 Hz~200 Hz,200 Hz~300 Hz 和 300 Hz~400 Hz 4 个频带[7-8]，分别将每个频带的能量均值作为后续处理的输入。区域 2 的频带划分时，一般具有声带振动特性的信号的低频能量较大，为了降低声带振动信息的干扰，准确检测中频区域语谱结构的变化，将区域 2 的范围划定为 1.3 kHz~5.5 kHz，并将其划分为 1.3 kHz~1.5 kHz,1.5 kHz~

1.7 kHz,1.7 kHz~2 kHz,2 kHz~2.5 kHz,2.5 kHz~3.0 kHz,3.0 kHz~3.7 kHz,3.7 kHz~4.5 kHz 和 4.5 kHz~5.5 kHz 8 个频带，分别将每个频带的能量均值作为后续处理的输入。区域 3 的范围划定为 6 kHz~8 kHz，这样划分的原因是语音信号的高频部分变化集中体现在该范围内，为了更准确地捕捉高频区域的变化，将其划分为 6.0 kHz~6.5 kHz, 6.5 kHz~7.0 kHz,7.0 kHz~7.5 kHz 和 7.5 kHz~8 kHz 4 个频带,分别将每个频带的能量的均值作为后续处理的输入。本文在低频时频率划分较为细致，而在高频时划分较为粗糙，原因是人的听觉感知过程对信号的低频部分较为敏感，高频部分则不是如此，本文为模拟人的听觉感知过程，进行如此处理。

2 音素边界检测

基于语谱能量的音素边界检测主要分为 3 个阶段：区域的音素边界检测、边界筛选以及边界融合。首先根据每个区域的语谱能量向量的欧氏距离检测出每个区域的音素边界，然后对每个区域边界检测的结果进行二次筛选，最后将 3 个区域检测的边界融合。

2.1 区域的音素边界检测

在语音的语谱结构中，同一音素的语音帧具有相似的语谱能量分布，而当音素发声开始与结束时语谱能量的结构都发生很大变化，因此可以通过计算相邻语音帧之间语谱能量的差异来检测边界。本文采用语谱能量向量的欧氏距离作为判定语音帧之间差异性的标准。如果某相邻语音帧之间的欧氏距离超过给定阈值，则判定该时间点为边界候选，否则，判定该时间点不为边界候选。

假设给定区域 1 的连续 2 帧语音的语谱能量向量为 $\boldsymbol{G}_k$ 和 $\boldsymbol{G}_{k+1}$，$\boldsymbol{G}_k$ 表示为 $\boldsymbol{G}_k=(x_{k,1},x_{k,2},x_{k,3},x_{k,4})$，并假定这 2 个向量的欧氏距离为 $ED(\boldsymbol{G}_k,\boldsymbol{G}_{k+1})$，如式(1)所示：

$$ED(\boldsymbol{G}_k,\boldsymbol{G}_{k+1})=\sqrt{\sum_{n=1}^{4}(x_{k,n}-x_{k+1,n})^2} \tag{1}$$

式中 $x_{k,n}$ 对应区域 1 的第 k 帧语音的第 n 个频带的语谱能量，并用 P_{G} 作为区域 1 的语谱能量向量欧氏距离的判定阈值。实验发现 $P_{\mathrm{G}}=0.7$ 时检测效果最好。

根据式(2)对给定的语音数据进行事件 $Th_{\mathrm{G}}(k)$ 判决。

$$Th_{\mathrm{G}}(k)=\begin{cases}1, & \text{if } ED(\boldsymbol{G}_k,\boldsymbol{G}_{k+1})>P_{\mathrm{G}}\\ 0, & \text{else}\end{cases} \tag{2}$$

式中,当 $Th_{\mathrm{G}}(k)=1$时,表示区域 1 在第 k 帧时有语谱能量突变事件发生,即判定该时间点为候选音素边界 $BD_{\mathrm{G}}(m)$，m 表示第 m 个边界候选。在区域 1 的事件检测过程中，若超过阈值的语音帧之间的时间间隔小于 7 ms，则判定该连续数帧信号的时间范围只存在一个候选边界，并取其中欧氏距离最大的语音帧为候选边界的位置。

2.2 边界筛选

在进行区域边界检测时，为了得到较高的检测率，本文将欧氏距离阈值设置得偏低，提高了区域边界的检测率，降低了删除错误率，但也使插入错误大大增加，影响了后续处理的效果，因此本文为了剔除候选中的错误边界，对边界检测的结果进行二次筛选。

假设区域 1 中的第 k 帧语音事件 $Th_{\mathrm{G}}(k)=1$，记为边界候选 $BD_{\mathrm{G}}(m)$，且该事件为一个插入错误，即表示第 m 个候选边界的两侧语音区域处于同一音素的发音范围，由于同一音素的发音特征具有相似性，同时体现在语谱结构上的相似性，因此可以通过判别候选边界所划分的语音段之间的差异性来剔除插入错误。本文采用语音段间语谱能量均值向量的欧氏距离作为判别的主要方式，见图 3。

假设候选边界 $BD_{\mathrm{G}}(m+1)$ 和 $BD_{\mathrm{G}}(m)$ 之间的语音段的语谱能量均值向量为 $\boldsymbol{G}(m+1,m)$，候选边界 $BD_{\mathrm{G}}(m)$ 和 $BD_{\mathrm{G}}(m-1)$ 之间的语音段的语谱能量均值向量为 $\boldsymbol{G}(m,m-1)$，这 2 个均值向量的欧氏距离为 $ED(\boldsymbol{G}(m,m-1),\boldsymbol{G}(m+1,m))$，如式(3)所示。

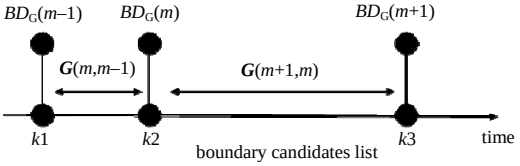

Fig.3 Graph for double-check on boundaries of region
图 3 区域音素边界筛选示意图

$$ED(\boldsymbol{G}(m,m-1),\boldsymbol{G}(m+1,m))=\sqrt{\sum_{i=1}^{4}(e_{m,i}-e_{m+1,i})} \tag{3}$$

式中 $e_{m,i}$ 为候选边界 $BD_{\mathrm{G}}(m)$ 和 $BD_{\mathrm{G}}(m-1)$ 之间的语音段的能量均值向量 $\boldsymbol{G}(m,m-1)$ 的第 i 个频带能量均值，如式(4)所示。

$$e_{m,i}=\sum_{n=k1}^{k2}x_{n,i}/(k2-k1) \tag{4}$$

式中：$k2$ 为候选边界 $BD_{\mathrm{G}}(m)$ 所处于的语音帧位置；$k1$ 为候选边界 $BD_{\mathrm{G}}(m-1)$ 所处于的语音帧位置；$x_{n,i}$ 表示的是第 n 帧语音的第 i 个频带。语音段欧氏距离阈值设为 Q_{G}，通过实验发现 $Q_{\mathrm{G}}=5.5$ 时边界筛选效果最好。

2.3 边界融合

将 3 个区域检测得到的边界信息进行融合，融合准则如下：

1) 相邻边界的时间间隔不得小于 10 ms。由于英语音素的最短持续时间不小于 10 ms[4]，因此如果相邻 2 个边界时间间隔小于 10 ms，则取 2 个边界时间均值作为检测边界时间。

2) 如果区域边界 1 与区域边界 2 的时间间隔小于 20 ms，并且区域 1 边界的前面为静音区域，则以区域 1 的边界为融合边界；如果区域 1 的边界后面为静音区域，则以区域 2 的边界为融合边界。由于强辅音发声开始时，区域 1 和区域 2 的语谱能量都会发生变化，在 TIMIT 语料库的人工标注中是以声带开始振动作为发音的开始，因此以区域 1 的边界为准确边界；强辅音发声结束时，区域 2 的语谱能量立即变化，而区域 1 的语谱能量仍将持续一段时间，根据 TIMIT 语料库的人工标注特点，本文以区域 2 的边界作为发音的结束位置。

3) 对于时间间隔大于 20 ms 的区域边界很难用发音知识判断边界的真实性，因此本文采用了与区域的音素边界筛选相同的方法对融合后的边界进行筛选，通过判定相邻的语音段之间的差异性来剔除错误候选，处理完成后得到最终的音素边界检测结果。

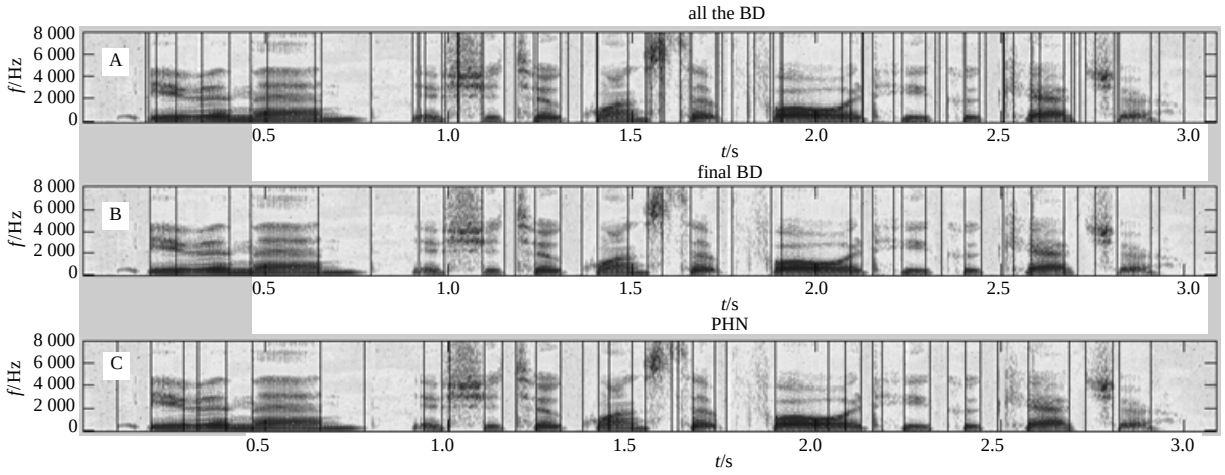

Fig.4 Results for boundaries detection

图 4 边界检测结果

边界检测结果见图 4，其中图 A 为 3 个区域边界融合后得到的边界，图 B 为经过融合准则判定后的结果，图 C 为 TIMIT 语料库 PHN 文件人工标注的边界。

2.4 检测结果的评测

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相距在 $t$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。

3　实验配置及结果

3.1 实验配置

实验采用 TIMIT 语料库，该语料库最初是由 MIT 等多家单位共同开发，共有 63 000 句语句，分为 TRAIN 和 TEST 两个集合。本文采用 TIMIT 中的 TRIAN 中的 DR1,DR2 和 DR3 作为测试语料。其中 DR1 中的 FCJF0 作为实验的观测数据集，用于观察调整信号处理中的参数设置。并分别从 DR1,DR2 和 DR3 中选取了 288,616 和 575 句语句作为测试语句，为了得到不同语句的边界检测性能，不采用 SA1 和 SA2 语句。

3.2 音素边界检测结果

边界检测性能评测分别在训练集 DR1,DR2 和 DR3 共 1 479 个语句上进行,相应人工标注的 PHN 文件上共有 55 738 个边界点。音素边界检测结果见表 1。

表 1 基于语谱能量的边界检测结果

Table1 Results for boundaries detection on spectrum

tolerance error/ms	detection rate/%	deletion rate/%	insertion rate/%
20	81.75	18.25	31.52
30	89.15	10.85	26.24
40	92.85	7.15	18.89

表 2 本文和其他文献的检测结果

Table2 Results for this paper and other paper

detection method	tolerance error/ms	detection rate/%
paper[4]	20	77.80
this paper	20	81.75

另外，在相关的边界检测实验中，通常选择容错范围为 20 ms，此处将本文和文献[4]的检测结果进行比较，见表 2。文献[4]首先利用时间延迟神经网络(Recurrency and Time-delayed Neural Network，RTDNN)模型检测得到音位属性及其后验概率，然后利用音位属性进行音素边界检测，作为一种基于统计模型的检测方法，其计算量庞大，耗时长，且当输入信号变化时，模型调整代价巨大。

3.3 音素边界检测结果分析

本文提出的检测方法仍存在着音素边界漏检的情况，其主要由元音的边界漏检造成。当某些元音、双元音或半元音等音素发声时，对应的语谱结构变化不明显，在区域 1、区域 2、区域 3 都无法有效地检测。图 5 中，左侧 A 图为 TIMIT 语料库人工标注的音素边界，圈中区域为“even”单词的发音，中间区域为“iy”的发音，由语谱图可以看出，整个单词的发音过程中语谱结构变化很小，很难通过语谱结构的变化对其进行准确检测，左侧 B 图则是本文检测的结果。另外，TIMIT 语料库的人工标注的音素边界存在这样的情况，某个音素的结束与下个相邻音素的开始之间标注了两个边界，由于标注边界间的时间间隔较短，且在语谱结构上变化不明显，在实际的检测过程时只是以其中的某个边界或者这 2 个边界之间的某个时间点为音素边界，造成删除错误，如图 5 所示，右侧 A 图为 TIMIT 语料库人工标注的音素边界，右侧 B 图则是本文方法的检测结果。

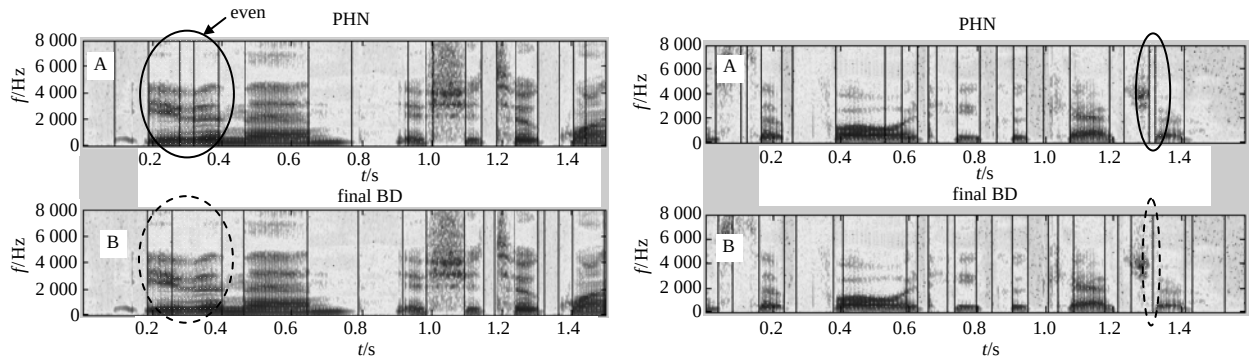

Fig.5 Graph for deletion error

图 5 删除错误示意图

4 结论

本文根据语音发音特征提出了一种基于语谱能量的边界检测方法，该方法将语音信号的频域分为 3 个区域，并通过分别计算每个区域相邻帧语谱能量向量间的欧氏距离,选定超过阈值的且在某个时间范围内极大值点为候选的边界，然后通过筛选算法，去除极大值点中的虚假边界，从而将正确的边界点保留下来。再通过边界融合方法将区域边界融合，得到音素的边界信息。实验证明，该方法具有较好的性能。

参考文献：

[1] Malfrere F,Dutiot T. High-quality speech synthesis for phonetic speech segmentation[C]// Proceedings of the European Conference on Speech Communication and Technology. Rhodes,Greece:[s.n.], 1997:2631–2663.

[2] Wang Yih-Ru. A Two-Stage Sample-based Phone Boundary Detector using Segmental Similarity Features[C]// 12th Annual Conference of the International Speech Communication Association. Florence,Italy:[s.n.], 2011:413–416.

[3] 张宝奇,张连海,屈丹. 基于听觉事件检测的汉语语音声韵切分[J]. 声学学报, 2010,35(6):701–707. (ZHANG Baoqi, ZHANG Lianhai,QU Dan. The detection of Chinese phone boundaries based on the auditory events[J]. ACTA ACUSTICA, 2010,35(6):701–707.)
[4] 许友亮,张连海,牛铜. 基于音位属性和边界信息的音素识别[J]. 数据采集与处理, 2012,28(2):79–87. (XU Youliang, ZHANG Lianhai,NIU Tong. Phoneme recognition based on the phonological properties and boundary information[J]. Data Acquisition and Processing, 2012,28(2):79–87.)
[5] Stevens K N. Toward a model of lexical access based on acoustic landmarks and distinctive features[J]. Journal of the Acoustical Society of America, 2002,111(4):1372–1891.
[6] Thomas F Quatieri. Discrete-Time Speech Signal Processing: Principles and Practice[M]. S.l:Prentice Hall PTR, 2001.
[7] Liu Anne Sharlene. Landmark Detection for Distinctive Feature-Based Speech Recognition[J]. Journal of Acoustical Society of America, 1996,100(5):3417–3430.
[8] Park Chiyoun. Consonant Landmark Detection for Speech Recognition[EB/OL]. [2012–05–02]. http://dspace.mit.edu/bitstream/handle/1721.1/44905/297548228.pdf?sequence=1. 2008:75–77.

作者简介:

李立永(1987–),男,河北省邢台市人,在读硕士研究生,主要研究方向为连续语音识别. email:forlly@126.com.

张连海(1971–),男,山东省单县人,副教授,主要研究方向为语音信号处理.

冯志远(1988–),男,河南省周口市人,在读硕士研究生,主要研究方向为语音信号处理.

(上接第 931 页)

[6] 樊玉伟. 磁绝缘线振荡器及其相关技术研究[D]. 长沙:国防科技大学, 2007. (FAN Yuwei. Investigation of Magnetically Insulated Transmission Line Oscillator and Correlative Technologies[D]. Changsha: National University of Defense Technology, 2007.)
[7] 周传明,刘国治,刘永贵. 高功率微波源[M]. 北京:原子能出版社, 2007. (ZHOU Chuanming,LIU Guozhi,LIU Yonggui. High power microwave sources[M]. Beijing:Atomic Energy Press, 2007.)
[8] Ginzburg N S,Novozhilova N Y,Zotova I V,et al. Generation of powerful subnanosecond microwave pulses by intense electron bunches moving in a periodic backward wave structure in the superradiative regime[J]. Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top., 1999,60(3):3297–3304.)

作者简介:

孙会芳(1974–),女,山东省德州市人,副研究员,现从事高功率微波技术研究.email: sun_huifang@iapcm.ac.cn.

李瀚宇(1980–),男,四川省资阳市人,助理研究员,从事高功率微波相关工作,主要方向为计算电磁学.

周海京(1970–),男,江苏省建湖市人,副研究员,研究方向为高功率微波技术、超宽带天线、复杂电磁环境.

姜幼明(1942–),男,上海市人,研究员,从事自由电子激光以及高功率微波物理等领域的理论研究.

董　烨(1981–),男,西安市人,助理研究员,主要研究方向为高功率微波技术、全电磁粒子模拟方法、太赫兹微电真空器件理论与数值模拟.

董志伟(1962–),男,河北省滦县人,研究员,研究方向为高功率微波技术、脉冲功率技术、太赫兹技术.