广播新闻语料识别中的自动分段和分类算法

吕 萍 颜永红
(中国科学院声学研究所中科信利实验室 北京 100080)

摘 要 该介绍了中文广播新闻语料识别任务中的自动分段和自动分类算法。提出了 3 阶段自动分段系统。该方法通过粗分段、精细分段和平滑 3 个阶段，将音频流分割为易于识别的音频段。在精细分段阶段，文中提出两种算法：动态噪声跟踪分段算法和基于单音素解码的分段算法。仿效说话人鉴别中的方法，文中提出了基于混合高斯模型的分类算法。该算法较好地解决了音频段的多类判决问题。在“新闻联播”测试数据中的实验结果表明，该文提出的自动分段和分类算法性能与手工分段分类性能几乎相当。

关键词 语音识别，自动分段，自动分类
中图分类号：TP391.42 **文献标识码**：A **文章编号**：1009-5896(2006)12-2292-04

Audio Segmentation and Classification in a Broadcast News Task

Lü Ping Yan Yong-hong
(*Zhongke Xinli Speech Lab, Institute of Acoustics, Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract This paper describes the work on the development of an audio segmentation and classification system applied to a broadcast news task for Chinese language. Three-phase automatic audio segmentation algorithm is provided. Audio stream is cut to audio segments (or sentences) by simply segmentation, fine segmentation and smoothing. Two different fine segmentation algorithms are given. They are dynamic noise tracking segmentation algorithm and segmentation based on mono-phone decoder algorithm respectively. Classifier based on mixture Gaussian model is used to classify audio segment into four groups: noise, music, male and female. The experiments on “Xin Wen Lian Bo” broadcast news show the performance of automatic segmentation and classification is almost equivalent to that of manual segmentation and classification.

Key words Speech recognition, Automatic segmentation, Automatic classification

1 引言

从 1996 年以来，美国国防部(DARPA)在世界范围内举行了广播新闻语料(Broadcast News，BN)[1]识别系统的评测(HUB-4 测试)[2]。这一举措将语音识别的研究领域引向现实环境，并提高了语音识别的研究水平。进入新千年后，DARPA 进一步提高评测难度，测试内容不仅仅涉及声音到文字的转换，还要求对“主题”或“新闻”进行检测和跟踪[3]。在 863 计划的推动下，国内语音识别的研究水平正逐步向国际先进水平靠近。本文在这方面做了一些初步的探索。

识别或翻译广播新闻语料，对大词汇连续语音识别系统而言是一项非常有挑战性的研究工作[4]。与以往的识别任务相比，广播新闻语料的识别要面临如下问题 [5-6]：(1)句子边界未知。一个电视频道的内容是连续不断的，而不是以句子为单位的听写输入。(2)多样的和快速变化的声学环境。一个典型的广播新闻语料可能包含来自不同通信信道的数据，如含有背景音乐、广告或背景噪声、现场采访等等。还可能包含各种口音的语音信号。(3)自然语音。与朗读式语音相比，自然语音不受约束变化较多，会含有不同的情感风格(如激动等)，并且说话人发生变化的时间是未知的。(4)自然语言。其难度在于对话主题的转换是不可预见的，并且对话中的自然反应(如“嗯”，“这个”等习惯性的口头语)会影响语言模型的连贯性。

与传统的语音识别相比，广播新闻识别系统面对的第 1 个挑战是，如何将整段的音频流分割成一个一个的音频段(即句子)。第 2 个挑战是，如何检测信道、背景以及说话人的转变，并减小或适应这种变化带来的影响。本文从这两点出发，研究了音频信号的自动分段、分类问题。本文主要以新闻联播语料为研究对象，选用 2004 年 5 月 22 日至 6 月 10 日中的 5 天新闻联播数据作为测试数据。本文安排如下：第 2 节介绍了本文提出的 3 阶段自动分段算法，基于混合高斯模型的音频分类算法在第 3 节给出，第 4 节介绍自动分段、分类算法在广播新闻识别系统上的实验结果，最后给出本文结论。

2 自动分段算法

广播新闻识别系统的首要任务是将连续输入的音频流分割为适于识别的音频段或句子。以距离度量为基础(metric-based)[7, 8] 的分段算法是使用较广泛的分类方法。其基本思想是，由于不同条件下的音频信号在统计上有很大差异，那么检测到差异就可以得到分段点。该类分段算法的基

2005-04-06 收到，2005-09-20 改回
中国科学院百人计划(G13BR01)和国家 973 计划(2004CB318106)资助课题

本做法是，首先对输入语音提取特征；然后按一定窗长将语音特征分成片断，并计算相邻窗间的距离；接着以一定步长滑动窗口，并计算新的相邻窗间距离；以此类推，直到所有特征计算完毕。相邻窗间距离越大，表明相邻窗的声学差异越大。因此上面所得距离的局部最大点即为分段点。Metric-Based分段算法性能受特征选取、距离测度选取、窗长、滑动步长以及分割准则等因素的影响[7]。

与 metric-based 分段算法的基本思想不同，本文考虑到人们在说话过程中，总有停顿的时候(也称为静音)，而且这种停顿基本上反映了语义信息或者是说话人的变化。在广播新闻语料中，“停顿”或“静音”同样存在。与正常说话相比，静音时的波形幅度很低，它基本反映了背景噪声的情况。在广播新闻语料中，虽然有信道变化、背景噪声的变化，但突变的情况几乎没有，并且在同一主题的新闻中背景噪声基本是平稳的。因此，利于时域能量可以对广播新闻语料进行粗分段。

2.1　基于时域能量的自动分段

对于输入的音频数据流，首先进行分帧。每帧长度为25ms，相邻帧间重叠 10ms。依下式计算每帧的时域能量：

$$e = \text{const} \times \sqrt{\sum_{t=0}^{T} x^2(t)} \tag{1}$$

其中 $x(t)$为第 t 个采样点，T 为每帧包含的总采样点数目，const 为一常量。

若连续 N 帧数据的时域能量小于某一域值时，则认为出现了一段静音。相邻两段停顿间的数据即为一个音频段(audio segment)或一个句子。而“静音”部分能量很低，仅体现了背景噪声不包含有效的语音信息，可以直接丢弃不用。

对 5 天的新闻联播数据，进行了能量自动分段。自动分段后，共得到了 2852 个音频段(或句子)。表 1 列出了音频段的时长统计信息。

表 1　音频段的时长统计信息

Tab.1 The statistic of length of audio segments

	[0,1s]	(1s,2s)	(2s,6s)	(6s,10s)	(10s,20s]	>20s
音频段数目	410	811	1505	74	39	13
占总长度的比例(%)	4.62	17.66	59.61	6.92	6.38	4.81

表 1 显示在 2852 个音频段中，有 1221 个音频段的持续时间较短(小于 2s)，而长度大于 10s 的音频段有 52 个。分段的最终目的是，将连续的音频流分割为易于识别的句子。就识别而言，待识别的句子在语义上最好是完整而独立的。过长的句子存在语义上的转折点，而过短的句子很可能语义不完整。句子的长度最好在 2～10s 之间。从表 1 可以看出，用时域能量自动分段算法得到的句子中仅有 66.53%满足这个要求。因此，还需要进一步的处理。对于过短的句子，应与邻近段进行合并。对于较长的段，则需要用其它方法进一步分割。下面讨论两种精细分段算法：动态噪声跟踪分段算法和基于音素识别的分段算法。

2.2　动态噪声跟踪自动分段算法

较长的音频段中并非没有停顿，而是由于信道或环境发生变化使得背景噪声发生了变化。若背景噪声大于能量分段时所设定的域值，自然就无法将这时候的“停顿”检测出来了。增加能量域值，无疑可以解决这一问题。但是在整个广播新闻语料中，信道和环境变化都是未知的。事先很难确定满足各类情况的域值。本节效仿 VAD 中的处理方法，通过动态估计噪声能量来动态更新域值。

尽管背景噪声是时刻变化的，但是只要变化不是太剧烈，那么音频信号幅度最小处基本上反映的就是噪声信号。因此，可以通过跟踪信号能量并检测其最小值来动态估计背景噪声[9]。

假设 $\bar{x}(n)$ 是含噪语音信号，其中 $x(n)$ 是干净语音，$w(n)$ 是噪声，并且 $x(n)$ 和 $w(n)$ 互不相关且均值等于零；含噪信号 $\bar{x}(n)$ 和噪声信号 $w(n)$ 都是短时平稳的。这里，不仅仅考虑当前帧的能量值，将历史信息也包含进来。含噪信号能量的更新公式为

$$e_{\bar{x}}(t) = \alpha_{\bar{x}} \cdot e_{\bar{x}}(t-1) + (1-\alpha_{\bar{x}}) \cdot e_{\bar{x},\text{new}}(t) \tag{2}$$

其中 $e_{\bar{x},\text{new}}(t)$ 是用式(1)计算出来的含噪信号第 t 帧的实际能量。参数 $\alpha_{\bar{x}}$ 为平滑系数，取值范围是 $0.45 \le \alpha_{\bar{x}} \le 0.95$ (本文取值 0.8)。式(2)中实际包含了第 t 帧以及第 t 帧之前的所有帧的能量信息，只是比例有所不同。

与式(2)类似，噪声信号能量 $e_w(t)$ 的更新公式为：

$$e_w(t) = \alpha_w \cdot e_w(t-1) + (1-\alpha_w) \cdot e_{w,\text{new}}(t) \tag{3}$$

其中 $e_{w,\text{new}}(t)$ 的取值标准为：

$$\left.\begin{array}{l} \text{若} e_{\bar{x}}(t-1) < e_{\bar{x}}(t), \text{并且} e_{\bar{x}}(t-1) < e_{\bar{x}}(t-2)\text{L}, \\ \text{并且} e_{\bar{x}}(t-1) < 2e_w(t-1), \text{则} e_{w,\text{new}}(t) = e_{\bar{x}}(t-1); \\ \text{否则} e_{w,\text{new}}(t) = e_w(t-1) \end{array}\right\} \tag{4}$$

即若 t–1 帧的含噪信号能量为局部最小值且小于 t–1 帧时的噪声能量，则 $e_{w,\text{new}}(t)$ 等于 t–1 帧的含噪信号能量，否则等于 t–1 帧时的噪声能量。换而言之，当检测到含噪信号能量的局部最小值时，更新噪声能量。该算法能够很快地跟踪噪声变化，延时仅有一帧。

估计出动态噪声能量后，可以方便对音频流数据进行分割。动态噪声跟踪分割算法的实施步骤如下：

(1)　估计初始噪声能量 $e_w(0)$ 和信噪比 SNR。

(2)　依式(2)和式(3)，计算含噪信号能量 $e_{\bar{x}}(t)$ 和噪声能量 $e_w(t)$。

(3)　若连续 M 帧信号满足：　$e_{\bar{x}}(t) < \text{SNR} \cdot e_w(t)$，则判定该 M 帧数据为一段“静音”。

(4)　相邻两段“静音”间的数据为一个音频段。

2.3　基于音素识别的分段算法

在连续的音频流中，可以通过区分语音数据和非语音数据对句子进行切分。音素识别器可以用于检测语音和非语音。此时识别的目的主要是区分语音与非语音，并不关心语音部分的具体内容。所以声学模型由上下文无关的单音素 HMM 模型和一个静音(或噪声)HMM 模型构成。静音模型由

所有非语音信号训练得到。CI 音素 HMM 模型表示了所有的语音信息。表 2 是解码结果的一个例子。

表 2 单音素 Viterbi 解码结果

Tab.2 The recognized result of mono-phone decoder

音素	aa	i	c	i	aa	en
起始帧	259	263	270	278	287	292
结束帧	262	269	277	286	291	304
音素	zh	E	sil	t	i	
起始帧	305	312	323	340	345	
结束帧	311	322	339	344	353	

第 1 行是识别出的音素，第 2、3 行分别是该音素的起止时间。其中音素“sil”为静音模型。若静音模型的持续时间足够长，则可认定它确实是一个静音段。相邻两个静音段之间的数据是一个音频段。因此通过检测识别结果中的静音段，我们将输入音频流分割为句子。静音段将被丢弃，不参加后面的处理。

与动态噪声跟踪分割算法相比，基于音素识别的方法要复杂一些。前者只利用了时域能量，计算量很小；而后者需要对音频信号提取特征、训练音素及静音模型，并且进行 Viterbi 解码。基于识别的方法虽然复杂，但它对语音与非语音的描述更精细准确。这一点可以从后面的实验结果中得到验证。

2.4 平滑

对于过短音频段(小于 2s 的句子)，本节用合并的方法减少它们的数目。合并算法的实施步骤如下：

(1)提取每个音频段的 MFCC 特征。

(2)计算每段音频的统计信息：均值和对角协方差矩阵(因为有些音频段数据量较少，可能不足以估计一个稳定的全协方差矩阵)。

(3)对于时长小于 2s 的音频段 S，分别计算它与左右两个相邻音频段间的 KL2 距离 $d_{\mathrm{KL2}}(\boldsymbol{S}_L,\boldsymbol{S})$ 和 $d_{\mathrm{KL2}}(\boldsymbol{S},\boldsymbol{S}_R)$：

$$d_{\mathrm{KL2}}(\boldsymbol{S}_1,\boldsymbol{S}_2)=\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}\left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+\frac{1}{2}\mathrm{tr}\left(\Sigma_1\Sigma_2^{-1}+\Sigma_1^{-1}\Sigma_2-2\boldsymbol{I}\right) \tag{5}$$

其中 $\boldsymbol{I}$ 为单位矩阵。假设 $d_{\mathrm{KL2}}(\boldsymbol{S}_L,\boldsymbol{S})<d_{\mathrm{KL2}}(\boldsymbol{S},\boldsymbol{S}_R)$，若 $d_{\mathrm{KL2}}(\boldsymbol{S}_L,\boldsymbol{S})$ 同时小于某一门限，则说明音频段 $\boldsymbol{S}$ 与 $\boldsymbol{S}_L$ 属于同一类声音，将 $\boldsymbol{S}$ 与 $\boldsymbol{S}_L$ 合并成一段。反之亦然。

(4)遍历所有的音频段，直到没有一个时长小于 2s 或者没有可以合并的段为止。

2.5 分段算法性能评价

分段算法主要产生有两种错误：一类是语音被误判为非语音并丢弃了，丢弃的语音不会被识别，造成不可恢复的删除错误；另一类错误是，由于句子边界判断错误，使得某些词部分被丢失(words cut)。

对于 5 天的新闻联播数据，总长度为 8837s。不同分段算法的性能见表 3。表 3 所用分段算法都经过了粗分割、精细分割和平滑 3 个阶段，只是在精细分割阶段所用算法有所不同。为了叙述上的简便，表 3 中用精细分割阶段所用算法名称代表整个自动分割算法。

作为比较，表 3 同时给出了基于距离的分段算法性能。可见，本文提出的分段算法优于传统的基于距离的分段算法，而基于音素识别的分段算法性能又优于动态噪声跟踪分段算法。

表 3 分段算法性能表

Tab.3 The performance of automatic segmentation

算法	丢弃的语音(s)	Words cut(%)
动态噪声跟踪分段算法	80	0.12
基于音素识别的分段算法	32	0.07
基于距离的分段算法	120	0.25

注：Words cut = 被截断的字数/总字数

3 音频分类算法

对于切分出的音频段或句子，标记或判别其声音类型可以提供更丰富的信息。这些信息对于聚类、自适应等都很有用，能有效地提高识别系统性能。最简单的一个例子就是，将句子分成“男声”、“女声”两大类，并且训练相应的声学模型。识别时根据句子的“性别”选择相应的模型。

高斯混合模型(Gaussian Mixture Model, GMM)可以拟合任意概率分布形式，在说话人识别[10]和语音识别中都有广泛应用。本节利用GMM对音频段进行自动类型判别。图 1 显示的是基于GMM的自动分类过程。首先对输入的音频段提取特征。然后计算音频特征在不同GMM模型下的似然值，最后用最大似然判决准则得到其声音类型。

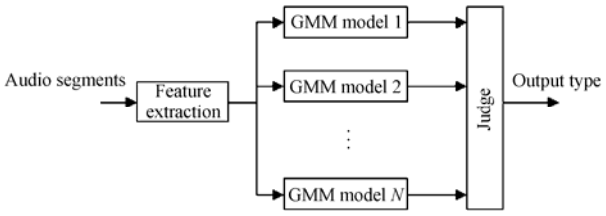

图 1 基于 GMM 的自动分类系统

Fig.1 The audio classification system based on GMM

本文根据广播新闻语料中声音的特点，将声音分成 4 类：男声(male)，女声(female)，音乐(music)和噪声(noise)。对于这种多类问题，通常用混淆矩阵(见表 4)来评估其算法性能。表 4 所用音频段为动态噪声跟踪分割算法切分出的音频段。

表 4 音频分段混淆矩阵 (%)

Tab.4 Confusion matrix of automatic classification (%)

GMM 类型	实际类型				
	male	female	noise	music	混合段
male	98.30	0.42	—	—	1.25
female	0.52	98.04	—	0.26	1.17
noise	9.09	—	72.73	—	18.18
music	—	—	—	80.00	20.00

其中混合段是指分段算法生成的含有两个或两个以上声音类型的音频段。用本节的分类算法无法修正混合段。表 4 显示基于 GMM 的分类方法能够有效地将“男声”和“女声”区别开来。在有混合段干扰的情况下，这两类数据的分类正确率也都在 98%以上。

4 实验结果

本文采用的广播新闻语料识别系统的完整流程如下。首

先对输入的音频流信号进行自动分段。自动分段算法包含 3 个阶段，利用时域能量的粗分段、利于动态噪声或单音素解码的精细分段，以及最后的平滑阶段。接着，自动分类算法对已分割好的音频段进行声学类型判别。判断为噪声或音乐的非语音段将被丢弃，而语音段则得到其“性别”(即男声或女声)。性别相关的 HMM 模型对相应的音频段进行识别，得到文本信息。自适应可以进一步提高系统性能。依据自动分类得到的类信息，对每一个音频类用最大似然线性回归[11](MLLR)算法进行无监督自适应，更新声学模型。最后用自适应后的声学模型重新识别音频段的内容。上面所述的自适应过程可以根据需要重复 1～2 次。

本节，我们在广播新闻识别系统上，测试自动分段、分类算法得到的句子(音频段)的识别性能。实验结果见表 5。作为比较，同时测试了手工分段分类后句子的识别性能。

表 5 显示基于音素的分段算法优于动态噪声跟踪分段算法。另一方面，动态噪声跟踪分段算法在仅仅利用能量信息的情况下，也取得了 12.25%的错误率。表 5 中，手工分段分类后句子的错误率为 11.32%，略优于本文提出的两种分段算法。

表 5　自动分段和分类算法下的系统识别率(%)

Tab.5 The performance of automatic segmentation and classification (%)

算法	正确率	替代率	删除率	插入率	错误率
动态噪声跟踪分段算法+分类算法	88.86	8.75	2.47	1.03	12.25
基于音素识别的分段算法+分类算法	89.25	8.87	1.98	1.01	11.86
手工分段+手工分类	90.04	8.24	1.80	1.28	11.32

经过自动分类算法后，含有语音的音频段被分为两类：“男声”和“女声”。对这两类音频段分别进行无监督的 MLLR 自适应，用自适应后的声学模型重新识别音频段。实验结果见表 6。

表 6 中最后一列给出了自适应后系统误识率比自适应前(见表 5)的误识率相对下降情况。手工分段分类下的误识率相对下降为 5.83%，而本文提出的自动分类分段算法下的误识率下降分别为 5.22%和 6.49%。该结果表明，自动分类的性能与手工分类性能几乎相当。

表 6　无监督自适应后的系统识别率(%)

Tab.6 The performance of unsupervised adaptation (%)

算法	正确率	替代率	删除率	插入率	错误率	错误率相对下降
动态噪声跟踪分段算法+分类算法	89.51	8.31	2.26	1.04	11.61	5.22
基于音素识别的分段算法+分类算法	89.93	8.41	1.75	0.93	11.09	6.49
手工分段+手工分类	90.66	7.76	1.66	1.23	10.66	5.83

5　结束语

本文介绍了中文广播新闻语料识别系统中的自动分段、自动分类算法。本文提出了 3 阶段分段算法，即时域能量粗分段、精细分段和平滑。该算法取得了较好的分段性能。精细分段时采用基于单音素识别的分段算法，识别误识率仅比手工分段高 0.54%。基于 GMM 的分类算法自动将音频段区分为噪声、音乐、男声和女声，其中男声和女声的分类正确率在 98%以上。对新闻联播数据的测试表明，本文提出的算法较好地解决了广播新闻语料的自动分段和分类问题。

参 考 文 献

[1] David Graff. An overview of broadcast news c*orpora. Speech Communication*, 2002, 37(1): 15−26.

[2] Pallett D S. A look a NIST’s benchmark ASR tests: Past, present, and future. IEEE 2003 Automatic Speech Recognition and Understanding workshop, U S. Virgin Islands, 30 Nov.-3 Dec., 2003: 483 – 488.

[3] Wayne C. Mutilingual topic detection and tracking: Successful research enabled by corpora and evaluation. Language Resources and Evaluation Conference (LREC), Athens, Greece, 31 May-2 June, 2000: 1487−1494.

[4] David S. Automatic transcription of broadcast news data. *Speech Communication*, 2002, 37(1): 1−2.

[5] Robinson A J. Connectionist speech recognition of broadcast news. *Speech Communication*, 2002, 37(1): 27−45.

[6] Woodland P C. The development of the HTK broadcast news transcription system: An overview. *Speech Communication*, 2002, 37(1): 47−67.

[7] Hung Jeih-Weih. Automatic metric-based speech segmentation for broadcast news via principal component analysis. In International Conference on Spoken Language Processing (ICSLP) 2000, Beijing China, October 16-20, 2000, (4): 121−124.

[8] Cheng Shi-sian. A sequential metric-based audio segmentation method via the Bayesian information criterion. EuroSpeech 2003, Geneva, Switzerland, Sep. 1–4, 2003: 945−948.

[9] Lin L. Speech enhancement for nonstationary noise environment. Asia-Pacific Conference on Circuits and Systems, 2002, Singapore, Oct. 28–31, 2002, Vol(1): 177–180.

[10] Yamamoto H. Parameter sharing and minimum classification error training of mixtures of factor analyzers for speaker identification. IEEE International Conference on Acoustics Speech and Signal Processing 2004, Montreal Canada, May 17-21, 2004, Vol(1): 17−21.

[11] Legetter C J. Maximum likelihood linear regression for speaker adaptation of continuous density HMM’s. *Computer Speech and Language*, 1995, 9(2): 171−186.

吕　萍：　女，1974 年生，助理研究员，研究方向为语音信号处理、大词表连续语音识别系统、人机界面.

颜永红：　男，1967 年生，中科院声学所中科信利实验室主任，研究员，研究方向为音频信号处理、嵌入式口语对话系统、大词表连续语音识别系统、人机交互.