基于说话人的对话语音分割

邓英

（中国科学院声学研究所）

Conversation Speech Segmentation Based On Speakers

Ying Deng

(Institute of Acoustics, Chinese Academy of Sciences)

1 引言

随着说话人识别研究的深入，多人对话的多个说话人识别受到人们的重视，已成为目前研究的热点[1-3]。从已发表的研究[1-5]来看，方法主要有基于模型的分割方法、基于距离尺度的方法，和基于假设检验的方法。本文提出了对对话语音首先在计算计盒维数的基础上进行语音音节端点检测，然后使用 BIC 检测说话人改变点，最后使用 BIC 进行说话人聚类，并根据聚类结果改进说话人改变点检测结果的方法。

2 基于计盒维数和 BIC 的对话语音分割

本实验的算法主要包括三个部分，首先通过使用计算计盒维数的方法检测出语音信号的静音、噪音和纯语音内容，然后在此基础上使用贝叶斯准则获得对话语音种不同说话人的跳变点，最后使用聚类的方法改进跳变点的检测结果。

2.1 基于计盒维数的静音检测

根据语音信号的声学特征，作者对 8kHz，8Bit 采集的语音进行计盒维数的计算，取窗宽为 128，窗移为 64，图(1)显示了对一段语音进行计算的结果，其中上半部分为声音信号，下半部分为计盒维数曲线。

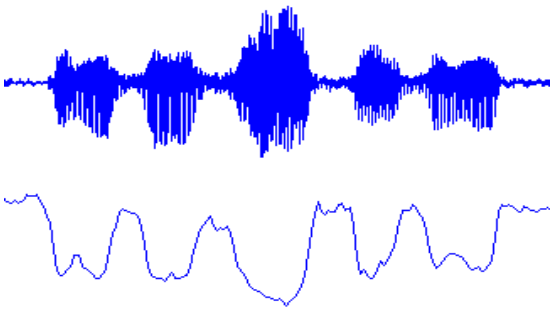


图 1 每个语音窗计算一个计盒维数值，窗移为窗宽的一半

从计算结果可以看出，利用计盒维数可以明显的区分静音、噪音以及语音信号，因此可以使用计盒维数进行静音（噪音）的检测，而且这种方法具有较好的抗噪功能，满足实际应用的需求。

2.2 基于 BIC 的对话语音分割

假设语音信号样本 X 符合多元高斯分布，对在时间 i 是否发生说话人改变作假设检验：

H_0： $(x_1,\ldots,x_n) \sim N(\mu_x, \Sigma_x)$ 整个语音属于同一个说话人，没有发生说话人改变，使用单一的多元高斯模型描述。

H_1：$(x_1,\ldots,x_i) \sim N(\mu_x, \Sigma_x)$ 以及 $(x_i,\ldots,x_n) \sim N(\mu_x, \Sigma_x)$，语音属于两个不同的说话人，在时间 i 发生说话人改变，分别用两个多元高斯模型描述。

H_0 和 H_1 的最大似然比定义为：

$$R(i)=1/2(N_x\log|\Sigma_x|-N_{x1}\log|\Sigma_{x1}|-N_{x2}\log|\Sigma_{x2}|) \quad (1)$$

其中 $\Sigma_x,\Sigma_{x1},\Sigma_{x2}$ 分别为总样本 X，子样本 $X_1=\{x_1,\ldots,x_i\}$ 和子样本 $X_2=\{x_i,\ldots,x_n\}$ 的协方差矩阵，μ_x，μ_{x1}，μ_{x2} 为对应的均值，N_x,N_{x1},N_{x2} 分别为样本数。因此模型的 H_0 和 H_1 的 BIC 值的差等于：

$$\triangle BIC=-R(i)+\Lambda P \quad (2)$$

其中 $p=1/2(p+1/2p(p+1))\log N_x$，$p$ 是样本空间的维数，Λ 是惩罚因子。如果 $\triangle BIC<0$，则表明 H_1 假设成立，即在 i 时刻说话人发生改变，否则 H_0 假设成立，即在 i 时刻说话人未改变。

另外，关于 $|\Sigma_x|$ 的计算，当采用 LPCC 作为语音的特征时，认为各个特征向量无关，可以用对角阵取代协方差矩阵，以减少计算量，但检测效果略有下降，实验中使用协方差矩阵进行计算，以提高检测的准确度。

本文首先用（2.1）的方法对对话语音进行语音信号检测，得到候选分割点集 S={s1,s2,..sn}，然后使用 BIC 依次对所有的候选分割点进行确认和放弃。方法如下：

对于候选分割点 s_{i+1}，需要计算语音段 s_i~s_{i+1} 和语音段 s_{i+1}~s_{i+2} 之间的 BIC 值，使用的特征参数是 13 维的 LPCC 参数。计算结果有两种可能：

(1) $\triangle BIC<0$，s_{i+1} 是真正的语音分割点，继续计算语音段 s_{i+1}~s_{i+2} *和语音段* s_{i+2}~s_{i+3} 的 BIC 值。

(2) $\triangle BIC>=0$，s_{i+1} 不是真正的语音分割点，合并语音段 s_i~s_{i+1} 和 s_{i+1}~s_{i+2}，然后对语音段 s_i~s_{i+2} 以及 s_{i+2}~s_{i+3} 进行 BIC 值计算。

按照以上步骤，依次对每个候选分割点进行确认和放弃，最后得到该段对话语音的说话人改变点结果集 $S=\{s1,s2,\ldots,sl\}$。

2.3 基于 BIC 的说话人聚类

完成对话语音的语音分割后，进一步的工作是要判断哪些语音段是同一个说话人所说的，也就是要进行说话人聚类。本文采用了基于 BIC 的自底向上[1,5]聚类算法，也就是先把每一个语音段都单独作为一个聚类，然后进行迭代合并，每次迭代把两个最符合合并条件的聚类合并成为一个新的聚类，直到算法终止。假设 $S=\{s1,s2,\ldots,sl\}$ 是语音分割的结果，设 $P=\{p1,p2,\ldots,pc\}$ 是聚类集。算法描述如下：

（1）根据语音分割的结果，每一个语音段单独作为一个聚类，也就是 $c=l$ 以及 $pl=sl$。

（2）计算聚类集中的聚类两两之间的 BIC 距离，得到距离矩阵。

（3）找出距离矩阵中最大的 BIC 距离，假设是 $BIC(i,j)$。

（4）如果 $BIC(i,j)>0$，则合并聚类 i 和 j 为新的聚类，更新聚类集，跳转到步骤（2）。

（5）当前聚类集为最终的说话人聚类集，算法停止。

在步骤 2.3 得到的实验结果中，如果两个相邻的语音段属于同一个聚类，那么显然这两个语音段之间的说话人改变点是不存在的，从而可以根据聚类的信息来改进 2.2 中语音分割的结果。这一改进是非常有效的，因为在聚类中，衡量两个聚类之间的距离使用了更多的数据，因此相似性判断的准确性也就更高。本文在实验结果中给出了使用聚类信息改进前后的语音分割结果对比。

3 实验及结果

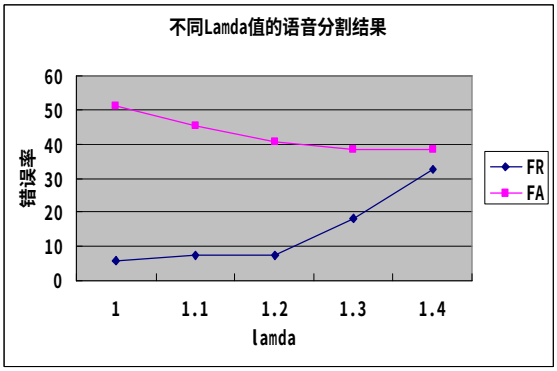
本文的实验数据是在实验室条件下采用 8kHz，8Bit 录得的长度为 687 秒的对话录音，共有 6 男 4 女进行对话，其中说话人的跳变点个数为 136，在进行语音数据处理之前，先对数据进行端点检测、预加重和加汉明窗处理，预加重系数为 0.96。进行语音特征提取时，语音窗取为 128 个采样点（16ms），窗移为

表 1 计盒维数进行静音检测的结果

候选分割点个数	包含分割点个数	命中率[%]	错误率[%]
317	136	100	57.1

64 个采样点（8ms），提取的特征是 13 维的 LPCC 特征。

表 1 显示了使用计盒维数进行静音检测的结果，从中可以看出计盒维数的方法可以很好的区分静音、噪音和语音，从而获得一系列候选的分割点，在此基础上，使用全矩阵和不同 λ 值进行 BIC 计算所得到的实验结果如图 2 所示，使用聚类的方法改进后的实验结果如图 3 所示。



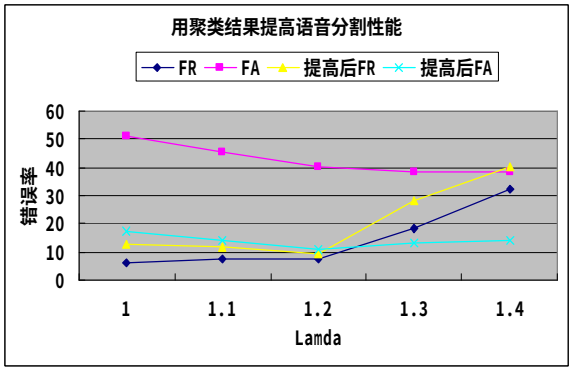


图 2 聚类前采用不同 λ 值的分割结果　　图 3 聚类后采用不同 λ 值的分割结果

表 2 本文方法与其他文献所的结果比较

方法	计盒维数静音检测+BIC	GLR+BIC	音频特征检测+BIC	VQ
$F-measure$	89.78	82.84	81.17	89.30

4 结论

多人对话的语音分割存在很大难度，本文采用了基于计盒维数和贝叶斯准则相结合的方法进行多人对话语音段的分割。方法简单并得到比较好的结果。本文的实验所采用的语音材料是在实验室环境下使用麦克风录制的，而对电话语音可能会出现更复杂的情况，这将作为今后的研究工作。

5 参考文献

[1] Kazumasa MORI and Seiichi NAKAGAWA. Speaker Change Detection and Speaker Clustering using VQ Distortion for Broadcast News Speech Recognition. In Proceedings ICASSP'01, pages 413-416, 2001.

[2] André G. Adami, Sachin S. Kajarekar, Hynek Hermansky, “A NEW SPEAKER CHANGE DETECTION METHOD FOR TWO-SPEAKER SEGMENTATION”, ICASSP2000.

[3] Lu, G. and Hankinson, T, “An Investigation of Automatic Audio Classification and Segmentation”, Proceedings of ICSLP2000, 776-781, 2000.

[4] Lie Lu, Hong-Jiang Zhang, Hao Jiang, "Content Analysis for Audio Classification and Segmentation". IEEE Trans. on Speech and Audio Processing, Vol.10, No.7, pp.504-516, Oct. 2002.

[5] Chen, S.S., Gopalakrishnan, P.S. Clustering via the bayesian information criterion with applications in speech recognition. In: Proceedings of the ICASSP98, Vol. 2, Seattle, Washington: IEEE, 1998. 645~648.

第一作者简介：邓英，中科院声学所研实员，主要参与的工作有语音信号处理、计算机网络及通信。