一种两步判决的说话人分割算法

杨继臣 贺前华 李艳雄 王伟凝
(华南理工大学电子与信息学院 广州 510640)

摘 要：为了提高说话人分割(SS)准确率，该文综合考虑了静音信息和性别信息在 SS 中的作用，提出了一种两步判决的 SS 算法。在从音频流中分离出语音段的基础上，采用两步判决的方法进行 SS。第 1 步采用基频信息为主、性别模型为辅的策略进行 SS，将相邻说话人基频差异大的说话人改变检测出来；第 2 步采用基于性别的改进 T^2 判决公式进行 SS，实现相邻说话人基频差异小的同性别 SS，为此，该文提出了一个基于块的潜在说话人改变点检测算法。实验结果表明，本文算法提高了分割准确率，F_1 度量值可达 85.14%。对于短时长(<2 s)语音段的 SS，该算法和传统的贝叶斯信息判决算法相比，漏检率减少了 16%。

关键词：语音信号处理；两步判决；说话人分割；基频信息；性别信息
中图分类号：TN912.3 **文献标识码**：A **文章编号**：1009-5896(2010)08-2006-04
DOI: 10.3724/SP.J.1146.2009.01072

A Two-step Criterion Algorithm of Speaker Segmentation

Yang Ji-chen He Qian-hua Li Yan-xiong Wang Wei-ning
(*School of Electronic and Information Engineering*, *South China Univercity of Technology*, *Guangzhou* 510640, *China*)

Abstract: To improve the precision of Speaker Segmentation (SS), this paper propose a two-step SS algorithm by making use of silence and gender information. Two-step criterion is used to decide the Speaker Change Point (SCP) within detected speech segmentations. In the first step, pitch difference between different speakers and gender model are used to locate the SCP within neighboring speech segments; In the second step, a gender-based modified T^2 criterion formula is used to locate SCP among the same gender speakers, and potential speaker change point is detected based on chunk. The experiment results show that the proposed algorithm improved SS precision and F_1 can reach 85.14%. For SS with duration less than 2 s, the algorithm can reduce missed detection rate of about 16%, compared with Bayesian information Criterion.

Key words: Speech signal processing; Two-step criterion; Speaker Segmentation (SS); Pitch information; Gender information

1 引言

说话人分割(Speaker Segmentation,SS)是把音频流分割成具有相同声学性质的多个音频段，每个音频段只包含一个说话人的语音信息[1]。SS 是音频索引、说话人跟踪、自动标注的必需步骤[2]。

SS 主要是在音频流中通过说话人改变检测(Speaker Change Detection, SCD),寻找说话人改变点(Speaker Change Point, SCP)实现的[2]。常用的 SCD 方法主要有基于能量、基于模型和基于距离的[3−5] 3 种。在实际应用中，一般是把基于距离和基于模型的方法混合使用[6,7]。

在现有的 SCD 方法中[2,3,6,8−10]，有下面 3 个特点：(1)说话人语音之间的静音所起的作用不是很大，一般都是在前端处理中把静音去掉。一般情况下，大部分的 SCP 都是出现在静音段处，而出现在一句话内部是比较少的；因此，主要的工作是判断静音段处两端的数据是否属于同一个说话人，当然也需要对时长较长的，有可能中间存在 SCP 的语句段进行检测。(2)性别信息一般很少被考虑；由于男性和女性特征有差异，若把不同性别的说话人分别对待，可以提高 SS 的准确率。(3)对于较短时长(<2 s)的说话人改变(Speaker Change, SC)，检测效果一般不理想[3]；对于有性别改变的短时长的 SC，可以通过使用性别改变来检测 SC，从而可以提高 SS 的准确率。

本文综合考虑了静音信息和性别信息，提出了一种两步判决的 SS 算法。首先通过前端处理，得到短时语音段(Short Duration Speech Segment,

2009-08-10 收到，2009-12-01 改回
国家自然科学基金(60972132，60602014)资助课题
通信作者：杨继臣 nisonyoung@yahoo.cn

SDSS)。然后，采用两步判决策略对这些 SDSS 进行 SS：第 1 步以基频为主、性别模型为辅；第 2 步对男性、女性说话人语音分别采用不同比例系数构成的判决公式。

2　数据库介绍

由于 SS 的实验数据库方面，没有统一的数据库，比如，文献[2]采用 TIMIT，文献[9]采用 CNN 和 CCTV 新闻。本文选用中央电视台的新闻联播。选用新闻联播的原因有 3 个：(1)从研究的角度，它有最简单的场景(新闻提要部分基本上无噪音)，也有比较复杂的场景(比如战事报道，暴风雨报道等)。(2)从影响的范围来看，它是全国收视率最高的新闻节目，也是世界上观众最多的新闻节目，影响范围甚广。(3)从存档的角度看，因它是国内外大事的真实记录者，内容涵盖政治、经济、科技、社会、军事、外交、文化、体育等方面，因此它是最有可能存档的新闻节目。在我们的数据库中，总共有 8 天将近 4 个小时的数据。

3　算法介绍

3.1 算法思想

该算法主要有两大部分构成：前端处理和 SS。前端处理的目的是把音频流分割成 SDSS，为后续 SS 作前端处理；它包括按照静音分割和去除非语音两部分。

在我们的数据库中，绝大部分(98%)的不同说话人之间的语音都是不重叠有静音存在；其余不同说话人之间的语音界限不明显，可能是后续说话人在前一个说话人尚未结束之前就开始发言或者前后两说话者的语音连接在一起，也有可能是语音之间填充了背景声。音频流的前端处理方法是，根据短时能量的大小在音频流中找出静音段的位置从而确定了非静音音频信号段。然后采用一个事先训练好的基于KNN和LSP-VQ的分类器[9]去除非静音音频信号段中的非语音信号。经过前端处理后，就得到短时语音段(SDSS)。

根据实验数据统计，绝大部分小于 0.3 s 的 SDSS 与其后的 SDSS 属于同一个说话人，所以本文把小于 0.3 s 的 SDSS 和其后面的 SDSS 合并。

本文根据 SCP 发生的位置，把 SCP 分成两类。第 1 类发生在 SDSS 之间，占了绝大多数(98%)，且这类 SCP 前后的 SDSS 时长比较短，大部分都小于 3 s；第 2 类发生在 SDSS 内部，这类 SCP 对应的 SDSS 时长一般比较长，大部分都大于 3 s。所以可根据 SDSS 的时长是否大于 3 s，以判断其中是否存在潜在说话人改变点(Potential Speaker Change Point, PSCP)，然后再利用基频信息进行进一步的判决。

SS 按照两步判决的方法进行。第 1 步利用基频信息和性别模型将性别改变以及相邻说话人基频差异大的 SCP 判决出来；首先判断出每个 SDSS 的性别，并完成时长大于 3 s 的 SDSS 是否存在 SC 的判决。第 2 步对男、女说话人语音分别对待，首先使用本文提出的基于块的 PSCP 检测算法，然后使用不同比例系数构成的、依据性别的改进 T^2 判决公式；其中特征方面选用 24 维的美尔倒谱系数(Mel-Frequency Cepstral Coefficients, MFCC)。

3.2 基频信息为主性别模型为辅的 SS

不同性别的说话者有不同的基频分布，男性的一般为 60–220 Hz，而女性的一般为 180–550 Hz。因此，对于一段 SDSS，其基频平均值(Pitch Mean, PM)可以用来作为性别判断的依据。本文使用 Praat 软件[11]提取基频值。

一般情况下，男性的 PM 会小于 200 Hz，女性的 PM 会大于 200 Hz。但有时也会出现男性的 PM 大于 200 Hz 和女性的 PM 小于 200 Hz 的情况，根据实验数据的统计分析， PM 在[180，220] Hz 的情况约有 15%。为了减少错误率，本文把 PM 大于 220 Hz 的说话人视为女性和小于 180 Hz 的说话人看成男性。对于 PM 在[180,220] Hz 的 SDSS，使用两个事先训练好的高斯混合模型(Gaussian Mixture Model，GMM)判别其性别：首先对该 SDSS 提取 24 维的 MFCC，其次计算该 SDSS 在这两个 GMM 的概率，取概率大的性别为最后结果。

对于相邻的两 SDSS，如果有性别改变，说明有 SC；另外对于相邻且同性别的两 SDSS，根据实验数据统计：如果它们的 PM 的差大于 87.6 Hz，说话人肯定发生了改变，否则，说话人可能发生也可能没发生改变。根据这两点，对于相邻的两 SDSS，假设它们的 PM 分别为 μ_{P1} 和 μ_{P2}，本文提出一个说话人改变判决规则：

(1)若 μ_{P1} 和 μ_{P2} 所对应的性别不同，说话人发生改变；

(2)若 μ_{P1} 和 μ_{P2} 所对应的性别相同，且 $|\mu_{P1}-\mu_{P2}| > 87.6$ Hz，说话人发生改变。

对于相邻且同性别的两 SDSS，若它们的 PM 差的绝对值小于 87.6 Hz，为了判断这些 SDSS 之间是否有说话人发生改变，进入第 2 阶段的说话人改变判决。

以上介绍的是对说话人语音无重叠的情形的确定方法；下面介绍说话人语音之间没有静音的情形的处理方法：

(1)把时长大于 3 s 的 SDSS 作为处理对象。

(2)把要处理的 SDSS 分成 3 段，取第 1 s 为第 1 段，最后 1 s 为第 3 段；若第 1 段和第 3 段的 PM 差的绝对值大于 87.6 Hz 或第一段和第三段对应的性别不同，说明在该 SDSS 内部有 SCP，进入第三步；否则结束。

(3)把该 SDSS 重新平均分段(大于 3 s 小于 4 s 的分成 5～6 段，大于 4 s 小于 5 s 的分成 6～7 段，大于 5 s 的分成 8 段)，重新计算每段的 PM。把差异最大的两相邻段中间的点作为分割点。

3.3 依据性别的改进 T^2 判决公式的 SS

3.3.1 潜在说话人改变点检测 在这一步，本文使用 T^2 公式[8]进行 PSCP 检测。T^2 公式如下：

$$T^2=\frac{n_1\times n_2}{n_1+n_2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) \tag{1}$$

使用 T^2 公式进行 PSCP 检测时，假设要检测的同性别之间的 SDSS 的个数为 n，如图 1 所示。

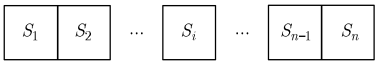

图 1 要检测的同性别之间的短时语音段

本文把每次计算 T^2 的两语音段简称为块，且分别为 a 和 b，每次使用 T^2 公式时，a 和 b 内包含的 SDSS 的个数是不固定的，可能为 1 个也可能为 2 个。假设要检测的同性别之间的 SDSS 个数为 n，本文提出了一个基于块的 PSCP 检测算法，该算法描述如下：

(1)如果 n=2，两个 SDSS 之间的点为 PSCP。

(2)如果 n=3：(a)令第 1 个 SDSS 为 a 和第 2，第 3 个 SDSS 的合并为 b，计算 T^2 值；(b)令第 1、2 个 SDSS 的合并为 a 和第 3 个 SDSS 为 b，计算 T^2 值；(c)取两次计算 T^2 值中较大的那个相对应的点为 PSCP。

(3)如果 n>3：(a)取最前面 3 个 SDSS，进入步骤(2)；(b)首先把前一步得到的 PSCP 前面的 SDSS 抛掉，然后把 PSCP 后面的所有 SDSS 放在一起，重新令 SDSS 的个数为 n，若 n=2，进入步骤(1)；若 n=3，进入步骤(2)；若 n>3，进入步骤(3)。

3.3.2 说话人改变判决 虽然贝叶斯信息判决(BIC)成了现在 SC 判决的主流[2,8,9,10]，但是在不同的声学环境下需要调节 BIC 中的惩罚因子的值[10]。本文没有采用 BIC 公式，而是通过修改 T^2 公式，提出了一个 T_μ^2 公式：

$$T_\mu^2=\lambda\lg(10^n\times T^2)+\beta\lg((\mu_1-\mu_2)'(\mu_1-\mu_2)) \tag{2}$$

这个公式由两部分构成；若有SC，第2部分值一般比较大，若无SC，第2部分值一般比较小。对这两部分都取对数，是因为对数函数有压缩数据的特性。其中，T^2 由公式(3)定义，n代表自然数，这是因为计算得到的 T^2 值一般都很小，为了能有效地把它们区别开来，就需要放大T^2的值。μ_1 和 μ_2 分别代表PSCP两端语音段特征矢量的平均值；λ 和 β 分别代表第1和第2部分的比例系数值，且满足

$$0<\lambda<\beta<1,\quad \lambda+\beta=1 \tag{3}$$

其中 β 大于 λ 是为了在SC时，使第2部分较大。

对于男性和女性，通过设置不同的比例系数 λ 和 β，公式(2)可以演变成下面两个公式：

$$(T_\mu^2)_M=\lambda_M\lg(10^n\times T^2)+\beta_M\lg((\mu_1-\mu_2)'(\mu_1-\mu_2)) \tag{4}$$

$$(T_\mu^2)_F=\lambda_F\lg(10^n\times T^2)+\beta_F\lg((\mu_1-\mu_2)'(\mu_1-\mu_2)) \tag{5}$$

式(4)和式(5)分别用于男性和女性的说话人改变判决。

4 算法评估

4.1 实验设计

将 4 个小时的实验数据分成两部分，2 个小时的数据用于训练，另外 2 个小时的数据用于实验评估。所有的实验数据都是单声道的 wav 格式、16 kHz 的采样率和 16 bit 的量化精度。

实验评估使用准确率(PRC),召回率(RCL)，F_1 度量(PRC 和 RCL 的调和平均值)和漏检率(MDR)[10]对算法进行评估。

4.2 实验结果与分析

通过训练得到的实验参数最优值如表1，Th_M、Th_F 分别代表男性和女性的判决门限值。

表 1 训练得到的实验参数最优值

n	λ_M	β_M	λ_F	β_F	Th_M	Th_F
25	0.25	0.75	0.2	0.8	6.16	4.45

表 2 是采用本文算法做实验得到的结果和其他算法得到的结果准确率的比较。

表 2 本文算法和其他算法准确率比较(%)

	本文	文献[3]	文献[1]
PRC	83.75	85.38	63.4
RCL	86.57	70.70	92.2
F_1	85.14	77.27	73.8

从表 2 中可以看出，本文提出的算法和其他算法相比，在 F_1 度量值方面，比文献[3]和文献[1]都要好。

对于小于 2 s 的 SCD，分别采用本文算法和传统的 BIC 算法，在漏检率方面，本文算法比传统的 BIC 算法减少约 16%(分别为 4.48%和 21.50%)。

图 2 是不把男女分别进行处理，不断改变式(2)中比例系数 λ 和 β ($\lambda+\beta=1$)得到的 PRC、RCL 和 F_1 的变化曲线。

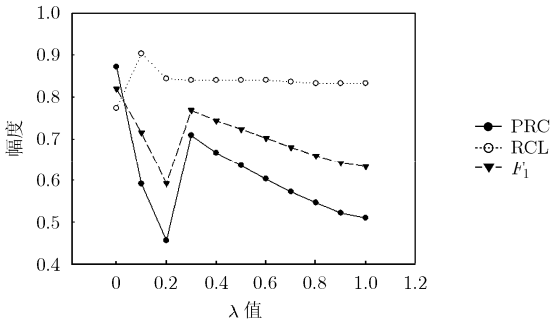

图 2 男女统一处理得到的 PRC、RCL 和 F_1 的变化曲线

由式(2)可以看出，当 λ 等于 1 时，式 T_μ^2 就变成 $\lg(T^2)$公式；由图 2 可以看出，不把男女分开处理，无论怎样改变 λ 和 β 值，得到的结果没有把男女分开处理得到的结果好，这也说明本文提出把男女分开处理可以提高分割准确率的正确性。

5　结束语

本文综合利用了静音信息和性别信息，提出了一个两步判决的说话人分割算法。

实验结果表明，把男女分开处理比放在一起处理，说话人分割性能有了显著提高。该算法 F_1 度量值可达 85.14%，比 BIC 算法提高了将近 8%。另外对于短时长(<2 s)语音段的说话人分割，在漏检率方面，该算法和传统的 BIC 算法相比，减少了约 16%；成功地解决了较短时长、有性别改变的语音段的说话人分割。

参 考 文 献

[1] Sinha R, Tranter S E, Gales M J F, and Woodland P C. The cambridge university March 2005 speaker diarisation system. In proceeding of the European Conference Speech Communication and Technology. Lisbon, Portugal, 2005: 2437–2440.

[2] Kotti M, Benetos E, and Kotropoulos C. Computationally efficient and robust BIC-Based speaker segmentation [J]. *IEEE Transactions on Speech and Audio Processing*, 2008, 16(5): 920–933 .

[3] Chen S and Gopalakrishnan P S. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. Proc. DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, Feb. 1998: 127–132.

[4] El-Khoury E, Senac C, and Pinquier J. Improved speaker diarization system for meetings. In ICASSP2009, Taipei, April, 2009: 4097–4100.

[5] Christoph Boehm and Franz pernkopf. Effective metric-based speaker segmentation in the frequency domain. In ICASSP2009, Taipei, April 2009: 4081–4084.

[6] Kwon S and Narayanan S. Unsupervised speaker indexing using generic models [J]. *IEEE Transactions on Speech and Audio Processing*, 2005, 13(5): 1004–1013.

[7] 郑铁然, 李海峰等. 基于预分割的说话人分割方法. 通信学报, 2009, 30(2): 118–123.
Zheng Tie-ran and Li Hai-feng, *et al.*. Method of speakers segmentation based on pre-segmentation. *Journal of Communcation*, 2009, 30(2): 118–123.

[8] Zhou B and Hansen H L. Efficient audio stream segmentation via the combined T^2-statistics and Bayesian information criterion [J]. *IEEE Transactions on Speech and Audio Processing*, 2005, 13(4): 467–474.

[9] Lu Lie, Zhang Hong-jiang, and Jiang Hao. Content analysis for audio classification and segmentation [J]. *IEEE Transactions on Speech and Audio Processing*, 2002, 10(7): 504–516.

[10] Kotti M, Moschou V, and Kotropoulos C. Speaker segmentation and clustering [J]. *Journal of Signal Processing*, 2008, 88(5): 1091–1124.

[11] Boersma P and Weenink D. Paraat: Doing phonetics by computer. Available: http:/ /www. praat.org/

杨继臣：　男，1980年生，博士生，研究方向为语音信号处理.
贺前华：　男，1965年生，教授，博士生导师，研究方向为语音及音频信号处理、嵌入式系统开发.
李艳雄：　男，1980年生，博士生，研究方向为语音信号处理.