

一种自适应建模的 VAD 方法

腾潇琦¹, 冯 祥², 张翼飞^{2,3}

(1. 北京市互联网信息办公室, 北京 100062;

2. 讯飞智元信息科技有限公司, 安徽 合肥 230088;

3. 上海大学 机电工程与自动化学院, 上海 200072)

摘 要:语音活动检测 (Voice Activity Detection, VAD) 是语音前端特征处理的一个重要环节, 它直接影响到后续处理的效果和效率。主流模型 VAD 对训练数据的依赖度过高, 在不同场景下需要重新训练不同的模型, 这带来的数据标注的工作量是非常惊人的。一种自适应建模的 VAD 方法结合了能量 VAD 和模型 VAD 的优点, 成功地解决了这个问题。它对每一条语音在线地训练出语音和非语音模型, 根据每一帧在模型上的似然度得分给它们打上标签, 经过平滑后就可以很好地找到语音的起点和终点。实验结果表明, 该方法取得了很好的效果, F_1 指标相比传统能量 VAD 提升了 0.031, 说话人分离错误率下降了 0.45%。

关键词:语音活动检测; 能量 VAD; 模型 VAD; 自适应建模

中图分类号: TP301

文献标识码: A

文章编号: 1673-629X(2016)09-0026-04

doi:10.3969/j.issn.1673-629X.2016.09.006

An Voice Activity Detection of Adaptive Modeling

TENG Xiao-qi¹, FENG Xiang², ZHANG Yi-fei^{2,3}

(1. The Office of Internet Information, Beijing 100062, China;

2. Iflytek Intelligent System Co., Ltd., Hefei 230088, China;

3. School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200072, China)

Abstract: Voice Activity Detection (VAD) is an important part of speech front-end features processing which directly affects the effectiveness and efficiency of subsequent processing. Because of over-dependence on training data, the model VAD must train different model in different scenarios that will bring many tasks of data labeling. A VAD method of adaptive modeling, which combines with the advantages of energy VAD and model VAD, solves the problem successfully. It trains speech model and non-speech model online to each voice and labels each frame according to the likelihood score of different model, then the endpoint of voice can be get. The experiments show that this method has achieved a good result. It makes the F_1 parameters increased 0.031 and error rate of speaker separation decreased by 0.45% compared with the traditional energy VAD.

Key words: voice activity detection; energy VAD; model VAD; adaptive modeling

0 引 言

端点检测 (Endpoint Detection) 又称语音活动检测 (Voice Activity Detection, VAD), 是指从一段包含语音的信号中确定出语音的起止点。它广泛应用于通信系统、语音编码等领域, 在语音识别中更是不可或缺的环节。语音信号端点的有效检测不仅能减少语音信号后期处理的运算量, 而且对后续识别的效果有极大的促进作用^[1]。传统的 VAD 方法主要有基于短时能量、过零率、谱熵, 基于混合高斯模型以及基于隐马尔

可夫模型等方法, 它们大体可分为基于能量的 VAD^[2-4] 和基于模型的 VAD^[5-7] 两种。

能量 VAD 是使用能量以及过零率来判断语音和非语音, 该方法优点是简单、速度快, 但是由于它无法滤除噪音和一些非语音信息的声音, 所以效果并不是太理想。模型 VAD 比较复杂, 它是利用语音的统计特性对有效语音、静音、噪音等进行建模, 比较测试语音在各种模型上的得分实现分类。这种方法在效果上要优于能量 VAD, 但是需要大量的人工标注过的数据

收稿日期: 2015-06-02

修回日期: 2015-10-15

网络出版时间: 2016-08-23

基金项目: 北京市科技计划项目 (Z141100006014002)

作者简介: 腾潇琦 (1983-), 女, 硕士, 研究方向为新闻传播。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.tp.20160823.1112.010.html>

进行训练,并且在测试语音和训练语音信道不匹配时可能会引起效果的下降。

文中提出了一种新的自适应 VAD 方法,它结合了能量 VAD 和模型 VAD 的优点,采用了自适应在线建模的方法,解决了测试语音和训练语音信道不匹配的问题,并且不需要离线的训练数据,简化了传统的模型 VAD 方法,效果上可以达到模型 VAD 的水准。

1 传统的能量 VAD

除去静音外,任何一段语音都是一段能量脉冲。一般来说,有效语音拥有相对较高的能量,因此可以采用划门限的方法来检测语音段。最经典的能量 VAD 算法如图 1 所示。

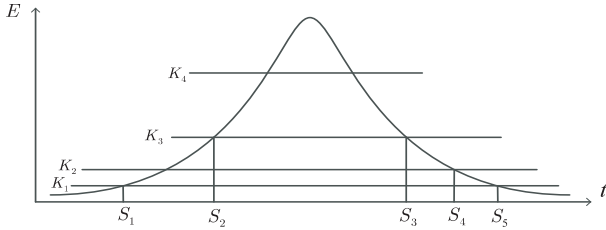


图 1 四门限能量 VAD

从语音的开始到结束阶段,会有一个能量上升和下降的过程,可以利用这个过程建立一套检测语音段的规则。首先对待测语音进行 K -means 聚类,得到四个能量阈值 K_1 、 K_2 、 K_3 和 K_4 。当能量脉冲到来时, E 上升到大于 K_1 ,并且不会再降低到 K_1 之下而是逐渐增大到大于 K_2 时,语音的起点就定为 S_1 ,如果 S_1 到 S_2 的距离过长起点就定为 S_2 ;同样的,在能量下降阶段,当 E 小于 K_2 ,并且不会再上升到 K_2 之上而是逐渐减小到小于 K_3 时,语音的终点就为 S_4 ,如果 S_3 到 S_4 的距离过长终点就定为 S_3 。当峰值能量低于 K_4 时,该段能量脉冲被丢弃,当能量脉冲的持续时间太短时,能量脉冲也被丢弃^[8]。

这种方法在信噪比高的环境下,可以准确检测出语音的起点和终点。但是其固有的缺点还是容易引入较大能量的噪声,包括一些持续时间较长的噪声能量脉冲,影响了语音段标注的准确率。

2 传统的模型 VAD

2.1 混合高斯模型

一个混合高斯模型 (Gaussian Mixture Model, GMM) 由多个高斯概率密度函数加权求和得到,如式 (1):

$$p(x|\lambda) = \sum_{i=1}^M w_i N_i(x) \quad (1)$$

其中, M 为高斯混合模型的混合度; x 为一个 D 维随机向量; w_i 为每个高斯函数的混合权重; $N_i(x)$

为一个 D 维的联合高斯概率分布,见式 (2):

$$N_i(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right\} \quad (2)$$

其中, μ_i 为均值矢量; Σ_i 为协方差矩阵。

至此,整个混合高斯模型 λ 可由 $\{w_i, \mu_i, \Sigma_i\}$ 来描述。

2.2 模型训练

对于 T 个训练矢量 $X = \{x_t, t=1, 2, \dots, T\}$,在用 K -means 聚类确定了初始 λ 的参数后,可以通过经典的 EM (Expectation Maximization) 算法迭代出一个新的混合高斯模型。其中:

$$\bar{w}_i = \frac{1}{T} \sum_{t=1}^T p(\ell_i | x_t, \lambda) \quad (3)$$

$$\bar{\mu}_i = \frac{\sum_{t=1}^T x_t p(\ell_i | x_t, \lambda)}{\sum_{t=1}^T p(\ell_i | x_t, \lambda)} \quad (4)$$

$$\bar{\Sigma}_i = \frac{\sum_{t=1}^T p(\ell_i | x_t, \lambda) (x_t - \bar{\mu}_i) (x_t - \bar{\mu}_i)^T}{\sum_{t=1}^T p(\ell_i | x_t, \lambda)} \quad (5)$$

其中, \bar{w}_i 、 $\bar{\mu}_i$ 、 $\bar{\Sigma}_i$ 分别是新模型的第 i 个高斯的权重、均值和方差; $p(\ell_i | x_t, \lambda)$ 为 x_t 属于第 i 个高斯的后验概率^[9]。

在建立模型前需要大量经过人工标注选出的语音和非语音片段,标注完成后用语音片段训练出一个混合高斯模型 λ_{speech} ,用非语音片段训练出另一个混合高斯模型 $\lambda_{\text{nonpeech}}$ 。比较测试帧在这两种模型上的得分即可实现语音帧和非语音帧的分类,再加入平滑就可以很容易找到语音的端点。

由于可以将噪声片段加入非语音片段中训练 $\lambda_{\text{nonpeech}}$,模型 VAD 可以很好地解决能量 VAD 不能解决的高能量噪声问题。但是此方法的缺点也是很明显的,首先是需要大量人工标注过的数据,其次它对模型的依赖性很高,模型的好坏决定了最终 VAD 的效果,所以对于不同的语音背景环境,需要针对性地重新训练出相应的模型,才能保证结果的准确性。最近几年比较流行的模型 VAD 是基于 DNN (Deep Neural Network) 的 VAD^[10-11],该方法使用 DNN 来建立模型,相比 GMM 模型复杂度更高,效果更佳,但是此方法面临着和传统模型 VAD 一样的问题。

3 自适应建模 VAD

文中提出了一种自适应建模的 VAD 方法,该方法通过在线训练出语音段和非语音段的混合高斯模型,有效去除了静音段以及能量较低的噪音段,而且不像

传统模型 VAD 那样需要大量的训练数据,在信噪比高的环境下取得了较好的效果。流程如图 2 所示。

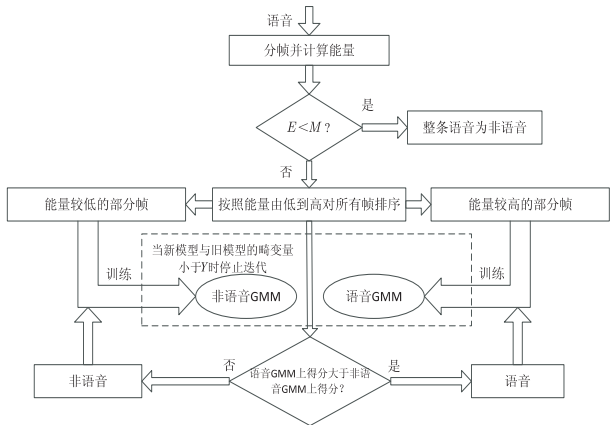


图 2 自适应建模流程图

算法具体步骤如下：

(1) 将待测语音分帧后计算能量,能量最高的帧标记为 A,能量最低的帧标记为 B,计算 $E = (E_A - E_B) / E_B$,将 E 与门限值 M 相比较,若小于 M 则认为此条语音整段都是静音或者噪音,若大于 M 则需要进行第二步。

(2) 如图 3 所示,将每一帧按照能量高低排序,抽取能量较低的一部分帧用以训练出初始的 $\lambda_{nonpeech}$,抽取能量较高的一部分帧用以训练初始的 λ_{speech} 。

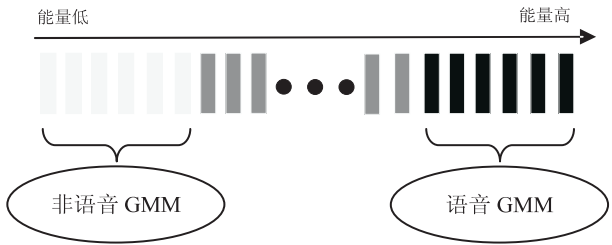


图 3 初始模型的训练

(3) 将语音的所有帧在 $\lambda_{nonpeech}$ 和 λ_{speech} 上计算得分,通过比较两种模型上的得分高低给每一帧数据打上语音或非语音的标签。

(4) 用打上非语音标签的所有帧数据训练一个新的 $\lambda_{nonpeech}$,同样用打上语音标签的所有帧数据训练一个新的 λ_{speech} 。

(5) 重复步骤(3)、(4)若干次,直到新模型相当于上一个模型的畸变量小于 Y 时停止循环。

(6) 再执行一次步骤(3),将每一帧数据都打上语音或非语音的标签。

(7) 使用平滑策略去掉其中的毛刺点。

经过以上七步,可以很容易地标记出语音起始点与结束点。该方法的训练是在线进行的,虽然在效率上相比传统模型 VAD 会有所下降,但是省去了繁琐的离线训练过程。实验结果表明,该方法在信噪比较高的环境下效果数据。

4 实验

实验数据采用的是电话信道下的移动客服数据,一共 3 000 条,都为两人电话中的对话,信噪比较高。其中陕西移动、安徽移动、黑龙江移动的数据各 1 000 条。将陕西移动和安徽移动数据作为开发集用作调参,黑龙江移动的数据作为测试集使用。

基线系统采用传统的四门限能量 VAD 和传统的 GMM 模型 VAD,新系统采用上文介绍的自适应建模 VAD。其中,特征选用 39 维的 MFCC 特征(经过 RASTA 和二阶差分),新系统中在线训练时所用的畸变量 Y 取 5%,M 取 10,GMM 的混合度在下面的开发集实验中选取。

使用的评测指标是 F_1 和 VAD 后的语音进行说话人分离^[12-13]的错误率。其中：

$$F_1 = \frac{2 \times \text{Recall Rate} \times \text{Precision Rate}}{\text{Recall Rate} + \text{Precision Rate}} \tag{6}$$

其中,Recall Rate 表示语音的召回率;Precision Rate 表示语音的正确率。

首先看开发集中不同高斯混合度下的几组测试结果,见表 1。

表 1 不同高斯混合度下的结果

数据集合	GMM 混合度	F_1	说话人分离错误率/%
陕西移动数据	32	0.902	4.56
	64	0.921	4.47
	128	0.923	4.41
安徽移动数据	32	0.882	5.39
	64	0.915	5.30
	128	0.919	5.26

从表 1 可以看出,128 混合度的 GMM 无论是在 F_1 指标还是说话人分离错误率上都取得了最好的效果,但是相比较 64 混合度的 GMM 提升并不明显,然而 128 混合度的 GMM 在运算量上大约是 64 混合度 GMM 的两倍。为了兼顾效率,实验后面的测试选用混合度为 64 的 GMM,表 2 是 1 000 条测试集在三种不同策略系统上的对比。

表 2 新老系统结果对比

VAD 算法	F_1	说话人分离错误率/%
四门限能量 VAD	0.894	4.13
自适应建模 VAD	0.925	3.68
GMM 模型 VAD	0.928	3.70

表 2 的统计结果表明,由于结合了模型 VAD 的优点,自适应建模 VAD 系统的 F_1 指标要好于采用基于传统能量 VAD 方法的系统,并且在后续的降低说话人分离错误率上有明显的优势,而在与传统模型 VAD 的对比中效果略有下降。这是因为自适应建模的 VAD

系统并没有在自适应训练中将高能量的噪音加入到非语音模型的训练中,但是在高能量噪音很少的环境中,效果上几乎和传统模型 VAD 没有区别,而且自适应建模 VAD 的便利性和环境适应性弥补了效果上的不足。

5 结束语

文中提出了一种自适应建模的 VAD 方法,该方法结合了能量 VAD 和模型 VAD 的优点,采用了在线自适应训练 GMM 的方法,避开了传统模型 VAD 中繁杂的人工数据标注和线下模型训练的工作,并且不用担心不同场景下的信道以及背景音不同等问题。该方法在实验中取得了很好的效果, F_1 指标比传统能量 VAD 提高了 0.031,说话人分离错误率也比传统能量 VAD 降低了 0.45%。但是该方法还存在一些不足,首先它对高能量噪音的过滤能力并不好,必须在较高的信噪比环境下才能很好地工作,其次由于是在线的训练模型,所以在运算速度上要弱于传统的能量 VAD 和传统的模型 VAD,这些都是后续需要解决的问题。

参考文献:

[1] 孙战先,储飞黄,王江. 一种自适应语音端点检测算法[J]. 计算机工程与应用,2014,50(1):206-210.

[2] Lamel L,Rabiner L,Rosenberg A,et al. An improved endpoint detector for isolated word recognition[J]. IEEE Transactions on Acoustics Speech & Signal Processing,1981,29(4):777-785.

[3] 张仁志,崔慧娟. 基于短时能量的语音端点检测算法研究

[J]. 电声技术,2005(7):52-54.

[4] 周明忠,吉立新. 基于平均幅度和加权过零率的 VAD 算法及其 FPGA 实现[J]. 信息工程大学学报,2010,11(6):713-718.

[5] Wu J,Zhang X L. An efficient voice activity detection algorithm by combining statistical model and energy detection[J]. Journal on Advances in Signal Processing,2011(2):150-154.

[6] 雷建军,杨震,刘刚,等. 基于复高斯混合模型的鲁棒 VAD 算法[J]. 天津大学学报,2009,42(4):353-356.

[7] 朱杰,韦晓东. 噪声环境中基于 HMM 模型的语音信号端点检测方法[J]. 上海交通大学学报,1998,32(10):14-16.

[8] 章钊,郭武. 话者识别中结合模型和能量的语音激活检测算法[J]. 小型微型计算机系统,2010,31(9):1914-1917.

[9] 郭武. 复杂信道下的说话人识别[D]. 合肥:中国科学技术大学,2007.

[10] Zhang X L,Wu J. Denoising deep neural networks based voice activity detection[C]//Proc of international conference on acoustics, speech, and signal processing. [s. l.]:[s. n.],1988:853-857.

[11] 黎林,朱军. 基于小波分析与神经网络的语音端点检测研究[J]. 电子测量与仪器学报,2013,27(6):528-534.

[12] Reddy A M,Raj B. Soft mask methods for single-channel speaker separation[J]. IEEE Transactions on Audio Speech & Language Processing,2007,15(6):1766-1776.

[13] 张策. 电话信道下说话人分离及识别研究[D]. 北京:中国科学院大学,2013.

(上接第 25 页)

intelligent vehicles symposium. [s. l.]: IEEE,1996:323-326.

[2] 赵娜,袁家斌,徐晗. 智能交通系统综述[J]. 计算机科学,2014,41(11):7-11.

[3] 赵金亮. 自适应交通路口控制系统设计与实现[J]. 太原理工大学学报,2013,44(4):531-535.

[4] Robertson D I,Bretherton R D. Optimizing networks of traffic signals in real time-the SCOOT method[J]. IEEE Transactions on Vehicular Technology,1991,40(1):11-15.

[5] 陈龙,彭森第,魏明. SCATS 优选配时方案的研究[J]. 硅谷,2010(12):77-77.

[6] 宋依青,张润. 自适应交通灯控制系统的设计与实现[J]. 计算机测量与控制,2008,16(4):497-499.

[7] 李清泉,刘濒铎. 基于无线传感器网络交通红绿灯控制系统研究[J]. 科协论坛:下半月,2010(6):71-73.

[8] Yousef K M,Al-Karaki J N,Shatnawi A M. Intelligent traffic

light flow control system using wireless sensors networks[J]. Journal of Information Science and Engineering,2010,26(3):753-768.

[9] 叶文斌. 基于红绿灯优化城市交通控制设计与仿真[D]. 上海:华东师范大学,2015.

[10] Milanés V,Villagra J,Godoy J,et al. An intelligent V2I-based traffic management system[J]. IEEE Transactions on Intelligent Transportation Systems,2012,13(1):49-58.

[11] 魏赞,鲁怀伟,何朝晖. 基于 OPNET 的智能交通系统简单场景下的 V2I 通信性能研究[J]. 自动化与仪器仪表,2015(1):8-10.

[12] 王建强,吴辰文,李晓军. 车联网架构与关键技术研究[J]. 微计算机信息,2011,27(4):156-158.

[13] 孙龙,李孟良,徐达. OBD 技术的应用及其发展[J]. 汽车工程师,2011(10):54-58.

[14] 蒲泓全,贾军营,张小娇,等. ZigBee 网络技术研究综述[J]. 计算机系统应用,2013,22(9):6-11.