GMM-UBM 语种识别技术在无线电监管中的应用*

田 昕[1] 唐 皓[2] 余 江[1] 蔡光卉[1] 肖文珂[1]
(1.云南大学信息学院 昆明 650091;2.云南省无线电监测中心 昆明 650031)

摘 要：提出了一种基于 GMM-UBM 算法的语种识别应用系统，系统结合云南边境无线电监管特点，根据频率、接收地点等的不同，采用不同的识别子库来对越界信号进行识别，能够较好的改善因监测到越界信号质量不佳而导致单纯运用 GMM-UMB 算法识别率不高的问题。经过对河口、芒市监测到实际越界信号的识别实验，识别效果明显提高，证明了该系统的有效性。

关键词：语种识别；边境无线电监测；越界信号
中图分类号：TN912.34 **文献标识码**：A **国家标准学科分类代码**：510.4040

Application of GMM-UBM language identification technique in radio regulation

Tian Xin[1] Tang Hao[2] Yu Jiang[1] Cai Guanghui[1] Xiao Wenke[1]
(1. School of Information Science and Enginnering, Yunnan University, Kunming 650091, China.
2. Yunnan Radio Management Committee. Kunming 650031, China)

Abstract: This paper presents a language identification system based on GMM-UBM algorithm, the system combines the border of Yunnan radio regulatory characteristics, depending on the frequency, receiving location, etc., using different identification sub-libraries for the identification of cross-border signal can be better improvement of cross-border due to poor signal quality monitoring to lead to a simple algorithm using GMM-UMB recognition rate not high. After Mans monitored for cross-border recognition experiment actual signal, identifying the effect of significantly improved, demonstrate the effectiveness of the system.

Keywords: language recognition; border radio monitoring; cross-border signal

1 引 言

随着社会经济的高速发展，各国无线电技术的运用越来越广泛，边境无线电越界问题日益突出。各国通过对无线电信号进行探测、搜索、截获并对其进行分析、识别、监视，以得到其技术参数、特征和辐射位置等技术信息加强边境无线电资源管理，提高无线电监测能力，其中对越界信号的分析和判断国别是一项重要内容。当前主流的语种辨识方法主要包括：

1) 基于声学建模的方法，如混合高斯-背景模(gaussianmixture model-universal background model, GMM-UBM)等。

2)基于音素识别的方法，如基于音素识别的语言模型(phone recognition followed bylanguagemodel, PRLM)等[1]。

对于边境广播信号，GMM-UBM 方法易实现，效率高，实用性强。但单纯的运用 GMM-UBM 对信号进行语种识别，其效果并不理想。

传统的 GMM 模型利用多个高斯分布来拟合特征矢量的空间分布，要求语言训练集比较充分，且能包含该种语言的发音特点[2]。但无论在实际应用中还是在实验过程中训练集都难以符合要求且对噪声鲁棒性比较差，不适合广播语音的语种识别[3]。目前语种识别一般采用高斯混合模型—通用背景模型(GMM-UBM)。高斯混合模型—通用背景模型(GMM-UBM)是由很多不同语种语料训练得到，各种语种的语料具有互相的独立性[4]。

边境越界模拟信号多以广播电视信号、对讲机信号为主，监测到信号噪声较大。同时针对云南边界地区具有语言相似度较高的特点，为了解决这些问题，本文运用多个子库，并结合接收到的无线电信号频率和接收位置等因素来建立语种识别的系统。该系统应用简单，运用该系统对越界信号进行识别，能够有效地减少因为东南亚国家语种相似性较大而造成的识别效果不佳的问题。

收稿日期：2014-10
*项目基金：云南省重点区域(航路、边境)无线电监测技术能力研究、国家自然科学基金(61162004)、云南大学研究生科研题目(ynuy201368)资助项目

2　GMM-UBM 模型

高斯混合模型(GMM)用于说话人识别是由麻省理工学院林肯实验室的 D. A. Reynolds 等人于 1995 年提出的,它是一个状态的连续隐马尔科夫模型,该模型用多个高斯分布的密度函数的组合来描述特征矢量在概率空间的分布情况。在语种识别系统中用高斯混合模型的参数来描述某种语言的语音信号特征矢量的概率分布,第 i 个 N 维的高斯函数可以表示为:

$$b_i(x)=\frac{1}{2\pi/s\mid\sum_i\mid^{1/2}}\exp(-\frac{1}{2}(x-u_i)')\cdot\sum_i^{-1}(x-\mu_i),i=1,\cdots,M \tag{1}$$

x 是由 λ 所得到的概率为:

$$p(x\mid\lambda)=\sum_{i=1}^{M}p^ib_i(x) \tag{2}$$

式中:M 为混合度,p^i 为混合权重,约束条件为 $\sum_{i=1}^{M}p^i=1$。

2000 年又提出 GMM-UBM 说话人识别技术,改善了 GMM 模型对噪声的鲁棒性比较差,不适用于广播语音的语种识别的特点。统计理论表明,用多个高斯概率密度函数的线性组合可以逼近任意分布。因此,GMM 可以对任意的语音成分分布进行精确的描述,这是 GMM 相对其他模型如 VQ 的一个很大的优点。基于以上优点,本文采用 GMM-UBM 进行语种识别。

GMM-UBM 模型由多种语言在不同环境下的语音信号数据训练得到一个高混合度的 GMM 来获得,且数据量越大最终识别效果越好,我们采用 MFCC 来提取特征,Mel 频率的倒谱系数(MFCC)是在语音识别中常用的语音特征[5-8]。但就本文而言,由于客观条件限制,语料的收集难度比较大,语料库有待充足。该模型一般有两种设计方案:一种是将所有训练语料直接训练成一个 GMM 模型(图 1);另一种为阶段化的训练方法,将训练语料进行分类训练,最后将所有模块结合起来形成 GMM 模型(图 2),语种识别系统如图 3 所示。

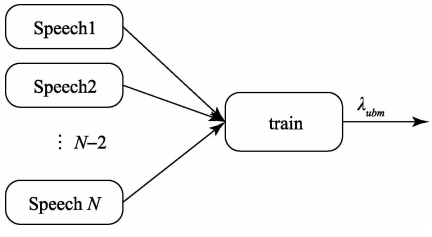

图 1　直接训练 GMM 模型

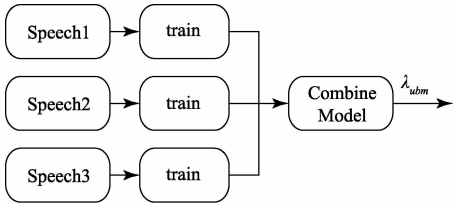

图 2　分几段训练 GMM 模型

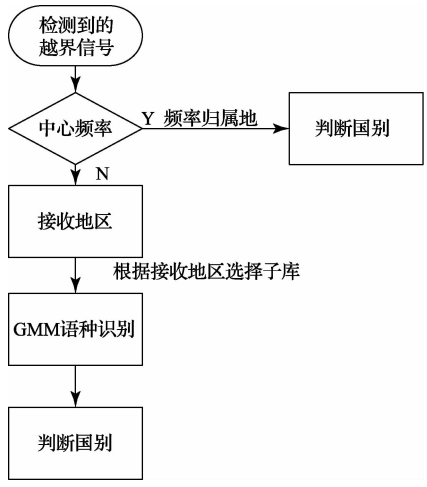

图 3　越界信号国别判别

3　基于云南边境现状的 GMM 识别系统

当语言相似性较高时,通用背景模型的 GMM 识别方法容易出现误判[7-10]。因此根据边境无线电监测的应用特点,可以结合接收到的无线电信号频率、接收地点等因素来进行判别,选择最佳的子库模型,使得最终识别正确率得到提高,以达到能准确判断信号国别的目的。

在日常监测中,监测部门能够统计出一个地区所有非中文信号的频率,例如红河州接收到的越界频率,有 91 MHz、925.94 MHz、95.205 MHz、95.6 MHz、97.06 MHz等,而根据地理环境来看,红河州主要与越南接壤,那么当我们在红河州边境地区监测到越界信号并要对其进行识别时,就可以排除缅甸语、老挝语,进而从语言库中识别时,就能更准确地识别出信号国别。于是可以构想以下思路来进行信号识别。

建立多个子库,根据接收到信号的地区,中心频率的不同情况,用不同的子库来进行识别。例如,在红河地区接收到的信号,语料库中就不包含缅甸语。根据云南地区的特点,东南亚各国语言从发音的角度来说相似性略高,用联合子库的 GMM 的方法来识别时,很多时候各种语言得到的评分都比较高,有时会对单个文件导致识别出错,导致最终识别率不高。用不同的子库来识别,可以有效地避免这类情况的发生,从而提高整个系统的识别率。根据监测部门语种识别的要求,采用的全库暂包含有以下语言:中文、英语、越南语、缅甸语。可以根据实际需求进行扩充。

在图 3 越界信号识别系统中,当监测到一个越界信号,首先判断信号的中心频率是否在以往监测到的数据库中,若在,则可以直接判断其国别。如果不在,则根据接收地区,选择相应的子库,进行 GMM 语种识别,根据评分高低,来判断最终国别。对于子库的选择采用以下 3 个原则:

1)与一个邻国接壤的站点,就选择包含这个国家语料,排除其他接壤国家的语料的子库。

2)与两个国家接壤的站点,同样就选择包含这两个国

家的语料，排除其他接壤国家的语料的子库。

3)根据已有频率库，不同国家边境广播的频点也有所不同，根据接收到的信号的频率，加上合理的浮动，在频率库中进行比对，选择此频率附近处有可能的国家。

4　实验结果及讨论

以模拟德宏地区为例，当确定越界信号的接收地为德宏地区时，可以判断信号不会为库中的越南语，则运用的子库则应包含中文、英语、缅甸语。实验中，需要将信号中语音部分进行切割，尽量排除了全音乐等干扰，每一小段的时常为 10 s，用全库和对应子库对上述信号同时进行识别，二者的识别率结果如表 1 和 2 所示。

表 1　德宏地区信号全库识别结果

语种	匹配个数	概率
缅甸语	6	60%
越南语	2	20%
英语	1	10%
汉语	1	10%

表 2　德宏地区信号子库识别结果

语种	匹配个数	概率
缅甸语	8	80%
英语	2	20%
汉语	0	0%

由上述实验可以看出，通过对语音子库的合理选择，在越界信号噪音较大的情况下，仍然有着较好的识别率，且对于缅甸语的识别率有了明显提高，达到 80%，由此可以判断此越界信号来自缅甸。

为了进行验证该实验的可靠性，又对另一组越界信号进行测试。结果如表 3 和 4 所示。

表 3　芒市机场站信号全库识别结果

语种	匹配个数	概率
缅甸语	6	85.71%
英语	1	14.28%
汉语	0	0

表 4　芒市机场站信号子库识别结果

语种	匹配个数	概率
缅甸语	5	71.42%
越南语	2	28.57%
英语	0	0
汉语	0	0

表 3～4 为芒市实测越界信号语种识别结果，此次实验音频无背景音乐，只伴随有信号本身带有的噪声，音频质量整体好于上个实验。从结果中依然可以看出，通过对子库的预先判别筛选，可以较大的提高识别率，并且信号伴随噪声越小时，得到的识别率越高。此结果，可以为监测部门对越界信号国别的判断提供更准确的结论，大大改善了可能出现的误判情况。

5　结　　论

本文提出了一种基于 GMM-UBM 算法并结合云南边境无线电监管特点的语种识别应用系统。运用该系统，对越界信号的识别率明显提高，可以使得监管部门对截获无线电信号的分析、识别、监视等能力加强，能更加有效地应对越界信号对我国边境无线电安全造成的威胁。

参考文献

[1] 仲海冰，宋彦，戴礼荣. 基于因素识别的便是方法中的因子分析[J]. 模式识别与人工智能，2012，25：105-110.

[2] 吴慧玲，杜成东，毛鹤. 基于 GMM 的说话人识别算法的研究与应用[J]. 现代计算机(专业版)，2014(14)：31-35.

[3] 李思一，戴蓓蒨，王海洋. 基于自带 GMM-UBM 广播语音多种识别[J]. 数据采集与处理，2007，22(1)：14-18.

[4] 徐永华. 基于 GMM-UBM 模型的语种识别[D]. 昆明：云南大学，2010.

[5] 武光利. 基于 GMM 的少数民族自动语种识别系统设计[J]. 自动化与仪器仪表，2013(6)：61-63.

[6] 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.

[7] 李晓阳，吾守尔·斯拉木. 基于 GMM-UBM/SVM 的维吾尔语电话语音监控系统[J]. 计算机应用与软件，2012，29(1)：46-48.

[8] 张丽，杨镇西，吉立新. 语种识别算法中 GSV 计算的定点仿真与实现[J]. 计算机工程与设计，2012，33(2)：679-683.

[9] 曾秀花，杨鉴，徐永华. 语种辨识的多特征信息应用[J]. 计算机工程与应用，2010，46(25)：146-148.

[10] 许辉，热依曼·吐尔逊，吾守尔·斯拉木. 基于 HMM 和 GMM 的维吾尔语联机手写体识别研究[J]. 计算机工程与应用，2014，50(11)：202-205，222.

作者简介

田昕，1989 年出生，在读研究生。主要研究方向为无线电监测。

E-mail：178377213@qq.com