一种用于城市交通的优化声音识别仿真

郑 皓,赵庶旭,屈睿涛
(兰州交通大学 电子与信息工程学院,甘肃 兰州 730070)

摘 要: 随着机动车违法鸣笛现象日益严重,汽车鸣笛声识别可以识别违法鸣笛车辆,并对该行为给出科学有力的证据,因此对城市交通治理有着重要意义。传统方法主要包含基于 GMM-HMM 的概率模型算法、支持向量机等。但其准确率较低,且过程麻烦,给交管部门进行人工复核造成了很大困难。针对此问题,以城市交通汽车鸣笛声识别为背景,结合深度信念网络(DBNs)强大的非线性建模和特征提取能力,提出了一种优化的声音识别方法。该方法采用汽车鸣笛声信号的梅尔频率倒谱系数(MFCC)以及其一二阶差分作为特征参数,用于 DBN 网络的输入,对样本数据进行建模并提取更深层的特征,最后加入 softmax 分类器来实现汽车鸣笛声信号的匹配和识别。该方法获得比 GMM-HMM 更好的识别效果。并通过仿真实验证明了该方法的有效性。

关键词: 神经网络; 深度信念网络; 特征提取; 梅尔频率倒谱系数; 汽车鸣笛声识别
中图分类号: TP391.9 **文献标识码:** A **文章编号:** 1673-629X(2019)02-0060-05
doi: 10.3969/j.issn.1673-629X.2019.02.012

An Optimized Voice Recognition Simulation for Urban Traffic

ZHENG Hao,ZHAO Shu-xu,QU Rui-tao
(School of Electronics and Information Engineering,Lanzhou Jiaotong University,Lanzhou 730070,China)

Abstract: With the increasingly serious whistle of motor vehicles,the whistle recognition of the car can identify the whistle-blowing vehicle and give strong scientific evidence to the behavior,which is of great significance to the urban traffic control. The traditional methods mainly include probabilistic model algorithm based on GMM-HMM,support vector machine and so on,but their accuracy is low and the process is troublesome,which makes it difficult for the traffic control department to carry out manual review. In order to solve this problem,based on the recognition of whistle sound in urban traffic vehicles,we present an optimized sound recognition method in combination of the powerful nonlinear modeling and feature extraction capabilities of DBNs (deep belief networks). The MFCC (Mel-frequency cepstrum coefficients) and its first and second order differences are used as the characteristic parameters for the input of the DBN network,and the sample data are modeled and further features are extracted, Finally softmax classifier is introduced to achieve the car whistle signal matching and identification. This method gives better recognition than GMM-HMM,and its effectiveness is proved by the simulation.

Key words: neural networks; depth belief network; feature extraction; Mel-frequency cepstrum coefficients; car whistle sound recognition

1 概 述

从 20 世纪末开始,城市居民的机动车持有量和驾驶员的数量在大幅增加,城市交通呈日渐繁忙的趋势,交通拥堵现象也变得越来越严重。众所周知,除了机动车尾气排放对大气环境造成的影响,机动车的鸣笛声也给城市环境造成了越来越大的影响。再加上部分驾驶员环境保护意识淡薄,以至于汽车的违法鸣笛在城市区域频繁发生,对市民的正常生活造成了恶劣的影响。保持道路交通的安全、畅通、有序,遏制汽车违法鸣笛行为的上升势头,成为交管和环境保护部门函待解决的重要问题。因此,相关执法部门也颁布了相应的法律法规来禁止或者限制机动车在医院、居民区、学校等城市人口密集区鸣笛,但是这一规定在具体执行中却遭到了巨大的困难,人工执法来判断违法鸣笛车辆不仅效率低,且浪费了大量的人力物力,换来的是较低的可靠性和大量的误判,可行性不高。

同济大学孙懋珩团队 2010 年提出了一种基于麦克风阵列和闭式球形差值算法的汽车鸣笛声识别系

收稿日期: 2018-02-12 **修回日期:** 2018-06-13 **网络出版时间:** 2018-11-15
基金项目: 甘肃省自然科学基金(1504GKCA018)
作者简介: 郑 皓(1994-),男,硕士研究生,研究方向为机器学习; 赵庶旭,教授,博导,CCF 会员(27584M),研究方向为智能交通与机器学习。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20181115.1047.022.html

统,用于协助相关执法部门准确有效地锁定违法鸣笛车辆的位置[1-3]。

近年来国内部分城市也开始引入汽车违法鸣笛识别系统对违法鸣笛行为进行监控,例如北京协和医院西门的"电子警察监管"系统、沈阳的"声呐电子警察"系统等。但是这些只能单方面地确定目标车辆的位置,且准确性不高,存在较多误判。特别是针对部分司机拒不承认鸣笛行为的现象,不能给出有力的证据。所以要对鸣笛识别系统的识别结果进行人工复核,将识别到的目标车辆鸣笛声与鸣笛声识别系统捕捉到的声音信号加以匹配和鉴别,进一步提升汽车鸣笛声定位的准确性。

传统的鉴别方法主要有基于 GMM-HMM 的概率模型算法[4]、支持向量机等。

在深度学习的研究上,2011 年,Hinton 等[5-6] 提出了使用深度信念网络在电话语音中提取深层特征,对电话语音进行识别,给出仅 19.4%的错误率。2012 年,Andrew 等[7] 提出一种顺序深度信念网络模型(SD-BN),其在所有隐藏层及输出层使用固有的顺序模型,使得潜在变量可以模拟长距离现象,在 TIMIT 电话识别实验中效果较好。2015 年,周晓敏等[8] 提出了一种基于小波矩和 BP 网络的声音识别方法,用于噪声环境下的声音识别,并取得了不错的效果。2017 年,陈秋菊等[9] 提出了一种基于优化正交匹配追踪和深度置信网络的声音识别方法,用于不同环境声音影响下的声音识别,并且能够有效地识别各种不同环境下的声音事件。因此深度学习技术在声音识别问题中有着巨大的优势。

GMM 模型作为一个使用时间较长的传统模型,存在很多缺点,例如: 需要对数据的状态做出假设; 需人工选择特征参数且不易于多种特征的组合; 不能接受较大维度的数据输入,计算复杂; 判决准确率不高。

因此,文中提出一种基于深度信念网络的识别方法。结合深度神经网络强大的非线性建模能力和特征提取能力,对识别到的目标车辆加以鉴别,以提高识别的准确性,从而提高人工复核的效率,对鸣笛识别系统的识别结果提供科学有力的支持。

2　数据预处理

2.1　预处理

通过麦克风阵列采集到的汽车鸣笛声音频信号是模拟信号,需要对其进行采样和量化,获得可用于计算机处理的数字音频信号。此过程通过麦克风和计算机声卡完成,采用 Cool Edit Pro 2.1 软件完成对汽车鸣笛声信号的采集并获得音频信号片段,用于后期的处理。

2.2　特征提取

2.2.1　梅尔频率倒谱系数

梅尔频率倒谱系数[10] (Mel-frequency cepstrum coefficients,MFCC) 的分析是根据人的听觉机理,基于滤波器组的频率分析,滤波器组的带宽间隔约为临界子带的间隔。期望获得更好的声音特性,MFCC 分析依据的听觉机理有两个。

(1) 人的主观感知频域的划定不是线性的,如式 1 所示:

$$F_{\text{mel}} = 1\ 125\log(1 + f/700) \tag{1}$$

其中,F_{mel} 是以梅尔(Mel) 为单位的感知频率; f 是以 Hz 为单位的实际频率。

(2) 临界带(critical band)。频率群相应于人耳基底膜分成许多很小的部分,每一部分对应一个频率群,对应于同一频率群的那些频率的声音,在大脑中是叠加在一起进行评价的。按临界带的划分,将声音在频域上划分成一系列频率群组成了滤波器组,即 Mel 滤波器组。

2.2.2　MFCC 的提取过程

MFCC 的提取过程如图 1 所示。

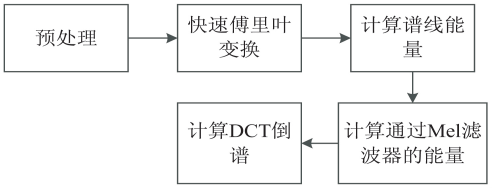

图 1　MFCC 提取流程

(1) 信号预处理。

预处理的部分包括预加重、分帧和加窗函数三个步骤。

预加重: 预加重处理其实就是将音频信号通过一个高通滤波器,目的是提升信号的高频部分,使信号的频谱变得更平坦,保持在低频到高频的整个频带中,能用同样的信噪比求频谱。高通滤波器的传递函数一般设为:

$$H(z) = 1 - \alpha z^{-1} \tag{2}$$

其中,α 的值介于 0.9~1.0 之间,通常取 0.97。

分帧: 现在成熟的信号处理技术往往分析的都是时不变信号即稳态信号,由于音频信号是一个准稳态信号,把它切分成较短的帧,在每帧中可将其看作稳态信号,可以用处理稳态信号的方法来处理。同时,为了使一帧与下一帧之间的参数能较平稳地过渡,在相邻两帧之间有互相重叠的部分。

加窗函数: 加窗函数的目的是为了减少信号在频域中的泄漏。对比矩形窗和汉明窗可以看出,矩形窗的主瓣宽度小于汉明窗,具有较高的频谱分辨率,但是矩形窗的旁瓣峰值较大,因此各谱线之间的相互干扰

较为严重。相比之下,虽然汉明窗的主瓣宽度较宽,约为矩形窗的一倍多,但是它的旁瓣衰减较大,具有更平滑的低通特性,能在较高的程度上反映短时信号的频率特性,适合处理频谱表现复杂,存在多个频率分量的信号。所以文中选择对每一帧信号加汉明窗。信号 $\chi(n)$ 经过预处理后为 $\chi_i(m)$,其下标 i 表示分帧后的第 i 帧。

(2) 快速傅里叶变换(FFT)。

对每一帧信号进行 FFT 变换,从时域数据转变为频域数据,如下:

$$X(i,k) = \mathrm{FFT}[\chi_i(m)] \tag{3}$$

(3) 计算谱线能量。

对每一帧 FFT 后的数据计算谱线能量,如下:

$$E(i,k) = [X(i,k)]^2 \tag{4}$$

(4) 计算通过 Mel 滤波器的能量。

将求出的每帧谱线能量通过 Mel 滤波器,并计算在 Mel 滤波器中的能量。在频域中相当于把每帧的能量谱 $E(i,k)$ (其中 i 表示第 i 帧,k 表示频域中的第 k 条谱线)与 Mel 滤波器的频域响应相乘并相加,如下:

$$S(i,m) = \sum_{k=0}^{N-1} E(i,k)\, H_m(k),\ 0 \leqslant m < M \tag{5}$$

(5) 计算 DCT 倒谱。

把 Mel 滤波器的能量取对数后计算 DCT,如下:

$$\mathrm{mfcc}(i,n) = \sqrt{\frac{2}{M}} \sum_{m=0}^{M-1} \log[S(i,m)] \cos\frac{\pi n(2m-1)}{2M} \tag{6}$$

其中,$S(i,m)$ 是由上一步求出的 Mel 滤波器能量;m 是指第 m 个 Mel 滤波器(共有 M 个);i 表示第 i 帧;n 是 DCT 后的谱线。

2.2.3 梅尔频率倒谱系数差分系数

上一节得到的是静态的 MFCC 系数,只能表征音频谱的即时信息。实验表明,音频谱的动态信息中也包含了很重要的成分,可以对声音信号加以区分,可以提高对汽车鸣笛声音的识别准确率。音频倒谱的动态信息可以用来表征声音信号的特征参数随时间变化的规律。这里引入 MFCC 的一阶差分和二阶差分系数,来表示音频信号的动态特征。然后将原始的 MFCC 特征和 MFCC 动态特征组合作为特征向量进行训练和识别,可以明显提高识别性能。

2.3 特征表示

由特征参数的曲线表示可以看出,同一辆汽车鸣笛声不同帧的特征参数存在一定的相似性,不同汽车鸣笛声的特征参数很明显具有一定的差异性,通过这种差异性,将不同汽车的鸣笛声加以区分,并将同一辆汽车的鸣笛声作匹配处理,从而对现有的汽车违法鸣笛识别系统做一个准确性的验证,以进一步提高其识别的准确性。

3 DBNs 网络模型

3.1 人工神经网络的组成

与生物神经网络类似,人工神经网络的基本计算单元是人工神经元,一般是由多个输入和一个输出组成的非线性单元,可以有反馈输入和阈值参数,同时能够与其他神经元连接。在人工神经网络中,神经元常被称为“处理单元”;为了方便对人工神经网络的描述,在整个网络中,把单个的人工神经元称为网络“节点”。人工神经元是对生物神经元的一种形式化描述,它抽象地模仿了生物神经元的信息处理过程,并用数学语言来描述生物神经元处理信息的过程。为了直观地对生物神经元的结构和功能进行模拟,采用结构图的方式,将人工神经元模型结构表示出来。

3.2 网络的拓扑结构

文中用到的深度信念网络(deep belief networks,DBNs)相比于传统的 GMM-HMM 模型有很多优点,包括不需要对输入数据的状态做假设,易于多种特性的组合,使用更多的数据来约束一个参量,以取得更好的识别结果。整个网络结构包含多个隐层,即由多个 RBM 堆叠而成,相邻两层节点之间的关系为全连接,采用无监督的预训练方法对权值 w 进行初始化,最后一个隐层和输出层之间用 softmax 网络连接,通过误差反向传播算法(back propagation,BP)对整个网络参数进行调整。

RBM(restricted Boltzmann machines)是一种随机神经网络[11-13],它借鉴了统计物理学思想。神经元状态变化引入了统计概率。图2为RBM的结构,由一个输入层和一个隐层组成,之间由权值 w 连接。

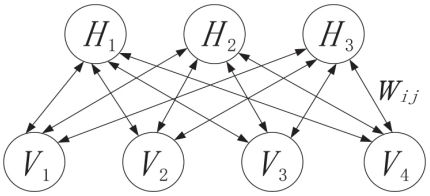

图2 RBM 结构

单个神经元结构如图3所示,包含多个输入和一个输出,参数包括和每个输入所连接的权值 w 以及偏量 b。而单个神经元的输入和输出所对应的激活函数为 $P_j(b_j, W_{ij})$,如下:

$$P_j = \sigma\left(b_j + \sum_{i=1}^{n} W_{ij} v_i\right) \tag{7}$$

其中

$$\sigma(x) = \frac{1}{1+\mathrm{e}^{-x}} \tag{8}$$

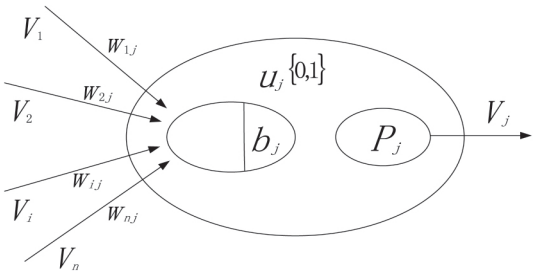

图 3　神经元结构

3.3　网络模型的训练

(1) 训练 RBM,训练的目的是由初始化权值 w 到确定性权值 w,由输入数据结合权值和偏量计算出 p 的值,当 p 的值大于一个随机值时,神经元 u 更新为 1,否则为 0,即神经元的开和关。

(2) 根据隐层 H 的值计算出重建数据,希望输入数据和重建数据相差越来越小,即两者做差,更新权值 w。

(3) 只有当第一个 RBM 完成训练之后,得到了 H 层的值,才可以进行下一个 RBM 的训练。接下来就是用有监督的反向传播算法微调每层之间的权值 w,如图 4 及式 9-11 所示。

$$P(u_j = 1 \mid v) = \sigma\left(b_j + \sum_{i=1}^{n} W_{ij} v_i\right) \tag{9}$$

$$P(v_i = 1 \mid h) = \sigma\left(b_j + \sum_{i=1}^{n} W_{ij} h_j\right) \tag{10}$$

$$\Delta w_{ij} = \varepsilon \frac{\partial \log p(v)}{\partial w_{ij}} \approx \varepsilon\left(< v_i h_j >_{\text{data}} - < v_i h_j >_{\text{recon}}\right) \tag{11}$$

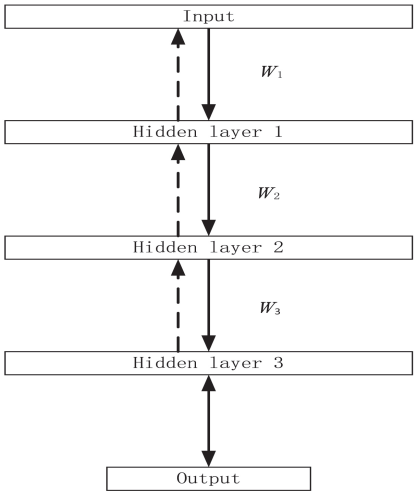

图 4　DBN 网络结构

图 4 为文中所用的网络结构图。首先由 2.2 节所述方法提取汽车鸣笛声的 MFCC 特征参数,以及它的一、二阶差分; 将该特征参数作为网络的输入层(Input) 输入网络,对第一层 RBM 进行训练,得到第一个隐层(Hidden layer 1) 的参数,包括连接各神经元的权重以及到各个神经元的偏量。将第一个隐层作为第二个 RBM 的输入,逐层进行训练。最后在第三个隐层下方加入 softmax 网络作为输出层(Output) 对输入数据进行分类。

4　实　验

4.1　实验过程

共采集了 10 辆不同汽车个体的鸣笛声样本。样本数据来自兰州军鼎车辆检验中心,数据采集的先后顺序随机,分别是:

(1) 宇通(一),大客;
(2) 大众朗逸,轿车;
(3) 金杯,轻型封闭货车;
(4) 中联重型载货专项作业车;
(5) 大众桑塔纳,轿车;
(6) 宇通(二),大客;
(7) 丰田,中型普通客车;
(8) 大众帕萨特(一),轿车;
(9) 江淮,中型普通客车;
(10) 大众帕萨特(二),轿车。

将每辆汽车的鸣笛声数据视为一个类,因此实验中数据包含 10 个类,每一类均包含 1 000 条样本数据,该样本数据均出自同一辆汽车,即所有样本个数为 1 000* 10=10 000 个。

样本分帧: 单个数据样本时长均为 0.05 s,约2 230 个采样点(采样频率为 44.1 kHz,数据位数 16 位,单声道),256 点分为一帧,两帧之间重复长度为 128 点,即单个样本共分为 16 帧,即构成一个连续 16 帧的特征块,将这些特征块作为深度神经网络训练样本的输入。

MFCC 参数: 以每帧为单位提取特征参数,取前 12 阶,即每帧数据的特征参数为 12 维,所以每个样本共包含 12* 16=192 维特征参数,作为 DBN 网络的输入层。

标签的选择: 文中用到的数据种类可分为 10 个类,即需要十个标签对样本数据进行标注。

训练模型:

DBN 输入层: 192 维特征参数;

DBN 输出层: 10 维(10 个标签);

训练集: 每个类取 800 条样本数据作为训练集,共 800* 10=8 000 条样本;

验证集: 每个类取 100 条样本数据作为验证集,共 100* 10=1 000 条样本;

测试集: 每个类取 100 条样本数据作为测试集,共 100* 10=1 000 条样本。

将 MFCC 及其一阶差分的组合作为特征参数进行输入,进行上述实验。

将 MFCC 和其一阶差分,二阶差分的组合作为特

征参数进行输入，进行上述实验。

4.2 实验结果与分析

实验结果可信性标准，用识别错误率(EER)来衡量，EER 越低，表示识别效果越好。EER 的计算方法如下:

$$P_{\mathrm{EER}} = \frac{N_{\mathrm{f}}}{N} \tag{16}$$

其中，N_{f} 表示识别错误数据条数; N 表示实验数据总数。

分别将 MFCC，MFCC+一阶差分，MFCC+一、二阶差分作为特征进行识别，第二行代表每种特征参数识别的错误率，可以看出 MFCC 及其一阶差分，二阶差分的组合特征准确率较高。

在选取的特征统一的情况下，选取的特征块所包含的帧数对识别准确率也存在一定的影响，当特征块帧数在 16 帧时识别效果较好。

图 5 为不同特征组合与不同的测试声音长度对识别错误率的影响。可以看出，当输入特征选择 36 维 MFCC，即 MFCC 与其一阶，二阶差分的组合时，测试声音长度越长，识别的错误率越低，识别效果越好。

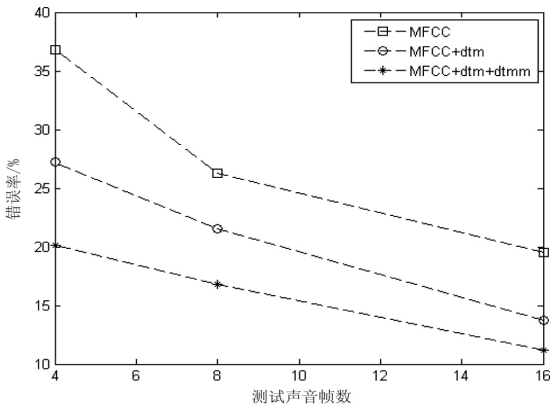

图 5 特征组合与测试长度对准确率的影响

在选取 16 帧的情况下，MFCC+dtm+dtmm 组合特征为输入数据的实验结果，与同样帧数下 GMM 模型的错误率进行比较，错误率明显降低，识别效果显著提高。结果见表 1。

表 1 DBN 与传统 GMM 的准确率对比 %

模型	错误率
GMM	32.9
DBN	11.2

所以基于 DBN 模型的识别方法可以解决传统模型的缺点且显著提高识别效果。

5 结束语

提出了一种基于深度信念网络的声音识别方法，用于城市交通汽车鸣笛声识别。该方法通过深度信念网络，提取和学习汽车鸣笛声 MFCC 特征参数中的深层特征，从而建立更精确的网络模型来对汽车鸣笛声加以匹配和鉴别。实验结果表明，该方法能准确地将不同汽车个体的鸣笛声加以匹配，相比于传统 GMM 模型，效率更高，识别效果更好。解决了现有的汽车违法鸣笛识别系统中存在大量误判的现象，且为人工复核提供了科学有力的支撑，对今后的各种汽车违法鸣笛识别系统的准确性验证具有重要的意义。

参考文献:

[1] 孙懋珩，俞莹婷.汽车鸣笛声定位系统仿真[J].声学技术，2009，28(5)：640-644.

[2] 施珂毅，孙懋珩.汽车鸣笛声定位算法研究及系统实现[J].中国新通信，2009，11(1)：64-67.

[3] 徐 静，李卫红，孙懋珩，等.基于麦克风阵列的车辆鸣笛嗅探器[J].数据采集与处理，2012，27：262-266.

[4] 吕霄云.基于 MFCC 和 GMM 的异常声音识别算法研究[D].成都：西南交通大学，2010.

[5] MOHAMED A R，SAINATH T N，DAHL G，et al.Deep belief networks using discriminative features for phone recognition[C]//IEEE international conference on acoustics，speech and signal processing.Prague，Czech Republic：IEEE，2011：5060-5063.

[6] MOHAMED A，DAHL G E，HINTON G.Acoustic modeling using deep belief networks[J].IEEE Transactions on Audio，Speech，& Language Processing，2011，20(1)：14-22.

[7] ANDREW G，BILMES J. Sequential deep belief networks[C]//IEEE international conference on acoustics，speech and signal processing.Kyoto，Japan：IEEE，2012：4265-4268.

[8] 周晓敏，李 应.基于小波矩和 BP 网络的声音识别[J].计算机工程与应用，2015，51(3)：192-196.

[9] 陈秋菊，李 应.基于优化正交匹配追踪和深度置信网的声音识别[J].计算机应用，2017，37(2)：505-511.

[10] 蒋翠清，邵宏波.基于 MFCC 与改进 ACF 的汽车声音识别算法研究[J].计算机技术与发展，2015，25(2)：140-143.

[11] 张 娟，蒋 芸，胡学伟，等.基于快速持续对比散度的卷积受限玻尔兹曼机[J].计算机工程，2016，42(9)：174-179.

[12] NOROUZI M，RANJBAR M，MORI G.Stacks of convolutional restricted Boltzmann machines for shift-invariant feature learning[C]//IEEE conference on computer vision and pattern recognition.Miami，FL，USA：IEEE，2009：2735-2742.

[13] MOCANU D C，MOCANU E，NGUYEN P H，et al.A topological insight into restricted Boltzmann machines[J].Machine Learning，2016，104(2-3)：243-270.