



基于卷积神经网络的电子变调语音检测算法

徐宏伟, 严迪群, 阳帆, 王让定, 金超, 向立
(宁波大学, 浙江 宁波 315211)

摘要: 提出了一种基于梅尔倒谱系数统计特征和卷积神经网络的电子变调语音检测算法。首先提取待测语音的梅尔倒谱系数及其差分系数, 并将上述系数的统计特征进行有针对性的构造, 作为卷积神经网络的输入。从卷积核尺寸、卷积核个数以及池化层尺寸等方面, 对 24 种不同网络结构进行了测试评估, 最终确定了可有效用于变调检测的卷积神经网络结构。实验结果表明, 所提出的算法能够有效地检测出电子变调的痕迹, 并可准确估计出电子变调语音经过的具体伪造操作, 为电子变调语音的检测提供了一种新的方法。

关键词: 电子变调语音; 卷积神经网络; 梅尔倒谱系数

中图分类号: TP391

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2018041

Detection algorithm of electronic disguised voice based on convolutional neural network

XU Hongwei, YAN Diqun, YANG Fan, WANG Rangding, JIN Chao, XIANG Li
Ningbo University, Ningbo 315211, China

Abstract: An electronic disguised voice detection algorithm based on the statistical features of MFCC and the convolution neural network was proposed. Firstly, the statistical features of MFCC were extracted and reconstructed as the input of convolution neural network. Considering the convolution kernel size, the number of convolution kernels and the pooling size, 24 different network structures were evaluated in this work. Finally, the convolution neural network structure which could be effectively used for electronic disguised voice detection was determined. The experimental results show that the proposed algorithm can effectively detect the trace of electronic disguising. Meanwhile, the specific forgery operation of the electronic disguised voice can also be estimated.

Key words: electronic disguised voice, convolutional neural network, MFCC

收稿日期: 2017-06-15; 修回日期: 2017-10-02

基金项目: 国家自然科学基金资助项目 (No. 61300055, No. 61672302); 浙江省自然科学基金资助项目 (No. LZ15F020002, No. LY17F020010); 宁波市自然科学基金资助项目 (No. 2017A610123)

Foundation Items: The National Natural Science Foundation of China (No.61300055, No.61672302), Natural Science Foundation of Zhejiang Province of China (No. LZ15F020002, No. LY17F020010), Natural Science Foundation of Ningbo City of China (No. 2017A610123)

1 引言

对语音进行变调处理^[1]是犯罪分子用来掩盖自己身份的一种行为。语音变调从实现的方式上大体可以分为两类：非电子变调和电子变调。通过捏鼻、捂嘴或在嘴里放置异物等物理操作来改变说话人的语音特征，是典型的非电子变调行为。电子变调则是通过变调算法改变说话人的语音特征。与非电子变调相比，电子变调可以实现更具迷惑性的伪装。电子变调可在保持变调后语音自然度的情况下，改变原有说话人语音特征，从而达到掩盖说话人身份信息的目的。

目前电子变调语音造成的威胁并未得到足够的重视，针对电子变调语音的检测研究还相对较少。Wu 等人^[2, 3]提出了一种基于梅尔倒谱系数统计特征和支持向量机的电子变调语音检测算法，并针对该算法做了交叉验证实验，证明该算法能对典型变调软件产生的电子变调语音进行检测。Cao 等人^[4]在参考文献[3]研究的基础上，对噪声环境下的电子变调语音检测算法的顽健性做了研究。他们在电子变调语音中加入了不同信噪比和不同类型的环境噪声，并采用了线性倒谱系数和共振峰的组合特征作为检测特征，使得检测算法的顽健性得到了一定程度的提高。这些方法都使用了一种由 10 个 SVM 分类器组成的集成分类器，分类模型较为复杂。同时，这种分类器并不能判断出所变音调的种类，即原始语音是经过了升调操作，还是经过了降调操作。

本文在已有研究的基础上，提出一种新的电子变调语音检测算法，以梅尔倒谱系数 (Mel-scale frequency cepstral coefficient, MFCC)^[5]的统计量作为检测特征，并引入了卷积神经网络 (convolutional neural network, CNN)^[6, 7]作为分类器。实验对比了 24 种不同的网络结构的实验结果，最终确定了一种仅有 4 个网络层的 CNN 作为分类器。在实验中，使用 3 种常用的语音处理软

件 Audacity^[8]、CoolEdit^[9]和 RTISI^[10, 11]作为电子变调语音的伪造工具，并进行了交叉验证，最终的检测准确率达到 96% 以上。本算法使用一种非常简单的 CNN 结构作为分类器，取代了原有的复杂的集成分类器的同时，同时可以有效地判断出变调语音是经过了升调处理还是降调处理。

2 相关研究工作

2.1 电子变调语音

对语音信号进行电子变调操作，其本质是通过改变原始语音的音调来伪装说话人的语音特征，从而达到掩盖说话人身份信息的目的。最容易实现的电子语音变调是直接改变语音的播放速度。这种方式通过加快语音的播放速度或者减慢语音的播放速度，实现语音的升调或者降调。但是，通过这种方式产生的电子变调语音韵律不自然，通常不被伪装者所采用。另一种常用的电子伪装语音，是在保证语音时长不变的情况下，实现对语音音调的升高或者降低。这种变调不变速方法产生的伪装语音，在音色、音调等方面更加逼真自然，不易被人们察觉。本文正是针对这种变调不变速的电子变调语音进行研究。

变调不变速的语音变调方法主要有时域插值法和频域法。经典的时域插值法^[12]通过在相邻两个数据点之间线性插入新数据点来达到上采样的目的，也可以通过每隔一定时间抽取一些点来达到下采样的目的。上采样可以实现频谱的压缩，从而起到降调的作用；下采样能够实现频谱的扩展，从而有升调的作用。对语音信号进行变采样率处理后，在变调的同时也会改变语音信号的播放时间，需要对变调后的信号进行时长规整^[13]，将语音信号规整到原来的时长。

本文使用半音来衡量语音变调的幅度。在乐音体系中^[14]，通常采用十二平均律将一个纯八度分成 12 个均等的半音，相邻两个半音间的物理振动频率相差 $2^{1/12}$ 倍，也就是说各半音间的



振动频率成等比关系。根据十二平均律和标准音,如果将信号中的所有频率成分升高或降低 $2^{1/12}$ 倍,就能使音调升高或降低一个半音。假设原信号频率为 f_0 ,变调后频率为 f ,二者的关系应该满足:

$$f = f_0 \times 2^{a/12} \quad (1)$$

其中, $a = \pm 1, \pm 2, \pm 3, \dots, \pm 11$ 。当 $a > 0$ 时,升调;当 $a < 0$ 时,降调。 a 每变化1,音调就升高或降低一个半音。本文使用 $+k$ 表示将原始语音升高 k 个半调;使用 $-k$ 表示将原始语音降低 k 个半调,并将 $+k$ 和 $-k$ 称为变调系数。

2.2 参考文献[3]的工作

Wu 等人^[3]首先对语音的变调幅度进行了量化,对每段语音共进行了10种变调系数(+4, +5, +6, +7, +8, -4, -5, -6, -7, -8)的变换。然后提取每段语音的梅尔倒谱系数(MFCC)及其一阶差分和二阶差分,并分别提取其统计特征(均值、相关系数)作为最终的输入特征,将其放入一个由多个SVM分类器组成的集成分类器中进行训练,最终输出测试语音是否为变调语音。实验采用了不同的语音数据库进行语音库内和语音库之间的交叉测试,同时采用了不同的语音处理软件进行变声工具之间的交叉测试,最

终的检测平均正确率在90%以上,平均虚警率低于10%,说明所提出的检测算法对电子变调语音的检测有较好的效果。

在上述算法中,使用了一个由10个SVM分类器组成的集成分类器,如图1所示。训练时每个单独的SVM分类器只针对一种变调系数的语音和相对应的原始语音,测试时将待测语音依次通过10个分类器,最后对10个SVM分类器的输出结果进行决策。这种集成分类器很好地综合了每个单独分类器的性能,使得实验的检测率达到很好的效果,并展现出了很好的泛化效果。但是集成分类器过于复杂,训练和测试时都要分别针对每个单独的SVM分类器进行处理操作。并且集成分类器无法对变调语音的伪造操作做出具体的判断,即无法判断出电子变调语音是经过了降调处理还是经过了升调处理。本文针对这种情况使用卷积神经网络作为分类器,在获得较好的正确率和虚警率的同时,使得模型得到一定的简化,并且能够对电子变调语音的具体伪造类型做出进一步的判断。

3 基于CNN的电子变调语音检测算法

目前卷积神经网络已经被成功地应用到语音

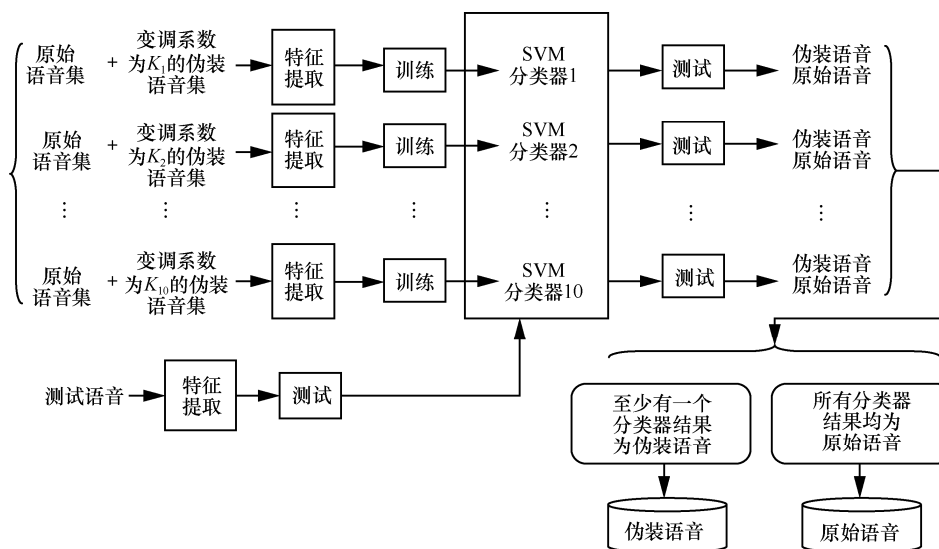


图1 参考文献[3]中的检测算法流程

识别、说话人识别等语音相关的应用中。它是一种多层的有监督学习网络,主要由输入层、卷积层、下采样层、池化层和输出层组成。相比于传统的全连接神经网络(多层感知机等),CNN使用了局部感知野、权值共享等技术有效地降低了网络中的神经元个数,从而使可训练网络参数大规模下降。

本文使用了卷积神经网络来实现分类任务,将分类任务看成一个多分类问题,把原始语音和两种类型的电子变调语音(升调的变调语音和降调的变调语音)组合在一起进行多分类。使用的输入特征与参考文献[3]使用的输入特征相同,但是在输入CNN之前对特征进行了重新组织,使其更加适合CNN。本部分内容先介绍MFCC统计特征的提取过程,再对其进行重新组织,最后给出最终使用的卷积神经网络的具体结构和相关参数。

3.1 MFCC统计特征

本文使用了MFCC及其差分的两种统计特征(均值和相关系数)作为最终的输入特征。MFCC是在Mel标度的频率域内提取出来的倒谱参数。在该提取过程之先要对语音信号进行预处理(包括端点检测、加窗分帧),然后再对数据进行傅里叶变换,并将变换后的数据通过一组Mel滤波器组。再将数据进行对数变换,并进行离散余弦变换(DCT)降维,可得到一组MFCC系数。对得到的MFCC系数进行一阶差分和二阶差分可得到 Δ MFCC和 $\Delta\Delta$ MFCC。因此,每帧语音会由3部分特征组成:MFCC、 Δ MFCC和 $\Delta\Delta$ MFCC。分别计算每部分特征的均值和相关系数,再将它们组合在一起得到最终的输入特征。

假设从每帧语音信号中可以提取出 L 维MFCC。记 v_{ij} 为某段语音第 i 帧MFCC系数的第 j 个分量, V_j 表示所有帧对应的第 j 维MFCC分量的集合,此时 V_j 可以表示为:

$$V_j = \{v_{1j}, v_{2j}, v_{3j}, \dots, v_{Nj}\}, j=1, 2, 3, \dots, L \quad (2)$$

$$m_j = E(V_j), j=1, 2, 3, \dots, L \quad (3)$$

$$c_{jj'} = \frac{\text{cov}(V_j, V_{j'})}{\sqrt{\text{VAR}(V_j)}\sqrt{\text{VAR}(V_{j'})}}, j, j'=1, 2, \dots, L, j \neq j' \quad (4)$$

式(2)中 L 一般为12, N 为此段语音的帧个数。每个特征分量集合 V_j 的均值 m_j ,特征分量集合 V_j 和特征分量集合 $V_{j'}$ 的相关系数 $c_{jj'}$,可分别由式(3)和式(4)计算得出。 m_j 和 $c_{jj'}$ 将会组成MFCC对应的统计特征,可以表示为:

$$W_{\text{MFCC}} = [m_1, m_2, \dots, m_L, c_{12}, c_{13}, \dots, c_{L-1L}] \quad (5)$$

通过同样的方法计算出MFCC一阶差分和二阶差分的统计特征 $W_{\Delta\text{MFCC}}$ 和 $W_{\Delta\Delta\text{MFCC}}$,将 W_{MFCC} 、 $W_{\Delta\text{MFCC}}$ 和 $W_{\Delta\Delta\text{MFCC}}$ 结合起来得到最终的输入特征 W :

$$W = [W_{\text{MFCC}}, W_{\Delta\text{MFCC}}, W_{\Delta\Delta\text{MFCC}}] \quad (6)$$

最终经过特征提取过程,会得到一个由3部分组成的输出特征,这3部分相互独立,在时间上并没有前后关系。本文提取的MFCC的维度为12,也就式(2)~式(5)中的 L 取12。因此,每部分特征均值的维数为12维,相关系数的维数为66维,最终得到特征 W 中每部分特征维数为78维。最后输入特征 W 的总维数为234维。

3.2 卷积神经网络(CNN)

CNN最初主要应用在图像识别中,考虑到图像的空间结构和空间相关性与语音信号有很大的差异,因此需要考虑输入特征的结构。由于上述提取的输出特征 W 主要由3部分组成: W_{MFCC} 、 $W_{\Delta\text{MFCC}}$ 和 $W_{\Delta\Delta\text{MFCC}}$,这3部分在时间上并不是前后关系,而是并列关系。如果直接将最终特征 W 作为一维向量输入CNN中,并使用一维的卷积核进行卷积,就会忽略 W 中每部分特征间的并列关系。因此本文将 W 分成3行,每行代表一部分特征,从上到下依次为 W_{MFCC} 、 $W_{\Delta\text{MFCC}}$ 和 $W_{\Delta\Delta\text{MFCC}}$ 。最终输入CNN的特征为一个 3×78 的矩阵。此过程如图2所示,图2中(a)、(b)、(c)分别代表 W_{MFCC} 、



$W_{\Delta MFCC}$ 和 $W_{\Delta \Delta MFCC}$ 。

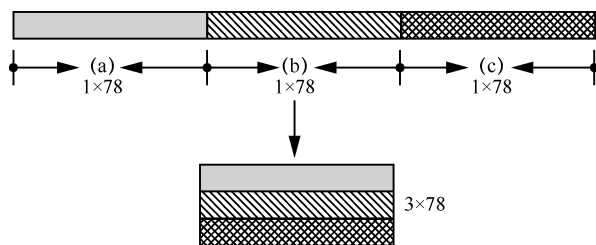


图2 输入特征的结构

LeNet^[15]是一种典型的 CNN 结构,这种最早由 Yan 提出的 7 层网络结构在手写体识别中取得了较好的识别效果。鉴于本文的输入特征尺寸与这种网络结构的尺寸相当,本文在此网络结构的基础上进行了修改,针对不同的网络参数做了优化,该优化过程在第 4 节实验中给出。最终确定了一种 4 层的卷积神经网络作为最终分类器。除去输入层和输出层,主要由卷积层、卷积层、全连接层和 softmax 分类层组成。另外,在第二个卷积层和全连接层后分别加入了一个 dropout 层^[16],用于提高网络泛化能力,减少网络过拟合。网络主要结构如图 3 所示,具体参数见表 1。表 1 中网络参数为从第一层开始主要的可调整参数,每层可能有多个参数。例如,表 1 中第一个参数

32@2×2 表示第一个卷积层的卷积核个数为 32,卷积核形状为 2×2,经过卷积后数据的结构状为 (2, 77, 32);第二个参数表示第一个卷积层的激活函数为 Relu。

表 1 本文使用的 CNN 结构具体参数表

网络参数名称	参数	输出数据结构
卷积	32@2×2	(2, 77, 32)
激活函数	Relu	(2, 77, 32)
卷积	32@2×2	(1, 76, 32)
激活函数	Relu	(1, 76, 32)
dropout	0.25	(1, 76, 32)
全连接	128	128
激活函数	Relu	128
dropout	0.25	128
softmax	-	3
激活函数	Relu	3

3.3 基于 CNN 的电子变调语音检测算法

本文提出的检测算法基于 MFCC 的统计特征和卷积神经网络 (CNN),具体流程由图 4 给出。

在训练阶段,训练集包含 1 组原始语音集和 2 组变调语音集。2 组变调语音分别为升调语音集

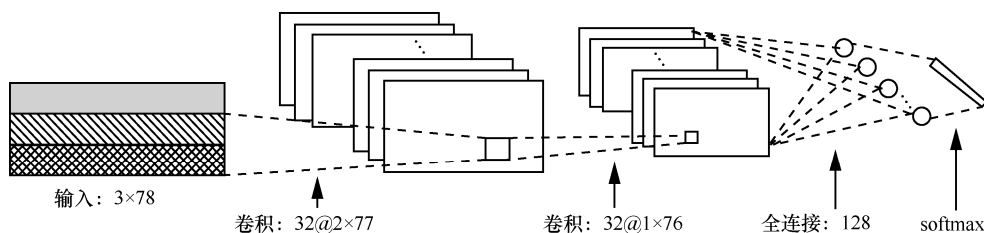


图3 本文提出的 CNN 结构

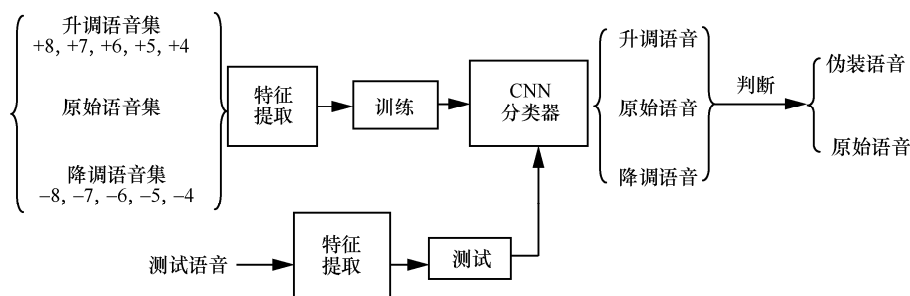


图4 基于 CNN 的电子变调语音检测方法

和降调语音集。其中,升调语音集包含有5种数量相等、升调幅度不同的语音子集,分别对应的变调系数为+4、+5、+6、+7、+8;降调语音集也有5种数量相等,降调幅度不同的语音子集,分别对应的变调系数为-4、-5、-6、-7、-8。因此,训练集中包含3种类型的语音数据,即原始语音、升调语音和降调语音。CNN分类器被用来训练成可分辨上述3类语音数据的多分类器。

在测试阶段,将测试语音的特征向量输入CNN中以后,CNN会将测试语音判定为升调语音、降调语音或原始语音的其中一种。为了与已有的参考文献进行对比,需要得到原始语音和电子变调语音的二分类检测率。因此,需要再对CNN输出的三分类结果进行重新判断。判断策略为:如果CNN输出的结果认为该测试语音为原始语音,就将该测试语音判定为原始语音;如果CNN的输出结果为原始语音以外的其他类别,则将该测试语音判定为变调语音。

4 实验结果

4.1 实验准备

本文使用的语料库是TIMIT语音库。它由630个不同说话人的6300段语音组成。语音的格式为WAV格式,采样率为16 kHz,量化位数为16 bit,声道为单声道。在实验中,语音库被分为两部分:包含500个说话人共5000段语音的训练集TIMIT-1和包含剩下130个说话人共1300段语音的测试集TIMIT-2。

实验中使用3种较为常用的语音处理软件来对语音库进行变调,即:Audacity、CoolEdit和RTISI。对每段语音升5个半调,分别记为+4、+5、+6、+7、+8;同样对每段语音降5个半调,分别记为-4、-5、-6、-7、-8。对于一种变调软件,需要制作包含10种变调系数的语音集,每个变调系数对应6300个语音样本,共63000个语音样

本。这样可得到3种包含63000个语音样本的变调语音库:Audacity变调语音库、CoolEdit变调语音库和RTISI变调语音库。

在训练时,原始语音共有5000个训练样本。为了保持样本数的平衡,需要分别从5种升调语音中随机抽取1000个升调语音样本,共5000个升调的样本。同理可获得5000个降调的样本。因此,训练样本数共有15000个,对应3个类别,每个类别5000个样本。在测试时,原始语音共有1300个测试样本。为了保持样本数的平衡,需要分别从5种升调系数语音中随机抽取260个升调样本,共1300个升调的样本。同理可得到1300个降调的样本。因此,测试样本数共有3900个,对应3个类别,每个类别1300个样本。

由于存在多种语音处理软件和变调方法,犯罪分子所使用的变调方法和训练分类器时使用的变调方法可能并不一样。因此,本实验会进行变调软件之间的交叉验证。即使用原始语音和一种语音变调软件产生的语音去训练分类器,在测试时会同时测试3个不同的语音集。每个测试语音集中的变调语音分别由Audacity、CoolEdit和RTISI产生。交叉验证实验的检测结果将会在下一节中同时给出。

在实验结果部分,对于三分类结果本文会给出三分类的平均检测准确率 A_1 。对于二分类结果本文会给出电子变调语音的检测准确率 A_2 和原始语音的虚警率FAR。假设用 Y_D 和 Y_S 分别表示测试集中电子变调语音的个数和原始语音的个数,用 Y_D' 和 Y_S' 分别表示分类结果中分类正确的电子变调语音的个数和原始语音的个数。那么, A_2 可以表示为 Y_D'/Y_D ;FAR可以表示为 Y_S'/Y_S 。

目前,存在多种开源的深度学习框架,训练深度神经网络的过程也正在进一步被简化。本文采用的卷积神经网络是基于开源框架Keras^[17]的,



并使用 GPU 作为后端的计算工具。

4.2 CNN 网络结构选择实验

由于卷积神经网络存在多种层次结构,并且卷积核尺寸、卷积核个数、池化层尺寸等参数都会对分类效果产生影响,因此需要确定一种相对合理的网络结构。在实验过程中发现网络层次结构的变化会对检测效果产生较大的影响,对于层数固定的 CNN,其卷积核尺寸、卷积核个数,甚至激活函数等参数的修改并没有对检测效果产生较大的影响。因此,先对网络层结构进行调整,即先确定卷积单元的个数以及池化层的个数和位置,再对其余的网络参数(卷积核个数、全连接层神经元个数、卷积核形状)进行调整。另外,最后一层 softmax 分类层,参数较为固定,不需要进行参数调整。因此本部分实验并没有考虑最后一层, CNN 的网络结构表 2 中也并没有给出最后一层。

4.2.1 卷积单元对比实验

首先要需要确定卷积单元的个数。输入特征每经过一个网络层,特征尺寸都会变得更小。由于输入特征的尺寸较小,即每个样本的输入特征尺寸仅为 3×78 ,因此 CNN 的网络层数不可能设置得太深。实验对比了两个卷积单元和一个卷积单元对检测结果的影响,它们的具体结构由表 2 中的网络 a-1 和网络 a-2 给出。网络 a-1 只含有一个卷积单元,即一个卷积层和一个池化层。网络 a-2 中含有两个卷积单元,即两个卷积层和两个池化层。

网络 a-1 和网络 a-2 的检测结果显示在表 3 中给出。表中同时给出了三分类结果 A_1 、二分类结果 A_2 和 FAR。此外,还给出了交叉验证的结果。例如,训练语音变调方法为 Audacity,测试语音变调方法有 3 种: Audacity、CoolEdit 和 RTISI。可以看出在大多数实验结果中,两种网络结构的检测效果相差不大,网络 a-2 有部分结果优于网络 a-1。为此,本文给出了网络 a-1 和网络 a-2 在训

表 2 CNN 的网络结构

网络名称	实验	网络结构
a-1	卷积单元个数	卷积层:32×2×2 池化:1×2 全连接层:128
a-2		卷积层:32×2×2 池化:1×2 卷积层:32×2×2 池化:1×2 全连接层:128
b-1	池化层个数和位置	卷积层:32×2×2 卷积层:32×2×2 池化:1×2 全连接层:128
b-2		卷积层:32×2×2 池化:1×2 卷积层:32×2×2 全连接层:128
b-3		卷积层:32×2×2 卷积层:32×2×2 全连接层:128
c-1	卷积核个数	卷积层:8×2×2 卷积层:8×2×2 全连接层:128
c-2		卷积层:16×2×2 卷积层:16×2×2 全连接层:128
c-3		卷积层:64×2×2 卷积层:64×2×2 全连接层:128
c-4		卷积层:8×2×2 卷积层:16×2×2 全连接层:128
c-5		卷积层:16×2×2 卷积层:32×2×2 全连接层:128
c-6		卷积层:32×2×2 卷积层:64×2×2 全连接层:128
d-1	全连接层神经元数	卷积层:32×2×2 卷积层:32×2×2 全连接层:48
d-2		卷积层:32×2×2 卷积层:32×2×2 全连接层:64
d-3		卷积层:32×2×2 卷积层:32×2×2 全连接层:96
d-4		卷积层:32×2×2 卷积层:32×2×2 全连接层:160
d-5		卷积层:32×2×2 卷积层:32×2×2 全连接层:256
e-1	卷积核形状	卷积层:32×2×2 卷积层:32×1×2 全连接层:128
e-2		卷积层:32×2×2 卷积层:32×1×3 全连接层:128
e-3		卷积层:32×2×2 卷积层:32×2×1 全连接层:128
e-4		卷积层:32×2×1 卷积层:32×2×2 全连接层:128
e-5		卷积层:32×3×1 卷积层:32×1×2 全连接层:128
e-6		卷积层:32×3×1 卷积层:32×1×3 全连接层:128
e-7		卷积层:32×3×1 卷积层:32×1×5 全连接层:128
e-8		卷积层:32×2×1 卷积层:32×1×5 全连接层:128

练过程中 loss 曲线的变化情况,如图 5 所示。从图 5 中可以看出,随着训练次数的增加,网络 a-1 在训练集和测试集上的 loss 相差越来越大,网络 a-2 在训练集和测试集上的 loss 却变化不大。这说明网络 a-1 在训练过程中,随着训练次数的增加,

表3 不同的CNN结构对应的实验结果

网络名称	训练方法	测试方法分类结果 A_1/A_2						FAR
		Audacity		CoolEdit		RTIS		
a-1	Audacity	98.87%	99.35%	98.77%	99.19%	98.67%	98.77%	2.08%
	CoolEdit	94.59%	93.12%	98.92%	99.62%	97.74%	97.38%	2.46%
	RTISI	97.37%	97.20%	98.21%	99.32%	98.97%	98.42%	1.84%
a-2	Audacity	99.02%	99.31%	98.26%	98.15%	97.41%	97.00%	1.54%
	CoolEdit	95.03%	94.12%	98.74%	99.69%	97.89%	97.77%	3.15%
	RTISI	97.23%	97.11%	98.67%	99.07%	98.51%	98.21%	1.92%
b-1	Audacity	99.28%	99.42%	98.92%	98.88%	97.56%	97.30%	1.00%
	CoolEdit	95.46%	94.35%	99.10%	99.81%	97.46%	97.00%	2.31%
	RTISI	97.92%	98.96%	98.20%	98.08%	98.31%	97.54%	1.15%
b-2	Audacity	99.26%	99.38%	98.87%	98.81%	98.64%	98.46%	1.00%
	CoolEdit	96.31%	96.69%	98.85%	99.50%	98.23%	98.07%	2.46%
	RTISI	96.56%	96.04%	97.87%	97.11%	99.36%	99.23%	0.92%
b-3	Audacity	99.54%	99.77%	98.87%	98.77%	98.33%	98.62%	0.92%
	CoolEdit	97.10%	96.46%	99.36%	99.85%	98.84%	98.80%	1.62%
	RTISI	99.23%	99.18%	99.69%	99.57%	99.26%	99.83%	0.85%
c-1	Audacity	99.41%	99.69%	99.00%	99.03%	98.59%	99.58%	1.08%
	CoolEdit	95.05%	93.46%	99.13%	99.58%	98.26%	98.38%	1.77%
	RTISI	97.18%	95.13%	97.51%	96.62%	98.87%	98.73%	1.00%
c-2	Audacity	99.51%	99.77%	98.91%	99.00%	98.92%	98.92%	0.92%
	CoolEdit	96.03%	94.69%	99.41%	99.77%	98.38%	98.15%	1.31%
	RTISI	96.54%	96.15%	97.51%	96.58%	99.18%	99.04%	0.63%
c-3	Audacity	99.51%	99.81%	99.00%	99.04%	98.64%	98.54%	1.08%
	CoolEdit	96.38%	95.12%	99.51%	99.81%	98.36%	98.19%	1.08%
	RTISI	97.54%	96.00%	98.03%	97.27%	99.46%	99.35%	1.31%
c-4	Audacity	99.28%	99.65%	98.90%	99.08%	98.46%	98.27%	1.46%
	CoolEdit	95.67%	94.58%	99.21%	99.88%	98.00%	98.12%	2.15%
	RTISI	96.41%	95.91%	97.87%	97.00%	99.05%	98.89%	0.62%
c-5	Audacity	99.49%	99.58%	98.79%	98.54%	98.59%	98.35%	0.69%
	CoolEdit	96.36%	95.08%	99.44%	99.69%	98.21%	97.69%	1.08%
	RTISI	97.05%	96.77%	99.97%	97.42%	99.36%	99.23%	0.78%
c-6	Audacity	99.49%	99.81%	98.95%	99.00%	98.67%	98.69%	1.15%
	CoolEdit	96.62%	95.64%	99.44%	99.88%	98.41%	98.23%	1.46%
	RTISI	97.59%	96.12%	97.89%	97.23%	99.23%	99.04%	0.78%
d-1	Audacity	99.51%	99.77%	99.13%	99.19%	98.26%	98.27%	1.00%
	CoolEdit	95.21%	93.62%	99.28%	99.73%	98.43%	98.27%	1.62%
	RTISI	96.53%	96.63%	97.87%	97.07%	99.38%	99.42%	1.09%
d-2	Audacity	99.36%	99.62%	98.69%	98.62%	98.48%	98.23%	1.15%
	CoolEdit	96.59%	95.42%	99.49%	99.77%	98.23%	98.04%	1.08%
	RTISI	97.77%	97.26%	97.54%	96.19%	99.13%	99.04%	1.09%
d-3	Audacity	99.38%	99.58%	99.00%	98.96%	98.41%	98.27%	0.92%
	CoolEdit	95.85%	94.62%	99.31%	99.81%	98.10%	97.92%	1.69%
	RTISI	96.67%	97.59%	98.35%	96.58%	99.21%	99.12%	0.61%
d-4	Audacity	99.44%	99.58%	98.90%	98.77%	98.67%	98.38%	0.85%
	CoolEdit	96.36%	95.08%	99.51%	99.81%	98.26%	97.96%	1.08%
	RTISI	97.97%	97.26%	97.77%	96.08%	99.44%	99.35%	0.78%
d-5	Audacity	99.36%	99.81%	98.97%	99.23%	98.69%	98.65%	1.54%
	CoolEdit	96.23%	94.92%	99.49%	99.81%	98.31%	98.19%	1.15%
	RTISI	97.59%	97.72%	97.11%	96.85%	99.15%	98.96%	0.85%



续表

网络名称	训练方法	测试方法分类结果 A_1/A_2						FAR
		Audacity		CoolEdit		RTIS		
e-1	Audacity	99.26%	99.73%	99.00%	99.31%	98.49%	98.27%	1.62%
	CoolEdit	95.46%	93.88%	99.28%	99.62%	98.21%	97.85%	1.38%
	RTISI	97.08%	97.59%	98.54%	96.69%	99.15%	99.00%	0.64%
e-2	Audacity	99.46%	99.62%	98.87%	98.69%	98.41%	98.54%	0.77%
	CoolEdit	95.87%	94.35%	99.43%	99.69%	98.03%	98.46%	1.08%
	RTISI	98.53%	98.28%	98.08%	97.65%	99.15%	98.96%	0.78%
e-3	Audacity	99.44%	99.77%	99.10%	99.27%	98.48%	98.35%	1.23%
	CoolEdit	95.41%	93.96%	99.28%	99.78%	98.28%	98.35%	1.69%
	RTISI	97.89%	97.69%	97.54%	96.81%	99.31%	99.12%	0.89%
e-4	Audacity	99.33%	99.69%	99.05%	99.31%	98.64%	98.50%	1.38%
	CoolEdit	96.59%	95.73%	99.31%	99.81%	98.33%	98.04%	1.69%
	RTISI	97.85%	96.95%	96.65%	95.65%	99.38%	99.35%	0.78%
e-5	Audacity	99.46%	99.77%	99.08%	99.19%	98.56%	98.58%	1.15%
	CoolEdit	96.28%	95.50%	99.18%	99.84%	98.15%	97.73%	2.15%
	RTISI	97.05%	96.38%	97.85%	96.35%	99.31%	99.19%	0.85%
e-6	Audacity	99.41%	99.69%	98.90%	98.88%	98.69%	98.62%	1.08%
	CoolEdit	95.08%	95.08%	99.36%	99.81%	98.38%	98.08%	1.54%
	RTISI	97.15%	97.67%	97.54%	96.81%	99.36%	99.31%	0.78%
e-7	Audacity	99.49%	99.81%	99.05%	99.15%	98.31%	97.85%	1.15%
	CoolEdit	94.97%	93.46%	99.21%	99.81%	99.18%	99.50%	2.00%
	RTISI	97.23%	96.00%	98.62%	97.12%	99.27%	99.04%	0.85%
e-8	Audacity	99.49%	99.65%	99.08%	99.04%	98.28%	98.23%	0.85%
	CoolEdit	96.28%	94.92%	99.46%	99.69%	98.31%	98.23%	1.00%
	RTISI	97.13%	96.46%	97.12%	97.31%	99.33%	99.27%	0.53%

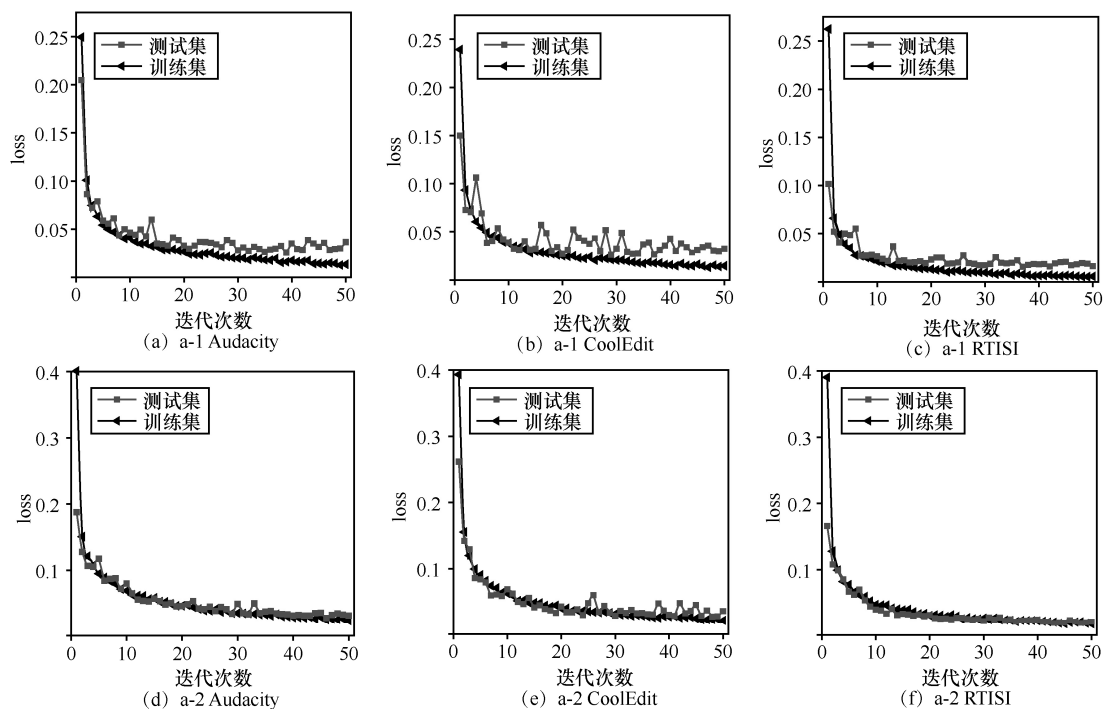


图5 不同卷积单元的 loss 变化曲线

出现了较大过拟合的情况。因此,选定网络 *a-2* 作为接下来实验的基本结构,在此网络的基础上对比池化层对检测结果的影响。

4.2.2 池化层对比实验

在网络 *a-2* 的基础上通过实验对比不同的池化层对检测结果的影响。分别去掉网络 *a-2* 中第一个池化层和第二个池化层得到网络 *b-1* 和网络 *b-2*,同时将两个池化层去掉得到网络 *b-3*。加上网络 *a-2*,得到4种池化层不同、其他参数都相同的卷积神经网络。具体网络结构见表2,对应的检测结果在表3中给出。通过实验结果的对比发现,只有两个卷积层、不包含池化层的网络 *b-3* 的检测结果最好。网络 *b-3* 的检测结果中,不仅虚警率 FAR 明显低于以上几种网络,而且在大多数二分类结果和三分类结果上都超过了以上几种网络。因此,选定网络 *b-3* 作为接下实验的基本结构。现在已经确定了主要的网络层结构:卷积层、卷积层、全连接层。

4.2.3 其他网络参数对比试验

在网络 *b-3* 的基础上,本文给出了6种卷积核个数不同、其他参数都相同的网络结构,它们分别是 *c-1*、*c-2*、*c-3*、*c-4*、*c-5* 和 *c-6*;同样在网络 *b-3* 的基础上,给出了5种全连接层神经元个数不同、其他网络参数都相同的网络结构,它们分别是 *d-1*、*d-2*、*d-3*、*d-4*、*d-5*;在网络 *b-3* 的基础上,给出了8种卷积核形状不同、其他网络参数都相同的网络结构,它们分别是 *e-1*、*e-2*、*e-3*、*e-4*、*e-5*、*e-6*、*e-7*、*e-8*。以此来对比卷积核个数、全连接层的神经元个数以及卷积核形状对检测结

果的影响。以上网络具体结构在表2中给出,它们对应的检测结果在表3中给出。

对比表3中上述网络结构的实验结果,会发现它们的检测效果相差不大。这些检测结果中只有个别实验结果优于网络 *b-3* 的实验结果,其他绝大部分实验结果都没有超过网络 *b-3* 的实验结果。这说明相比于网络层结构,这些网络参数的调整对检测结果的影响相对较小。网络 *b-3* 是可有效用于本文变调检测的卷积神经网络结构。因此,最终选定网络 *b-3* 作为本实验的分类器结构,以网络 *b-3* 给出的实验结果作为本文最终的实验结果。

4.3 算法测试结果和分析

通过本部分的实验最终确定了网络 *b-3* 对应的卷积神经网络作为本文最终的分类器。从网络 *b-3* 对应的实验结果中可以看出,它的三分类的平均准确率 A_1 均达到了96%以上,电子变调语音的检测准确率 A_2 均达到96%以上,原始语音的虚警率 FAR 也都低于2%。为了与已有参考文献进行对比,本文重复了参考文献[3]中的部分实验,得到的实验结果见表4。从表4中可以看出,本文给出的部分二分类结果与参考文献[3]中给出的结果相当,部分检测准确率优于参考文献[3],但原始语音的虚警率都有明显的下降。

从卷积核尺寸、卷积核个数以及池化层尺寸等方面,本文对比了24种不同的网络结构。对于每种网络结构,使用3种不同的变调软件进行了两两之间的交叉测试。即对于每一种网络结构,

表4 与参考文献[3]实验结果对比

分类器类型	训练方法	测试方法/检测准确率 A_2			FAR
		Audacity	CoolEdit	RTISI	
SVM-参考文献[2]	Audacity	99.83%	99.27%	97.75%	1.83%
	CoolEdit	97.29%	99.81%	98.62%	3.35%
	RTISI	98.75%	99.35%	99.80%	1.75%
CNN-本文	Audacity	99.77%	98.77%	98.62%	0.92%
	CoolEdit	96.46%	99.85%	98.80%	1.62%
	RTISI	99.18%	99.57%	99.83%	0.85%



都进行了 9 次交叉测试。从实验结果中可以看出,当使用一种变调软件产生的变调语音和原始语音进行训练,使用其他变调软件产生的变调语音进行测试时,变调语音的检测准确率都达到了 96% 以上。这说明本文的检测算法是有效的,并且具有很好的泛化能力。与此同时,本文算法给出的三分类检测准确率也达到了 96% 以上,这说明本文算法可以有效地判断出待测语音是原始语音、升调语音还是降调语音。与参考文献[3]的算法相比,本文算法使用一种结构较为简单的 CNN 替代了复杂的 SVM 集成分类器,使得检测结果有明显的提升,并可以对变调语音的种类做出进一步的判断。

5 结束语

电子变调语音正在被越来越多的犯罪分子用来伪装自己的身份。本文针对这一问题提出一种基于 MFCC 统计特征和卷积神经网络的检测算法。相比于已有的研究,本算法将复杂的分类器模型简化为一个仅有 4 层网络结构的卷积神经网络,并且可以进一步判断出电子变调语音经过的具体伪造操作。实验结果表明,本算法在有效检测出电子变调语音的同时,可以对电子变调语音进行进一步判断。但是,本文中 CNN 使用的输入特征仍是经过处理的 MFCC 的统计特征,卷积神经网络自学习特征的能力并没有得到充分体现。在未来的工作中,将尝试直接将得到的 MFCC 矩阵输入 CNN 中,让 CNN 自发地去学习 MFCC 矩阵的特征并进行分类。

参考文献:

- [1] RODMAN R. Speaker recognition of disguised voices: A program for research [C]// Consortium on Speech Technology in Conjunction with the Conference on Speaker Recognition by Man and Machine: Directions for Forensic Applications, Oct 8-11, 1998, Ankara, Turkey. [S.l.: s.n.], 1998: 9-22.
- [2] WU H, WANG Y, HUANG J. Blind detection of electronic disguised voice [C]// 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, May 26-31, 2013, Vancouver, Canada. Piscataway: IEEE Press, 2013: 3013-3017.
- [3] WU H, WANG Y, HUANG J. Identification of electronic disguised voices [J]. IEEE Transactions on Information Forensics and Security, 2014, 9(3): 489-500.
- [4] CAO W, WANG H. Identification of Electronic Disguised Voices in the Noisy Environment [C]// International Workshop on Digital-forensics and Watermarking, Sep 17-19, 2016, Beijing China. [S.l.: s.n.], 2016: 75-87.
- [5] ZHENG F, ZHANG G, SONG Z. Comparison of different implementations of MFCC [J]. Journal of Computer science and Technology, 2001, 16(6): 582-589.
- [6] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]// Advances in neural information processing systems, Dec 3-8, 2012, Lake Tahoe, USA. New York: ACM Press, 2012: 1097-1105.
- [7] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv:1409.1556, 2014.
- [8] ROUCOS S, WILGUS A. High quality time-scale modification for speech [C]//ICASSP'85: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Apr 26-29, 1985, Florida, USA. Piscataway: IEEE Press, 1985: 493-496.
- [9] ZHU X, BEAUREGARD G, WYSE L. Real-time signal estimation from modified short-time Fourier transform magnitude spectra [J]. IEEE Transactions on Audio Speech & Language Processing, 2007, 15(5): 1645-1653.
- [10] Time-scale/pitch modification [EB/OL]. (2009-11-24) [2017-09-27]. <http://cn.mathworks.com/matlabcentral/fileexchange/25880-time-scale-pitch-modification>.
- [11] ZHU X, BEAUREGARD G T. Real-time signal estimation from modified short-time Fourier transform magnitude spectra [J]. IEEE Transactions on Audio Speech & Language Processing, 2007, 15(5): 1645-1653.
- [12] TREHUB S E, COHEN A J, THORPE L A. Development of the perception of musical relations: semitone and diatonic structure [J]. Journal of Experimental Psychology Human Perception & Performance, 1986, 12(3): 295.
- [13] Audacity: Free Audio Editor and Recorder[EB/OL]. (2016-01-20)[2017-03-27]. <http://www.audacityteam.org/>.
- [14] Cool Edit Pro is Now Adobe Audition [EB/OL]. (2012-11-08) [2017-03-27]. <http://www.adobe.com/products/audition.html>.
- [15] LECUN Y, BOTTOU L, BENGIO Y. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [16] SRIVASTAVA N, HINTON G E, KRIZHEVSKY A. Dropout: a

simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[17] CHOLLET F. Keras [EB/OL]. (2016-09-16) [2016-11-24]. <https://github.com/fchollet/keras>.

[作者简介]



徐宏伟 (1990-), 男, 宁波大学信息科学与工程学院硕士生, 主要研究方向为多媒体通信与信息安全等。



严迪群 (1979-), 男, 博士, 宁波大学信息科学与工程学院副教授、硕士生导师, 主要研究方向为多媒体通信、信息安全、基于深度学习的数字语音取证等。



阳帆 (1991-), 男, 宁波大学信息科学与工程学院硕士生, 主要研究方向为多媒体通信与信息安全等。



王让定 (1962-), 男, 博士, 宁波大学高等技术研究院教授、博士生导师, 主要研究方向为多媒体通信与取证、信息隐藏与隐写分析、智能抄表及传感网络技术。

金超 (1990-), 男, 宁波大学信息科学与工程学院博士生, 主要研究方向为多媒体通信与信息安全等。

向立 (1994-), 男, 宁波大学信息科学与工程学院硕士生, 主要研究方向为多媒体通信与信息安全等。