基于手工特征提取与结果融合的 CNN 音频隐写分析算法

李敬轩 胡润文 阮观奇 项世军*

(暨南大学信息科学技术学院/网络空间安全学院, 广州 510632)

摘 要 随着互联网技术的快速发展，出现了基于 IP 的语音传输技术，给人们带来方便的同时也带来了许多安全隐患，如不法分子利用 VoIP 压缩域语音传输协议传送秘密信息。因此，针对基于 G.729A 编码的基音隐写算法和互补邻居顶点的量化索引调制音频隐写算法，本文提出了一种基于手工特征提取与结果融合的卷积神经网络音频隐写分析算法。通过将手工提取特征与卷积神经网络相结合，可以实现在 VoIP 压缩域同时对基于基音的隐写算法和互补邻居顶点的量化索引调制音频隐写算法进行有效检测。实验结果表明，在同时对基音隐写算法和互补邻居顶点的量化索引调制音频隐写算法进行检测时，本文所提出的基于手工特征提取与结果融合的卷积神经网络音频隐写分析算法的检测准确率可以达到 86.2%(嵌入率为 100%、音频样本时长为 0.1s)。与现有隐写分析算法相比，在音频时长较短时，本文所提算法取得了优异的检测结果。

关键词 隐写分析；G.729A；卷积神经网络；手工特征提取； 结果融合

中图法分类号 TP309

A CNN Based Audio Steganalysis Algorithm By Manual Feature Extraction and Result Merging

LI Jing-Xuan HU Run-Wen RUAN Guan-Qi XIANG Shi-Jun*

(College of Information Science and Technology/College of Cyber Security, Jinan University, Guangzhou 510632)

Abstract With the rapid development of Internet technology, IP-based voice transmission technology has emerged. While bringing convenience to people, it also brings many security risks. The criminals using VoIP voice transmission protocol in compressed domains to transmit secret information has brought great challenges to social security. In this paper, for the pitch steganography algorithm and the quantized index modulation audio steganography algorithm of complementary neighbor vertex based on G.729A encoding, an audio steganalysis algorithm based on manual feature extraction and convolutional neural network is proposed. By combining manually extracted features with convolutional neural networks, it is possible to achieve effective detection of both the quantized index modulation audio steganography algorithm of complementary neighbor vertex and the pitch-based steganography algorithm in the VoIP compressed domain. Specifically, the algorithm proposed in this paper firstly extracts manual features from the G.729A speech segment (including two manual features extracted by the pitch steganography algorithm and three manual features extracted by the quantized index modulation audio steganography algorithm with complementary neighbor vertex). After using audio steganography algorithm to steganography audio samples, the five extracted manual features have been changed to vary degrees. Therefore, these five manual features can be used as one of the basis for judging whether the audio samples contain secret information. Then, after extracting the five manual features, this paper designs two

本课题得到国家自然科学基金(No.61772234)、广东省科技创新战略专项资金(No.pdjh2020a0060)资助.李敬轩，博士研究生,主要研究领域为信息隐藏、机器学习、隐写分析.E-mail: 2446288716@qq.com. 胡润文，硕士研究生,主要研究领域为多媒体信息安全、可逆鲁棒水印技术.E-mail: 916049256@qq.com. 阮观奇，硕士研究生,主要研究领域为多媒体信息安全、可逆信息隐藏.E-mail: 3092703779@qq.com. 项世军（**通信作者**），博士，教授，计算机学会(CCF)会员（10542S）,主要研究领域为信息隐藏、多媒体信息安全、人工智能安全.E-mail: shijun_xiang@qq.com.

different convolutional neural networks for the pitch steganography algorithm and the quantized index modulation audio steganography algorithm with complementary neighbor vertex. The two extracted manual features for the pitch steganography algorithm and the three manual features for the quantized index modulation audio steganography algorithm based on complementary neighbor vertex are input into the two different convolutional neural networks, respectively. Immediately afterwards, the two convolutional neural networks will further extract and discriminate the input manual features, and obtain the steganalysis results based on the pitch audio steganography algorithm and the quantized index modulation audio steganography algorithm with complementary neighbor vertex, respectively. Finally, according to a designed fusion rule, the network merges the two discriminant results to obtain the final discriminant result, that is, the network discriminates whether the input audio sample contains steganographic information. In summary, the algorithm proposed in this paper extracts features manually from the audio samples encoded by G.729A, and combines the manually extracted features with the convolutional neural network, which can effectively perform steganalysis and detection on the pitch audio steganography algorithm and the quantized index modulation audio steganography algorithm with complementary neighbor vertex in the VoIP compression domain. The experimental results show that in detecting both the pitch steganography algorithm and the quantized index modulation audio steganography algorithm with complementary neighbor vertex at the same time, the detection accuracy rate of the proposed audio steganalysis algorithm based on manual feature extraction and the convolutional neural network proposed in this paper can reach 86.2% (when the embedding rate is 100% and the audio sample duration is 0.1s). Compared with the existing excellent steganalysis algorithms, the algorithm proposed in this paper has achieved state-of-the-art detection results when the audio duration is shorter.

Key words steganalysis; G.729A; convolutional neural network; manual feature extraction; result merging

1 引言

随着计算机网络技术的日新月异，互联网技术早已走入千家万户，给人们的工作和生活带来了极大的便利。但科学技术是一把“双刃剑”，互联网技术在给人们生活带来便利的同时，也出现了一些不法分子利用互联网技术非法窃取用户隐私信息，甚至是国家机密信息的非法行为。因此，为了保护国家安全和用户的个人隐私信息不受侵犯，出现了信息加密技术。人们利用加密技术对明文信息进行加密之后再通过信道发送给接收方，接收方在接收到加密信息后再进行解密，恢复明文信息。但是，在信息的传输过程中，一旦经过加密之后的信息被不法分子截获并破解、解密，那么加密技术就无法再对明文信息起到有效的保护作用，这正是加密技术最大的弱点之一。信息在信道传输的过程中如果能够隐藏通信的行为，该问题即可迎刃而解。因此，作为加密技术的补充形式，出现了信息隐藏技术。信息隐藏技术主要包括隐写术[1-4]、数字水印[5-7]、隐写分析[8-10]等。作为信息隐藏技术的一个重要分支，隐写术是指发送方将需要秘密传递的信息隐藏在多媒体载体文件中，通过公开信道进行传输，接收方在接收到含有秘密信息的多媒体文件后从中提取出秘密信息。隐写术与加密技术最大的不同之处在于，加密技术只能保证在信道传输过程中载体本身的安全，而一旦解密之后，秘密信息无法得到有效的保护。与此同时，加密技术在信道传输的过程中，一旦被不法分子截获，更容易引起不法分子的怀疑，甚至传输的密文信息被不法分子破译。而隐写术则不同，隐写术能够隐藏通信行为，达到隐蔽通信的目的。故即使含有秘密信息的载体在信道传输的过程中被不法分子截获，不法分子一般也不会注意到载体文件中是否含有秘密信息，从而达到隐蔽通信的目的。但是，在军事领域，若敌方通过隐写术传递军事命令，则将会使己方遭受重大损失。因此，针对隐写术，出现了隐写分析技术。隐写分析技术主要是指通

过隐写分析算法判定多媒体文件中是否含有秘密信息，甚至提取出多媒体文件中的秘密信息。

目前，大多数的学者对非压缩域中的语音隐写算法进行了研究，并取得了较为理想的结果。但是，近年来，随着网络技术的快速发展，出现了基于 IP 语音传输(Voice over Internet Protocol, VoIP)的压缩域语音隐写技术，大量基于 IP 语音传输的压缩域语音隐写算法的出现，也促进了针对 VoIP 语音隐写分析技术的发展。针对 VoIP 压缩域隐写技术，世界各国的研究学者提出了许多相应的隐写分析算法，但目前对于 VoIP 压缩域的隐写分析算法，当隐写音频时长较短时，大多数的隐写分析算法无法取得较为理想的检测结果。

Ren 等人[11]通过计算基音延迟的二阶差分特征来计算马尔科夫转移概率矩阵。同时，为了提高隐写分析检测的准确率，Ren 利用校准的方法估计载体图像的二阶差分马尔科夫转移概率矩阵并使用支持向量机作为隐写分析的分类器。实验表明，该算法在嵌入率为30%时能够达到85%以上的检测准确率。Yang 等人[12]结合滑动窗口检测算法和卷积神经网络(Convolutional Neural Networks, CNN)提出了一种基于多通道卷积滑动窗口的实时 VoIP 语音隐写分析算法。由于量化索引调制(Quantization Index Modulation, QIM)隐写算法可以隐藏 VoIP 语音流中的秘密信息，因此，针对 QIM 音频隐写算法，Lin 等人[13]提出了一种有效的在线音频隐写分析算法。利用循环神经网络(Recurrent Neural Network, RNN)寻找 VoIP 语音流中的 4 个强相关性的码字特征，并提出了一种分类模型对提取出的特征进行分类进而分辨出含有隐写信息的语音。Yang 等人[14]结合 CNN 网络和双向长短时记忆循环神经(Bi-directional Long Short-Term Memory, Bi-LSTM)网络，提出了一种基于 QIM 隐写算法的音频隐写分析算法。实验结果表明，在时长为0.1s 的音频文件、嵌入率为20%的情况下，该算法的检测准确率可达到 61%以上。同样是针对 QIM 音频隐写算法，根据语音生成理论和音素分布特征，Li 等人[15]指出线性预测编码滤波器系数分离向量化(Vector Quantization, VQ)码字的相关特征在 QIM 音频隐写前后发生了变化。基于这个发现，Li 等人[15]提出了基于相邻语音帧分割的量化码字相关模型(Quantization Codeword Correlation Network, QCCN)并使用支持向量机进行分类。而文献[16]首先分析了量化索引调制 QIM 隐写算法对 G.729A 码流造成的影响，设计并实现了针对线性预测编码(Linear Predictive Coding, LPC)滤波器量化索引分布特性的量化特征抽取隐写分析算法，实现了针对 QIM 音频隐写的快速有效检测。针对基音隐写算法，文献[17]中作者发现基音隐写算法在隐写之后将会导致压缩语音流中相邻语音帧自适应码书的关联特性发生改变，因此，文献[17]提出了一种基于码书关联网络的基音调制信息隐藏检测算法并与支持向量机相结合构建音频隐写分析检测器。Tian 等人[18]利用脉冲对的特性表征经过自适应多速率(Adaptive Multi-rate, AMR)编码之后的语音特征，将脉冲对的概率分布作为长期分布特征，提取脉冲对的马尔科夫转移概率矩阵作为短期不变特征，并引入自适应增强机制以优化特征集并降低其维度。文献[19]利用长短期记忆(Long Short-Term Memory, LSTM) 网络和卷积神经网络提出了一种新颖的检测 VoIP 压缩域语音的异构并行隐写分析算法，该算法通过提取 G.729 编码中的线谱对(Linear Spectrum Pair, LSP)码字和基音延迟特征，能够同时对文献[20]和文献[21]中的音频隐写算法进行检测。在同时对两种音频隐写算法进行检测时(音频样本时长为 0.1s、隐写嵌入率为 100%的情况下)，该算法的检测准确率可以达到 76.8%；此外，在音频样本时长为 1s、嵌入率为 20%的情况下，该算法的隐写分析检测准确率能够达到 77.6%。

综上所述，以上算法虽然可对基于基音隐写的隐写算法和基于 QIM 隐写算法进行有效的隐写分析检测，但同时针对两种音频隐写算法的检测还未达到理想的结果。因此，针对文献[20]和文献[21]所提出的音频隐写算法，本文提出了一种基于基音隐写和基于 QIM 的音频隐写分析算法。本文通过手工分别提取待测音频基音特征和 QIM 隐写音频的特征，并与 CNN 相结合，可以同时对基音隐写算法和 QIM 隐写算法进行有效的隐写分析检测。实验结果表明，本文所提出的基于手工特征提取与结果融合的 CNN (CNN based on manual feature extraction and result merging， CMFERM)音频隐写分析算法取得了较好的检测结果。该 CNN 隐写分析模型荣获了第一届全国信息隐藏大赛(The 1st Chinese Information Hiding Competition, CIHC2019)音频组第一名。

2 两种隐写算法简要介绍

文献[20]中，作者提出了一种基于码本划分的互补邻居顶点(Complementary Neighbor Vertex, CNV)算法。该算法将码字表示为多维空间中的顶点，码字与码字之间的关系表示为边，而边的长度表示两个码字之间的欧氏距离。作者假设码字的集合为 V，码字 $X=(x_1,x_2,\cdots,x_m)$ 和码字 $Y=(y_1,y_2,\cdots,y_m)$ 满足 $X,Y\in V$，$D(XY)$ 表示边 XY 的长度，$D(XY)$ 的定义如公式(1)所示[20]。

$$D(XY)=(\sum_{i=1}^{m}(x_i-y_i)^2)^{\frac{1}{2}} \tag{1}$$

由公式(1)可知，即任取一个码本中的码字 $x_i\in X$，在码字集合 V 中寻找到距离 x_i 最近的码字 $y_i\in Y$，将 x_i、y_i 分别划分成不同的组表示秘密信息，即完成码本划分。例如，在进行码本划分时，Xiao 等人[20]首先使用经典的图广度优先遍历算法将图划分成若干个子图，然后使用图理论转换成 2 色图并标记“0”和“1”的数量。如图 1 所示[20]，若黑色点和白色点分别代表“0”和“1”，黑色点与白色点之间的连线代表欧氏距离，当嵌入的秘密信息与该码字所代表的秘密信息相同时，则将秘密信息嵌入该码字中；否则，将秘密信息嵌入至该码字最邻近的码字中。在文献[20]中，每一个码字都有一个量化索引与之对应，因此本文将这种使用 CNV 码字划分方法的音频隐写算法称为 CNV-QIM 音频隐写算法。

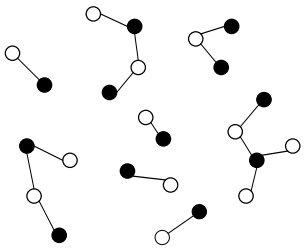

图 1 CNV 码本划分[20]

在文献[21]中，作者提出了一种利用 G.723.1 编码的低比特率 VoIP 语音隐写算法。在该算法中，作者以 G.723.1 编码为例对所提音频隐写算法进行说明。作者利用开环基音延迟和闭环基音延迟嵌入秘密信息 $M=[m_1,m_2,m_3,\cdots]$，假设秘密信息 M 嵌入至音频中第 m 帧中的 $F_i[m]$ 子帧，秘密信息的嵌入算法如下：

若嵌入的秘密信息 $m_i=0$，则第 $F_i[m]$ 子帧中基音周期 $l_i^{'}$ 的搜索范围为 $U_i^{'}(i=0,1,2,3)$[21]

$$\begin{aligned}
&i=0,\\
&l_0^{'}\in U_0^{'}=\begin{cases}\{L_{OLA_0}\},\ if\ \ \mathrm{mod}(L_{OLA_0},2)==0\\ \{L_{OLA_0}-1,L_{OLA_0}+1\},\ if\ \ \mathrm{mod}(L_{OLA_0},2)==1\end{cases}\\
&i=1,\\
&l_1^{'}\in U_1^{'}=\begin{cases}\{L_0,L_0+2\},\ if\ \ \mathrm{mod}(L_0,2)==0\\ \{L_0-1,L_0+1\},\ if\ \ \mathrm{mod}(L_0,2)==1\end{cases}\\
&i=2,\\
&l_2^{'}\in U_2^{'}=\begin{cases}\{L_{OLA_2}\},\ if\ \ \mathrm{mod}(L_{OLA_2},2)==0\\ \{L_{OLA_2}-1,L_{OLA_2}+1\},\ if\ \ \mathrm{mod}(L_{OLA_2},2)==1\end{cases}\\
&i=3,\\
&l_3^{'}\in U_3^{'}=\begin{cases}\{L_2,L_2+2\},\ if\ \ \mathrm{mod}(L_2,2)==0\\ \{L_2-1,L_2+1\},\ if\ \ \mathrm{mod}(L_2,2)==1\end{cases}
\end{aligned} \tag{2}$$

其中，L_{OLA_i} 表示开环基音周期，L_i 表示闭环基音周期。若嵌入的秘密信息 $m_i=1$，则 m 帧中第 $F_i[m]$ 子帧中基音周期 $l_i^{'}$ 的搜索范围为 $U_i^{'}(i=0,1,$[21]

$$\begin{aligned}
&i=0,\\
&l_0^{'}\in U_0^{'}=\begin{cases}\{L_{OLA_0}\},\ if\ \ \mathrm{mod}(L_{OLA_0},2)==1\\ \{L_{OLA_0}-1,L_{OLA_0}+1\},\ if\ \ \mathrm{mod}(L_{OLA_0},2)==0\end{cases}\\
&i=1,\\
&l_1^{'}\in U_1^{'}=\begin{cases}\{L_0,L_0+2\},\ if\ \ \mathrm{mod}(L_0,2)==1\\ \{L_0-1,L_0+1\},\ if\ \ \mathrm{mod}(L_0,2)==0\end{cases}\\
&i=2,\\
&l_2^{'}\in U_2^{'}=\begin{cases}\{L_{OLA_2}\},\ if\ \ \mathrm{mod}(L_{OLA_2},2)==1\\ \{L_{OLA_2}-1,L_{OLA_2}+1\},\ if\ \ \mathrm{mod}(L_{OLA_2},2)==0\end{cases}\\
&i=3,\\
&l_3^{'}\in U_3^{'}=\begin{cases}\{L_2,L_2+2\},\ if\ \ \mathrm{mod}(L_2,2)==1\\ \{L_2-1,L_2+1\},\ if\ \ \mathrm{mod}(L_2,2)==0\end{cases}
\end{aligned} \tag{3}$$

文献[21]中所提出的音频隐写算法可以以较小的失真将秘密信息嵌入经过 G.723.1 编码的音频中，从而达到在音频中隐写秘密信息的目的。

3 隐写分析算法及手工特征提取

3.1 隐写分析算法

文献[21]中作者对经过 G.723.1 编码之后的音频文件进行隐写，而本文针对 VoIP 压缩域音频进行隐写分析检测，利用 G.729A 音频编码对待测音频样本进行编码并利用文献[20]和文献[21]中的音频隐写算法进行隐写，然后提取出经过编码、隐写后音频的 5 个手工特征，将这 5 个手工特征分别送入两个不同的 CNN 网络进行隐写分析检测。

针对文献[20]和文献[21]中提出的音频隐写算法，本文提出了一种基于手工特征提取与结果融合的 CNN 音频隐写分析算法。经过实验证明，与现有的音频隐写分析算法相比，本文所提出的基于手工特征提取与结果融合的 CNN 音频隐写分析算法能够同时对基音隐写算法和 CNV-QIM 隐写算法进行快速有效的隐写分析检测。本文所提算法流程图如图 2 所示。

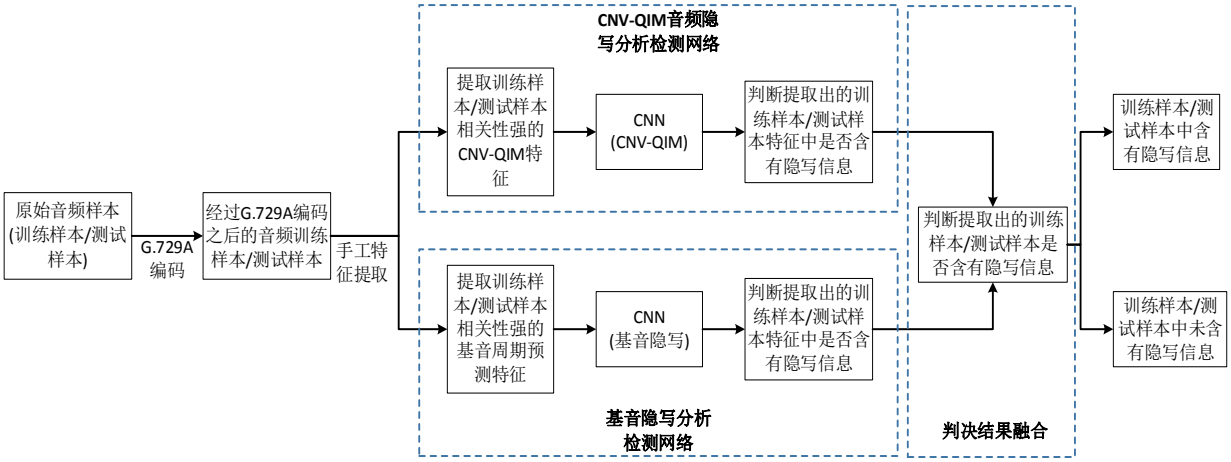

图 2 基于手工特征提取与结果融合的 CNN 音频隐写分析算法流程图

如图 2 所示，首先，使用 G.729A 编码对原始音频样本(训练样本/测试样本)进行编码，在编码的过程中同时对部分音频样本进行隐写，得到经过 G.729A 编码后的音频样本(训练样本/测试样本)；然后，对编码后的音频样本(训练样本/测试样本)进行手工特征提取，提取出编码后音频样本(训练样本/测试样本)每帧中的第一子帧基音延迟(P1_1)、第二子帧基音延迟(P2_1)、量化器第一级向量(L1_1)、LSP 量化器第二阶段低级向量(L2_1)和 LSP 量化器第二阶段高级向量(L3_1)五个手工特征，将提取出的五个手工特征分别送入 CNV-QIM 音频隐写分析检测网络和基音隐写分析检测网络，分别对 CNV-QIM 隐写算法和基音隐写算法进行隐写分析判别(其中，将每帧中量化器第一级向量(L1_1)、LSP 量化器第二阶段低级向量(L2_1)和 LSP 量化器第二阶段高级向量(L3_1)三个手工特征送入 CNV-QIM 音频隐写分析检测网络；将每帧中第一子帧基音延迟(P1_1)、第二子帧基音延迟(P2_1)两个手工特征送入基音隐写分析检测网络)，最后将两个 CNN 网络的隐写分析判别结果按照一定的规则进行融合，得到最终的隐写分析判别结果。本文所提出的基于手工特征提取与结果融合的 CNN 音频隐写分析算法中的 G.729A 编码、手工特征提取、CNV-QIM 音频隐写分析检测网络和基音隐写分析检测网络的详细结构设计以及结果融合规则将分别在 3.2、3.3、3.4 和 3.5 节做详细介绍。

3.2 G.729A编码

如图 2 所示，本文所提出的基于手工特征提取与结果融合的 CNN 音频隐写分析算法首先对原始音频样本(训练样本/测试样本)进行 G.729A 编码。G.729A 编码器的输入为 8kHz、经 16 比特采样量化的音频信号，通过线性预测(Linear Prediction, LP)滤波器得到量化之后的 LP 系数。在得到 LP 系数之后，需要将语音信号转换成在频域内具有更好量化特性的 LSP 系数，然后再对 LSP 系数进行矢量

量化。在矢量量化的过程中，利用滑动平均(Moving Average, MA)法预测当前帧的线谱频率(Linear Spectral Frequency, LSF)系数，并利用二级矢量量化器进行量化得到L1、L2和L3三个码本。

由于在 G.729A 编码中，每帧音频信号可以分为 2 个子帧，对于两个不同的子帧，码字 P1 和 P2 的计算方法略有不同。在第一子帧中码字 P1 的计算如公式(4)所示[22]：

$$P_1=\begin{cases}3(\mathrm{int}(T_1)-19)+frac-1, & T_1=(19,\ldots,85),\ frac=(-1,0,1)\\ \mathrm{int}(T_1-85+197), & T_1=(86,\ldots,143),\ frac=0\end{cases} \quad (4)$$

其中，int 表示整数部分，T_1表示 G.729A 编码中第一子帧基音延迟。与第一子帧中码字 P1 的计算略有不同，在第二子帧中码字 P2 的计算如公式(5)所示[22]：

$$P_2=3(\mathrm{int}(T_2)-t_{\min})+frac+2,\ frac=(-1,0,1) \quad (5)$$

其中，$t_{\min}$的取值与第一子帧有关，T_2表示G.729A编码中第二子帧基音延迟。由公式(4)和公式(5)可以得到 G.729A 编码过程中的码字$P1$和$P2$。

3.3 手工特征提取

为了提高音频隐写分析检测的准确率，在对原始音频样本(训练样本/测试样本)进行 G.729A 编码后，紧接着将进一步提取经过 G.729A 编码后的五个手工特征。本文所提的手工特征是指原始音频样本(训练样本/测试样本)经过 G.729A 编码后提取出每帧中的第一子帧基音延迟(P1_1)、第二子帧基音延迟(P2_1)、量化器第一级向量(L1_1)、LSP 量化器第二阶段低级向量(L2_1)和 LSP 量化器第二阶段高级向量(L3_1)五个码字特征。本文主要针对文献[20]和文献[21]中的音频隐写算法同时进行隐写分析，现有研究成果表明，针对文献[20]中的音频隐写分析算法目前来说已相对成熟，而对于文献[21]中提出的音频隐写算法，在进行隐写分析时有一定难度，主要原因是根据人体的发声原理，人体的声道具有易变性。不同个体的声道特征不尽相同，即使是同一个体在不同时间、不同状态下发出的声音的基音周期也不尽相同。例如同一个体在没有生病的情况下和重感冒的情况下发出声音的基音周期是截然不同的。此外，对同一个体而言，音调的高低也会对基音周期产生影响。例如同一个体正常说话时的声音和故意提高音调时发出的声音，两者的基音周期也不尽相同。因此，为了提高音频隐写分析检测的准确率，本文分别针对两种不同的音频隐写算法提取不同的手工特征输入至 CNN 网络。针对文献[21]，本文利用卷积神经网络提取特征的同时，另外增加手工提取的 2 个经 G.729A 标准编码的语音特征(每段音频经 G.729A 标准编码后的第一子帧基音延迟(P1_1)、第二子帧基音延迟(P2_1))，同时将手工提取的 2 个特征输入至 CNN 网络做进一步处理，以提高隐写分析算法检测的准确率。对于文献[20]提出的隐写算法，本文同样先手工提取出 3 个语音特征(每段音频经 G.729A 标准编码中量化器第一级向量(L1_1)、LSP 量化器第二阶段低级向量(L2_1)和 LSP 量化器第二阶段高级向量(L3_1))，然后再将手工提取的 3 个语音特征输入至 CNN 网络。如公式(6)所示，输入至神经网络的数据为经手工提取的特征矩阵 A，大小为$n\times 5$：

$$A=\begin{bmatrix}a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} & a_{1,5}\\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} & a_{2,5}\\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} & a_{3,5}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ a_{n,1} & a_{n,2} & a_{n,3} & a_{n,4} & a_{n,5}\end{bmatrix} \quad (6)$$

其中，$a_{ij}(1\le i\le n, 1\le j\le 5)$为手工提取的 5 个特征(分别为 G.729A 标准编码中的第一子帧基音延迟(P1_1)、第二子帧基音延迟(P2_1)、量化器第一级向量(L1_1)、LSP 量化器第二阶段低级向量(L2_1)和 LSP 量化器第二阶段高级向量(L3_1))。在对所有音频样本(训练样本/测试样本)进行手工提取特征后，将手工提取出的特征矩阵输入至 CNN 进行进一步的处理。

如图 3-图 7 所示，以时长为 1s 的音频为例，随机选取一段经过 G.729A 编码之后的隐写语音片段，与原始语音片段相应帧中的码字进行对比，分别提取出每帧中的第一子帧基音延迟(P1_1)、第二子帧基音延迟(P2_1)、量化器第一级向量(L1_1)、LSP 量化器第二阶段低级向量(L2_1)和 LSP 量化器第二阶段高级向量(L3_1)五个码字特征在隐写前后的变化。由图 3-图 7 可以看出，这 5 个码字特征均发生了不同程度的变化，尤其是基音隐写算法中的第二子帧基音延迟(P2_1)、CNV-QIM 隐写算法中量化器第一级向量(L1_1)、LSP 量化器第二阶段低级向量(L2_1)以及 LSP 量化器第二阶段高级向量(L3_1)码字在隐写前后均发生了明显变化，因此经过 G.729A 编码之后语音片段一帧中的第一子帧基音延迟(P1_1)、第二子帧基音延迟(P2_1)、量化器第一级向量(L1_1)、LSP 量化

器第二阶段低级向量(L2_1)和 LSP 量化器第二阶段高级向量(L3_1)这五个码字特征可以作为判断音频样本是否含有隐写信息的主要依据之一。故本文所提算法将音频样本中每帧中的第一子帧基音延迟(P1_1)、第二子帧基音延迟(P2_1)、量化器第一级向量(L1_1)、LSP 量化器第二阶段低级向量(L2_1)和 LSP 量化器第二阶段高级向量(L3_1)五个码字特征作为手工提取特征输入至 CNN 网络。

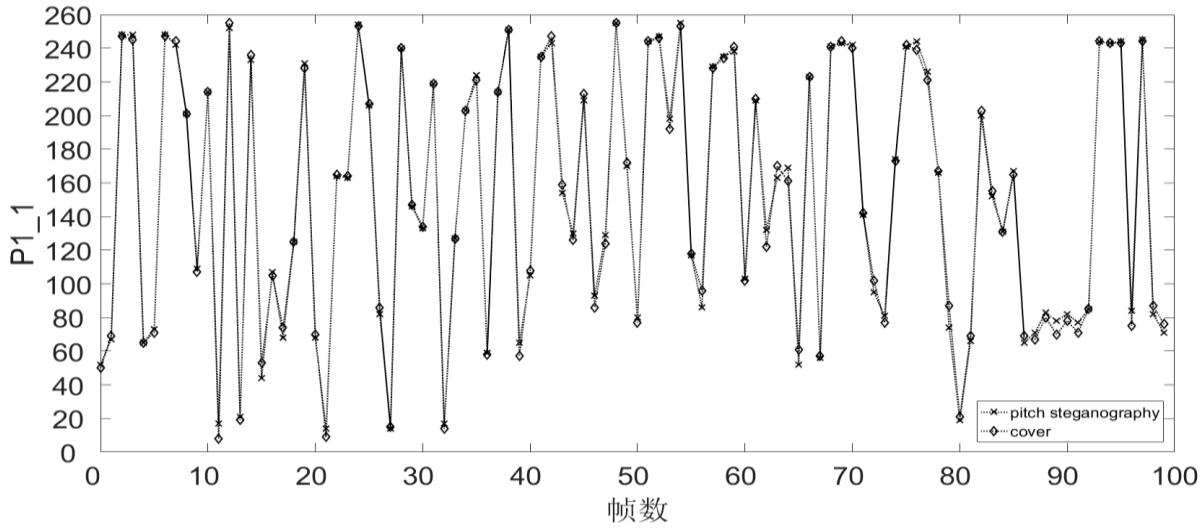

图 3 第一子帧基音延迟(P1_1)

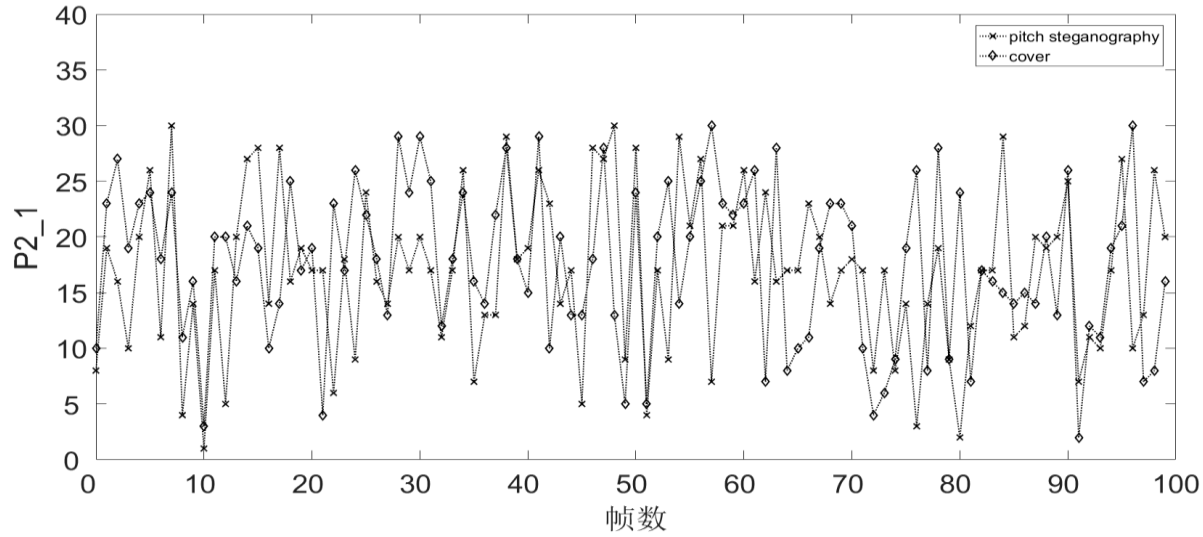

图 4 第二子帧基音延迟(P2_1)

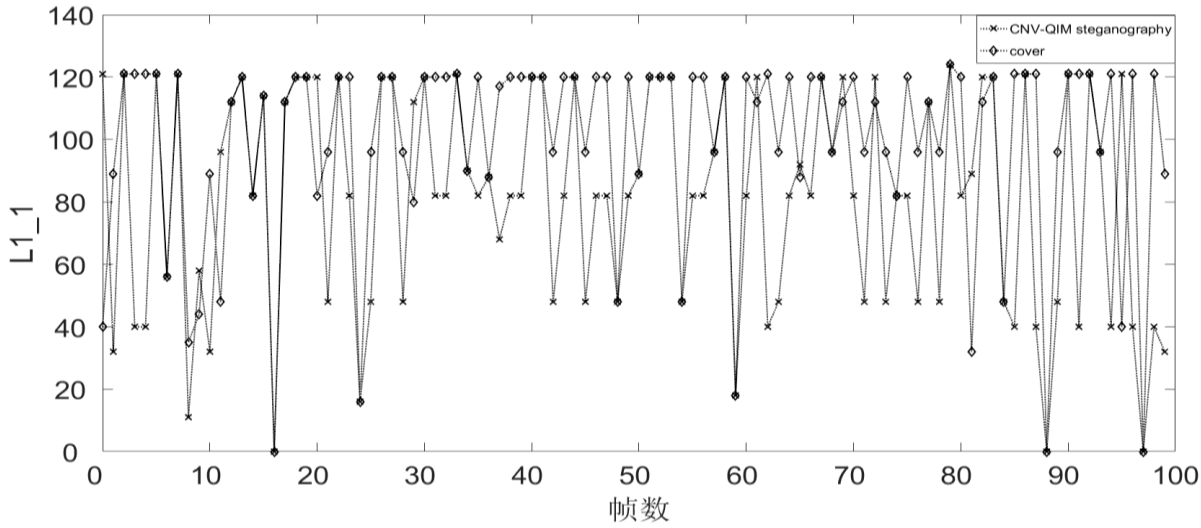

图 5 量化器第一级向量(L1_1)

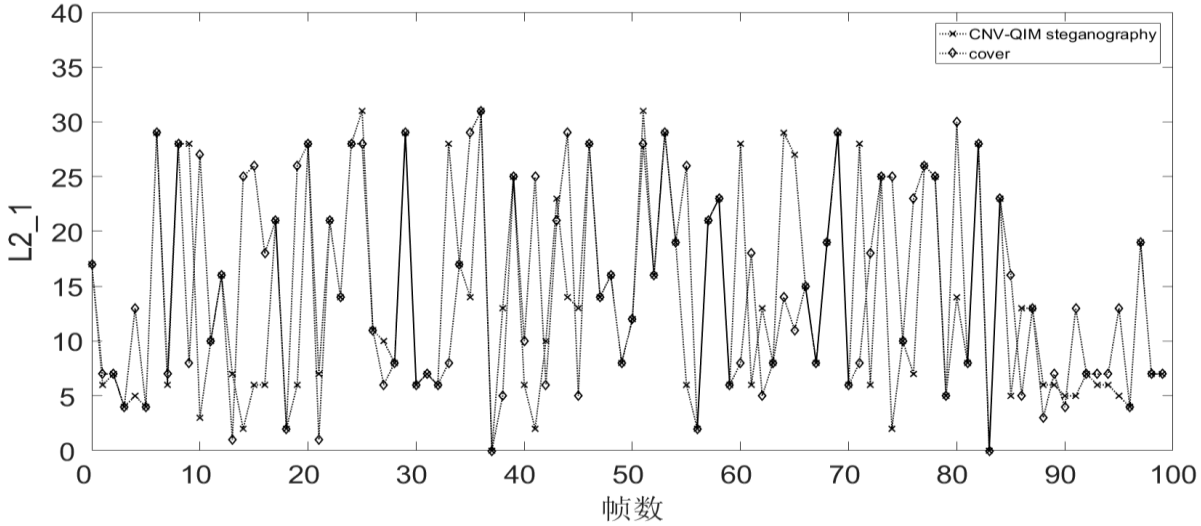

图 6 LSP 量化器第二阶段低级向量(L2_1)

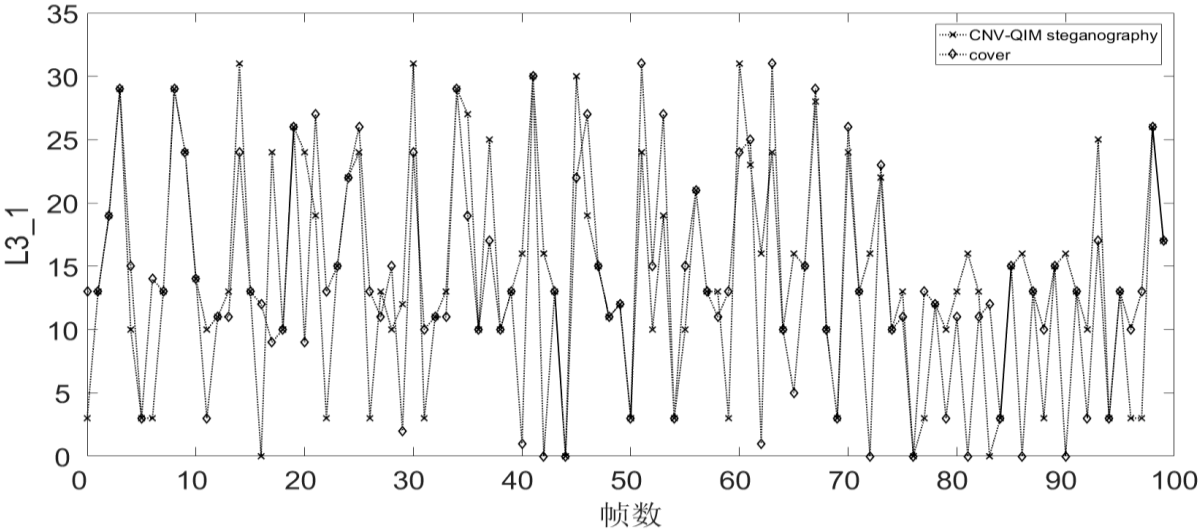

图 7 LSP 量化器第二阶段高级向量(L3_1)

3.4 CNN网络结构的设计

由前述可知，在原始音频样本(训练样本/测试样本)经过 G.729A 编码后，对经过 G.729A 编码的音频样本(训练样本/测试样本)进行手工特征提取，分别提取出每帧音频样本(训练样本/测试样本)中的第一子帧基音延迟(P1_1)、第二子帧基音延迟(P2_1)、量化器第一级向量(L1_1)、LSP 量化器第二阶段低级向量(L2_1)和 LSP 量化器第二阶段高级向量(L3_1)五个码字特征，将提取出的五个手工特征分别送入两个不同的 CNN 网络。针对文献[20]和文献[21]提出的两种不同的音频隐写算法，本文分别设计了两个不同的 CNN 网络。两个不同的 CNN 网络先对输入的手工特征分别进行判断是否含有隐写信息，然后将两个 CNN 网络的判别结果进行融合得到最终的判别结果，从而实现同时对文献[20]和文献[21]中的隐写算法进行隐写分析检测。其中，两个 CNN 网络的输入均为经过 G.729A 编码后手工提取的音频特征(针对 CNV-QIM 隐写算法的隐写分析网络的输入为经过 G.729A 编码后待测音频样本的 CNV-QIM 特征：每帧中的量化器第一级向量(L1_1)、LSP 量化器第二阶段低级向量(L2_1)和 LSP 量化器第二阶段高级向量(L3_1)；针对基音隐写算法的隐写分析网络的输入是经过 G.729A 编码后待测样本的基音特征：每帧中的第一子帧基音延迟(P1_1)、第二子帧基音延迟(P2_1))。本文所提算法两个不同的 CNN 网络结构如图 8 所示。

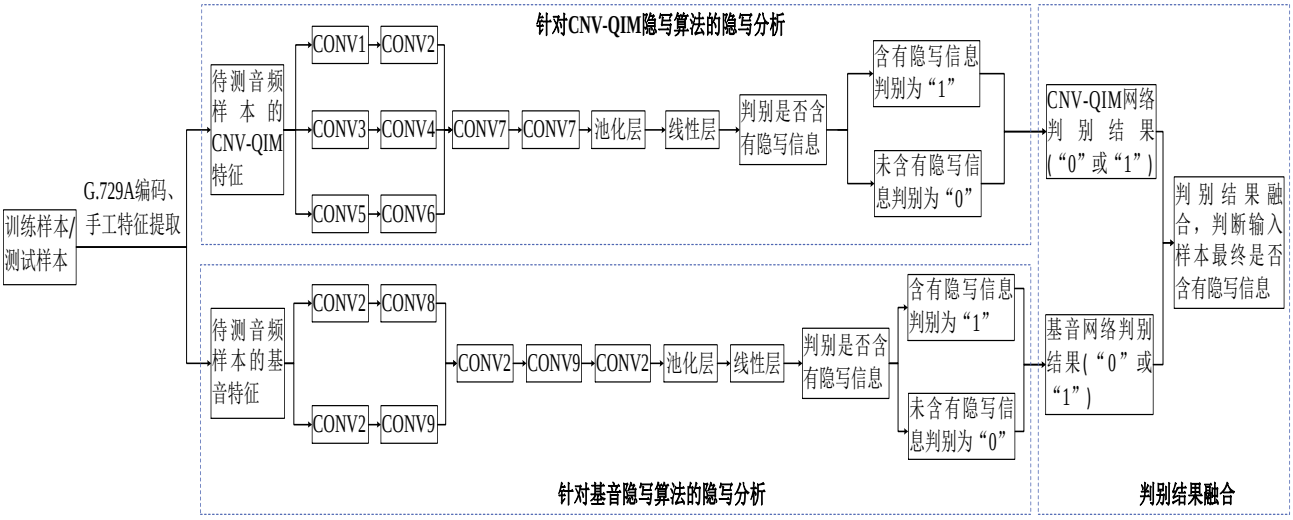

图 8 CNN 网络结构

其中，CONV 为卷积模块，每个卷积模块中均包含一个卷积核、批标准化操作(Batch Normalization)和激活函数(ReLu 函数)。ReLu 激活函数如公式(7)所示。

$$y = \mathrm{Re}Lu(x) = \max(0, x) \tag{7}$$

在本文所提算法中，池化层均采用自适应平均池化的方式进行池化。为了能够更加清楚的说明本文算法，下面将本文所提出的 CNN 网络模型分为针对 CNV-QIM 音频隐写分析算法和针对基音隐写分析算法两部分进行详细说明。首先，针对 CNV-QIM 音频隐写分析算法的详细 CNN 网络结构如图 9 所示。

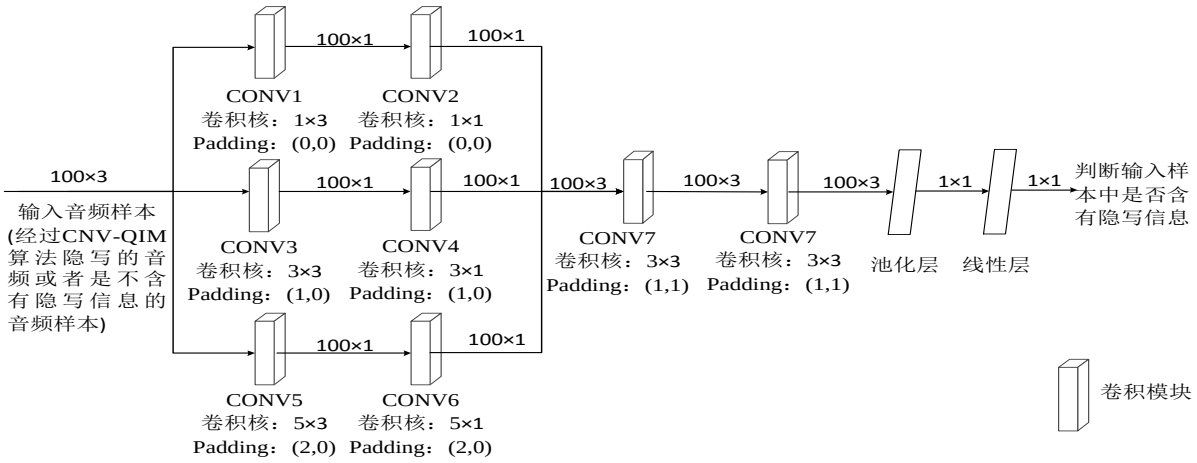

图 9 CNV-QIM 音频隐写分析算法(以时长为 1s 的音频为例)

由图 9 所示，以时长为 1s 的音频为例，对于经过 G.729A 编码后的音频样本，本文首先利用三组不同的卷积核(CONV1-CONV6)对其进行特征提取。输入的音频样本大小为 100×3，经过 CONV1-CONV6 后，三组不同的卷积核分别输出一个 100×1 的特征向量，可以表示为：

$$C1_{CNV}(i)=F_{CNV1}(L_Q,\theta_{Q1}(i)),\quad (i=1,2,3) \tag{8}$$

其中，$C1_{CNV}(i)(i=1,2,3)$表示卷积模块 CONV2、CONV4 和 CONV6 的特征输出结果，$F_{CNV1}(\square)$表示非线性函数，L_Q表示经 G.729A 编码后手工提取的 3 个待测音频样本码字特征(量化器第一级向量(L1_1)、LSP 量化器第二阶段低级向量(L2_1)和 LSP 量化器第二阶段高级向量(L3_1))，$\theta_{Q1}(i)$表示 CNN 网络参数。在本文算法中，输入的待测音频特征经过 CONV2、CONV4 和 CONV6 卷积模块卷积后分别输出一个 100×1 的特征向量，然后，将 CONV2、CONV4 和 CONV6 三个卷积模块输出的 100×1 的特征向量拼接成为一个 100×3 的特征矩阵 $C2_{CNV}$ 并再次使用两个 CONV7 卷积模块对其进行进一步的特征提取，利用非线性函数可以表示为：

$$C3_{CNV}=F_{CNV2}(C2_{CNV},\theta_{Q2}) \tag{9}$$

其中，$C3_{CNV}$表示经过两个 CONV7 卷积模块之后的特征输出，$F_{CNV2}(\square)$表示非线性函数，θ_{Q2}表示 CNN 网络参数。在经过两个 CONV7 后，CNN 网络输出一个 100×3 的特征矩阵，由于需要判别待测音频样本中是否含有隐写信息，因此需对 $C3_{CNV}$ 进行池化，将其池化成一个 1×1 的向量并输入线性层对待测音频进行判别。池化的过程可以表示为：

$$Out_{CNV-QIM}=P_{CNV}(C3_{CNV},\theta_{Q3}) \tag{10}$$

其中，$Out_{CNV-QIM}$表示特征矩阵经过自适应平均池化(AdaptiveMaxPool)层后的输出向量，大小为 1×1，$P_{CNV}(\square)$为池化函数，θ_{Q3}表示自适应平均池化层参数。在经过自适应平均池化层后，将输出的向量送入线性层，由线性层的输出结果判别待测音频样本中是否含有隐写信息，并输出判别结果(若待测样本中含有隐写信息，则输出“1”；否则，输出“0”)，从而完成对 CNV-QIM 音频隐写算法的隐写分析检测。

另一方面，针对经过 G.729A 标准编码的基音隐写算法，本文提出了另外一种基于 CNN 的网络结构进行隐写分析检测。针对经过 G.729A 标准编码的基音隐写算法的隐写分析检测算法 CNN 网络框图如图 10 所示。

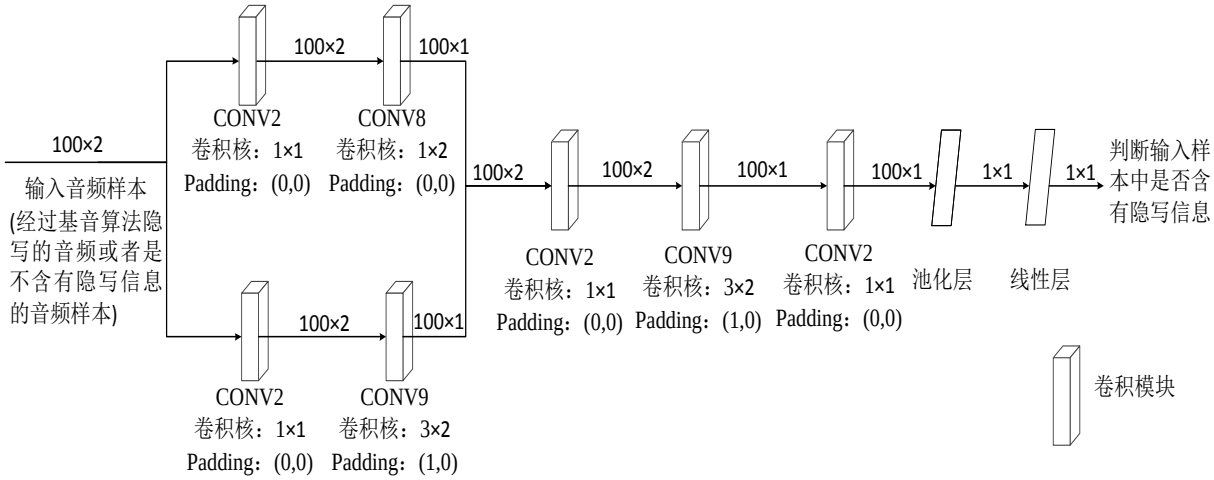

图 10 基音音频隐写分析算法(以时长为 1s 的音频为例)

由图 10 可知，输入基音音频隐写分析网络的音频样本为手工提取的 2 个经 G.729A 标准编码的语音特征(每段音频经 G.729A 标准编码中的第一子帧基音延迟(P1_1)、第二子帧基音延迟(P2_1))。同样以时长为 1s 的音频为例，对于经过 G.729A 编码后的音频样本，本文首先利用两组不同的卷积模块 (CONV2、CONV8 和 CONV2、CONV9)对其进行特征提取。在检测基音音频隐写算法时，输入的音频样本大小为 100×2(当音频时长为 1s 时，每段音频经过 G.729A 标准编码之后含有 100 帧，基音音频隐写分析网络的输入为手工提取的 2 个经 G.729A 标准编码的语音特征(每段音频经 G.729A 标准编码后的第一子帧基音延迟(P1_1)和第二子帧基音延迟(P2_1)，因此基音隐写分析网络中每段音频样本的输入大小为 100×2)，经过 CONV2、CONV8 和 CONV2、CONV9 卷积模块后，两组不同的卷积网络分别输出一个 100×1 的特征向量，可以表示为：

$$P1_{Pitch}(i)=F_{Pitch1}(L_P,\theta_{P1}(i)),\quad (i=1,2) \tag{11}$$

其中，$P1_{Pitch}(i)$ $(i=1,2)$ 表示卷积核 CONV2、CONV8 和 CONV2、CONV9 的特征输出结果，$F_{Pitch1}(\cdot)$ 表示非线性函数，L_P 表示手工提取的待测音频样本特征，$\theta_{P1}(i)$ 表示CNN网络参数。在本文算法中，输入的待测音频特征经过 CONV2、CONV8 和 CONV2、CONV9 卷积后分别输出一个 100×1 的特征向量，然后，将两个输出的 100×1 的特征向量拼接成为一个 100×2 的特征矩阵 $P2_{Pitch}$ 并依次使用 CONV2、CONV9、CONV2 卷积模块对其进行进一步的特征提取，利用非线性函数可以表示为：

$$P3_{Pitch}=F_{Pitch2}(P2_{Pitch},\theta_{P2}) \tag{12}$$

其中，$P3_{Pitch}$ 表示经过 CONV2、CONV9、CONV2 卷积之后的特征输出，$F_{Pitch2}(\cdot)$ 表示非线性函数，θ_{P2} 表示 CNN 网络参数。在依次经过 CONV2、CONV9、CONV2 卷积模块后，CNN 网络输出一个 100×1 的特征向量，由于需要判别待测音频样本中是否含有隐写信息，因此需对 $P3_{Pitch}$ 进行池化，将其池化成一个 1×1 的向量。池化层同样采用自适应平均池化的方式对 $P3_{Pitch}$ 进行池化，池化的过程可以表示为：

$$Out_{Pitch}=P_{Pitch}(P3_{Pitch},\theta_{P3}) \tag{13}$$

其中，Out_{Pitch} 表示特征向量经过自适应平均池化层后的输出向量，大小为 1×1，$P_{Pitch}(\cdot)$ 为池化函数，θ_{P3} 表示自适应平均池化层参数。在经过自适应平均池化层后，将输出的向量送入线性层，由线性层的输出结果判别待测音频样本中是否含有隐写信息，并输出判别结果(若待测样本中含有隐写信息，则输出“1”；否则，输出“0”)，从而完成对基音音频隐写分析的检测。

此外，在卷积神经网络的训练阶段，为了防止卷积神经网络出现过拟合现象，本文所提出的算法中使用了 Dropout 算法。本文中，网络的损失函数均采用交叉熵损失函数(Cross Entropy Error Function)，如式(14)所示。

$$L=\frac{1}{N}\sum_i -[y_i\cdot\log(p_i)+(1-y_i)\cdot\log(1-p_i)] \tag{14}$$

式(14)中，y_i 表示样本标签值，p_i 表示样本 i 预测为“1”的概率。在卷积神经网络的训练过程中，本文将 epoch 设置为1000，即对样本进行1000次迭代计算，以更好的调整网络参数，提高算法

的检测准确性。

3.5 CNN网络判别结果融合

由3.4节CNN网络结构的设计可知，将提取出的手工特征分别输入至CNV-QIM音频隐写分析检测网络和基音隐写分析检测网络，经过两个网络的线性层后，两个网络的线性层均分别输出一个1×1的判别结果，但是，众所周知，对于同一个输入的音频样本而言，只能出现一种判别结果，即音频样本中不含有隐写信息或者音频样本中含有隐写信息。因此，为了判别输入的音频样本中是否含有隐写信息，需要对CNV-QIM音频隐写分析检测网络和基音隐写分析检测网络线性层的输出结果进行融合判别，融合判别具体规则如下：若两个网络均判定输入的待测样本中未包含隐写信息，即CNV-QIM音频隐写分析检测网络和基音隐写分析检测网络中线性层的输出均为“0”，则最终判定该待测样本中不包含隐写信息(即最终判定该样本为“0”)；否则，判定该样本中含有隐写信息(即最终判定该样本为“1”)。

4 实验结果与分析

针对文献[20]和文献[21]中的音频隐写算法，本文算法采用G.729A标准编码器进行音频编码并分别嵌入隐写信息。与现有音频隐写分析算法相比，本文所提算法在检测较短时长的音频时取得了优异的检测结果。在同时对两种隐写算法进行检测时，本文所提出的基于手工特征提取与结果融合的CNN音频隐写分析算法的检测准确率可以达到86.2%(嵌入率为100%、音频样本时长为0.1s)，比文献[19]中的算法检测准确率高出9.4%(嵌入率为100%、音频样本时长为0.1s)。

4.1 数据集

本文所提算法是基于CNN的音频隐写分析算法，所采用的数据集包括训练集和测试集。数据集采用文献[19]提供的包含41小时的中文语音数据集。其中，训练集随机选取了文献[19]数据集中80%经过G.729A标准编码后的语音片段，测试集随机选取文献[19]数据集中20%经过G.729A标准编码后的语音片段。同时，本文在实验测试时，分别测试了在时长为1s、不同嵌入率下的检测准确率和在100%嵌入率、不同时长下的检测准确率，并对实验结果进行了对比分析。

4.2 实验结果分析及对比

本文实验测试结果在Intel(R) Core(TM) i9-9820X CPU @ 3.30GHz 3.31GHz、RAM为32.0G和Intel(R) Core(TM) i9-9900K CPU @ 3.60GHz 3.60GHz、RAM为32.0G的电脑上取得。本文所提算法采用G.729A编码，首先对待测音频进行手工特征提取(第一子帧基音延迟(P1_1)、第二子帧基音延迟(P2_1)、量化器第一级向量(L1_1)、LSP量化器第二阶段低级向量(L2_1)和LSP量化器第二阶段高级向量(L3_1))，然后将手工提取出的待测音频特征输入至CNN进行特征的进一步提取与判别。在CNN中，本文设置批处理量batchsize为800，在100%嵌入率、不同时长下对三种不同隐写算法的音频隐写分析检测准确率和在时长为1s、不同嵌入率下对三种不同隐写算法的音频隐写分析检测准确率分别如表1和表2所示。

由表1可以看出，在嵌入率均为100%的情况下，待测音频的不同时长对隐写分析检测的准确性有一定影响。由表1可知，在同时对CNV-QIM隐写算法和基音隐写算法进行检测时，本文所提出的基于手工特征提取与结果融合的CNN音频隐写分析算法的检测准确率可以达到86.2%(嵌入率为100%、音频样本时长为0.1s)，比文献[19]中的算法检测准确率高出9.4%。此外，本文所提算法在同时对两种音频隐写算法进行检测时，在嵌入率为100%、待测音频时长为2s的情况下，本文所提算法的检测准确率最高，可以达到96.6%。由表1可以看出，与其他音频隐写分析算法相比，本文所提算法在同时对两种隐写算法进行检测且音频样本时长大于0.1s时具有较高的检测准确率。由于本文算法在测试时采用与文献[19]相同的数据集，因此本文参考文献[19]得到表1和表2中SS-QCCN[15]、CCN[17]、Ensemble(SS-QCCN+CCN)[15]+[17]、RNN-SM[13]的实验结果。当同时对CNV-QIM隐写算法和基音隐写算法进行检测时，在嵌入率为100%、待测音频样本时长为0.1s的情况下，本文算法的检测准确率相较于SS-QCCN[15]、CCN[17]、Ensemble(SS-QCCN+CCN)[15]+[17]、RNN-SM[13]、SFFN[19]算法分别高出45.9%、45.2%、22.2%、17.7%、9.4%。此外，通常情况下，语音序列具有较强的时序性，使用循环神经网络(Recurrent Neural Network, RNN)处理时序序列具有较好的实验结果，但本文使用CNN网络同时对基于CNV-QIM音频隐写算法和基音隐写算法进行隐写分析检测仍

然取得了令人满意的实验结果，主要原因在于本文所提取出的 5 个手工特征。对同一个待测音频样本而言，提取出的 5 个手工特征在音频文件隐写前后均发生了不同程度的改变，正是因为这 5 个特征在隐写前后发生的改变，从而使 CNN 网络能够更为准确地对待测音频样本进行分析、判别，提高了 CNN 网络隐写分析的判别准确率。因此，本文所提算法优于 SS-QCCN[15]、CCN[17]、Ensemble(SS-QCCN+CCN)[15]+[17]、RNN-SM[13]、SFFN[19]算法。

表 1 嵌入率 100%时待测音频不同时长下对三种不同隐写算法(CNV[20]、Pitch[21]和 C-P[20]+[21]音频隐写算法)的音频隐写分析检测准确率

隐写分析算法	隐写算法	音频长度										
		0.1s	0.2s	0.3s	0.4s	0.5s	0.6s	0.7s	0.8s	0.9s	1s	2s
SS-QCCN[15]	CNV	68.5	86.7	92.4	93.5	95.3	97.6	97.6	97.9	98.8	99.0	99.9
	Pitch	8.7	8.9	6.1	5.0	3.8	3.8	2.9	2.3	1.9	0.8	0.1
	C-P	40.3	51.3	53.3	53.6	54.8	55.4	53.8	52.3	52.1	53.8	47.1
CCN[17]	CNV	27.1	29.8	26.1	25.4	23.7	19.3	19.1	15.0	14.4	14.1	8.2
	Pitch	56.1	74.4	79.6	84.4	87.5	91.6	93.1	91.7	94.1	94.7	99.1
	C-P	41.0	49.2	52.9	54.0	54.2	55.6	59.3	52.5	57.5	57.0	57.5
Ensemble (SS-QCCN+CCN) [15]+[17]	CNV	75.5	90.0	93.7	94.8	96.2	98.4	98.4	98.4	98.9	99.0	100.0
	Pitch	60.2	76.9	81.0	85.2	88.0	91.8	93.3	91.8	94.1	94.7	99.1
	C-P	64.0	74.6	80.2	78.7	79.4	81.6	81.7	77.2	80.6	81.8	78.0
RNN-SM[13]	CNV	88.2	93.6	96.5	96.2	96.4	98.7	98.7	98.3	99.1	98.8	99.6
	Pitch	48.9	61.7	59.8	72.8	73.2	73.1	77.0	78.1	72.8	71.1	81.3
	C-P	68.5	73.1	76.9	77.2	79.2	78.1	82.5	80.5	80.3	76.5	78.0
SFFN[19]	CNV	92.4	95.4	96.0	96.8	98.3	98.1	98.8	98.5	99.5	99.3	99.8
	Pitch	66.0	77.7	80.8	84.9	88.3	92.9	92.5	92.9	94.8	94.6	98.1
	C-P	76.8	78.5	78.3	75.1	83.7	83.4	83.7	79.4	83.4	85.8	84.6
CMFERM (Ours)	CNV	93.9	96.7	97.5	98.0	98.5	98.0	98.0	98.5	98.6	98.4	99.6
	Pitch	84.5	88.1	87.9	89.5	90.1	92.0	92.9	93.4	93.8	94.7	95.5
	C-P	**86.2**	**88.7**	**90.6**	**92.0**	**91.8**	**93.2**	**94.0**	**94.4**	**95.0**	**95.6**	**96.6**

如表 1 所示，C-P 表示同时对文献[20]中 CNV-QIM 隐写算法和文献[21]中基音音频隐写算法进行检测，CNV 表示仅仅对文献[20]中的 CNV-QIM 隐写算法进行检测，Pitch 表示仅仅对文献[21]中的基音隐写算法进行检测。表 2 测试了在音频时长均为 1 s、不同嵌入率下的隐写分析检测准确率。由表 2 可以看出，在待测音频样本时长均为 1s 的情况下，不同的隐写嵌入率对音频隐写分析检测算法的检测准确率也会产生影响。由表 2 可知，与文献[19]所提算法相比，本文所提算法在同时对 CNV-QIM 隐写算法和基音隐写算法进行检测时，在嵌入率大于 50%时，本文所提算法检测准确率较高，而在嵌入率低于 50%时，文献[19]由于结合了 LSTM 与 CNN 网络提取音频隐写特征，而 LSTM 网络在处理语音序列时，不仅能够提取出当前语音帧的相关特征，而且能够提取出与当前帧相邻的语音帧特征，从而将当前语音帧的相关特征与相邻语音帧的相关特征相结合以便提取出更多的语音特征，并送入后续 CNN 网络做进一步的特征提取与判别。而本文算法仅仅使用手工提取的经 G.729A 编码之后的五个码字特征与 CNN 网络相结合进行进一步的特征提取，无法提取出音频序列中与当前语音帧相邻的语音帧之间的特征，因此本文所提算法在嵌入率低于 50%时的低嵌入率下隐写分析检测效果欠佳。另一方面，对于同一段音频样本来说，隐写嵌入率越大，手工提取出的 5 个特征在音频样本隐写前后发生的变化也就越明显，故当嵌入率大于 50%时，本文所提出的基于手工特征提取与结果融合的 CNN 音频隐写分析算法取得了较好的检测结果。

由表 2 可知，在同时检测 CNV-QIM 音频隐写算法和基音音频隐写算法、待测音频样本时长为 1 s、嵌入率为 50%的情况下，本文所提算法的检测准确率可以达到 81.9%。在同时对两种隐写算法进

行检测时，在嵌入率为 50%的情况下，本文所提算法与 SS-QCCN[15]、CCN[17]、Ensemble(SS-QCCN+CCN)[15]+[17]、RNN-SM[13]、SFFN[19]算法相比，检测准确率分别提高了 31.1%、21.3%、0.3%、7.9%、1.0%。综上所述，本文所提算法在待测音频样本时长均为 1s 的情况下，对嵌入率大于 50%的音频隐写语音片段具有较好的检测效果。

表 2 待测音频时长为 1s 时不同嵌入率下对三种不同隐写算法(CNV[20]、Pitch[21]和 C-P[20]+[21]音频隐写算法)的音频隐写分析检测准确率

隐写分析算法	隐写算法	嵌入率(%)								
		20	30	40	50	60	70	80	90	100
SS-QCCN[15]	CNV	66.5	73.3	77.8	82.5	90.6	94.2	95.3	98.0	99.0
	Pitch	51.3	38.0	20.3	11.2	7.5	5.0	3.0	2.0	1.0
	C-P	61.3	56.3	52.8	50.8	51.5	52.4	53.6	52.6	52.6
CCN[17]	CNV	45.8	49.2	42.4	42.2	35.6	28.5	21.0	16.5	14.2
	Pitch	57.0	66.9	74.5	81.0	82.5	89.1	92.6	93.1	94.7
	C-P	51.6	58.6	57.7	60.6	60.2	56.9	58.0	58.6	57.4
Ensemble (SS-QCCN+CCN) [15]+[17]	CNV	85.7	88.0	85.8	88.3	93.1	94.9	95.9	98.0	99.1
	Pitch	83.5	80.8	80.3	84.2	84.0	89.6	93.0	93.2	94.7
	C-P	83.1	84.4	79.0	81.6	80.3	79.6	80.9	80.8	80.6
RNN-SM[13]	CNV	86.7	87.0	93.8	93.9	93.4	98.0	95.9	97.2	98.8
	Pitch	39.5	26.7	33.6	41.1	54.6	53.2	75.9	70.1	71.1
	C-P	68.9	63.2	74.7	74.0	70.4	85.6	80.8	75.1	76.5
SFFN[19]	CNV	86.7	91.3	90.6	94.0	96.7	97.5	98.5	98.8	99.3
	Pitch	67.2	70.1	75.3	80.4	84.7	88.6	92.0	93.6	94.6
	C-P	77.6	81.2	77.3	80.9	85.0	82.9	86.5	83.1	85.8
CMFERM (Ours)	CNV	77.6	83.5	87.5	91.1	94.1	95.5	97.5	99.0	98.4
	Pitch	63.6	71.1	76.8	81.8	86.0	89.1	91.2	93.2	94.7
	C-P	65.3	71.7	77.0	**81.9**	**86.2**	**89.2**	**91.6**	**94.0**	**95.6**

4.3 运行速度对比

本文所提算法能够同时对文献[20]和文献[21]中所提出的音频隐写算法进行有效的隐写分析检测，在同时对两种隐写算法进行检测时的运行速度如表 3 所示。本文在 Intel(R) Core(TM) i5-7200 U CPU @ 2.50GHz 电脑上测试运行速度时，未使用 GPU 加速，因此表 3 中的运行速度为使用 CPU 测试的运行时间。此外，表 3 中的平均检测时间是指在同时检测两种隐写算法时(0.1s、0.2s、…、1s、2s)每帧音频的平均检测时间。由于本文使用与文献[19]相同的数据集，因此，本文参照文献[19]得到表 3 中 SS-QCCN[15]、CCN[17]、Ensemble(SS-QCCN+CCN)[15]+[17]、RNN-SM[13]和 SFFN[19]的平均运行时间。

表 3 运行速度对比

隐写分析算法	平均检测时间(ms)
SS-QCCN[15]	0.36
CCN[17]	0.03
Ensemble(SS-QCCN+CCN)[15]+[17]	0.39
RNN-SM[13]	0.15
SFFN[19]	0.34
CMFERM (Ours)	**0.06**

由表 3 可知，本文所提算法在同时对两种音频隐写算法进行检测时每帧音频片段的平均检测时间仅仅为 0.06ms(其中每帧音频的平均手工特征提取时间为 0.03ms)，在保证音频隐写分析检测准确率的同时，检测速度较快，能够满足 VoIP 压缩域音频隐写分析实时性的检测需求。

根据上述实验结果及分析，在同时对文献[20]中 CNV-QIM 隐写算法和文献[21]中基音音频隐写算法进行检测的情况下，待测音频嵌入率均为 100%时，对于大于 0.1s 的语音片段，本文所提算法的音频隐写分析检测效果明显优于文献[19]；在待测音频时长均为 1s 时，对于嵌入率大于 50%的

语音片段，本文所提算法的音频隐写分析检测效果明显优于文献[19]。由以上实验结果和分析可知，本文所提算法不仅取得了优异的检测准确率，而且检测时间较快，能够满足 VoIP 压缩域音频隐写分析实时性的检测需求。

5 小结

本文提出了一种基于手工特征提取与结果融合的 CNN 音频隐写分析算法并荣获第一届全国信息隐藏大赛音频组第一名。通过将手工提取的 CNV-QIM 隐写特征和基音隐写特征分别输入至不同的 CNN 网络，最后对两个 CNN 网络的输出结果进行融合判别，实现了在 VoIP 压缩域同时对 CNV-QIM 隐写算法和基音隐写算法的有效检测，并分别与文献[13][15][17][19]中的实验结果进行对比，通过实验进一步验证了本文所提算法对 CNV-QIM 隐写算法和基音隐写算法进行隐写分析检测的可行性和有效性。值得一提的是，本文所提算法使用两个不同的 CNN 网络同时对两种隐写算法进行检测分析，最后再对两个 CNN 网络的检测结果按照一定的规则进行融合判决，而不是分别单独对 CNV-QIM 隐写算法或者基音隐写算法进行检测。实验结果进一步证明了通过将手工提取出的音频隐写特征与 CNN 网络相结合，可以更加有效地同时对 CNV-QIM 隐写算法和基音隐写算法进行有效检测。

在实验方面，本文采用文献[19]数据集对算法进行测试。实验结果表明，在满足实时性检测的条件下，当嵌入率为 100%、音频时长为 0.1s 时，本文所提算法隐写分析检测准确率相较于文献[19]高出 9.4%，取得了良好的隐写分析检测效果。在音频时长均为 1s 的情况下同时对 CNV-QIM 隐写算法和基音隐写算法进行检测，嵌入率越高，本文所提算法的隐写分析检测准确率也越高。在音频时长均为 1s，嵌入率为 50%时，本文所提算法相较于文献[19]高出 1.0%，取得了较好的检测结果。本文优越的检测结果主要归结于通过将手工提取的音频隐写特征与 CNN 网络相结合从而获得了较好的隐写分析检测结果。本文未来针对 VoIP 压缩域隐写音频检测分析工作的方向是：(1)考虑将手工提取特征、CNN 网络与 LSTM 网络相结合，进一步提高 VoIP 压缩域语音隐写分析检测的准确率；(2)寻求将传统的音频隐写分析算法与 CNN 网络进行结合，以便更加准确地对 VoIP 压缩域语音隐写进行分析检测。

参 考 文 献

[1] Luo Wei-Qi, Zhang Yue, Li Hao-Dong. Adaptive audio steganography based on advanced audio coding and syndrome-trellis coding//International Workshop on Digital Watermarking. Magdeburg, Germany. 2017: 177-186.

[2] Wu Jun-Qi, Chen Bo-Lin, Luo Wei-Qi, Fang Yan-Mei. Audio steganography based on iterative adversarial attacks against convolutional neural networks. IEEE Transactions on Information Forensics and Security, 2020, 15: 2282-2294.

[3] Kheddar Hamza, Bouzid Merouane, Megias David. Pitch and fourier magnitude based steganography for hiding 2.4 kbps melp bitstream. IET Signal Processing, 2019, 13(3): 396-407.

[4] Tian Hui, Sun Jun, Chang Chin-Chen, Qin Jie. Hiding information into voice-over-ip streams using adaptive bitrate modulation. IEEE Communications Letters, 2017, 21(4): 749-752.

[5] Ko Hung-Jui, Huang Cheng-Ta, Horng Gwoboa, Wang Shiuh-Jeng. Robust and blind image watermarking in DCT domain using inter-block coefficient correlation. Information Sciences, 2020, 517: 128-147.

[6] Xiang Shi-Jun, He Jia-Yong. Database authentication watermarking algorithm in order preserving encrypted domain. Journal of Software, 2018, 29(12): 3837-3852 (in Chinese).
(项世军, 何嘉勇. 一种保序加密域数据库认证水印算法. 软件学报, 2018, 29(12): 3837-3852.)

[7] Xiang Shi-Jun, Yang Le. Robust and reversible image watermarking algorithm in homomorphic encrypted domain. Journal of Software, 2018, 29(04): 957-972 (in Chinese).
(项世军, 杨乐. 基于同态加密系统的图像鲁棒可逆水印算法. 软件学报, 2018, 29(04): 957-972.)

[8] Ghasemzadeh Hamzeh, H. Kayvanrad Mohammad. Comprehensive review of audio steganalysis methods. IET Signal Processing, 2018, 12(6): 673-687.

[9] Ghasemzadeh Hamzeh, Kayvanrad Mohammad H. Universal audio steganalysis based on calibration and reversed frequency resolution of human auditory system. IET Signal Processing, 2017, 11(8): 916-922.

[10] Yang Jie, Li Song-Bin. Steganalysis of joint codeword quantization index modulation steganography based on codeword bayesian network. Neurocomputing, 2018, 313: 316-323.

[11] Ren Yan-Zhen, Yang Jing, Wang Jin-Wei, Wang Li-Na. AMR steganalysis based on second-order difference of pitch delay. IEEE Transactions on Information Forensics and Security, 2017, 12(6): 1345-1357.

[12] Yang Zhong-Liang, Yang Hao, Hu Yu-Ting, Huang Yong-Feng, Zhang

Yu-Jin. Real-time steganalysis for stream media based on multi-channel convolutional sliding windows. arXiv preprint arXiv:1902.01286, 2019.

[13] Lin Zi-Nan, Huang Yong-Feng, Wang Ji-Long. RNN-SM: Fast steganalysis of VoIP streams using recurrent neural network. IEEE Transactions on Information Forensics and Security, 2018, 13(7): 1854-1868.

[14] Yang Hao, Yang Zhong-Liang, Huang Yong-Feng. Steganalysis of VoIP streams with CNN-LSTM network//Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. Paris, France, 2019: 204-209.

[15] Li Song-Bin, Jia Yi-Zhen, Kuo C-C.Jay. Steganalysis of QIM steganography in low-bit-rate speech signals. IEEE-ACM Transactions on Audio Speech and Language Processing, 2017, 25(5): 1011-1022.

[16] Li Song-Bin, Sun Dong-Hong, Yuan Jian, Huang Yong-Feng. A steganalysis method for G.729A compressed speech stream based on codeword distribution characteristics. Chinese Journal of Electronics, 2012, 40(04): 842-846 (in Chinese).
(李松斌，孙东红，袁键，黄永峰．一种基于码字分布特性的G.729A 压缩语音流隐写分析方法．电子学报，2012，40(04): 842-846.)

[17] Li Song-Bin, Jia Yi-Zhen, Fu Jiang-Yun, Dai Qiong-Xing. Detection of pitch modulation information hiding based on codebook correlation network. Chinese Journal of Computers, 2014, 37(10): 2107-2117 (in Chinese).
(李松斌，贾已真，付江云，戴琼兴．基于码书关联网络的基音调制信息隐藏检测．计算机学报, 2014, 37(10): 2107-2117.)

[18] Tian Hui, Wu Yan-Peng, Chang Chin-Chen, Huang Yong-Feng, Chen Yong-Hong, Wang Tian, Cai Yi-Qiao, Liu Jin. Steganalysis of adaptive multi-rate speech using statistical characteristics of pulse pairs. Signal Processing, 2017, 134: 9-22.

[19] Hu Yu-Ting, Li Hao-Yun, Huang Yi-Hua, Yang Zhong-Liang, Huang Yong-Feng. An effective steganalysis feature fusion network for heterogeneous parallel steganography detection. The 15th China Information Hiding Workshop. Xiamen, China, 2019.

[20] Xiao Bo, Huang Yong-Feng, Tang Shan-Yu. An approach to information hiding in low bit-rate speech stream//IEEE GLOBECOM 2008-2008 IEEE Global Telecommunications Conference. New Orleans, USA, 2008.

[21] Huang Yong-Feng, Liu Cheng-Hao, Tang Shan-Yu, Bai Sen. Steganography integration into a low-bit rate speech codec. IEEE Transactions on Information Forensics and Security, 2012, 7(6): 1865-1875.

[22] Xing Wei-Jing. The Research and Implementation of G.729 Annex A [M. S. dissertation]. Xidian University, Xi'an, 2015 (in Chinese).
(邢维静．G.729a 语音编解码算法的研究与实现[硕士学位论文].西安电子科技大学，西安, 2015.)

Li Jing-Xuan, Ph.D. candidate. His current research interests include information hiding, machine learning and steganalysis.

Hu Run-Wen, M.S. candidate. His current research interests include multimedia information security and lossless robust watermarking.

Ruan Guan-Qi, M. S. candidate. His current research interests include multimedia information security and reversible watermarking.

Xiang Shi-Jun, Ph.D., professor. His current research interests include information hiding, multimedia information security and artificial intelligence security.

Background

With the rapid development of computer technology, voice over Internet protocol transmission technology has emerged. People can use VoIP to transfer voice information, which brings great convenience to people's lives and work. However, science and technology are a double-edged sword. While VoIP technology brings convenience to people, it also provides opportunities for some criminals. Some criminals use the VoIP voice transmission protocol to transmit secret information, which poses a great challenge to social security. On the one hand, in the military, if the enemy uses VoIP compressed domain voice transmission technology to transmit secret information, it will cause incalculable losses to our side; on the other hand, in our daily life, if criminals conduct criminal activities by sending VoIP compressed audio files containing secret information to transmit secret instructions, it will pose an immeasurable threat to social security. Therefore, in order to detect whether the audio file in the VoIP

compressed domain contains secret information, a steganalysis detection technology for the VoIP compressed domain has appeared.

At present, research scholars from all over the world have proposed many steganalysis detection algorithms for speech steganography algorithms, but most of the research methods are directed to audio steganalysis algorithms in the uncompressed domain. In fact, in recent years a large number of audio steganography algorithms in the VoIP compression domain have appeared, and at the same time, the development of audio steganalysis technologies for the VoIP compression domain has also been promoted. Regarding the steganography algorithm in the VoIP compressed domain, scholars around the world have proposed many corresponding steganalysis algorithms. However, for the steganalysis algorithm in the VoIP compressed domain, when the steganography audio duration is shorter, most steganalysis algorithms cannot obtain ideal detection results.

This paper proposes a CNN audio steganalysis algorithm based on manual feature extraction and result merging, which can provide satisfactory performance for the VoIP steganography and won the first place in the audio group of the 1st Chinese information hiding competition (CIHC2019). By combining manually extracted audio features with CNN, the algorithm proposed in this paper can simultaneously detect CNV-QIM steganography algorithm and pitch steganography algorithm effectively. By comparing with other existing algorithms, experimental results have shown that the proposed algorithm have achieved state-of-the-art detection results when the steganographic audio duration is shorter.

This research is supported by the National Natural Science Foundation of China (No.61772234) and the Special Fund for Guangdong Science and Technology Innovation Strategy (No. pdjh2020a0060).