第 42 卷 第 1 期
2023 年 1 月

应用声学
Journal of Applied Acoustics

Vol. 42, No. 1
January, 2023

◇ 研究报告 ◇

基于Transformer编码器的合成语声检测系统*

万 伊[1,2] 杨飞然[1,2] 杨 军[1,2†]

(1 中国科学院声学研究所噪声与振动重点实验室 北京 100190)

(2 中国科学院大学 北京 100049)

摘要: 自动说话人认证系统是一种常用的目标说话人身份认证方案，但它在合成语声的攻击下表现出脆弱性，合成语声检测系统试图解决这一问题。该文提出了一种基于 Transformer 编码器的合成语声检测方法，利用自注意力机制学习输入特征内部的长期依赖关系。合成语声检测问题并不关注句子的抽象语义特征，用参数量较小的模型也能得到较好的检测性能。该文分别测试了 4 种常用合成语声检测特征在 Transformer 编码器上的表现，在国际标准的 ASVspoof2019 挑战赛的逻辑攻击数据集上，基于线性频率倒谱系数特征和 Transformer 编码器的系统等错误率与串联检测代价函数分别为 3.13% 和 0.0708，且模型参数量仅为 0.082 M，在较小参数量下得到了较好的检测性能。

关键词: 自动说话人认证；合成语声检测；Transformer 编码器

中图法分类号: TP302.1 **文献标识码:** A **文章编号:** 1000-310X(2023)01-0026-08

DOI: 10.11684/j.issn.1000-310X.2023.01.004

Transformer encoder-based spoofing countermeasure for synthetic speech detection

WAN Yi[1,2] YANG Feiran[1,2] YANG Jun[1,2]

(1 *Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing* 100190, *China*)

(2 *University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: The automatic speaker verification system is a commonly used solution for target speaker identity authentication, but it shows vulnerability under the attack of synthetic speech, which can be alleviated by a spoofing countermeasure system. In this paper, we introduce a synthetic speech detection method based on the Transformer encoder, which uses the self-attention mechanism to learn the long-term dependencies of the input features. Synthetic speech detection does not focus on the abstract semantic features of the sentences, and a model with small parameters can also perform well. This paper evaluated the performance of four commonly used synthetic speech detection features on Transformer encoders. On the evaluation set of the ASVspoof2019 challenge logical access scenario, the proposed system based on linear frequency cepstral coefficient features and Transformer encoder achieves an equal error rate (EER) of 3.13% and a tandem detection cost function (t-DCF) of 0.0708, respectively, and the parameters of the model is only 0.082 M, a better detection performance is obtained with a smaller model.

Keywords: Automatic speaker verification; Synthetic speech detection; Transformer encoder

2021-11-08 收稿; 2022-02-15 定稿

*国家自然科学基金项目 (62171438), 中国科学院青年创新促进会基金项目 (2018027), 中国科学院声学研究所自主部署“前沿探索”类项目 (QYTS202111)

作者简介: 万伊 (1995–), 女, 河北张家口人, 博士研究生, 研究方向: 合成语声鉴伪。

†通信作者 E-mail: jyang@mail.ioa.ac.cn

0 引言

语声是用户与智能设备之间的一种实用的交互方式。语声信号便于采集，易于获取，并且可以和其他生物特征相结合对用户进行个人身份验证。自动说话人认证系统 (Automatic speaker verification system, ASV) 作为一种高效便捷的身份验证方案，在电话银行、健康管理、智能家居等电话和网络接入的控制系统中有着广泛应用。但是与其他生物认证技术类似，ASV 系统容易受到欺骗攻击。

近年来，深度学习技术的快速发展促进了语声合成 (Text-to-speech, TTS) 与语声转换 (Voice conversion, VC) 技术的飞速发展。谷歌、百度等公司提出了 WaveNet、Tacotron 和 Deep Voice 等高效的语声合成技术 [1−3]，可以根据输入的任意文字来生成接近真人发声的高质量语声，而语声转换技术可以将输入的真实语声转换成目标说话人的语声。语声合成与转换技术的发展给人们的生活带来便利。但随着网络与社交媒体的发展，犯罪分子可以很容易取得用户发布在网络平台上的声频、视频数据并借助先进的语声合成与转换算法来生成合成语声，对用户的个人账户和设备进行攻击。

已有的研究表明，ASV 系统本身在合成语声攻击的场景下表现出脆弱性 [4−6]。为提高 ASV 系统的安全性，可以设计一个独立的合成语声检测系统 (Spoofing countermeasure system, CM)，专门用于检测欺骗攻击。独立系统的优势是不需要对原有的 ASV 系统进行大幅度改动，只需通过和 ASV 系统融合，就能得到对于输入语声的准确判断。如图 1 所示，合成语声检测系统与 ASV 系统可以通过串联和并联的方式进行融合 [7]。

ASVspoof2019 挑战赛的任务中包括了逻辑攻击 (Logical access, LA) 和物理攻击 (Physical access, PA) 两种场景，其中 LA 场景包括了语声合成和语声转换两种针对 ASV 系统的攻击方式，PA 场景则特指录声重放的攻击方式。本文的研究主要基于 LA 场景。ASVspoof2019-LA 任务 [8] 为合成语声检测任务提供了统一的数据库与评价标准，推动了合成语声检测技术的研究与发展。目前许多有效的合成语声检测系统都是基于声学特征与机器学习模型。

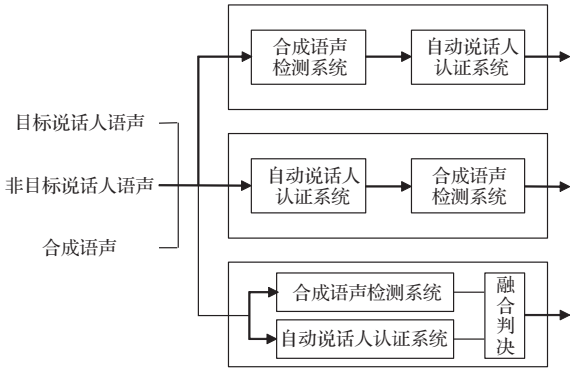

图 1　合成语声检测系统与自动说话人认证系统的 3 种融合方式

Fig. 1　Three combination methods of spoofing countermeasure system and automatic speaker verification system

Zhang 等 [9] 探究了功率谱特征在卷积神经网络 (Convolutional neural networks, CNN) 和循环神经网络 (Recurrent neural networks, RNN) 上的表现，并提出了 CNN 与 RNN 两种网络相结合的方法，用于合成语声检测任务。Todisco 等 [10] 提出了常数 Q 倒谱系数 (Constant-Q cepstral coefficients, CQCC)，作为合成语声检测的有效特征。CQCC 在深度神经网络 (Deep neural networks, DNN) 和残差神经网络 (Deep residual network, ResNet) 上都得到了较好的检测效果 [11−13]。Lavrentyeva 等 [14] 提出了基于线性频率倒谱系数 (Linear frequency cepstral coefficients, LFCC) 和轻量级卷积神经网络 (Light convolutional neural networks, LCNN) 的方法，并引入了基于角度的损失函数用于网络训练。Luo 等 [15] 提出了基于 LFCC 和胶囊网络 (Capsule network) 的合成语声检测方法。Zhang 等 [16] 提出了单分类的损失函数以提高检测算法的泛化能力。Sahidullah 等 [17] 针对合成语声检测中的常用特征进行了探究，认为声学特征中的高频特征、动态特征、相位特征以及特征间的长期依赖关系对于合成语声检测更为有效。

Transformer 模型 [18] 是一种基于自注意力机制的自然语言处理模型，最初由谷歌提出并用于机器翻译任务中，其中的 Transformer 编码器模型近年来在图像和语声的分类任务中也取得了较好的结果 [19−21]。Zhang 等 [22] 将 Transformer 编码器用于特征的进一步提取，并用 ResNet 模型作为分类器，用于合成语声检测任务中。

现有的合成语声检测系统大多模型结构复杂且参数量较大。本文提出了一种基于Transformer编码器模型的合成语声检测方法，利用自注意力机制，学习输入信号声学特征内部的相关性和长期依赖关系。相比于原始的Transformer编码器模型[18]，适当地减少了编码器层数与注意力操作次数，模型结构简单，参数量较小，实验结果表明本文所提出的模型在ASVspoof2019 数据集上取得了比大部分现有模型更优的性能。

1 合成语声检测系统结构

基于Transformer编码器的合成语声检测系统框架如图2所示，分为声学特征提取和分类器两部分，其中分类器采用了Transformer编码器结构。Transformer编码器[18]由N个相同的层堆叠而成，每个编码器层包含两个子层，分别是多头自注意力和前馈神经网络。编码器通过自注意力机制对输入特征进行非线性变换，能够更好地捕捉特征的内部相关性，学习长期依赖关系。在每个子层后都添加了残差连接和层归一化，以保证网络的快速收敛。

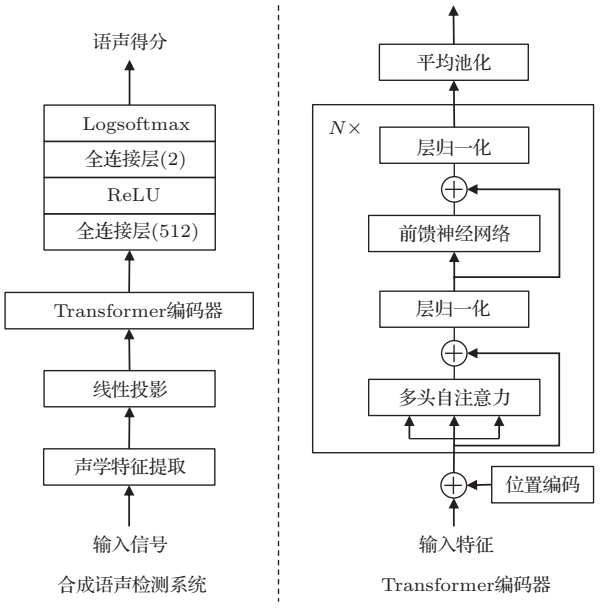

图2 基于Transformer编码器的合成语声检测系统

Fig. 2 Transformer encoder-based spoofing countermeasure system for synthetic speech detection

1.1 声学特征提取

本文选取了在合成语声检测中常用的4种声学特征，分别是对数功率谱(Spectrogram, Spec)、LFCC、CQCC和修正群延时(Modified group delay, MGD)特征，均为经过分帧将语声信号近似为平稳信号后计算得到的帧级别特征。基于时序的帧级别特征可以使得Transformer编码器更好地学习到特征内部的长期依赖关系。ASVspoof2019-LA数据集中语声数据的采样率为$f_s = 16$ kHz。

对数功率谱是对输入语声分帧加窗后，进行512点傅里叶变换，并计算功率谱，然后对功率谱进行对数尺度变换得到的[9]。所采用的是窗长为25 ms，帧移为10 ms的汉宁窗，如式(1)所示：

$$S_{\text{spec}}(t,\omega) = 20\lg\left(\frac{|X(t,\omega)|}{2\times 10^{-5}}\right), \tag{1}$$

其中，$|X(t,\omega)|$为输入语声信号的短时傅里叶变换幅度谱，t为帧标识，ω为角频率。

LFCC是根据ASVspoof2019挑战赛所提供的基线系统[8]计算得到的。使用窗长为20 ms、帧移为10 ms的汉明窗对语声进行分帧加窗，并做512点傅里叶变换，计算对数功率谱。然后利用一组线性三角滤波器处理后，再进行对数运算与离散余弦变换(Discrete Cosine transform, DCT)后得到倒谱系数，其中滤波器数量为20个。对20维的倒谱特征逐帧计算一阶与二阶动态系数，最终得到60维的特征向量。

常数Q倒谱系数是基于常数Q变换(Constant Q Transform, CQT)计算得到的倒谱特征，同样根据ASVspoof2019挑战赛所提供的基线系统[8]进行计算。CQT更符合人类的感知系统，在低频具有更高的频率分辨率，而在高频具有更高的时间分辨率，对于高频信息损失更少。选择的分析频段最高频率为ASVspoof2019数据集中语声数据采样频率的1/2，即$f_{\max} = 8$ kHz，将分析频段分割为10个八度音，每个八度音内再分割为96个频带，则最低频率定义为$f_{\min} = f_{\max}/2^{10} \simeq 7$ Hz。利用CQT计算得到频谱，然后计算对数功率谱，再进行重采样后，就可利用DCT变换计算得到CQCC特征。本文提取CQCC及其一阶和二阶动态系数，最终得到90维的特征向量。

MGD特征是傅里叶频谱的复数表示，同时包含了相位和幅度信息。给定输入信号$x(n)$，经过短时傅里叶变换得到复数谱$X(\omega)$，其实部和虚部分别为$X_R(\omega)$和$X_I(\omega)$，定义$nx(n)$的复数谱为$Y(\omega)$，其实部和虚部分别为$Y_R(\omega)$和$Y_I(\omega)$，则MGD特征

$\tau_{\rho,\gamma}(\omega)$ 可表示为 [23]

$$\tau_{\rho}(\omega)=\frac{X_R(\omega)Y_R(\omega)+Y_I(\omega)X_I(\omega)}{|S(\omega)|^{2\rho}},\tag{2}$$

$$\tau_{\rho,\gamma}(\omega)=\frac{\tau_{\rho}(\omega)}{|\tau_{\rho}(\omega)|}|\tau_{\rho}(\omega)|^{\gamma},\tag{3}$$

其中，$|S(\omega)|^2$ 是由 $|X(\omega)|^2$ 先做 DCT，取前 30 个系数后再进行逆 DCT 操作后得到的，相当于进行了谱平滑。ρ 和 γ 为控制复数谱形状的参数，本文中取 $\rho=0.3$、$\gamma=0.1$。

由于 Transformer 编码器是完全基于注意力机制，不含任何递归与卷积操作，因此不会对输入序列的顺序进行建模，需要通过添加位置编码 (Positional encoding, PE)，将输入特征时间帧的位置信息注入到输入序列中。本文中先将每段声频长度统一为 4 s，再提取相应的声学特征，对于较短的声频信号采用先重复再截取的方式，较长的声频信号则直接截取。将声学特征进行线性投影后得到相应的嵌入特征，再与一组可学习的位置编码相加，最终得到分类器的输入特征。

1.2　分类器结构

Transformer 编码器 [18] 将一组给定的输入序列 $\boldsymbol{X}=\left(\boldsymbol{x}_1^{\mathrm{T}},\cdots,\boldsymbol{x}_T^{\mathrm{T}}\right)^{\mathrm{T}}$ 映射到相同维度的输出序列，输入序列 $\boldsymbol{X}\in\mathbb{R}^{T\times D_{\mathrm{model}}}$，其中 T 为序列长度，D_{model} 为输入特征维度。自注意力机制是 Transformer 编码器的核心，对于每个输入向量 $\boldsymbol{x}_t$，自注意力函数可以描述为将查询向量和一组键值对向量映射到输出向量，根据查询向量和相应的键向量计算注意力得分，则输出向量可以表示为注意力得分与相应值向量的加权和。其中查询向量和键值对向量都是由 $\boldsymbol{x}_t$ 映射得到，查询向量和键向量的维度均为 D_k，值向量的维度为 D_v。

常用的注意力函数包括加性注意力函数和点积注意力函数，这两种方法在理论上复杂度是相似的，但点积注意力可以使用矩阵乘法来实现，因此在实际应用中效率更高。本文中使用的是缩放点积注意力函数 [18]。计算各向量组成的查询矩阵 $\boldsymbol{Q}$ 和键矩阵 $\boldsymbol{K}$ 的点积，并用 $\sqrt{D_k}$ 进行缩放，经过 softmax 函数计算注意力得分，并与值矩阵 $\boldsymbol{V}$ 进行加权求和后，即可得到注意力输出。注意力输出为 $\boldsymbol{A}=\mathrm{softmax}\left(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{D_k}\right)\boldsymbol{V}$，此处用 $\sqrt{D_k}$ 进行缩放，是为了避免输入值过大时，softmax 函数进入饱和区，导致输出值过小而使梯度无法更新。

多头自注意力机制则是在不同的子空间中，分别进行注意力操作。多头自注意力的输出为 $[\boldsymbol{A}_1,\cdots,\boldsymbol{A}_M]\boldsymbol{W}^O$，其中

$$\boldsymbol{A}_i=\mathrm{softmax}(\boldsymbol{Q}_i\boldsymbol{K}_i^{\mathrm{T}}/\sqrt{D_k})\boldsymbol{V}_i$$

为各个子空间得到的注意力输出，$\boldsymbol{Q}_i\in\mathbb{R}^{T\times D_k}$，$\boldsymbol{K}_i\in\mathbb{R}^{T\times D_k}$、$\boldsymbol{V}_i\in\mathbb{R}^{T\times D_v}$ 分别为第 i 个子空间中的查询矩阵、键矩阵与值矩阵，$\boldsymbol{W}_O\in\mathbb{R}^{MD_v\times D_{\mathrm{model}}}$ 为多头自注意力权重矩阵，$M=D_{\mathrm{model}}/D_k$ 为注意头的数目，代表进行注意力操作的次数。多头自注意力的输出可以看作是对各个子空间注意力输出进行加权求和的结果。

Transformer 编码器的第二个子层是一组前馈神经网络，由两层全连接层组成，其隐藏层单元数为 D_{ff}，两层全连接层中间采用 ReLU 激活函数进行非线性变换。前馈神经网络层的输出为 $\boldsymbol{O}=\max\left(0,\boldsymbol{z}_t\boldsymbol{W}_1+\boldsymbol{b}_1\right)\boldsymbol{W}_2+\boldsymbol{b}_2$，其中 $\boldsymbol{z}_t$ 为 $\boldsymbol{x}_t$ 经过多头自注意力层后，再进行残差连接与层归一化操作得到的输出向量，$\boldsymbol{W}_1\in\mathbb{R}^{D_{\mathrm{model}}\times D_{ff}}$、$\boldsymbol{W}_2\in\mathbb{R}^{D_{ff}\times D_{\mathrm{model}}}$ 分别为两层前馈神经网络的权值，$\boldsymbol{b}_1$、$\boldsymbol{b}_2$ 为相应的偏置值。

Transformer 编码器的输出经过一组前馈神经网络和 Logsoftmax 层，得到输入序列分别属于自然语声和合成语声的概率，并计算相应的分数，作为系统的最终输出。对于给定输入序列 $\boldsymbol{X}$，其得分根据其分别属于自然语声和合成语声概率的对数似然比得到，如式 (4) 所示：

$$L_{\mathrm{score}}=\lg P\left(\text{bona fide}|\boldsymbol{X}\right)-\lg P\left(\text{spoof}|\boldsymbol{X}\right).\tag{4}$$

2　实验设置

2.1　数据集

ASVspoof2019 数据集中的 LA 数据集是由英国爱丁堡大学语声技术研究中心发布的专门用于评估合成语声检测算法的数据集 [8]，本文所有实验都是在该数据集上设计验证的。表 1 给出了该数据集的划分方式和每个子集的组成方式。

ASVspoof2019-LA 数据集中包含自然语声 (Bona fide utterances) 与合成语声 (Spoofed utterances) 两部分。自然语声均录制于爱丁堡大学半消声室中，录声没有经过后期处理，且没有明显的背景噪声与信道噪声。合成语声由自然语声作为训练数据，通过多种不同的语声合成和语声转换算法得到。

需要注意的是，作为合成语声训练数据的自然语声与ASVspoof2019-LA数据集中的自然语声数据不重叠。

表1 ASVspoof2019-LA数据集组成

Table 1 Partitions of the ASVspoof2019-LA dataset

	自然语声	合成语声	
	语声数据数目	语声数据数目	攻击算法编号
训练集	2580	22800	A01–A06
验证集	2548	22296	A01–A06
测试集	7355	63882	A07–A19

ASVspoof2019-LA数据集划分为3个子集：训练集、验证集和测试集。训练集和验证集中的合成语声是由编号A01-A06的算法处理得到，其中包括4种语声合成算法和2种语声转换算法。测试集中的合成语声是由编号A07-A19的算法处理得到，其中包括7种语声合成算法和6种语声转换算法，且包含了两种在训练集中出现过的已知算法，其余均为未知算法。每个子集中自然语声与合成语声的比例约为1∶9。本文中训练集用于网络模型的训练，验证集用于模型选择，在验证集上性能最优的模型用于测试最终结果。

2.2 评价指标

合成语声检测系统的评价指标有两种，分别是串联检测代价函数(tandem detection cost function, t-DCF)和等错误率(Equal error rate, EER)[24]，根据分类器输出的语声得分情况来计算这两类评价指标。

t-DCF是基于ASVspoof2019挑战赛提出的一种新的评价指标，引入了风险决策的思想，可用于评估ASV系统和CM系统的综合性能。本文将t-DCF作为主要评价指标，为了便于计算通常使用其最小归一化的形式，可表示为

$$\text{t-DCF}_{\text{norm}}^{\min}=\min_{\tau_{\text{cm}}}\left\{\beta P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})+P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}})\right\}, \quad (5)$$

其中，参数β取决于经验参数和ASV系统的性能，包括输入语声属于自然语声和合成语声的先验概率、风险决策中损失值的选择，以及ASV系统的丢失率(Miss rate)和误报率(False alarm rate)。文献[25]中详细介绍了有ASVspoof2019挑战赛中t-DCF参数设置的详细信息。$P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})$和$P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}})$分别代表阈值$s=\tau_{\text{cm}}$时，CM系统的丢失率和误报率，如式(6)和式(7)所示：

$$P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})=\frac{\#\{\text{bona fide trials with CM score}\leqslant\tau_{\text{cm}}\}}{\#\{\text{Total bona fide trials}\}}, \quad (6)$$

$$P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}})=\frac{\#\{\text{spoof trials with CM score}>\tau_{\text{cm}}\}}{\#\{\text{Total spoof trials}\}}, \quad (7)$$

其中，#代表符合括号中所描述条件的语声数目。丢失率$P_{\text{miss}}^{\text{cm}}(s)$随着阈值$s$的增大而单调递增，误报率$P_{\text{fa}}^{\text{cm}}(s)$随着阈值增大而单调递减。

EER是使$P_{\text{miss}}^{\text{cm}}(s)$和$P_{\text{fa}}^{\text{cm}}(s)$同时最小的错误率，用于衡量单一CM系统的性能，本文中作为辅助评价指标。

2.3 训练策略及参数设置

本文设定线性投影后得到的嵌入特征维度为$D_{\text{model}}=60$，根据经验值设定前馈神经网络隐藏层单元数$D_{ff}=256$，约为嵌入特征维度的4倍。分别在编码器的多头自注意力层和全连接层后加入Dropout层以防止模型过拟合。

网络训练使用带权重的交叉熵作为损失函数，来消除训练集中自然语声与合成语声之间数据量不平衡带来的影响，权值设为9∶1。共训练500个周期，批处理大小为32。使用AdamW[26]优化器对模型进行优化，其中优化器参数$\beta_1=0.9$，$\beta_2=0.999$，权重衰减值设为0.01，每个训练周期学习率初始值设置为5×10^{-5}。

3 实验结果

3.1 Transformer编码器结构对合成语声检测系统性能的影响

本节探讨Transformer编码器注意头数目M和编码器层数N对本文出所提出的合成语声检测系统的影响。表2中给出了在输入特征为LFCC时，不同注意头数目和编码器层数的合成语声检测系统在ASVspoof2019-LA测试集上的结果及对应的模型参数量。Transformer编码器模型参数量不受注意头数目影响，但会受编码器层数影响，编码器层数增加，模型的参数量和训练成本都会随之增加。

从表2中可以看出，在注意头数目M相同的条件下，合成语声检测系统均在编码器层数$N=1$时

达到了最佳性能。由于合成语声检测是为了判断输入语声是自然语声还是合成语声，更关注输入特征的内部相关性与长期依赖性，而非抽象的语义信息，因此浅层的编码器对合成语声检测更加有效。而在编码器层数 N 相同的条件下，系统选择不同的注意头数目 M 对 EER 和 t-DCF 两种性能指标的影响并不显著。多头自注意力机制的目的是在不同的表示子空间中分别进行自注意力操作，使模型学习到来自不同表示子空间的信息。注意头数目代表了子空间的个数，注意头数目过少，则从特征中学习到信息不足，而过多则可能会学习到与合成语声检测无关的干扰信息。

根据表 2，Transformer 编码器选择注意头数目 $M=2$、编码器层数 $N=1$ 时，合成语声检测系统在 EER 和 t-DCF 两种指标上都取得了最小值，模型参数量也仅为 0.082 M。后文对合成语声检测系统的研究也都将基于此编码器结构。

表 2　不同结构编码器的合成语声检测系统在 ASVspoof2019-LA 测试集上的性能

Table 2　Results of spoofing countermeasure systems with different encoder architectures on the evaluation set of the ASVspoof2019-LA dataset

	EER/%				t-DCF				参数量
	$M=1$	$M=2$	$M=4$	$M=6$	$M=1$	$M=2$	$M=4$	$M=6$	
$N=1$	3.64	**3.13**	3.90	3.71	0.0750	**0.0708**	0.1009	0.0992	0.082 M
$N=2$	5.34	6.91	7.45	6.40	0.1324	0.1437	0.1415	0.1369	0.128 M
$N=3$	7.14	5.93	8.50	6.65	0.1525	0.1402	0.1595	0.1367	0.174 M
$N=4$	7.56	6.65	7.62	8.89	0.1649	0.1497	0.1548	0.1591	0.220 M
$N=5$	8.13	5.23	5.87	6.23	0.1800	0.1362	0.1551	0.1478	0.266 M
$N=6$	9.94	5.70	6.81	6.10	0.1735	0.1445	0.1596	0.1494	0.312 M

3.2　基于 Transformer 编码器的合成语声检测系统性能

表 3 中给出了在不同声学特征下，基于 Transformer 编码器的合成语声检测系统在 ASVspoof 2019-LA 数据集上得到的实验结果，其中参数量差异是由输入特征维度不同导致的。系统的名称由输入特征和分类器名称组成，仅考虑基于机器学习的模型的参数量。TE 代表本文所使用的 Transformer 编码器模型。其中 B1 和 B2 分别是 ASVspoof2019 挑战赛提供的两类基线模型 [8]，LFCC-LCNN [14] 是 ASVspoof2019 挑战赛 LA 场景中最佳的单一系统，即非融合系统。

从表 3 中可看出，本文提出的基于 Transformer 编码器的合成语声检测系统，在 LFCC 和 MGD 特征下的 t-DCF 指标均超过了基线系统 B1 和 B2，且 LFCC 特征在 Transformer 编码器模型上表现出了优势。同时，表 3 中还显示出了模型在验证集和测试集的性能差异，验证集上的 EER 和 t-DCF 指标都远低于测试集，说明未知的攻击算法会对系统的检测性能产生较大的影响。因此对于合成语声检测问题，需要提高系统的泛化性能，以应对未知算法的攻击。

得到本文最佳结果的为 LFCC+TE 系统，在测试集上的 EER 和 t-DCF 指标分别为 3.13% 和 0.0708，比 ASVspoof2019 挑战赛 LA 场景下的基线系统 B2 分别降低了 61.31% 和 66.54%，比最佳的 LFCC+LCNN 系统分别降低了 38.14% 和 29.2%，证明了 Transformer 编码器对于合成语声检测的有效性。同时，LCNN 模型中为了防止模型过拟合，加入了较大的全连接层，导致模型参数量较大，为 10.22 M，而本文所提出的 LFCC+TE 模型参数量为 0.082 M，仅为 LFCC+LCNN 模型的 0.78%。

为了验证本文所提系统的有效性，表 4 中对比了 LFCC+TE 系统与目前已有的、性能较好的单一合成语声检测系统的性能，系统名称分别由特征和分类器名称组成。其中 Zhang 等 [22] 提出的 LPS+TE-ResNet 系统，也是基于对数功率谱特征，但与本文所采用的参数略有不同，其在测试集上的表现优于本文的 Spec+TE 系统，但从表 4 中可以

表 3 不同合成语声检测系统在 ASVspoof2019-LA 数据集上的性能

Table 3 Results of different spoofing countermeasure systems on the ASVspoof2019-LA dataset

系统名称	参数量	验证集		测试集	
		EER/%	t-DCF	EER/%	t-DCF
B1: CQCC + GMM[8]		2.71	0.0663	9.57	0.2366
B2: LFCC + GMM[8]		0.43	0.0123	8.09	0.2116
LFCC + LCNN[14]	10.22M	0.16	0.0043	5.06	0.1000
LFCC + TE	**0.082M**	0.47	0.0134	**3.13**	**0.0708**
Spec + TE	0.094M	0.24	0.0071	10.81	0.2696
CQCC + TE	0.084M	0.31	0.0099	9.42	0.2159
MGD + TE	0.094M	0.0	0.0	7.60	0.1826

表 4 与现有的单一合成语声检测系统在 ASVspoof2019-LA 测试集上的性能比较

Table 4 Performance comparison with existing single systems on the evaluation set of the ASVspoof2019-LA dataset

系统名称	EER/%	t-DCF
MFCC + ResNet [12]	9.33	0.2042
CQCC + ResNet[12]	7.69	0.2166
Spec + ResNet [12]	9.68	0.2741
LPS + TEResNet [22]	6.02	—
LFCC + TE-ResNet [22]	8.58	0.2024
FFT + LCNN [14]	4.53	0.1028
CQT-MMPS + LCNN [27]	5.99	0.1760
CQT-MMPS + ResNet [27]	3.72	0.1190
STFT-MGD-GCRNN + PLDA [28]	3.85	0.0952
LFCC + SE-Res2Net50 [29]	2.87	0.0790
CQT + SE-Res2Net50 [29]	**2.50**	0.0743
LFCC + TE	3.13	**0.0708**

看出，本文最佳的LFCC+TE系统仍具有比其更低的EER。该系统对输入数据进行数据增强后，将Transformer编码器用于特征的进一步提取，但并未对编码器原始结构进行改动。根据文献[22]，本文还搭建了LFCC+TE-ResNet系统，并在表4中给出了其在ASVspoof2019-LA测试集上的性能，其系统整体参数量为27.62 M。除此之外，表4中系统的EER与t-DCF结果均来自原论文，原论文中未给出的结果用‘—’代表。从表4中可以看出，在ASVspoof2019-LA测试集上，本文所提的LFCC+TE系统比LFCC+TE-ResNet系统具有更低的EER与t-DCF。Li等[29]提出的基于LFCC特征的系统，其后端分类器选择了改进的残差网络，系统整体参数量为0.92M，在ASVspoof2019-LA测试集上具有更低的EER，但是其t-DCF指标均略高于本文提出的LFCC+TE系统，说明在考虑了决策风险的条件下，LFCC+TE系统可以在较小的参数量下获得更好的检测性能。

4 结论

本文提出了一种基于Transformer编码器的合成语声检测系统，通过自注意力机制，学习输入特征内部的长期依赖关系与时间相关性。本文还探讨了编码器结构对合成语声检测系统的影响，并据此在原始编码器模型基础上减少了编码器层数和注意力操作的次数，缩小了模型参数量。合成语声检测的目的是判别自然语声与合成语声，更注重于检测语声中的人工篡改信息，而非抽象的语义信息，使用浅层的Transformer编码器模型，可以在模型参数量较小的情况下，得到较好的合成语声检测效果。本文中提出的基于LFCC特征的系统，在ASVspoof2019-LA测试集上的EER和t-DCF指标分别可以达到3.13%和0.0708，证实了Transformer编码器结构对于合成语声检测问题的有效性，同时模型参数量仅为0.082 M。

参 考 文 献

[1] van den Oord A, Dieleman S, Zen H, et al. WaveNet: a generative model for raw audio[J]. arXiv Preprint, arXiv:

1609.03499, 2016.

[2] Wang Y, Skerry-Ryan R, Stanton D, et al. Tacotron: towards end-to-end speech synthesis[C]//Interspeech, 2017: 4006–4010.

[3] Arik S O, Chrzanowski M, Coates A, et al. Deep voice: real-time neural text-to-speech[J]. arXiv Preprint, arXiv: 1702.07825, 2017.

[4] Kinnunen T, Wu Z, Lee K, et al. Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2012: 4401–4404.

[5] de Leon P, Pucher M, Yamagishi J, et al. Evaluation of speaker verification security and detection of HMM-based synthetic speech[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(8): 2280–2290.

[6] Wu Z, Evans N, Kinnunen T, et al. Spoofing and countermeasures for speaker verification: a survey[J]. Speech Communication, 2015, 66: 130–153

[7] Kinnunen T, Lee K, Delgado H, et al. t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification[C]//Odyssey: The Speaker and Language Recognition Workshop, 2018: 312–319.

[8] Wang X, Yamagishi J, Todisco M, et al. ASVspoof 2019: a large-scale public database of synthesized, converted and replayed speech[J]. Computer Speech and Language, 2020, 64: 101114.

[9] Zhang C, Yu C, Hansen J. An investigation of deep-learning frameworks for speaker verification antispoofing[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(4): 684–694.

[10] Todisco M, Delgado H, Evans N. A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients[C]//Odyssey: The Speaker and Language Recognition Workshop, 2016: 283–290.

[11] Yu H, Tan Z, Ma Z, et al. Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(10): 4633–4644.

[12] Alzantot M, Wang Z, Srivastava M. Deep residual neural networks for audio spoofing detection[C]//Interspeech, 2019: 1078–1082.

[13] Das R, Yang J, Li H. Long range acoustic features for spoofed speech detection[C]//Interspeech, 2019: 1058–1062.

[14] Lavrentyeva G, Novoselov S, Tseren A, et al. STC antispoofing systems for the ASVspoof2019 challenge[C]//Interspeech, 2019: 1033–1037.

[15] Luo A, Li E, Liu Y, et al. A capsule network based approach for detection of audio spoofing attacks[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2021: 6359–6363.

[16] Zhang Y, Jiang F, Duan Z. One-class learning towards synthetic voice spoofing detection[J]. IEEE Signal Processing Letters, 2021, 28: 937–941.

[17] Sahidullah M, Kinnunen T, Hanilci C. A comparison of features for synthetic speech detection[C]//Interspeech, 2015: 2087–2091.

[18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, 2017: 5998–6008.

[19] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C]//International Conference on Learning Representations (ICLR), 2021.

[20] Liu Z, Lin Y, Cao Y. Swin transformer: hierarchical vision transformer using shifted windows[J]. arXiv Preprint, arXiv: 2103.14030, 2021.

[21] Gong Y, Chung Y, Glass J. AST: audio spectrogram transformer[J]. arXiv Preprint, arXiv: 2104.01778, 2021.

[22] Zhang Z, Yi X, Zhao X. Fake speech detection using residual network with transformer encoder[C]//ACM Workshop on Information Hiding and Multimedia Security, 2021: 13–22.

[23] Saratxaga I, Sanchez J, Wu Z, et al. Synthetic speech detection using phase information[J]. Speech Communication, 2016, 81: 30–41.

[24] Todisco M, Wang X, Vestman V, et al. ASVspoof 2019: Future horizons in spoofed and fake audio detection[C]//Interspeech, 2019: 1008–1012.

[25] Delgado H, Evans N, Kinnunen T, et al. ASVspoof 2021: automatic speaker verification spoofing and countermeasures challenge evaluation plan[J]. arXiv Preprint, arXiv: 2109.00535, 2021.

[26] Loshchilov I, Hutter F. Decoupled weight decay regularization[J]. arXiv Preprint, arXiv: 1711.05101, 2017.

[27] Yang J, Wang H, Das R, et al. Modified magnitude-phase spectrum information for spoofing detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1065–1078.

[28] Gomez-Alanis A, Peinado A, Gonzalez J, et al. A gated recurrent convolutional neural network for robust spoofing detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(12): 1985–1999.

[29] Li X, Li N, Weng C, et al. Replay and synthetic speech detection with Res2Net architecture[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2021: 6354–6358.