采用时域高斯协同模型的合成伪造语音检测方法

简志华，梁承涵，朱峰满
（杭州电子科技大学通信工程学院，浙江 杭州 310018）

摘　要：为降低静音分布不一致对合成伪造语音检测性能的影响，提升检测系统的准确性与泛化能力，提出了一种时域高斯协同模型的合成伪造语音检测方法。通过构建统一的高斯混合模型，从真实与伪造语音中提取对数高斯后验（LGP）特征，避免了传统独立建模带来的信息割裂与参数冗余问题，从而提升特征空间的可区分性。针对静音干扰问题，模型引入基于统计能量的阈值抑制机制，在保留潜在判别信息的同时自适应抑制低能量语音段。进一步地，采用二维卷积对时间–高斯分量二维结构进行联合建模，并在时间域和分量域分别构建图结构，通过异构图注意力网络学习跨域特征关系。实验结果表明，所提方法在 ASVspoof 2021 数据集上相较基线模型将等错误率降低 12.07%，串联检测代价函数降低 12.14%，验证了所提方法的有效性与泛化性能。

关键词：合成伪造语音检测；对数高斯后验特征；统一建模；阈值抑制
中图分类号：TP391.42
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2026054

Synthetic spoofing speech detection method using temporal Gaussian synergistic model

Jian Zhihua, Liang Chenghan, Zhu Fengman
School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: To mitigate the adverse impact of inconsistent silence distribution on synthetic spoofing speech detection performance and to improve detection accuracy and generalization capability, a temporal-Gaussian synergistic model (TGSM) for synthetic spoofing speech detection method was proposed. A unified Gaussian mixture model (GMM) was constructed to extract log Gaussian posterior (LGP) features from both bona fide and spoofed speech, thereby avoiding the information fragmentation and parameter redundancy caused by traditional independent modeling strategies and enhancing the discriminability of the feature space. To address silence-related interference, a statistical energy-based threshold suppression mechanism was introduced, which adaptively suppressed low-energy speech segments while preserving potentially discriminative information. Furthermore, two-dimensional convolution was employed to jointly model the temporal-Gaussian component structure, followed by graph construction in both the temporal and component domains. A heterogeneous graph attention network was then used to learn cross-domain feature interactions. Experimental results on the ASVspoof 2021 dataset demonstrate that the proposed method reduces the equal error rate (EER) by 12.07% and the tandem detection cost function (t-DCF) by 12.14% compared with the baseline model, validating the effectiveness and generalization capability of the proposed method.

Keywords: synthetic spoofing speech detection, log Gaussian posterior feature, uniformly constructing model, threshold screening

收稿日期：2025–11–12；**修回日期：**2026–02–15
通信作者：简志华，jianzh@hdu.edu.cn
基金项目：国家自然科学基金资助项目（No.61772166）
Foundation Item：The National Natural Science Foundation of China (No.61772166)

0　引言

自动说话人验证（automatic speaker verification, ASV）作为生物特征识别技术的重要组成部分，在身份认证、金融支付、安全监控等领域发挥着重要作用。ASV 通过分析语音的声学特征，判断输入语音是否来自目标说话人，从而实现个体身份的自动验证[1-2]。然而，近年来，随着语音伪造与合成技术的不断发展[3]，ASV 系统面临着日益严峻的安全挑战，容易受到合成伪造语音的攻击。由于合成语音凭借其高度可控性、生成成本低和合成质量的快速提升，已成为 ASV 系统最具威胁性的攻击手段[4]。因此，如何有效检测合成语音，提升 ASV 系统的抗欺骗攻击能力，已成为语音安全领域的重要研究课题，并在学术界和工业界引起了广泛关注。

目前，合成语音检测方法主要分为两类：基于人工设计特征的传统方法和基于端到端深度学习的方法。传统方法通常包括前端声学特征提取和后端分类器设计两个关键步骤，其中常见的特征提取方法包括感知线性预测（perceptual linear prediction, PLP）系数、Gammatone 频率倒谱系数（Gammatone frequency cepstral coefficient, GFCC）、梅尔频率倒谱系数（Mel-frequency cepstral coefficient, MFCC）和短时傅里叶变换（short-time Fourier transform, STFT）能量谱等。这些方法通过对语音信号的时频分布、频带能量结构和感知建模进行刻画，以挖掘真实与合成语音之间的细微差异。PLP 系数特征考虑了人类听觉的掩蔽效应和等响度曲线，适用于建模语音中的感知相关信息，但在复杂声学环境中稳定性不足[5]。GFCC 使用模拟人耳听觉模型的 Gammatone 滤波器组，能够增强语音信号在非线性频带下的建模能力，在嘈杂背景下具有一定鲁棒性[6]。MFCC 由于符合人耳听觉感知特性，在语音识别领域广泛应用，但其高频区域分辨率较低，无法有效捕捉合成语音在高频部分的异常特征[7]。STFT 能量谱可直接反映语音信号的局部频谱结构，尤其在分析合成语音瞬时能量变化方面具有一定优势，其性能易受分析窗长的影响，窗长过长会导致瞬时能量变化细节难以捕捉，过短则会降低频率分辨率，难以准确反映频谱结构[8]。

近年来，深度学习方法在语音反欺骗检测任务中取得了显著进展，并在 ASVspoof 系列挑战赛中展现出较强的特征学习能力。卷积神经网络（convolutional neural network, CNN）通过层级特征提取有效捕捉语音信号的局部模式，但受限于固定感受野，在长时依赖关系建模方面存在不足[9]。递归神经网络（recurrent neural network, RNN）及其变体如长短期记忆（long short-term memory, LSTM）网络、门控循环单元（gated recurrent unit, GRU）可建模时序动态，但计算复杂度较高，且在长序列任务中训练稳定性受限[10]。基于 Transformer 的方法依赖自注意力机制实现全局信息交互，在合成语音检测中表现出较强的建模能力，但对数据规模与计算资源依赖性较大，且在分布变化场景下易发生过拟合[11]。还有方法进一步引入自监督学习（self-supervised learning, SSL）特征或图结构建模策略，以增强模型对复杂语音结构的表达能力。例如，基于 SSL 表征的反欺骗模型通过预训练语音表示提升检测性能，AASIST 及其改进版本 AASIST2 通过谱–时联合图注意力与多尺度建模，在多种评测场景下取得了良好效果。尽管深度学习方法在语音欺骗检测任务中展现了一定优势，但仍存在诸多挑战，如模型易受攻击方式变化的影响、泛化能力较弱，以及在计算资源受限的环境中推理效率较低[12]。

同时，现有研究表明，语音中的静音片段会对合成语音检测模型的判别行为产生显著影响，部分模型可能在训练过程中过度依赖静音模式而非语音内容本身的欺骗特征，从而在语音数据分布发生变化时导致检测性能显著下降。文献[13]针对 ASVspoof 2019/2021 数据集的系统性分析研究进一步指出，真实与伪造语音在前导与尾随静音时长分布上存在明显不一致性，该分布偏置可能被模型无意利用，进而影响检测系统的泛化能力。为降低静音因素带来的干扰，已有研究提出了多种预处理或建模策略。例如，文献[14]通过静音裁剪的方式直接移除语音中的静音片段，以削弱静音信息对模型决策的影响，但该方法通常依赖人工规则设定，处理流程较为烦琐，且缺乏统一的建模框架。传统的话音活动检测（voice activity detection, VAD）方法[15]通过删除低能量语音片段抑制静音干扰，但该方法容易破坏语音的节奏与韵律结构，同时可能丢失包含异常波动信息的低能量语音片段，而这些片段在一定程度上反映了语音的真伪特性。此外，部分工作从模型训练角

度出发，通过损失函数重加权或偏置分析等方式缓解模型对静音信息的过度依赖，但此类方法往往需要额外的先验假设或复杂的训练策略，难以在不同数据分布条件下保持稳定效果。因此，如何在保留低能量语音中潜在判别信息的同时抑制静音分布不一致带来的偏置影响，仍然是合成语音检测领域亟待解决的问题。

另外，现有许多合成伪造语音检测方法通常采用独立建模策略，即分别对真实与伪造语音训练两个独立的模型，以增强不同类型之间的特征区分性[16]。这种建模方式在一定程度上保留了语音类别的独立性，但同时也带来了信息割裂与特征冗余的问题[17]。真实与伪造语音在部分特征维度上存在较大的空间分布重叠，独立建模会导致两种模型在表示空间中难以形成清晰的判别边界，从而影响模型的稳定性与泛化性。尤其在静音分布不一致或复杂背景条件下，模型对训练数据尤其敏感，缺乏自适应调节机制，会极大降低检测系统的性能。

为解决上述问题，本文从统计建模、静音抑制与结构协同建模3个相互关联的层面出发，提出了一种时域高斯协同模型（temporal-Gaussian synergistic model, TGSM）合成伪造语音检测方法。TGSM首先构建统一的高斯混合模型（Gaussian mixture model, GMM），提取真实与伪造语音的对数高斯后验（log Gaussian posterior, LGP）特征，避免了传统独立建模方法带来的信息割裂与参数冗余问题，提升了特征空间的可区分性。例如，Two-path GMM-ResNet等方法通过为真实与伪造语音分别构建独立的GMM统计建模分支，以增强类别间的差异，但该方法在一定程度上仍面临统计空间不一致与参数冗余的问题，尤其在静音分布不一致或数据分布变化场景下，其泛化能力受到限制。为减弱静音干扰，模型设计了基于GMM统计能量的阈值抑制机制，该机制并非简单裁剪或删除低能量语音片段，而是在统一建模框架下对低能量模式进行自适应抑制，从而在降低静音干扰的同时保留潜在判别信息。处理后的LGP特征采用二维卷积捕捉时序与高斯分量之间的耦合关系，再在时间域和分量域各自构建图结构，通过异构图注意力融合显著提升伪造语音检测性能。

1 TGSM合成伪造语音检测

1.1 整体框架

TGSM合成伪造语音检测系统的整体框架和处理流程如图1所示。输入的语音信号首先通过特征提取模块获得线性频率倒谱系数（linear frequency cepstral coefficient, LFCC），并结合统一建模的GMM生成LGP特征，捕捉语音的统计分布信息。随后，LGP特征被送入阈值处理模块，通过能量阈值机制自动减弱低能量或静音区域，保留具有判别价值的关键语段，从而降低无效信息干扰。最后，在后端分类模块中，系统采用时间–高斯分量跨域建模策略，融合二维卷积与异构图注意力机制，对特征进行深层次的交互建模与判别优化，最终完成真实与伪造语音的分类任务。下面依次对系统各模块的结构与原理进行详细介绍。

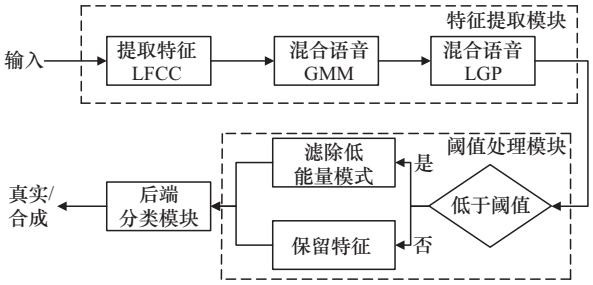


图1 TGSM合成伪造语音检测系统整体框架和处理流程

1.2 GMM统一建模与LGP特征提取

在合成伪造语音检测中，对语音特征的概率分布进行建模对提升系统判别性能具有重要作用。本文采用统一建模策略，利用单个GMM对输入的由真实与合成语音构成的混合语音进行联合建模，从而在统一表示空间内提取区分度更强的判别特征。系统首先提取混合语音库中各语音的LFCC特征，然后用GMM对LFCC特征空间进行建模。λ表示GMM参数集，其由期望最大化（expectation-maximization, EM）算法训练得到，表示为

$$\lambda=\left\{\omega_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right\},i=1,2,\cdots,N \tag{1}$$

其中，ω_i表示第i个高斯分量的混合权重，且满足$\sum_{i=1}^{N}\omega_i=1$；$\boldsymbol{\mu}_i$和$\boldsymbol{\Sigma}_i$分别表示第i个高斯分量的均值向量和协方差矩阵，N表示GMM的分量数目。

假定$\boldsymbol{x}_t$表示第t帧语音信号的LFCC特征矢量，求取$\boldsymbol{x}_t$在GMM各分量的后验概率，再取对数可得

$$\ln p(\boldsymbol{x}_t|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}\boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{x}_t + \boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i \tag{2}$$

其中，d是矢量$\boldsymbol{x}_t$的维数，T表示转置运算符。为了聚焦于判别性特征，只保留直接反映$\boldsymbol{x}_t$在不同分量下概率大小的项，即保留式(2)中与$\boldsymbol{x}_t$相关的项，得到$\boldsymbol{x}_t$在第i个高斯分量的LGP特征为

$$\mathrm{LGP}_{t,i} = -\frac{1}{2}\boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{x}_t + \boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i \tag{3}$$

因此，每帧语音就得到一个N维的LGP特征。在计算完所有语音帧的LGP特征矢量之后，就可以得到LGP特征矩阵。该矩阵在结构上具有两个维度，分别对应时间维度（语音帧序列所反映的时序动态）和高斯分量维度（高斯混合模型中各统计分量的响应分布）。

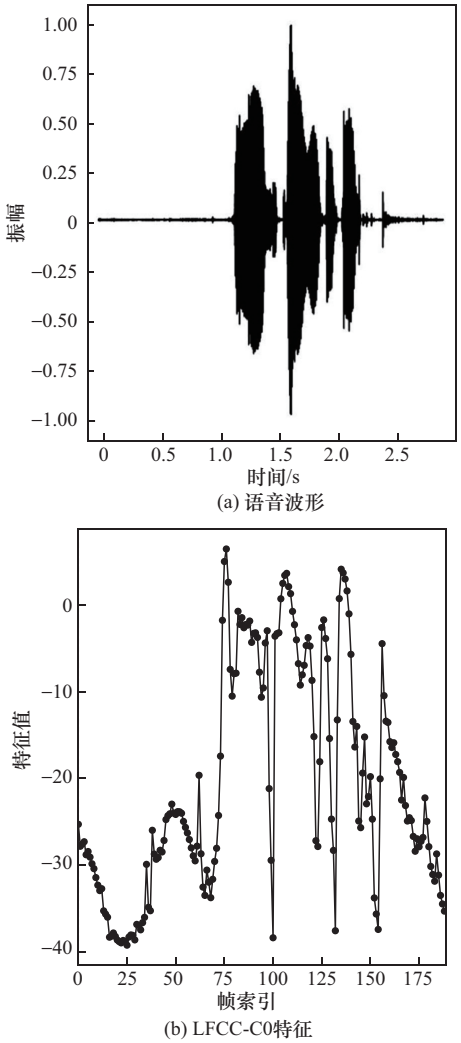
在后续的后端分类网络中，本文通过二维卷积操作对上述时间–高斯分量二维结构进行联合建模，以实现两个维度之间的协同特征学习与交互建模，使模型能够同时关注语音统计特征在时间演化过程中的动态变化以及在高斯分量维度上的分布特性。

需要说明的是，本文模型中“时域”的含义源于上述特征结构与建模方式：即模型在时间维度上对语音统计特征演化过程进行显式建模，并通过与高斯分量维度的协同优化来缓解静音与低能量语音段分布不一致所导致的信息割裂问题，而并非指传统意义上的时域信号处理方法。

1.3　低能量模式抑制

在合成伪造语音检测任务中，数据集各语音中静音片段的分布不均匀会导致模型对低能量模式的过度关注，当不同数据集存在静音分布差异时，会降低系统对真伪语音的判决性能，导致系统的泛化能力差。为此，利用GMM各分量均值向量中的能量信息$\boldsymbol{\mu}_i[0]$与阈值的比较来约束LGP特征分量，即低于阈值的GMM分量称为低能量模式，从特征空间层面对低能量模式进行抑制，从而降低静音和低能量语音片段对合成伪造语音检测系统性能的影响。低能量模式更多地体现在静音片段，抑制低能量模式可以使模型更加关注非静音的关键语音段的特性。

图2(a)为数据集任意一个语音样本的波形，可以明显观察到，静音片段通常分布于语音的首尾部分以及句子中间的停顿区域。这些静音片段虽不包含显著的语音内容，若采用VAD方法直接删除这些静音片段，不仅会破坏语音信号的时域结构和节奏韵律，同时也会丢失部分潜在的伪造痕迹，从而影响检测系统的判决能力。



(a) 语音波形

(b) LFCC-C0特征

图2　训练集任一语音信号的语音波形和LFCC-C0特征

在LFCC中，第0维系数（LFCC-C0）代表了该帧语音信号的整体能量信息，其计算式为

$$C_0 = \sum_{m=1}^{M}\ln\left(E_m\right) \tag{4}$$

其中，E_m表示LFCC求取过程中第m个滤波器输出的能量，M表示滤波器的数量。图2(b)展示了该段语音的LFCC-C0特征随帧索引的变化情况。从图2(b)中可以看出，LFCC-C0特征在语音内容区域数值大且表现出较大的波动，而在静音区域，其数值较小且波动范围明显受限。联合图2(a)和图2(b)分析

可得，LFCC-C0 能够有效表征语音信号中能量随时间的动态变化，LFCC-C0 的数值大小与局部语音能量呈明显的正相关关系，为后续基于能量阈值的抑制机制提供了合理的特征基础。结合 GMM 建模原理，$\mu_i[0]$作为第 i 个高斯分量均值向量的能量表征，其大小与建模语音帧的 LFCC-C0 数值密切相关，即 $\mu_i[0]$越小，该高斯分量更倾向于描述语音低能量模式，而这些模式通常与静音帧或背景噪声具有较高的匹配度。为了降低检测系统对这些低能量片段的关注度，设定能量阈值 θ，并依据该阈值按以下规则调整 LGP 特征。

$$\mathrm{LGP}_{t,i}=\begin{cases}\mathrm{LGP}_{t,i}, & \mu_i[0]\geqslant\theta\\ 0, & \mu_i[0]<\theta\end{cases} \tag{5}$$

阈值 θ 代表最低可接受的能量水平，低于此阈值的高斯分量将被抑制，即低能量模式被抑制。通过这种方式，系统可以在不破坏语音结构的前提下，减少模型对低能量模式的依赖，从而增强泛化能力。阈值大小的设定会影响系统对语音帧匹配程度的判断，从而影响系统整体检测性能，将在实验部分详细分析与讨论。

1.4 后端分类网络

后端分类网络是 TGSM 合成伪造语音检测系统的重要组成部分，它根据前端输入的特征参数对语音进行真伪判决。TGSM 后端分类网络由二维卷积、图注意力网络（graph attention network, GAT）和分类判别等模块组成，整体框架如图 3 所示，其中，H 表示输入图注意力网络模块的特征表示，即上一级的输出；Z 表示经图注意力机制更新后的输出特征，是下一级的输入。

每段语音的 LGP 特征序列经过前端处理后，为了确保不同时长的特征序列符合网络输入规范，将特征序列通过补零或裁剪等手段统一至 400 帧，这样可以保证系统基于固定维度进行扩展训练。输入的 LGP 特征序列经过一个由 4 个卷积残差块组成的浅层网络，每个残差块主要由 3×3 卷积核的二维卷积层、配合批归一化（batch normalization, BN）层和 SeLU 激活函数组成，该结构合理地表达了对特征静态分布的控制，优化了训练的稳定性。在 TGSM 中，二维卷积操作的引入具有非常重要的作用。与传统的一维卷积操作仅沿时间轴滑动不同，TGSM 在 LGP 特征的二维结构上（时间×高斯分量）应用 3×3 的二维卷积核，能够在局部区域内同时建模相邻时间帧与不同高斯分量之间的耦合关系。这种方式不仅能显式捕捉同一时间帧内不同高斯分量的协同变化，还能够识别相邻帧之间的跨分量动态模式，有效增强特征的表达能力。

在卷积编码之后，TGSM 分别沿时间维度和高斯分量维度执行最大池化操作，得到时间和分量特征图，并在每个维度上构建图结构，使用 GAT[18] 学习时间序列内部和分量内部的依赖关系。最终，

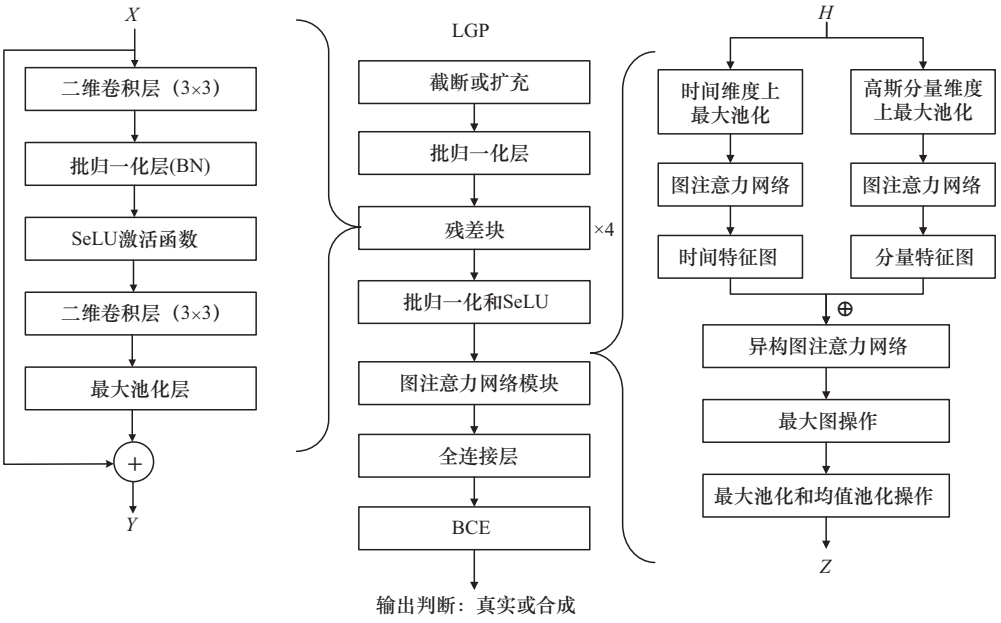


图 3 TGSM 后端分类网络整体框架

两类特征图融合为一个包含时间节点和高斯分量节点的异构图，通过异构图注意力网络（heterogeneous GAT, HGAT）[19]在跨域节点之间建立连接，提取代表性节点，这样既保留了时序和分量各自的结构，又利用跨域关系提升了对语音统计结构的表达能力和模式识别能力。

与AASIST、RawGAT-ST等同样引入图结构建模的合成伪造语音检测方法不同，TGSM并非基于时频表征直接构建图节点，而是以统一GMM提取的LGP特征为基础，在时间维度与高斯分量维度上分别构图，并进一步通过HGAT建模跨维度交互关系。该设计能够更有针对性地刻画语音统计结构在不同维度上的协同变化，从而有效缓解静音与低能量语音段分布不一致所带来的统计偏置。

在图注意力网络处理后，特征通过最大图池化操作，选择强分类性节点表征提取。之后通过最大池化和均值池化操作维持统一维度，将各图表征聚合为一个组合特征向量，输入全连接层进行分类。模型通过二元交叉熵（binary cross entropy, BCE）损失函数进行训练优化，其表达式为

$$L=-\frac{1}{K}\sum_{k=1}^{K}\left[y_k\ln\left(p_k\right)+\left(1-y_k\right)\ln\left(1-p_k\right)\right] \quad (6)$$

其中，K为语音样本数，y_k为第k个样本的真实标签，p_k为模型预测其为真实语音的概率。通过最小化该损失函数，模型能够对预测结果与真实标签之间的差异进行度量，从而引导网络学习区分真实与伪造语音的判别特征[20]。

2　实验与结果分析

2.1　数据集与实验设置

本文实验采用ASVspoof 2021[21]数据集中的逻辑访问（logical access, LA）子集进行训练与测试。该数据集专为语音欺骗检测任务构建，其中的真实与合成语音是该实验拟采用的语料，后者是利用神经波形模型、声码器、波形拼接等17种不同算法生成的高仿真欺骗合成语音[22]。数据集划分为训练集、开发集和评估集，各子集之间无重复样本，确保了模型训练与测试的独立性。其中训练集包括2 580条真实语音和15 200条合成语音，用于模型训练；开发集包括2 548条真实语音和14 864条合成语音，用于模型超参数调整优化；评估集则包括7 355条真实语音和49 140条合成语音，用于最终性能测试。

本文将高斯混合模型的分量数N设置为512，主要考虑了特征分辨能力与计算效率之间的平衡。一方面，较高的分量数能够提供更精细的语音特征空间分布概率建模能力，捕捉更丰富的局部统计特征，从而提升对伪造语音微弱差异的表征能力。另一方面，512个分量的设置在GMM大小与下游神经网络结构之间实现了较好的匹配，使时间维度（固定为400帧）与分量维度在尺度上具有一定的可比性，便于后续二维建模模块在时间–分量域空间中进行协同特征提取与结构建模。

实验在NVIDIA RTX 4090 GPU上进行训练，可以确保模型在大规模数据集上的高效计算能力。训练过程采用Adam优化器，该优化算法结合了一阶和二阶动量估计，以提升梯度更新的稳定性和收敛速度[23]。其中，学习率初始值设定为0.000 1，并使用余弦退火调度策略进行动态调整，使学习率在训练后期逐渐衰减至最小值0.000 005，以增强模型收敛的稳定性。Adam优化器超参数β_1和β_2分别设定为0.9和0.999，以控制梯度的一阶和二阶动量估计累积，防止梯度振荡。权重衰减参数设定为0.000 1，用于抑制过拟合，提高模型的泛化能力。

在训练过程中，批大小设定为72，以平衡计算效率与模型性能，并保证梯度更新的稳定性。整个训练过程持续100个Epoch，在此期间，模型不断优化参数以提升合成语音检测的性能。本文实验环境的优化配置旨在确保模型在高效计算资源支持下，达到最优的收敛状态。

2.2　性能评价指标

为了全面评估语音欺骗检测模型的性能，本文采用等错误率（equal error rate, EER）与串联检测代价函数（tandem detection cost function, t-DCF）作为主要评价指标。

EER是语音反欺骗系统中最常用的评价指标之一，定义为系统在错误接受率（false acceptance rate, FAR）与错误拒绝率（false rejection rate, FRR）相等时的错误率，即EER = FAR = FRR。该指标可用于衡量系统的整体判别能力，EER值越低，说明系统对真实与合成语音的判别能力越强。

然而，EER未考虑实际应用中ASV与反欺骗模块（spoofing countermeasure, CM）之间的协同作用。为此，ASVspoof挑战赛引入t-DCF指标，

以反映联合系统在真实场景下的性能表现[24]。t-DCF综合考虑了错误拒绝与接受的成本，并结合真实语音先验概率进行加权，能够更灵活地适应不同应用需求[25]，其定义为

$$\text{t-DCF}=C_{\text{FRR}}^{\text{CM}}\pi_{\text{tar}}P_{\text{FRR}}^{\text{CM}}+C_{\text{FAR}}^{\text{CM}}\pi_{\text{spoof}}P_{\text{FAR}}^{\text{CM}} \tag{7}$$

其中，$C_{\text{FRR}}^{\text{CM}}$和$C_{\text{FAR}}^{\text{CM}}$分别表示错误拒绝真实语音的代价和错误接受欺骗语音的代价，可根据应用场景调整；π_{tar}表示目标说话人语音的先验概率，π_{spoof}表示伪造语音的先验概率；$P_{\text{FRR}}^{\text{CM}}$和$P_{\text{FAR}}^{\text{CM}}$分别表示语音欺骗检测系统的错误拒绝率和错误接受率。本文严格遵循ASVspoof 2021评测任务的官方配置进行计算。t-DCF的具体参数设置如下：伪造语音先验概率$\pi_{\text{spoof}}=0.05$；目标说话人语音的先验概率$\pi_{\text{tar}}=0.940\,5$；CM漏报代价$C_{\text{FRR}}^{\text{CM}}=1$，误报代价$C_{\text{FAR}}^{\text{CM}}=10$。

因此，EER能有效反映模型的整体判别能力，而t-DCF能进一步评估其在实际系统部署下的适应性与实用性，二者互为补充，均在实验中被采用。

2.3　参数设置

为了分析高斯混合模型中分量数N对TGSM性能的影响，并为分量维度的选取提供更加系统的实验依据，本文在ASVspoof 2021数据集上对不同分量数进行了对比实验，分别设置N=64,128,256,512,1 024，在保持其他网络结构与训练策略不变的情况下，评估模型在EER和t-DCF指标上的性能变化，实验结果如表1所示。

表1　分量数N不同情况下的性能对比

分量数N	EER	t-DCF
64	3.00‰	9.8×10^{-3}
128	2.72‰	9.2×10^{-3}
256	3.26‰	10.4×10^{-3}
512	**2.04‰**	$\mathbf{6.7\times10^{-3}}$
1 024	3.53‰	11.6×10^{-3}

从实验结果可以观察到，当N较小时，模型性能相对较差，说明分量数小不足以充分刻画语音统计特征的多样性；当N增加至256时，性能有所提升，但仍未达到最优。

当分量数设置为N=512时，模型在EER和t-DCF两个指标上均取得最优结果，表明该配置能够在表达能力与模型复杂度之间取得较好平衡。进一步增加分量数至1 024后，模型性能反而出现下降，这是由于分量维度过高引入冗余统计分量，从而加重模型学习负担并削弱泛化能力。

结合本文关注的静音与低能量语音段分布不一致问题，可以认为适当数量的高斯分量有助于对不同能量模式进行细粒度建模，而过少或过多的分量都会削弱这种统计建模优势。因此，本文实验中统一采用N=512作为高斯混合模型的分量数配置。

同时，为了优化系统的检测性能，实验需对能量阈值参数θ进行设置，选取$\theta\in\{-40,-35,-30,-25,-20\}$5个不同的数值，对比不同阈值下模型在各种攻击类型上的EER与t-DCF性能表现，实验结果如表2和图4所示，表中加粗数值代表按行取最优值。

表2　不同阈值下各种攻击类型的EER

攻击类型	$\theta=-40$	$\theta=-35$	$\theta=-30$	$\theta=-25$	$\theta=-20$
A07	1.22‰	1.39‰	**0.99‰**	1.22‰	1.39‰
A08	3.02‰	**2.04‰**	3.26‰	5.70‰	5.47‰
A09	0.99‰	1.22‰	1.05‰	1.22‰	**0.58‰**
A10	3.43‰	**2.04‰**	2.44‰	3.02‰	2.68‰
A11	2.85‰	**2.21‰**	3.84‰	3.43‰	3.26‰
A12	1.22‰	1.39‰	1.39‰	1.80‰	**0.99‰**
A13	1.05‰	1.39‰	0.82‰	0.82‰	**0.58‰**
A14	**1.80‰**	1.87‰	1.87‰	2.21‰	**1.80‰**
A15	**2.04‰**	**2.04‰**	2.27‰	2.44‰	2.68‰
A16	4.89‰	4.48‰	**3.50‰**	4.89‰	4.72‰
Overall	2.46‰	**2.04‰**	2.34‰	3.13‰	3.26‰

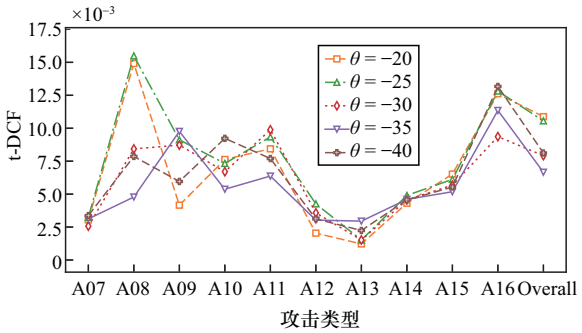


图4　不同阈值下各种攻击类型的t-DCF

总体来看，当$\theta=-35$时，EER和t-DCF两个指标均取得最优结果，分别是2.04‰和6.66×10^{-3}，说明该阈值$\theta=-35$的设定，在保留关键语音特征与抑制低能量干扰之间实现了良好平衡，设置合理的阈

值可有效增强模型对真实与伪造语音的判别能力。因此，本文后续实验中，TGSM中的θ参数取值为−35。

进一步分析不同攻击类型的性能表现可知，对于A08、A10、A11等攻击类型，较低的阈值（如θ=−40或−35）往往带来更优的检测结果。这些伪造语音大多由基于复杂的神经网络结构或融合模型（如GAN与vocoder）生成，其语音波形中常包含高仿真度的低能量噪声或微弱过渡段，容易混淆判别模型[26]。适当的能量筛选能够在保留主要语音特征的同时，有效削弱伪造语音的“伪真性”成分，提升模型对边界样本的识别能力。

在A09、A12、A13等依赖频谱合成的传统声码器攻击类型中，较低的阈值对检测表现的提升相对有限。这是因为这些语音样本本身能量分布较集中，不需要太强的低能量抑制策略就能很好地削弱低能量区域的干扰，让模型更聚焦于承载语音内容的主要区段，从而提高模型的分辨能力。

从上述数据和分析可以看出，阈值机制在不同攻击类型下均展现出一定的适应能力，合理设定的θ不仅能有效缓解静音分布不均对模型判别造成的干扰，也为系统引导注意力聚焦于高能量区域，进一步增强了特征的可分性与检测性能。

2.4　与现有方法对比

为了评估TGSM的优越性，将其与主流语音欺骗检测模型进行了性能对比。为确保实验的可控性、公平性与稳定性，所有模型在本地计算环境中部署和运行。EER和t-DCF的实验结果分别如表3和表4所示，其中，TGSM和ResNet(1d)[27]的输入特征为LGP，AASIST[28]、RawGAT-ST[29]、RawNet2[30]和AASIST2的输入特征为原始语音波形。

在训练配置方面，为保证不同模型之间对比的公平性，本文在可行范围内对各模型采用统一的训练策略。除模型结构差异外，所有模型采用Adam优化器，并在相同训练数据划分下进行训练。默认批大小设置为72，当模型规模较大、显存占用较高（如AASIST与AASIST2）时，适当降低批大小以满足显存约束。所有模型训练100个Epoch，并采用早停策略防止过拟合。对于对比模型的学习率、网络深度及正则化参数设置，本文严格遵循原论文及其公开源码中的推荐配置，仅在必要情况下根据硬件资源进行微调，以确保实验结果的可复现性与对比的公平性。

从表3可以看出，TGSM在整体检测性能上优于当前主流的语音伪造检测模型。TGSM的总体EER为2.04‰，显著优于AASIST（2.32‰）、RawGAT-ST（3.81‰）以及RawNet2（12.10‰）等模型，说明本文方法在真实与伪造语音的区分能力上具有明显优势。其中，相较于AASIST，TGSM将等错误率从2.32‰降低至2.04‰，性能提升了12.07%。相较于最新提出的AASIST2，TGSM在多数攻击类型及整体性能上表现出具有竞争力的检测能力，表明本文方法在不依赖大规模原始波形建模的情况下，依然能够有效刻画伪造语音的统计差异特征。同时，与采用相同LGP特征输入的一维ResNet(1d)相比，EER下降约70.61%，TGSM在特征建模与判别性能上均展现出更优表现，验证了本文提出的二维建模策略在复杂语音分布方面具有明显优势，这方面也会在后续的t分布随机邻域嵌入（t-distributed stochastic neighbor embedding, t-SNE）可视化分析中得到进一步验证。

表3　各模型在不同攻击类型下的EER对比

攻击类型	TGSM	AASIST	RawGAT-ST	ResNet (1d)	RawNet2	AASIST2
A07	1.39‰	2.45‰	4.24‰	2.04‰	11.82‰	**0.98‰**
A08	2.04‰	**1.39‰**	2.85‰	4.65‰	16.88‰	3.43‰
A09	1.22‰	**0**	0.58‰	0.17‰	1.22‰	0.23‰
A10	**2.04‰**	3.10‰	5.53‰	3.91‰	13.45‰	3.02‰
A11	**2.21‰**	2.61‰	3.09‰	**2.21‰**	11.41‰	2.61‰
A12	1.39‰	2.21‰	4.07‰	**0.82‰**	16.70‰	1.80‰
A13	1.39‰	1.05‰	2.28‰	0.99‰	7.57‰	**0.81‰**
A14	1.87‰	1.05‰	1.05‰	11.99‰	8.73‰	**0.81‰**
A15	2.04‰	2.04‰	3.09‰	10.59‰	10.83‰	**0.81‰**
A16	4.48‰	3.67‰	6.11‰	6.69‰	15.90‰	**1.05‰**
Overall	**2.04‰**	2.32‰	3.81‰	6.94‰	12.10‰	2.18‰

表4列出了各模型在不同攻击类型下的t-DCF对比情况，虽然AASISIT和ResNet(1d)在部分攻击类型下得到了相较于TGSM更佳的性能，但在整体上依然是TGSM获得了最优的检测性能6.66×10^{-3}，相较于AASIST（7.58×10^{-3}），在t-DCF性能上提升了12.14%，相较于RawGAT-ST（12.06×10^{-3}）性能提升了44.77%。

从各种攻击类型的对比实验结果来看，TGSM在A10和A11攻击场景中均实现了最低EER和t-DCF，其中A10攻击类型下EER和t-DCF仅为

表 4　各模型在不同攻击类型下的 t-DCF 对比

攻击类型	TGSM	AASIST	RawGAT-ST	ResNet(1d)	RawNet2	AASIST2
A07	**3.12×10^{-3}**	5.37×10^{-3}	11.04×10^{-3}	5.79×10^{-3}	32.89×10^{-3}	3.18×10^{-3}
A08	4.78×10^{-3}	**3.85×10^{-3}**	7.17×10^{-3}	12.39×10^{-3}	43.82×10^{-3}	11.43×10^{-3}
A09	9.76×10^{-3}	**0**	4.39×10^{-3}	0.81×10^{-3}	10.63×10^{-3}	10.54×10^{-3}
A10	**5.37×10^{-3}**	8.96×10^{-3}	14.64×10^{-3}	9.38×10^{-3}	37.11×10^{-3}	9.51×10^{-3}
A11	6.37×10^{-3}	6.94×10^{-3}	8.24×10^{-3}	**5.54×10^{-3}**	32.40×10^{-3}	8.09×10^{-3}
A12	3.03×10^{-3}	4.77×10^{-3}	11.16×10^{-3}	**1.48×10^{-3}**	47.26×10^{-3}	5.76×10^{-3}
A13	2.95×10^{-3}	2.54×10^{-3}	5.54×10^{-3}	**2.03×10^{-3}**	21.76×10^{-3}	2.96×10^{-3}
A14	4.60×10^{-3}	**2.45×10^{-3}**	2.86×10^{-3}	32.48×10^{-3}	23.24×10^{-3}	1.97×10^{-3}
A15	**5.18×10^{-3}**	5.43×10^{-3}	8.28×10^{-3}	29.73×10^{-3}	30.01×10^{-3}	2.49×10^{-3}
A16	11.35×10^{-3}	**8.96×10^{-3}**	14.81×10^{-3}	17.44×10^{-3}	44.27×10^{-3}	3.39×10^{-3}
Overall	**6.66×10^{-3}**	7.58×10^{-3}	12.06×10^{-3}	22.45×10^{-3}	41.02×10^{-3}	7.23×10^{-3}

2.04‰ 和 5.37×10^{-3}，显著低于其他模型。这表明 TGSM 对多样化语音伪造技术具有良好的适应能力，在多种攻击场景下均具备较强的泛化能力。

为了进一步直观分析模型对真实与伪造语音的区分能力，引入 t-SNE 方法对高维特征进行降维可视化分析。t-SNE 是一种非线性降维技术[31]，能够在二维空间中保持高维数据的局部结构，使同类样本聚集、异类样本分散，广泛应用于深度特征可视化任务。本文使用 t-SNE 分别展示不同模型下真实与合成语音在特征空间中的分布情况，实验结果如图 5 所示。

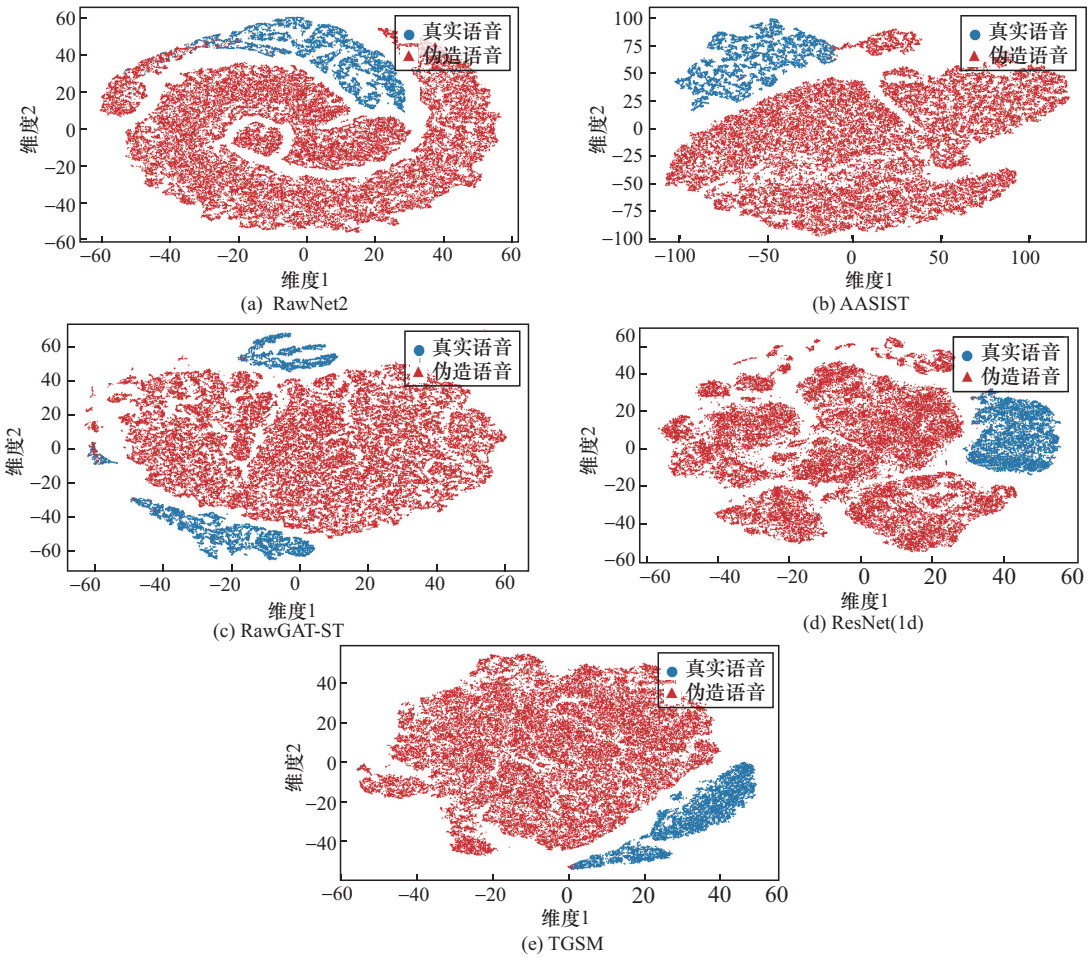


图 5　t-SNE 可视化分析对比

从图5可以看出，RawNet2的特征分布呈现出大面积的混合区域，真实与伪造语音在多个区域高度重叠，表明其提取的特征判别性较弱，难以有效分隔不同类型的语音。RawGAT-ST虽在结构上引入了注意力机制，但从图5(c)中可见，其将真实语音分为两个明显不连续的簇，特征一致性欠佳，会影响分类边界的稳定性。相比之下，AASIST模型的特征聚类效果较好，真实与伪造语音大致形成两个独立的簇，但也出现了一定程度的混淆。同时也可以看到，ResNet(1d)中的真实与伪造语音边界模糊，且语音整体特征分布较为稀疏和离散。这说明仅基于一维LGP特征序列进行建模难以充分捕捉时间与高斯分量之间的协同变化信息，导致特征表达能力受限。相比之下，TGSM采用时间–高斯分量二维建模策略，从t-SNE可视化分析中可见，TGSM所提取的真实与伪造语音特征簇分布清晰、界限分明，几乎无明显混淆区域，验证了二维建模策略的有效性。可视化与定量实验结果相互印证，充分说明TGSM在语音伪造检测任务中具备良好的判别能力与泛化性。

2.5　消融实验

2.5.1　能量阈值

为验证所提能量阈值抑制机制在语音伪造检测中的有效性，本文设计了消融实验，对比了无阈值抑制机制的系统（w/o θ）与阈值为θ= −35时系统的检测性能，具体实验结果如表5所示。

表5　　能量阈值消融实验对比

攻击类型	w/o θ		θ=−35	
	EER	t-DCF	EER	t-DCF
A07	**1.39‰**	3.46×10^{-3}	**1.39‰**	$\mathbf{3.12\times10^{-3}}$
A08	6.69‰	18.18×10^{-3}	**2.04‰**	$\mathbf{4.78\times10^{-3}}$
A09	**1.05‰**	$\mathbf{7.90\times10^{-3}}$	1.22‰	9.76×10^{-3}
A10	3.02‰	8.55×10^{-3}	**2.04‰**	$\mathbf{5.37\times10^{-3}}$
A11	3.02‰	7.68×10^{-3}	**2.21‰**	$\mathbf{6.37\times10^{-3}}$
A12	1.87‰	4.61×10^{-3}	**1.39‰**	$\mathbf{3.03\times10^{-3}}$
A13	1.46‰	3.51×10^{-3}	**1.39‰**	$\mathbf{2.95\times10^{-3}}$
A14	**1.87‰**	4.76×10^{-33}	**1.87‰**	$\mathbf{4.60\times10^{-3}}$
A15	**0.20‰**	5.54×10^{-3}	2.04‰	$\mathbf{5.18\times10^{-3}}$
A16	**3.26‰**	$\mathbf{8.71\times10^{-3}}$	4.48‰	11.35×10^{-3}
Overall	3.34‰	11.31×10^{-3}	**2.04‰**	$\mathbf{6.66\times10^{-3}}$

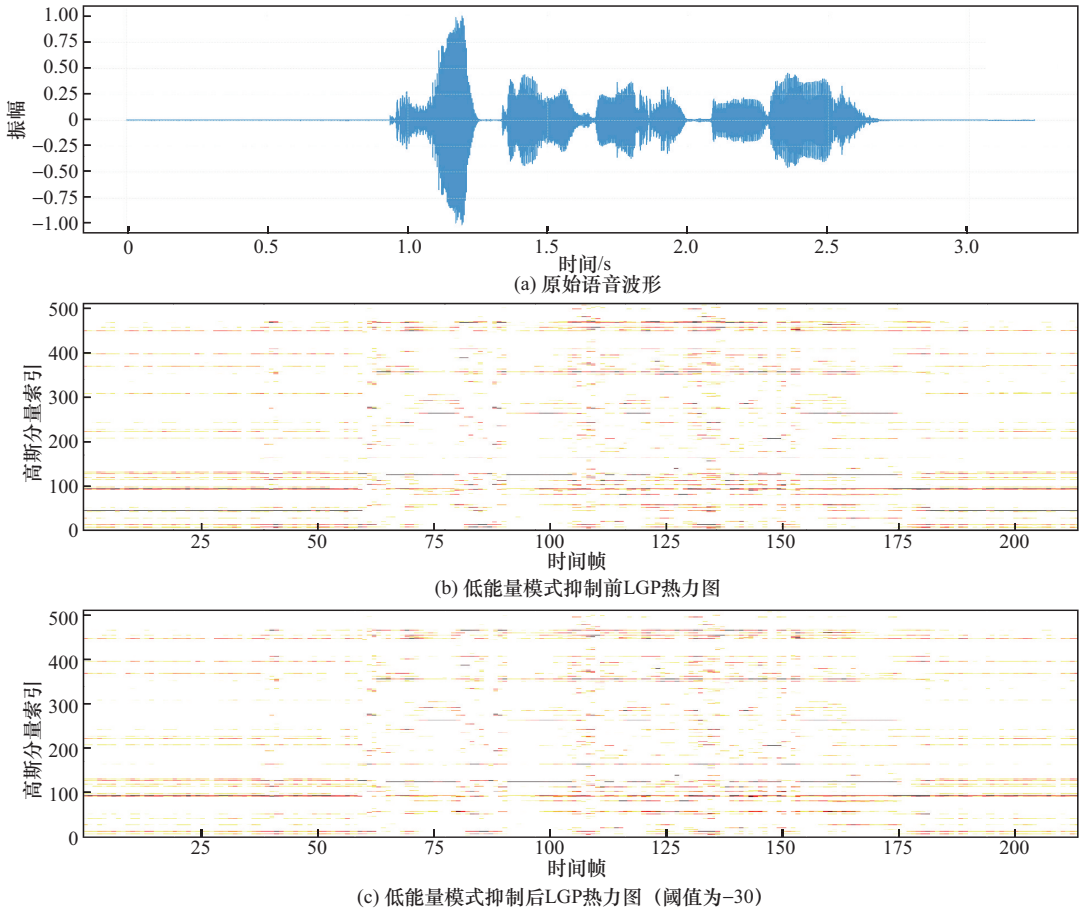
从总体数据来看，加入阈值后，系统的EER从3.34‰降至2.04‰，降低了38.92%，t-DCF从11.31×10^{-3}显著下降至6.66×10^{-3}，降低了41.11%，表明该机制在系统整体检测性能上具有明显提升作用。

从各攻击类型来看，大部分攻击类型在引入阈值后均表现出EER和t-DCF的不同程度下降。例如，A08的EER从6.69‰降至2.04‰，降低了69.51%，A12的t-DCF从4.61×10^{-3}降至3.03×10^{-3}，表明引入能量阈值对于降低静音干扰、增强真实与伪造语音的区分能力具有非常重要的作用。

为了进一步分析能量阈值抑制机制的机理，随机从语音库中选取一段语音，画出该段语音的LGP特征序列在经能量阈值抑制前后的对比图，如图6所示。从上至下依次为原始语音波形、未经能量阈值抑制处理的LGP特征序列，以及经过能量阈值抑制（阈值为−35 dB）后的LGP特征序列。由图6(b)可知，静音段和有声段在不同高斯分量上的LGP特征值拥有显著差异。在静音段，部分对应于低能量模式的分量表现出高强度的轨迹响应，数值较大；在有声段，同一分量的响应则显著衰减，基于趋于0。这是因为这些分量属于低能量模式，静音段在这些分量上的后验概率大，对应的LGP数值就大，对系统的性能有较大的影响，而有声段在这些分量上的后验概率小，对应的LGP数值就小。在引入能量阈值抑制后，将低于阈值的LGP特征进行抑制，有效减少了静音段中虚假强响应的出现，如图6(c)所示。有声段在这些分量上LGP数值极小，即使被抑制掉，也几乎不会受到影响，从而降低了模型对静音段的干扰，增强了模型对有声段关键信息的关注度，提升了系统的检测能力。

此外，本文还对TGSM w/o θ进行了t-SNE可视化分析，结果如图7所示。由图7与图5(e)对比可知，尽管两个系统均能在整体上形成较为清晰的真实与伪造语音分布边界，但TGSM w/o θ在伪造语音的主要聚集区域末端仍存在较多混杂现象。相比之下，引入能量阈值抑制机制后，该部分混淆显著减少，进一步验证了能量阈值抑制机制在抑制干扰和增强特征区分性方面的有效性。

通过上述消融实验可见，能量阈值抑制机制能有效减少静音段对模型分类能力的干扰，提升系统检测性能，是TGSM系统具有优越性能的重要保证。



(a) 原始语音波形

(b) 低能量模式抑制前LGP热力图

(c) 低能量模式抑制后LGP热力图（阈值为−30）

图6　能量阈值抑制效果

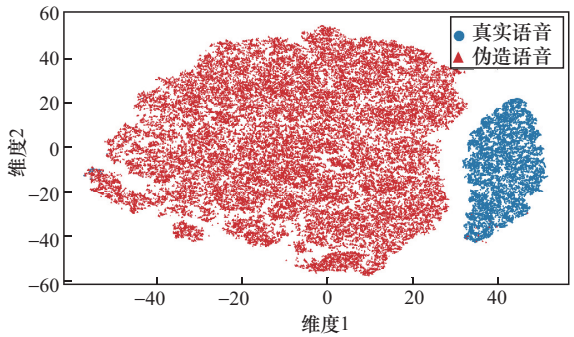


图7　TGSM w/o θ的t-SNE可视化分析

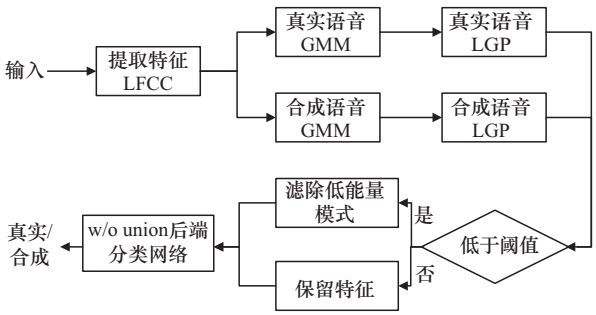


图8　采用TGSM w/o union模型构建的伪造语音检测系统框架

2.5.2　统一建模

为了进一步验证统一建模策略在语音伪造检测中的有效性，本文设计了统一建模消融实验。对照模型TGSM w/o union采用独立建模策略，即针对真实与伪造语音分别建立独立的GMM，构建两个通道。两个通道分别提取对应的LGP特征，并通过结构一致但参数不共享的残差卷积模块与图建模模块进行特征学习，最后在后端分类网络拼接融合，通过全连接层完成分类。该系统的整体框架如图8所示。

消融实验分别在“无能量阈值（no）”以及5个不同能量阈值（θ= −40, −35, −30, −25, −20）的情况下进行，图9显示了TGSM和对应的消融模型w/o union在EER和t-DCF检测性能上的差异。从图9可以看出，在未引入能量阈值时，w/o union的性能要优于TGSM，但随着能量阈值抑制机制的引入，TGSM的EER与t-DCF两项指标在整体上表现出更优性能，特别是在θ=−35时均达到最优值，即EER为2.04‰，t-DCF降至6.66×10^{-3}。这一结果表明，TGSM统一建模下的特征分布更具结构一致

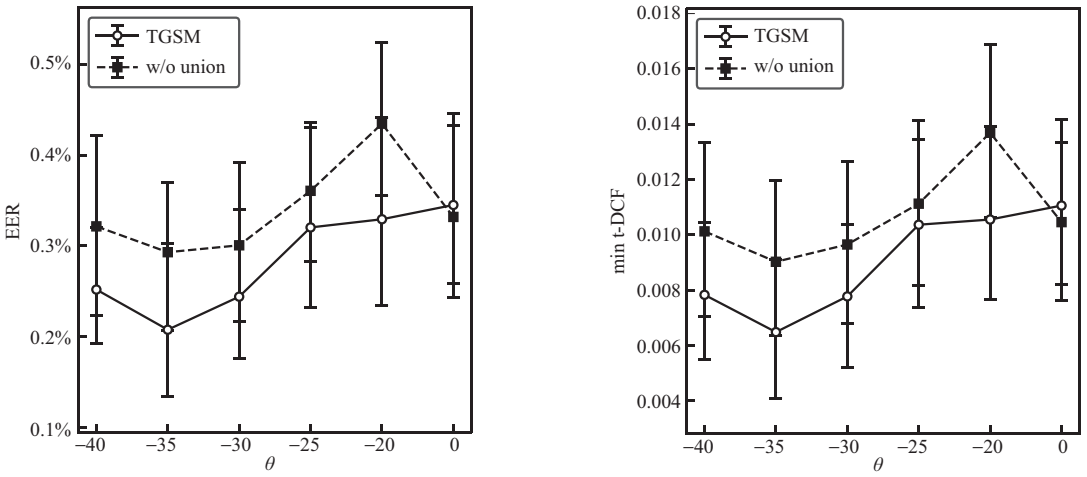


图 9　TGSM 统一建模消融实验性能对比

性，配合能量阈值抑制机制，能够更有效地抑制低能量干扰，增强判别性特征的建模能力。同时，TGSM 所采用的参数共享策略在保持性能优势的同时，也具备更高的计算效率与更强的泛化能力。因此，TGSM 不仅在判决准确度上优于独立建模结构的 w/o union，同时在实际部署中也更具实用价值。由图 9 还可以看出，不同阈值下整体性能呈现“先降后升”的变化规律，在 $\theta = -35$ 附近达到最优。这一现象从侧面验证了能量阈值抑制机制的有效性，过低的阈值无法完全过滤冗余信息，而过高的阈值又容易丢失关键特征，唯有在合理设定下，才能实现系统性能的协同提升。

为进一步从特征分布角度分析统一建模与独立建模策略之间的差异，本文对 TGSM w/o union 模型所学习到的高层特征进行了 t-SNE 可视化分析，其结果如图 10 所示。该可视化是基于后端分类网络输出的判别特征，通过非线性降维方式将高维特征映射至二维空间，以直观观察真实与伪造语音在特征空间中的分布情况。

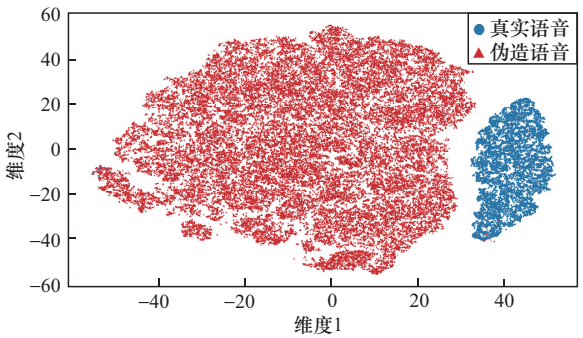


图 10　TGSM w/o union 的 t-SNE 可视化分析

对比图 10 与图 5(e)可以发现，在采用独立建模策略的 TGSM w/o union 模型中，真实与伪造语音在主要聚集区域的边缘位置存在更为明显的重叠与混杂现象，表明不同通道独立学习得到的统计空间在融合阶段仍存在一定的分布不一致性。相比之下，采用统一建模策略的 TGSM 在特征空间中呈现出更加集中且边界更为清晰的分布形态，有助于突出高判别性区域之间的差异，从而为后续分类阶段形成更稳定的决策边界提供支持。

从上述分析可以看出，TGSM 采用统一建模策略，将真实与伪造语音共同输入同一个 GMM，从而构建共享的高斯空间表示。它不仅避免了通道冗余带来的计算开销，还提升了模型对数据分布整体的建模能力，增强了系统的性能。

2.5.3　后端分类模块消融实验

为进一步分析 TGSM 后端分类网络中各模块对整体检测性能的贡献，并验证所提出结构设计的必要性，本文在 ASVspoof 2021 数据集上设计了多组消融实验，分别考察二维卷积、时间图、高斯分量图以及 HGAT 对模型性能的影响。所有消融实验在保持前端特征提取方式、训练策略及能量阈值设置一致的条件下进行，仅对后端结构进行调整。

表 6 给出了不同后端分类模块组合配置下 TGSM 在 EER 与 t-DCF 两个指标上的检测性能，其中，√表示该模块被保留在当前模型配置中，×表示该模块被移除或未被采用，1#为完整 TGSM 后端结构，包含二维卷积、时间图、高斯分量图及

HGAT模块，2#~5#分别为移除或仅保留部分模块，以分析单一结构或简化结构下模型性能的变化趋势。

表6 后端分类模块消融实验性能对比

实验编号	二维卷积	时间图	高斯分量图	HGAT	EER	t-DCF
1#	√	√	√	√	**2.04‰**	**6.7×10^{-3}**
2#	√	×	×	×	3.40‰	10.9×10^{-3}
3#	√	√	×	×	3.40‰	11.4×10^{-3}
4#	√	×	√	×	3.13‰	10.6×10^{-3}
5#	√	√	√	×	3.67‰	11.9×10^{-3}

从实验结果可以观察到，当同时引入二维卷积与时间图、高斯分量图并通过HGAT进行跨域融合时，模型能够取得最优性能（EER为2.04‰，t-DCF为6.7×10^{-3}）。当仅保留二维卷积而不引入图结构建模时（2#），模型性能明显下降，说明仅依赖局部卷积特征难以充分刻画语音的全局统计关系。

进一步地，对比3#与4#可以发现，仅在时间域或高斯分量域构建图结构均无法达到完整模型的性能水平，表明两种图结构在建模语音时序动态与统计分量分布方面具有互补作用。此外，当引入时间图与高斯分量图但不采用HGAT进行跨域建模时（5#），模型性能同样出现退化，验证了HGAT在融合不同结构特征和增强判别能力方面的关键作用。

综上所述，消融实验结果表明，TGSM后端分类网络中各模块并非简单叠加，而是在时间维度与高斯分量维度的协同建模框架下相互配合，共同提升模型对复杂语音伪造模式的建模能力与检测性能。

3 结束语

为了增强合成伪造语音检测系统的性能，降低训练语料中静音分布不一致对系统性能的影响，提出了一种采用能量阈值抑制机制、具有统一建模架构的合成伪造语音检测方法TGSM。实验结果表明，在EER和t-DCF两方面，TGSM都具有更好的性能，展现出更高的准确率和泛化性。该方法通过统一的高斯混合模型对真实与伪造语音进行建模，有效避免了独立建模造成的信息割裂与参数冗余问题，提升了语音特征表示的准确性与有效性。其引入的能量阈值抑制机制显著抑制了静音、背景噪声等低能量区域对检测性能的干扰，增强了模型在不同语料和场景下的泛化性。同时，结合时间–高斯分量二维结构建模策略与异构图注意力机制，TGSM能够充分挖掘跨域特征的内在关联，在保持低计算复杂度的前提下实现高效且准确的合成伪造语音检测。

参考文献：

[1] 杨震, 王天朗, 郭海燕, 等. 跨域注意力特征融合的说话人确认方法[J]. 通信学报, 2023, 44(8): 89-98.
Yang Z, Wang T L, Guo H Y, et al. Speaker verification method based on cross-domain attentive feature fusion[J]. Journal on Communications, 2023, 44(8): 89-98.

[2] Song Z D, Cai H, Chen X, et al. Improving speaker verification robustness with multilingual phonetic information and feature decorrelation[J]. IEEE Transactions on Audio, Speech and Language Processing, 2025, 33: 3494-3507.

[3] 简志华, 章子旭. 采用表示分离自编码器的任意说话人语音转换[J]. 通信学报, 2024, 45(2): 162-172.
Jian Z H, Zhang Z X. Any-to-any voice conversion using representation separation auto-encoder[J]. Journal on Communications, 2024, 45(2): 162-172.

[4] Gupta P, Patil H A, Guido R C. Vulnerability issues in automatic speaker verification (ASV) systems[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2024, 2024: 10.

[5] Hanilçi C. Linear prediction residual features for automatic speaker verification anti-spoofing[J]. Multimedia Tools and Applications, 2018, 77(13): 16099-16111.

[6] Javed A, Malik K M, Malik H, et al. Voice spoofing detector: a unified anti-spoofing framework[J]. Expert Systems with Applications, 2022, 198: 116770.

[7] Tracey B, Volfson D, Glass J, et al. Towards interpretable speech biomarkers: exploring MFCCs[J]. Scientific Reports, 2023, 13: 22787.

[8] Zhang Y X, Li Z, Lu J Z, et al. The impact of silence on speech anti-spoofing[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 3374-3389.

[9] Tian X H, Xiao X, Chng E S, et al. Spoofing speech detection using temporal convolutional neural network[C]//Proceedings of the 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Piscataway: IEEE Press, 2016: 1-6.

[10] Gomez-Alanis A, Peinado A M, Gonzalez J A, 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.

[11] Zhang Q, Zhang X W, Sun M, et al. A transformer-based deep learning approach for recognition of forgery methods in spoofing speech attribution[J]. Applied Soft Computing, 2025, 171: 112798.

[12] Khan A, Malik K M, Ryan J, et al. Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures[J]. Artificial Intelligence Review,

2023, 56(S1): 513-566.

[13] Müller N M, Dieckmann F, Czempin P, et al. Speech is silver, silence is golden: what do ASVspoof-trained models really learn?[PP]. arXiv (2021-06-23) [2025-11-12]. arXiv:arXiv.2106.12914.

[14] Shim H J, Sahidullah M, Jung J W, et al. Beyond silence: bias analysis through loss and asymmetric approach in audio anti-spoofing[PP]. V2. (2024-08-25) [2025-11-12]. arXiv:arXiv.2406.17246.

[15] 龙华, 杨明亮, 邵玉斌. 基于特征流融合的带噪语音检测算法[J]. 通信学报, 2020, 41(4): 134-142.
Long H, Yang M L, Shao Y B. Noisy voice detection algorithm based on feature stream fusion[J]. Journal on Communications, 2020, 41(4): 134-142.

[16] Lei Z C, Yang Y G, Liu C H, et al. Siamese convolutional neural network using Gaussian probability feature for spoofing speech detection[C]//Proceedings of the Interspeech 2020. Piscataway: IEEE Press, 2020: 1116-1120.

[17] Lei Z C, Yan H, Liu C H, et al. Two-path GMM-ResNet and GMM-SENet for ASV spoofing detection[C]//Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 6377-6381.

[18] Liu Z Y, Zhou J. Graph attention networks[M]//Introduction to Graph Neural Networks. Berlin: Springer, 2020: 39-41.

[19] Wang X, Ji H Y, Shi C, et al. Heterogeneous graph attention network[C]// Proceedings of the World Wide Web Conference. New York: ACM Press, 2019: 2022-2032.

[20] Al-Tairi H, Javed A, Khan T, et al. DeepLASD countermeasure for logical access audio spoofing[J]. Scientific Reports, 2025, 15: 20839.

[21] Liu X C, Wang X, Sahidullah M, et al. ASVspoof 2021: towards spoofed and deepfake speech detection in the wild[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2507-2522.

[22] Delgado H, Evans N, Kinnunen T, et al. ASVspoof 2021: automatic speaker verification spoofing and countermeasures challenge evaluation plan[PP]. V1. (2021-09-01) [2025-11-12]. arXiv:arXiv.2109.00535.

[23] Hassan E, Shams M Y, Hikal N A, et al. The effect of choosing optimizer algorithms to improve computer vision tasks: a comparative study[J]. Multimedia Tools and Applications, 2023, 82(11): 16591-16633.

[24] Guan Y, Ai Y, Li Z L, et al. Recursive feature learning from pre-trained models for spoofing speech detection[C]//Proceedings of the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2025: 1-5.

[25] Kinnunen T, Delgado H, Evans N, et al. Tandem assessment of spoofing countermeasures and automatic speaker verification: fundamentals[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2195-2210.

[26] Zhang B W, Cui H, Nguyen V, et al. Audio Deepfake detection: what has been achieved and what lies ahead[J]. Sensors, 2025, 25(7): 1989.

[27] Lei Z C, Yan H, Liu C H, et al. GMM-ResNet2: ensemble of group resnet networks for synthetic speech detection[C]//Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 12101-12105.

[28] Zhang Y X, Lu J Z, Shang Z Q, et al. Improving short utterance anti-spoofing with Aasist2[C]//Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 11636-11640.

[29] Tak H, Jung J W, Patino J, et al. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection[PP]. arXiv (2021-07-28) [2025-11-12]. arXiv: arXiv.2107.12710.

[30] Tak H, Patino J, Todisco M, et al. End-to-end anti-spoofing with RawNet2[C]//Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 6369-6373.

[31] Grisci B I, Inostroza-Ponta M, Dorn M. Assessing feature scorer results on high-dimensional datasets with t-SNE[J]. Neurocomputing, 2025, 652: 130561.

[作者简介]

简志华（1978– ），男，江西新余人，博士，杭州电子科技大学副教授、硕士生导师，主要研究方向为伪造语音检测、语音转换、语音数据隐私保护。

梁承涵（2001– ），男，湖南娄底人，杭州电子科技大学硕士生，主要研究方向为伪造语音检测、语音安全。

朱峰满（2003– ），男，浙江温州人，杭州电子科技大学硕士生，主要研究方向为伪造语音检测、智能语音处理。