计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(8): 2095−2110, 2026

面向源说话人验证的语音转换伪造数据集与模型

张国富[1,2,3] 田天才[1] 苏兆品[1,3] 王垚飞[1,3] 段宇衡[1] 周康健[1] 罗永琪[1]

[1]（合肥工业大学计算机与信息学院 合肥 230601）

[2]（合肥工业大学交叉科学研究院 合肥 230009）

[3]（智能互联系统安徽省实验室（合肥工业大学） 合肥 230009）

（zgf@hfut.edu.cn）

Voice Conversion Spoofing Dataset and Model for Source Speaker Verification

Zhang Guofu[1,2,3], Tian Tiancai[1], Su Zhaopin[1,3], Wang Yaofei[1,3], Duan Yuheng[1], Zhou Kangjian[1], and Luo Yongqi[1]

[1]（*School of Computer Science and Information Engineering*, *Hefei University of Technology*, *Hefei* 230601）

[2]（*Institute of Interdisciplinary Science*, *Hefei University of Technology*, *Hefei* 230009）

[3]（*Intelligent Interconnected Systems Laboratory of Anhui Province* (*Hefei University of Technology*), *Hefei* 230009）

Abstract With the rapid development of deep learning-based voice conversion (VC) technology, attackers can compromise the security of automatic speaker verification (ASV) systems by tampering with the source speaker's speech signal to mimic the target speaker's acoustic characteristics. The existing anti-spoofing countermeasures primarily focus on distinguishing between genuine and spoofed speech, yet underutilize the implicit source speaker identity information concealed within the spoofed speech. This limitation hinders their ability to meet practical demands in scenarios such as forensic investigation and fraud traceability. Currently, research on source speaker identity verification and discrimination remains insufficient, and existing methods suffer from limited generalisation capabilities. To address this, we construct a new voice conversion spoofing dataset specifically designed for source speaker verification, named VCSD-SSV (voice conversion spoofing dataset for source speaker verification). This dataset encompasses six VC algorithms (i.e., Seed-VC, FACodec, DDDM-VC, Diff-HierVC, TriAAN-VC, and Free-VC) and includes 400 speakers, totaling 327 677 speech samples (46 811 genuine samples and 280 866 converted samples). Building upon this, we propose a feature generalisation enhancement-based source speaker verification model, FGE-SSV (source speaker verification based on feature generalised enhancement). This model extracts multi-level source speaker timbre features through a feature enhancement module and cascaded instance-batch normalization squeeze-and-excitation blocks. Experimental results demonstrate that FGE-SSV maintains consistently low equal error rate (*EER*) (0.2%~1.0%) in within-domain results across all six VC methods in VCSD-SSV dataset and also achieves relatively low *EER* across the six VC methods in SSTC2024 dataset. Furthermore, FGE-SSV demonstrates strong generalization performance on unseen voice conversion methods, with its *EER* values outperforming those of comparative methods.

Key words source speaker verification; voice conversion spoofing dataset; feature generalised enhancement; adversarial learning; timbre feature

收稿日期: 2025-09-10; **修回日期**: 2026-03-12

基金项目: 教育部人文社会科学研究规划基金项目(24YJA870011); 国家自然科学基金项目(62302146); 安徽省自然科学基金项目(2208085MF166); 中央高校基本科研业务费专项资金(PA2025GDSK0078, PA202511SL0104)

This work was supported by the Foundation on Humanities and Social Sciences, Ministry of Education (24YJA870011), the National Natural Science Foundation of China (62302146), the Anhui Provincial Natural Science Foundation (2208085MF166), and the Fundamental Research Funds for the Central Universities (PA2025GDSK0078, PA202511SL0104).

通信作者: 苏兆品(szp@hfut.edu.cn)

摘 要 随着基于深度学习的语音转换(voice conversion, VC)技术的快速发展,攻击者能够通过篡改源说话人的语音信号来模仿目标说话人的声学特征,这严重威胁自动说话人验证(automatic speaker verification, ASV)系统的安全性。现有反欺骗防御机制主要聚焦于区分真实语音和伪造语音,却未能充分利用伪造语音中隐含的源说话人身份信息,导致其在司法取证、欺诈溯源等实际应用场景中难以满足需求。目前,针对源说话人身份验证与辨别的研究尚不充分,且已有模型泛化性能有限。基于此,构建了一个专门用于源说话人验证的新语音转换伪造数据集(voice conversion spoofing dataset for source speaker verification, VCSD-SSV)。该数据集涵盖 Seed-VC, FACodec, DDDM-VC, Diff-HierVC, TriAAN-VC, Free-VC 共 6 种语音转换算法,包含 400 个说话人,总计 327 677 条语音样本(其中真实语音 46 811 条、转换语音 280 866 条)。在此基础上,设计了一种基于特征泛化增强的源说话人验证(source speaker verification based on feature generalised enhancement, FGE-SSV)模型。该模型通过特征增强模块和级联的实例-批量归一化挤压激励块实现多层次源说话人音色特征的提取。实验结果表明,FGE-SSV 在 VCSD-SSV 数据集全部 6 种 VC 模型中域内结果保持稳定的低等错误率(*EER*)(0.2%~1.0%),在 SSTC2024 数据集 6 种 VC 模型上也保持较低的 *EER*;此外,FGE-SSV 在未见过的 VC 模型上也展现出良好的泛化性能,其 *EER* 值均优于对比模型。

关键词 源说话人验证;语音转换伪造数据集;特征泛化增强;对抗学习;音色特征

中图法分类号 TP181; TP309

DOI: 10.7544/issn1000-1239.202550673 **CSTR:** 32373.14.issn1000-1239.202550673

在语音识别技术日益普及的背景下,自动说话人验证(automatic speaker verification, ASV)作为核心身份验证手段,已深度融入金融安全、智能设备认证等关键领域。然而,语音转换(voice conversion, VC)技术能够在保留原始语义内容的前提下,将源说话人的发音特性转换为目标说话人的音色特征。攻击者可借此篡改源说话人的声学特征,生成高度逼真的伪造语音,对 ASV 系统的安全性构成了严重威胁。

当前,多数反欺骗防御机制主要聚焦于伪造语音检测[1],即区分真实语音与伪造语音,却忽略了伪造语音中潜藏的源说话人身份信息。这导致其未能覆盖源说话人验证(source speaker verification, SSV)这一细粒度安全要求,在司法取证、欺诈溯源等实际应用场景中难以满足现实的迫切需求。

SSV 不仅需提取语音中的常规声学特征,更需关注那些即使经过语音转换仍可能保留的源说话人特征,其与伪造语音检测的关系如图 1 所示。具体而言,伪造语音检测属于二分类任务,常利用深度学习算法区分真实语音和伪造语音[2-4]。该领域常用数据集包括:ASV2019[5], ASV2021[6], FoR(fake or real)[7], SITW(speakers in the wild)[8], FakeAVCeleb(fake audios and videos of celebrities)[9] 等。

SSV 的核心目标是从转换后的语音中准确提取源说话人信息。从技术层面看,现有语音转换技术普遍存在源说话人信息残留现象:由于声学特征解耦不完全或生成模型的固有约束,转换语音的频谱、韵律等特征中仍可能携带源说话人的生物标识,这为溯源攻击者身份提供了理论依据。

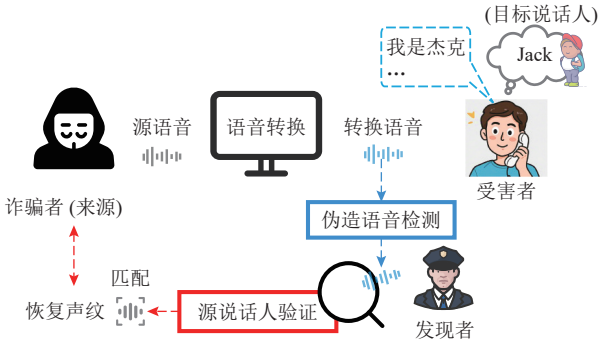


Fig. 1 Illustration of sproofing speech detection and source speaker verification

图 1 伪造语音检测与源说话人验证示意图

虽然已有模型[10-13]证明了在训练中引入标记源说话人身份的转换语音数据可在一定程度上实现源说话人识别,且利用多种不同 VC 模型生成的数据集训练有助于提升对未知转换模型的识别性能。但需要指出的是,不同 VC 模型在改变语音特征时对源说话人信息的保留和改变方式各有不同。面对性能不断提升的 VC 算法,如何高效准确地提取源说话人的音色特征信息,提升源说话人验证的泛化性,仍面临着严峻挑战。

基于上述背景,本文提出了面向源说话人验证的语音转换伪造数据集与模型。主要工作如下:

1)基于开源的 LibriSpeech 数据集[14]和 VoxCeleb1

数据集[15]，采用最新发布的 Seed-VC[16]，FACodec[17]，DDDM-VC[18]，Diff-HierVC[19]，TriAAN-VC[20]，FreeVC[21]等 6 种 VC 模型，构建了一个大规模源说话人验证语音转换伪造数据集（voice conversion spoofing dataset for source speaker verification，VCSD-SSV）。

2）提出了一种基于特征泛化增强的源说话人验证（source speaker verification based on feature generalised enhancement，FGE-SSV）模型。该模型利用特征增强模块和级联的实例-批量归一化挤压激励块（instance-batch normalization with sqeeze-and-excitation block，IBN-SEBlock）[22]实现多层次特征提取，并通过实验验证了所提模型的有效性。

1 相关工作

本节从 VC 和 SSV 这 2 个方面对已有工作进行对比分析。

1.1 VC 技术

VC 旨在改变语音的说话人身份（如共振峰、频谱包络、基频 F0），同时保留语义内容，其架构包含训练与转换 2 阶段。训练阶段主要分析源/目标说话人语音，提取特征，训练映射模型建立两者关系；转换阶段聚焦提取源语音特征，利用映射模型生成目标特征，最终重构目标语音波形。现有 VC 技术主要分为以下 3 类：

1）编码-解码类。与传统语音编解码技术[23]不同的是，VC 多采用 Seq2Seq 架构（编码器提取源特征，解码器生成目标语音），常需特征对齐。在这方面，FACodec[17]引入分解向量量化和分解扩散模型，实现人类自然度的语音生成。TriAAN-VC[20]使用 3 重自适应注意力归一化块，结合 siamese 损失，有效平衡源内容保持与目标说话人之间的相似性。FreeVC[21]基于 VITS 转换技术，利用信息瓶颈和数据增强从 WavLM 特征分离内容信息，实现高质量、鲁棒的一次性转换。

2）解耦类。侧重分离语音的语义内容与说话人身份，无需严格内容对齐。例如，DDDM-VC[18]将去噪过程分解为并行步骤来处理特定属性，提升可控性与质量。Diff-HierVC[19]分层系统分解为 DiffPitch 生成目标 F0 和 DiffVoice 转换语音。

3）扩散模型类。借鉴图像生成，通过噪声添加与去噪来学习复杂映射。例如，Seed-VC[16]基于零样本框架，引入音色变换器减少泄露，采用扩散变换器增强音色表征。Diff-HierVC[19]利用扩散模型生成 F0 并进行语音转换。

1.2 SSV

SSV 是应对 VC 攻击下身份安全威胁的关键技术。与传统 ASV 主要依据声纹特征识别目标说话人不同，SSV 旨在从经过深度转换的语音中鉴别出原始的源说话人身份，其核心科学问题是如何在目标说话人特征主导的信号中有效捕捉与分离微弱的源身份残留特征。当前该研究方向尚处于发展初期，仍面临公开数据集覆盖不足、VC 模型泛化能力弱等挑战。

近年来，国内外学者已尝试从不同技术路径推进 SSV 的发展，主要包括以下 3 类：

1）基于编码器设计与监督训练。例如，VC-SID 模型[10]采用 ResNet 编码结构，直接利用转换语音及其对应的源身份标签进行训练，从而学习源说话人特征表示。VC-SID 模型在相同 VC 条件下表现良好，但跨模型泛化能力有限。

2）侧重于特征解耦与表示学习。例如，REVELIO 框架[11]通过差分纠正算法抑制目标说话人特征，进而提取与源身份相关的正交表示。MeMo 模型[12]利用自监督学习与多头多模式知识蒸馏框架实现高效特征提取。

3）引入注意力机制，用于增强对源信息的捕捉能力。例如，SSTC 基准模型[13]利用注意力机制聚焦源特征并减弱目标说话人干扰。

虽然上述研究初步验证了 SSV 的可行性，但在特征表示的鲁棒性、跨 VC 泛化能力以及可解释性等方面仍存在明显不足，反映出 SSV 任务本身的艰巨性。与上述已有工作不同的是，本文构建了一个大规模源说话人验证语音转换伪造数据集 VCSD-SSV，提出了一种基于特征泛化增强的源说话人验证模型 FGE-SSV，旨在进一步提升源说话人验证的泛化性。

2 VCSD-SSV 数据集构建

2.1 数据准备

我们使用开源的 LibriSpeech[14]和 VoxCeleb1 数据集[15]构建我们的 VCSD-SSV 数据集。LibriSpeech 数据集采集自有声书网站，包含 7 个子集。VoxCeleb1 数据集则包含 1 251 位名人的超过 100 000 个语音片段，并提供开发集和测试集。

本文选用 LibriSpeech 作为源语音数据集，VoxCeleb1 作为目标语音数据集。为兼顾说话人的多样

性与代表性，我们从 LibriSpeech 的 train-clean-100 子集中分别选取 100 名男性说话人（11 553 条语音）和 100 名女性说话人（11 529 条语音），这些说话人具有不同的年龄、地域和语气特征，分别保存至 LM（LibriSpeech_Male）和 LF（LibriSpeech_Female）文件夹。同样，从 VoxCeleb1 开发集选取 100 名男性说话人（10 890 条语音）和 100 名女性说话人（9 110 条语音）作为目标说话人，分别保存至 VM（VoxCeleb1_Male）和 VF（VoxCeleb1_Female）文件夹。

2.2 数据生成

我们使用 Seed-VC, FACodec, DDDM-VC, Diff-HierVC, TriAAN-VC, FreeVC 共 6 种 VC 模型生成转换数据。根据源说话人与目标说话人的性别组合，转换方式分为 4 类：男到男（male to male, M2M）、男到女（male to female, M2F）、女到女（female to female, F2F）和女到男（female to male, F2M）。每种 VC 模型的数据生成流程如图 2 所示。

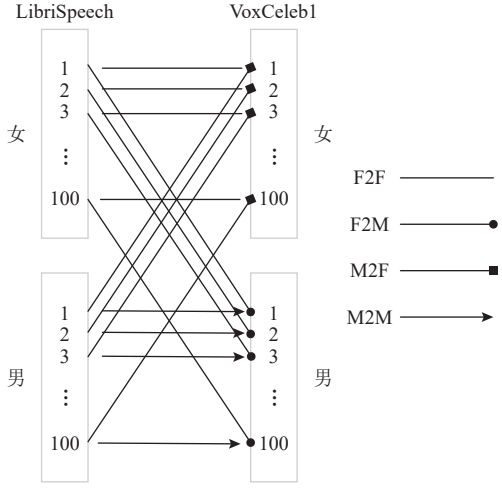


Fig. 2 Illustration of data generation

图 2 数据生成示意图

数据生成的具体方法描述如下：

1）M2M。将 LM 文件夹中的说话人语音，使用 VC 模型依次向 VM 文件夹中的说话人转换。即，LM1 向 VM1 转换，LM2 向 VM2 转换，以此类推，共形成 100 对源-目标说话人组合。对于每一对组合，使用源说话人的所有语音作为源输入，并从目标说话人中随机选取一条语音作为参考语音（目标语音），应用指定的 VC 模型进行转换。

2）M2F, F2M, F2F。均遵循与 M2M 相同的生成方式，分别对应 LM→VF, LF→VM, LF→VF 的转换路径。

2.3 数据标注

在数据生成阶段，针对每种 VC 模型（Seed-VC, FACodec, DDDM-VC, Diff-HierVC, TriAAN-VC, FreeVC），我们创建一个独立的文件夹用于存储其生成的转换语音。该文件夹命名为“VC 名称-LibriSpeech-VoxCeleb1”。每个文件夹内进一步划分为 4 个子文件夹，分别存放不同性别转换类型的语音：M2M、M2F、F2M 和 F2F。所有语音文件均采用.wav 格式。

在每个子文件夹中，每个源-目标说话人对的转换语音存储在以源说话人 ID-目标说话人 ID 命名的目录下。例如，目录名 103-id10270 表示源说话人 ID 为 103，目标说话人 ID 为 id10270，即由说话人 103 向 id10270 转换的结果。

以使用 FreeVC 将 LM 中的男性说话人 103 向 VM 中的男性说话人 id10270 转换为例，生成的语音文件路径结构如下：

FreeVC-LibriSpeech-VoxCeleb1/M2M/103-id10270,
├── 103-1240-0001-id10270.wav,
├── 103-1240-0002-id10270.wav,
├── 103-1240-0003-id10270.wav,

其中 103 是源说话人 ID，1240 是章节编号，0001~0003 是语音序列号，id10270 是目标说话人 ID。

基于上述规范构建的 VCSD-SSV 数据集结构如表 1 所示。该数据集包含 6 种语音转换模型（Seed-VC, FACodec 等）生成的子集和原始语音库。所有 VC 子集均基于相同的源数据（100 名源/目标说话人、23 082 条源语音），生成等量的转换语音（各 145.6 h、45 558 条），且转换语音数为源语音的 2 倍；原始语音库则分为独立的源说话人子集（72.8 h、100 人）和目标说话人子集（85.2 h、100 人），前者用于对比实验或基线训练，后者用于说话人特征提取或验证任务，两者均未参与转换过程。

Table 1 VCSD-SSV Dataset Structure

表 1 VCSD-SSV 数据集结构

子集	时长/h	源说话人数（男/女）	目标说话人数（男/女）	源说话人语音数	VC 语音数
Seed-VC-LibriSpeech-VoxCeleb1	145.6	100/100	100/100	23 082	45 558
FACodec-LibriSpeech-VoxCeleb1	145.6	100/100	100/100	23 082	45 558
DDDM-VC-LibriSpeech-VoxCeleb1	145.6	100/100	100/100	23 082	45 558
Diff-HierVC-LibriSpeech-VoxCeleb1	145.6	100/100	100/100	23 082	45 558
TriAAN-VC-LibriSpeech-VoxCeleb1	145.6	100/100	100/100	23 082	45 558
FreeVC-LibriSpeech-VoxCeleb1	145.6	100/100	100/100	23 082	45 558
Source Speaker	72.8	100/100		23 082	
Target Speaker	85.2		100/100		

总之，我们的 VCSD-SSV 数据集[①]包含两大模块：VC 转换语音库（6 种模型×4.55 万条语音）和原始语音库（2.3 万条源语音+目标语音独立库）。这种统一数据基底+多模型对比+原始数据隔离的实验设计，为 VC 和 SSV 任务提供了多样化的数据支持。

3 基于特征泛化增强的 SSV 模型

本文提出的基于特征泛化增强的源说话人验证方法 FGE-SSV，其总体框架如图 3 所示。FGE-SSV 首先对输入语音（自然语音与转换语音）提取 Fbank 特征[24]，并通过特征融合模块生成融合特征。该融合特征随后输入并行双分支进行处理。

1）特征增强分支。依次通过特征增强模块（feature enhancement module, FEM）[25]、空间与通道重建卷积（spatial and channel reconstruction convolution, SCConv）模块[26]提取源说话人信息。

2）泛化学习分支。通过级联的 IBN-SEBlock[22]、SCConv[26]以及双分支注意力融合（attention merge）模块实现多层次特征提取。

最终，融合 3 个 IBN-SEBlock 的输出特征输入到源说话人验证网络进行身份验证。

下面我们将详细介绍一些关键模块的设计。

3.1 特征融合模块

输入语音（包括自然语音与转换语音）首先经过 Fbank 特征[24]提取。随后，通过特征融合模块对 Fbank 特征进行处理，具体包含 1 维卷积（Conv1d）、ReLU 激活函数以及批量归一化（batch normalization, BN）操作，最终生成高阶融合特征。

3.2 特征增强模块

源说话人身份还原任务具有高度复杂性，模型常面临学习源说话人信息不足的挑战。为更有效地捕捉转换语音中的源说话人信息并获取更丰富的特征，FGE-SSV 引入了 FEM[25]。FEM 模块旨在从 2 个维度对源说话人信息进行强化，其整体结构如图 4 所示。

FEM 的核心设计采用多分支卷积架构，以挖掘多样化的说话人信息，从而提升特征的丰富程度。同时，为拓展感受野并捕获更丰富的局部上下文信息，模块中引入了空洞卷积。这种设计的优势在于，能够在保持模型轻量化的前提下，使生成的特征图蕴含更丰富的源说话人信息。

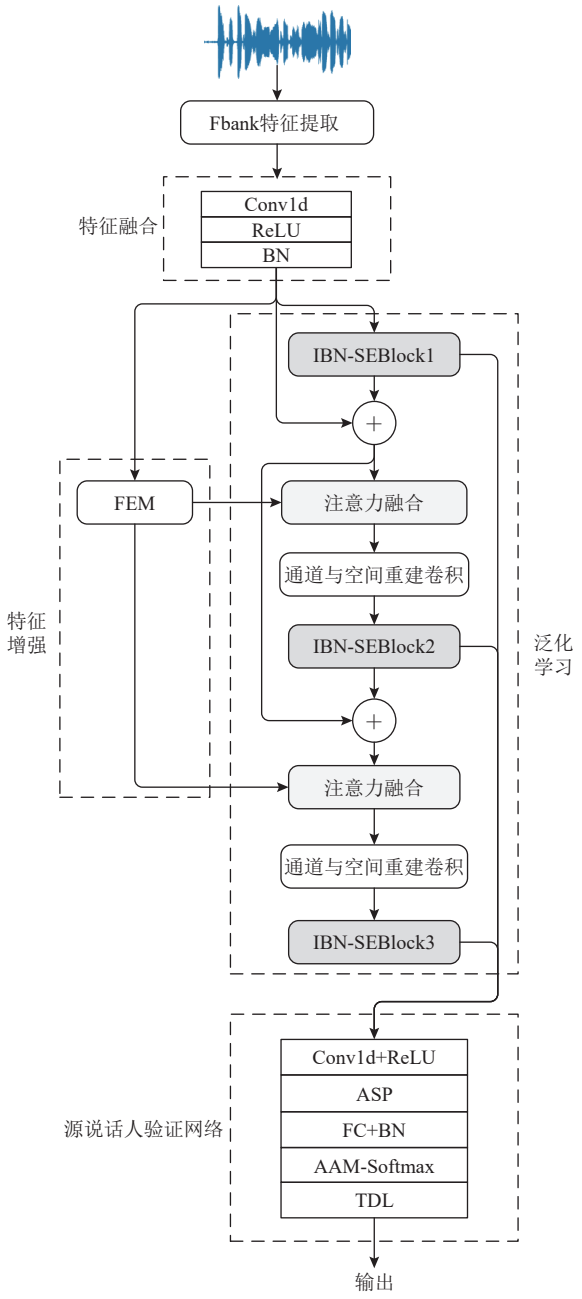


Fig. 3 The overall framework of FGE-SSV model

图 3 FGE-SSV 模型总体框架

在图 4 中，FEM 的计算过程由式（1）~（4）描述：

$$\boldsymbol{X}_1 = f_{\text{conv}}^{3\times3}[f_{\text{conv}}^{1\times1}(\boldsymbol{F})]\,, \tag{1}$$

$$\boldsymbol{X}_2 = f_{\text{atconv}}^{3\times3}\{f_{\text{conv}}^{3\times1}\{f_{\text{conv}}^{1\times3}[f_{\text{conv}}^{1\times1}(\boldsymbol{F})]\}\}\,, \tag{2}$$

$$\boldsymbol{X}_3 = f_{\text{atconv}}^{3\times3}\{f_{\text{conv}}^{1\times3}\{f_{\text{conv}}^{3\times1}[f_{\text{conv}}^{1\times1}(\boldsymbol{F})]\}\}\,, \tag{3}$$

$$\boldsymbol{Y}_{\text{FEM}} = Cat(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3) \oplus f_{\text{conv}}^{1\times1}(\boldsymbol{F})\,, \tag{4}$$

其中 $f_{\text{conv}}^{1\times1}$，$f_{\text{conv}}^{1\times3}$，$f_{\text{conv}}^{3\times1}$，$f_{\text{conv}}^{3\times3}$ 分别代表使用 1×1, 1×3, 3×1, 3×3 卷积核的标准卷积运算；$f_{\text{atconv}}^{3\times3}$ 表示空洞率为 5 的空洞卷积运算；$Cat(\cdot)$ 为特征图拼接操作；$\oplus$ 表示特征

① VCSD-SSV 数据集已在线发布共享：https://figshare.com/articles/dataset/FEG-SSV/29235527

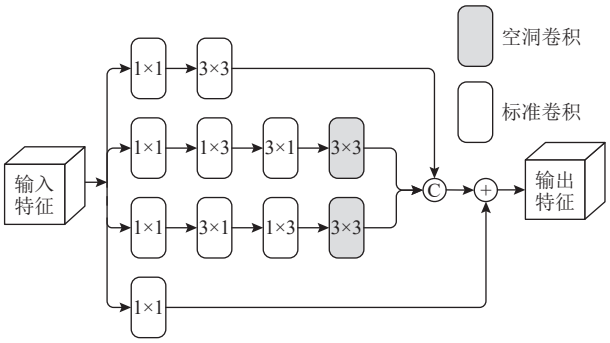


Fig. 4 The FEM module structure

图 4 FEM 模块结构

图的逐元素相加; $\boldsymbol{F}$ 为输入特征图; $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$ 分别为前 3 个分支经标准卷积及空洞卷积处理后的输出特征图; $\boldsymbol{Y}_{\mathrm{FEM}}$ 是 FEM 模块的输出特征图。

3.3 泛化学习模块

为降低模型对特定训练数据集的依赖, 提升其在未见数据上的泛化能力, FGE-SSV 引入了以下关键组件: IBN-SEBlock[22], SCConv[26] 及双分支 Attention Merge 模块。

3.3.1 IBN-SEBlock 模块

在包含批 BN 的 CNN 中, 外观变化引起的特征差异主要集中于浅层, 而深层特征虽富含说话人信息, 浅层也具备内容区分能力。设计 IBN-SEBlock 模块的目的就是为了保护深层区分能力, 避免在网络末端添加实例归一化(instance normalization, IN), 同时保留浅层说话人信息, 在残差路径中保留部分 BN 特征。

如图 5 所示, 残差路径的首个卷积层后, 对一半通道应用 BN, 另一半应用 IN。其中, IN 应置于残差路径而非恒等路径, 以保持恒等路径的优化特性; 通道分割策略(BN/IN 各半)旨在平衡泛化能力与说话人信息保留。这是因为 IN 有助于学习外观不变特征以利用数据集内差异大的音频, 同时对其适度引入能有效保留关键说话人信息。为最大化 IN 的泛化潜力, 将其置于残差连接后, 确保外观信息的同时在残差和恒等路径中保留。此外, IBN-SEBlock 模块集成了 SENet[27], 通过对通道特征进行自适应加权, 缓解模型对特定通道的过度依赖, 降低过拟合风险, 进一步增强泛化能力与特征表达能力。

3.3.2 SCConv 模块

SCConv 模块旨在减少 CNN 特征中的空间和通道冗余, 从而压缩模型以及降低复杂度与计算成本, 同时提升特征表示能力。这有助于模型学习源说话人的本质特征, 提高对未知语音转换模型的源说话

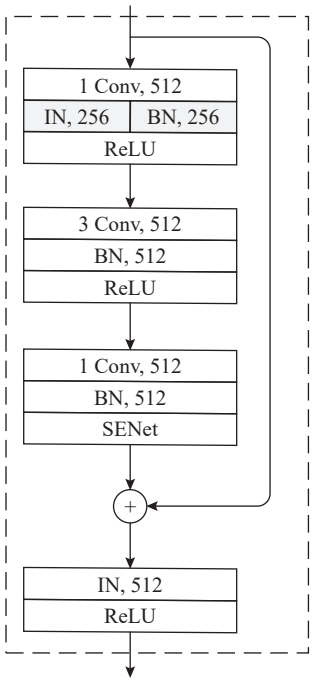


Fig. 5 Structure of IBN-SEBlock

图 5 IBN- SEBlock 结构

人身份还原能力和泛化能力。如图 6 所示, SCConv 模块主要包含空间重建单元(spatial reconstruction unit, SRU)和通道重建单元(channel reconstruction unit, CRU)两个核心单元。

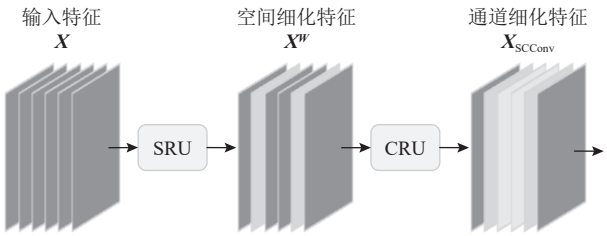


Fig. 6 The overall structure of SCConv

图 6 SCConv 总体结构

1) SRU

如图 7 所示, 利用组归一化(group normalization, GN)层中的缩放系数 $\boldsymbol{\beta}$ 量化特征信息量(信息丰富区 $\boldsymbol{\beta}$ 值大)。通过计算权重 $\boldsymbol{W}_{\beta}$ 和门控阈值处理输出权重分量 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$, 将输入特征 $\boldsymbol{X}$ 解耦为富含信息的特征 $\boldsymbol{X}_1^{W}$ 和信息匮乏(冗余)的特征 $\boldsymbol{X}_2^{W}$。

$$\boldsymbol{W}_{\beta}=\boldsymbol{\beta}_i/\sum_{j=1}^{C}\boldsymbol{\beta}_j,\ i=1,2,\cdots,C, \tag{5}$$

$$\boldsymbol{W}=Gate(Sigmoid(\boldsymbol{W}_{\beta}(GN(\boldsymbol{X})))), \tag{6}$$

$$\begin{cases}\boldsymbol{X}_1^{W}=\boldsymbol{W}_1\otimes\boldsymbol{X},\\ \boldsymbol{X}_2^{W}=\boldsymbol{W}_2\otimes\boldsymbol{X},\end{cases} \tag{7}$$

其中 *Gate* 为阈值函数, C 为通道维度, $\otimes$表示逐元素

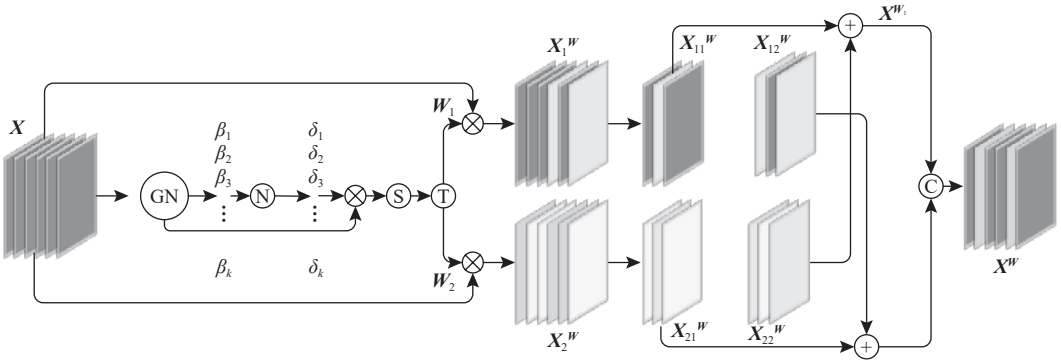


Fig. 7 The structure of SRU

图 7 SRU 结构

相乘。

随后采用交叉重构操作而非简单相加，有效融合不同信息密度的特征，增强信息交互，生成空间细化特征 $\boldsymbol{X}^{W}$：

$$\begin{cases} \boldsymbol{X}_{11}^{W} \oplus \boldsymbol{X}_{22}^{W} = \boldsymbol{X}^{W_1}, \\ \boldsymbol{X}_{21}^{W} \oplus \boldsymbol{X}_{12}^{W} = \boldsymbol{X}^{W_2}, \\ \boldsymbol{X}^{W_1} \cup \boldsymbol{X}^{W_2} = \boldsymbol{X}^{W}, \end{cases} \tag{8}$$

其中⊕表示逐元素相加，∪表示拼接。

2）CRU

如图 8 所示，对空间细化特征 $\boldsymbol{X}^{W}$ 进行通道维度的冗余削减。采用拆分-变换-融合策略，利用轻量级卷积提取代表性特征 $\boldsymbol{X}_{\mathrm{SCConv}}$，并高效处理冗余特征以降低计算成本。SCConv 模块通过有效消除空间和通道维度的冗余信息，显著抑制转换语音中目标说话人的影响，提升模型泛化能力。

3.3.3 Attention Merge 模块

Attention Merge 模块负责融合来自不同分支（如 FEM 和泛化学习分支）的特征张量。其核心是使用注意力机制动态调整输入特征的权重，以更灵活高效地综合信息，如图 9 所示。

大致流程描述如下：

1）通过独立线性投影层将增强特征 $\boldsymbol{X}_{\mathrm{enh}}$ 和泛化特征 $\boldsymbol{X}_{\mathrm{gen}}$ 的每个通道在时间维度上映射为标量注意力分数 $\boldsymbol{S}_{\mathrm{enh}}$ 和 $\boldsymbol{S}_{\mathrm{gen}}$：

$$\boldsymbol{S}_{\mathrm{enh}} = \frac{L(\boldsymbol{X}_{\mathrm{enh}})}{\sqrt{t}}, \tag{9}$$

$$\boldsymbol{S}_{\mathrm{gen}} = \frac{L(\boldsymbol{X}_{\mathrm{gen}})}{\sqrt{t}}, \tag{10}$$

其中 $L(\cdot)$ 是线性映射函数，t 为特征的时间维度。

2）使用 Softmax 函数对 $\boldsymbol{S}_{\mathrm{enh}}$ 和 $\boldsymbol{S}_{\mathrm{gen}}$ 进行归一化，得到各通道的注意力权重 $\boldsymbol{W}_{\mathrm{enh}}$ 和 $\boldsymbol{W}_{\mathrm{gen}}$ 反映其重要性。

3）将注意力权重与原始特征 $\boldsymbol{X}$ 相乘进行加权，并通过全局平均池化压缩时间维度，得到加权值 $\boldsymbol{V}_{\mathrm{enh}}$ 和 $\boldsymbol{V}_{\mathrm{gen}}$。

4）根据融合权重 $\boldsymbol{V}_{\mathrm{enh}}$ 和 $\boldsymbol{V}_{\mathrm{gen}}$ 对输入特征进行加权融合，并经过线性变换得到最终融合输出 $\boldsymbol{X}_{\mathrm{merge}}$：

$$\boldsymbol{X}_{\mathrm{merge}} = L(\boldsymbol{V}_{\mathrm{enh}} \times \boldsymbol{X}_{\mathrm{enh}} + \boldsymbol{V}_{\mathrm{gen}} \times \boldsymbol{X}_{\mathrm{gen}})。 \tag{11}$$

Attention Merge 模块能自适应调整融合策略，考虑多尺度特征信息，提升模型从语音信号中提取源说话人信息的能力。此外，Attention Merge 模块的使用增强了模型的鲁棒性，有助于防止过拟合。

3.3.4 模块协同与数据流

FGE-SSV 模型的泛化学习分支按以下流程处理特征：

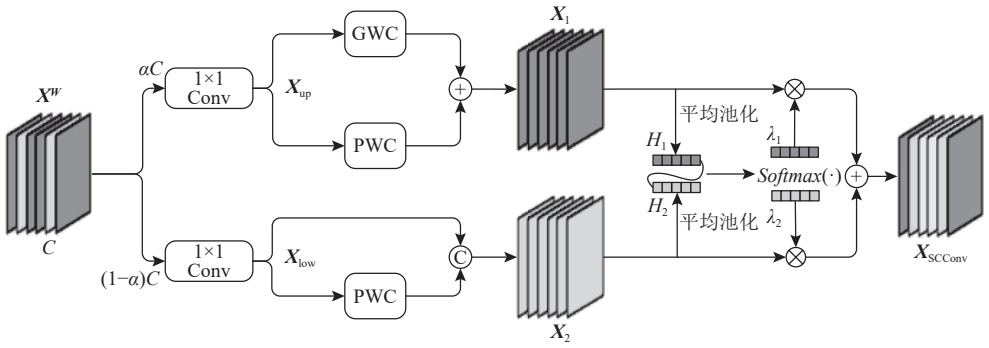


Fig. 8 The structure of CRU

图 8 CRU 结构

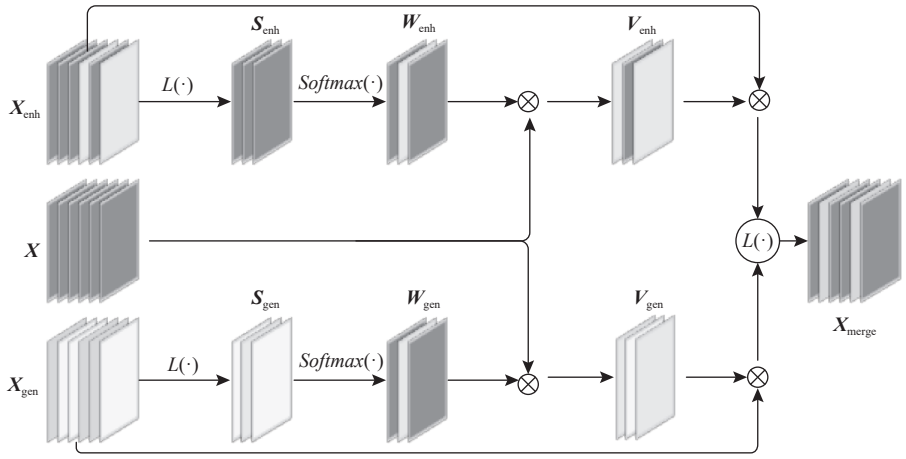


Fig. 9 The structure of Attention Merge

图 9 Attention Merge 结构

1)特征融合模块的输出分别输入 FEM 模块和 IBN-SEBlock 模块。

2)IBN-SEBlock 模块的输出输入到 SCConv 模块进行空间和通道冗余削减。

3)SCConv 模块的输出与 FEM 模块的输出一同输入 Attention Merge 模块进行自适应加权融合。

4)融合后的特征可继续输入后续的 SCConv 模块和 IBN-SEBlock 模块进行进一步处理。

5)处理后的特征通过源说话人验证网络输出结果。

3.4 源说话人验证网络

在图 3 中，源说话人验证网络对输入特征执行以下后处理流程，以基于提取的说话人信息进行身份判别：

1)为捕获不同层次的特征信息，将 IBN-SEBlock 模块中 3 个子模块的输出在通道维度上进行拼接。

2)利用 1 维卷积对拼接后的特征进行深度融合。

3)采用注意力统计池化(attentive statistics pooling, ASP)层处理融合特征的通道维度，生成更具判别性的统计特征表示。

4)通过全连接层(fully connected layer, FC)将聚合后的特征映射为一个 192 维度的说话人嵌入向量。

5)使用 AAM-Softmax 损失函数进行模型训练。该损失在标准 Softmax 中引入了角度裕度，以增大类间差异和增强分类判别能力，从而优化模型性能。

6)在类别判别层(type discrimination layer, TDL)中计算生成的嵌入向量与目标说话人注册向量之间的余弦距离，并依据该距离判断输入转换语音是否来源于目标说话人。

4 实验方案

本节阐述用于评估我们提出的 FGE-SSV 模型性能的实验方案。

4.1 实验设置

为验证 FGE-SSV 模型的有效性，我们选取了 VCSID[10]，REVELIO[11]，SSTC 基准模型[13]3 种模型进行对比实验分析。为了增强训练数据的多样性和对抗性，提高模型对合成语音，特别是多种转换技术所生成欺骗语料的判别能力，我们引入 SSTC2024 官网数据集[13]。该数据集通过 16 种不同的 VC 模型构建，在 LibriSpeech 和 VoxCeleb2 的基础上合成了大规模、多方法、多说话人配对的高仿真语音语料。SSTC 基准模型[13]在训练阶段使用了超过 32 万条转换语音，开发和测试集也分别包含 1.4 万和 1.3 万条样本。除了上述 VC 模型，我们选取了 SSTC2024 官网数据集 train1~train6 共 6 种 VC 模型进行对比实验：TriAAN-VC[20]，FreeVC[21]，AGAIN-VC[28]，MediumVC[29]，StyleTTS[30]，VQMIVC[31]。实验配置环境见表 2。

Table 2 Experimental Configuration Environment

表 2 实验配置环境

项目名称	配置环境
CPU	Intel Core i5-12600KF
RAM/GB	8
GPU	NVIDIA GeoForce RTX 4060Ti
CUDA	Cuda 12.6
深度学习框架	PyTorch 2.4
操作系统	Windows 10
编程语言	Python 3.9

FGE-SSV 各模块的输入输出维度配置如表 3 所示。需特别说明的是，当模块间存在维度转换需求时，需要通过特征增维或降维操作确保数据维度的兼容性。FGE-SSV 采用 Adam 优化器，初始学习率设为 0.001，配合 0.97 的权重衰减系数，并采用分阶段学习率调整策略以提升训练稳定性。AAM-SoftMax 损失函数中的超参数损失函数缩放因子 *scale*=30，损失函数间隔参数 *margin*=0.2。

Table 3 Input-Output Dimension Configuration for Modules in FGE-SSV Model

表 3 FGE-SSV 模型模块的输入输出维度配置

模块	输入维度	输出维度
FEM	(B,512,8,12)	(B,512,8,12)
SCConv1	(B,512,8,12)	(B,512, 8,12)
IBN-SEBlock1	(B,512,157)	(B,512,157)
SCConv2	(B,512,1,157)	(B,512,157)
IBN-SEBlock2	(B,512,157)	(B,512,157)
SCConv3	(B,512,1,157)	(B,512,1,157)
IBN-SEBlock3	(B,512,157)	(B,512,157)

4.2 实验数据配置

为确保实验的可靠性，采用分层抽样方法对 VCSD-SSV 数据集和 SSTC2024 官网数据集进行实验数据划分：从每个 VC 数据集中随机选取 20 组样本进行训练，包含 F2F, F2M, M2F, M2M 这 4 种类型各 5 组。训练集包含源语音、目标语音和转换语音各约 200 条/人，涵盖 20 个说话人。

在测试阶段，所有样本均以语音对形式呈现，每对包含转换语音和源语音各 1 条。测试集保证各 VC 算法的数据量均衡，并采用二元标注机制：当源语音与转换语音的说话人 ID 一致时，标注为 1；当说话人 ID 不一致时，标注为 0。

最终在每个 VC 算法上构建的训练集规模约为 5 000 条样本，测试集为 2 000 对样本（1 对样本对应 1 条源语音和 1 条转换语音）。

为明确本文所设计实验的适用场景，本文将实验划分为以下 2 类：

1）闭集设定。训练集与测试集共享相同的说话人集合，所有测试说话人均已在训练阶段出现。本文 5.1~5.4 节中所报告的域内及泛化实验均基于此设定，旨在验证模型对已知说话人身份的溯源性能。

2）开集设定。训练集与测试集的说话人 ID 互不重叠，模型需针对未见过的说话人，判断其是否是某段转换语音的源说话人（该部分实验结果将在本文 5.5 节进行介绍）。

4.3 评价指标

采用等错误率（equal error rate, *EER*）作为源说话人验证的核心评价指标。*EER* 是当误识率（false acceptation rate, *FAR*）与拒识率（false rejection rate, *FRR*）相等时的性能度量，能够综合反映检测模型的鲁棒性和平衡性。

$$FAR(\theta) = \frac{N_{\mathrm{FA}}(\theta)}{N_{\mathrm{sf}}}, \tag{12}$$

$$FRR(\theta) = \frac{N_{\mathrm{FR}}(\theta)}{N_{\mathrm{bf}}}, \tag{13}$$

$$EER = FAR(\theta) = FRR(\theta), \tag{14}$$

其中 θ 为判定阈值，N_{sf} 和 N_{bf} 分别表示伪造语音和真实语音的样本总数，$N_{\mathrm{FA}}(\theta)$ 为模型检测分数大于 θ 的伪造语音样本数，$N_{\mathrm{FR}}(\theta)$ 为模型检测分数小于或等于 θ 的真实语音样本数。*EER* 值越小，表明模型的检测性能越优。

5 实验结果与分析

在本节中，我们将从多个角度汇报并讨论实验结果。

5.1 消融实验分析

为量化评估 FGE-SSV 模型中各关键模块对消除转换语音中目标音色干扰及提升模型泛化能力的作用，表 4 给出了 FGE-SSV 各模块在 VCSD-SSV 数据集上的消融实验的 *EER* 结果。

Table 4 Ablation Experimental *EER* Results of FGE-SSV Modules on VCSD-SSV Dataset

表 4 FGE-SSV 模块在 VCSD-SSV 数据集上消融实验的 *EER* 结果 %

模型	Seed-VC	FACodec	DDDM-VC	Diff-HierVC	TriAAN-VC	FreeVC
FGE-SSV	**0.7**	**8.4**	**15.6**	**14.7**	**12.3**	**16.9**
w/o FEM	3.5	9.6	17	15.6	12.9	18.3
w/o SCConv	0.9	10.5	18.2	20.1	20.7	17.6
w/o IBN-SEBlock	1.2	11.5	18.3	20.7	22.2	23

注：黑体数值表示最优值。

从表 4 可以看出，移除 FEM 模块后，模型在 Seed-VC（域内）和其他数据集上的 *EER* 分别上升 2.8 个百分点和 1.2~1.4 个百分点。这表明 FEM 模块显著提升了模型对语音特征的表示能力，尤其在域内数据上表现更为突出（*EER* 升幅最大达 2.8 个百分点）。这一结果验证了特征增强分支的双重作用：既防止模型过度偏向泛化性优化，又确保了域内性能的稳

定性。其次，禁用 SCConv 模块导致域内 *EER* 小幅上升 0.2 个百分点，而在跨 VC 数据集上的 *EER* 升幅更为显著(2.1~8.4 个百分点)。其中，在 Diff-HierVC 和 TriAAN-VC 数据集上的 *EER* 分别上升 5.4 个百分点和 8.4 个百分点，表明 SCConv 模块对复杂语音转换场景(如层级扩散模型和对抗生成模型)的泛化能力具有关键作用。此外，移除 IBN-SEBlock 模块后，模型在域内 *EER* 仅上升 0.5 个百分点，但在跨数据集上的性能退化尤为明显(*EER* 升幅 3.1~9.9 个百分点)。例如，在 TriAAN-VC 和 FreeVC 数据集上，*EER* 分别上升 9.9 个百分点和 6.1 个百分点，说明 IBN-SEBlock 模块通过特征归一化和通道注意力机制，有效增强了模型对异构语音转换方法的适应能力。

总之，FGE-SSV 在所有数据集上均取得最低 *EER*，而任何模块的移除均会导致性能退化，其中 IBN-SEBlock 模块的缺失对跨数据集影响最大。

5.2 音色特征有效性分析

为定量评估 FGE-SSV 模型从转换语音中提取源说话人音色特征的能力，我们基于 VCSD-SSV 数据集设计了特征空间分布对比实验。每种 VC 模型选取 4 组典型说话人对(M2M, F2M, M2F, F2F)，每组说话人对包含 10 条源语音及其对应转换语音样本。图 10 展示了 FGE-SSV 与 VC-SID 在 6 种 VC 数据集上的特征空间分布 UMAP[32] 可视化结果，其中红色表示 M2M 转换，蓝色代表 F2M 转换，绿色表示 M2F 转换，紫色代表 F2F 转换。

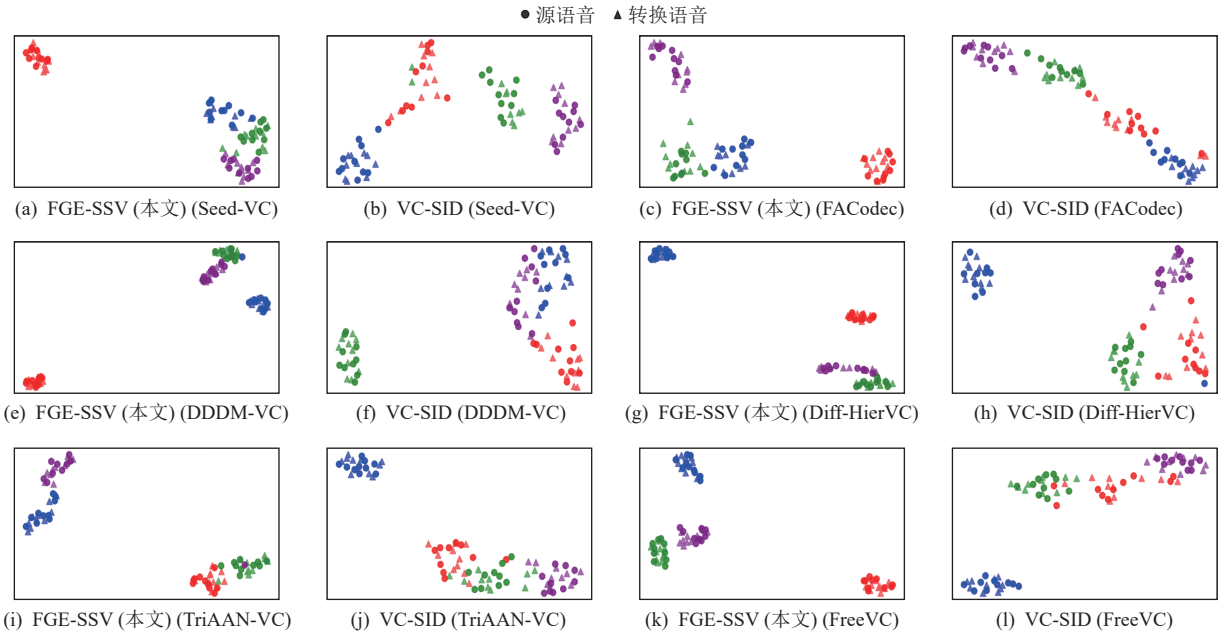


Fig. 10 Visualization results of feature space distributions for FGE-SSV and VC-SID models across different datasets

图 10 FGE-SSV 与 VC-SID 模型在各数据集上的特征空间分布可视化结果

由图 10 可以看出，FGE-SSV 中同颜色点集呈现显著聚类特征，而 VC-SID 存在明显离散现象，尤其在 Diff-HierVC 数据集上出现多个离群点。其次，FGE-SSV 成功分离 4 类转换场景，在性别转换场景(F2M/M2F)仍保持良好聚类，证明其能有效克服音高变异干扰，而 VC-SID 在跨性别转换中出现特征重叠，尤其是在 TriAAN-VC 数据集的 M2F 转换中出现特征漂移。

上述实验结果表明，FGE-SSV 通过多尺度特征提取机制，有效分离了语音转换过程中注入的目标说话人特征与保留的源说话人特征。即使在性别转换场景下，FGE-SSV 模型仍能捕捉源说话人的声道等生理特征，这解释了其在 F2M/M2F 任务中的优异表现。面对不同 VC 算法(特别是基于扩散模型的 Diff-HierVC)，FGE-SSV 表现出更强的适应性，这与消融实验中 IBN-SEBlock 模块的作用相印证。

5.3 域内测试分析

为验证 FGE-SSV 模型在域内场景下的有效性，我们选取训练集与测试集均来自同一 VC 模型生成的转换语音。表 5 给出了 4 种模型在 VCSD-SSV 数据集上的域内 *EER* 对比。由表 5 可以看出，FGE-SSV 保持稳定的低 *EER*(0.2%~1.0%)，在全部 6 种 VC 模型上的 *EER* 平均下降 70 个百分点，在 DDDM-VC 上优势最显著。REVELIO 表现波动剧烈(9.9%~43.8%)，尤其在 TriAAN-VC 上异常退化(*EER* 为 43.8%)。对于 Diff-HierVC，因其转换痕迹明显，4 种模型表现较

Table 5 In-Domain Experimental Results for *EER* of Different Models on VCSD-SSV Dataset

表 5 各模型在 VCSD-SSV 数据集的域内 *EER* 实验结果 %

VC 模型	FGE-SSV（本文）	VC-SID	REVELIO	SSTC 基准模型
Seed-VC	0.7	1	28.5	17.1
FACodec	0.4	1.2	33.2	26.2
DDDM-VC	0.2	2.1	28.4	14.3
Diff-HierVC	0.7	1.8	9.9	11.3
TriAAN-VC	1	3.4	43.8	19.9
FreeVC	0.7	2.3	21.2	26.5
平均	0.6	2	27.5	19.2

优。可能的原因是，FGE-SSV 中的 FEM 模块能够有效捕捉 VC 模型的特定伪影（如 Diff-HierVC 的频带失真），IBN-SEBlock 模块抑制了生成式 VC（Diff-HierVC/TriAAN-VC）的过度平滑效应和对抗性扰动。

5.4 泛化性测试分析

为测试所提出 FGE-SSV 模型的泛化性，用其中一个 VC 模型的转换语音进行训练，而在另外 5 种 VC 模型上进行测试，实验结果如表 6 和表 7 所示。

由表 6 可以看出，与已有方法相比，所提 FGE-SSV 模型在未见过的语音转换方法上展现出良好的泛化性能。具体来说，采用 Seed-VC 模型进行训练时，在其他 5 种模型的 *EER* 平均为 13.58%，均低于 VC-SID 的 29.72% 和 REVELIO 的 46.52% 以及 SSTC 基准模型的 24.32%；采用 FACodec 模型进行训练时，其他 5 种模型的 *EER* 平均为 29.56%，均低于 VC-SID 的 37.74% 和 REVELIO 的 57.96% 以及 SSTC 基准模型的 33.38%；采用 DDDM-VC 模型进行训练时，其他 5 种模型的 *EER* 平均为 15.04%，均低于 VC-SID 的 24.66% 和 REVELIO 的 32.24% 以及 SSTC 基准模型的 24.52%；采用 Diff-HierVC 模型进行训练时，其他 5 种模型的

Table 6 *EER* Results of Different Models on VCSD-SSV Dataset in Generalisation Experiment

表 6 各模型在 VCSD-SSV 数据集上的泛化实验 *EER* 结果 %

网络模型	数据集	Seed-VC	FACodec	DDDM-VC	Diff-Hier VC	TriAAN-VC	FreeVC
VC-SID	Seed-VC	-	17.6	30.3	32.2	40.8	27.7
	FACodec	39.0	-	39.5	42.5	41.8	25.9
	DDDM-VC	30.8	22.0	-	4.7	36.4	33.8
	Diff-HierVC	31.6	28.8	2.5	-	52.2	36.1
	TriAAN-VC	37.6	32.7	39.2	41.5	-	27.8
	FreeVC	43	45.2	48.2	47.1	45.3	-
REVELIO	Seed-VC	-	35.3	50.8	48.9	51.6	46.0
	FACodec	42.4	-	60.9	61.9	60.9	63.7
	DDDM-VC	35.0	31.0	-	31.3	31.4	32.5
	Diff-HierVC	19.7	15.7	13.6	-	21.3	24.4
	TriAAN-VC	52.1	51.9	51.8	51	-	51.6
	FreeVC	38.8	37.4	43.2	41.4	43.0	-
SSTC 基准模型	Seed-VC	-	21.1	29.5	25.3	23.1	22.6
	FACodec	33.3	-	38.2	31.8	36.4	29.7
	DDDM-VC	27.9	20.2	-	17.8	29.8	26.9
	Diff-HierVC	28.7	22.9	15.0	-	21.1	33.9
	TriAAN-VC	30.2	27.7	25.5	20.1	-	31.6
	FreeVC	39.9	35.4	45.6	30.2	39.2	-
FGE-SSV（本文）	Seed-VC	-	8.4	15.6	14.7	12.3	16.9
	FACodec	26.1	-	32.7	34.5	27.9	26.6
	DDDM-VC	17.5	13.2	-	1.2	20.5	22.8
	Diff-HierVC	27.6	20.4	3.8	-	29.3	27.4
	TriAAN-VC	24.6	19.1	25.9	26.6	-	21.9
	FreeVC	33.7	26.7	39.5	39.2	30.0	-

注：- 表示域内实验结果，在前文已单独讨论，此表内不予给出。

Table 7 *EER* Results of Different Models on SSTC Dataset in Generalisation Experiment

表 7 各模型在 SSTC 数据集上的泛化实验 *EER* 结果 %

网络模型	数据集	AGAIN-VC	FreeVC	MediumVC	StyleTTS	TriAAN-VC	VQMIVC
VC-SID	AGAIN-VC	24.54	47.66	47.66	45.84	48.11	47.05
	FreeVC	48.11	17.52	37.82	49.32	45.69	43.87
	MediumVC	46.44	40.39	25.80	49.02	46.29	42.06
	StyleTTS	55.22	47.36	45.66	21.3	46.75	48.71
	TriAAN-VC	48.71	41.91	42.21	48.11	16.16	42.97
	VQMIVC	56.58	43.72	39.79	49.02	44.93	23.09
REVELIO	AGAIN-VC	40.3	48.3	50.8	48.9	51.6	56.0
	FreeVC	42.4	39.7	50.9	51.9	50.9	53.7
	MediumVC	48.4	49.3	38.6	48.2	45.5	47.9
	StyleTTS	49.2	47.7	41.6	40.5	42.8	54.4
	TriAAN-VC	47.5	53.4	52.3	51.5	40.6	53.2
	VQMIVC	55.1	58.5	55.4	52.8	51.7	43.2
SSTC 基准模型	AGAIN-VC	22.66	48.26	47.95	45.47	46.47	44.98
	FreeVC	41.6	17.55	38.16	39.03	36.16	36.16
	MediumVC	44.78	40.39	18.46	42.09	39.03	45.04
	StyleTTS	45.08	45.84	47.16	23.9	45.69	50.08
	TriAAN-VC	42.97	36.91	40.21	40.54	17.7	43.12
	VQMIVC	47.5	44.48	46.61	44.63	45.08	33.13
FGE-SSV（本文）	AGAIN-VC	17.82	44.32	46.92	43.12	43.17	46.6
	FreeVC	38.71	14.24	38.61	37.2	34.78	35.39
	MediumVC	46.14	39.32	15.78	38.11	35.84	45.69
	StyleTTS	42.84	39.47	42.11	20.7	38.26	46.6
	TriAAN-VC	35.69	40.53	39.09	35.84	15.03	43.27
	VQMIVC	46.6	39.92	39.32	37.66	46.6	21.73

EER 平均为 21.70%，低于 VC-SID 的 30.16% 以及 SSTC 基准模型的 24.32% 但略高于 REVELIO 的 18.94%；采用 TriAAN-VC 模型进行训练时，其他 5 种模型的 *EER* 平均为 23.62%，均低于 VC-SID 的 35.76% 和 REVELIO 的 51.68% 以及 SSTC 基准模型的 27.02%；采用 FreeVC 模型进行训练时，其他 5 种模型的 *EER* 平均为 33.82%，均低于 VC-SID 的 45.76% 和 REVELIO 的 40.76% 以及 SSTC 基准模型的 38.06%。

由表 7 可以看出，所提 FGE-SSV 模型在未见过的语音转换方法上也展现出良好的泛化性能。具体来说，采用 AGAIN-VC 模型进行训练时，其他 5 种模型的 *EER* 平均为 44.83%，均低于 VC-SID 的 47.26% 和 REVELIO 的 51.12% 以及 SSTC 基准模型的 46.62%；采用 FreeVC 模型进行训练时，在其他 5 种模型的 *EER* 平均为 36.94%，均低于 VC-SID 的 44.96% 和 REVELIO 的 49.96% 以及 SSTC 基准模型的 38.22%；采用 MediumVC 模型进行训练时，其他 5 种模型的 *EER* 平均为 41.02%，均低于 VC-SID 的 44.84% 和 REVELIO 的 47.86% 以及 SSTC 基准模型的 42.26%；采用 StyleTTS 模型进行训练时，其他 5 种模型的 *EER* 平均为 41.85%，均低于 VC-SID 的 48.74% 和 REVELIO 的 46.94% 以及 SSTC 基准模型的 47.14%；采用 TriAAN-VC 模型进行训练时，在其他 5 种模型的 *EER* 平均为 38.88%，均低于 VC-SID 的 44.78% 和 REVELIO 的 53.00% 以及 SSTC 基准模型的 40.75%；采用 VQMIVC 模型进行训练时，在其他 5 种模型的 *EER* 平均为 42.02%，均低于 VC-SID 的 46.81% 和 REVELIO 的 54.70% 以及 SSTC 基准模型的 45.66%。

需要指出的是，本文 FGE-SSV 模型在 SSTC2024 官网数据集上的实验结果普遍比在 VCSD-SSV 数据集上的性能下降，这可能是因为数据集的构造方式不同。VCSD-SSV 数据集是成对式构造，即 1 个源说话人对应 1 个目标说话人，而 SSTC2024 官网数据集是针对每段目标语音，随机选取 3 个源语音样本进行

语音转换，模拟 3 个不同攻击者对同一目标语音发起的攻击。而实验中所选数据是根据源说话人来进行选择，SSTC2024 官网数据集中所选实验数据中 1 个源说话人中的语音样本会对应多个目标说话人，因此 1 对 1（1 个源说话人对应 1 个目标说话人）的语音样本会大幅减少，而有效数据较少可能导致模型训练从转换语音中提取源说话人信息的效果降低。

5.5 开集有效性分析

为测试所提 FGE-SSV 模型在开集设定下的有效性，用开集设定下的测试集进行测试，实验结果如表 8 所示。

从表 8 可以看出，与表 6 相比，在开集条件下，所提 FGE-SSV 模型无论在域内还是在未见过的语音转换方法上，泛化性能都出现明显的下降，表明 SSV 方法在开集条件下仍具有很大的挑战性。但与已有方法相比，所提 FGE-SSV 模型在域内依然保持着较低的 *EER*，且依然对未见语音转换方法保持最优平均泛化性能。具体来说，采用 Seed-VC 模型进行训练时，在其他 5 种模型的 *EER* 平均为 42.28%，均低于 VC-SID 的 47.1% 和 REVELIO 的 57.40% 以及 SSTC 基准模型的 45.78%；采用 FACodec 模型进行训练时，在其他 5 种模型的 *EER* 平均为 42.64%，均低于 VC-SID 的 49.3% 和 REVELIO 的 66.24% 以及 SSTC 基准模型的 46.12%；采用 DDDM-VC 模型进行训练时，在其他 5 种模型的 *EER* 平均为 39.22%，均低于 VC-SID 的 44% 和 REVELIO 的 51.3% 以及 SSTC 基准模型的 43.7%；采用 Diff-HierVC 模型进行训练时，在其他 5 种模型的 *EER* 平均为 40.84%，低于 VC-SID 的 49.84% 和 REVELIO 的 42.5% 以及 SSTC 基准模型的 45.5%；采用 TriAAN-VC 模型进行训练时，在其他 5 种模型的 *EER* 平均为 39.08%，均低于 VC-SID 的 43.88% 和 REVELIO 的 57.42% 以及 SSTC 基准模型的 47.18%；采用 FreeVC 模型进行训练时，在其他 5 种模型的 *EER* 平均为 45.12%，均低于 VC-SID 的 54.86% 和

Table 8 *EER* Results of Different Models on VCSD-SSV Dataset in Open-Set Generalisation Experiment

表 8 各模型在 VCSD-SSV 数据集上的开集泛化实验 *EER* 结果 %

网络模型	数据集	Seed-VC	FACodec	DDDM-VC	Diff-Hier VC	TriAAN-VC	FreeVC
VC-SID	Seed-VC	17.6	45	50.7	50.2	49.2	40.4
	FACodec	54.9	20.3	52.5	51.1	48.6	39.4
	DDDM-VC	45.9	44	17.4	43.7	39.7	46.7
	Diff-HierVC	51.2	43.9	45.3	16.1	61.6	47.2
	TriAAN-VC	50.0	41.1	46.1	45.0	17.8	37.2
	FreeVC	59.6	56.4	55.1	52.6	50.6	13.3
REVELIO	Seed-VC	34.8	61.5	63.2	50.3	58.5	53.5
	FACodec	56.1	40.7	66.5	66	69.7	72.9
	DDDM-VC	53.2	53.3	38.6	54.2	48.8	47.0
	Diff-HierVC	37.2	35.8	44.7	13.3	52.0	42.8
	TriAAN-VC	55.0	56.2	57.6	55.3	46.8	63.0
	FreeVC	53.2	55.0	58.8	55.6	59.8	33.3
SSTC 基准模型	Seed-VC	21.0	48.7	49.0	44.7	44.2	42.3
	FACodec	51.2	29.5	45.3	43.0	51.5	39.6
	DDDM-VC	46.5	38.8	20.1	49.8	49.6	33.8
	Diff-HierVC	47.0	40.5	41.1	14.5	57.3	41.6
	TriAAN-VC	52.0	41.8	54.2	48.2	29.3	39.7
	FreeVC	52.5	40.3	51.6	50.9	48.1	32.3
FGE-SSV（本文）	Seed-VC	14.9	39.8	47.9	46.3	36.6	40.8
	FACodec	42.6	17.5	44.2	42.3	48.5	35.6
	DDDM-VC	45.2	34.3	17.5	47.9	37.1	31.6
	Diff-HierVC	43.8	35.2	48.0	15.4	41.8	35.4
	TriAAN-VC	40.3	33.7	42.4	50.9	11.8	28.1
	FreeVC	48.1	38.9	54.6	42.2	41.8	11.5

REVELIO 的 56.48% 以及 SSTC 基准模型的 48.68%。

值得注意的是, 所提 FGE-SSV 模型在开集情况下的 *EER* 明显提高, 可能的原因是: 受限于单卡 RTX 4060Ti(8 GB)显存, 训练时仅使用 20 名说话人(约 5 000 条语音)进行训练, 所见说话人特征空间远小于常规闭集方案。说话人先验分布的稀疏性导致模型在遇见开集全新说话人时, 无法充分匹配残留源特征, 从而 *EER* 会明显上升。

5.6 讨 论

本文提出的 FGE-SSV 模型在 VCSD-SSV 与 SSTC 2024 官网数据集上均取得显著性能优势, 其在训练效果、泛化能力及与基线方法的比较中表现突出。这主要得益于以下设计: FEM 模块增强了源说话人特征的表达能力, IBN-SE 与 SCConv 模块提升了模型泛化性能, 而双分支注意力融合机制则有效兼顾了对关键声纹信息的聚焦与泛化特征的融合, 从而确保模型能够提取区分度更高的源说话人特征。

然而, FGE-SSV 模型在部分特定场景下仍存在局限, 主要表现在以下 2 方面:

1)低信噪比环境。由于所用训练数据 VCSD-SSV 基于高质量纯净语音构建, 模型在噪声干扰较强的真实场景中表现可能下降, 说明其对声学环境变化的鲁棒性仍有提升空间。

2)音色极度相似的说话人。当源说话人与目标说话人音色高度接近(如双胞胎、同性别同年龄段)时, 模型判别性能显著降低。分析表明, 此类转换语音中目标说话人特征占据主导, 源说话人声学痕迹残留极少, 导致 FEM 模块难以有效捕捉和分离出微弱的源特征信息。

这些情况反映出当前模型在特征解耦与弱痕迹提取方面仍存在挑战, 尤其是在声纹信息被严重掩蔽或覆盖的场景下, 模型架构对极微弱源特征的感知与增强机制尚需进一步优化。

6 结 论

本文围绕语音转换数据集构建与源说话人验证方法展开探索性工作, 尝试将人工智能技术应用于源说话人的验证工作中。主要工作包括: 采集并构建了语音转换数据集 VCSD-SSV, 并对其进行了细致的标注、转录和清洗等预处理工作; 提出了基于特征泛化增强的源说话人验证方法 FGE-SSV。实验结果表明, FGE-SSV 模型在 VCSD-SSV 数据集和 SSTC 2024 官网数据集上的表现相对优于现有模型, 其源说话人验证效果在准确度和泛化性方面都取得了一定提升。具体而言, FGE-SSV 方法在 VCSD-SSV 数据集全部 6 种 VC 模型中域内结果保持稳定的低 *EER*(0.2~1.0 个百分点), 相较于对比方法 *EER* 平均下降 70 个百分点, 展现出明显的优势; 泛化性方面, FGE-SSV 模型在未见过的语音转换方法上展现出良好的泛化性能, 均优于对比方法。FGE-SSV 方法在 SSTC2024 官网数据集 6 种 VC 模型上也保持较低的 *EER*(15.03~21.73 个百分点), 相较于 SSTC2024 基准模型平均下降 20 个百分点; 泛化性方面, 也具有明显优势。这些初步尝试在源说话人验证的高准确度和泛化性方面取得了一定进展, 为自动说话人验证系统的保护提供了新的技术思路。

但是, 当前方法侧重于在可控环境下验证模型核心架构的有效性, 仍有一些问题需要进一步验证和完善。首先, VCSD-SSV 数据集的规模相对有限, 在方言、低信噪比及跨语言场景下的验证不足, 可能影响模型的训练效果和泛化能力。其次, 现有数据集对最新发展的 VC 模型覆盖可能不够全面, 因此模型在遇到从未见过的 VC 模型的泛化性仍有待改进。此外, 目前对模型性能的评估方法还不够系统, 缺乏更科学、更细致的评价体系。这些局限性都可能对研究结果的全面性产生一定影响。针对当前研究的不足, 未来工作拟从以下 2 个方面进行改进: 首先, 计划将 VCSD-SSV 数据集规模扩展至 10 个 VC 模型以上, 并开展方言、低信噪比及跨语言场景下的验证, 以提升模型的泛化能力; 其次, 将探索泛化性更强的源说话人验证方法, 进一步提高 FGE-SSV 遇到未知 VC 模型时的泛化性。我们希望通过这些改进, 能够逐步提升 FGE-SSV 的性能, 使其更好地保护自动说话人验证系统。

作者贡献声明: 张国富、田天才、苏兆品负责方案整体设计并撰写论文; 王垚飞提出指导意见并修改论文; 段宇衡、周康健、罗永琪负责部分数据集的制作。

参 考 文 献

[1] Li Zeyu, Zhang Xuhong, Pu Yuwen, et al. A survey on multimodal deepfake and detection techniques[J]. Journal of Computer Research and Development, 2023, 60(6): 1396−1416 (in Chinese)
(李泽宇, 张旭鸿, 蒲誉文, 等. 多模态深度伪造及检测技术综述[J]. 计算机研究与发展, 2023, 60(6): 1396−1416)

[2] Zhang Xiaohui, Yi Jiangyan, Tao Jianhua, et al. Elastic orthogonal weight modification continual learning algorithm in the context of synthetic speech detection[J]. Journal of Computer Research and Development, 2025, 62(2): 336−345 (in Chinese)
(张晓辉, 易江燕, 陶建华, 等. 面向合成语音检测的弹性正交权重修正连续学习算法[J]. 计算机研究与发展, 2025, 62(2): 336−345)

[3] Liu Xuechen, Wang Xin, 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

[4] Fan Cunhang, Ding Minging, Tao Jianhua, et al. Dual-branch knowledge distillation for noise-robust synthetic speech detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 2453−2466

[5] Wang Xin, Yamagishi J, Todisco M, et al. ASVspoof2019: A large-scale public database of synthesized, converted and replayed Speech[J]. Computer Speech & Language, 2020, 64: 101114

[6] Yamagishi J, Wang Xin, Todisco M, et al. ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection[C]// Proc of the 4th Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. Red Hook, NY: Curran Associates Inc, 2021: 47−54

[7] Pierno A D, Guarnera L, Allegra D, et al. End-to-end audio deepfake detection from RAW waveforms: A RawNet-based approach with cross-dataset evaluation[C]//Proc of Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2025: 1−8

[8] McLaren M, Ferrer L, Castan D, et al. The speakers in the wild (SITW) speaker recognition database[C]//Proc of the 17th Annual Conf of the Int Speech Communication Association. Red Hook, NY: Curran Associates Inc, 2016: 818−822

[9] Khalid H, Tariq S, Woo S S. FakeAVCeleb: A novel audio-video multimodal deepfake dataset[C]//Proc of the 35th Conf on Neural Information Processing Systems Track on Datasets and Benchmarks. Red Hook, NY: Curran Associates Inc, 2021: 1−14

[10] Cai Danwei, Cai Zexin, Li Ming. Identifying source speakers for voice conversion based spoofing attacks on speaker verification systems[C]//Proc of the 48th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2023: 1−5

[11] Deng Jiangyi, Chen Yanjiao, Zhong Yinan, et al. Catch you and I can: Revealing source voiceprint against voice conversion[C]//Proc of the 32nd USENIX Security Symp. Berkeley, CA: USENIX Association, 2023: 5163−5180

[12] Ma Xinlei, Lu Wenhuan, Zhang Ruiteng, et al. Distillation-based feature extraction algorithm for source speaker verification[C]//Proc of IEEE Spoken Language Technology Workshop. Piscataway, NJ: IEEE, 2024: 1240−1246

[13] Li Ze, Lin Yuke, Yao Tian, et al. The database and benchmark for the source speaker tracing challenge 2024[C]//Proc of IEEE Spoken Language Technology Workshop. Piscataway, NJ: IEEE, 2024: 1254−1262

[14] Panayotov V, Chen Guoguo, Povey D, et al. LibriSpeech: An ASR corpus based on public domain audio books[C]//Proc of the 19th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2015: 5206−5210

[15] Nagrani A, Chung J S, Zisserman A. VoxCeleb: A large-scale speaker identification dataset[C]//Proc of the 18th Annual Conf of the Int Speech Communication Association. Red Hook, NY: Curran Associates Inc, 2017: 2616−2620

[16] Liu Songting. Zero-shot voice conversion with diffusion transformers[J]. arXiv preprint, arXiv: 2411.09943, 2024

[17] Ju Zeqian, Wang Yuancheng, Shen Kai, et al. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models[C]// Proc of the 41st Int Conf on Machine Learning. Norfolk, MA: JMLR. org, 2024: 22605−22623

[18] Choi H, Lee S, Lee S. DDDM-VC: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 17862−17870

[19] Choi H, Lee S, Lee S. Diff-HierVC: Diffusion-based hierarchical voice Conversion with robust pitch generation and masked prior for zero-shot speaker adaptation[C]//Proc of the 24th Annual Conf of the Int Speech Communication Association. Red Hook, NY: Curran Associates Inc, 2023: 2283−2287

[20] Park H J, Yang S W, Kim J S, et al. TriAAN-VC: Triple adaptive attention normalization for any-to-any voice conversion[C]//Proc of the 48th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2023: 1−5

[21] Li Jingji, Tu Weiping, Xiao Li. FreeVC: Towards high-quality text-free one-shot voice conversion[C]//Proc of the 48th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2023: 1−5

[22] Pan Xingang, Luo Ping, Shi Jianping, et al. Two at once: Enhancing learning and generalization capacities via IBN-Net[C]//Proc of the 15th European Conf on Computer Vision. Cham: Springer, 2018: 464−479

[23] Zhou Junzuo, Yi Jiangyan, Tao Jianhua, et al. Neural speech codec method based on mel spectrogram and squeeze-excitation-weighted quantization[J]. Journal of Computer Research and Development, 2025, 62(11): 2725−2736 (in Chinese)
(周俊佐, 易江燕, 陶建华, 等. 基于梅尔谱与压缩激励加权量化的语音神经编解码方法[J]. 计算机研究与发展, 2025, 62(11): 2725−2736)

[24] Su Zhaopin, Wu Zhangqian, Yue feng, et al. Source cell-phone identification under background noise based on low-dimensional deep features[J] Acta Electronica Sinica, 2021, 49(4): 637−646 (in Chinese)
(苏兆品, 吴张倩, 岳峰, 等. 自然环境背景噪声下基于低维深度特征的手机来源识别[J]. 电子学报, 2021, 49(4): 637−646)

[25] Zhang Yin, Ye Mu, Zhu Guiyi, et al. FFCA-YOLO for small object detection in remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5611215

[26] Li Jiafeng, Wen Ying, He Lianghua. SCConv: Spatial and channel reconstruction convolution for feature redundancy[C]//Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition.

Piscataway, NJ: IEEE, 2023: 6153−6162

[27] Hu Jie, Shen Li, Sun Gang. Squeeze-and-excitation networks[C]// Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7132−7141

[28] Chen Y, Wu D, Wu T, et al. Again-VC: A one-shot voice conversion using activation guidance and adaptive instance normalization[C]// Proc of the 46th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2021: 5954−5958

[29] Gu Yewei, Zhang Zhenyu, Yi Xiaowei, et al. MediumVC: Any-to-any voice conversion using synthetic specific-speaker speeches as intermedium features[J]. arXiv preprint, arXiv: 2110.02500, 2021

[30] Li Y A, Han Cong, Mesgarani N. StyleTTS: A style-based generative model for natural and diverse text-to-speech synthesis[J]. IEEE Journal of Selected Topics in Signal Processing, 2025, 19(1): 283−296

[31] Wang Disong, Deng Liqun, Yeung Y T, et al. VQMIVC: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion[C]// Proc of the 22nd Annual Conf of the Int Speech Communication Association. Red Hook, NY: Curran Associates Inc, 2021.1344−1348

[32] Healy J, McInnes L. Uniform manifold approximation and projection[J]. Nature Reviews Methods Primers, 2024, 4(1): 82



Zhang Guofu, born in 1979. PhD, professor. His main research interests include artificial intelligence and audio security.

张国富, 1979 年生。博士, 教授。主要研究方向为人工智能、语音安全。



Tian Tiancai, born in 2001. Master candidate. His main research interest includes deepfake audio detection and source tracing.

田天才, 2001 年生。硕士研究生。主要研究方向为深伪语音检测与溯源。



Su Zhaopin, born in 1983. PhD, associate professor. Her main research interest includes network and information security.

苏兆品, 1983 年生。博士, 副教授。主要研究方向为网络与信息安全。



Wang Yaofei, born in 1996. PhD, associate professor. His main research interests include information security and artificial intelligence security.

王垚飞, 1996 年生。博士, 副教授。主要研究方向为信息安全、人工智能安全。



Duan Yuheng, born in 2004. Undergraduate. His main research interest includes network and information security.

段宇衡, 2004 年生。本科生。主要研究方向为网络与信息安全。



Zhou Kangjian, born in 2005. Undergraduate. His main research interest includes network and information security.

周康健, 2005 年生。本科生。主要研究方向为网络与信息安全。



Luo Yongqi, born in 2005. Undergraduate. His main research interest includes network and information security.

罗永琪, 2005 年生。本科生。主要研究方向为网络与信息安全。