中图法分类号:TP18;TP183　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)01-0167-10

论文引用格式:Su Z P, Wei Y Y, Zhang G F, Lian C S and Yue F. 2026. Target speaker verification method for cloned speech. Journal of Image and Graphics, 31(1):0167-0176(苏兆品, 魏玉洋, 张国富, 廉晨思, 岳峰. 2026. 面向克隆语音的目标说话人鉴别方法. 中国图象图形学报, 31(1):0167-0176)[DOI:10.11834/jig.240686]

面向克隆语音的目标说话人鉴别方法

苏兆品[1,2,3],魏玉洋[1],张国富[1,2,3*],廉晨思[3,4],岳峰[1,2]

1. 合肥工业大学计算机与信息学院,合肥　230601; 2. 智能互联系统安徽省实验室(合肥工业大学),合肥　230009;
3. 音视频智能防识联合实验室,合肥　230000; 4. 安徽省公安厅物证鉴定管理处,合肥　230000

摘　要: **目的**　随着文本到语音(text to speech,TTS)、语音转换(voice conversion,VC)等克隆语音技术的快速发展,如何在司法实践中准确识别克隆语音,即克隆语音是否来源于目标说话人特征,成为一个极具挑战性的难题。虽然现有说话人识别技术可以通过声纹特征比对确认自然语音的说话人身份,但由于克隆语音不仅与目标说话人音色相似,且又包含源说话人的特点,使得传统说话人识别技术难以去除源说话人音色的干扰,难以直接应用于深度克隆语音。基于此,研究了一种面向克隆语音的目标说话人鉴别方法。**方法**　基于Res2Block设计组渐进信道融合模块(group progressive channel fusion, GPCF),以有效提取自然语音与克隆语音之间的公共有效声纹特征信息;采用基于K独立的动态全局滤波器(dynamic global filter, DGF),以有效抑制源说话人的影响,提高模型表征和泛化能力;利用基于多尺度层注意力的特征融合机制,以有效融合不同层次GPCF模块和DGF模块的深浅层特征;使用注意力统计池(attentive statistics pooling,ASP)层,进一步增强表示特征张量中的目标说话人信息。**结果**　在所设计的数据集上与3种较新的方法进行实验比较,相对于其他3种方法,本文方法等错误率(equal error rate,EER)分别降低了1.38%、0.92%和0.61%,最小检测代价函数(minimum detection cast function,minDCF)分别降低了0.012 5、0.006 7和0.044 5。**结论**　在FastSpeech2(fast and high-quality end-to-end text to speech)、TriAANVC(triple adaptive attention normalization for any-to-any voice conversion)、FreeVC(high-quality text-free one-shot voice conversion)和KnnVC(nearest neighbors voice conversion)共4种语音克隆数据集上的对比实验结果表明,所提方法在处理面向克隆语音的声纹认定任务时更具有优势,可以有效提取克隆语音中的目标说话人特征,为克隆语音的声纹认定提供方法指导。

关键词: 克隆语音;声纹认定;组渐进信道融合(GPCF);动态全局滤波器(DGF);多尺度层注意力机制

Target speaker verification method for cloned speech

Su Zhaopin[1,2,3], Wei Yuyang[1], Zhang Guofu[1,2,3*], Lian Chensi[3,4], Yue Feng[1,2]

1. School of Computer and Information Technology, Hefei University of Technology, Hefei　230601, China;
2. Intelligent Interconnected Systems Laboratory of Anhui Province (Hefei University of Technology), Hefei　230009, China;
3. Joint Laboratory of Intelligent Prevention and Recognition of Audio and Video, Hefei　230000, China;
4. Department of Physical Evidence Identification, Anhui Public Security Department, Hefei　230000, China

Abstract: **Objective** The rapid development of artificial intelligence technology, especially the revolutionary progress in the field of deep learning, has boosted voice cloning technologies such as text to speech (TTS) and voice conversion (VC)

收稿日期:2024-11-12;**修回日期**:2025-02-20;**预印本日期**:2025-02-27

*通信作者:张国富　zgf@hfut.edu.cn

基金项目:教育部人文社会科学研究规划基金项目(24YJA870011);安徽省重点研究与开发计划项目(202104d07020001)

Supported by: MOE(Ministry of Education in China)Project of Humanities and Social Sciences(24YJA870011); Anhui Provincial Key R&D Program, China(202104d07020001)

beyond traditional limitations. This innovation enables the rapid generation of digital voices that "sound like one's own voice", which offers users unprecedented convenience and enjoyment. However, malicious attackers may use their cloned voices to commit unlawful activities such as fraud, forgery, dissemination of false information, targeted harassment, and illegal access to personal sensitive information. These behaviors will undoubtedly pose an unprecedented challenge and threat to personal voice rights. Therefore, under the clear guidance of the law, the protection of voice rights as an important part of personal identity has become the focus of attention from various sectors of society. Accurately determining the sound source of the cloned speech in judicial practice, that is, ascertaining whether the cloned speech originates from the characteristics of the target speaker, has become a challenging dilemma. Although the existing speaker identification technology can confirm the speaker identity of human speech through voiceprint feature comparison, the cloned speech is not only similar to the timbre of the target speaker but also contains the characteristics of the source speaker. Therefore, the interference of the timbre of the source speaker makes the traditional speaker recognition technology difficult to be directly applied to the cloned speech. Against this backdrop, the target speaker verification method for cloned speech is proposed in this study. **Method** Specifically, the group progressive channel fusion (GPCF) module is first designed based on Res2Block to extract the common effective voiceprint features between human speech and cloned speech. This module is critical for distinguishing subtle differences between human speech and cloned speech that are generally indistinguishable to the human ear alone. Subsequently, the K-independent dynamic global filter (DGF) module is adopted to suppress the influence of the source speaker, which improves the representation and generalization capability of the model. The DGF plays a key role in filtering out the unwanted features of the source speaker, which ensures that the voiceprint of the target speaker remains the focus of the analysis. Thereafter, a feature aggregation mechanism is presented based on multi-scale layer attention to fuse deep and shallow features from different levels of the GPCF and DGF modules. This mechanism is designed to capture the intricate details of the voiceprint that may be missed by traditional methods, which provides a more comprehensive analysis of the speech features. Finally, the attentive statistics pooling (ASP) layer is used to further enhance the target speaker information in the representation feature tensor. The ASP layer is instrumental in refining the feature set. This refinement ensures that the most relevant and discriminative features are emphasized, which improves the accuracy of the voiceprint identification results. **Result** Experiments on the AISHELL3 human speech dataset and the cloned speech dataset, which are created based on four speech synthesis and conversion algorithms——FastSpeech2, TriAANVC, FreeVC, and KnnVC——how that the equal error rates are reduced by 1.38%, 0.92%, and 0.61%, while the values for the minimum detection cost function are decreased by 0.012 5, 0.006 7, and 0.044 5, compared with the three advanced methods. **Conclusion** These comparative experimental results show that the proposed method can significantly reduce the error rate associated with voiceprint identification, which leads to more reliable and accurate results. They also show the advantages of the proposed method in solving the task of cloned speech voiceprint identification. The proposed method effectively extracts the features of the target speaker from cloned speech, which offers a methodological guide for the voiceprint identification of cloned speech.

Key words: cloned speech; voiceprint identification; group progressive channel fusion (GPCF); dynamic global filter (DGF); multi-scale layer attention

0 引　言

随着人工智能技术的迅猛发展,特别是深度学习领域的革命性进步,文本到语音(text to speech, TTS)、语音转换(voice conversion, VC)等语音克隆技术已经跨越了传统技术的界限(李嘉欣 等,2023;许裕雄 等,2024),能够迅速生成"听起来像自己的声音"的数字语音,不仅为用户带来了前所未有的便捷与乐趣,同时也为不法分子利用其伪造的声音,进行诈骗、伪造身份、散布虚假信息、恶意骚扰以及非法获取个人敏感信息等带来了可乘之机,对个人声音权构成了前所未有的挑战与威胁(乔喆,2023)。在《中华人民共和国民法典》的明确指引下,声音权(尹懋锟和李云滨,2024)作为个人身份的重要组成部分,其保护已经成为社会各界关注的焦点。

在司法实践以及其他应用中,通常会对音频做伪造检测来识别内容的真实性以及防范各种语音攻

击手段对说话人验证系统的攻击，但是尚未有对已确定是克隆语音的目标说话人的声纹认定研究，难以解决民法典中的声音侵权问题。从技术角度看，克隆语音的声纹认定需要判定克隆语音中的声纹特征是否与目标说话人音色特征高度相似，而说话人识别技术提供了可借鉴的解决思路。然而，传统的说话人识别技术，如特征参数分析、动态时间规整（Booth 等，1993）、矢量量化（Gersho 和 Gray，1992）以及深度学习（江铃燚 等，2023；Huang 等，2017）等，均是面向自然语音，判定待检测的自然语音是否来源于目标说话人。而对于基于 TTS、VC 等方法的克隆语音，由于其在合成过程中融入目标说话人的音色特征，不仅与目标说话人音色相似，且又包含源说话人的特点，使得传统说话人识别技术难以去除源说话人音色的干扰，捕捉到克隆语音与自然语音之间的音色共同点差异。因此，如何准确判断克隆语音的声纹已成为亟需解决的难点问题。为此，本文在深入分析说话人识别和克隆语音合成技术的基础上，提出一种面向克隆语音的目标说话人鉴别方法，以深入挖掘并有效区分自然人声与克隆语音之间的公共及特有声纹特征，提高克隆语音声纹认定的准确性和鲁棒性。

1　研究动机

首先介绍已有的说话人识别技术，然后分析克隆语音的合成过程及其声纹特点，以说明研究工作的必要性和挑战性。

1.1　说话人识别技术

随着深度学习技术的发展，说话人识别技术利用深度神经网络强大的特征提取能力，提高说话人识别系统的准确性与稳定性。

Huang 等人（2017）设计了一种密集卷积网络架构用于提取声纹特征，让每一层网络都能够整合并累积之前所有层级的输出，实现了特征信息的高效传递和最大化利用。Malykh 等人（2017）引入深度残差网络（deep residual network，ResNet），应用于说话人语谱图的深度学习训练，专门用于捕捉和提炼说话人的独特性特征。Snyder 等人（2017）设计了一种基于时延神经网络（time delay neural network，TDNN）的说话人识别方法，能够将不同时间长度的输入语音帧转换为具有一致维度的语句级特征表示，增强了短时长语音在说话人识别任务中的稳定性和可靠性。Lee 和 Nam（2017）以及 Gao 等人（2019）提出利用复杂网络结构，提取与说话人身份密切相关的深层次特征，也指出浅层特征映射有助于更稳健的说话人嵌入。Desplanques 等人（2020）提出 ECAPA-TDNN（emphasized channel attention，propagation and aggregation in TDNN based speaker verification）模型，融合了通道注意力机制和 SE（squeeze-excitation）模块（Hu 等，2018），利用压缩激励技术与多尺度特征的加权融合策略，有效地提升了说话人识别的精确度。Li 等人（2024）提出一种基于全局感知滤波器（global filter，GF）的双流 TDNN（dual-stream TDNN，DS-TDNN）架构，通过两个分支并行提取并融合局部与全局特征。Yu 和 Li（2020）提出一种密集连接时延神经网络（densely connected TDNN，D-TDNN），通过引入瓶颈层和密集连接，减少参数数量的同时，保持了识别精度提升。Wang 等人（2023）提出上下文感知屏蔽（context-aware masking++，CAM++）架构，结合上下文感知掩码和多粒度池化的高效网络，适用于对推理速度和资源有限制的场景。

1.2　克隆语音的声纹特征分析

克隆语音技术主要包括 TTS 和 VC 两大类。TTS 是通过特定的深度学习算法对文本内容进行处理、分析，通过语音合成系统将文字转换成目标说话人的语音技术，代表算法是 FastSpeech2（Ren 等，2022），原理如图 1 所示。

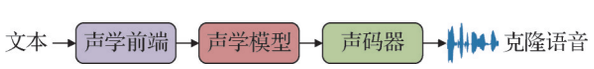

图 1　TTS 原理图

Fig. 1　TTS schematic

VC 利用深度学习技术，保留原始语音的内容，同时改变说话人的声学特征，如音色、音高和音长等，以达到模仿另一个说话人的声音的效果，转换原理如图 2 所示，代表算法主要有 TriAANVC（Park 等，2023）、FreeVC（Li 等，2023）和 KnnVC（Baas 等，2023）。

为了分析克隆语音中的声纹特征，本文首先基

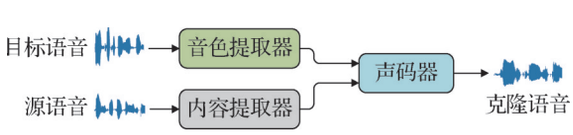

图 2　VC 原理图

Fig. 2　VC schematic

于AISHELL3(Shi等,2021)高保真中文语音数据库中130个说话人的自然语音,分别利用FastSpeech2、TriAANVC、FreeVC和KnnVC共4种算法,为每位说话人合成100条克隆语音。然后利用梅尔滤波倒谱系数(Mel-scale frequency cepstral coefficients, MFCC)声纹特征,分别提取克隆语音特征、源说话人特征以及目标说话人特征,分别计算每条克隆语音与源语音和目标语音的余弦相似度。每种克隆语音与源语音和目标语音的平均相似度如表1所示。从表1可以看出,由于FastSpeech2属于TTS算法,不需要源语音的参与,其与目标语音的平均相似度为76.48%。TriAANVC、FreeVC和KnnVC属于VC方法,克隆语音与源语音的相似度在60%左右,而与目标语音的平均相似度在80%左右,特别是TriAANVC达到86.27%。因此,克隆语音虽然"听起来很像是"目标说话人,包含较多的目标说话人的特征信息,但同时也包含了源说话人的特征信息。综上,不同于自然语音只包含一种说话人特征,克隆语音作为人工智能的产物,可能包括多个互相干扰的说话人特征信息,使得现有针对自然语音的说话人识别技术难以直接应用于克隆语音的声纹认定。

表1　克隆语音与源语音和目标语音相似度分析

Table 1　Analysis of the similarity between cloned voices, source voices, and target voices

/%

算法	源语音	目标语音
FastSpeech2	-	76.48
TriAANVC	59.37	86.27
FreeVC	58.26	79.19
KnnVC	60.79	81.33

注:"-"表示无相关信息。

2　方　法

面向克隆语音的目标说话人鉴别方法(target speaker verification method for cloned speech, TSVCS)整体结构如图3所示。先对输入的语音进行MFCC特征提取,接着将输出的特征张量交替输入到组渐进信道融合(group progressive channel fusion, GPCF)模块和动态全局滤波器(dynamic global filter, DGF)模块,每次的输出结果也同样输入到下一层,然后将训练过程中经过的所有GPCF模块和DGF模块按先后顺序拼接起来经过多尺度特征聚合层(multiscale feature aggregation, MFA)得到说话人表示,并将其输送给注意力统计池(attentive statistics pooling, ASP)(Okabe等,2018)加权,最后通过全连接层得到192维的说话人嵌入特征。在进行说话人认定时,对于待检测的一条自然语音和一条克隆语音,借助训练完成的模型提取说话人特征,再计算余弦相似度,对结果与初设的阈值进行比较,若大于阈值则可判断输入的克隆语音源自于目标说话人,否则判定为不是。

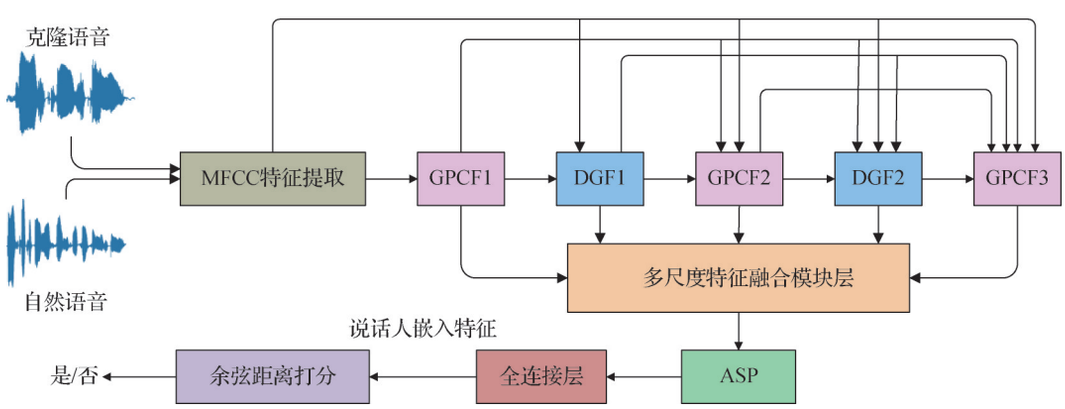

图3　TSVCS方法总体结构图

Fig. 3　Overall structure of the TSVCS method

2.1　GPCF模块

为了更有效地提取自然语音与克隆语音之间的公共有效的声纹特征信息,本文在Res2Block的基础上,提出一种组渐进信道融合模块(GPCF)结构,能够同时获得不同大小的感受野,获得更多的相关特征信息,在避免过拟合的同时,将传感器空间从局部扩展到深度块中的整个特征图,如图4所示。

图4中,$\boldsymbol{Z}$表示输入的特征张量,其在通道上被

拆分成s组，$\boldsymbol{Z}=[z_1,z_2,\cdots,z_s]$，$z_i$表示每一个被分割后的特征张量，$i\in\{1,2,\cdots,s\}$表示初始组别，$s$表示初始组别最大值。经过初始分组处理后每组有$a$个通道，$\lambda$表示整体输入通道数量，则$\lambda=a\times s$。

对于每一组通道C_i的输出$\boldsymbol{y}_i$均可表示为

$$\boldsymbol{y}_i=\begin{cases}C_i(z_i+z_{i+1}) & i=1\\ C_i(\boldsymbol{y}_{i-1}+z_i+z_{i+1}) & 1<i<s\\ C_i(\boldsymbol{y}_{i-1}+z_i) & i=s\end{cases}\tag{1}$$

也就是说，对于当前分组z_i，与相邻的下一个分组z_{i+1}和相邻的前一个分组的输出$\boldsymbol{y}_{i-1}$进行合并，应用ResNet34的stage1和stage3模块分别进行卷积操作，用以学习到不同感受野大小的局部特征，最后按照原始分组的顺序依次对所有的分组进行拼接，形成完整的特征张量。

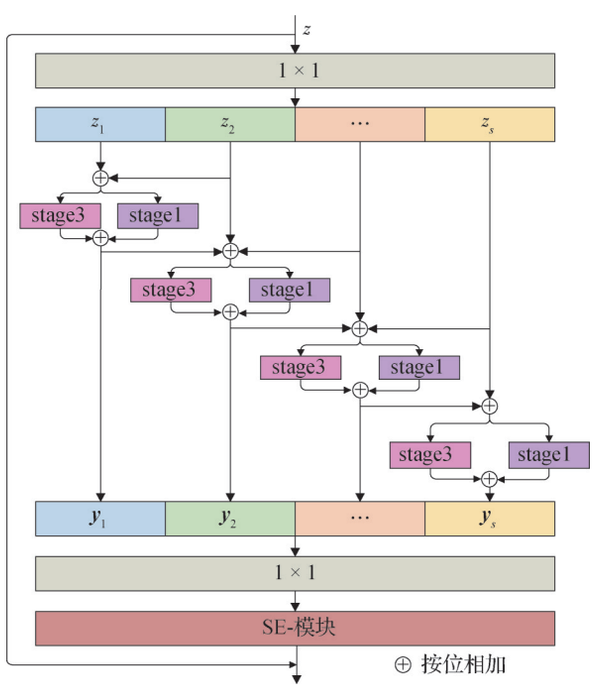

图4　GPCF结构图

Fig. 4　Structure diagram of GPCF

由于SE模块能够学习特征权重，并能够增大有效特征权重、减小无效或效果小的特征权重，因此，将特征张量经过GPCF结构之后再通过SE模块去关注更有效的信息，得到最终的GPCF模块特征。对于不同层次的GPCF层，使用渐进信道融合策略（Zhao等，2023），让模型在前向传播中逐渐融合信道关系。实际的GPCF模块层参数如表2所示，其中λ和数字分别表示输入通道大小、卷积核大小、分块数目和膨胀大小。

表2　GPCF内部参数结构

Table 2　Internal parameters of GPCF

层级	结构
GPCF1	(λ, 3, 8, 1)
GPCF2	(λ, 3, 4, 2)
GPCF3	(λ, 3, 2, 3)

2.2　DGF模块

为了有效去除源说话人特征的影响，优化克隆语音的特征表示，本文利用了动态全局滤波器（DGF）模块，其中包括动态滤波器和正则化操作。

2.2.1　动态滤波器

克隆语音是人工智能的产物，来源和质量错综不一，声纹特征分布存在很大的差异，而且声纹特征与语言内容、情绪和通道等多种信息相互交织在一起，采用静态的全局感知滤波器和空间域滤波器难以去除源说话人特征的影响，无法提取有效的声纹特征。因此，本文采用基于K独立的动态滤波器策略，能够动态地适应输入，提高表征和泛化能力。具体而言，首先在调制期间应用K独立的全局感知滤波器$\boldsymbol{F}_j(j\in\{1,2,\cdots,K\})$，并使用元素求和进行组合，具体为

$$\tilde{\boldsymbol{X}}_r=\boldsymbol{F}_1\odot\boldsymbol{X}_r+\boldsymbol{F}_2\odot\boldsymbol{X}_r+\cdots+\boldsymbol{F}_K\odot\boldsymbol{X}_r\tag{2}$$

式中，$\boldsymbol{X}_r$是令牌$\boldsymbol{X}$在每个通道上分别执行一维快速傅里叶变换得到的，$\odot$表示按位相乘。然后，利用一维通道注意力函数生成一系列动态分数$\boldsymbol{m}=[m_1,m_2,\cdots,m_K]$，它可以适应输入令牌$\boldsymbol{X}$，具体为

$$\boldsymbol{m}=\underbrace{softmax(FC_2(ReLU(FC_1(GAP}_{AttentionFN}(\boldsymbol{X})))))\in\mathbf{R}^{1\times K}\tag{3}$$

式中，GAP是全局平均池，FC_1和FC_2分别代表两个完全连接层，且FC_1的神经元数与FC_2相等，设为$K(K=8)$。

最后，使用动态分数$\boldsymbol{m}=[m_1,m_2,\cdots,m_K]$对过滤器组合进行加权，具体为

$$\tilde{\boldsymbol{X}}_r=m_1\boldsymbol{F}_1\odot\boldsymbol{X}_r+m_2\boldsymbol{F}_2\odot\boldsymbol{X}_r+\cdots+m_K\boldsymbol{F}_K\odot\boldsymbol{X}_r\tag{4}$$

2.2.2　Dropout正则化

虽然动态滤波器可以生成更加通用和鲁棒的说话人表示，但随着优化对象的增多，不仅容易使模型陷入局部最小值，增加优化的难度，而且增加了参数数量，使模型更容易过拟合。为此，在使用动态滤波器方法的基础上使用了Dropout正则化（Hinton等，

2012)技术来调节DGF。具体而言,在每个DGF层之后都加入了一个Dropout层,并设置其比率为p。这意味着,在每次前向传播过程中,大约有占比为p的神经元输出被随机置为零,这样做有助于防止神经元之间过度依赖,从而提高模型对新数据的适应能力。值得注意的是,Dropout正则化只在训练阶段为了防止模型过拟合而进行。

2.3 MFA结构

为了有效利用不同层次GPCF模块和DGF模块提取的特征,采用多尺度特征融合(MFA)模块,将经过训练网络输出的每一层特征映射,同样按顺序拼接起来,利用多层信息的互补方法得到先前所有层的输出,最后将其投入注意力统计池计算特征权重。

假设一个GPCF模块视为L,一个DGF模块视为G,则多尺度特征聚合机制为

$$\boldsymbol{H} = Concat(\boldsymbol{Y}_{L1},\boldsymbol{Y}_{G1},\boldsymbol{Y}_{L2},\boldsymbol{Y}_{G2},\boldsymbol{Y}_{L3}), dim = 1 \tag{5}$$

式中,$\boldsymbol{Y}_{L1}$,$\boldsymbol{Y}_{G1}$,$\boldsymbol{Y}_{L2}$,$\boldsymbol{Y}_{G2}$,$\boldsymbol{Y}_{L3}$表示每一个GPCF层和DGF层的输出,$dim = 1$表示层维度,$\boldsymbol{H}$表示在层维度上进行叠加的新特征。

2.4 注意力统计池

注意力统计池(ASP)(Okabe等,2018)是对注意力机制和统计池化思想的结合,旨在更有效地表示和提取特征张量中的关键信息。ASP通过为不同层次的特征分配不同的权重,使得每一个层次对最终的特征表达做出不同贡献度,从而得到结合深浅层特征的最优结果。

具体而言,首先采用ASP对$\boldsymbol{H}$的各帧级特征$\boldsymbol{H}_t \in \mathbf{R}^{5\lambda \times 1}$进行加权,提取更鲁棒的说话人嵌入进行验证,具体为

$$e_t = \boldsymbol{v}^{\mathrm{T}} \tanh(\boldsymbol{W}\boldsymbol{H}_t + \boldsymbol{b}) + n \tag{6}$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{\tau=1}^{N}\exp(e_\tau)} \tag{7}$$

式中,$t \in \{1, 2, \cdots, N\}$,$N$为总帧数,$\boldsymbol{W} \in \mathbf{R}^{5\lambda \times 5\lambda}$,$\boldsymbol{v} \in \mathbf{R}^{5\lambda \times 1}$,$\boldsymbol{b} \in \mathbf{R}^{5\lambda \times 1}$,$n$是一个可学习的参数。

然后对分数进行归一化操作,再将其作为池化层权重,计算加权平均向量和加权标准差,具体为

$$\tilde{\boldsymbol{\mu}} = \sum_{t=1}^{N}\alpha_t \boldsymbol{H}_t \tag{8}$$

$$\tilde{\boldsymbol{\sigma}} = \sqrt{\sum_{t=1}^{N}\alpha_t \boldsymbol{H}_t \odot \boldsymbol{H}_t - \tilde{\boldsymbol{\mu}} \odot \tilde{\boldsymbol{\mu}}} \tag{9}$$

最后将加权平均值$\tilde{\boldsymbol{\mu}}$和加权标准差$\tilde{\boldsymbol{\sigma}}$连接起来则可得到ASP的输出。这样将加权标准差和加权平均向量中所提取的话语级特征结合起来,会在重要的帧上具有更高的权重,从而具有更强的判别效果。此外,由于最后的输出包含不同层次的语音数据特征,在一定程度上也能加快网络的收敛,缩短训练时间。

3 实 验

为了验证本文提出的TSVCS方法的有效性,与ECAPA-TDNN(emphasized channel attention, propagation and aggregation-TDNN)、DS-TDNN(dual-stream TDNN)、CAM++(context-aware masking++)等3种目前性能较好的说话人识别方法进行对比分析。

3.1 数据集与实验设置

面向克隆语音的声纹鉴定研究,目前为止没有可利用的公开数据集。因此,本文首先基于AISHELL3中文数据集构造克隆语音数据集CS-AISHELL3,如表3所示。

表3 CS-AISHELL3数据集

Table 3 CS-AISHELL3 dataset

来源	方法	训练集		测试集	
		数量/条	人数	数量/条	人数
AISHELL3	FastSpeech2	100	130	100	40
	TriAANVC	100		100	
	FreeVC	100		100	
	KnnVC	100		100	

对于训练数据集,通过使用FastSpeech2(fast and high-quality end-to-end text to speech)、TriAANVC(triple adaptive attention normalization for any-to-any voice conversion)、FreeVC(high-quality text-free one-shot voice conversion)、KnnVC(nearest neighbors voice conversion)等4种算法进行克隆语音合成,分别得到对应的克隆语音数据集,其中通过每个算法得到的克隆语音数据集的语音数据数量均相同,保证训练过程中不会因为数据的过多或过少而产生偏向性学习,然后将AISHELL3的自然语音数据集与得到的克隆语音数据集组合在一起作为训练数据集进行模型训练。训练集中共包含130个说话人,每个说话人除了AISHELL3中原有的自然语音

外，还包含400条克隆语音，共计101 428条。

测试数据集包含40个说话人，以一对的形式进行输入，一对是由一条克隆语音和一条自然语音组成，测试组共计23 720对。且若自然语音与克隆语音的id相同（说话人身份相同）则为正样本，标签设置为1；若自然语音与克隆语音的id不同（说话人身份不同）则为负样本，标签设置为0。此外，测试集与训练集不重叠，且测试集也保证每个算法克隆的语音数据数量均相同。

使用MFCC作为初始声纹特征，采用汉明窗，每帧25 ms，偏移10 ms，并对每个话语提取80维MFCC特征，且不进行语音活动检测。网络架构中的通道大小和瓶颈尺寸分别为512和256，说话人嵌入向量为192维。使用AAM-softmax（Deng等，2019；Xiang等，2019）作为loss函数，并将margin和scale分别设置为0.2和30，使用Adam（Kingma和Ba，2017）作为优化器，学习率呈指数递减。

在测试时，将整条语音输入到说话人编码器中获得说话人嵌入，之后使用余弦相似度计算分数，且说话人的话语之间的平均得分用于说话人确认的最终结果。

标准SV的评价指标，即等错误率（equal error rate，EER）和最小检测代价函数（minimum detection cost function，minDCF）两种（Wang等，2023），可以直接评估模型在区分不同说话人方面的能力，能够有效反映模型在声纹认定过程中的关键性能，包括对目标说话人特征的捕捉能力、对干扰因素（如克隆语音中的源说话人特征残留）的抵抗能力等，这与声纹认定的本质需求高度契合，故本文采用EER和minDCF作为评价指标。其中，EER值越小，minDCF值越小，模型的性能越优。

3.2 GPCF模块有效性分析

为了验证GPCF模块的有效性，通过在全模型上替换GPCF模块为Res2Block模块（w/o GPCF）进行测试，结果如表4所示。可以看出，相比较于Res2Block模块，GPCF模块使EER降低了0.69%，minDCF降低了0.005 8。这是由于GPCF模块采用了基于不同感受野的特征融合策略，有效地捕获了语音信号中的不同级别和不同位置的局部特征，增强了模型对说话人特征的表达能力。因此，GPCF模块在说话人识别任务中具有更优越的性能，验证了其在提高识别精度和降低误识率方面的有效性。

表4 GPCF模块与Res2Block模块对比

Table 4 Comparison between GPCF and Res2Block modules

模型	EER/%	minDCF
全模型	9.24	0.672 2
w/o GPCF	9.93	0.678 0

3.3 DGF模块有效性分析

为了验证DGF模块的有效性，在搭建的全模型上分别去除动态滤波器（-动态滤波器）以及去除正则化部分（-Dropout）进行测试，结果如表5所示。可以看出，去除动态滤波器的模型与全模型相比，EER升高0.71%，minDCF升高0.024。此外，去除DGF模块中的正则化部分后的模型在EER和minDCF上也逊色于全模型，其中EER升高0.30%，minDCF升高0.008 2。这是由于DGF模块的移除导致模型失去了深度特征的全局感知能力，这种能力对于精确捕捉说话人的独特特征至关重要。正则化部分的去除可能减弱了模型对过拟合的抵抗力，影响了模型在未知数据上的泛化能力。因此，DGF模块的引入是为了弥补这些损失，它通过增强特征的表达力和提高模型的泛化性来提升整体性能。这表明DGF模块及其正则化部分对于维持和提升说话人识别系统的准确性和鲁棒性起着关键作用。

表5 DGF模块及Dropout模块有效性验证

Table 5 Validation of DGF module and dropout module

模型	EER/%	minDCF
全模型	9.24	0.672 2
-动态滤波器	9.95	0.696 2
-Dropout	9.54	0.680 4

注："-"表示去除。

3.4 消融实验分析

为了验证各模块的性能，在所构建的数据集上对各模块分别进行消融实验，如表6所示。其中RBM表示仅采用3层Res2Block堆叠形成Res2Block模块（Res2Block module，RBM），GPCF表示将Res2Block模块替换成GPCF模块，GPCF + DGF表示在使用GPCF模块的基础上加上DGF模块，GPCF + DGF + MFA表示在使用GPCF和DGF模块的基础上加上MFA模块，GPCF + DGF + MFA + ASP表示同时

使用GPCF模块、DGF模块、MFA模块和ASP模块。由表6可以看出，本文方法中使用的各模块对整体的说话人确认效果都有提升作用。在原本的基础上，最终的网络模型的训练结果显示EER降低了2.12%，minDCF降低了0.082 1。此外，多尺度特征聚合模块和注意力统计池模块也在一定程度上优化了网络性能。由整体结构的结果表明，全模型性能在EER和minDCF上都得到了一定改善，即本文提方法是合理有效的。

表6 消融实验分析

Table 6 Analysis of ablation experiments

方法	EER/%	minDCF
RBM	11.36	0.754 3
GPCF	11.19	0.737 6
GPCF + DGF	10.53	0.704 7
GPCF + DGF + MFA	9.90	0.695 9
GPCF + DGF + MFA + ASP	**9.24**	**0.672 2**

注：加粗字体表示各列最优结果。

3.5 对比实验分析

为了验证本文提出TSVCS的有效性，将其与说话人验证领域的ECAPA-TDNN、DS-TDNN和CAM++优秀方法在克隆语音数据集上进行对比实验，对比结果如表7所示。可以发现，本文TSVCS方法在评价指标EER和minDCF中具有更好的效果。其中，对于EER，相对于其他3种方法，EER分别降低了1.38%、0.92%和0.61%，minDCF分别降低了0.012 5、0.006 7和0.044 5，表明本文方法在处理面向克隆语音的声纹认定任务时更具有优势。这是因为TSVCS可以聚焦说话人的关键身份信息，增强目标说话人特征信息的相关性，抑制源说话人特征信息，因此能够提取到更具表现力的目标说话人特征。

3.6 泛化性能分析

为了检测TSVCS方法的泛化性，使用未曾训练过的基于扩散模型的语音转换算法DDDM-VC（decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion）（Choi等，2024）在AISHELL3中文语音数据集上进行克隆语音合成，得到新数据集DDDM-AISHELL3。表8是训练好的TSVCS模型

表7 对比实验结果

Table 7 Comparative experimental results

方法	EER/%	minDCF
ECAPA-TDNN	10.62	0.684 7
DS-TDNN	10.16	0.678 9
CAM++	9.85	0.716 7
TSVCS（本文）	**9.24**	**0.672 2**

注：加粗字体表示各列最优结果。

进行泛化性能测试的结果。从实验结果可以看出，由于在训练过程中并未对DDDM-VC进行训练，在各个评价指标方面与训练过的语音转换算法之间存在一些差距，但仍在可接受的范围内。因此，TSVCS方法对未知语音转换算法具有一定的泛化能力，同时也表明，即使是对扩散模型等新型模型的语音转换算法，TSVCS方法依然可以对其进行分析且具有一定的效果。

表8 泛化性能测试对比结果

Table 8 Generalization performance test comparison results

数据集	EER/%	minDCF
DDDM-AISHELL3	18.91	0.902 6
CS-AISHELL3	9.24	0.672 2

4 结 论

针对日益趋于成熟的语音克隆技术可能带来的声音侵权问题，本文提出一种面向克隆语音的目标说话人鉴别方法，在语音克隆技术快速发展且可能被不法利用的背景下，有助于准确判断克隆语音来源，保护个人声音权，为司法实践提供技术支持，同时也为后续相关研究提供了借鉴思路。

本文提出的TSVCS方法先对输入语音进行MFCC特征提取，接着将特征张量交替输入到GPCF模块和DGF模块，经过多尺度特征聚合（MFA）层得到说话人表示，再经注意力统计池（ASP）加权，最后通过全连接层得到说话人嵌入特征，通过计算余弦距离打分来判定克隆语音是否源自于目标说话人。本文构建了克隆语音数据集CS-AISHELL3，并与ECAPA-TDNN、DS-TDNN和CAM++等方法进行对

比。实验结果表明，TSVCS 在评价指标 EER 和 minDCF 上表现更优，能聚焦说话人关键身份信息，增强目标说话人特征相关性，抑制源说话人特征，提取更具表现力的目标说话人特征。此外，本文使用未训练过的语音转换算法测试模型的泛化性，结果表明 TSVCS 方法对未知语音转换算法也具有一定的适应能力。

由于语音克隆技术采用的模型和方法不同，其克隆的语音也存在一定的差异，如何进一步提高 TSVCS 方法对各种克隆语音尤其是新范式克隆语音模型的泛化性是下一步研究的重点。

参考文献（References）

Baas M, van Niekerk B and Kamper H. 2023. Voice conversion with just nearest neighbors//Interspeech 2023. Dublin, Ireland: ISCA: 2053-2057 [DOI: 10.21437/interspeech.2023-419]

Booth I, Barlow M and Watson B. 1993. Enhancements to DTW and VQ decision algorithms for speaker recognition. Speech Communication, 13(3/4): 427-433 [DOI: 10.1016/0167-6393(93)90041-I]

Choi H Y, Lee S H and Lee S W. 2024. DDDM-VC: decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 17862-17870 [DOI: 10.1609/aaai.v38i16.29740]

Deng J K, Guo J, Xue N N and Zafeiriou S. 2019. ArcFace: additive angular margin loss for deep face recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4685-4694 [DOI: 10.1109/cvpr.2019.00482]

Desplanques B, Thienpondt J and Demuynck K. 2020. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification//Interspeech 2020. Shanghai, China: ISCA: 3830-3834 [DOI: 10.21437/Interspeech. 2020-2650]

Gao Z F, Song Y, McLoughlin I, Li P C, Jiang Y H and Dai L R. 2019. Improving aggregation and loss function for better embedding learning in end-to-end speaker verification system//Interspeech 2019. Graz, Austria: ISCA: 361-365 [DOI: 10.21437/interspeech.2019-1489]

Gersho A and Gray R M. 1992. Vector Quantization and Signal Compression. New York, USA: Springer [DOI: 10.1007/978-1-4615-3626-0]

Hinton G E, Srivastava N, Krizhevsky A, Sutskever I and Salakhutdinov R R. 2012. Improving neural networks by preventing co-adaptation of feature detectors [EB/OL]. [2024-11-01]. https://arxiv.org/pdf/1207.0580.pdf

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7132-7141 [DOI: 10.1109/cvpr.2018.00745]

Huang G, Liu Z, van der Maaten L and Weinberger K Q. 2017. Densely connected convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2261-2269 [DOI: 10.1109/cvpr.2017.243]

Jiang L Y, Zheng Y F, Chen C, Li G H and Zhang W J. 2023. Review of optimization methods for supervised deep learning. Journal of Image and Graphics, 28(4): 963-983 (江铃燚, 郑艺峰, 陈澈, 李国和, 张文杰. 2023. 有监督深度学习的优化方法研究综述. 中国图象图形学报, 28(4): 963-983) [DOI: 10.11834/jig.211139]

Kingma D P and Ba J. 2017. Adam: a method for stochastic optimization [EB/OL]. [2024-11-01]. https://arxiv.org/pdf/1412.6980.pdf

Lee J and Nam J. 2017. Multi-level and multi-scale feature aggregation using sample-level deep convolutional neural networks for music classification [EB/OL]. [2024-11-01]. https://arxiv.org/pdf/1706.06810.pdf

Li J X, Zhang L H and Li Y T. 2023. Speaker feature extraction based on timbre consistency for voice cloning. Journal of Signal Processing, 39(4): 719-729 (李嘉欣, 张连海, 李宜亭. 2023. 基于音色一致的语音克隆说话人特征提取方法. 信号处理, 39(4): 719-729) [DOI: 10.16798/j.issn.1003-0530.2023.04.013]

Li J Y, Tu W P and Xiao L. 2023. Freevc: towards high-quality text-free one-shot voice conversion//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: 1-5 [DOI: 10.1109/icassp49357.2023.10095191]

Li Y X, Li B, Tan S Q and Huang J W. 2024. Research progress on speech deepfake and its detection techniques. Journal of Image and Graphics, 29(8): 2236-2268 (许裕雄, 李斌, 谭舜泉, 黄继武. 2024. 语音深度伪造及其检测技术研究进展. 中国图象图形学报, 29(8): 2236-2268) [DOI: 10.11834/jig.230476]

Malykh E, Novoselov S and Kudashev O. 2017. On residual CNN in text-dependent speaker verification task//Proceedings of the 19th International Conference on Speech and Computer. Hatfield, UK: Springer: 593-601 [DOI: 10.1007/978-3-319-66429-3_59]

Okabe K, Koshinaka T and Shinoda K. 2018. Attentive statistics pooling for deep speaker embedding//Interspeech 2018. Hyderabad, India: ISCA: 2252-2256 [DOI: 10.21437/interspeech.2018-993]

Park H J, Yang S W, Kim J S, Shin W and Han S W. 2023. TriAAN-VC: triple adaptive attention normalization for any-to-any voice conversion//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.1109/icassp49357.2023.10096642]

Qiao Z. 2023. Risks and countermeasures of artificial intelligence gener-

ated content technology in content security governance. Telecommunications Science, 39(10): 136-146 (乔喆. 2023. 人工智能生成内容技术在内容安全治理领域的风险和对策. 电信科学, 39(10): 136-146) [DOI: 10.11959/j.issn.1000-0801.2023190]

Ren Y, Hu C X, Tan X, Qin T, Zhao S, Zhao Z, et al. 2022. FastSpeech 2: fast and high-quality end-to-end text to speech [EB/OL]. [2024-11-01]. https://arxiv.org/pdf/2006.04558.pdf

Shi Y, Bu H, Xu X, Zhang S J and Li M. 2021. AISHELL-3: a multi-speaker mandarin TTS corpus//Interspeech 2021. Brno, Czechia: ISCA: 2756-2760 [DOI: 10.21437/interspeech.2021-755]

Snyder D, Garcia-Romero D, Povey D and Khudanpur S. 2017. Deep neural network embeddings for text-independent speaker verification//Interspeech 2017. Stockholm, Sweden: ISCA: 999-1003 [DOI: 10.21437/interspeech.2017-620]

Wang H, Zheng S Q, Chen Y F, Cheng L Y and Chen Q. 2023. CAM++: a fast and efficient network for speaker verification using context-aware masking//Interspeech 2023. Dublin, Ireland: ISCA: 5301-5305 [DOI: 10.21437/Interspeech.2023-1513]

Xiang X, Wang S, Huang H J, Qian Y M and Yu K. 2019. Margin matters: towards more discriminative deep neural network embeddings for speaker recognition//Proceedings of 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Lanzhou, China: IEEE: 1652-1656 [DOI: 10.1109/apsipaasc47483.2019.9023039]

Xu Y F, Gan J P, Lin X D, Qiu Y Q, Zhan H J and Tian H. 2024. DS-TDNN: dual-stream time-delay neural network with global-aware filter for speaker verification. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32: 2814-2827 [DOI: 10.1109/taslp.2024.3402072]

Yin M K and Li Y B. 2024. Private law regulation of sound infringement in the era of artificial intelligence. Journal of Harbin University, 45(4): 73-77 (尹懋锟, 李云滨. 2024. 人工智能时代声音侵权的私法规制. 哈尔滨学院学报, 45(4): 73-77) [DOI: 10.3969/j.issn.1004-5856.2024.04.016]

Yu Y Q and Li W J. 2020. Densely connected time delay neural network for speaker verification//Interspeech 2020. Shanghai, China: ISCA: 921-925 [DOI: 10.21437/interspeech.2020-1275]

Zhao Z D, Li Z, Wang W C and Zhang P Y. 2023. PCF: ECAPA-TDNN with progressive channel fusion for speaker verification//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.1109/icassp49357.2023.10095051]

作者简介

苏兆品,女,副教授,硕士生导师,主要研究方向为克隆语音的声纹检测及鉴定。E-mail: szp@hfut.edu.cn

张国富,通信作者,男,教授,硕士生导师,主要研究方向为语音安全。E-mail: zgf@hfut.edu.cn

魏玉洋,男,硕士研究生,主要研究方向为克隆语音的声纹认定。E-mail: 2022171232@mail.hfut.edu.cn

廉晨思,女,高级工程师,主要研究方向为声纹鉴定。
E-mail: lchsi324@163.com

岳峰,男,副研究员,主要研究方向为AIGC内容检测。
E-mail: yuefeng@huft.edu.cn