刻意伪装场景下的说话人确认

覃晓逸[1,2]，励 泽[1,2]，刘 东[2]，李 明[1,2]
（1. 武汉大学计算机学院，武汉 430072；
2. 昆山杜克大学苏州市多模态智能系统重点实验室，江苏 215316）

摘 要：刻意伪装的说话人确认任务的难点在于说话人刻意隐藏自己的身份而改变音色成为其他人。本文将这一任务视为一人饰演多角的场景，并为该任务提出了 CN-Moives 训练集和 TheSound-test 测试集。 CN-Moives 数据通过对中文电影中的演员及配音演员进行人物匹配、人脸检测、人脸识别、唇动识别和语音活动片段检测，获取了一人多部戏的多个角色的语音片段。该数据集包含了演员原声和对应的配音演员，利用演员和配音演员为塑造角色而有意改变自己音色的特性，实现了刻意伪装中一人多角数据的采集。同时，利用 TheSound 节目中配音演员刻意隐藏自身身份不被识破的节目特性，提出刻意伪装场景的测试集 TheSound-test。本文通过联合以上领域挖掘的数据，提出采用孪生网络建模，在 VoxMoives 测试集和 TheSound-test 集上均取得了说话人验证性能的显著提升。

关键词：说话人确认；刻意伪装；孪生网络
中图分类号：TP183;TN912.3　**文献标志码：**A

语音伪装通常是指伪装者故意改变自己的声音以绕过系统检测。伪装者通常采用自然的方式（用刺耳的声音说话以避免被发现，模仿他人的声音等）或使用语音转换系统（Voice Conversion，VC）[1]以实现绕过和攻击说话人系统的目的。根据 FarrÚs [2] 的研究介绍，声音伪装可以分为四类：

- 非刻意和非电子：这类声音伪装是无意识的，比如声音嘶哑、情绪变化、中毒以及老化。
- 非刻意和电子：这种情况主要是由于信道引起的语音失真，例如电话信道传输和麦克风录制造成的失真。
- 刻意和非电子。这种伪装是自然的，取决于说话者如何刻意改变自己的发音习惯。
- 刻意和电子。这种刻意伪装通常由 VC 技术实现，利用变声模型改变原始音色。

近年来，随着深度语音合成[3]技术的迅猛发展，语音真假难辨，例如 2023 年最新推出的 VALL-E [4]。因此大部分的声音伪装主要集中在刻意和电子场景，如近些年火热的 ASVSpoof [5]以及 ADD [6] 系列比赛都是为探究电子语音伪装检测而举办的研讨会和比赛。而这些比赛都是对机器合成语音的检测，属于二分类，只需要判断音频是否是伪造而成的，属于语音鉴伪领域。

本文重点研究刻意场景中的非电子实现，即识别用自然方式改变声音的伪装者身份。不同于电子方法的实现，以自然的方式掩盖自己或模仿他人的声音来源于真人实现，不容易被发现。其中自然的方式根据目的的不同可以分为：模仿者和伪装者。模仿者的目的是在不借助计算机技术辅助的前提下模仿他人声音以攻击验证系统或欺骗真人。对于系统而言，模仿者会增加假阳率（False Accept Rate, FAR）；而伪装者的目的是刻意改变自己的声音来深藏自己的身份。对于系统而言，伪装者的出现会降低真阳率（True Accept Rate，TAR）。二者都会对说话人识别系统造成挑战，模仿者和伪装者都可以认为是刻意伪装场景下的说话人识别，如何识别出伪装者是谁是本文的目的之一。

目前对伪装者的研究主要集中于法医领域。在刑事和诈骗案件中，法医需要通过分析伪装者原始和伪装后的声音进行举证和推断。Kuunzel 基于自己在 2000 年的研究 [7]，分析了自然语音伪装对 UBM-MAP 适应的 GMM 法医自动说话人识别系统准确性的影响，在 2004 年的 odyssey 上 Kuunzel 等 [8]通过在一个涵盖了 100 名德国说话人的数据集上，分析了由提高音调、降低音调和噪声压缩引起的系统性能降级情况。仅当参考群体由正常讲话组成时，等错误率（Equal Error Rate, EER）才表现出显著的降低，而在测试过包含相同类型伪装的参考群体时，降低情况大大减轻，这表明了在处理此类语音伪装时，基于谱的系统的鲁棒性是不足的。Kajarekar 等人 [9] 分析了故意改变讲话风格（改变音调、时长或模仿口音）对 FISHER 数据库上的 13 个基于 GMM-MFCC 的说话人识别系统的影响。结果显示，EER 从 0.05%（使用正常语音测试）增加到 7.46%（使用伪装语音测试）。

类似的研究还包括了 Zhang [10] 以及 Tan [11] 在一种名为法庭自动说话人识别系统（FASRS）的

发展系统中研究了 10 种语音伪装的影响，该系统与 20 名男学生录制的正常语音一起使用。结果表明，由于语音伪装，说话人识别的性能大大降低，不同类型的伪装产生了不同的影响，唯独外语口音具有高度的抗干扰性。这是因为 FASRS 被设计成一种语言和方言无关的系统——低语和口罩是对系统影响最大的两种伪装类型。最近，González-Hautamäki 等人 [12] 报告称，大约 70% 的伪装音频中的 F0 和至少一个共振峰有显著差异，导致了自动说话人确认系统中 EER 的增加。

然而目前大部分的研究主要集中于刻意伪装的发声基理，关于如何对抗刻意伪装的研究并不多。Zheng 等人 [13] 广泛研究了自动语音伪装（Automatic Voice Disguise，AVD）和自动说话人验证（Automatic Speaker Verification, ASV）之间的相互关系，提出了一种将伪装声音恢复为其原始版本的方法。然而这种方法针对 AVD 有所提升，却没有解决真人的语音伪装问题。而由于真人语音伪装数据的缺乏，对于如何确认伪装者原声和伪装后的声音的研究寥寥无几。因此本文采用一种领域数据挖掘的方法，通过分析中文电影和特定综艺场景，以影视剧中一人多角的场景线索为基础，利用演员为不同人物性格塑造而改变音色的特性，结合音视频多模态的说话人追踪定位，采集并构建了 CN-Movies 训练集和 TheSound-test 测试集。CN-Movies 数据集来源中文电影， TheSound-test 数据集则来源于中国的配音综艺节目。通过对数据的深入分析，本文提出了一种基于孪生网络建模学习获取对抗伪装说话人表征向量的解决方案。最终模型性能在 VoxMovies 和 TheSound-test 测试集上获得了一致性提升。

收稿日期：2024-05-24

基金项目：国家自然科学基金面上项目(62171207)；

苏州市科技项目(SYC2022051)

作者简介：覃晓逸（1994—），男，博士研究生；

励泽（2001－），男，硕士研究生；

刘东（1997—），男，硕士研究生

通信作者：李明，副教授，E-mail:

ming.li369@dukekunshan.edu.cn

1 CN-Movies 和 TheSound-test 数据集构建

本小节将介绍如何构造 CN-Movies 训练集和 TheSound-test 测试集。由于针对刻意伪装场景获取数据较为困难，而本文受到 VoxMovies [14] 的启发，在电影场景中，演员为了符合角色形象会刻意改变自己的说话和发音方式。VoxMovies 通过结合 VoxCeleb [15,16] 数据，将名人在电影中的片段截出，得到名人在电影片段中的数据，通过电影片段数据和采访数据构造，以达到识别电影场景中说话人的身份的目的。然而 VoxMovies 大部分都是外国电影且以英文发音为主，并且训练集仅有 382 人及其对应的 8122 条音频。因此，为了进一步研究伪装者场景的说话人识别，本文基于中文电影数据以及《声临其境》（TheSound）综艺进行研究。针对中文电影，虽然目前已经有 CN-Celeb [17] 中国名人说话人识别数据库，但是由于中文电影中存在大量的配音以及 CN-Celeb 数据库的自动化采集流程，导致该数据库不可避免地会引入噪声标签。因此，本文通过重新过滤中文电影并依据百度百科对电影中的角色进行演员和配音的定位。通过百度百科中演员的生活照和剧照对电影中的人脸进行检测和识别，利用自动语音识别（Automatic Speech Recognition, ASR）技术提取有声部分，从而构建了一个大规模的中文电影说话人识别数据集，称为 CN-Movies。该数据集不仅含有演员原声并且有配音演员身份。

其次，尽管电影场景中的角色语言情景非常丰富，但演员并未完全刻意伪装自己的声音。为了真实的模拟刻意伪装的场景，本文基于《声临其境》综艺（该综艺邀请明星进行配音并且不让观众猜出明星身份）提出刻意伪装场景的说话人识别测试集 TheSound-test。由于其节目的特性，该数据天然地包含了刻意伪装场景。本文使用与 CN-Movies 相同的提取方法，并加入说话人聚类以剔除误标的异常样本。下面将具体介绍两个数据库的采集流程。

1.1 CN-Movies 数据集

由于训练数据主要来源于中文电影数据，因此该数据集被命名为 CN-Movies。整个数据采集过程通过自动化脚本完成并经过人工随机抽查以确保数据质量。CN-Movies 数据采集流程如图 1 所示。整个流程分为 5 部分：

• 步骤 1：视频数据收集。首先，基于豆瓣平台爬取所有华语电影各项榜单的前 500 部电影，组成备

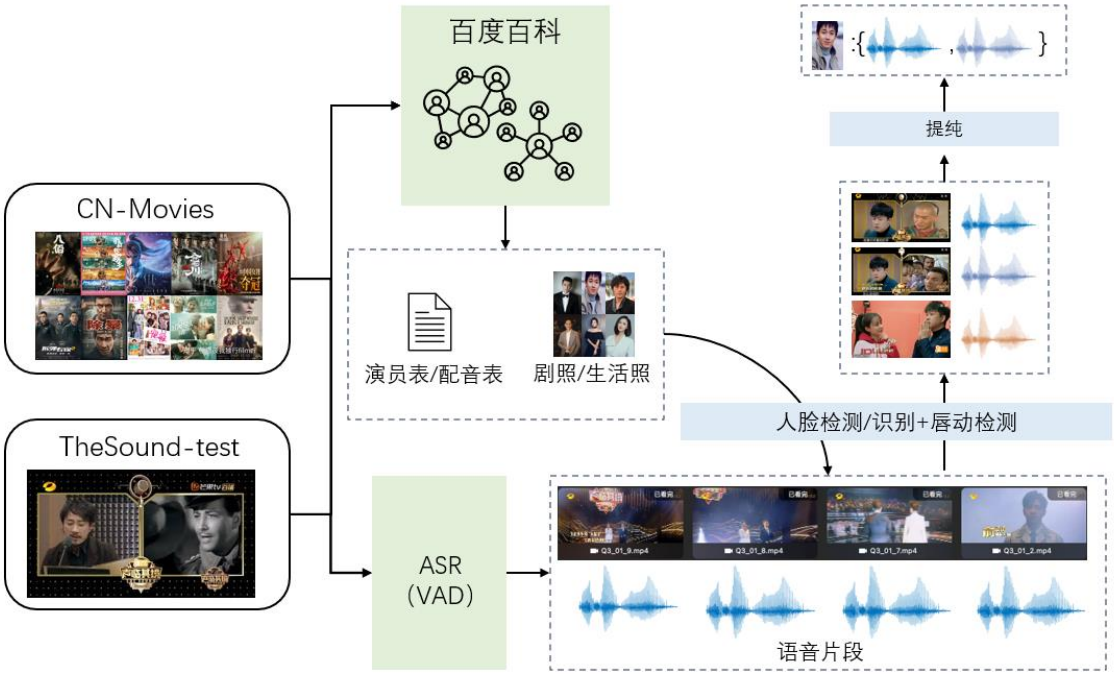


图 1 CN-Movies 及 TheSound-test 数据采集流程

选电影列表，并根据电影名从 YouTube 平台上进行检索和下载。

• 步骤 2：视频角色元信息的收集。根据电影名称反查百度百科中的电影对应的演员及其配音信息。对每位演员采集 20 张以上的生活照和剧照。并利用这些照片进行演员的人脸注册。

• 步骤 3：语音活动检测（Voice Activate Detection, VAD）。由于电影时长较长且包含大量音乐声，采用一般基于能量的 VAD 方法得到的语音片段中会包含很多音乐部分，而逐一遍历数据又耗时巨大。因此，本文采用基于 ASR 的 VAD，利用开源语音识别工具 Whisper [1]进行电影的语音识别，并提取出语音片段所在的时间戳。通过这种方式，获得了只有语音文本的电影片段。

• 步骤 4：音频数据提纯。由于自动化采集难免会引入噪声标签，因此本文对音频数据进行了提纯。首先，将人脸对应的音频进行聚类，选择数量最多的聚类结果作为注册音频。然后，利用这些注册音频对其他音频进行说话人识别打分。由于在一部剧中说话人的音色变化较少，这一方法能够有效过滤大量的噪声标签。

• 步骤 5：人脸检测与验证。利用基于 ASR 的 VAD 得到的语音片段，本文通过人脸识别和人脸验证技术对语音对应的视频片段进行人脸比对。每部电影都有其对应的演员人脸库，本文利用人脸检测技术，通过与数据库中的人脸进行验证，并通过与预先设定阈值进行比对，高于阈值的分配给某一个特定说话人，低于阈值的则丢弃。

• 步骤 6：人工随机复查。最后通过人工进行随机抽查检测。检测结果显示，该流程的准确率达到 98%以上，基本满足使用需求。

具体实现细节如下：

视频数据收集: 为了确保电影数据能够涵盖大量演员并满足每位演员在多部电影中出演的需求，本文首先基于豆瓣平台，按照电影分类收集所有华语电影的 Top500 榜单，组成备选电影列表。随后在 YouTube 平台上进行检索和下载。考虑到并非所有电影都有对应的下载资源，本文对检索结果的前 10 项进行时长验证。当视频时长和电影时长接近时，选择搜索排序第一的电影进行下载，并进行人工检查，确保下载数据为电影而非其他类型的视频。最终，共计收集 992 部电影，涵盖演员 2025 人，配音演员 330 人。

视频角色元信息的收集: 获取可用电影资源后，本文通过电影名对百度百科进行反查，利用百度百科中对电影的正则化表示，提取对应的主要演员及其配音演员信息。在获得电影对应的演员姓名和该角色对应的配音演员后，收集电影中的演员剧照和生活照（每位演员采集 20 张以上），构建了 CN-Movies-Face 人脸数据集，该数据集包含 6302 名演员共计 121128 张人脸数据，并利用这些人脸数据

[1] https://github.com/openai/whisper

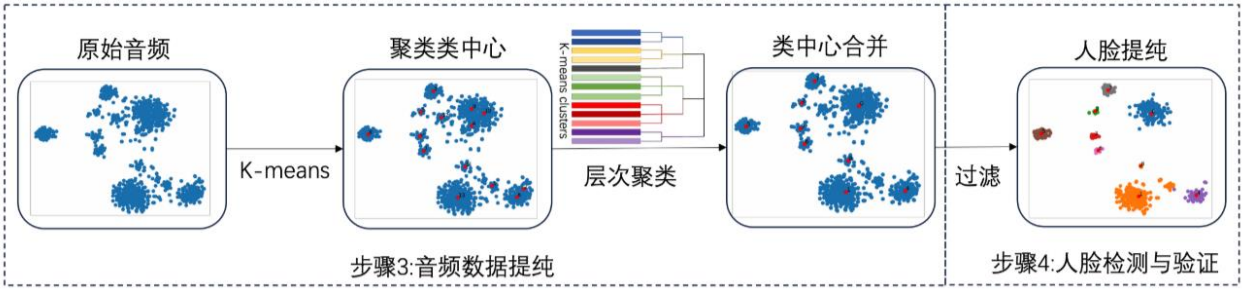


图 3 音频数据提纯及人脸过滤

进行注册。然而由于现有开源的人脸资源主要针对欧美人种，在测试过程中通过采用 Deepface [18]、Face Recognition [2]、InsightFace [3] 等开源工具对人脸进行比对打分后发现，虽然这些模型在 LFW 数据集上的准确率在 98% 以上，但在本文采集的人脸数据测试集上的准确率仅为 80%。因此本文采用迁移学习的策略对人脸识别模型进行中国明星人脸的微调（Finetune，FT）。本文采用 InsightFace 中的 ResNet50 作为预训练模型，微调前后的结果如表 1 所示。随后，本文利用微调后的模型对所有演员的生活照和剧照提取人脸表征向量，并对这些向量进行平均，最终得到人脸注册样本库。

表 1 人脸识别模型结果

方法	数据集	LFW	CN-Movies-test
ResNet50	WebFace	0.9836	0.8432
+ FT	CN-Movies-Face	0.9721	0.9559

语音活动检测: 通常，VAD 采用基于统计的方法，如能量门限法（Energy-Based VAD）、过零率法（Zero Crossing Rate-Based VAD）、基于概率的 VAD 等，或者基于深度学习的方法。由于电影时长较长且包含大量音乐声，采用一般基于能量的 VAD 方法会导致得到的语音片段很多都是音乐。因此本文采用语音识别作为语音活动检测的方法，利用开源语音识别工具 Whisper 进行电影的语音识别，并给出语音片段所在的时间戳（如图 2）。本文通过对比 Whisper 中的 base，large，medium，small 和 tiny 四种模型的识别率和漏词率，最终选择 large 模型作为语音识别模型。另外考虑到语音识别会将一个连续的语音片段切成很多小片段，本文采用相近音频片段合并的原则，将低于 2 秒的相近音频片段进行合并，最终得到大于 2 秒的音视频语音片段。

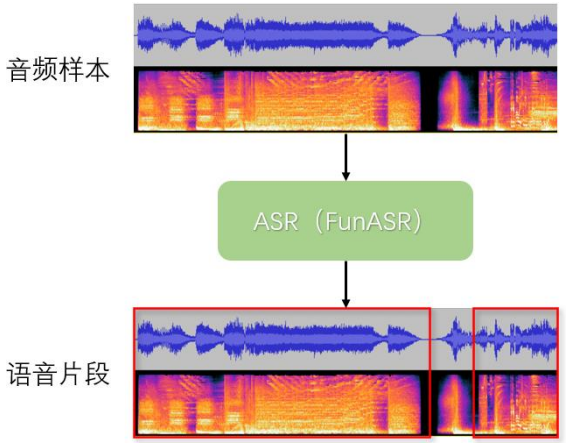


图 2 基于 ASR 的 VAD 流程图

音频数据提纯：由于自动化采集难免引入噪声标签，例如电影中人化妆浓重导致标签误判给其他人的音频说话人和视频人脸不对应问题。考虑到在一部剧中说话人的音色很少突变，因此本文采用音频数据提纯策略。具体流程如图 3 所示，首先根据电影主演表中的演员数量初始化聚类数，对一部电影的所有音频数据进行 K-means 聚类。由于聚类类数通常超过实际类，因此进一步对 K-means 的中心簇进行层次聚类以实现类合并。而后设定相似度阈值，计算每个簇中心与该簇所有音频的相似度矩阵，剔除低于阈值的音频以确保该类说话人的纯净度。

人脸检测与验证：通过基于 ASR 的 VAD 的处理和音频提纯后得到了音频数据对应的伪类，但是这些伪类无法直接确定说话人的准确标签。因此本文采用视频角色元信息收集阶段训练的中国明星人脸识别模型对说话人的人脸进行特征提取作为测试人脸。而后将测试人脸和该电影场景下的注册人脸库进行打分对比。本文采用 RetainFace [19] 作为人脸检测和定位工具。最终，通过预先设定的阈值，

[2] https://github.com/ageitgey/face_recognition
[3] https://github.com/deepinsight/insightface

高于阈值的分配给某一个说话人，低于阈值的则丢弃。如果一段视频中只出现了一个人那么该音频片段分配给唯一的说话人。如果一段视频中出现了多个人，则统计该视频出现人脸次数最多的人作为该音频的标签。最后通过统计每个类中出现最多的人脸确定该类的说话人标签。

考虑到一人多角色场景，本文只选择经以上提取后在 3 部或以上电影中出现的演员和配音演员以构建一个大规模中文电影演员和配音演员数据库，具体信息统计如表 2 所示：

表 2 CN-Movies 统计结果

数据集	CN-Movies
演员（配音演员）数量	694 (134)
话语数量	205,649
每个说话人的平均话语数量	248
每句话语的平均时长（秒）	2.1
每个电影的平均演员数量	5.2

1.2 TheSound-test 数据集

TheSound-test 是刻意伪装场景说话人识别的验证集，数据采集流程和 CN-Movies 基本一致。但是由于测试集对标签严苛的高要求，因此在具体实施中有以下几点不同：

• 在步骤 3 和步骤 4 之间增加了音视频说话人活动检测（Audio-Viual Active Speaker detection，AV-ASD）。鉴于 TheSound 数据中频繁的场景切换和镜头切换，许多视频片段可能存在音视频不对应的情况。因此，本文采用音视频联合的说话人活动检测，以实现语音和人脸的精准对齐。

• 当为所有音视频分配好标签后，不同于训练集 CN-Movies 的人工随机抽查，这里采用人工遍历检查，通过观察音视频确保每一个片段都是正确的。

音视频说话人活动检测：在电影和综艺场景中，通常存在一个视频镜头内有多个说话人的场景。为了确定在多说话人的音视频场景中谁在讲话，本文采用 AV-ASD 技术将音频和视频中的说话人进行对齐。通常情况下，确定语音视频片段中哪个说话人讲话，可以采用唇动技术判断声音是否与唇动对应。TalkNet [20] 是一种考虑了短期和长期特征的结构，其包括了音频和视觉的时间编码器用于帧级别特征表示，音视频交叉注意机制用于跨模态交互，以及一个自注意机制用于捕获长期说话证据。本文采用 TalkNet 作为语音和活动人脸定位模块，以确定多人脸和多说话人场景中目标说话人音频位置。最终处理后的效果示意图如图 4 所示。

图 4 AV-ASD 处理后的效果示意图

如图 4 所示，该音频片段包含两位不同的说话人。通过 AV-ASD 技术可以精确定位不同说话人所在的时间片段。根据这一规则，红色框内的说话人属于目标说话人，绿色框内的说话人属于非目标说话人，于是保留红色框内容并剔除绿色框内容。尽管这一方法不可避免的存在判决过严的问题，但它确保了标签的准确性。然而，由于该步骤处理需要花费大量的时间（为原视频时长的 3 倍），且剔除了过多的数据，考虑到测试集的准确性，因此仅在 TheSounds 数据集上进行了使用。最终，本文收集到了 104 位配音演员在刻意伪装场景下的数据，并利用该数据构建了 TheSound-test 测试样本对。

2 联合域数据挖掘的孪生网络

2.1 伪装场景下说话人表征向量分析

首先本文从 TheSound-test 中选取了一期节目的音频，使用 t-SNE 方法观察不同说话人的表征向量分布，如图 5(a) 所示。所有提取说话人表征向量的说话人模型均基于 VoxCeleb 数据训练。可以发现，尽管不同说话人之间存在一定的区分度，但整体的说话人表征向量分布较为松散且说话人间的界限并不明显。相比之下，观察 VoxMovies 的 t-SNE 图（如图 5(b)所示），可以发现同一个说话人的数据明显分为两类，一类是采访场景，另一类是电影场景，其中电影场景又细分出很多子类。这表明电影场景和采访场景属于不同域，且电影场景之间也存在域差异。因此，单纯使用对抗学习方法（如生成对抗网络、梯度反转等）难以有效解决此类问题。

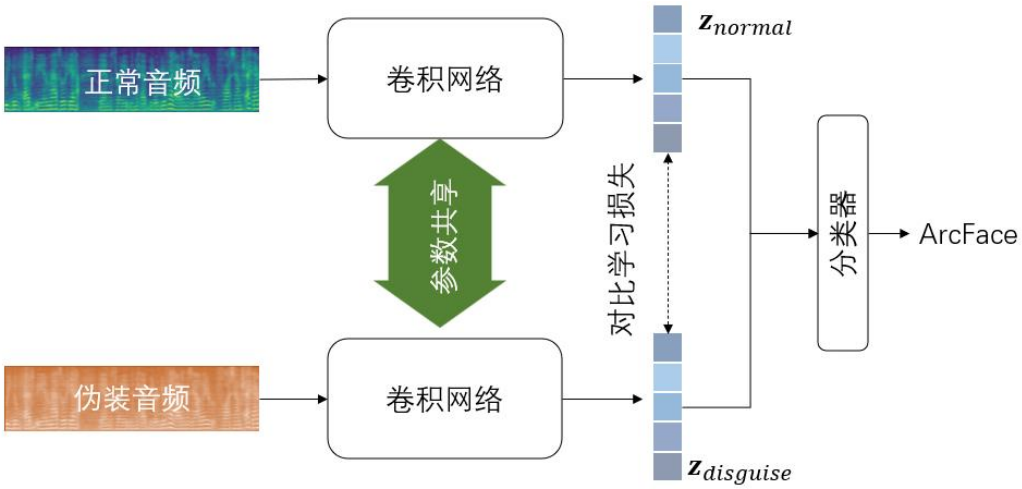
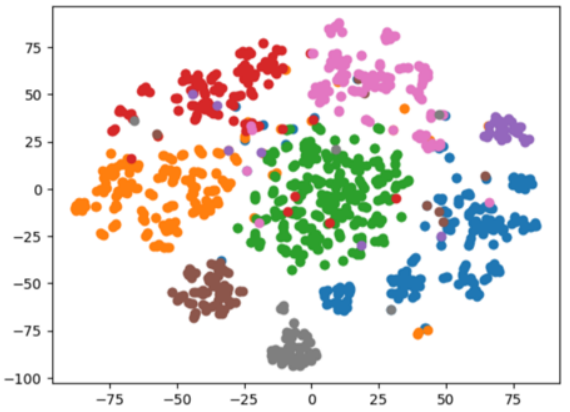
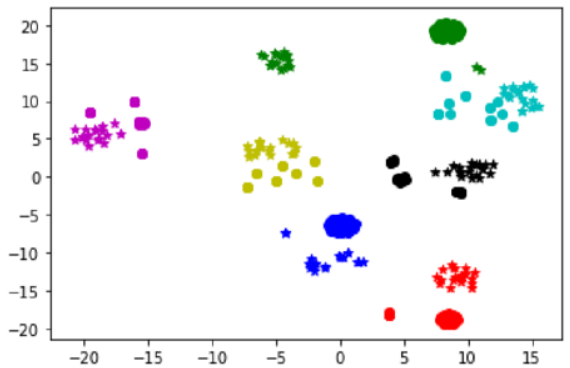


图 6 联合域数据挖掘的孪生网络框架



(a) TheSound-test 中的不同说话人表征向量的 t-SNE 图



(b) VoxMovies 中的不同说话人表征向量的 t-SNE 图。
("*"表示采访场景，"·"表示电影场景)

图 5 伪装场景下说话人表征向量的 t-SNE 图

从本质上探究，电影场景中的说话人往往因角色需要刻意改变音色，而在采访场景中则是正常讲话。类似电影场景，刻意伪装的场景中，伪装者为了避免自己的声音被识别，会刻意地改变自己的音色。只不过这种刻意的伪装相比于电影场景更难处理。因此，需要利用神经网络对同一人不同域的数据进行声学特征的共同建模，挖掘原声和伪装声特征的共性。

2.2 学习域不变的孪生网络

孪生网络是一种用于处理成对数据任务的神经网络架构。该方法最早由 Yann LeCun 等人 [21] 在 1994 年提出，用于签名验证。孪生网络的核心思想是通过共享权重来处理输入对。它包含两个完全相同的子网络，这两个子网络共享相同的权重和参数。这两个子网络分别接受输入数据同一类不同域两个音频，并通过网络的前向传播过程将它们映射到一个共享的特征空间。本文受笑声说话人识别研究 [22] 启发，该任务采用一种孪生网络通过输出说话人正常音频和对应的笑声音频，以实现说话人笑声表征向量和说话人正常语音表征向量的一致表示，显著提升了识别效果。

因此本文采用相同的架构，如图 6 所示。利用神经网络对不同域的数据进行声学特征建模，挖掘原声和伪装声特征的共性。将正常音频 x_{normal} 和伪装音频 $x_{disguise}$ 同时输入到同一个卷积网络 $E(\cdot)$ 中，获得正常音频表征向量 z_{normal} 和伪装音频表征向量 $z_{disguise}$：

$$z_{normal} = E(x_{normal}). \quad (1)$$

$$z_{disguise} = E(x_{disguise}). \quad (2)$$

为了令伪装语音表征向量和正常语音表征向量取得一致表示，本文采用对比学习损失，公式如下所示：

$$L_{contrastive} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{sim(z_{normal},z_{disguise})/\tau}}{\sum_{k=1,k\neq i}^{N}e^{sim(z_i,z_k)/\tau}}. \quad (3)$$

其中 $sim(\cdot)$ 表示两个输入向量之间的余弦相似度，τ 表示超参数。然而，仅仅使用对比学习会导致说话人识别网络在有限说话人数据的基础上陷入局部最优，进而导致泛化性下降。因此，本文采用 ArcFace [23] 和交叉熵损失作为损失函数指导学习到的表征向量保证说话人识别性能。最终的损失函数为：

$$L_{ce} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{s\cdot cos(\theta_{y_i}+m)}}{e^{s\cdot cos(\theta_{y_i}+m)}+\sum_{j=1,j\neq y_i}^{N} e^{s\cdot cos\theta_j}} \quad (4)$$

$$L = \lambda \cdot L_{ce} + L_{contrastive}. \quad (5)$$

其中 L_{ce} 表示基于 ArcFace 的交叉熵损失, s 为缩放因子用于加快模型训练，m 为附加的角裕度惩罚旨在增强类内紧度的同时增大类间差异 ，λ为超参数，设置为 0.5。

3 实验设置

3.1 数据使用

本文选用 VoxCeleb2 开发集和 VoxMovies-Train 作为训练数据进行前期的验证工作，并在 VoxMovies-Test 和 TheSound-test 上进行测试。随后通过加入大规模的 CN-Movies 中配音演员的数据训练进行进一步的提升。

表 3 展示了 VoxMovies 数据集的组成，包括训练集和测试集。测试集的具体设置如表 4 所示。VoxMovies 的研究目标是评估现有流行的说话人识别模型在电影语音片段上的表现，由于演员在扮演角色时通常会故意改变自己的声音，这一特点与本文的研究方向高度契合。VoxMovies 数据集涵盖了来自近 4000 个电影片段的 856 位说话人的语音，这些语音片段表现出不同情感、口音和背景噪声，与目前的说话者人数据集（如 VoxCeleb）中的以采访为主、情感平静的话语存在显著差异。此外，VoxMovies 还提供了一系列用于域自适应的评估集，并在这些评估集上对最先进的说话人识别模型进行了基准测试。实验结果表明，无论在说话人确认还是识别任务中，这些模型的性能都显著下降。

表 3 VoxMovies 数据统计

数据集	说话人来源	说话人数	话语条数
VoxMovies-Test	VoxCeleb1	485	4,943
VoxMovies-Train	VoxCeleb2	371	3,962

表 4 VoxMovies 不同测试集统计。其中 DM 表示电影场景数据，DI 表示交互场景（即 VoxCeleb 中的数据）

场景	正样本来源	负样本来源	话语条数	正样本对/负样本对
E1	DM (same)	DM	20,572	10,286/10,286
E2	DI, DM	DI, DM	46,578	23,289/23,289
E3	DI, DM	DI	46,804	23,402/23,402
E4	DI, DM	DM	46,866	23,433/23,433
E5	DM (diff)	DM	41,090	20,545/20,545

3.2 模型使用

本文选用 ResNet34 [24] 作为骨干网络，其残差块的通道数依次设置为 {32, 64, 128, 256}。全局统计池化层用于计算输出特征图的均值和标准差。在池化层之后，接入全连接层以输出 128 维的说话人表征向量。我们采用基于 ArcFace 的分类器（s=32，m=0.2）用于身份分类，旨在增加说话人类间距离，同时保证类内的紧凑性。声学特征方面，使用 80 维的对数梅尔谱能量，帧长为 25 毫秒，帧移为 10 毫秒，提取的特征在送入深度说话人网络之前进行了均值归一化处理。

3.3 训练细节

由于 VoxMovies 以英文数据为主，虽然场景和本文提出的场景相似，但是也存在着语种不匹配的问题。利用孪生网络可以挖掘英语语系的说话人发音共性，但是对于中文语系可能提升有限。因此整个实验分为两阶段进行：

• 第一阶段预训练，以 VoxCeleb2 开发集和 VoxMovies-Train 作为训练数据，VoxMovies 和 TheSound-test 作为测试数据，验证网络结构的有效性。首先以 VoxCeleb2 开发集为训练集预训练我们的基线系统，并在 VoxMovies 上进行测试。而后采用混合微调训练策略，以 VoxCeleb2 开发集 和 VoxMovies-Train 作为训练数据微调模型。

• 第二阶段微调，在一阶段模型的基础上，以 CN-Movies 为训练数据，TheSound-test 为测试数据对模型进行微调，实现领域自适应。

模型采用 SGD 作为优化器进行训练，在第一阶段的初始学习率为 0.1，每 10 个 epoch 进行一次学习率衰减，衰减因子为 0.1。当学习率达到 1e-4 时，结束训练；第二阶段的学习率固定为 0.001。训练时，采用 MUSAN [25] 和 RIR_NOISE [26] 用于数据增强。

表 5: 刻意模仿场景下的说话人识别结果

方法	VoxMovies-Test										TheSound-test	
	E1		E2		E3		E4		E5			
	EER	mDCF	EER	mDCF	EER	mDCF	EER	mDCF	EER	mDCF	EER	mDCF
ResNet34	7.797%	0.594	8.669%	0.626	9.764%	0.845	9.627%	0.780	12.105%	0.790	18.704%	0.923
+VoxMovies-Train	6.854%	0.572	8.150%	0.608	8.978%	0.841	9.431%	0.789	11.185%	0.759	17.541%	0.957
++孪生网络	5.989%	0.503	6.909%	0.551	7.166%	0.773	8.915%	0.809	9.875%	0.751	17.475%	0.942
+++CN-Movies	3.914%	0.334	4.521%	0.350	5.032%	0.578	5.124%	0.547	7.003%	0.543	11.576%	0.712

4 实验结果及分析

本文采用等错误率（Equal Error Rate, EER）和最小检测成本函数（minimum Detection Cost Function, mDCF）作为评估模型的性能指标。

表 5 展示了我们构建的任务的难点和所提出方法的有效性。首先，我们观察基线系统的性能，在 VoxMovies 的测试集中，从 E1 到 E5 的 EER 逐级增加。性能最好的场景为 E1 即注册和测试在同一电影中，由于在同一电影中，因此说话人的音色不会有太大的变化。其次是 E2 场景，可以观察到 E2 中多了 DI 的数据，即说话人正常讲话场景。通过比较 E2 和 E4 场景，可以发现当注册和测试来自不同场景时，说话人确认性能迅速下降，E4 仅剔除了 DI 作为测试，其 EER 从 8.67% 上升到了 9.63%。最难的两个场景来自 E4 和 E5， 其中 E5 的 DM 电影来自不同的电影片段，说话人为符合角色身份进行了明显的音色改变，导致说话人性能的急剧下降。本文提出的 TheSound-test 集涵盖了 E4 和 E5 场景，但不同的是，VoxMovies 中的说话人并没有特意伪装，而是基于原本音色的自然改变。此外，由于训练数据语种的不匹配，导致基线系统的在 TheSound-test 上表现很差。

而后，通过为训练集加入 VoxMovies 训练集的数据发现可以有效提升 VoxMovies 测试集的结果，并在一定程度上提升 TheSound-test 测试集上的结果。这说明挖掘领域数据可以有效提升模型在该复杂场景下的性能。

此外，通过采用本文提出的孪生网络，在 VoxMovies 上获得了接近 15% 的性能提升。通过比较 VoxMovies 数据的 t-SNE 图前后对比变化（如图 7 所示）可以观察到：在不进行任何处理的情况下，说话人表征向量的类内分布是松散的。而经过域数据挖掘的孪生网络后，同一个说话人在不同的交互和电影场景中的说话人表征更为紧凑，并且整体的类内分布相对于基线系统更为紧致。表 5 和图 7 证明了本文提出方法的有效性，其中提升最明显的来自 VoxMovies 的 E5 场景，从最初的 11.2% 的 EER 降低至 9.87%。然而，由于跨语种问题的存在，孪生网络无法解决跨语种的正常音和伪装音的共性，因此在 TheSound-test 上的提升有限。

最终，在加入 CN-Movies 领域数据和孪生网络的结合后，系统性能在所有测试集上得到了 20% 的提升。如图 8 所示。在加入 CN-Movies 前，TheSound-test 的说话人类内分布同样是松散的且说话人间没有明显的分界。加入 CN-Moives 后，一方面减少了跨语种的影响，说话人类间有明显的分界（如说话人 4，一开始混杂在其他说话人中，加入 CN-Movies 后自成一簇）；另一方面，说话人的类内更为紧致（如说话人 7 和说话人 3），且有些说话人实现了类内不同子类的合并（如说话人 2）。TheSound-test 的 EER 从 17.475% 降低到了 11.576%。但是值得关注的是，尽管采用了 CN-Movies 和域数据挖掘的孪生网络，TheSound-test 上的 EER 依然很高，这表明该问题依然具有挑战性。

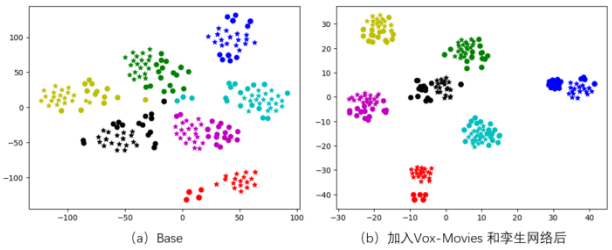


图 7: VoxMovies 上 t-SNE 对比图。不同颜色标注表示不同说话人标签，其中“•”表示交互场景，“⋆”表示电影场景。

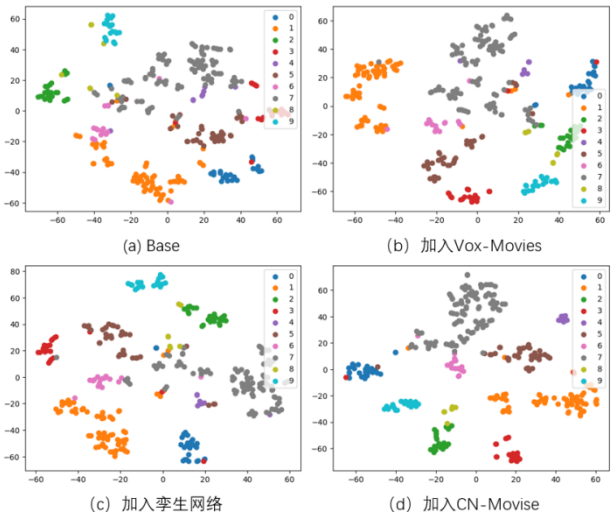


图 8: TheSound-test 上 t-SNE 对比图。不同颜色标注表示不同说话人标签。

5 总结

在很多电信诈骗和法庭举证场景中，嫌疑人往往会通过刻意改变自己的音色防止被定位和定罪。针对这一全新课题，本文通过分析刻意伪装场景，发现刻意伪装可以是一种说话人多角色扮演的场景，因此提出 CN-Movies 和 TheSound-test 用于训练和测试。CN-Movies 来自中文电影和 TheSound 则是中国的综艺《声临其境》，其场景完美的切合了本文提出的刻意伪装。本文通过联合域数据挖掘（增加更多的目标域数据）和孪生网络的方法（挖掘原声和伪装声特征共性），在 VoxMovies 和 TheSound-test 上实现了一致性的提升。

参考文献

[1] Hansen J H L, Hasan T. Speaker recognition by machines and humans: A tutorial review[J]. IEEE Signal processing magazine, 2015, 32(6): 74-99.

[2] Farrús M. Voice disguise in automatic speaker recognition[J]. ACM Computing Surveys (CSUR), 2018, 51(4): 1-22.

[3] Tan X, Qin T, Soong F, et al. A survey on neural speech synthesis[J]. arXiv preprint arXiv:2106.15561, 2021.

[4] Wang C, Chen S, Wu Y, et al. Neural codec language models are zero-shot text to speech synthesizers[J]. arXiv preprint arXiv:2301.02111, 2023.

[5] Liu X, 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.

[6] Yi J, Tao J, Fu R, et al. Add 2023: the second audio deepfake detection challenge[J]. arXiv preprint arXiv:2305.13774, 2023.

[7] Künzel H J. Effects of voice disguise on speaking fundamental frequency[J]. International Journal of Speech Language and the Law, 2000, 7(2): 150-179.

[8] Künzel H J, Gonzalez-Rodriguez J, Ortega-García J. Effect of voice disguise on the performance of a forensic automatic speaker recognition system[C]//ODYSSEY04-The Speaker and Language Recognition Workshop. 2004.

[9] Kajarekar S S, Bratt H, Shriberg E, et al. A study of intentional voice modifications for evading automatic speaker recognition[C]//2006 IEEE Odyssey-The Speaker and Language Recognition Workshop. IEEE, 2006: 1-6.

[10] Zhang C, Tan T. Voice disguise and automatic speaker recognition[J]. Forensic science international, 2008, 175(2-3): 118-122.

[11] Tan T. The effect of voice disguise on automatic speaker recognition[C]//2010 3rd International Congress on Image and Signal Processing. IEEE, 2010, 8: 3538-3541.

[12] Hautamäki R G, Sahidullah M, Hautamäki V, et al. Acoustical and perceptual study of voice disguise by age modification in speaker verification[J]. Speech Communication, 2017, 95: 1-15.

[13] Zheng L, Li J, Sun M, et al. When automatic voice disguise meets automatic speaker verification[J]. IEEE Transactions on Information Forensics and Security, 2020, 16: 824-837.

[14] Brown A, Huh J, Nagrani A, et al. Playing a part: Speaker verification at the movies[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2021: 6174-6178.

[15] Nagrani A, Chung J S, Zisserman A. Voxceleb: a large-scale speaker identification dataset [J]. arXiv preprint arXiv:1706.08612, 2017.

[16] Chung J S, Nagrani A, Zisserman A. Voxceleb2: Deep speaker recognition[J]. arXiv preprint arXiv:1806.05622, 2018.

[17] Fan Y, Kang J W, Li L T, et al. Cn-celeb: a challenging chinese speaker recognition dataset[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7604-7608.
[18] Serengil S I, Ozpinar A. Hyperextended lightface: A facial attribute analysis framework[C]//2021 International Conference on Engineering and Emerging Technologies (ICEET). IEEE, 2021: 1-4.
[19] Deng J, Guo J, Ververas E, et al. Retinaface: Single-shot multi-level face localisation in the wild[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5203-5212.
[20] Tao R, Pan Z, Das R K, et al. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection[C]//Proceedings of the 29th ACM international conference on multimedia. 2021: 3927-3935.
[21] Bromley J, Guyon I, LeCun Y, et al. Signature verification using a" siamese" time delay neural network[J]. Advances in neural information processing systems, 1993, 6.
[22] Lin Y, Qin X, Cui H, et al. Laugh Betrays You? Learning Robust Speaker Representation From Speech Containing Non-Verbal Fragments[J]. arXiv preprint arXiv:2210.16028, 2022.
[23] Deng J, Guo J, Xue N, et al. Arcface: Additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4690-4699.
[24] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[25] Snyder D, Chen G, Povey D. Musan: A music, speech, and noise corpus[J]. arXiv preprint arXiv:1510.08484, 2015.
[26] Ko T, Peddinti V, Povey D, et al. A study on data augmentation of reverberant speech for robust speech recognition[C]//2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017: 5220-5224.

Speaker verification in deliberately disguised scenarios

QIN Xiaoyi[1,2], LI Ze[1,2], LIU Dong[2], LI Ming[1,2]
(1. School of Computer Science, Wuhan University, Wuhan 430072, China;
2. Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Duke Kunshan University, Jiangsu 215316, China)

Abstract: The challenge in the task of deliberately disguised speaker verification lies in the speaker intentionally altering their voice to become someone else and thereby concealing their identity. This paper views this task as a scenario where one person plays multiple roles and proposes the CN-Movies training set and TheSound-test testing set for this task. The CN-Movies dataset is constructed by matching characters, detecting faces, recognizing faces, lip movement recognition, and voice activity detection in Chinese movies featuring actors and voice actors. This dataset includes the original voices of actors and their corresponding voice actors, leveraging the characteristic of actors and voice actors intentionally altering their voice to portray different roles, thus facilitating the collection of multi-role data for deliberate disguise. Additionally, utilizing the feature of the program TheSound, where voice actors intentionally hide their identities to avoid being recognized, this paper proposes the TheSound-test as a testing set for deliberate disguise scenarios. By combining the data mined from the above fields, this paper proposes using a Siamese network model, achieving significant improvements in speaker verification performance on both the VoxMovies test set and TheSound-test set.

Key words: speaker verification; deliberate disguise; siamese network