DOI: 10.16300/j.cnki.1000-3630.24011601 CSTR: 32055.14.sxjs.1000-3630.24011601
引用格式：郑立通, 洪峰, 郑婉, 等. 基于迁移学习和多尺度损失的短语音说话人识别方法[J]. 声学技术, 2025, **44**(4): 565-574. [ZHENG Litong, HONG Feng, ZHENG Wan, et al. A short speech speaker recognition method based on transfer learning and multi-scale loss[J]. Technical Acoustics, 2025, **44**(4): 565-574.]

基于迁移学习和多尺度损失的短语音说话人识别方法

郑立通[1,2]，洪 峰[1]，郑 婉[1,2]，许伟杰[1]
(1. 中国科学院声学研究所东海研究站, 上海 201815；2. 中国科学院大学, 北京 100190)

摘要：在面向门禁或考勤等说话人识别的应用场景中，中文短数字串语料能够提高用户使用体验，然而代价是其性能下降明显。为此，文章提出了一种基于短语音的说话人识别框架，该框架包含模型预训练阶段和迁移学习阶段。首先，提出了一种改进的预训练模型，通过特征增强和预热网络有效提高了文本无关说话人识别模型的泛化能力；其次，提出了一种多重子空间交叉熵说话人分类损失，有效提高了迁移学习阶段从源域到目标域的适配能力；最后，提出了一种长短语音嵌入码相对熵损失，通过将短语音嵌入码分布映射到音色信息更丰富的长语音分布上，从而提高性能。在中文短语音数据集 SHAL 上的实验结果表明，提出的预训练模型具有较高的泛化能力，多重子空间交叉熵损失和长短语音嵌入码相对熵损失组成的联合损失也能有效提高模型的性能。

关键词：说话人识别；短语音；迁移学习；联合损失
中图分类号：TN912.34 **文献标志码**：A **文章编号**：1000-3630(2025)-04-0565-10

A short speech speaker recognition method based on transfer learning and multi-scale loss

ZHENG Litong[1,2], HONG Feng[1], ZHENG Wan[1,2], XU Weijie[1]
(1. *Shanghai Acoustics Laboratory*, *Chinese Academy of Sciences*, *Shanghai* 201815, *China*;
2. *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

Abstract: In speaker recognition application scenarios such as access control or time and attendance oriented, Chinese short digit string corpus can improve the user experience. However, at the cost of recognition performance degradation is obvious. Therefore, this paper proposes a short speech-based speaker recognition framework which consists of a model pre-training phase and a transfer learning phase. Firstly, an improved pre-training model is proposed, which effectively improves the generalization ability of the text-independent speaker recognition model through feature enhancement and preheating network. Secondly, this paper proposes a multi-subspace cross-entropy speaker classification loss, which effectively improves the adaptation ability from the source domain to the target domain in the transfer learning stage. Finally, a long and short speech embedding code relative entropy loss is proposed to improve the performance by mapping the short speech embedding code distribution to the long speech distribution which is richer in timbre information. Experimental results on the Chinese short speech dataset SHAL show that the pre-trained model proposed in this paper has high generalization ability, and the joint loss consisting of multi-subspace cross-entropy loss and long and short speech embedding code relative entropy loss can also effectively improve the performance of the model.

Key words: speaker recognition; short speech; transfer learning; joint loss

0 引 言

说话人识别，也称作声纹识别，属于生物特征识别技术的一种[1]，其在远程身份认证中具有便捷、隐私保护等优势，引起了广泛的研究兴趣。根据应用场景说话人识别可分为说话人确认和说话人辨认[2]。说话人确认为判断当前用户是否是待验证用户，是一对一的行为。说话人辨认是从诸多用户中返回该用户的真实身份，是多对一的行为。此外，说话人识别技术从内容上可以分为文本相关和文本无关两个方向[3]。文本相关的说话人识别内容是事先约定好的。文本无关的说话人识别对识别文

收稿日期：2024-01-16; **修回日期**：2024-02-22
基金项目：上海市自然科学基金项目 (22ZR1475700)、中国科学院声学研究所自主部署“前沿探索”项目 (QYTS202114)、中国科学院青年创新促进会项目资助 (Grant No. 2021022)
作者简介：郑立通 (1998—)，男，浙江金华人，硕士研究生，研究方向为说话人识别。
通信作者：洪峰，E-mail: hongfeng@mail.ioa.ac.cn

本内容没有限制要求。虽然文本无关的说话人识别更加困难，但其应用更广泛，得益于深度学习技术的进步，说话人识别技术目前可应用于智能唤醒、金融支付、侦查破案、门禁考勤等领域[4]。

本研究聚焦基于中文数字串语料的文本无关说话人识别，主要面临以下三个挑战：(1) 文本无关任务对模型的泛化能力要求高；(2) 迁移学习时目标域对源域出现性能退化的域偏移现象；(3) 对音素信息有限的短语音识别效果差[5]。2017 年，Snyder 等[6]提出了 X-vector 方法。X-vector 方法使用时延神经网络模块[7]作为解码网络，并使用时间池化层将帧级别表示转换到句子级别表示，使得神经网络可以将不定长的语音转换成定长的说话人向量。同时得益于 Kaldi[8]的广泛使用，使得基于 Kaldi 实现的 X-vector 方法成为声纹识别领域知名度最高的方法。2018 年，文献[9]提出使用速度扰动等数据增强手段，进一步提高了识别性能。2018 年，Chung 等发布了 Voxceleb2 数据集[10]，该数据集包含了非常丰富的说话人数量，也具有语调、背景噪声等多样性，是较理想的文本无关说话人识别数据集。2020 年，Villalba 等引进了 Additive Angular Margin Softmax[11]替代 Softmax 损失函数，进一步提高了性能。2020 年，Desplanques 等提出了基于通道注意力增强与特征传播聚合的时延神经网络 (emphasized channel attention propagation aggregation time delay neural network, ECAPA-TDNN)[12]，引进通道注意力[13]模块和注意力统计池化[14]，其强大的性能使其迅速成为学术界的研究热点。2021 年，Thienpondt 提出了大角度微调 (large margin fine-tune, LMFT) 技术[15]，该技术在经过一定次数的训练后，减少学习率，提高损失函数的偏移量再次训练，可以进一步提高准确率，其核心思想在于提高对困难样本的学习能力。2022 年，快商通团队将 Subcenter 子空间技术引入损失分类层，同时也提出了 inter-topK 惩罚方法[16]，对最接近的几个非标签项结果进行惩罚，有效提高了类内和类间差异区分能力。文献[17]提出了基于 Wasserstein 生成对抗网络的说话人确认系统，联合不同准则的损失函数对网络进行优化，将短语音嵌入码映射为更具有区分性的嵌入码。

上述方法应用于 VoxCeleb1-O 等相关测试集上一定程度提高了短语音的说话人识别性能，并通过扩增数据集和优化算法等提高模型的泛化能力、缓解域偏移现象。然而在实际应用场景下，如 1~4 字以内的中文短数字串语料任务中，会出现明显的迁移学习性能退化，具体存在以下两个问题：(1) 常规的迁移学习不能完全发挥源域大规模数据训练出来的网络性能；(2) 当字数较短时，如“8173”“817”“81”这样的短语音测试性能会比“8173259604”的中文数字长语料下降很多。

过长的验证语音会降低用户体验，而过短的验证语音会影响系统性能。为此，本文提出了一种面向中文数字短语音的文本无关说话人识别框架。在验证阶段，用户仅需朗读一组 1~4 个字的中文数字词即可完成身份验证。为提高性能指标，本文从两方面进行优化改进：(1) 模型预训练方面，提出了一种改进的端到端模型框架，提升网络的泛化能力和网络对短语音的特征提取能力。(2) 迁移学习微调方面，提出了一种基于交叉熵和相对熵的多尺度损失方法，有效提高了短语音说话人识别的性能。

1 总体框架

针对预训练模型泛化能力差、迁移学习至中文数字短语音数据集后性能退化的问题，本文设计一种基于迁移学习和多尺度损失的文本无关短语音说话人识别框架，总体框架如图 1 所示。总体框架分为 3 个阶段，首先获得预训练网络，其次进行迁移学习微调，最后在相应数据集上测试。

本文设计的模型训练方法如图 1(a) 所示。首先从对应的训练数据集中取出训练语音，结合 MUSAN 噪声数据集[18]和 RIR 混响数据集[19]对其添加噪声、添加混响等常见的数据增强手段。其次增强后的语音信号经过预加重、分帧、加窗、端点检测等预处理，然后进行 Fbank 特征[20]提取，利用说话人嵌入网络提取出的说话人嵌入码计算训练损失，通过误差反向传播和梯度下降算法对网络参数进行优化。

该说话人识别整体框架包含的 3 个阶段如图 1 (b) 所示。

第 1 阶段是获取预训练模型阶段。本文利用大型的开源 VoxCeleb2 数据集，并使用子图 1(a) 中的训练方法，结合一些先进的和本课题组自己提出的特征增强手段、优化类内类间差异的损失函数、优化后的 ECAPA-TDNN 网络、微调方法等获得一个具有更强泛化能力的预训练模型，为后续的迁移学习提供更好的保障。

第 2 阶段是迁移学习及微调阶段。该阶段将预训练模型迁移至中文短语音数字数据集 SHAL 上。在该阶段，将 SHAL 数据集划分成长语音和短语音，定义 SHAL 原始数据集中 10 个字的中文数字串语音为长语音，并将 SHAL 长语音切分为 1~4

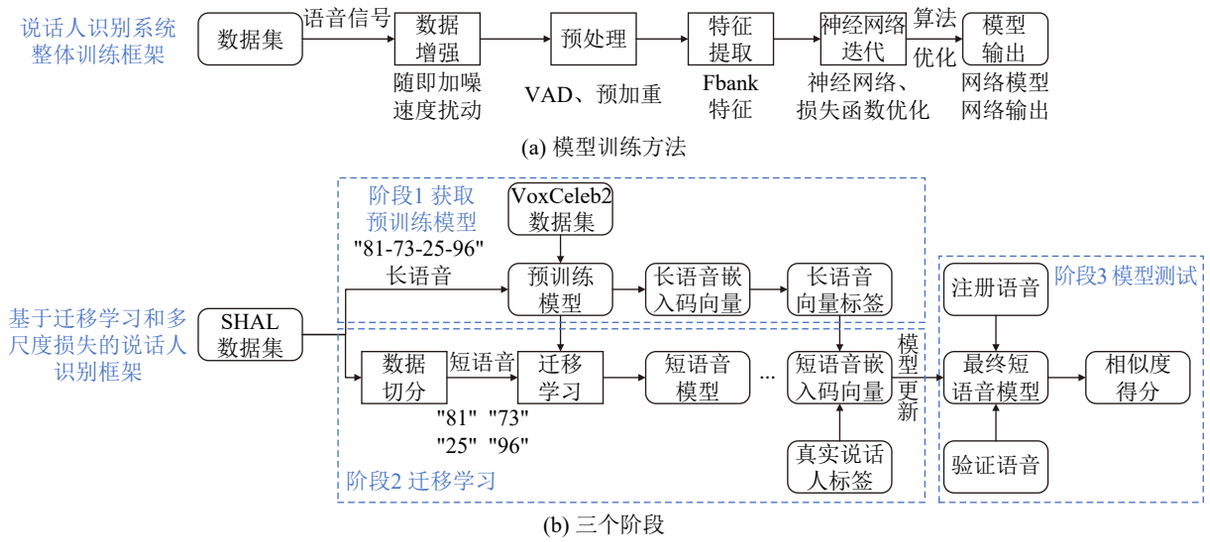

图 1　本文所提说话人识别方法的总体框架

Fig.1　General framework of speaker recongnition method proposed in this papaer

个字的语音切片定义为短语音。预训练模型的参数作为初始参数输入到短语音训练模型，训练损失采用本文提出的多重子空间交叉熵损失函数，能够有效提高模型区分困难样本的能力。此外还引入了一项和长语音相关的相对熵损失，长语音直接利用不变的预训练模型，得到长语音嵌入码，该嵌入码经过一定的处理，得到长语音嵌入码向量标签。优化长语音嵌入码标签和短语音嵌入码的相对熵损失，强制短语音嵌入码逼近表达信息更加丰富的长语音嵌入码向量。

第 3 阶段是模型测试阶段，测试不同模型在不同字数长度的语音测试集下的性能。模型测试时，按照注册语音和验证语音对的形式进行测试。注册语音为长语音，验证语音按照不同测试任务，选取不同字数长度的语音。通过模型测试，能够清晰地评估不同模型或方案对模型泛化能力和长短语音适配能力的提升效果。

本文提出的基于中文数字串的文本无关说话人识别框架的创新和优势是通过优化网络和算法调整优化得到了一个具有高泛化能力的预训练说话人识别系统。其次提出的多重子空间交叉熵损失函数与长短语音嵌入码相对熵损失函数联合优化方法，能够有效缓解短语音性能退化的现象，并解决迁移学习中目标域无法很好继承源域性能的问题。

2　基于迁移学习和多尺度损失的短语音说话人识别框架

本节将具体介绍第 1 节中提及的中文数字串文本无关说话人识别框架的模型预训练阶段和迁移学习阶段。以上阶段的模型具体结构细节如图 2 所示。

2.1　模型预训练

在说话人识别领域，一个稳健的预训练模型是十分必要的[21]。为了提高预训练模型的泛化能力，本文提出的模型预训练阶段，迁移学习阶段，多重空间损失函数分别如图 2(a)、2(b)、2(c) 所示。图 2 中，Utterances 表示多条语音输入，Mini batch 表示最小批次，num_class 表示目标分类数，Loss 表示损失函数。下面从数据增强、特征提取、网络优化、损失函数、微调方法等几个方面阐述。

训练的数据为 VoxCeleb2 数据集该数据集共包含 1 092 009 个语句和 5 994 位发言人。由于数据增强能使系统更加稳健，在这里首先采用 RIRs 和 MUSAN 进行随机添加噪声。在进行噪声添加时，信噪比从 0~20 dB 中随机产生，共分为 5 个添加噪声选项：混响噪声、音乐噪声、演讲噪声、环境噪声、不添加噪声，每种噪声选项概率保持一致。

在本文中，我们对数据增强后的信号进行特征提取。特征提取的步骤如下：(1) 对语音信号预加重，即通过一个高通滤波器改善高频分量；(2) 对给信号进行分帧加窗，减少语音信号非平稳特性的影响；(3) 对其进行短时傅里叶变换，得到时频谱；(4) 将该时频谱输入到一组三角滤波器组中，该滤波器为梅尔滤波器，能够校正人耳对语音信号的非线性；(5) 将梅尔谱取对数，即得到 Fbank 特征，若采用 80 个三角窗滤波器，则最后的特征维度为 80。在本文中，时间帧维度为 200。该特征为说话人识别领域常用的特征，本文不过多叙述。

在得到 Fbank 特征之后，还进行了时频遮掩的特征增强。传统的时频遮掩特征增强仅对某个时频

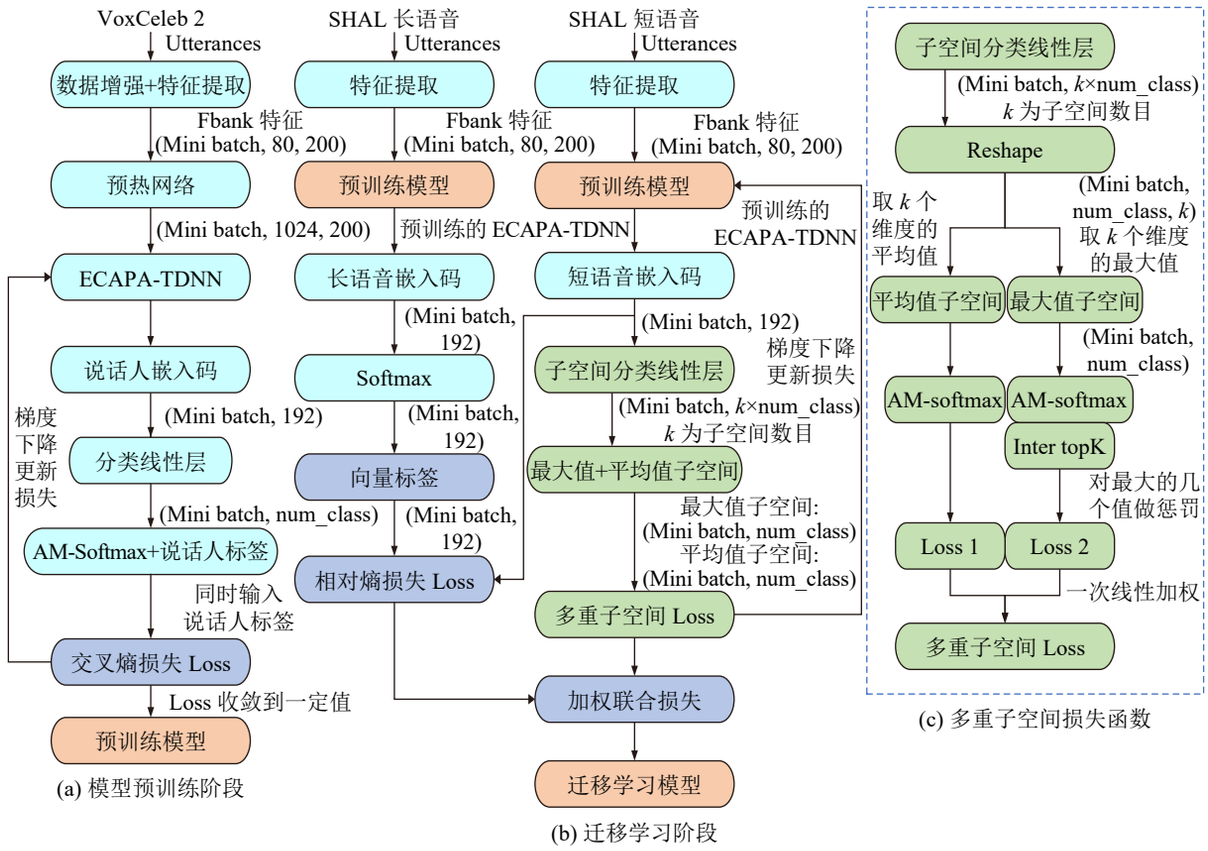

图 2 基于迁移学习和多重空间损失的短语音说话人识别框架

Fig.2 Short speech speaker recognition framework based on transfer learning and multi-space loss

段内进行遮掩，本文提出的时频遮掩范围在特征图的全时频段内随机产生。根据时间帧维度和特征维度可以计算得到特征图的总点数大小，随后产生一个从 0~10.00% 的随机遮掩密度，根据随机遮掩密度的大小计算出需要遮掩的点数，随后随机地在特征图上进行对应遮掩点数遮掩，将其特征值置零。从机理上分析，由于本文的数据集规模小，训练时容易出现过拟合现象，为了缓解过拟合现象，我们对一段音频的 Fbank 特征进行整个时频段的随机遮掩，将一张特征图上的随机数个点置零，这样使得模型每次训练的数据都有细微差别，即在一定程度上减少了过拟合。

在获得增强过的特征后，并未将其直接放入 ECAPA-TDNN 神经网络之中，我们将其先放入一个预热网络，该预热网络由三个卷积块组成，如图 3 所示。图 3 中 b 为批数据大小。

这种在模型初期使用大卷积核操作有助于扩大感受野，建立更加鲁棒和综合的特征表示，起到网络预热的作用。预热网络主要由 3 个不同卷积核大小的卷积层组成，为了获得更大的感受野，卷积核的尺寸应该尽量大。最后并将 3 个分支得到的结果相加，最终的输出作为 ECAPA-TDNN 网络的输入。损失函数采用加性边界损失函数 (additive margin softmax, AM-Softmax)：

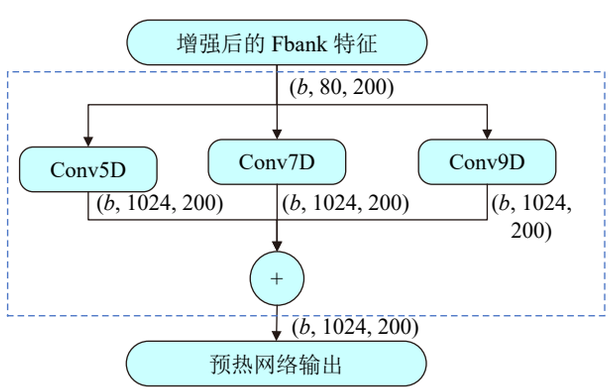

图 3 预热网络的组成

Fig.3 Composition of the pre-heating network

$$L_{\mathrm{AM}}=-\frac{1}{N}\sum_{i=1}^{N}\ln\frac{\mathrm{e}^{s(\cos(\theta_{y_i})-m)}}{\mathrm{e}^{s(\cos(\theta_{y_i})-m)}+\sum\limits_{j=1,j\neq y_i}^{n}\mathrm{e}^{s(\cos(\theta j))}} \tag{1}$$

其中：N为批处理样本数；n为总类别数；y_i表示第i条语音对应的类标签；θ_i表示第i条语音的嵌入码和第j类对应的权重向量之间的夹角；s为缩放因子；m为边缘角度偏移量。

本文采用 LMFT 微调方法，该做法能够提高模型区分困难样本的能力，具体步骤是：(1) 在获

得正常训练的模型后加载该模型，然后适当调低学习率。(2) 将训练的语音帧长从 200 帧调整到 600 帧，不足的部分将语音的前端部分补足到后端。(3) 将损失函数的 Margin 值从 30 提高到 50[16]。随后训练较少的轮次 (epoch) 即可完成微调，最终得到迁移学习所需的预训练模型。

2.2　迁移学习阶段

为了提高模型的泛化能力，我们进行了基于 VoxCeleb2 数据集的模型预训练。随后将该模型迁移学习至中文短数字串数据集 SHAL 上。常规的迁移学习时常出现性能退化等域偏移现象，为了使目标域更好地继承源域的泛化性能，本文提出了一种多重子空间一次加权的交叉熵损失函数。此外，本文研究的数字串数据在语音文本字数长度变短时，性能出现明显的下降，为此本文提出了联合长语音嵌入码相对熵损失函数优化的方法，如图 2(b) 所示。进行短语音数据集迁移学习时，为了更好地监测本文方法对模型性能的提升，在迁移学习阶段没有使用任何数据增强手段。除损失函数外，迁移学习的超参数配置和预训练模型阶段保持一致。

2.2.1　多重子空间交叉熵损失函数

本文迁移学习采用的多重子空间如图 2(c) 所示。首先加载预训练模型的网络参数，并对短语音进行嵌入码提取，提取到的短语音嵌入码为$x\in\mathbb{R}^{192\times1}$，其中 192 为嵌入码维度。得到短语音嵌入码后，将该嵌入码经过一个线性分类层，一般的分类层维度为训练数据集的说话人人数n，则该线性层的权重为$W\in\mathbb{R}^{n\times192}$，该权重与语音嵌入码相乘则可得到相应的余弦值，记为S，$S\in\mathbb{R}^{n\times1}$。余弦值越大则说明该类别的预测概率越大。在本文的多重子空间损失中，将其维度大小调整成说话人数目的 k 倍，因此经过线形层的权重为$W_k\in\mathbb{R}^{(n\cdot k)\times192}$，其中 k 为子空间的数目。此时得到的余弦值记为S'，$S'\in\mathbb{R}^{n\times k}$，通过对子空间维度 k 进行压缩，可将S'的维度压缩成与S一致。多重子空间示意图如图 4 所示。

本文考虑两个子空间：平均值子空间和最大值子空间。平均值子空间为求取k个子空间的平均值，并将其空间压缩至原来大小计算公式为

$$\cos(\theta_{i,j})=\underset{1\leqslant k\leqslant K}{\text{average}}(\|x_i\|\cdot\|W_{j,k}\|) \tag{2}$$

其中：x_i为短语音嵌入码；$W_{j,k}$为 k 个子空间的权重；average(·) 代表求取平均子空间的操作，得到的结果可以代入式 (1) 进行损失函数计算，作为平均值子空间的损失部分，把该值记为$L_{\text{AM-ave}}$。

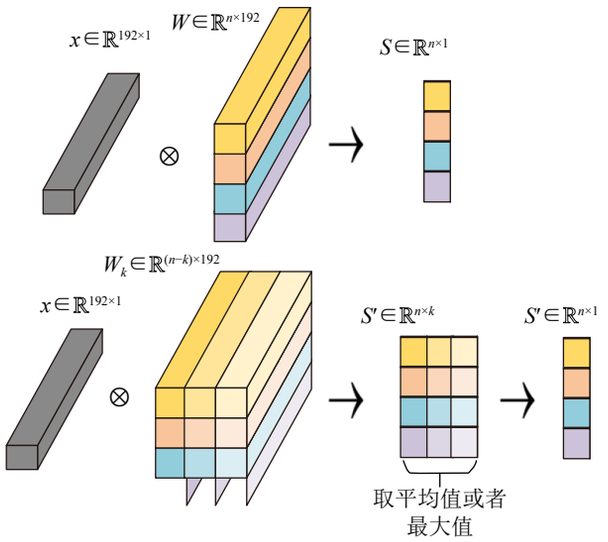

图 4　多重子空间示意图

Fig.4　Schematic diagram of the multi-subspace

最大值子空间为求取 k 个子空间的最大值，同样将其维度压缩，计算公式为

$$\cos(\theta_{i,j})=\max_{1\leqslant k\leqslant K}(\|x_i\|\cdot\|W_{j,k}\|) \tag{3}$$

其中：max(·) 代表求取子空间的最大值。此外在计算其 AM-Softmax 损失时，还引入一个惩罚项：

$$\phi\left(\theta_{i,j}\right)=\begin{cases}\cos\theta_{i,j}+p & ,j\in\arg\operatorname{top}K(\cos\theta_{i,n})\\ \cos\theta_{i,y_i} & ,\text{其他}\end{cases} \tag{4}$$

其中，p为惩罚项，添加在数个相似度较高的邻近负样本处。总体而言，添加惩罚项后的损失如式 (5) 所示：

$$L'_{\text{AM}}=-\frac{1}{N}\sum_{i=1}^{N}\ln\frac{\mathrm{e}^{s[\cos(\theta_{y_i})-m]}}{\mathrm{e}^{s[\cos(\theta_{y_i})-m]}+\sum\limits_{j=1,j\neq y_i}^{n}\mathrm{e}^{s[\phi(\theta_{i,j})]}} \tag{5}$$

其中：$\phi(\theta_{i,j})$如式 (4) 所定义。对$\cos(\theta_n)$进行排列，将损失项添加到最大的几个负样本处，得到的最大值子空间损失记为$L_{\text{AM-max}}$。

平均值子空间基于所有样本的平均状态。这种方法特别适用于模型训练的早期阶段，因为它能为模型提供一个稳定的、基于整体趋势的学习基准。随着训练的深入，我们引入了最大值子空间，它从所有样本空间中提取最大值。最大子空间结合相应的惩罚项，有效地提升了正负样本概率的余弦值，从而推动模型朝着更高性能的方向发展。最大值子空间适用于训练的后期阶段，能帮助模型更有效地区分难以识别的样本，从而显著提高模型的整体识别能力。将平均值子空间损失和最大值子空间损失一次线性加权得到最终的交叉熵说话人分类损失：

$$L_{\text{CE}}=\alpha_1 L_{\text{AM-ave}}+\beta_1 L_{\text{AM-max}} \tag{6}$$

其中：α_1为与训练轮次 E 负相关的常数，β_1为与训练次数正相关的常数。

2.2.2 长短语音嵌入码相对熵损失函数

在实验中发现，长语音在迁移学习方面表现更佳，测试性能上显著优于短语音。长语音包含更丰富的音素和语音特征，能够提供更全面的声音信息，由于其包含更多样化的语音特征，长语音能够更有效地帮助模型学习并适应新的语音任务或环境。这种广泛的特征覆盖使得模型在迁移到不同的语音任务时更加鲁棒和灵活。基于这些观察，本文提出了一种新的长短语音嵌入码相对熵损失方法。该方法通过计算长短语音嵌入码之间的相对熵损失 (如图 2(b) 的迁移学习阶段左侧部分所示) 来优化短语音的嵌入表示。在迁移学习训练时，同时输入某一说话人的长语音和短语音。长语音经过特征提取，经过梯度固定不变的预训练模型得到长语音嵌入码$x_l \in \mathbb{R}^{192\times1}$，其中 192 为嵌入码维度，随后进行 Softmax 操作，得到的结果为长语音嵌入码向量标签，记为$x'_l \in \mathbb{R}^{192\times1}$。同时短语音经过一个梯度不固定的预训练网络，得到短语音嵌入码，同样对其进行特定维度归一化操作。最后将长语音嵌入码作为真实值标签，将短语音嵌入码作为预测值输入到相对熵损失函数中。本文的长短语音相对熵损失函数L_{KL}为

$$L_{\mathrm{KL}} = \sum_{i=1}^{192} x'_l(i) \ln \frac{x'_l(i)}{x'_s(i)} \tag{7}$$

其中：$x'_s(i)$为短语音嵌入码经过 Softmax 之后的结果，i为嵌入码的维度。相对熵也称为库尔贝克-莱布勒 (Kullback-Leibler) 散度，是一种衡量两个概率分布差异的变量。在本文中，这个变量被用来量化短语音与长语音之间概率分布的差异。这种方法本质上建立了一种短语音到长语音的映射关系，通过最小化相对熵损失，实际上使短语音的概率分布尽可能接近长语音的分布。由于长语音含有更加丰富的信息，这种接近过程有助于丰富短语音的表征，使之包含更多的上下文信息和说话人特征。这不仅提高了短语音的表现力，也使其在声音处理任务中的表现更加准确和可靠。最终的损失函数结合了前文的多重子空间交叉熵分类损失，以及长短语音嵌入码相对熵损失，联合优化的损失为

$$\mathrm{Loss} = \alpha_2 L_{\mathrm{CE}} + \beta_2 L_{\mathrm{KL}} \tag{8}$$

其中：α_2和β_2分别为固定的常数；L_{CE}是前文的多重子空间交叉熵损失函数；L_{KL}是长短语音嵌入码相对熵损失函数。

通过上述提出的联合多重子空间交叉熵损失函数和长短语音嵌入码相对熵损失函数的多尺度损失函数方法，能够有效地将源域强大的泛化能力迁移至目标域，同时使目标域在面对一些难以区分的困难样本时具有更稳健的性能，在一定程度上减轻了域偏移，适应了目标域分布。

3 数据集和实验设置

3.1 数据集

预训练数据集为 VoxCeleb2，该数据集共包含 5 994 个说话人，共 1 092 009 个语句，其中 61.00% 的说话人为男性，涵盖了不同种族、口音、语言的语音数据。

迁移学习的数据集为中国科学院声学研究所东海研究站自行录制的数字串数据集 SHALCAS dataset(简称 SHAL)，是截至目前少有的开源免费中文数字串数据集。该数据集包含时长约 72.30 h 的音频，有 46 583 条，侧重年龄在 10~40 岁之间的说话者，并兼顾性别。录制文本内容和语言条数如表 1 所示。

表 1 SHAL 数据集录制的文本信息
Table 1 Text messages of the SHAL dataset recording

文本标签	文本内容	语音条数
d001	8|1|7|3|2|5|9|6|0|4	1 200
d002	8-1-7-3|2-5-9-6|-0-4	1 440
d003	8-1-7|-3-2-5|-9-6-0|-4	1 500
d004	8-1|-7-3|-2-5|-9-6|-0-4	1 500
d005	9-4-0-5|3-7-2-6|-8-1	1 500
d006	9-4-0|-5-3-7|-2-6-8|-1	1 500

注：“|”表示录制时的停顿节奏。

我们从 SHAL 数据集中挑选了 60 个说话人，其中 50 个说话人的数据用于训练，剩下 10 人的语音数据用于测试，每个说话人的每种文本标签 (d00x) 语料数量在 20~25 条之间，每条语料时长约 4 s。我们将 SHAL 数据集公开在了 openSLR 网站上，网址为：http://www.openslr.org/138/。

在模型测试方面，通过对一条语音进行语音识别和双门限法端点检测，进行了限定字数的切分。区别于传统采用不同时间长度语音数据来测试模型性能的方法，本文采用不同数字长度语音来测试模型性能。

本文中提及的“长语音”，即为任一 d00x 包含 10 个中文数字串的语音。本文研究的“短语音”，是从字数上定义的，SHAL 数据集录制时具有一定的停顿节奏，因此可以根据该节奏进行语音

的切分，切分得到的数据质量较高，可作为短语音数据集。如将某个说话人的一条 d001 语音进行切分，得到 10 条文本内容从“0”到“9”的语音数据，如把 d002 切分成“8173”，“2596”4 个字的语音数据。切分后的子数据集称为 SHAL-S，具体信息如表 2 所示。

表 2　SHAL-S 短语音数据集信息
Table 2　Information on the SHAL-S short voice data set

文字数量	文本内容	语音条数
1字	0, 1, 2, 3, 4, 5, 6, 7, 8, 9	3810
2字	81, 73, 25, 96, 04	3565
3字	817, 325, 960, 940, 537, 268	6370
4字	8173, 2596, 9405, 3726	4964
混合	1字，2字，3字，4字	24649

注：“混合”子数据集为1~4字数据集的混合。

3.2　实验设置

长语音嵌入码通过计算某个说话人的 50 条长语音嵌入码的平均值得到，该方法能够提高长语音嵌入码标签的稳健性。本文的一次测试分为注册语音和验证语音，通过注册-验证语音样本对的形式进行模型测试打分，计算注册-验证样本对两条语音的相似度，并根据真实标签进行性能测试。评价指标采用说话人识别领域常用的等错误率 (equal error rate，EER)。EER 指当假接受率 (false acceptance rate, FAR) 和假拒绝率 (false rejection rate, FRR) 相等时的数值。在这一点上，系统对错误接受和错误拒绝的容忍程度是相同的。EER 越低，表示说话人识别系统的整体性能越好。

在实验中，将语音的采样率降采样至 16 kHz，截取 2 s 的语音，不足时进行补零。Fbank 特征提取的傅里叶变换点数为 512，窗长为 400，窗移为 160，添加汉明 (hanmming) 窗，提取 80 维的特征。

网络层面采样 ECAPA-TDNN，说话人嵌入码维度为 192。说话人分类损失为 AM-Softmax，m 为 0.20(迁移学习时为 0.50)，s 为 30.00。优化器为 Adam，初始学习率为 0.001(迁移学习时为 0.000 1)，每 1 个 epoch 下降 3.00 个百分点，共训练 80 个 epoch。

4　实验结果与分析

4.1　预训练模型消融实验

为了提高模型的泛化能力，本文使用 VoxCeleb2 数据集进行模型的预训练，采用 ECAPA-TDNN 网络，同时为了得到一个稳健的声纹识别系统，在模型训练时采用许多策略：数据增强、特征增强和预热网络。针对以上各个模块进行了消融实验。测试集为 VoxCeleb1-O 数据集和 SHAL-S 数据集的 1 字，2 字，3 字，4 字，混合测试集。SHAL-S 数据集的测试注册语音嵌入码为 5 条长语音的平均嵌入码，验证语音为每一条不同字数长度的语音。消融实验结果如表 3 所示。

表 3　预训练模型消融实验
Table 3　Pre-training model ablation experiments

训练模型	EER/%				
	VoxCeleb1 数据集	1字	2字	3字	4字
基线	1.28	9.32	5.00	1.29	0.67
+	1.20	10.00	5.16	1.20	0.56
++	1.15	8.47	5.65	1.17	1.02
+++	1.01	7.46	4.35	1.06	0.48

注：“+”为基线基础上添加数据增强模块，“++”为在“+”基础上添加时频遮掩特征增强模块，“+++”为在“++”基础上添加预热网络模块。

表 3 的消融实验表明，各个模块在 VoxCeleb1-O 测试集上的性能都有一定的提升。基线网络是目前说话人识别领域最主流的 ECAPA-TDNN 网络，我们将在其基础上进行优化，并与基线网络结果对比。传统的数据增强和本文提出的全时频段的时频遮掩方法都提高了模型在 VoxCeleb1-O 数据集上的性能。从结果上来看，在“++”实验中添加的时频遮掩模块，在时频谱图的不同频率和时间段上引入噪声或屏蔽部分频谱来模拟各种环境下的语音变化，增加模型对于环境噪声和其他干扰的鲁棒性。性能提升最大的是添加了预热网络的“+++”模型，EER 性能最佳达到了 1.01%，相比于传统方法直接输入特征，通过一个预热网络相当于提前对特征进行了训练，能加速后面模型收敛，提高泛化性能。然而在短语音数据集 SHAL 上，“+”与“++”模型的性能不稳定，这是因为训练集和测试集存在域失配的现象。VoxCeleb1-O 测试集更接近训练集 VoxCeleb2，因此一些对于数据加噪的手段能够提高模型的鲁棒性，然而 SHAL 短语音数字数据集时长更短，抗噪声的鲁棒能力相对较差，因此采用数据增强和时频遮掩方法训练得到的模型直接运用在 SHAL 测试集上时会出现性能退化的现象。添加了预热网络模块之后的模型在两个测试数据集上都达到了最佳性能。这是因为预热网络在模型的初始阶段就通过多个大卷积块相加的放大获得一个较大的全局感受野，同时也弥补了数据加噪和时频遮掩带来的部分弊端，使模型具有良好的泛化能力和鲁棒性。因此我们选择“+++”模型作为预训练模型，该模型的 EER 为 1.01%，

相比于基线模型，在表 3 所示的各个测试集上性能分别提升了 21.1%，20.0%，13.0%，17.8%，28.4%。从消融实验的整体上看，本文的数据增加、特征增强和预热网络的方法，有效提高了模型的分类性能，这为后续的迁移学习做好了充足的准备。

4.2 迁移学习实验

4.2.1 多重子空间交叉熵损失函数测试实验

在迁移学习阶段，使用前文得到的综合最优的预训练模型“+++”。首先确定数种迁移学习的策略：(1) 重新训练预训练模型所有参数；(2) 冻结预训练的所有参数；(3) 冻结预训练的所有参数，添加两个线性层重新训练。针对以上 3 个策略，首先探究了单独使用最大值子空间损失和平均值子空间损失时不同迁移学习策略的性能。结果如表 4 所示。表 4 中加粗的数字表示最优值，下同。

表 4 不同子空间损失及不同迁移学习策略实验的 EER 结果

Table 4 EER results of the experiments with different subspace losses and different transfer learning strategies

模型	策略	EER/%				
		混合	1字	2字	3字	4字
Ave	策略(1)	**4.67**	8.36	4.41	**0.81**	**0.45**
Ave	策略(2)	8.52	7.46	**4.35**	1.06	0.48
Ave	策略(3)	10.31	23.90	13.28	4.38	3.12
Max	策略(1)	6.34	**8.08**	6.77	0.89	0.69
Max	策略(2)	8.52	7.46	**4.35**	1.06	0.48
Max	策略(3)	10.15	23.22	12.83	4.73	3.12

注: Ave 代表模型只使用平均值子空间损失; Max 代表只使用最大值子空间损失; 策略 (i) 表示迁移学习方法, i=1, 2, 3。

由表 4 可知，策略 (3) 是冻结所有参数后新增两个线性层再训练的迁移学习方法。由于新增的线性层无法享受到预训练模型的高泛化能力，因此其测试结果较差。策略 (2) 是冻结所有参数，由于梯度并未更新，其结果与预训练模型消融实验中“+++”模型的结果保持一致。对于策略 (1)，模型所有参数全部重新训练的策略，在测试集为“3 字”“4 字”“混合”时，平均值子空间损失的结果最优，在测试集为“1 字”时，最大值子空间损失的结果最优。这可能是因为当字数较多时，平均值子空间通过调大 Margin 值，能够正常的进行模型迁移学习，使得目标域从源域处较好地提升了性能。而当字数较少时，可看作困难样本增多，而最大值子空间损失部分包含一定的惩罚项，以及相比于平均值子空间损失具有更好的针对困难样本的类内类间区分能力，因此其在字数较少时性能有提升。

基于上述实验，本文得出了初步的假设和结论：在模型训练前期令平均值子空间的损失为损失主要部分，在模型训练中后期令最大值子空间的损失为损失的主要部分。因此结合平均值子空间和最大值子空间，形成多重子空间的交叉熵损失函数，并赋予两者在模型训练不同时间段不同的权重，参数全部重新训练，相关实验如表 5 所示。

表 5 多重子空间损失不同权重组合实验的 EER 结果

Table 5 EER results of the experiments on combining different weights for multiple subspace losses

(α_1,β_1)	EER/%				
	混合	1字	2字	3字	4字
(E/80,1-E/80)	3.80	8.17	2.58	0.68	0.55
(E/40,1-E/80)	3.60	8.08	2.90	0.61	0.45
(E/20,1-E/80)	3.55	7.97	**2.10**	0.60	0.36
(1-E/80,E/80)	3.60	8.17	2.58	**0.34**	0.35
(1-E/80,E/40)	**3.54**	**7.68**	**2.10**	**0.34**	0.33
(1-E/80,E/20)	3.60	8.08	3.60	0.42	**0.29**

注: (α_1,β_1)为式 (6) 中的权重系数, E 代表模型训练时的 epoch。

受表 4 的启发，使平均值子空间和最大值子空间损失的系数与训练次数成正相关或负相关。由表 5 可得，平均值/最大值子空间损失随着训练次数增加而递减/递增的设置总体上优于随着训练次数增加而递增/递减的设置。综合最优的(α_1,β_1)系数设置为 (1-E/80, E/40)，此时的模型性能在“混合”“1 字”“2 字”“3 字”测试集上都达到了最优。在该设置下，模型在训练次数较小时，平均值子空间损失函数为主要部分，保证了模型梯度下降时整体方向趋势的正确，在训练次数较大时，最大值子空间损失函数为主要部分，提高了模型在后期优化困难样本的能力。

4.2.2 长短语音嵌入码相对熵损失函数测试实验

在上述实验中，依然存在语音长度较短时，模型性能退化快，短语音性能与长语音差距较大的问题。为此我们提出并引入了长短语音嵌入码相对熵损失函数，相关实验结果如表 6 所示。

表 6 长短语音嵌入码相对熵损失函数的 EER 结果

Table 6 EER results of the relative entropy loss function for long and short speech embedding codes

(α_2,β_2)	EER/%				
	混合	1字	2字	3字	4字
(1,1)	**3.17**	8.14	2.40	**0.34**	**0.22**
(1,10)	3.50	7.87	2.31	0.34	0.24
(1,100)	3.37	**6.95**	**1.81**	0.34	0.27

注：(α_2,β_2)为式(8)中的权重系数。

在表 5 中的“(1-E/80, E/40)”多重子空间损失权重系数设置下，引入了长短语音嵌入码相对熵损

失。同样赋予其一定的系数，在实验过程中，发现该损失的数值一般较小，前文的多重子空间损失初始值约为 20，长短语音嵌入码相对熵损失的初始值约为 0.5。因此三组实验的相对熵损失权重系数分别为 1，10，100。从表 6 可知，当该相对熵损失的权重系数β_2较小，为 1 时，“3 字”“4 字”的性能达到了最优，而随着相对熵权重系数β_2的增大，“4 字”测试集的性能反而下降。相反的是，当β_2为 1 时，“1 字”“2 字”的测试集性能比β_2为 100 时的性能差，且具随着β_2增大，EER 有明显的下降趋势。这可能由于“3 字”“4 字”的测试语音具有一定的文本长度，因此其不需要过多的相对熵损失函数帮助，也能达到一个不错的结果。然而对于字数较少的“1 字”“2 字”数据集，其短语音的缺陷更加明显，因此增大相对熵损失部分，能够使模型学习到一个更好的程度。

总体而言，在 SHAL-S 数据集上的测试结果，本文方法相比于基线模型，“1 字”“2 字”“3 字”“4 字”测试子集上 EER 性能分别提升了 25.43%，63.80%，73.64%，67.16%。相比于迁移学习用到的预训练模型，在“1 字”“2 字”“3 字”“4 字”测试子集上性能分别得到了 6.84%，58.39%，67.92%，43.75% 的性能提升。由此可见，当字数为 2 或者 3 时，本文方法性能提升明显，较好地验证了我们的假设。当字数为 4 字时，性能也有较高的提升，但当字数为 1 时，由于其字数过少带来的缺陷，性能提升相对较少。从性能的绝对值看，“3 字”与“4 字”的 EER 指标远远优于“2 字”与“1 字”。考虑在实际的应用中，会受到多种噪声和信道的干扰，因此相关性能会比实验结果更低，因此验证时的语音字数也不能过短，以确保验证的成功率。“3 字”的短数字串语音不仅包含一定数量的音素，同时也能够提高用户的使用体验。

本文对采用的多重子空间损失以及长短语音相对熵损失进行了记录，最终模型的损失函数结果如图 5 所示。图 5(a) 为平均值子空间的损失函数，由于其权重系数在模型的训练前期较大，因此其下降的趋势更陡。图 5(b) 为最大值子空间的损失函数，由于其权重系数在模型的中后期较大，因此其下降的趋势相比平均值子空间，更加平稳。图 5(c) 为长短语音嵌入码相对熵损失函数，图 5(d) 为总体损失函数。当总体损失函数接近收敛时，相对熵损失函数仍处于不平稳阶段，这是因为采用的训练数据集 SHAL 规模不是很大，在短语音嵌入码向长语音嵌入码映射的过程中，已有的数据暂时无法

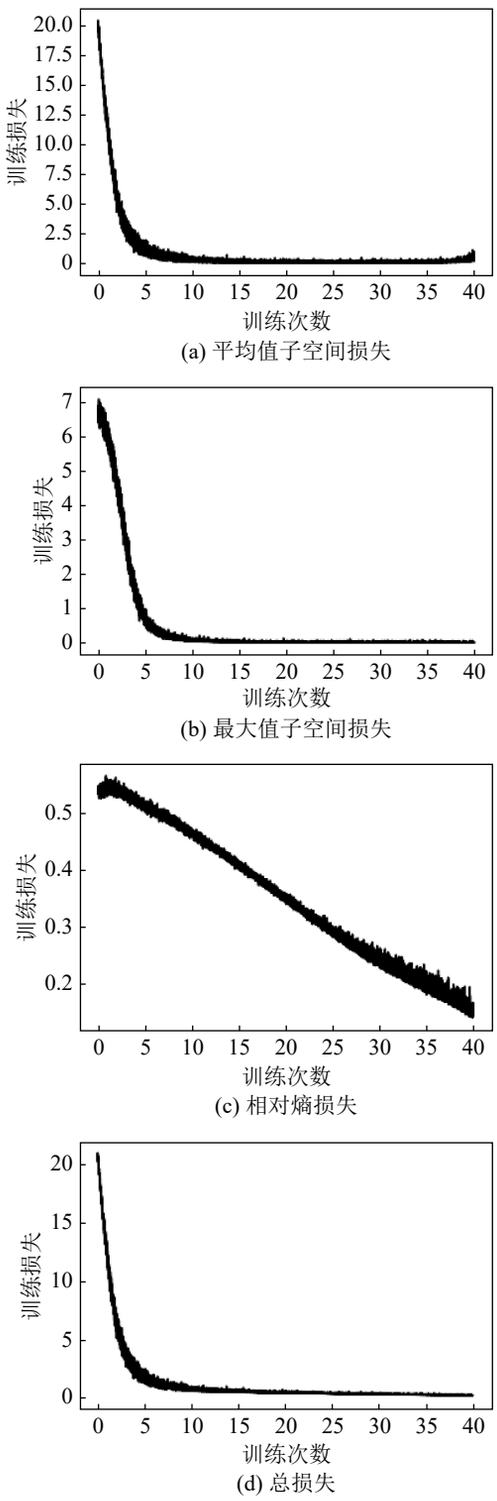

图 5　损失函数图

Fig.5　Loss function plot

达到完全收敛的平稳阶段，但是其稳步下降的数值以及模型性能的提升证明了该方法的有效性。

相比于基线模型，本文方法的最终的结果在“1 字”“2 字”“3 字”“4 字”测试子集上 EER 性能绝对数分别提升了 2.37，3.19，0.95，0.45 个百分点。相比于本文提出的预训练模型，最终的结

果在“1 字”“2 字”“3 字”“4 字”测试子集上 EER 性能分别提升了 0.51，2.54，0.72，0.26 个百分点。

总体来说，本文最后的引入相对熵损失的模型，能够较好应对 2~3 字的短语音状况，然而其还存在一定的缺陷：受 ECAPA-TDNN 网络影响，模型的稳定性一般，在模型参数搜索方面需要花费较大的功，这也是当前说话人识别领域的重要研究方向之一。

5 结 论

本文从说话人识别相关应用的用户使用体验出发，针对中文短数字串语料展开了相关研究。为了解决预训练模型泛化能力不高以及迁移学习时源域与目标域的域失配等问题，本文提出了一种基于中文数字串短语音的说话人识别框架。该框架由模型预训练阶段和模型迁移学习阶段组成。首先，在模型预训练阶段提出了一种时频遮掩方法和预热网络，对特征图全时频段的时频随机遮掩，以及在网络前端添加多个大卷积核模块组成的预热网络，通过消融实验证明以上方法能够提高预训练网络的泛化性和鲁棒能力。其次，在迁移学习阶段提出了一种多尺度联合损失。该损失由多重子空间交叉熵损失和长短语音嵌入码相对熵损失组成。通过实验证明了在模型训练的不同时间段赋予平均值子空间和最大值子空间不同的权重分配能够提升迁移学习的性能。最后的相对熵损失权重实验证明了该部分损失在权重较小时对数字较多的语音性能提升明显，在权重较大时对数字较少的语音性能提升明显。我们最优的模型在 SHAL-S 1 字，2 字，3 字，4 字的测试集上性能提升分别为 25.43%，63.80%，73.64%，67.16%。

参 考 文 献

[1] BAI Z X, ZHANG X L. Speaker recognition based on deep learning: an overview[J]. Neural Networks, 2021, **140**: 65-99.

[2] FURUI S. Recent advances in speaker recognition[J]. Pattern Recognition Letters, 1997, **18**(9): 859-872.

[3] TU Y Z, LIN W W, MAK M W. A survey on text-dependent and text-independent speaker verification[J]. IEEE Access, 2022, **10**: 99038-99049.

[4] 郑方, 李蓝天, 张慧, 等. 声纹识别技术及其应用现状[J]. 信息安全研究, 2016, **2**(1): 44-57.
ZHENG Fang, LI Lantian, ZHANG Hui, et al. Overview of voiceprint recognition technology and applications[J]. Journal of Information Security Research, 2016, **2**(1): 44-57.

[5] CHEN W D, HUANG J, BOCKLET T. Length- and noise-aware training techniques for short-utterance speaker recognition[C]//Interspeech 2020. ISCA: ISCA, 2020: 3835-3839.

[6] SNYDER D, GARCIA-ROMERO D, POVEY D, et al. Deep neural network embeddings for text-independent speaker verification[C]//Interspeech 2017. ISCA: ISCA, 2017: 999-1003.

[7] WAIBEL A, HANAZAWA T, HINTON G, et al. Phoneme recognition using time-delay neural networks[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1989, **37**(3): 328-339.

[8] POVEY D, GHOSHAL A, BOULIANNE G, et al. The Kaldi speech recognition toolkit[C]//IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society, 2011.

[9] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-vectors: robust DNN embeddings for speaker recognition[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB. IEEE, 2018: 5329-5333.

[10] CHUNG J S, NAGRANI A, ZISSERMAN A. VoxCeleb2: deep speaker recognition[EB/OL]. 2018: 1806.05622. https://arxiv.org/abs/1806.05622v2.

[11] VILLALBA J, CHEN N X, SNYDER D, et al. State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the Wild evaluations[J]. Computer Speech & Language, 2020, **60**: 101026.

[12] DESPLANQUES B, THIENPONDT J, DEMUYNCK K. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification[C]//Interspeech 2020. ISCA: ISCA, 2020: 1-5.

[13] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT. IEEE, 2018: 7132-7141.

[14] OKABE K, KOSHINAKA T, SHINODA K. Attentive statistics pooling for deep speaker embedding[C]//Interspeech 2018. Hyderabad, India, ISCA, 2018: 2252-2256.

[15] THIENPONDT J, DESPLANQUES B, DEMUYNCK K. The idlab voxsrc-20 submission: large margin fine-tuning and quality-aware score calibration in DNN based speaker verification[C]//ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada. IEEE, 2021: 5814-5818.

[16] ZHAO M, MA Y F, DING Y W, et al. Multi-query multi-head attention pooling and inter-topk penalty for speaker verification[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore. IEEE, 2022: 6737-6741.

[17] LIU K, ZHOU H. Text-independent speaker verification with adversarial learning on short utterances[C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain. IEEE, 2020: 6569-6573.

[18] SNYDER D, CHEN G G, POVEY D. MUSAN: a music, speech, and noise corpus[EB/OL]. 2015: 1510.08484. https://arxiv.org/abs/1510.08484v1.

[19] 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). New Orleans, LA. IEEE, 2017: 5220-5224.

[20] 王泉. 声纹技术: 从核心算法到工程实践[M]. 北京: 电子工业出版社, 2020.

[21] CHEN Z Y, HAN B, XIANG X, et al. Build a SRE challenge system: lessons from VoxSRC 2022 and CNSRC 2022[EB/OL]. 2022: 2211.00815. https://arxiv.org/abs/2211.00815v2.