E在达里语语音中应用混合深度神经网络

Jawid Ahmad Baktash[1] 和 Mursal Dawodi[2]

LIA，阿维尼翁大学，阿维尼翁，法国[1]
jawid.baktash1989@gmail.com，[2]mursal.dawodi@gmail.com

摘要

本文是我们此前发表的会议论文的扩展。近年来，研究人员对开发和改进语音识别系统的兴趣日益浓厚，旨在促进并增强人机交互。如今，自动语音识别（ASR）系统已无处不在，应用范围从游戏到翻译系统、机器人等。然而，针对低资源语言的语音识别系统仍有大量研究工作亟待开展。本文重点研究达里语单词识别，采用梅尔频率倒谱系数（MFCCs）特征提取方法，并使用三种不同的深度神经网络模型：卷积神经网络（CNN）、循环神经网络（RNN）和多层感知机（MLP），以及两种结合CNN与RNN的混合模型。我们使用自行创建的达里语孤立词语料库对这些模型进行评估，该语料库包含20个简短达里语术语的1000条话语。研究取得了令人瞩目的98.365%平均准确率。

关键词

达里语，深度神经网络 网络、语音识别、循环神经网络、多层感知机、卷积神经网络
l 网络

1. INTRODUCTION

人类通过语音相互交流。随着技术进步和智能手机的广泛使用，自动语音识别（自动语音识别）工具极大地促进了人机交互。自动语音识别系统有许多应用，包括人机对话系统、机器人技术、翻译工具等。然而，针对许多地区性语言和亚洲语言的自动语音识别研究仍面临诸多挑战，例如资源匮乏和方言差异。

达里语是阿富汗的官方语言之一，全球约有3200万人将其作为第一语言或第二语言使用。尽管许多达里语使用者生活在阿富汗境外，但由于难以获取现代技术设备和应用程序，他们往往面临沟通障碍。此外，针对达里语及其他区域语言的自动语音识别（ASR）研究仍处于起步阶段，迄今为止仅有少数相关研究。

本文旨在通过采用五种不同的深度学习技术，开发一种更加稳健且错误率更低的达里语自动语音识别系统，以实现达里语孤立词自动语音识别。这些技术包括多层感知机（MLP）、卷积神经网络（CNN）、长短期记忆网络（LSTM）、CNN与LSTM的混合模型，以及CNN与双向长短期记忆网络（BLSTM）的混合模型。随后，我们对不同模型的性能进行比较，以评估其效率。本研究通过比较五种最先进的深度学习方法的有效性，为达里语语音识别领域带来了新的研究成果。

据我们所知，尚无已发表文章比较不同模型在达里语自动语音识别上的性能。开发达里语自动语音识别系统的主要挑战是缺乏可用的数据集以及方言的多样性。我们通过开发和使用我们构建的孤立达里语词语料库来应对这一挑战，该语料库由针对20个简短达里语术语的1000条话语组成。

Dawodi 等人 [3] 建立了一个针对达里语的自动语音识别系统，该系统采用梅尔频率倒谱系数进行特征提取，并使用卷积神经网络进行词分类，达到了 88.2% 的平均准确率。然而，当前研究在他们的研究基础上进一步开展。

通过实施三种额外的深度学习技术开展工作，并使准确率提高了 10% 以上。下一节将讨论该领域的一些相关研究，第 3 节将简要介绍语料库结构。随后，第 4 至第 6 节将概述达里语语音识别，包括梅尔频率倒谱系数特征提取和深度神经网络模型。第 7 节将阐述本研究的结果与讨论。最后，第 8 节将总结本文并报告未来工作。此外，作者将说明，本文是其先前会议论文的扩展。

2. Related Works

多项研究使用机器学习和自然语言处理技术来执行语音识别任务，近期重点则放在利用深度学习技术开发自动语音识别系统上。然而，大多数研究都聚焦于英语等非区域性语言。Mitra 等人 [8] 提出了一种用于对声学空间和发音空间进行建模的混合卷积神经网络（HCNN）。与卷积神经网络（CNN）/深度神经网络（DNN）相比，HCNN 模型表现更佳，词错误率也更低。类似地，Grozdic 和 Jovicic [6] 开发了一种新框架，该框架基于自动深度解码器编码器系数和 Teager 能量存储系数。他们指出，在描述耳语方面，基于 Teager 能量的倒谱特征比梅尔频率倒谱系数（MFCC）和高斯混合模型—隐马尔可夫模型（GMM-HMM）更具优势；在耳语场景下，其准确率高出 31%，词语识别率达到 92.81%。

近十年来，一些研究聚焦于更多区域性语言。Dua 等人开发了一种旁遮普语自动语音识别系统[4] ，使用梅尔频率倒谱系数进行特征提取，并使用 HTK 工具包进行识别。他们通过 8 位旁遮普语母语者的语音，准备了包含 115 个旁遮普语词条的数据集。总体平均准确率达到 94%–96%。Bakht Zada 和 Rahimullah [17] 开发了一种普什图语孤立数字自动语音识别系统，用于检测从零到九的普什图语数字。他们使用梅尔频率倒谱系数提取特征，并使用卷积神经网络对数字进行分类。所用数据集包含每个数字的 50 条语音。他们提出的模型包含四个卷积层，后接修正线性单元和最大池化层。随后，他们获得了 84.17% 的平均准确率。Dahl 等人[2] 运用深度神经网络技术，开发了一种用于语音识别的新型上下文相关模型。他们证明，所提出的模型优于以往的上下文相关方法。同样，Abdel-Hamid[1] 在语音识别领域采用卷积神经网络，并证明其能够降低错误率、增强鲁棒性。Graves 等人[5] 提出了一种结合双向长短期记忆网络循环神经网络和权重噪声的混合模型。他们在 TIMIT 音素识别基准上评估了该模型。结果显示，错误率最多降低了 17.7%。

近期发表的少数文章重点研究了使用不同模型进行波斯语自动语音识别。Sameti 等人[12] 实现了一个波斯语连续语音识别系统。他们采用经过一定修改的梅尔频率倒谱系数来学习语音信号特征，并使用基于模型的技术，以及频谱减法和维纳滤波等语音增强方法，以获得理想的鲁棒性。同样，Hasanabadi 等人 [7] 使用了一个包含波斯语孤立词的数据库，并于 2008 年开发了波斯语自动语音识别系统。最后，他们创建了一个轮式移动机器人，使其能够利用波斯语语音命令进行导航。他们研究了用于捕获属性的简单快速傅里叶变换（FFT），以及用于分类模式的多层感知机。S. Malekzadeh 等人使用梅尔频率倒谱系数从波斯语语音中提取特征，并使用多层感知机检测元音和辅音字符。所使用的数据集包含来自 10 个人话语的 20 个声学类别。所提出的模型实现了 61% - 87% 的转换准确率。S. Malekzadeh 采用深度多层感知机模型识别波斯语音，以改善语音信号处理。同样， [9] 建立了一个波斯语字母和词语到语音系统。他们在网络模型的第一层采用基于规则的方法，在其他层采用多层感知机。字母和词语预测分别取得了 60.7% 和 83% 的平均值准确率。最近，Veisi 和 Mani [15]采用深度置信网络（DBN）混合模型提取特征，并使用带有连接时序分类（CTC）输出层的深度双向长短期记忆网络创建声学模型。研究表明，与传统模型相比，深度双向长短期记忆网络在波斯语音素识别中具有更高的准确率。

本文介绍了一种针对达里语孤立词的自动语音识别系统设计。本文所 प्रस्तुत的工作基于三种不同的深度学习技术。

3. Dari Word Corpus

本研究使用了我们此前创建的数据集 [3]。该数据集包含1000段音频，代表20名不同性别、来自不同方言区的达里语母语者说出的20个简短达里语术语。

说话者使用智能手机录音机，在家中和办公室的无噪声环境下录制了这些话语。随后，我们使用USB线将录制的音频传输到个人电脑，并使用Adobe Audition软件去除背景噪声，将每个文件的时长限制为一秒。最后，我们将所有音频文件保存为.wav扩展格式，每个词单独保存为一个以其名称命名的文件。与某个术语相关的所有语音片段文件都存放在同一个文件夹中；例如，所有说话者说出的“ Salaam”一词的音频记录都存放在“第20折”文件夹中。因此，共有20个独立文件夹，每个文件夹都包含与一个术语相关的话语。由于部分文件已损坏，最终我们获得了1000条可用话语。

4. 达里语自动语音识别系统

本研究提出了一种达里语自动语音识别的先进方法。所提出的系统以音频信号为输入，使用梅尔频率倒谱系数从语音信号中提取特征，并利用五种独立的深度学习模型识别话语。此外，我们按不同比例将数据集划分为训练集和测试集，以评估这些架构的性能。每种方法都经过训练和测试，所有模型均能准确检测并输出相应词语。

特征提取是分析数据并识别不同对象（如音频信号）之间关系的关键步骤。模型无法直接识别话语，因此需要将语音信号转换为可理解的格式。此外，必须从原始波形中提取特征，以尽量减少语音信号的变化，并生成适合建模的输入。本研究采用梅尔频率倒谱系数，这是提取关键属性最常用的方法[16]。信号的梅尔频率倒谱系数是一组能够表征频谱包络整体形状的特征。

5. **NEURAL NETWORK** 模型

在我们的会议论文中，我们使用了三种深度神经网络模型。我们的深度卷积神经网络模型由四层组成，每层均包含二维卷积层，后接最大池化层，最后一层除外，该层采用全局平均池化。随后，在最后设置了一个全连接层作为输出层。第一个卷积层的输入是一个维度为 40174 的二维张量（表示梅尔频率倒谱系数特征），该张量与 16 个卷积核进行卷积运算。下一个卷积层与 32 个滤波器进行卷积运算。在后续各层中，我们将卷积核的数量加倍。前一层的输出作为后一层的输入。每个卷积滤波器都会根据输入生成一个特征图。例如，第一层生成 16 个特征图，而最后一层生成 128 个特征图。每个卷积层都使用 2 个滤波器进行卷积运算。我们采用线性激活函数来解决梯度消失问题[8]。在本研究中，我们使用了广为人知的修正线性单元（ReLU）激活函数。在下一步中，我们对特征应用最大池化，通过将特征图划分为较小的矩形区域（通常称为窗口）来减少特征图的采样量，这些区域彼此之间不重叠（Ide与Kurita，2017）。研究表明，与其他池化类型相比，最大池化更加有效（Ide与Kurita，2017）。因此，在前三个模块中，我们使用池化大小为 2 的最大池化，而不是其他池化层。通常，模型的性能会受到训练过程中噪声样本的影响。因此，我们采用 Dropout 来解决这一问题，并防止神经网络出现过拟合 [13]。我们尝试了多种 Dropout 概率值（范围为 0 至 2.5）。结果表明，最有效的概率值为 2.0，该值能够提高准确率并降低损失，这对于缓解每个术语对应的数据集规模有限所带来的影响具有重要作用。输出维度与类别总数相同，即 20 个类别，每个类别对应一个特定的单词。在训练过程中，我们考察了不同训练轮次和批次大小对该模型的影响。因此，当训练轮次为 80、批次大小为 84 时，获得了最优结果。

多层感知机模型由两个隐藏层组成，每个隐藏层后均接修正线性单元（ReLU）激活函数。除输出层外，每一层均包含 300 个神经元。随后，应用 2.0 的 Dropout，以便在每个训练周期中随机排除部分节点，从而降低模型在训练数据上发生过拟合的可能性。该模型的输入形状为一个包含 40 个特征的一维特征向量。输出层包含 20 个节点，每个节点代表一个类别，并且与卷积神经网络（CNN）模型一样，后接 SoftMax 激活函数。选择 SoftMax 作为最后一层激活函数的目的是使输出值之和为 1。最后，该模型采用 32 的批次大小进行训练，参数总数为 108,620。

所实现的循环神经网络模型的结构是一个序列模型。它由一个大小为 64 的长短期记忆网络（LSTM）架构块组成，后接一个丢弃率为 0.2 的 Dropout 层。LSTM 模型使用 Tanh 激活函数激活单元状态，并使用 sigmoid 激活函数生成节点输出。下一层是 Flatten 层，它将输入的空间维度压缩到通道维度中。Flatten 层的输出随后传递到输出层，该层由 20 个节点组成。最后，输出层后接 SoftMax 激活函数，这与多层感知机和卷积神经网络模型类似。使用 100 个训练轮次和 64 的批次大小时获得了最佳性能。我们使用多个不同的训练轮次和批次大小训练每个模型，并选择取得最高准确率的最合适值。

6. H**YBRID** 混合模型

我们实现了两种不同的混合模型，以评估混合深度学习技术在达里语自动语音识别系统中的性能。第一个模型由四个二维卷积隐藏层和一个长短期记忆网络隐藏层组成，此外还包含一个全连接层和一个输出层。第二个混合模型与第一个类似，但我们用双向长短期记忆网络技术替换了长短期记忆网络，并在将其输出馈入输出层之前进行了展平。每个隐藏层后均接修正线性单元激活函数。此外，仿照深度卷积神经网络结构，每个二维卷积层后都接一个二维最大池化层。两个模型中长短期记忆网络的容量均为 64。我们手动微调了层数、Dropout 率、训练轮次和批次大小，以达到最佳值。因此，我们在最大池化层后采用了 20% 的 Dropout。最终，两个模型分别以批次大小 64 训练了 100 个训练轮次，以获得最佳结果。

7. RESULTS 和 EXPERIMENTS

我们开发了一个用于达里语的自动语音识别（ASR）系统，以识别20个孤立词。为了构建该系统，我们使用了此前研究中的一个数据集，其中每个词包含50条话语。我们采用梅尔频率倒谱系数（MFCC）进行特征提取，并使用多层感知机（MLP）、卷积神经网络（CNN）、长短期记忆网络（LSTM）、CNN+LSTM和CNN+BLSTM进行语音分类。模型采用10折交叉验证进行训练和评估，以分类交叉熵作为损失函数，以准确率作为评估指标，并使用Adam优化器。该系统使用Python 3.6.9在搭载Windows 10操作系统的计算机上实现，计算机配备Intel Core i7(TM) 2.80 GHz 2.90 GHz处理器。特征提取使用Librosa 0.8.0版本完成，而模型的实现则使用TensorFlow 2.3.1版本和Keras 2.4.3版本。

表 1 展示了每个模型在训练和测试期间的平均值准确率和损失。CNN+BLSTM 模型取得了最高的平均值测试准确率 98.365%，其次是 CNN+LSTM 和 CNN，准确率分别为 98.184% 和 98.037%。尽管将 CNN 与 LSTM 或 BLSTM 结合对准确率和损失的影响很小，但这增加了该方法的复杂度和运行时间。因此，对于准确率或损失维度的细微改进，我们建议使用 CNN+BLSTM 方法。尽管 LSTM 具有非线性特性，适用于许多 NLP 应用，但由于数据集规模较小且结构简单，其平均值测试准确率与 MLP 相当。然而，在所有模型中，MLP 的平均值测试损失值最高。

表 2-6 展示了每个模型 10 个混淆矩阵的总和。LSTM 模型的平均训练损失最低，训练准确率最高，但平均测试准确率较低。CNN 在第 8 折中达到了 99.63% 的最高准确率，但在第 6 折中降至 95.63%。CNN+BLSTM 模型在第 4、6 和 7 折中达到了 99.27% 的最高准确率，并在第 9 折中达到 97.45% 的最低准确率。

训练和测试一折的多层感知机模型大约需要 11 秒，而训练和测试长短期记忆网络架构 100 次迭代大约需要 2 分 32 秒。由于卷积神经网络模型的复杂性，其耗时要长得多（80 个周期大约需要 14 分 12 秒）。混合模型 CNN+LSTM 和 CNN+双向长短期记忆网络的执行时间更长，分别在 100 个训练轮次中花费了约 19 分钟和 15 分 45 秒用于训练和测试模型。

表1. 训练和测试的准确率与损失

模型	平均值 训练 准确率	平均值 训练 **Loss**	平均值 测试 准确率	平均值 测试损失
MLP	99.73	0.028	97.02	0.129
CNN	99.93	0.018	98.037	0.105
LSTM	100	0	97.712	0.09
CNN+ LSTM	99.95	0.006	98.184	0.079
CNN+ 双向长短期记忆网络	99.98	0.006	98.365	0.07

在本研究中，我们旨在通过调整各种参数（如全连接层、池化层、Dropout、测试与训练迭代次数、批次大小以及卷积核大小），确定性能最佳的模型。根据获得的结果，我们为本研究选择了最优方案。在评估的五种模型中，CNN+BLSTM 模型表现最为出色，其训练阶段的平均准确率达到 99.98%，测试阶段达到 98.365%，甚至优于 LSTM 模型。尽管 LSTM 模型在训练阶段的平均准确率更高（100%），但其测试阶段的准确率较低，如表 1 所示。

我们的研究在达里语孤立词语音识别方面优于以往研究。例如，Dawodi 等人 [3] 采用梅尔频率倒谱系数和卷积神经网络开发了一个用于孤立词的达里语自动语音识别系统，但我们的模型在测试数据上的准确率高出10%以上。此外，与其他近期的自动语音识别研究相比，我们的方法还表现出更低的评估损失和更高的准确率。表7对比了我们的研究与一些最新且创新的自动语音识别系统。

8. 结论

本文是我们会议论文的扩展，研究了五种深度神经网络模型在识别达里语孤立词方面的有效性，即多层感知机、卷积神经网络、循环神经网络、CNN+长短期记忆网络和CNN_双向长短期记忆网络。这些模型通过梅尔频率倒谱系数提取音频信号特征进行识别。结果表明，卷积神经网络和混合卷积神经网络模型在达里语自动语音识别方面优于其他模型。尽管本研究为达里语自然语言处理提供了良好的起点，但仍需进一步研究，运用混合模型和新技术开发更准确、连续的达里语自动语音识别模型。然而，这需要更大且更多样化的数据集。

表 2. 多层感知机模型的混淆矩阵

多层感知机：	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	11 12	3	**14**	**15**	16 17		**18**	**19**	**20**
1	169 0.		0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
2	0	93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	5	0	220 0		0	0	1	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	135 0		0	2	0	0	0	0	0	0	0	0	3	0	0
5	0	0	0	0	131 0		0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	135 0		0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	227 0		0	0	0	0	0	0	0	0	1	0
8	0	0	2	0	0	0	0	108 0		0	0	2	4	0	0	2	0	0
9	0	0	0	0	0	0	0	0	132 0		0	0	0	0	0	0	0	0
10	0	2	0	0	0	2	0	1	0	120 1		0	3	0	0	5	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	88	0	2	0	0	0	0	0	2	0
12	4	0	0	0	0	0	0	0	0	0	0	84	0	2	0	0	0	0	0	0
13	0	0	0	0	0	0	1	2	0	3	2	0	162 0		0	0	0	0	0	2

14	0	0	2	0	0	0	0	0	0	0	0	0	0	112 0		0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	152 0		0	0	0	0
16	0	0	0	0	0	2	3	0	0	1	2	0	1	0	0	88	0	1	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140 0		0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	136 0		0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	128 0	
20	2	0	0	0	0	0	1	0	0	2	0	0	1	0	0	0	0	0	1	111

表 3. 混淆矩阵长短期记忆网络模型

长短期记忆网	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	174 0		0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	90	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	218 2		0	0	0	0	0	0	0	0	0	4	2	0	0	0	0	0
4	0	0	0	136 0		0	2	0	0	0	0	0	0	0	0	0	2	0	0	0
5	0	0	0	0	131 0		0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	135 0		0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	228 0		0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	114 0		0	0	0	2	2	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	132 0		0	0	0	0	0	0	0	0	0	0
10	2	0	0	0	0	0	0	2	0	122 0		0	2	2	0	0	2	0	0	2
11	0	0	0	0	0	0	0	0	0	0	92	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	90	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	2	2	0	0	168 0		0	0	0	0	0	0
14	0	2	0	0	0	0	0	2	0	0	0	0	2	108 0		0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	152 0		0	0	0	0
16	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	94	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140 0		0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	136 0		0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	128 0	
20	0	0	4	0	0	0	2	0	0	4	2	0	2	0	2	0	0	0	0	102

表4。混淆矩阵卷积神经网络模型

卷积神经网	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	171	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	1	0	0	0
2	0	93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	213 2		0	0	2	0	0	2	0	3	2	0	0	0	0	0	1	1
4	0	0	0	140	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	131 0		0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	135	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	2	0	226 0		0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	114	0	2	0	0	0	0	0	0	0	0	0	2
9	0	0	0	0	0	0	0	0	132 0		0	0	0	0	0	0	0	0	0	0
10	0	1	2	0	0	0	3	0	0	128 0		0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	92	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	2	0	88	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	172 0		0	0	0	0	0	0
14	0	0	0	0	0	0	4	0	0	0	0	0	4	106 0		0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	152 0		0	0	0	0
16	0	0	0	0	0	0	0	0	0	2	0	2	0	0	0	94	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140 0		0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	136 0		0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	128 0	
20	0	0	0	0	0	0	6	0	0	2	0	0	0	2	0	0	0	0	0	108

表5. 混淆矩阵卷积神经网络+双向长短期记忆网络模型

卷积神经网络+B	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	174	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
2	0	93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	2	2	218	2	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
4	0	0	0	140	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	131	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	3	0	0	0	0	132	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	228	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	116	2	0	0	0	0	0	0	0	0	0	0	0

9	0	0	0	0	0	0	0	0	132	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	130	0	0	0	0	0	2	0	0	0	2
11	0	0	0	0	0	0	0	0	0	0	92	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	88	0	0	0	2	0	0	0	0
13	0	0	0	0	0	0	0	2	0	0	0	0	170	0	0	0	0	0	0	0
14	0	2	0	0	0	0	0	4	0	0	0	0	0	108	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	152	0	0	0	0	0
16	0	0	0	0	0	0	2	0	0	2	0	0	0	0	0	92	0	0	0	2
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	136	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	128	0
20	2	0	2	0	0	0	0	2	0	0	0	0	0	0	0	2	0	2	0	108

表6。混淆矩阵 CNN+双向长短期记忆网络模型

卷积神经网络+	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	172 0		2	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
2	0	93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	2	0	217 0		0	0	0	2	0	0	0	1	0	0	2	0	0	0	0	2
4	0	0	0	138 1		0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
5	0	0	0	0	131 0		0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	135 0		0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	228 0		0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	114 0		0	0	0	2	0	0	0	2	0	0	0
9	0	0	0	0	0	0	0	0	132	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	2	0	125 0		2	0	2	0	2	0	0	1	0
11	0	0	0	0	0	0	0	0	0	0	92	0	0	0	0	0	0	0	0	0
12	0	0	2	0	0	0	0	0	0	0	0	86	0	0	0	0	2	0	0	0
13	0	0	2	0	0	0	2	0	0	0	0	0	166 2		0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	114 0		0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	152 0		0	0	0	0
16	0	0	0	0	0	0	0	0	0	2	0	0	0	2	0	92	0	2	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	140 0		0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	136 0		0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	128 0	
20	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	112

表7. 当前研究与其他最先进方法的比较

论文	技术 语言 数据集 识别率			
[17]	梅尔频率倒谱系数 +卷积神经网络	普什图语	普什图语数字 数据库	84.17%
[19]	前馈 TDNN	英语	Tedlium2	词错误率（WER）为 7.6 语音识别 语音情感预测的准确率为 71.7% 来自语音
[16]	梅尔频率倒谱系数 +人工神经网络	阿拉伯语	发音 3 个阿拉伯语 字母	92.42%
[6]	Deep 去噪 自编码器 + 逆 滤波	塞尔维亚 语	Whi-Spe	92.81%
[18]	带 注意力	英语	Google 语音 命令	91.4% （20条命令识别 任务)
[14]	梅尔频率倒谱系数 + 线性 预测 编码（LPC）	卡纳达语	卡纳达语 元音数据集	40%
[11]	CNN	英语	Switchboard-1	97.0%
[10]	连接主义	英语	TIMIT	核心测试集上的音素错误率为 18.2%

	时间性的 分类 （连接时序分类）+ 卷积神经网络			set
[8]	混合式 卷积的 神经网络 （混合卷积 神经网络）	英语	英语 孤立词 语音语料库 以及 TVs	-
[15]	梅尔频率倒谱系数+ 深度 信念网络 （DBN）长短期记忆网络， 双向长短期记忆网络， 深度长短期记忆网络， 深度双向长短 期记忆网络	波斯语 以及	Farsdat	HMM：75.2%，长短期记忆网络：77% LSTM-DBN：78%，双向长短期记忆网络：79.3% Karel-DNN：80.2%，DLSTM：80.3% 深度双向长短期记忆网络：82.9%，深度双向长短期记忆网络-深度神经网络：83.2%
本研究	梅尔频率倒谱系数，卷积神经网络， 长短期记忆网络、多层感知机， 卷积神经网络+双向长短期记忆网络， 卷积神经网络+长 短期记忆网络	Dari	我们内置的 Dari 语音 数据集	平均值准确率：卷积神经网络：98.037 %， 长短期记忆网络：97.712 %，多层感知机： 97.02% 卷积神经网络+长短期记忆网络：98.184 %，卷积神经网络+双向长短期记忆网络： 98.365%

REFERENCES

[1] O. Abdel-Hamid、A. Mohamed、H. Jiang、L. Deng、G. Penn 和 D. Yu，《用于语音识别的卷积神经网络》，IEEE/ACM音频、语音与语言处理汇刊，第22卷，第10期，第1533-1545页，2014年10月，doi：10.1109/TASLP.2014.2339736。

[2] G. E. Dahl、Dong Yu、Li Deng 和 A. Acero，“用于大词汇量语音识别的上下文相关的预训练深度神经网络”，IEEE 音频、语音与语言处理汇刊，第 20 卷，第 1 期，第 30–42 页，2012 年 1 月，doi：10.1109/TASL.2011.2134090。

[3] M. Dawodi、J. A. Baktash、T. Wada、N. Alam 和 M. Z. Joya，《使用深度卷积神经网络进行达里语语音分类》，2020年IEEE国际物联网、电子与机电一体化会议（IEMTRONICS），第1–4页，2020年12月，DOI：10.1109/IEMTRONICS51293.2020.9216370。

[4] M. Dua、R. K. Aggarwal、V. Kadyan 和 S. Dua，《使用 HTK 的旁遮普语自动语音识别》，《国际语音技术期刊》，第 9 卷，第 4 期，第 6-12 页，2012 年 12 月，doi：10.1007/s10772-012-9171-1。

[5] A. Graves、A. Mohamed 和 G. Hinton，“基于深度循环神经网络的语音识别”，2013年IEEE国际声学、语音与信号处理会议，第6645–6649页，2013年5月，doi: 10.1109/ICASSP.2013.6638947。

[6] D. T. Grozdic 和 S. T. Jovicic，“使用深度去噪自编码器和逆滤波进行耳语语音识别”，IEEE/ACM音频、语音与语言处理汇刊，第 25 卷，第 12 期，第 2313–2322 页，2017 年 12 月，doi: 10.1109/TASLP.2017.2738559。

[7] H. Hasanabadi、A. Rowhanimanesh、H. T. Yazdi 和 N. Sharif，“一种简单且稳健的波斯语语音识别系统及其在机器人技术中的应用”，2008年高级计算机理论与工程国际会议，第239-245页，2008年10月，doi：10.1109/ICACTE.2008.125。

[8] V. Mitra、G. Sivaraman、H. Nam、C. Espy-Wilson、E. Saltzman 和 M. Tiede，《用于基于发音和声学信息的语音识别的混合卷积神经网络》，《语音通信》，第 89 卷，第 103–112 页，2017 年。DOI：10.1016/j.specom.2017.03.003。

[9] M. Namnabat 和 M. Homayounpour，“使用神经网络实现波斯语声音系统的一封信”，载于 2006 年第 8 届国际信号处理会议，2006 年，第 4128933 页。DOI：10.1109/ICOSP.2006.345518。

[10] D. S. Park、W. Chan、Y. Zhang、C.-C. Chiu、B. Zoph、E. D. Cubuk 和 Q. V. Le，《SpecAugment：一种用于自动语音识别的简单数据增强方法》，发表于 Interspeech 2019，2019 年，第 2613-2617 页。DOI：10.21437/Interspeech.2019-2680。

[11] R. Price、K. Iso 和 K. Shinoda，《明智的教师能够训练出更好的深度神经网络声学模型》，《EURASIP 音频、语音与音乐处理期刊》，第 2016 卷，第 1 期，第 10 页，2016 年。DOI：10.1186/s13636- 016-0088-7。

[12] H. Sameti、H. Veisi、M. Bahrani、B. Babaali 和 K. Hosseinzadeh，《Nevisa：一种波斯语连续语音识别系统》，载于《计算机科学与工程进展》，H. Sarbazi-Azad、B. Parhami、S.-G. Miremadi 和 S. Hessabi 编。Springer，2009年，第485–492页。doi：10.1007/978-3-540- 89985-3_60。

[13] N. Srivastava、G. Hinton、A. Krizhevsky、I. Sutskever 和 R. Salakhutdinov，《Dropout：一种防止神经网络过拟合的简单方法》，《机器学习研究期刊》，第 15 卷第 56 期，第 1929–1958 页，2014 年。

[14] A. H. Unnibhavi 和 D. S. Jangamshetti，《基于线性预测编码的卡纳达语元音语音识别》，2017年电气、电子、通信、计算机与优化技术国际会议（ICEECCOT），第1-4页，2017年。doi：10.1109/ICEECCOT.2017.8284582。

[15] H. Veisi 和 A. Haji Mani，《使用深度学习进行波斯语语音识别》，《国际语音技术期刊》，2020年。doi：10.1007/s10772-020-09768-x。

[16] E. S. Wahyuni，《使用梅尔频率倒谱系数特征提取和人工神经网络分类进行阿拉伯语语音识别》，2017年第2届信息技术、信息系统与电气工程国际会议（ICITISEE），第22–25页，2017年。doi：10.1109/ICITISEE.2017.8285499。

[17] B. Zada 和 R. Ullah，《使用深度卷积神经网络进行普什图语孤立数字识别》，Heliyon，第6卷，第2期，第e03372页，2020年。doi：10.1016/j.heliyon.2020.e03372。

[18] D. C. de Andrade、S. Leo、M. L. D. S. Viana 和 C. Bernkopf，《用于语音命令识别的神经注意力模型》，arXiv:1808.08929 [cs，eess]，2018年8月，访问日期：2020年11月22日。[在线]。可访问：http://arxiv.org/abs/1808.08929

[19] S. Zhou 和 H. Beigi，《一种从自动语音识别进行语音情感识别的迁移学习方法》，arXiv:2008.02863 [cs, eess]，2020年8月，访问日期：2020年11月22日。[在线]。可访问：http://arxiv.org/abs/2008.02863