附件

第十八届“挑战杯”全国大学生课外学术科技作品竞赛

作品申报书

**作品名称： 基于深度学习的手语视频与语音互译系统**

**学院全称： 软件学院**

**申报者姓名**

**（集体名称）： 邹璟辰**

类别：

▉自然科学类学术论文

□哲学社会科学类社会调查报告和学术论文

□科技发明制作A类

□科技发明制作B类

说 明

1．申报者应在认真阅读此说明各项内容后按要求详细填写，表格各项内容不得为空。

2．申报者在填写申报作品情况时，需根据个人项目或集体项目情况填写A1或A2表，根据作品类别（自然科学类学术论文、哲学社会科学类社会调查报告和学术论文、科技发明制作）分别填写B1、B2或B3表（作品类别必须细化到最末一级类别标题）。所有申报者可根据情况填写C表。

3．表内项目填写时一律用钢笔或打印，字迹要端正、清楚，此申报书可复制。

4．学术论文、社会调查报告及所附的有关材料必须是中文（若是外文，请附中文本），请以4号楷体打印在A4纸上，附于申报书后，学术论文及有关材料在8000字以内，社会调查报告在15000字以内（文章版面尺寸14.5×22cm）。

A1．申报者情况（个人项目）

说明：

1.必须由申报者本人按要求填写，申报者情况栏内必须填写；

2.个人作品的第一作者（承担申报作品60%以上的工作者）；

3.本表中的学籍管理部门签章视为对申报者情况的确认。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 申报者情况 | 姓 名 | 邹璟辰 | | | | 性别 | | | 男 | | 出生年月 | | | 2001年4月 | |
| 学校全称 | 软件学院 | | | | | | | 现学历 | | 本科 | | | | |
| 专 业 | 物联网工程 | 年级 | | 大四 | | | 学制 | | | 四年 | 入学时间 | | | 2019年9月 |
| 作品全称 | 基于深度学习的手语视频与语音互译系统 | | | | | | | | | | | | | |
| 通讯地址 | 江西省南昌航空大学前湖校区 | | | | | | | | | 邮政编码 | | 330103 | | |
| 单位电话 | |  | | |
| 合作者情况 | 姓 名 | 性别 | | 年龄 | | | 学历 | | | 所在单位 | | | | | |
|  |  | |  | | |  | | |  | | | | | |
|  |  | |  | | |  | | |  | | | | | |
| 资 格 认定 | 学校学籍  管理部门  意 见 | 以上作者是否为2022年6月1日前正式注册在校的全日制非成人教育、非在职的本科高校学生（含专科生、本科生和硕士研究生）。  ▉是 □否  若是，其学号为： 19207333    2022年 11 月 29 日 | | | | | | | | | | | | | |
| 学校组织协调机构意 见 | 本作品是否为课外学术科技或社会实践活动成果  □是 □否  负责人签名： （学校团委代章）  　　　 年 月 日 | | | | | | | | | | | | | |

B1．申报作品情况（自然科学类学术论文）

说明：

1．作品分类请按作品的学术方向或所涉及的主要学科领域填写；

2．硕士研究生作品不在此列。

|  |  |
| --- | --- |
| 作品全称 |  |
| 作  品  分  类  （细化到最末一级类别选项，仅能填报1项） | 请参考以下选项，填写类别： B1 （如A1）  **机械与控制**（A1：机械，A2：仪器仪表，A3：自动化控制，A4：工程，A5：交通，A6：建筑）  **信息技术**（B1：计算机，B2：电信，B3：通讯，B4：电子）  **数理**（C1：数学，C2：物理，C3：地球与空间科学）  **生命科学**（D1：生物，D2：农学，D3：药学，D4：医学，D5：健康，D6：卫生，D7：食品）  **能源化工**（E1能源，E2材料，E3石油，E4化学化工，E5生态环保） |
| 作品撰写的  目的和基本  思路 | 手语作为听力障碍患者进行日常交流的主要方式，通过手和手臂的特定运动、嘴唇变化和面部表情等的组合来传达相应的语义信息。现今，我国手语体系完备，各项手语规范标准已经制定完毕，为手语的学习带来了便捷。然而，因手语学习较为耗时，且学习成本较高，大多数健听人士若不经过系统的手语训练，仍然无法与听障患者进行正常的交流，这给听障患者带来了许多不便。为了消除听障患者与正常人之间的交流障碍，尝试使用深度学习等技术对手语视频与文字或语音进行双向翻译。 |
| 作品的科学性、先进性及独特之处 | 作品科学性在于，论文数据全部来源于官方渠道或者根据真实实验结果所得出，并且论文参考了大量的国内外前沿文献质料，论文中所提手语翻译模型均可实现运行。  作品先进性在于，手语和语音互译系统在国外研究较少，且现有模型对于翻译的效果不佳。于此相反，本文所提出并实现的手语到语音互译系统不仅实现了手语到语音的翻译，同时还能够实现语音到手语的翻译，且性能效果不错。  作品独特性在于，提出并设计一个以ResNet18为编码器，以LSTM为解码器的神经网络模型将手语视频转换为文本数据，并利用一个端到端的语音合成框架Tacotron2，将上一步的得到的文字数据转换为语音信号，最后提出了一种语音到手语的翻译方法，首先使用一个端到端的模型将语音转换为手语编码序列，其后将手语编码序列与手语基础元动画进行匹配，再利用视频合成技术生成完整的手语视频动画。 |
| 作品的实际应用价值和现实意义 | 1、据世界卫生组织统计结果显示，全球约有4.3亿人患有残疾性听力损失，占全球人口5%以上，而根据第二次全国残疾人抽样调查结果显示，我国是世界上听力残疾人数最多的国家，听力残疾人数已达2780万，占全国残疾人数的30%以上。听力障碍患者与健听人士之间的沟通障碍，使其难以融入社会工作中，给他们在日常生活中带来了很大的不便，而手语视频与语音互译系统的实现将给听障患者融入社会带来希望。  2、党的十八大以来，以习近平同志为核心的党中央对残疾人这一特殊群体高度重视，并提出改善残疾人的生活品质，为残疾人谋福祉是推动当前残疾人事业高质量发展的重点所在，手语视频与语音互译系统的实现符合党的政策方针，提高了听障患者的沟通质量和生活品质。 |
| 学  术  论  文  文  摘 | 手语是聋哑人与健听人士进行交流的重要手段，手语主要通过手势的特定运动和面部表情等的组合来表达语义信息。然而，手语学习较为困难且学习成本较高，导致大多数健听人士对于手语的掌握程度不高，仅仅处于了解层面，无法与聋哑人进行正常的交流，聋哑人与正常人之间仍然存在沟通障碍。近几年来，随着深度学习在手语识别领域的应用，具备低时延、高可靠性和高准确性的手语与语音互译系统有望实现。  本文主要针对手语与语音互译系统的实现进行研究。首先，提出了一种基于残差神经网络和长短期记忆网络构成的连续手语识别模型，实现手语视频到文字的转换。然后，利用一个端到端的语音合成神经网络模型Tacotron2实现文字到语音的转换。最后，提出一种从语音序列到手语编码序列的端到端模型，再将手语编码序列与手语基准动作视频进行匹配生成完整的手语动画视频。 |
| 作品在何时、何地、何种机构举行的会议或报刊上发表及所获奖励 |  |
| 鉴定结果 |  |
| 请提供对于理解、审查、评价所申报作品具有参考价值的现有技术及技术文献的检索目录 | [1] 刘远,张万洪.习近平法治思想中的特定群体权利保障理论[J].求是学刊,2022,49(05):18-27.  [2] World Health Organization. Deafness and hearing loss[EB/OL].(2021-04-01) [2022-11-19].https://www.who.int/zh/news-room/fact-sheets/detail/deafness-and-hearing-loss.  [3] 央视新闻. 中国是世界上听力残疾人数最多的国家[EB/OL].(2020-09-27)[2022-11-19]. https://www.sohu.com/a/421233308\_701427.  [4] 陶唐飞,刘天宇.基于手语表达内容与表达特征的手语识别技术综述[J/OL].电子与信息学报:1-19[2022-11-28].  [5] He K，Zhang X，Ren S，et al. Deep residual learning for image recognition[C]//IEEE. 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). New York：IEEE，2016：7780459.  [6] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.  [7] SHEN J,PANG R,WEISS R,et al.Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[C]//International Conference on Acoustics,Speech,and Signal processing,2018:4779-4783.  [8] ZHOU H, ZHOU W, LI H. Dynamic Pseudo Label Decoding for Continuous Sign Language Recognition[C]// 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai: IEEE, 2019:1282-1287.  [9] 杨淑莹,田迪,郭杨杨,赵敏.端到端手语翻译研究[J/OL].系统科学学报:1-7[2022-11-28].  [10] GUO D, ZHOU W, LI H, et, al. Hierarchical LSTM for sign language translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans: AAAI, 2018: 6845-6852.  [11] HUANG J，ZHOU W，ZHANG Q，et al.Video-based sign language recognition without temporal segmenta tion[C]//Proceedings of the AAAI Conference on Artifi cial Intelligence，New Orleans，USA，2018：2257-2264.  [12] KAN Jichao, HU Kun, HAGENBUCHNER M, et al. Sign language translation with hierarchical spatio-temporal graph neural network[C]. Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, USA, 2022: 2131–2140. doi: 10.1109/ WACV51458.2022.00219.  [13] LI H，GAO L，HAN R，et al.Key action and joint CTC Attention based sign language recognition[C]//ICASSP 2020—2020 IEEE International Conference on Acoustics， Speech and Signal Processing（ICASSP），Barcelona，Spain， 2020：2348-2352.  [14] PU J，ZHOU W，LI H.Iterative alignment network for continuous sign language recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recogni tion（CVPR），Long Beach，USA，2019.  [15] GAO Liqing, LI Haibo, LIU Zhijian, et al. RNN Transducer based Chinese sign language recognition[J]. Neurocomputing, 2021, 434: 45–54. doi: 10.1016/j.neucom.2020.12.006. |
| 申报材料清单（申报论文一篇，相关资料名称及数量） | 申报论文《基于深度学习的手语视频与语音互译系统》一篇 |

D. 参赛作品打印处

**基于深度学习的手语视频与语音互译系统**

邹璟辰

(南昌航空大学，南昌 330063)

**摘 要：**手语是聋哑人与健听人士进行交流的重要手段，手语主要通过手势的特定运动和面部表情等的组合来表达语义信息。然而，手语学习较为困难且学习成本较高，导致大多数健听人士对于手语的掌握程度不高，仅仅处于了解层面，无法与聋哑人进行正常的交流，聋哑人与正常人之间仍然存在沟通障碍。近几年来，随着深度学习在手语识别领域的应用，具备低时延、高可靠性和高准确性的手语与语音互译系统有望实现。本文主要针对手语与语音互译系统的实现进行研究。首先，提出了一种基于残差神经网络和长短期记忆网络构成的连续手语识别模型，实现手语视频到文字的转换。然后，利用一个端到端的语音合成神经网络模型Tacotron2实现文字到语音的转换。最后，提出一种从语音序列到手语编码序列的端到端模型，再将手语编码序列与手语基准动作视频进行匹配生成完整的手语动画视频。

**关键词：**手语；手语与语音互译；深度学习

**Video and Speech Translation of Sign Language Based on Deep Learning**

Zou Jingchen

(Nanchang Hangkong University, Nanchang 330063)

**Abstract:** Sign language is an important means of communication between deaf mute and normal people. Sign language mainly expresses semantic information through specific movements of gestures and facial expressions. However, sign language learning is difficult and the cost of learning is high, which leads to the fact that most hearing people do not have a good command of sign language. They are only at the level of understanding and cannot communicate with the deaf mutes normally. There is still a communication barrier between the deaf mutes and the normal people. In recent years, with the application of deep learning in sign language recognition, a sign language and speech translation system with low delay, high reliability and high accuracy is expected to be realized.This paper focuses on the implementation of sign language and speech translation system. Firstly, a continuous sign language recognition model based on 3D convolutional neural network and short-term memory network is proposed to realize the conversion of sign language video to text. Then, an end-to-end speech synthesis neural network model, Tacotron, is used to achieve text to speech conversion. Finally, an end-to-end model from speech sequence to sign language coding sequence is proposed, and then the sign language coding sequence is matched with the sign language benchmark action video to generate a complete sign language animation video.

**Key words:** Sign language; Sign language and speech translation; Deep learning

# 1 引言

自新中国成立以来，我国始终重视残疾人的权益保障问题，尤其是党的十八大以来，以习近平同志为核心的党中央对残疾人这一特殊群体高度重视，并提出改善残疾人的生活品质，为残疾人谋福祉是推动当前残疾人事业高质量发展的重点所在[1]。据世界卫生组织统计结果显示[2]，全球约有4.3亿人患有残疾性听力损失，占全球人口5%以上，而根据第二次全国残疾人抽样调查结果显示，我国是世界上听力残疾人数最多的国家，听力残疾人数已达2780万，占全国残疾人数的30%以上[3]。听力障碍患者与健听人士之间的沟通障碍，使其难以融入社会工作中，给他们在日常生活中带来了极大的不便。手语作为听力障碍患者进行日常交流的主要方式，通过手和手臂的特定运动、嘴唇变化和面部表情等的组合来传达相应的语义信息。现今，我国手语体系完备，各项手语规范标准已经制定完毕，为手语的学习带来了便捷。然而，因手语学习较为耗时，且学习成本较高，大多数健听人士若不经过系统的手语训练，仍然无法与听障患者进行正常的交流，这给听障患者带来了许多不便。随着科学技术的不断发展，尤其是近年来人工智能等技术在许多领域取得了令人惊叹的成绩，越来越多的科研人员尝试利用深度学习等技术对手语视频与文字或语音进行双向翻译，以此消除听障患者与正常人之间的交流障碍。

手语视频翻译的核心为手语识别(Sign Langugae Recognization, SLR)技术，手语识别通过深度学习、计算机视觉等技术将手语视频转换为文字信息。根据手语数据的采集方式不同手语识别技术可以分为以下两种形式[4]：（1）基于穿戴式的手语识别，利用安装于手部的位置、加速度等各类传感器或通过穿戴数据手套进行手语数据的采集，使用该方法的SLR具有较高的识别准确性和较低的时延。但是，由于传感器等设备穿戴不便且价格较为昂贵，使其很难满足听障患者的正常需求。（2）基于计算机视觉的手语识别，利用摄像头等设备录制手语视频，后通过神经网络模型将手语视频直接转换为文字或语音，与基于穿戴式的SLR相比，该方法所需设备较少且不需各式传感器，在大规模手语数据集上的识别准确率高、泛化能力强。如今，大多数研究人员的关注重点在于手语到语音的单向翻译而对语音到手语方向的翻译研究较少。为了能够满足听障患者日常基本的交流需求，本文提出了一种手语视频与语音互译系统的实现，其主要研究和工作内容可以概述为以下三点：

1. 使用帧间差分算法提取手语视频中的关键帧图像，以去除视频冗余信息、降低模型计算复杂度，而后将得到的关键帧图像按时序拼接成一个四维张量传入一个以ResNet18为编码器，以LSTM为解码器的神经网络模型中，得到该手语视频所表达的文字信息。
2. 利用一个端到端的语音合成框架Tacotron2，将上一步的得到的文字数据转换为语音信号。
3. 提出一种语音到手语的翻译方法，首先使用一个端到端的模型将语音转换为手语编码序列，其后将手语编码序列与手语基础元动画进行匹配，再利用视频合成技术生成完整的手语视频动画。

# 2 网络模型分析

**2.1 残差神经网络**

残差神经网络（ResNet）是由我国青年学者何恺明[5]等人于2015年末首次提出，旨在解决当神经网络层数过深时而出现的训练困难问题。在ResNet基础架构中残差块（Residual Block）是其中最重要的组成部分，残差块内部结构如图1所示。

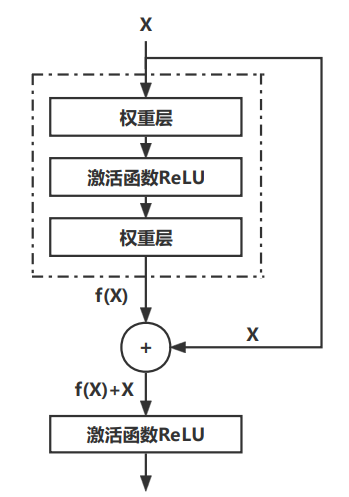


图1残差块内部结构图

通过在残差块中使用跳跃连接，使ResNet具有恒等映射功能。大量实验结果表明，ResNet在深层神经网络中能够十分有效的解决模型退化问题，大幅提高模型的准确性。在本次实验中使用ResNet18作为编码器的组成部分之一，用于对关键帧图像进行特征提取，ResNet18的结构图如图2所示。

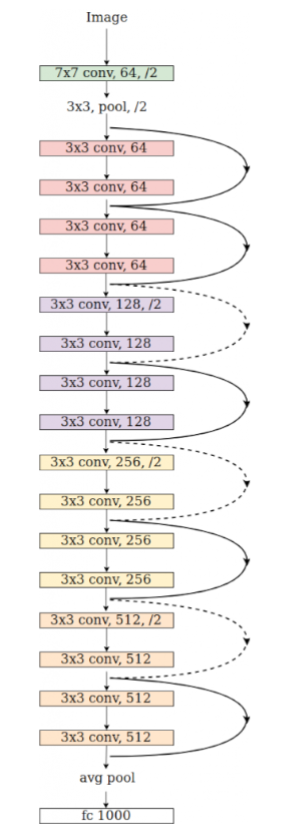


图2 ResNet18结构图

**2.2 长短期记忆网络**

长短期记忆（Long-Short-Term-Memory, LSTM）网络[6]是循环神经网络中的一种特殊网络模型，LSTM能够有效的解决神经网络在训练过程中出现的梯度消失问题和梯度爆炸问题，同时可以高效的传递和保存长时间序列数据中的有效信息，而不会使这些有效信息遗忘。LSTM内部结构如图3所示，主要由遗忘门、输入门和输出门组成，其中遗忘门可以根据当前阶段的输入和上一阶段的输出来选择性遗忘先前存储的部分信息，遗忘门使LSTM具有长期记忆历史信息的能力，且记忆能够随着输入进行动态调整；输入门用于从输入数据中提出有效信息，并对有效信息进行评分，评分越高就会有越多的记忆进入单元状态；输出门用于计算当前时刻的输出值，首先将输入值与上一个时刻的输出值传入sigmoid函数中，之后更新单元状态并通过tanh函数，最后将输入的两个向量按元素进行乘积，得到下一个时刻的新隐藏状态。在本次实验中使用LSTM作为解码器，用于提取和保存关键帧图像中的时序信息。

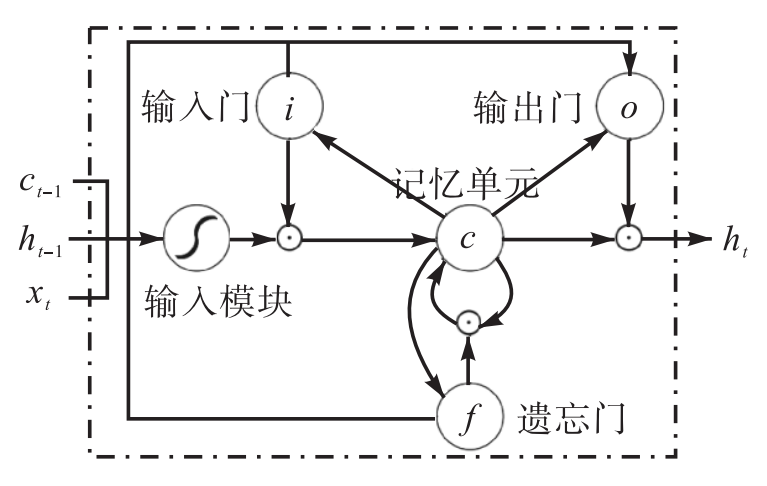


图3 LSTM内部结构

**2.3 语音合成模型**

语音合成（Text To Speech，TTS）可以将文本内容转换为接近于人声的语音信号。现今许多科研人员在TTS上使用深度学习技术并取得了令人满意的效果，其中2017年由Google Brain提出的Tacotron2[7]模型是一个端到端的条件自回归模型，该模型性能优越，模型的平均主观评分（Mean Opinion Score, MOS）达到4.526分，MOS是评价语音质量最重要的标准之一，MOS最高分为5分，当MOS的值大于等于4时，语音质量被评定为优，而当MOS低于3.6分时，语音质量较差。

Tacotron2模型结构如图4所示，其主要由声谱预测网络和声编码器两部分组成，将文本数据先转换为频谱图，再将其转换为波形并输出。利用Tacotron2语音合成模型，可以将手语识别模型生成的文字数据转换为语音数据，给用户带来更多的选择和更好的体验。

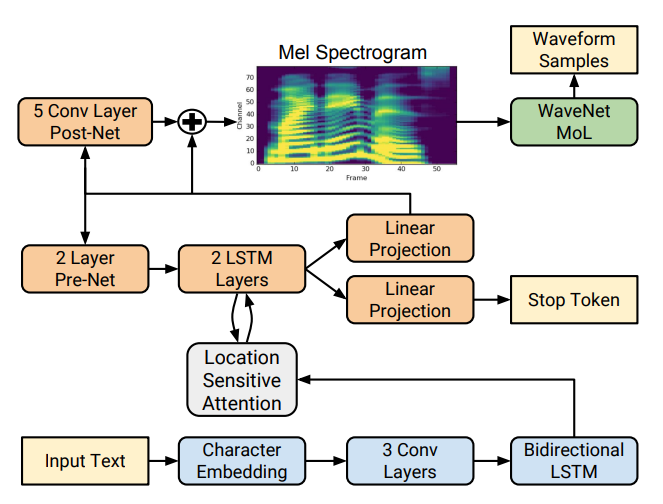


图4 Tacotron2结构图

# 3 手语识别数据处理和模型设计

**3.1 手语识别评价指标**

手语识别效果的好坏需由手语识别评价指标来衡量，手语识别领域常用的评价指标有错词率、准确率、平均均度均值、杰卡德系数、精度等，而使用最为广泛的是错词率和准确率。

（1）错词率（Word Error Rate, WER）用来评估生成文本G与真实文本R间的词错误率，该指标越低表明生成的文本序列越接近于真实的文本序列。错词率可根据(1)式计算得出：

(1)

其中，是真实文本R的序列长度，、和分别为生成文本转换成真实文本所需要进行插入、替换和删除某些词的次数。

（2）准确率（Accuracy, ACC）用来评估生成文本G与真实文本R完全匹配的样本数占总样本数的比例，准确率越高，模型识别的性能就越好。准确率可以根据(2)式计算得出：

(2)

其中，为总样本数，为总样本中生成文本与真实文本完全匹配的样本的总数。

**3.2 数据集获取**

实验使用中国科学技术大学视觉研究小组采集并公开发布的中国手语识别（Chinese Sign Language Recognition, CSL）数据集[8]，该数据集中包含两个不交叉的子数据集，分别为孤立手语数据集和连续手语数据集。在孤立手语数据集中每段视频只包含单个词语的手语演示，而在连续手语数据集中每段视频都包含一段句子的手语演示，为了使训练出来的手语识别系统功能更加强大，在此使用连续手语数据集作为本次实验所用的训练和测试数据集。

连续手语数据集由50名手语演示者，通过Kinect 2.0深度摄像头录制100句常用句子，并且每名手语演示者对每个句子演示5遍。因此，该数据集含有100个不同句子类别，每个类别包含250个手语演示视频，数据集总共包含25000个手语演示视频。

**3.3 基于帧差法的关键帧提取算法**

视频作为图像在时间上的连续序列，在某一时间段中若视频内目标没有发生明显运动，则相邻帧之间的差距微乎其微，若在此之间目标发生明显的运动，则相邻帧之间会有较大的区别。为了去除手语视频中存在的大量冗余数据，在此使用帧差法提取手语视频中的关键帧图像。帧差法提取视频关键帧过程如下：

步骤1：初始化关键帧集合；

步骤2：对手语视频进行逐帧提取并依次进行灰度处理操作，得到由张二维图像组成的帧序列；

步骤3：对帧序列进行二值化操作，并将序列中的每个按像素进行求和，得数值序列;

步骤4：将数值序列中的相邻元素（从第二个元素开始）进行两两相减并取其绝对值，得差分序列；

步骤5：计算阈值，并依次对差分序列中得每个元素进行判断，如果发现，则进行操作；

步骤6：集合即为手语视频的关键帧集合。



图5 帧差分获取的关键帧图片示例

**3.4 关键帧图像预处理**

每个手语视频经过帧差分法得到的关键帧图像数目并不相等，为了后续处理方便，本实验固定每个手语视频的关键帧为48张。因此，需要对每个手语视频提取出的关键帧数量进行统一，对于关键帧数目超过48张的进行随机剔除，而对于关键帧数目低于48张的，则在关键帧序列末尾填充全白图像，之后使用OpenCV将关键帧图像大小固定为128×128。

为了进一步提高模型识别性能，需要对关键帧图像进行高斯滤波以突出人物手臂等边缘信息，再对其进行直方图均衡化提高图像全局对比度。

**3.5 手语识别模型设计**

本文提出并设计的ResNet18-LSTM由两部分组成，结构图如图6所示，其中ResNet18能够提取关键帧图片的特征，且ResNet网络结构能够有效避免训练过程中的梯度爆炸或消失问题，LSTM能够提取关键帧间的时序信息。由于ResNet18-LSTM模型的输出结果为文字，因此最后还需要使用Tacotron2语音合成模型将文字转换为语音，转换过程如图7所示。模型运行步骤如下：

1. 使用帧间差分提取手语视频中的关键帧，将提取好的关键帧图片输入至ResNet18网络中，经过卷积、池化和全连接等操作后，获得帧序列特征向量；
2. 将输入到LSTM中，得到未归一化的词汇预测概率序列；
3. 对进行softmax概率归一化并输出翻译出的文字数据；
4. 最后将文字数据输入到Tacotron2语音合成模型并转化为语音信号。

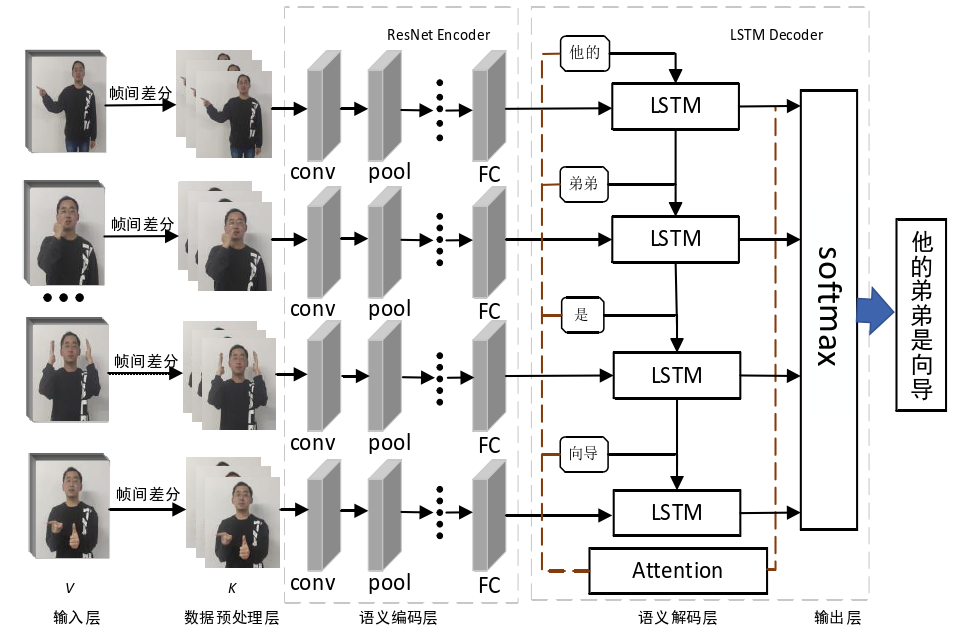


图6 ResNet18-LSTM结构图

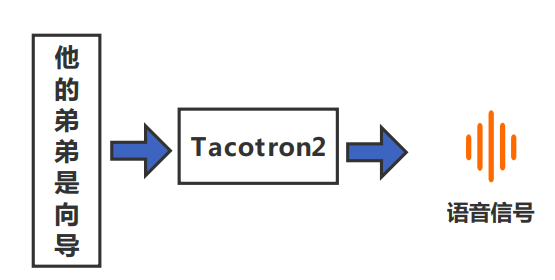


图7 语音合成

# 4 手语识别实验结果与分析

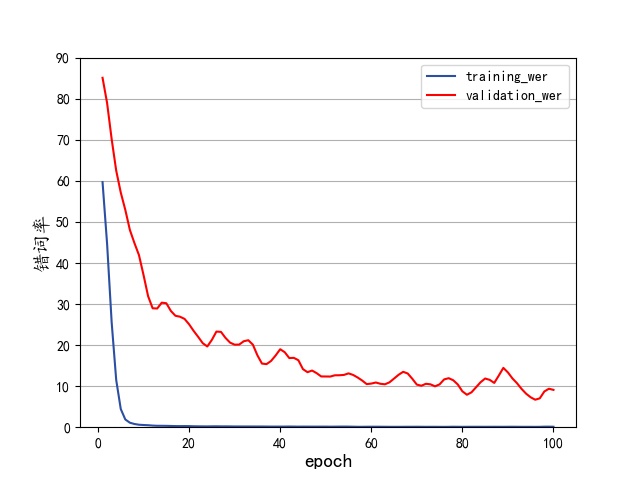
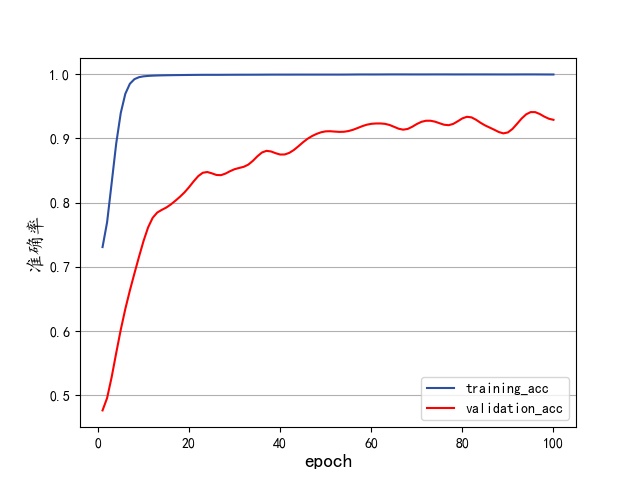
将连续手语数据集进行关键帧图像提取和预处理后，对ResNet18-LSTM手语识别模型进行测试。本文实验代码基于PyTorch实现，硬件环境：显卡型号GeForce RTX 3090，24G显存、处理器型号AMD EPYC 7601，16核64G内存；软件环境：Ubuntu 20.04.3 LTS、Python 3.8、PyTorch 1.10.0、CUDA 11.3.0。

实验模型中优化器、损失函数和学习速率等的定义如表1所示，其中Adam优化器的权重衰减参数设置为，Dropout参数设置为0.5，初始学习率设置为,学习率随训练次数以指数的形式进行衰减。除此之外，实验其他参数设置如下，训练轮数epoch值为100，样本批次batch值为16，CPU核心数num\_workers值为16。

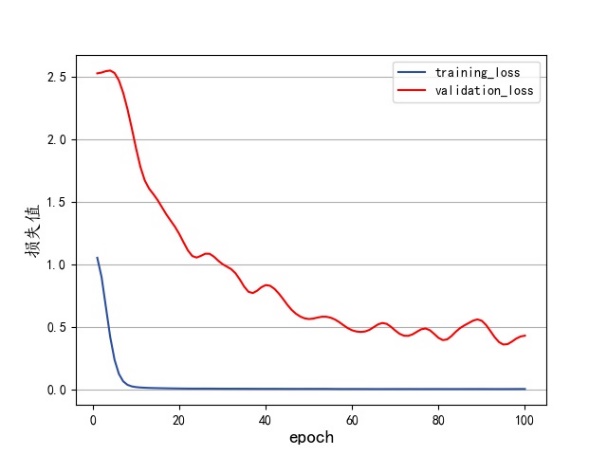
表1 模型指标定义

|  |  |
| --- | --- |
| 类别 | 指标 |
| 损失函数 | 交叉熵Cross-Entropy |
| 优化器 | Adam优化器 |
| 学习率 | 固定学习率衰减 |
| 防止过拟合 | L2正则和Dropout |

首先将ResNet18-LSTM模型在数据集上进行训练，模型训练过程中准确率、错词率和损失值的变化如图8所示，其中蓝色曲线代表模型在训练集上的性能表现，红色曲线代表模型在验证集上的性能表现。从图中可以看出，在经过大约10个epoch后，模型在训练集上快速收敛，此时其在训练集上准确率稳定在99.8%左右，错词率稳定在0.1%左右，损失值稳定在0.003左右；在经过大约60个epoch后，模型在验证集上收敛，此时其在验证集上准确率最高为96.37%，错词率最低为5.36%，损失值最低为0.205。



(a) 准确率变化 (b) 错词率变化



(c) 损失值变化

图8 ResNet18-LSTM模型训练结果图

之后将本文ResNet18-LSTM模型与HLSTM-Attention、Two-Stream CNN、HST-GNN、CNN-LSTM、3D-ResNet和RNN-Transducer等6种手语识别模型在同一手语数据集CSL下进行测试比较。由表2可知，HST-GNN模型在准确率上表现较好，但错词率较高，很难满足实际的手语翻译需求；RNN-Transducer模型具有较低的错词率，性能较好；而其他四种模型都在错词率或准确率上表现不佳。本文模型对比其他6种模型中的最好结果，在错词率上最少降低了0.7%，在准确上最少提高了3.7%，比较结果充分表明本文手语识别模型与其他模型相比存在较大优势。

表2 手语识别模型对比

|  |  |  |
| --- | --- | --- |
| 模型方法 | 错词率 | 准确率 |
| HLSTM-Attention[10] | 41.90% | 85.10% |
| Two-Stream CNN[11] | 23.70% | 82.70% |
| HST-GNN[12] | 19.50% | 92.6% |
| CNN-LSTM[13] | 59.40% | — |
| 3D-ResNet[14] | 32.70% | — |
| RNN-Transducer[15] | 6.10% | ­­­— |
| **ResNet18-LSTM** | **5.4%** | **96.30%** |

# 5 语音到手语的翻译

**5.1 语音到手语翻译模型**

手语识别技术实现了手语到语音的翻译，满足了健听人士理解听障患者手语的需求，为了更好的消除听障患者与健听人士的交流障碍，还需实现语音到手语的翻译功能，以满足听障患者理解健听人士语音信息的需求。现今，世界各国科研人员对于语音到手语的翻译研究较少，大多的语音到手语翻译模型存在着性能较差，生成的手语视频效果不好等缺点[9]，这也成为了手语与语音互译系统实现市场化所面临的最大瓶颈之一。

**5.2 手语基准动画的制作**

手语动作的呈现形式是语音到手语翻译的关键，本文使用3D动画建模为每个手语基础词汇制作一个手语动画视频，并将其称之为手语基准动画。建模软件使用Autodesk Maya，该软件是一款性能优越的三维动画建模软件，具备先进的三维建模、运动匹配以及目标渲染等功能。手语基准动画中人物的手语动作将依照2018年公布的《国家通用手语常用词表》建立，如表3所示为“我爱你们”的手语图示。

表3“我爱你们”的手语图示

|  |  |  |
| --- | --- | --- |
| **文字** | **动作** | **图示** |
| 我 | 右手食指指向胸部（或者右手手掌拍向胸部）。 |  |
| 爱 | 左手伸出大拇指，右手抚摸左手大拇指背面。 |  |
| 你们 | （一）右手食指指向正前方，手背向上。  （二）右手掌心向下，从胸前顺时针平移至身体右侧。 |  |

**5.3 手语动画编码**

每个手语基准动画都具有一个唯一的编码序列，该序列依据基准动画中手部特征产生，编码序列长度固定为7位或14位，其中每七位的前两位表示左右手以及掌间指向，后五位表示每个手指的指向。基准视频中若人物使用左手则第一位编码为R，使用右手则第一位编码为L，掌尖以及每个手指的指向都由一个十进制数表示，因此指向共有10种类别，分别为“右”、“右上”、“上”、“左上”、“左”、“左下”、“下”、“右下”、“正前”以及“正后”，并依次从0编号到9，即“右”序号为0，“正后”序号为9。编码示例如表4 所示。

表4 编码示例

|  |  |  |  |
| --- | --- | --- | --- |
| **文字** | **首编码** | **尾编码** | **整体编码** |
| 我 | R1 | 11111 | R111111 |
| 爱 | R8/L9 | 18888/29999 | R188888L929999 |
| 你 | R9 | 48999 | R948999 |
| 们 | R8 | 88888 | R888888 |

**5.4 语音到手语模型建立**

为此本文提出一种基于编码器解码器结构的端到端语音到手语翻译模型，该模型的结构如图9所示，语音信号通过模型转换为一串编码数据，之后通过与手语基准动画进行匹配并合成完整的手语动画视频。

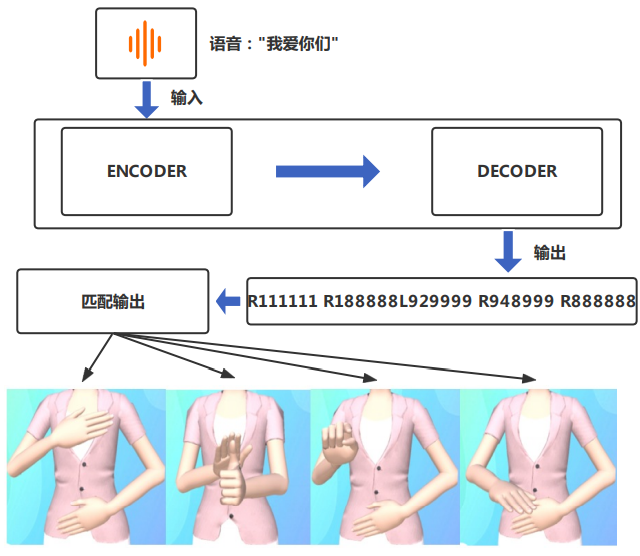


图9 语音到手语翻译模型

# 6 结束语

本文首先提出了一种手语识别模型ResNet18-LSTM，该模型可以将手语视频直接翻译为文本，实验结果表明其在验证集上准确率最高为96.37%，错词率最低为5.36%，与其他手语识别模型相比具有较大优势，能够满足正常的手语识别需求，之后利用语音合成模型Tacotron2将文本转换位语音信号，为用户提供多种翻译选择以提高服务质量，最后针对语音到手语的翻译，使用Autodesk Maya动画建模软件为每个手语词汇制作手语基准动画，并依据动画人物手势特征进行编码，之后利用一个编码器解码器结构的端到端模型将输入的语音转换为编码序列，编码序列与基准动画进行匹配以生成完整的手语动画。于此同时，手语识别模型ResNet18-LSTM依然存在一些不足之处，比如对于复杂背景下的手语视频翻译能力不佳，对于较大的手语视频需要消耗较长的时间。但是，随着各国科研人员对于手语翻译的关注，相信会有更大更全的手语数据集出现，以此数据集训练出的模型泛化能力将大幅增强。

# 参考文献

1. 刘远,张万洪.习近平法治思想中的特定群体权利保障理论[J].求是学刊,2022,49(05):18-27.
2. World Health Organization. Deafness and hearing loss[EB/OL].(2021-04-01)[2022-11-19]. https://www.who.int/zh/news-room/fact-sheets/detail/deafness-and-hearing-loss.
3. 央视新闻. 中国是世界上听力残疾人数最多的国家[EB/OL].(2020-09-27)[2022-11-19]. https://www.sohu.com/a/421233308\_701427.
4. 陶唐飞,刘天宇.基于手语表达内容与表达特征的手语识别技术综述[J/OL].电子与信息学报:1-19[2022-11-28].
5. He K，Zhang X，Ren S，et al. Deep residual learning for image recognition[C]//IEEE. 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). New York：IEEE，2016：7780459.
6. HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.
7. SHEN J,PANG R,WEISS R,et al.Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[C]//International Conference on Acoustics,Speech,and Signal processing,2018:4779-4783.
8. ZHOU H, ZHOU W, LI H. Dynamic Pseudo Label Decoding for Continuous Sign Language Recognition[C]// 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai: IEEE, 2019:1282-1287.
9. 杨淑莹,田迪,郭杨杨,赵敏.端到端手语翻译研究[J/OL].系统科学学报:1-7[2022-11-28].
10. GUO D, ZHOU W, LI H, et, al. Hierarchical LSTM for sign language translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans: AAAI, 2018: 6845-6852.
11. HUANG J，ZHOU W，ZHANG Q，et al.Video-based sign language recognition without temporal segmenta tion[C]//Proceedings of the AAAI Conference on Artifi cial Intelligence，New Orleans，USA，2018：2257-2264.
12. KAN Jichao, HU Kun, HAGENBUCHNER M, et al. Sign language translation with hierarchical spatio-temporal graph neural network[C]. Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, USA, 2022: 2131–2140. doi: 10.1109/ WACV51458.2022.00219.
13. LI H，GAO L，HAN R，et al.Key action and joint CTC Attention based sign language recognition[C]//ICASSP 2020—2020 IEEE International Conference on Acoustics， Speech and Signal Processing（ICASSP），Barcelona，Spain， 2020：2348-2352.
14. PU J，ZHOU W，LI H.Iterative alignment network for continuous sign language recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recogni tion（CVPR），Long Beach，USA，2019.
15. GAO Liqing, LI Haibo, LIU Zhijian, et al. RNN Transducer based Chinese sign language recognition[J]. Neurocomputing, 2021, 434: 45–54. doi: 10.1016/j.neucom.2020.12.006.