Modeling and Simulation 建模与仿真, 2024, 13(1), 365-376
Published Online January 2024 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2024.131035

Hans汉斯

基于喉部振动的语音自动识别系统的设计

陆心怡，卜朝晖*

上海理工大学生物医学工程研究所，上海

收稿日期：2023年11月14日；录用日期：2023年11月30日；发布日期：2024年1月16日

摘　要

现有且成熟的语音识别系统基本局限于健康群体及主流语言，并不适用于声带受损的患者。因此，本文研究设计了一款基于喉部振动的语音自动识别系统，旨在为声带受损患者及言语障碍残疾群体的康复训练与正常生活提供一种可行的方案。采用智能数字听诊器Mintti Smartho-D2对喉部软骨振动信号进行检测，借助于主流的语音识别深度学习算法：卷积神经网络模型、卷积长短时记忆神经网络模型、卷积递归神经网络模型，对喉振信号数据集分别进行多次训练，以期实现喉部软骨振动信号到正常语音信号的转换。通过对比实验，得出三种模型的测试字错率分别为0.1572、0.2018、0.06787，其中识别效果最佳为卷积递归神经网络模型，实现了字错率在安静环境下低于0.07的效果。本文可初步验证该设计的可行性及CRNN模型能够在效率和识别效果上取得较好的性能。

关键词

喉部软骨振动，深度学习，语音识别，卷积神经网络

Design of Automatic Speech Recognition System Based on Throat Vibration

Xinyi Lu, Zhaohui Bu*

Institute of Biomedical Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Nov. 14th, 2023; accepted: Nov. 30th, 2023; published: Jan. 16th, 2024

Abstract

Existing and mature speech recognition systems are basically limited to healthy groups and mainstream languages, and are not applicable to patients with impaired vocal cords. Therefore, in this paper, an automatic speech recognition system based on laryngeal vibration is designed

*通讯作者。

文章引用: 陆心怡, 卜朝晖. 基于喉部振动的语音自动识别系统的设计[J]. 建模与仿真, 2024, 13(1): 365-376.
DOI: 10.12677/mos.2024.131035

with the aim of providing a feasible solution for the rehabilitation training and normal life of patients with impaired vocal folds and speech-impaired disabled groups. Intelligent digital stethoscope Mintti Smartho-D2 was used to detect the vibration signal of laryngeal cartilage with the help of mainstream speech recognition deep learning algorithms: The convolutional neural network model, convolutional short-duration memory neural network model and convolutional recurrent neural network model were trained several times on laryngeal vibration signal data set respectively, in order to realize the conversion of laryngeal cartilage vibration signal to normal speech signal. Through comparison experiments, it is concluded that the test word error rates of the three models are 0.1572, 0.2018, and 0.06787, respectively, among which the best recognition effect is the convolutional recurrent neural network model, which realizes a word error rate of less than 0.07 in a quiet environment. This paper can initially verify the feasibility of the design and the CRNN model can achieve better performance in terms of efficiency and recognition effect.

Keywords

Laryngeal Cartilage Vibration, Deep Learning, Speech Recognition, Convolutional Neural Network

1. 引言

语言交流是人类社交和表达的基本手段之一，而我国聋哑症的发病率约为 2‰，每年总数可达 5700 万，听损伤的发病人数约为 17 万，并伴有言语障碍[1]。导致言语障碍的原因包括疾病、事故、嗓音滥用、医疗手术、老龄化、环境污染等，嗓音丧失的风险不断增加[2]。为语音识别技术开发新的方向和思路迫在眉睫，语言障碍人士日常生活最基本的交流需求亟待解决。

目前，市面上已有的产品如手语识别设备、唇语识别技术等解决方式，较难普及且经济成本较高、实际使用中限制较大[3] [4]。邱浩海和韦创军开发的“声活”APP 可以通过云服务平台“云来”让聋哑人发出“声音”，将聋哑人的手语信息转化为语音信息[4]。美国两位大学生研发了一款名叫“SignAloud”的手语翻译手套，利用手语识别技术翻译手语并转化成语音[5]。此外，孙科等人开发的基于前置摄像头的唇读输入“Lip-Interact”技术可以捕捉用户的嘴部动作并识别命令[6]。

然而，使用基于视觉的方法时，获取与手、唇有关的信息比较困难，需要进行复杂的图像处理，而且对象的形状识别会受到背景条件和照明敏感度的影响，且用户总是需要摄像头，日常使用更加不便[7]。

因此，本课题设计了一款基于喉部振动的语音自动识别系统，结合高效的深度学习算法，达到转换出正常语音的目的，避免了受镜头和环境的影响，也帮助患者重新用清晰可懂的语音进行交流，满足了患者对于日常生活最基本的沟通需求，重新推动其未来的人际关系和社交发展。

此外，本系统设计具有以下主要创新点：

1) 从无法正常发声交流、声带受损的患者角度出发，借助现有的语音识别神经网络模型，以期为患者生成相应的、可懂度高的清晰语音。

2) 具有可迁移至移动设备开发 app 的潜力，并且便于开发成可穿戴式的辅助设备。

3) 收集到的喉部软骨信号比较纯净，信噪比高，不易受环境噪声影响。

2. 研究方法

2.1. 语音自动识别系统基本结构

喉部软骨振动信号是指在人的喉部产生的一种声音信号，当气流从肺部冲击喉部区域，随着声带振动，产生声音，与喉部其他结构如喉软骨等共振，形成喉部软骨振动信号[8]。

自动语音识别系统(Automatic Speech Recognition, ASR)，被认为是将语音信号的声学微观结构连续地转化为其隐含的语音宏观结构[9]。

本系统结合现有的语音自动识别系统流程思路，对患者喉部软骨振动信号进行训练学习，达到转换出正常语音的目的，如图 1 为 ASR 系统应用于喉部软骨振动信号识别的结构流程图。

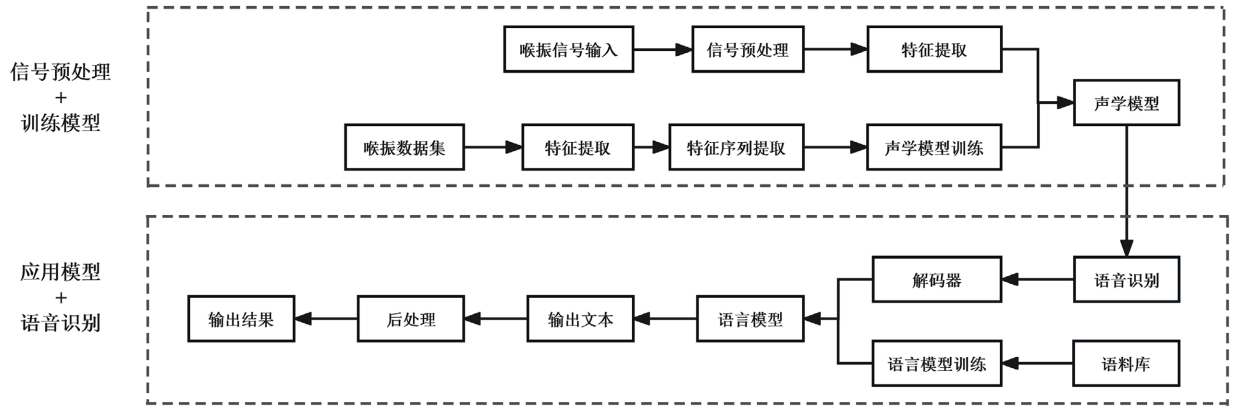

Figure 1. Structural flowchart of the automatic recognition system of throat vibration signals
图 1. 喉振信号自动识别系统的结构流程图

2.2. 神经网络算法

经过半个多世纪的努力，自动语音识别系统在减少单词错误率(Word Error Rate, WER)方面得到了显著改善，增加了应用的适用性[10]。其中，循环神经网络 RNN 模型和长短期记忆网络 LSTM 等都是经实验发现效果比较好的模型。因此，本文针对将喉振信号转换为正常语音信号的目的，选择三种算法并进行对比研究，其中预处理方法、算法的选用以及模型的解码处理如下表 1 所示。

Table 1. Selection of three deep learning algorithmic solutions
表 1. 三种深度学习算法方案的选用

方案	数据预处理方法	神经网络模型	解码器
1	MFCC	CNN	ctc_greedy
2	Fbank	CRNN	ctc_beam_search
3	Fbank	CNN-LSTM	ctc_beam_search

2.3. 系统整体流程

本系统采用智能数字听诊器 Mintti Smartho-D2 进行喉部软骨振动数字信号的采集，并通过数字界面输出到计算机、手机或平板电脑等设备上进行分析和诊断。Mintti Smartho-D2 是一种高度符合人体工程学的专用辅助设备，采集微弱人体信号可达到 100 倍放大[11]。

通过设计可训练学习患者喉振表述模式的深度学习算法程序，完成对喉振信号识别的样本训练、实时显示与语音播报。系统整体流程如图 2 所示。

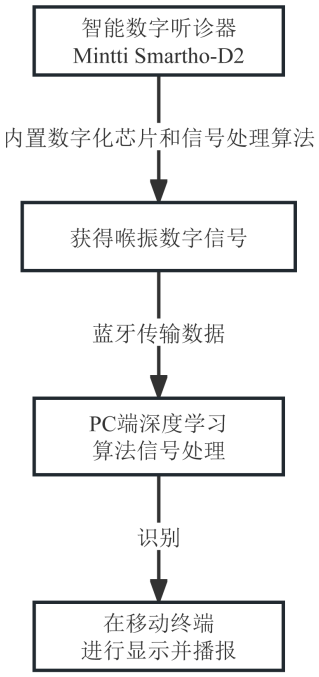

Figure 2. The flow chart of the overall system scheme
图 2. 系统整体方案流程图

3. 算法及理论基础

3.1. 预处理

在语音自动识别系统的开发中，预处理被认为是第一步，用于区分浊音或清音信号并创建特征向量。预处理对输入信号 $x(n)$ 进行调整或修改，使其更适合特征提取分析[12]。

对前期采集的喉部振动信号进行预处理工作如图 3 所示，包括：对信号去噪、端点检测、预加重、分帧以及加窗等[13]。

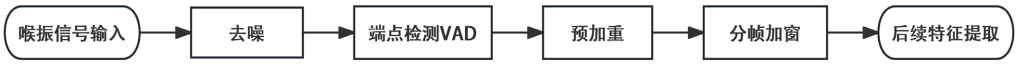

Figure 3. Flow chart of throat vibration signal pre-processing
图 3. 喉振信号预处理流程图

1) 去除噪声：

环境噪声在数据预处理中起着重要的影响，对喉振信号进行去噪可以提高信号质量，增强自动识别系统的鲁棒性[14]。

2) 端点检测：

数字喉振信号由喉部软骨振动信号、无声段和各种背景噪声组成，端点检测旨在确定有效信号的开始和结束点。不仅可以减少系统处理时间从而实现系统的实时处理，还可以消除无声段的噪声干扰从而提高后续过程的识别性能[15]。

3) 预加重滤波器：

预加重旨在提高信噪比，增强信号的清晰度和可辨识度，补偿采样过程中的高频的必要衰减，进而使得参数分析以及频谱分析等过程更加方便[13]。

预加重滤波器是一个一阶差分滤波器，其输出信号 $y(n)$ 和输入信号 $s(n)$ 之间的关系可以用方程表示：

$$y(n)=s(n)-as(n-1) \tag{1}$$

其中 a 为预加重系数。本系统设计的预加重系数取值为 0.97。

4) 分帧加窗(倒谱提升窗口)：

输入的喉部软骨振动信号每一帧都被处理并可视化为一个单一的特征向量，且都有一个锥形窗口用于消除喉振信号中的不连续因素。在为每帧数据计算出特征参数后，为这个系数分别乘以不同的权系数。实际上就是一个短的窗口，本设计选择汉明窗：

$$\overline{c_m}=w_m c_m \tag{2}$$

$$w_m=1+\frac{K}{2}\sin\left(\frac{\pi m}{K}\right),1\le m\le K \tag{3}$$

3.2. 卷积神经网络 CNN

卷积神经网络(Convolutional Neural Network, CNN)是一种经典的深度学习算法，输入数据在被送入深度神经网络之前，要经过卷积层、非线性层、池化层和全连接层。在二维特征值被扁平化后，通过对特征图逐行取值转换成一维数列，构建出整个模型的输出结果，具体步骤如下：

1) 输入喉部软骨振动信号 $x(t)$，其中 t 表示时间。使用短时傅里叶变换(Short-Time Fourier Transform, STFT)将信号 $x(t)$ 分解为不同的频率分量：

$$X(m,k)=\sum_{n=0}^{N-1}x(n)w(n-m)\mathrm{e}^{-j2\pi nk/N} \tag{4}$$

2) CNN 包含多个卷积层和池化层，其中每个层都有一组可学习的卷积核或滤波器，进行卷积操作并应用非线性激活函数 ReLU 并进行最大池化：

$$y(i,j)=\sigma\left(\sum_{k=1}^{K}\sum_{l=1}^{L}w(k,l)x(i+k,j+l)+b\right) \tag{5}$$

$$f(x)=\max(0,x) \tag{6}$$

$$y(i,j)=\max_{k,l}x(i\times s+k,j\times s+l) \tag{7}$$

其中，σ 是激活函数，$w(k,l)$ 是卷积核的权重，$x(i+k,j+l)$ 是输入信号的一个元素，k 和 l 是卷积核的大小，s 是池化核的大小。

3) 全连接层将输出映射到一个预定义的输出维度并使用 softmax 函数作为激活函数：

$$y_i=\frac{\mathrm{e}^{z_i}}{\sum_{j=1}^{C}\mathrm{e}^{z_j}} \tag{8}$$

3.3. 递归神经网络 RNN

递归神经网络(Recurrent Neural Network, RNN)，是一种能够处理序列数据的神经网络，通过在每个时间步将当前输入和前一时间步的状态作为输入，来计算当前时间步的输出和状态，这种状态的循环传递使得 RNN 可以将之前的信息存储在状态中，并随着输入的处理而更新状态[16]。

在图 4 中，在每个时间步，都要先将当前的输入 x_t 和上一个时间步的隐藏状态 h_{t-1} 作为 RNN 的输入，然后计算当前的隐藏状态 h_t。随后，使用 h_t 生成相应的输出 y_t。

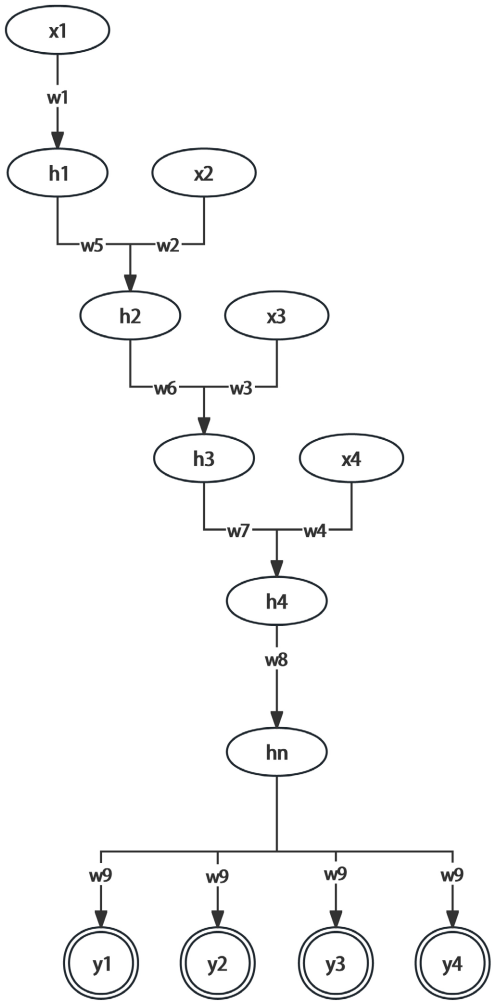

Figure 4. RNN time unfolding diagram
图 4. RNN 时间展开图

3.4. 长短时记忆神经网络 LSTM

长短时记忆网络(Long Short-Term Memory, LSTM)是一个强大的递归神经系统，通过引入特殊的“记忆单元”来记住长期的上下文信息，并通过“门控”机制来控制信息的流动，从而有效地解决了传统循环神经网络(RNN)在长序列数据处理中的梯度消失问题[17]。

LSTM 和 RNN 都是循环神经网络的类型，在本文系统应用中，LSTM 是 RNN 的一种改进，更适合处理长时间序列数据。二者的对比分析如下表 2 所示：

Table 2. Comparative analysis of LSTM and RNN algorithms
表 2. LSTM 与 RNN 算法对比分析

项目名称	RNN	LSTM
架构	简单，递归结构	复杂，包括递归结构和三个门控结构
记忆单元	有限的短期记忆	长期记忆和短期记忆单元
训练时间	快速	较慢
遗忘机制	没有遗忘机制	通过门控结构遗忘信息

续表

防止梯度消失问题	容易发生	通过门控结构和跨单元连接解决问题
准确性	可能欠佳	相对准确
应对长序列	效果欠佳	适应长序列
优点	简单、易于理解	能够处理长期依赖关系，并且能够遗忘信息

在模式识别领域，LSTM 模型已经得到广泛应用，本文基本思想是将喉振信号作为时间序列输入，经过多个 LSTM 层的处理，输出对应的文本结果。为了更好地训练 LSTM 模型，通常会使用交叉熵损失函数，其计算公式如下：

$$L(\theta)=-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i}\log P\left(y_{i,t}\mid x_i,\theta\right) \tag{9}$$

其中，N 为训练数据集的大小，T_i 为第 i 个样本的时序长度，$y_{i,t}$ 为第 i 个样本在第 t 时刻的真实字符标签，θ 为 LSTM 模型的参数。

3.5. 解码算法

连接时序分类(Connectionist Temporal Classification, CTC)是常用于序列学习的技术，用以识别序列中的目标标签，即本系统喉部软骨振动信号识别中的词语。通过训练深度神经网络来学习喉振特征与目标标签之间的关系，并将学到的特征映射到目标标签[18]。

对两种解码算法进行对比分析，分别为贪婪策略解码(CTC_Greedy Search)、集束搜索解码(CTC_Beam Search)。

贪婪策略解码(CTC_Greedy Search)在计算 CTC 损失时以最大概率解码出输出序列。在每一时刻选择当前最有可能的输出，将其作为输出序列的一部分。输出结果来自概率最高的路径，得到识别出的文本。

集束搜索解码(CTC_Beam Search)是对前馈神经网络(FNN)和循环神经网络(RNN)的输出进行语音识别的常用方法。与贪婪策略相比，它考虑多种路径可能导致的误识别，在搜索空间中生成多种路径，搜索到最终时刻，找到识别句子概率最大的路径[19]。因此，集束搜索通常比贪心策略更准确。

4. 实验结果与分析

4.1. 数据集准备

由于训练数据应为喉部软骨振动信号，目前还没有公开的数据集可供使用，本系统设计使用数字听诊器采集得到的来自 40 位健康人的喉部软骨振动信号构建数据集进行实验，具体信息见表 3，作为训练数据集进行后续处理和特征提取。

Table 3. Statistics on the source information of the training dataset
表 3. 训练数据集来源信息统计

数据集	TMDATA
语种	汉语(普通话)
录制环境	静音
时长	约 48 小时
句子数目	约 500 句
说话人数目	40
文本类型	普通话测试材料

4.2. 模型训练配置

4.2.1. 实验配置

本系统实验平台和软硬件情况如表 4 所示。

Table 4. Experimental configurations
表 4. 实验配置

硬件情况	测试环境	测试数据
CPU i7-13700、GPU RTX 4070 Ti	Cuda11.6 安装 GPU 版的 PaddlePaddle 以及 Pytorch 的 Pycharm (Python 环境)	TMDATA

4.2.2. 模型训练参数

在神经网络算法中，有两类参数：参数和超参数。参数是训练神经网络最终要学习的目标，如权重。超参数是控制模型结构、功能、效率等的调节途径，本实验中使用的三种神经网络算法方案在实际训练中需要根据情况不断调节参数，因此，本系统采用的主要参数含义与值见表 5~7。

Table 5. CNN model configurations
表 5. CNN 模型参数

参数名	参数含义	方案 A
batch_size	单轮训练的样本数目	32
learning_rate	学习率(控制模型训练过程中更新权重的步长)	0.001
data_mean	预处理过程中进行归一化处理时的均值	−3.017657
data_std	数据集中所有样本的标准差	51.585384

Table 6. CRNN model configurations
表 6. CRNN 模型参数

参数名	参数含义	方案 B
alpha	语言模型分数的权重调整	1.2
batch_size	单轮训练的样本数目	32
beam_size	选择保留的最佳输出序列的数量	10
decoder	序列建模方法：贪心解码	ctc_greedy
num_conv_layers	CNN 中卷积层的数量	2
num_proc_bsearch	并行执行 Beam_Search 的进程数	8
num_rnn_layers	RNN 中循环层的数量	3
rnn_layer_size	RNN 中循环层的神经元数量	1024

Table 7. CNN-LSTM model configurations
表 7. CNN-LSTM 模型参数

参数名	参数含义	方案 C
alpha	语言模型分数的权重调整	2.2
batch_size	单轮训练的样本数目	16

续表

decoder	序列建模方法：集束搜索	beam_search
r_num_blocks	双向 LSTM 中反向 LSTM 的块数	3
num_rnn_layers	RNN 中循环层的数量	5
num_blocks	模型深度	12
Input_layer	使用 CNN 作为特征提取器的输入层	conv2d
learning_rate	学习率(控制模型训练过程中更新权重的步长)	0.001

4.3. 系统测试结果与分析

4.3.1. 评估指标

本系统使用字错误率(Character Error Rate, CER)和训练损失(Loss Function, LOSS)并辅以准确率作为评估指标，其中 CER 表示识别结果与参考文本之间不匹配的字数占总字数的比例，LOSS 表示训练过程中神经网络输出与目标输出之间差异的度量，主要用作是否正常收敛的参考。

当涉及到字符错误率(CER)时，通常使用的公式如下：

$$\mathrm{CER} = \frac{S + D + I}{N} \tag{10}$$

其中，S 表示替换错误的字符数，D 表示删除错误的字符数，I 表示插入错误的字符数，N 表示参考字符总数。

4.3.2. 测试结果与分析

在系统测试时，选择待识别的、实时采集的喉振信号文件，最终输出识别结果，如图 5 所示。测试人表述原话为：“作为一个普通中学的普通教师，我的旅游机会是少而又少的。”

```
执行预测时间：7620ms
解码消耗时间：28ms
识别总时间：8976ms，识别结果：所为一个普同中学的普同教师我得旅有机会是少用效大，得分：61.919655
```

Figure 5. Schematic identification results
图 5. 识别结果示意

最终总结三种方案准确率如表 8 所示。

Table 8. Performance of the model on the training set
表 8. 模型在训练集上的表现

模型	准确率(真实信号实时测试)
CNN	48.6%
CRNN	79.2%
CNN-LSTM	66.7%

训练过程中记录各种关键指标和错误信息，并保存在日志文件中，包括训练损失、准确率、学习率、验证损失、验证准确率等信息，进而便于监测并进行模型调试和优化。三种模型方案在测试集上的具体表现如图 6~8 所示。

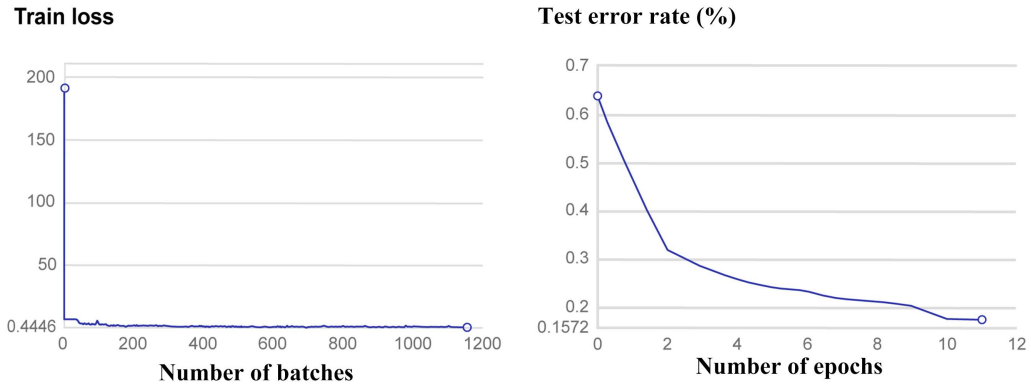

Figure 6. CNN model
图 6. CNN 模型

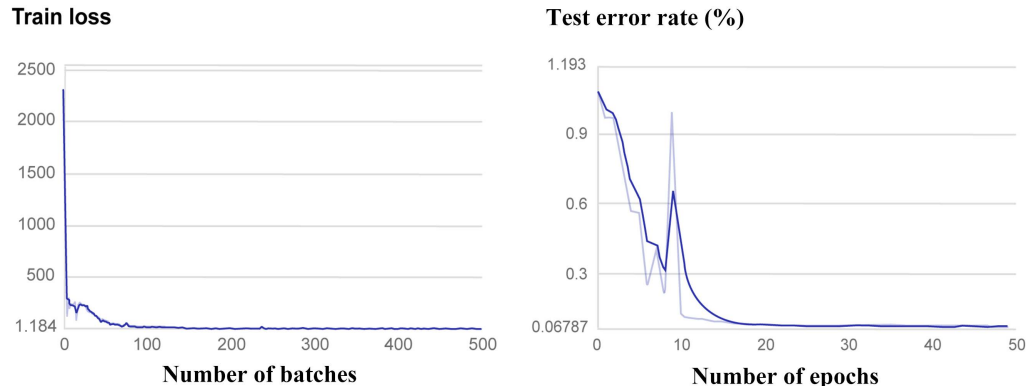

Figure 7. CRNN model
图 7. CRNN 模型

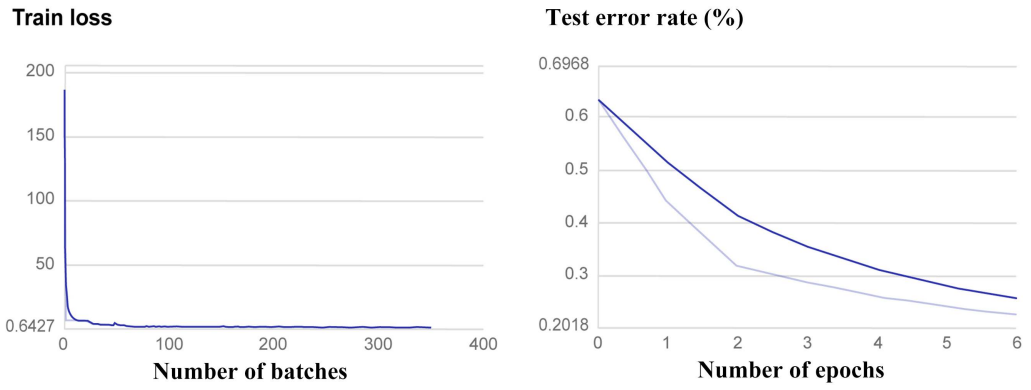

Figure 8. CNN-LSTM model
图 8. CNN-LSTM 模型

最终得到模型各指标结果总结如表 9 所示，在实际训练过程中，CRNN 模型在综合效率表现上更佳。

CNN-LSTM 模型在训练中时间成本较大，计算复杂，效率过低，未能在实验中体现出应有的准确率。分析原因为 LSTM 层仍然存在计算复杂度高、参数数量多、难以并行化等缺点。CNN 模型效率高、计算较快、结构简单，但是识别效果不尽如人意。原因为 CNN 模型只能提取局部特征，忽略了喉部振动信号的时序性和上下文相关性。

因此，通过综合对比分析，方案二的 CRNN 模型既能有较低的时间成本，又能有较好的识别效果。

Table 9. Performance of the model on the testing set
表 9. 模型在测试集上的表现

测试指标参数	参数含义	方案一 CNN	方案二 CRNN	方案三 CNN-LSTM
Learningrate	学习率	9×10^{-5}	4.876×10^{-8}	9×10^{-4}
Trainloss	损失函数值	0.4446	1.184	0.6427
Testcer	测试集字符错误率	0.1572	0.06787	0.2018
训练时间	消耗的时间成本	16 个小时	48 小时以上	168 小时以上

5. 结论

本文实现了基于喉部振动的语音自动识别系统的设计和测试，并提出了评判系统的指标，根据实验结果得出了相对最佳方案。其中，CRNN 模型实现了字错率在安静环境下低于 0.07 的效果，说明其在喉振信号识别任务上具有较优的性能，能基本识别出语音，有一定语义错误。

在本文实验中，针对该系统做了训练和测试的部分，需要在 PC 端完成，日后研究可改善其实用性，迁移至可移动端。同时，声带受损患者等实验对象的喉部软骨振动信号的采集并由此构建公开的大型数据集是非常有必要的工作。

参考文献

[1] 福州弋元信息技术有限公司. 一种帮助听障人和非听障人交流的智能交互设备商业计划书[EB/OL]. https://www.docin.com/p-2248375934.html, 2023-03-23.

[2] Lee, W., Seong, J.J., Ozlu, B., Shim, B.S., Marakhimov, A. and Lee, S. (2021) Biosignal Sensors and Deep Learning-Based Speech Recognition: A Review. *Sensors*, **21**, Article No. 1399. https://doi.org/10.3390/s21041399

[3] 李帅, 吴玉蓉. 面向聋哑人群的无障碍交流辅助系统设计研究[J]. 物联网技术, 2022, 12(11): 113-116. https://doi.org/10.16667/j.issn.2095-1302.2022.11.034

[4] 冯成龙, 刘桢. 人工智能技术在聋哑人沟通交流方面的应用[J]. 智库时代, 2021(7): 256-257.

[5] Joshi, A., Sierra, H. and Arzuaga, E. (2017) American Sign Language Translation Using Edge Detection and Cross Correlation. 2017 *IEEE Colombian Conference on Communications and Computing* (*COLCOM*), Cartagena, 16-18 August 2017. https://doi.org/10.1109/ColComCon.2017.8088212

[6] Sun, K., *et al*. (2018) Lip-Interact: Improving Mobile Device Interaction with Silent Speech Commands. *Proceedings of the* 31*st Annual ACM Symposium on User Interface Software and Technology*, Berlin, 14 October 2018, 581-593. https://doi.org/10.1145/3242587.3242599

[7] Ahmed, M.A., Zaidan, B.B., Zaidan, A.A., Salih, M.M. and Lakulu, M.M.B. (2018) A Review on Systems-Based Sensory Gloves for Sign Language Recognition State of the Art between 2007 and 2017. *Sensors*, **18**, Article No. 2208. https://doi.org/10.3390/s18072208

[8] Nishimura, T. (2020) Primate Vocal Anatomy and Physiology: Similarities and Differences between Humans and Nonhuman Primates. In: Masataka, N., Ed., *The Origins of Language Revisited*, Springer, Singapore. https://doi.org/10.1007/978-981-15-4250-3_2

[9] Ghai, W. and Singh, N. (2012) Literature Review on Automatic Speech Recognition. *International Journal of Computer Applications*, **41**, 42-50. https://doi.org/10.5120/5565-7646

[10] Lu, X., Li, S. and Fujimoto, M. (2020) Automatic Speech Recognition. In: Kidawara, Y., Sumita, E., Kawai, H., Eds., *Speech-to-Speech Translation*. *SpringerBriefs in Computer Science*, Springer, Singapore. https://doi.org/10.1007/978-981-15-0595-9_2

[11] Electronic User Guide Stethoscope. https://minttihealth.com/wp-content/uploads/2022/06/Digital-Stethoscope-User-Manual.pdf

[12] 谭磊, 余欣洋, 罗伟洋, 等. 基于深度学习的移动端语音识别系统设计[J]. 单片机与嵌入式系统应用, 2020, 20(9): 28-31+35.

[13] Lee, S.J. and Kwon, H.Y. (2020) A Preprocessing Strategy for Denoising of Speech Data Based on Speech Segment Detection. *Applied Sciences*, **10**, Article No. 7385. https://doi.org/10.3390/app10207385

[14] Labied, M., Belangour, A., Banane, M., *et al*. (2022) An Overview of Automatic Speech Recognition Preprocessing Techniques. 2022 *International Conference on Decision Aid Sciences and Applications* (*DASA*), Chiangrai, 23-25 March 2022, 804-809. https://doi.org/10.1109/DASA54658.2022.9765043

[15] Zhang, T., Shao, Y., Wu, Y., *et al*. (2020) An Overview of Speech Endpoint Detection Algorithms. *Applied Acoustics*, **160**, Article 107133. https://doi.org/10.1016/j.apacoust.2019.107133

[16] Caterini, A.L., Chang, D.E., Caterini, A.L., *et al*. (2018) Recurrent Neural Networks. In: *Deep Neural Networks in a Mathematical Framework*. *SpringerBriefs in Computer Science*, Springer, Cham. 59-79. https://doi.org/10.1007/978-3-319-75304-1_5

[17] 王毅, 谢娟, 成颖. 结合 LSTM 和 CNN 混合架构的深度神经网络语言模型[J]. 情报学报, 2018, 37(2): 194-205.

[18] 陈戈, 谢旭康, 孙俊, 等. 使用 Conformer 增强的混合 CTC/Attention 端到端中文语音识别[J]. 计算机工程与应用, 2023, 59(4): 97-103.

[19] Seki, H., Hori, T., Watanabe, S., *et al*. (2019) Vectorized Beam Search for CTC-Attention-Based Speech Recognition. 20*th Annual Conference of the International Speech Communication Association*: *Crossroads of Speech and Language*, *INTERSPEECH* 2019, Graz, 15-19 September 2019, 3825-3829.