Computer Science and Application 计算机科学与应用, 2021, 11(5), 1315-1324
Published Online May 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.115133

Hans汉斯

基于多模态的端到端语音识别

谭振宇，吴怡之

东华大学，信息科学与技术学院，上海
Email: tan_zy0102@163.com

收稿日期：2021年4月17日；录用日期：2021年5月11日；发布日期：2021年5月18日

摘　要

为了去除复杂的音频切分和强制对齐过程，并在噪音环境下充分发挥说话人发音过程中发音器官的视觉作用，本文提出了一种融合唇部特征的端到端的多模态语音识别算法。本文首先对说话人视频进行处理得到对应图像集，使用基于回归树的人脸对齐算法对图像集中发音的主要视觉部分进行特征提取，并与说话人的声学特征进行对齐融合得到新的特征，然后使用支持变长输入的端到端双向长短期记忆网络模型(DeepBiLstmCtc)对特征进行处理，输出对应的音素序列。实验结果表明该算法能有效地识别出视听觉信息中的音素序列，在噪声情况下也有一定的识别率提升。

关键词

多模态，端到端，语音识别，双向长短期记忆网络

End-to-End Speech Recognition Based on Multimode

Zhenyu Tan, Yizhi Wu

College of Information Science and Technology, Donghua University, Shanghai
Email: tan_zy0102@163.com

Received: Apr. 17th, 2021; accepted: May 11th, 2021; published: May 18th, 2021

Abstract

In order to remove the complex audio segmentation and forced alignment process, and give full play to the visual effect of the speaker's articulatory organs in the speaker's pronunciation process in a noisy environment, this paper proposes an end-to-end multi-modal speech recognition that incorporates lip features algorithm. This paper first processes the speaker's video to obtain the corresponding image set, uses the regression tree-based face alignment algorithm to extract the

文章引用: 谭振宇, 吴怡之. 基于多模态的端到端语音识别[J]. 计算机科学与应用, 2021, 11(5): 1315-1324.
DOI: 10.12677/csa.2021.115133

features of the main visual parts of the voice in the image set, and aligns and fuses it with the speaker's acoustic features to obtain new features, and then uses the end-to-end bidirectional long and short-term memory network model (DeepBiLstmCtc) that supports variable-length input to process the features and output the corresponding phoneme sequence. The experimental results show that the algorithm can effectively identify the phoneme sequence in the audiovisual information, and it also has a certain improvement in the recognition rate in the case of noise.

Keywords

Multimode, End-to-End, Speech Recognition, BiLstmCtc

1. 引言

语音是人与人之间最重要的交流方式之一，而自动语音识别(ASR)技术[1]实现了人和计算机之间的交互功能，计算机通过训练以及识别，将语音信息转换成相对应的命令或文本。人类对语音的识别过程实质上是一个多模态过程[2]，其中的特征除了声音外，还涉及说话人口型、面部表情、手势等多个部分。在很多情况下唇部视觉信息与声学信息就有很强的互补性，具体地体现在一些音位的对比上，例如辅音/n/与元音/u/在声学信息的表征力较小[3]，但是在唇部视觉信息的表征上更加显著。

传统语音识别技术比较单一地关注声学信息的研究，特别明显的一个缺点就是单模态语音识别鲁棒性比较差，在比较复杂的环境下，特别是噪音的情况下识别基本处于不可用状态[4]。事实上通过唇动视觉信号和音频信号进行融合，多模态语音识别能较好地应对噪声环境并提高识别的正确率[5]。

本文融合了视觉特征和声学特征，使用基于连接时序分类(CTC)算法[6]的端到端 LSTM [7]模型实现了对音素序列的多模态识别，避免了时域上复杂的切分和强制对齐过程。实验结果表明，本算法可以在避免切分和强制对齐的情况下，有效地完成音素序列的识别功能，并一定程度提高了识别正确率，在复杂环境下也有更强的鲁棒性。

2. 融合唇部特征的语音识别模型

2.1. 实验数据集

本文使用的数据集为 GRID 数据集，该数据集是一个支持语音感知联合计算－行为研究的大型多语言视听句子语料库。由 34 名说话者(18 名男性，16 名女性)每人说出 1000 句话的高质量音频和视频(面部)录音组成。

数据集说话人视频为 mpg 格式，分辨率为 360 × 288，帧率为 25 帧/秒。音频文件为 wav 格式，采样率为 50 kHz，音频长度与视频相同，标注文件为 align 格式，标注文件对单词进行了标注。

2.2. 音频数据预处理及特征提取

语音信号预处理是对语音信号进行转换，使其适合计算机处理，同时符合特征提取的要求。提取出能表示语音信号本质的特征参数，语音识别才可以高效地进行。语音信号预处理包括预加重、分帧和加窗[8]处理。

首先对音频信号进行预加重处理，通过高斯滤波器提高信号中的高频分量，以此提升信号的信噪比，传递函数如式(1)所示，其中μ值大小接近 1，通常取 0.97。然后将信号以 25 ms 帧长实行分帧操作，设置 10 ms 帧移长度，模拟出发音时的连续状态。分帧后帧边缘部分还依旧剧烈变化，因此进行加窗操作如式(2)所示，加窗减少了频谱泄露，使得相邻帧过渡平稳，增加了频谱平滑程度。

$$H(z)=1-\mu z^{-1} \tag{1}$$

$$C(n,a)=(1-a)-a\times\cos\left(\frac{2\pi n}{N-1}\right)\quad(0\le n\le N-1) \tag{2}$$

不同汉明窗 a 值不同，a 值大小通常取 0.46，N 为帧的大小，n 表示第 n 帧。

音频特征提取可以获得比原始语音信号更小的表征力特征。音频特征提取采用 MFCC (Mel Frequency Cepstral Coefficient) [9]算法，如式(3)所示，将线性频谱映射到梅尔尺度频谱上，可以很好的表示音色特征，反映人的听觉特征。

$$\mathrm{MFCC}(t,i)=\sqrt{\frac{2}{N}}\sum_{j=1}^{N}\lg\left[E_{mel}(t,j)\right]\cos\left[i(j-0.5)\frac{\pi}{N}\right] \tag{3}$$

N 为滤波器个数；E_{mel} 为 t 时刻第 j 个滤波器输出的能量；$\mathrm{MFCC}(t,i)\{i=1,2,\cdots,P\}$ 为 t 时刻对应的 MFCC 参数，P 为阶数。

对加窗过的信号进行快速傅里叶变换，得到频谱图以及三维时域图。经过梅尔滤波器组与过离散余弦变换，计算得到 13 维倒谱系数。MFCC 正是由 13 维倒谱系数、13 维一阶差分系数和 13 维二阶差分系数组成的 39 维特征向量，在此语音信息变为 39 维特征向量表征。

2.3. 视频预处理及特征提取

利用爬虫将视频数据集下载解压分类，部分视频存在质量不高，花屏情况，利用级联的残差回归树算法[10] (GBDT)检测视频中是否每帧都存有人脸信息，将质量不达标视频剔除，得到可用的视频数据集。视频预处理及特征提取结构如图 1 所示。

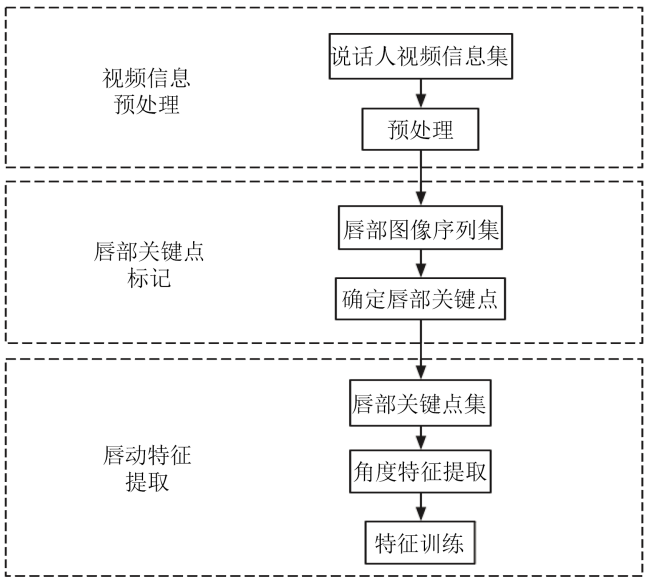

Figure 1. Video preprocessing and feature extraction structure
图 1. 视频预处理及特征提取结构

视频信息预处理按照 25 帧每秒的形式将视频切分为图像集合，对图像进行去噪处理，进行唇部定位，去除非唇部部分完成唇部图像的提取。人脸中唇部变化对语音的影响最为重要，因此选择唇部的参数作为视觉特征的参数，人脸检测(face detection) [11]可以将人脸位置从帧图像中识别出来，对唇部定位有很大帮助。

输入含有关键点信息的唇部图像序列集数据，对图像进行对齐操作，将其转换为特征点形式图像，并将其对齐到基准图像上。调用 dlib 库的 face_landmarks 模型对唇部特征点进行提取，最后得到嘴唇特征点位置坐标。

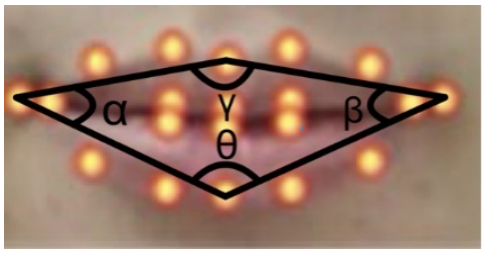

Figure 2. Outer lip angle
图 2. 外唇角度

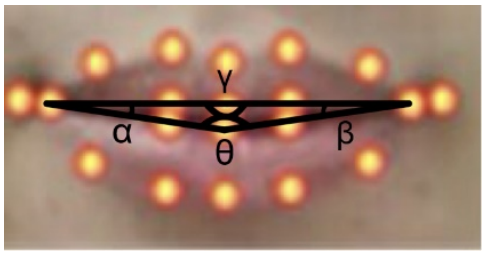

Figure 3. Angle of the inner lip
图 3. 内唇角度

若 α 角对应的点为 a，θ 角对应点为 b，β 对应点为 c，θ 的角度如式(4)所示。$\boldsymbol{ab}$ 表示 a 点指向 b 点的向量，$\boldsymbol{bc}$ 表示 b 点指向 c 点的向量。

$$\theta = \arccos\left(\frac{\boldsymbol{ab}\cdot\boldsymbol{bc}}{|\boldsymbol{ab}||\boldsymbol{bc}|}\right) \tag{4}$$

在唇部一共有 20 个特征坐标点，若全部提取相关信息作为视频特征，会造成特征维度过大，为了既能准确表征视频特征，又能降低特征维度，于是分别对外唇以及内唇提取特征角度，如图 2 与图 3 所示，外唇以及内唇左角度 α 和右角度 β 可以看作相等大小，所以统一取左边角度 α 表征。最后得到内外唇的 α、θ、γ 共 6 个角度表征唇部特征，将唇部图像的角度保存为特征矩阵形式。在噪声不明显的情况下，过多的视觉特征加入反而影响多模态的识别准确率，经过对视觉特征的多次调整，最后选择使用 6 个角度表征视觉特征。

2.4. 特征融合

在时域上将视觉特征和听觉特征进行融合，保证同一时刻音视频信息相互对应。在融合时又分为两种方式，第一种方式是先将特征融合继而统一进行归一化处理，第二种是先将特征分别归一化，最后再融合，两种方式带来的结果差别很大。本文采用的方法为后者，具体流程如图 4 所示。

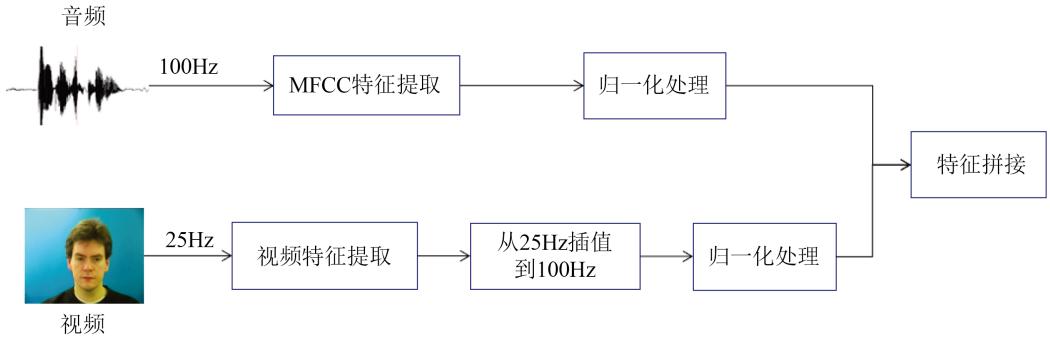

Figure 4. Interpolation fusion process
图 4. 插值融合流程

前文已完成特征提取以及降维操作，由于视频流的速率与音频流并不相等，需要对视频流进行插值操作，让音视频流具有同样的速率。对插值操作后将两个特征分别进行归一化处理，最后进行特征拼接。每一帧的音视频信息都可以分别表示为一个不同特征向量，在融合过程中需要保证两者同一时刻上是相互对应的，利用 numpy 模块的 column_stack 函数，将两个向量拼接为一个特征向量。将帧的个数看作向量的个数，特征拼接之后得到音视频融合的特征矩阵。特征融合很好地解决了时序异步情况，并且可以最大程度保留原始数据，让模型很好学习到两者的特征关系。

2.5. 端到端的多模态发音检测模型

总体模型网络基本流程如图 5 所示：

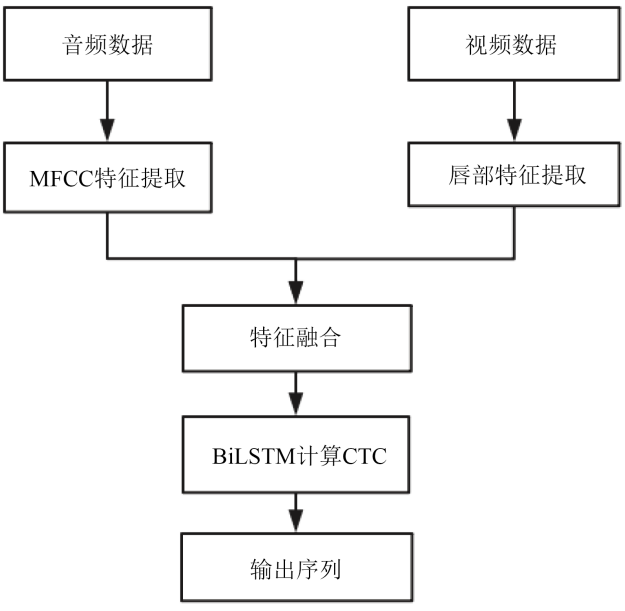

Figure 5. Overall detection model
图 5. 总体检测模型

本文引入循环神经网络，利用视频中人脸和语音的时间序列信息来得到更稳定的识别效果。而 LSTM 是一种特殊的循环神经网络，可以防止梯度消失和梯度爆炸，解决了传统 RNN 的长期依赖的问题。它的结构如图 6 所示。

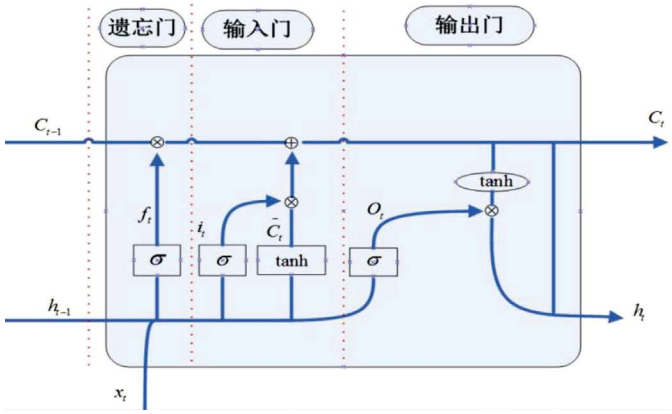

Figure 6. Long short-term memory
图 6. 长短期记忆网络

LSTM 主要由三部分组成，分别是遗忘门，输入门和输出门。其中遗忘门如式(5)所示，σ 表示取 Sigmoid 函数值，W_t 表示权重矩阵，h_{t-1} 表示上一层 LSTM 神经网络的输出，x_t 表示输入，b_f 表示偏置量。f_t 的元素取值范围是 0 到 1，表示遗忘的程度，0 表示全忘，1 表示全记住。输入门如式(6)、(7)所示，其中 i_t 表示输入状态量，c_t 表示对 i_t 的筛选。遗忘门和输入门决定了当前神经网络层的状态信息，即得到式(8)。输出门如式(9)、(10)所示。

$$f_t = \sigma\left(W_t\left[h_{t-1}, x_t\right] + b_f\right) \tag{5}$$

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{6}$$

$$c_t = \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \tag{7}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot c_t \tag{8}$$

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_i\right) \tag{9}$$

$$h_t = o_t \cdot \tanh\left(c_t\right) \tag{10}$$

双向长短期记忆网络(BiLSTM)结构是指在 LSTM 中当前时刻不仅连接到下一时刻还连接到上一时刻，通过信息的双向传递来更好地获得序列间的联系，经过 Alex 等验证在语音识别上取得了比普通单向 LSTM 更好的效果。在人脸和语音的多模态身份识别任务中，需要综合一段时间内的连续信息来进行识别，为了更好地获取一段时间序列中人脸特征和语音特征的上下文联系，本文也将采用 BiLSTM 对人脸语音融合特征序列学习。BiLSTM 结构如图 7 所示。

为了处理连续语音识别问题中的语音序列长短差距比较大的问题，视频与音频的融合特征交给传入到 BiLSTM 网络中，并且计算 CTC (连接时序分类)损失函数，最后根据 CTC 函数求得的值传入一个到全连接层，求得最大可能性的字符，这样就完成了融合特征序列到文本序列的一个转换，得到语音识别的结果。

CTC 主要解决的问题是，在给定输入序列 X 及其对应的标签序列 Y 前提下，如何将 X 和 Y 一一对应。CTC 损失函数的训练通常是利用 RNN 来实现，因为 RNN 对序列型数据训练的效果比较好。对于单个的输入输出匹配(X,Y)来说，CTC 的计算如式(11)所示，$\prod_{t=1}^{T} p\left(a_t \mid X\right)$ 是对单个路径概率进行计算，模型训练的目的就是将 $P\left(Y \mid X\right)$ 即 Y 的路径之和最大化。CTC 损失函数如式(12)所示。对于数据集 G，模型的优化的目的是将 CTC 损失函数最小化。

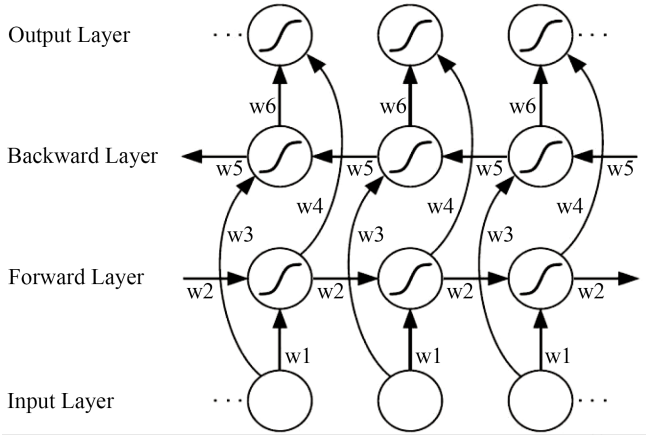

Figure 7. Bi-directional long-short term memory
图 7. 双向长短期记忆网络

$$P(Y \mid X) = \sum_{A \in A_{x,y}} \prod_{t=1}^{T} p(a_t \mid X) \tag{11}$$

$$Loss(\mathrm{CTC}) = \sum_{(X,Y) \in G} -\log p(Y \mid X) \tag{12}$$

3. 实验结果及分析

3.1. 实验配置

实验环境：1、win10+Ubuntu18.42、python3.63、tensorflow-gpu==1.14.0　4、gtx1080

本次实验采用的数据集为 GRID 数据集，语料库类型为标准英语，训练集与测试集的比例为 9:1，两者没有重叠部分。每层 LSTM 均含有 150 个单元，实验前对 2 到 5 层的 LSTM 组成的模型进行对比，发现选取 4 层效果最佳，后续模型统一为 4 层 LSTM。Batch size 大小设置为 64，epoch 设置为 500，learningrate 设置为 0.01。还有一个在噪声环境下的 GRID 数据集，对原有的数据集音频加上了的高斯白噪声。图 8 和图 9 分别是无噪声下的音频时域和频域图和高斯白噪声下的音频时域和频域图。

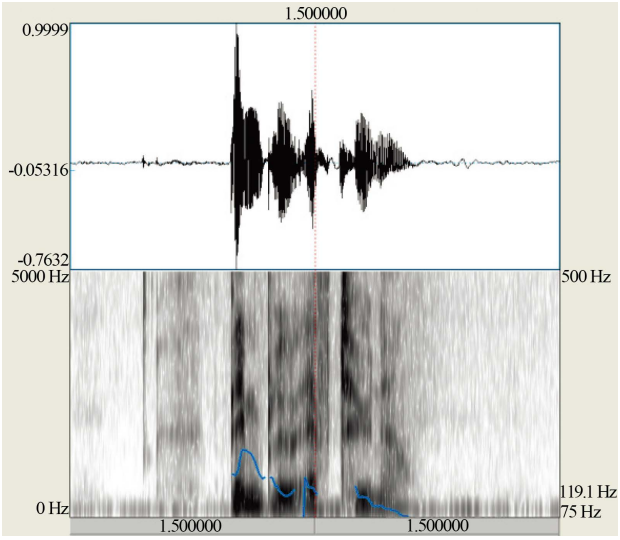

Figure 8. Audio in a noiseless environment
图 8. 无噪声环境下的音频

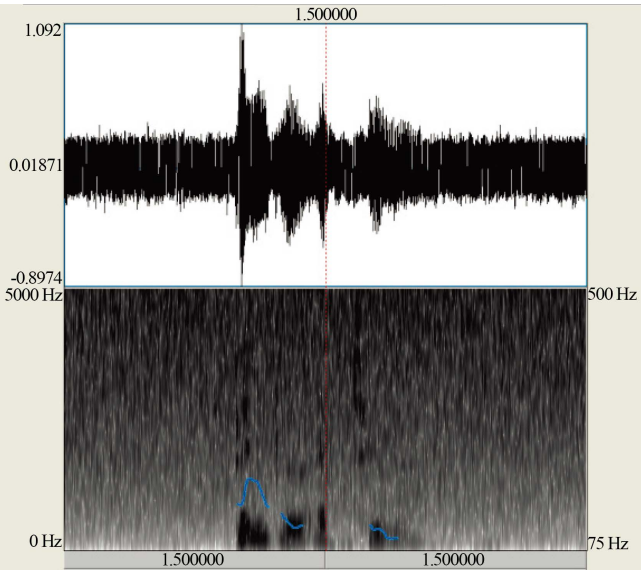

Figure 9. Audio under Gaussian white noise
图 9. 高斯白噪声下的音频

3.2. 评价指标

计算实际序列与预测序列之间的平均 Levenshtein 距离，作为实验中的准确性指标，在结果图中用 Ler 表示，如式(13)所示。

Levenshtein 距离，是指两个字符串之间，从一个字符串变成另一个字符串所需要的最小编辑操作次数。可以采用的编辑操作包括：插入操作、替换操作和删除操作。

$$\text{Ler} = \frac{\text{序列的Levenshtein距离}}{\text{实际序列长度}} \tag{13}$$

测试集的准确率为所有序列准确率的平均值，如式(14)所示。

$$\text{TestLer} = \frac{\sum \text{Ler}}{\text{n}} \times 100\% \tag{14}$$

3.3. 实验结果及分析

为了对比算法的结果，分别在有无噪音环境下将单模态的语音 mfcc 处理训练结果与多模态训练结果进行比较。为了验证归一化方式的不同，对两种归一化在无噪声情况下也进行了对比。由图 10 和图 11 可以得知先归一化后特征融合 loss 值下降地更快，并且拥有更低的错误率，更高的识别率，原因为两个特征差异巨大，直接融合再归一化会互相影响数据的准确性。如表 1 所示，融合了唇部特征的多模态 LSTM 无论在有噪音还是无噪音情况下都比音频 LSTM 错误率更低，拥有更高的识别率。

Table 1. Comparison of experimental results
表 1. 实验结果比较

有无噪声 声学模型	错误率(%)	
	无噪声	有噪声
单模态 LSTM 模型	3.72328	47.7366
多模态 LSTM 模型	2.97117	36.4322

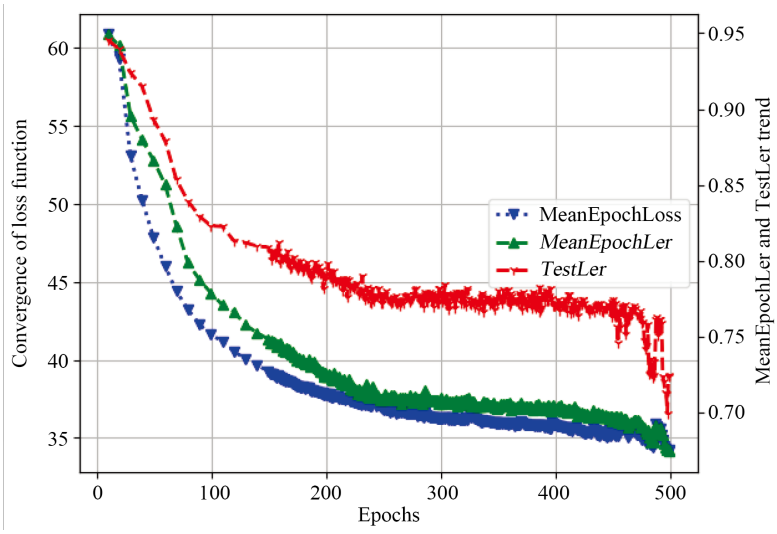

Figure 10. Normalize after fusion
图 10. 先特征融合后归一化

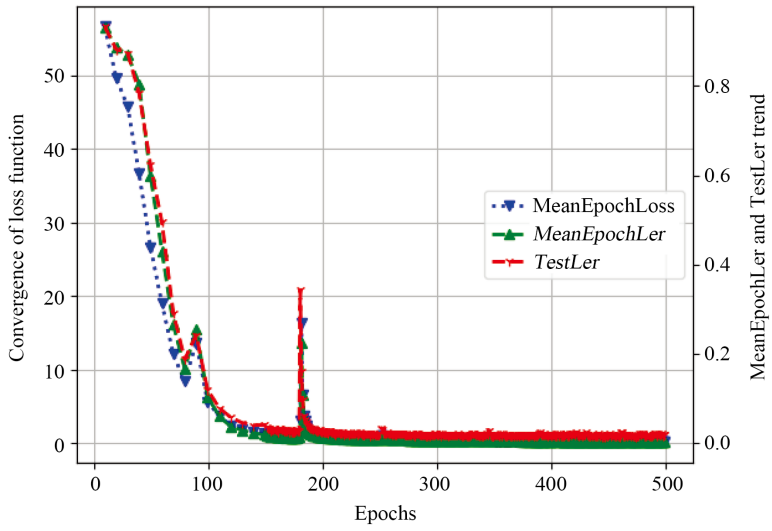

Figure 11. Normalize first and then merge
图 11. 先归一化再特征融合

4. 总结

本文提出了一种融合唇部特征的端到端的多模态语音识别算法，实验结果表明该算法能有效地识别出视听觉信息中的音素序列，在噪声情况下也有一定的识别率提升。该算法具有更强的鲁棒性，能应对更加复杂的环境，具有比较实际的意义，在以后的工作中会继续提高算法的识别准确率，并且加入纠错功能。

参考文献

[1] 王海坤，潘嘉，刘聪.语音识别技术的研究进展与展望[J]. 电信科学, 2018, 34(2): 1-11.

[2] 赵荣刚，贺庆民. 计算机人脸识别技术的应用[J]. 电子技术与软件工程, 2018(4): 137.

[3] 徐彦君，杜利民，侯自强. 面向未来的交互信息技术——听觉视觉双模态语音识别(AVSR) (上) [J]. 电子科技导

报, 1999(1): 26-30+34.

[4] Massaro, D.W. and Stork, D.G. (1998) Speech Recognition and Sensory Integration: A 240-Year-Old Theorem Helps Explain How People and Machines Can Integrate Auditory and Visual Information to Understand Speech. *American Scientist*, **86**, 236-244. https://doi.org/10.1511/1998.25.861

[5] 田春霖. 深度视音频双模态语音识别方法[D]: [硕士学位论文]. 西安: 中国科学院大学(中国科学院西安光学精密机械研究所), 2018.

[6] 李明浩. 基于深度神经网络的连续语音识别研究[D]: [硕士学位论文]. 长春: 吉林大学, 2018.

[7] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[8] Wongeun, O. (2020) Comparison of Environmental Sound Classification Performance of Convolutional Neural Networks According to Audio Preprocessing Methods. *The Journal of the Acoustical Society of Korea*, **39**, 143-149.

[9] 郭春霞, 裘雪红. 基于 MFCC 的说话人识别系统[J]. 电子科技, 2005(11): 55-58.

[10] 于维生. 最小残差绝对和回归模型参数的递推估计方法[J]. 中国管理科学, 1995(2): 49-55.

[11] 梁路宏, 艾海舟, 徐光祐, 张钹. 人脸检测研究综述[J]. 计算机学报, 2002, 25(5): 449-458.