



研究与开发

基于 SDBN 和 BLSTM 注意力融合的 端到端视听双模态语音识别

王一鸣, 陈恳, 萨阿卜杜萨拉木·艾海提拉木
(宁波大学信息科学与工程学院, 浙江 宁波 315211)

摘要: 提出一种端到端的视听语音识别算法。在该算法中, 通过具有瓶颈结构的深度信念网络 (deep belief network, DBN) 中引入混合的 $l_{1/2}$ 范数和 l_1 范数构建一种稀疏 DBN (sparse DBN, SDBN) 来提取稀疏瓶颈特征, 从而实现对数据的特征降维, 然后用双向长短期记忆网络 (bidirectional long short-term memory, BLSTM) 在时序上对特征进行模态处理, 之后利用一种注意力机制将经过模态处理的唇部视觉信息和音频听觉信息进行自动对齐、融合, 最后将融合的视听觉信息通过一个附加了 Softmax 层的 BLSTM 进行分类识别。实验表明, 该算法能有效地识别视听觉信息, 在同类算法中有很好的识别率和顽健性。

关键词: 端到端; 视听语音识别; 稀疏瓶颈特征; 注意力机制

中图分类号: TP391

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019290

End-to-end audiovisual speech recognition based on attention fusion of SDBN and BLSTM

WANG Yiming, CHEN Ken, ABUDUSALAMU Aihaiti
Institute of Communication Technology, Ningbo University, Ningbo 315211, China

Abstract: An end-to-end audiovisual speech recognition algorithm was proposed. In algorithm, a sparse DBN was constructed by introducing mixed $l_{1/2}$ norm and l_1 norm into Deep Belief Network with bottleneck structure to extract the sparse bottleneck features, so as to reduce the dimension of data features, and then a BLSTM was used to model the feature in time series. Then, a attention mechanism was used to align and fuse the lip visual information and audio auditory information automatically. Finally, the fused audiovisual information was classified and identified by a BLSTM with a Softmax layer attached. Experiments show that the algorithm can effectively identify visual and auditory information, and has good recognition rate and robustness in similar algorithms.

Key words: end-to-end, audiovisual speech recognition, sparse bottleneck features, attention mechanism

收稿日期: 2019-06-24; 修回日期: 2019-12-10

基金项目: 国家自然科学基金资助项目 (No.60972063); 宁波市自然科学基金资助项目 (No.2014A610065); 宁波大学科研基金 (理) /学科资助项目 (No.XKXL1308)

Foundation Items: The National Natural Science Foundation of China (No.60972063), The Natural Science Foundation of Ningbo of China (No.2014A610065), Scientific Research Foundation of Ningbo University (No.XKXL1308)



1 引言

在信息技术飞速发展的今天, 智能人机交互技术一直是人们所热衷研究的课题, 语言是人类交流最自然、有效的方式, 因此, 自动语音识别^[1] (automatic speech recognition, ASR) 是智能人机交互重要的组成部分。人与人之间的交流是多通道的, 说话人之间可以通过言语、口型以及肢体语言等来捕获对方表达的信息, 但主要通过声音信号进行交流。当在一些安静环境中, 声音信号承载了绝大部分信息, 在最近的 ASR 系统^[2]中, 识别率能达到 95% 以上。但是, 当在一些多说话人交叉或噪声环境中, 效果就会大打折扣, 此时, 唇动视觉信号就显得格外重要, 因为视觉信号可以不受噪声影响, 以此来“恢复”受损的音频信号。通过唇动视觉信号和音频信号融合进行双模态语音识别的系统, 又叫视听语音识别 (audio visual speech recognition, AVSR)。

传统的 AVSR 系统中, 主要分为 3 部分: 唇部视觉信号和音频信号的特征提取, 视听觉信息的融合以及分类识别。随着最近深度学习^[3]的大热, 很多使用深度学习的视听语音识别方法被提出。如 Noda 等人^[4]分别用 DBN 和 CNN 提取视听觉特征, 之后用一种多流隐马尔可夫模型 (MSHMM) 来对双模态特征进行融合、识别。参考文献[5]则使用一种循环时间多模态 RBM (RTMRBM) 来融合学习视听特征表示, 这种结构通过将每个 MRBM 转换为概率序列模型来实现对视听双模态序列的学习。参考文献[6]通过堆叠多个不同类型的网络层 (卷积层、池化层、瓶颈层) 形成的 CBN 分别对视听觉信息进行处理, 之后用 GMM-HMM 来融合、分类识别。虽然这些深度学习方法都很有效, 但 AVSR 是一个复杂的系统, 需要分别训练每个模块组件, 这使得任一组件训练的好坏都会影响最终结果, 所以近年来, 端到端的深度学习方法越来越受到青睐。如

牛津大学的 Chung 等人^[7]构建一种端到端的视听语音识别模型, 该模型用 LSTM 在时序上对视听特征进行编码, 之后再用一个附加了两个注意力模型的 LSTM 对特征进行融合、分类识别, 该方法取得了不错的效果, 但是在音频特征初提取用的是 MFCC 特征, 并没有使用原始信号或者音频频谱图, 所以信息有所缺失。参考文献[8]则使用多个 BGRU (bidirectional gated recurrent unit, BGRU) 来进行端到端的视听语音识别, 文中分别使用残差网络 (residual network, ResNet) 进行特征的提取, 之后用两个 BGRU 在时序上对特征进行模态处理, 最后用一个相同结构的 BGRU 来融合双模态视听信息。参考文献[9]也是用多个堆叠的网络层进行端到端视听语音识别, 其中堆叠层由一层全连接层和 Dropout 层组成, 在该端到端模型中, 分别用一个堆叠层对唇部信息和音频信息进行处理, 之后再用一个单独的堆叠层进行视听双模态信息的融合, 实验表明该模型有很好的识别率。在以上端到端的视听觉语音识别模型中, 其视听双模态信息的融合都是使用与特征处理层有相同结构的网络层进行融合, 这种融合方式为了达到音频序列和唇部视频序列的在时序上的同步, 都只是简单地对低帧率的视频流进行上采样以获得与音频相同的帧率^[10]。

本文提出一种基于注意力机制的视听双模态信息自动对齐融合的算法, 该算法首先在具有瓶颈结构的深度信念网络中引入 $l_{1/2}$ 范数和 l_1 范数, 通过引入混合范数来实现 SDBN, 这样不仅能更好地将高维数据转换为低维瓶颈特征表示, 还能使提取的特征具有更强的泛化性和顽健性, 再用 BLSTM 在时序上对特征进行模态处理, 视听流分别用 SDBN 和 BLSTM 处理之后, 引入一种注意力机制将视听觉信息在模态上自动对齐融合, 最后将融合的序列信息送入附加了 Softmax 层的 BLSTM 进行分类识别。实验表明, 该算法可有效地识别视听觉语音信息。

2 算法框架

本文算法框架模型如图 1 所示, 主要步骤如下。

步骤 1 预处理部分。首先对唇部视频帧序列进行预处理, 将视频帧裁剪为合适的统一尺寸, 并对其进行去均值和归一化处理; 同时对原始语音信号进行预加重、分帧、加窗、短时傅里叶变换得到其声谱图, 最后经过 Mel 滤波器组生成 Mel 频谱序列。

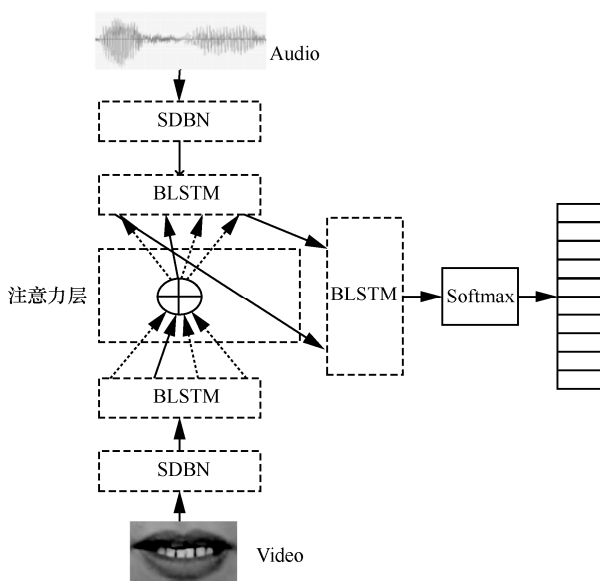


图 1 端到端视听双模态语音识别模型

步骤 2 提取稀疏瓶颈特征并进行模态处理。通过在具有瓶颈结构 DBN 的目标函数中引入混合的 $l_{1/2}$ 范数和 l_1 范数来构建 SDBN, 分别提取视听觉信息的稀疏瓶颈特征, 之后再各通过一个 BLSTM 在时序上对特征进行模态处理。

步骤 3 对视听觉信息进行融合。为将视听流 BLSTM 输出的时序不一致的视听双模态向量序列融合, 引入一种注意力机制, 通过计算每一时序上音频流状态向量与所有唇部视觉流输出向量的匹配度, 之后将其与视觉流 BLSTM 的输出做一种线性组合并与当前音频流 BLSTM 输出的序列用一连接层融合, 这样来自动对齐不同模态

的视听觉信息并融合, 以提取出更高级的融合向量序列, 便于后续分类识别。

步骤 4 分类模型的设计。使用附加了 Softmax 层并与步骤 2 具有相同结构的 BLSTM 对融合的序列信息进行分类识别。

3 基于 SDBN 和 BLSTM 注意力融合的视听觉语音识别

3.1 稀疏深度信念网络

3.1.1 DBN

深度信念网络 (deep belief network, DBN) 是由 Hinton 等人^[11]提出的, 它是由多个受限玻尔兹曼机 (restricted Boltzmann machine, RBM) 堆叠而成, 每个 RBM 又分为显层和隐层。整个 DBN 自下而上呈瓶颈结构, 顶层 (瓶颈层) 神经元最少, 其输出的特征又叫瓶颈特征 (bottleneck feature)^[12]。典型的 DBN 结构图如图 2 所示。

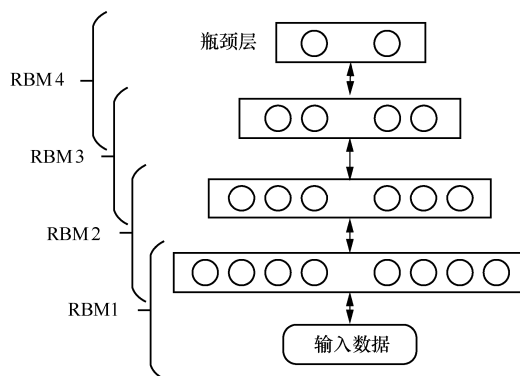


图 2 典型的 DBN 结构示意图

对于一个典型的 RBM, 能量函数定义如下:

$$E(v, h | \theta) = - \sum_{i=1}^I \frac{(v_i - \alpha_i)^2}{2\sigma_i^2} - \sum_{j=1}^J b_j h_j - \sum_{i=1}^I \sum_{j=1}^J \frac{v_i h_j}{\sigma_i} w_{ij} \quad (1)$$

其中, v_i 和 h_i 分别表示显层神经元和隐层神经元, $\theta \triangleq \{W, b, a\}$ 是一种模型连接参数 (可看作权重 w 和偏置量 a, b 的一种组合), σ 为一种高斯噪声标准差。

由 RBM 的能量函数可得显层与隐层神经元



之间的映射概率:

$$p(v_i = 1 | h) = \frac{1}{1 + \exp\left(-\sum_{j=1}^J w_{ij} h_j - \alpha_i\right)} \quad (2)$$

$$p(h_j = 1 | v) = \frac{1}{1 + \exp\left(-\sum_{i=1}^I w_{ij} \frac{v_i}{\sigma_i} - b_j\right)} \quad (3)$$

为求得模型连接参数 θ , 使用梯度上升法来极大化概率的对数似然函数:

$$\frac{\partial \log p(v, h | \theta)}{\partial \theta} = \left\langle \frac{\partial E_{\theta}(v, h)}{\partial \theta} \right\rangle_{\text{model}} - \left\langle \frac{\partial E_{\theta}(v, h)}{\partial \theta} \right\rangle_{\text{data}} \quad (4)$$

在式 (4) 中, $\langle \rangle_{\text{data}}$ 和 $\langle \rangle_{\text{model}}$ 表示一种期望。在实际中, 该期望难以计算, 一般采用对比散度算法近似计算, 模型连接参数中权重和偏置更新如下:

$$\Delta w_j = \frac{1}{L} \sum_{l=1}^L v^l p(h_j = 1 | v^l) - \tilde{v}^l p(h_j = 1 | \tilde{v}^l) \quad (5)$$

$$\Delta a_i = \frac{1}{L} \sum_{l=1}^L p(v_i^l = 1 | h) - p(\tilde{v}^l = 1 | h) \quad (6)$$

$$\Delta b_j = \frac{1}{L} \sum_{l=1}^L p(h_j = 1 | v^l) - p(h_j = 1 | \tilde{v}^l) \quad (7)$$

3.1.2 SDBN

神经学认为, 大脑皮层细胞对于刺激的处理过程本质上是稀疏的, 即大多数神经元在处理中处于不激活状态。基于这些研究, 研究者们通过模拟大脑皮层对于刺激的处理过程, 在深度学习的算法中加入稀疏惩罚项来实现神经网络的稀疏性。如 Ji 等人^[13]提出了一种 DBN 的变体, 通过在其定义的一种失真函数中引入 l_1 范数作为惩罚项, 使得能以较小的码率实现对图像特征的提取。参考文献[14]也是在 DBN 的目标函数中加入一种惩罚项 R (条件期望) 来实现其稀疏表示, 使得所由隐层神经元有一个较小的平均激活概率。还可以通过在卷积神经网络中引入一种 F 范数 (Frobenius norm) 作为其约束项^[15]。Khalid 等人^[16]则提出一种由方差和正态概率密度函数组成的正态稀疏惩罚项。通过在神经网络中加入稀疏惩罚项来实现对数据特征的稀疏表示, 这种稀疏表示

不仅可以有效规避维度灾难, 还能提高目标函数的泛化性, 使得网络对于输入的不同数据有更强的稳健性, 降低其结构复杂度。

受以上方法启发, 提出一种稀疏深度信念网络 (sparse deep belief network, SDBN), 通过在 DBN 的目标函数后引入混合的 $l_{1/2}$ 范数和 l_1 范数实现其稀疏表示。SDBN 的目标函数如下:

$$L = -\ln p(v) + \lambda_1 (\|p(h=1|v)\|_{l_{1/2}}) + \lambda_2 (p(h=1|v)\|_1) \quad (8)$$

其中, λ_1 和 λ_2 是正则化常数。

将非重叠组套索 (non-overlapping group lasso) 的方法用于 DBN, 即将 RBM 的隐层单元 h_j 均等地划分为 K 个互不重叠的组 P_k ($k=1, 2, \dots, K$)。对于式 (8) 等号右边第二项, 引入混合的 $l_{1/2}$ 范数实现对非重叠组级别上的惩罚, 即让大部分组中所有神经元不激活, 实现组级别的稀疏表示, 所以可将其写为:

$$\|p(h=1|v)\|_{\frac{1}{2}} = \sum_{k=1}^K \|p(h_{P_k}|v)\|_2 = \sum_{k=1}^K \sqrt{\sum_{i \in P_k} p(h_i=1|v)^2} \quad (9)$$

对于式 (8) 等号右边第三项, 以 l_1 范数对个体级别上的神经元进行惩罚, 实现在个体级别上的稀疏表示, 该项可写为:

$$\|p(h=1|v)\|_1 = \sum_{i=1}^L |p(h_i=1|v^l)| \quad (10)$$

之后如第 3.2.1 节一样, 为求模型连接参数 θ , 同样使用梯度上升法来极大化概率的对数似然函数, 所以稀疏惩罚项可改写为:

$$\begin{aligned} \frac{\partial}{\partial w_j} \left(\|p(h=1|v)\|_{\frac{1}{2}} \right) &= \frac{1}{2} \frac{1}{\sqrt{\sum_{m \in P_k} p(h_m=1|v^l)^2}} \cdot \\ 2p(h_m=1|v^l) \frac{\partial p(h_m=1|v^l)}{\partial w_j} &= \\ \frac{p(h_m=1|v^l)}{\|p(h_k|v^l)\|_2} \cdot \frac{\partial p(h_m=1|v^l)}{\partial w_j} &= \\ \frac{p(h_m=1|v^l)^2}{\|p(h_k|v^l)\|_2} \cdot p(h_m=0|v^l)v^l & \end{aligned} \quad (11)$$

$$\frac{\partial}{\partial w_j}(\|p(h=1|v)\|_1) = p(h_m=1|v'). \quad (12)$$

$$v'[1 - p(h_m=1|v')] = p(h_m=1|v')v'p(h_m=0|v')$$

由此可得 SDBN 的 θ , 更新计算式如下:

$$\Delta w_j = \frac{1}{L} \sum_{l=1}^L [v' p(h_j=1|v') + \lambda_1 v' \cdot \frac{p(h_j=1|v')p(h_j=0|v')}{\sqrt{\sum p(h_k=1|v')^2}} + \lambda_2 v' p(h_j=1|v')]. \quad (13)$$

$$p(h_j=0|v') - \tilde{v}' p(h_j=1|v')]$$

$$\Delta a_j = \frac{1}{L} \sum_{l=1}^L [p(h_j=1|v') + \lambda_1 \frac{p(h_j=1|v')p(h_j=0|v')}{\sqrt{\sum p(h_k=1|v')^2}} +$$

$$\lambda_2 p(h_j=1|v')p(h_j=0|v') - p(h_j=1|\tilde{v}')] \quad (14)$$

对于隐层的偏置 Δb_j 而言, 其更新公式与式 (14) 相同。

3.2 BLSTM

BLSTM^[17]是一种时间上的由正向序列和反向序列组成的双向递归神经网络 (BRNN), 它在 RNN 的基础上引入了记忆胞元, 该记忆胞元主要包括三部分: 输入门、输出门及遗忘门, 通过这三种门结构可以使得网络在时序上保持和更新胞元状态, 从而解决 RNN 的梯度消失问题, 而且双向的 LSTM 可以结合过去和未来的上下文信息, 实现优势互补, 使得模型能更好地对时序信息进行处理。

自前向后的正向隐层更新公式为:

$$\vec{h}_t = H(W_{xh}x_t + W_{hh}\vec{h}_t - 1 + b_h) \quad (15)$$

自后向前的反向隐层更新公式为:

$$\tilde{h}_t = H(W_{xh}x_t + W_{hh}\tilde{h}_t - 1 + b_h) \quad (16)$$

综合了网络的正反向隐层信息的输出层更新公式为:

$$y_t = W_{hy}\vec{h}_t + W_{hy}\tilde{h}_t + b_y \quad (17)$$

BLSTM 结合了 BRNN 与 LSTM 的优点, 它是由 BRNN 中的隐层单元由 LSTM 单元替换

即可组成。图 3 是沿着时间展开的 BLSTM 结构图。

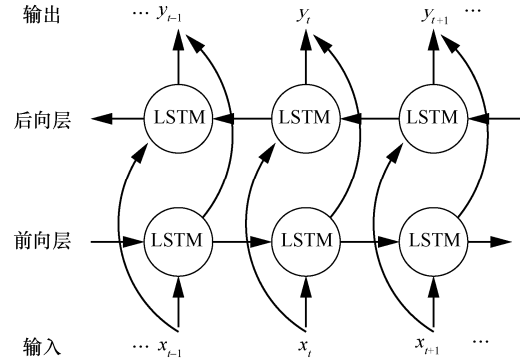


图3 BLSTM结构示意图

3.3 基于注意力的视听双模态信息融合机制

注意力 (attention) 机制源于人脑的视觉注意力机制, 它是指在某时刻人脑着重对有“吸引力”的区域分配更多的注意力资源, 也就是所谓的焦点, 利用这种机制人们能快速的定位感兴趣的区域。在 AVSR 中, 视听双模态信息融合中的异步问题一直是个难点, 因为在发声中, 唇部视觉信息总是先于语音信息, 在时序上它们是无法对齐融合的。基于此, 在视听双模态信息融合中引入一种基于注意力机制^[18]的模式融合方式, 这种融合方式发生在视听觉特征序列进行模态处理之后, 通过注意力机制从视觉流 BLSTM 的输出中“软选择”与当前音频流 BLSTM 隐层状态量相匹配的序列, 实现视听不同模态向量序列的自动对齐、融合。注意力层的结构如图 4 所示。

首先, 通过一种对齐函数将当前 j 时刻音频流 BLSTM 的隐层状态量 h_j 和唇部视觉流 BLSTM 的每一隐层状态量 h_i 进行“注意力对齐”, 得到对齐向量 (或者称注意力权重) a_i , 也可以说是视听流模态向量序列之间的一种匹配分数:

$$a_j(i) = \text{align}(h_j, h_i) = \frac{\exp(\text{score}(h_j, h_i))}{\sum_i \exp(\text{score}(h_j, h_i))} \quad (18)$$

其中:

$$\text{score}(h_j, h_i) = h_j^T W_a h_i \quad (19)$$

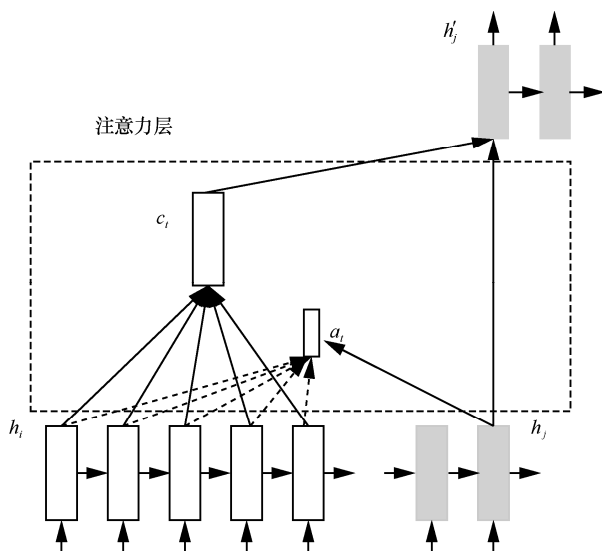


图4 注意力模型架构

之后得到文本向量 c_j :

$$c_j = \sum_{i=1}^I a_{ji} h_i \quad (20)$$

通过计算 j 时刻的音频流 BLSTM 的隐层状态量与所有视觉流 BLSTM 的输出作“匹配”，得到一种匹配分数 a_t ，之后将其与视觉流 BLSTM 的输出做一种线性组合，得到对应 j 时刻音频流 BLSTM 的文本向量 c_j 。在音频流每一时刻输出的向量序列中，都会通过对齐函数在视觉流 BLSTM 的输出序列找到与之匹配的文本向量，通过一个连接层 ($h'_j = \tan h(W_c[c_j, h_j])$) 融合文本向量和音频流的输出状态向量。通过这种注意力机制的方式将视觉流 BLSTM 的模式输出融合到音频流上，此时的音频流 BLSTM 的输出不只包含音频序列信息，它是一种更高级的视听融合序列表示，从而能更好地在分类模型中进行处理。

3.4 分类模型的设计

使用注意力机制融合不同模态的视听觉信息后，采用一个附加了 Softmax 层的 BLSTM 分类模型，其中 BLSTM 尺寸为 450。基于注意力融合后的向量序列 h' 送入 BLSTM 进行处理得到 y_t ，之后通过 Softmax 层，完成从输入 h' 到输出类别 Y

的映射。

$$p(Y = p | y) = \frac{\exp(y_t^p)}{\sum_{p'=1}^{10} \exp(y_t^{p'})} \quad (21)$$

其中， p 表示标签的种类 (OuluVS2 数据库中共有 10 类短语，所以标签取值范围为 1~10)，该概率公式表明在输入序列为 y 的情况下，输出类别为第 p 类的概率，最终取预测概率最大的标签种类作为分类的最终结果。

4 实验设计与分析

4.1 数据库

为验证算法的有效性，在芬兰奥卢大学所建的 OuluVS2 数据库^[19]中进行了实验。该数据库中共有 53 个受试者 (第 29 个受试者因为裁剪后的唇部视频中大部分时间看不见嘴部，所以不采用)，每个受试者说 10 个日常短语，重复 3 次，且分为 5 个角度同时进行拍摄，分别是 0° (正面)、30°、45°、60°、90°，所以共有 780 种嘴唇视频。数据库短语见表 1。

表1 OuluVS2 数据库短语

编号	短语
1	Excuse me
2	Good bye
3	Hello
4	How are you?
5	Nice to meet you
6	See you
7	I am sorry
8	Thank you
9	Have a good time
10	You are welcome

图 5 为数据库中 1 号受试者录制 5 个角度的唇部视频某 5 帧。



图5 OuluVS2 中 5 个角度的唇部视频帧

4.2 训练细节与评估标准

在 OuluVS2 数据库中,将受试者编号为 1~53,在实验中将数据库划分为 3 部分:训练集、验证集、测试集,其中随机选取的受试者视频个数分别为 35、5、12,具体细节见表 2。

表 2 OuluVS2 数据集的划分

数据集	受试者编号
训练集	1、2、3、5、7、10、11、12、14、16、17、18、19、20、21、23、24、25、27、28、31、32、33、35、36、37、39、40、41、42、45、46、47、48、53
验证集	4、13、22、38、50
测试集	6、8、9、15、26、30、34、43、44、49、51、52

首先,分别对音视频信息进行预处理。对于唇部视频,为了在训练时尽量得到其原始信息,所以先将视频裁剪为同一尺寸,之后计算所有视频帧的统计均值,让每个样本减之,以实现图像帧的去均值处理,最后将处理后的图像数据进行 Z-score 归一化处理,以减少唇部视频帧序列之间的共性,凸显个体样本差异。对于音频原始信号,将它由一维信号转换为包含丰富时频域信息的二维信号。首先进行预加重处理,将音频信号通过一个高通滤波器,滤出低频成分,突出高频信息,之后对其分帧、加窗处理,取帧长 40 ms,为了减少丢失音频信息,设置帧移 10 ms,即使相邻帧之间重叠 30 ms。然后在时域上堆叠进行了短时傅里叶变换(short-time Fourier transform, STFT)的每一帧,获取到频谱图,最后将频谱送入 Mel 滤波器组,得到相应的 Mel 频谱序列。

其次,使用 SDBN 和 BLSTM 分别对预处理后的视听觉信息进行稀疏瓶颈特征的提取以及模态处理。对于 SDBN,它由 3 个隐层以及线性瓶颈层(顶层)组成,其结构尺寸分别为 2 000、1 000、500、50,使用 ReLU 作为激活函数。通过非监督自下而上的逐层对 SDBN 进行预训练来获得其初始权重和偏置(共训练 20 个 epoch),学习率设

0.001^[20],动量因子从 0.5 增加到 0.9。正则化常数 λ_1 和 λ_2 设为 0.1、0.01,同时取非重叠组的尺寸为 10。模型中所有 BLSTM 的尺寸都设为 450,使用 Xavier 方法^[21]初始化权重。对于整个端到端模型中,我们使用 Adam 算法对整体进行训练,训练重复 10 次,每次训练总共迭代 20 个轮(epoch),每个 epoch 细分为 105 个批次(batch),批次大小(batchsize)为 10,整体学习率为 0.0001,为了防止过拟合,训练的时候使用了早停法(Early Stopping)。

使用识别率作为算法的评估标准,即分类正确的预测样本标签占总样本标签的比例,同时以混淆矩阵做辅助评估。所有实验都在 Ubuntu16.04LTS (64 bit)操作系统上完成,使用 Python 的发行版本 Miniconda 和 MATLAB R2016a 作为开发工具,整个视听语音识别模型用到了 Theano 深度学习库。

4.3 实验结果与分析

在表 1 中对 OuluVS2 数据库中的短语进行了 1 到 10 标号。在实验中,首先分别对以 0°、30°、45°、60°、90°的唇部视频及其对应的音频进行了单流的端到端模型的实验,之后取得单流中最佳模型(best model)中的参数,以此为基础进行各角度唇部视频和音频的视听双模态融合语音识别实验。图 6 单流及视听双流融合实验所得最终分类结果中的混淆矩阵(confusion matrices),将混淆矩阵以热力图(Heat Map)的形式呈现。在热力图中,纵轴 1 到 10 代表数据库中的真实标签,横轴代表实验所得的预测标签值,其中的对角线的亮度越高代表分类的结果越好。由热力图可看出,在单流唇部视频和音频的实验中,比较容易混淆的是“Hello”(3 号)和“Thank you”(8 号),此外会混淆的还有“Thank you”和“See you”(6 号)、“How are you”(4 号)和“Thank you”,而且当考虑双流视听融合时,亮度集中于热力图中的对角线,这表明混淆的数量减少,算法识别性能的提高。

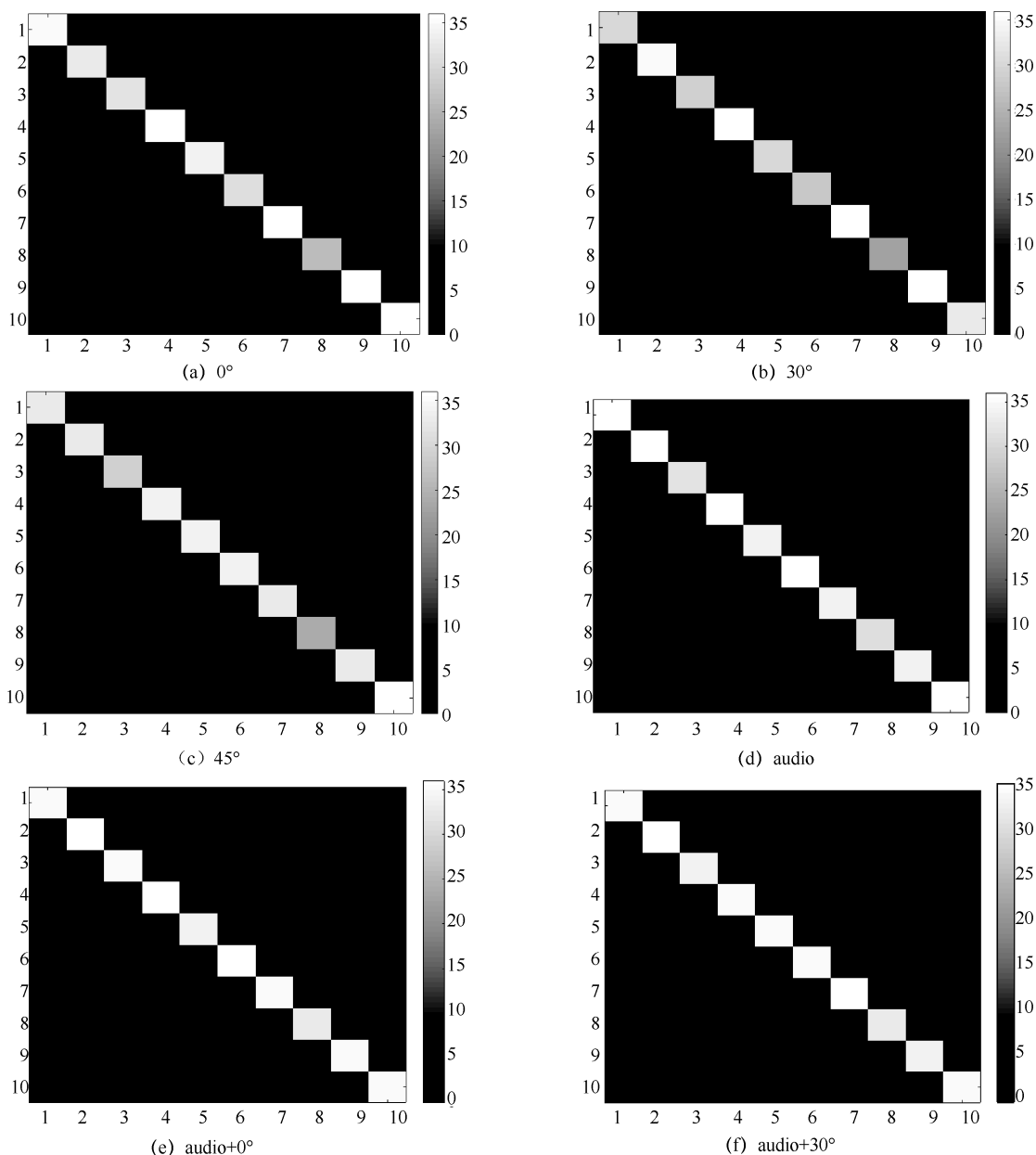


图6 热力图

本文算法在 5 个不同角度的单流唇部视频、音频及其双流融合后的结果见表 3(A 表示纯音频实验, A+0°表示两者融合实验)。对于单流模型实验, 音频的识别率最高, 优于所有角度的唇部视频识别结果, 这是因为音频承载了绝大部分的信息。5 个角度唇部视频中以正面录制的视频包含最多信息, 识别率最高。在视听融合实验中, 将提出的算法与当前标杆算法^[22,23]进行了比较, 可以看到即使单流音频结果与其接近, 但音频分别与

0°、30°、45°、60°、90°唇部视频进行注意力融合后的识别率均优于当前标杆算法, 其中音频与正面唇部视频的融合实验中识别率高达 0.991。这一实验结果也与热力图所得结果相符, 证实提出的视听融合识别模型的有效性。同时在图 7 中, 画出了在测试集上各受试者在 10 次重复训练实验中单流 0°唇部视频、音频及两者融合的双流实验的分类准确率, 可以观察到单流实验而言, 音频流识别率均优于 0°唇部视频, 其中 9 号与 30 号受试

表 3 本文算法与其他算法在单流及双流上的识别率比较

算法	0°	30°	45°	60°	90°	A	A+0°	A+30°	A+45°	A+60°	A+90°
参考文献[22]算法	94.7%	89.7%	90.6%	87.5%	93.1%	98.5%	98.6%	98.7%	98.3%	98.6%	98.9%
参考文献[23]算法	90.0%	89.2%	89.7%	87.2%	84.2%	86.39% (-5 dB)	89.7%	90.3%	89.4%	86.7%	85.3%
本文算法	95.1%	90.0%	90.3%	87.8%	93.7%	98.4%	99.1%	98.9%	98.6%	98.8%	98.9%

者的语音识别达到 100%。同时,除了 30 号,所有受试者进行注意力融合后的双模态视听语音识别结果均优于单流模型识别结果,而且融合后的结果均高于 95%,其中 9 号、34 号和 44 号受试者融合实验分类率达到 100%。

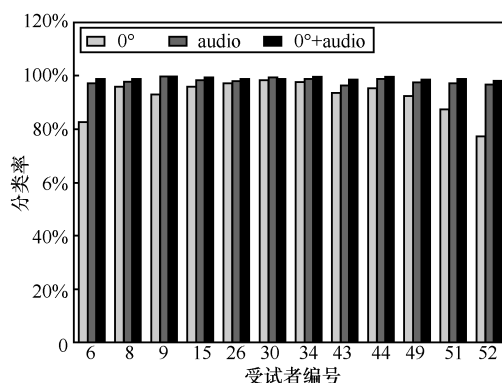


图 7 本文算法在测试集上各受试者的识别性能对比

此外,如表 4 所示,又分别将各个角度单流唇部视觉语音识别中 10 次运行结果中最优识别率列出,并与该数据库中经典算法作比较,可以看出,本文算法在以 0°、30°、60°、90°为拍摄角度录制的唇部视频中,算法识别率均优于经典算法,在 45°时接近参考文献[27]的结果,其中正面唇部视频识别率达到 95.1%,这表明算法模型在各单

流唇部视觉语音识别中的优越性,这也是为视听双模态融合实验打下了很好的基础。本文中并未对音频做特殊处理,例如对原始音频加噪声,而是将重点放在对单流的唇部视觉语音识别方法和视听双模态语音的同步融合识别方法中做突破,通过在这两方面的突破来最大化提升视听信息融合识别的效果。由以上分析可以看出,端到端视听双模态语音融合识别模型具有有效性。

5 结束语

本文提出了一种基于 SDBN 和 BLSTM 注意力融合的端到端视听双模态语音识别算法,该算法通过分别将音视频流信息送入 SDBN 和 BLSTM 进行处理,通过 SDBN 将高维音视频信息进行降维,提取稀疏瓶颈特征,之后通过 BLSTM 对特征序列进行模态处理,然后通过一种注意力机制将经过模态处理的音视频流向量序列进行自动对齐、融合,最后将融合后的向量序列送入一个附加了 Softmax 层的 BLSTM 进行分类识别。实验表明了端到端模型在单流及视听双流融合识别中的有效性。因为视听流模型一致,所以后续可以对音频流与多个视频流进行多流视听语音识别实验,以期得到更好的结果。

表 4 仅单流唇部视频在 5 个角度最优识别率比较

算法	0°	30°	45°	60°	90°
参考文献[24]算法	85.6%	82.5%	82.5%	83.3%	80.3%
参考文献[25]算法	87.6%	—	—	—	—
参考文献[26]算法	81.1%	80.0%	76.9%	69.2%	82.2%
参考文献[27]算法	94.7%	89.7%	90.6%	87.5%	93.1%
本文算法	95.1%	90.0%	90.3%	87.8%	93.7%



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.
WANG H K, PAN J, LIU C. Research development and forecast of automatic speech recognition technologies[J]. Telecommunications Science, 2018, 34(2): 1-11.
- [2] CHIU C C, SAINATH T N, WU Y, et al. State-of-the-art speech recognition with sequence-to-sequence models[C]//ICASSP 2018-2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 15-20, 2018, Calgary, AB, Canada. Piscataway: IEEE Press, 2018.
- [3] 王志宏, 杨震. 人工智能技术研究及未来智能化信息服务体系的思考[J]. 电信科学, 2017, 33(5): 1-11.
WANG Z H, YANG Z. Research on artificial intelligence technology and the future intelligent information service architecture[J]. Telecommunications Science, 2017, 33(5): 1-11.
- [4] NODA K, YAMAGUCHI Y, NAKADAI K, et al. Audio-visual speech recognition using deep learning[J]. Applied Intelligence, 2015, 42(4): 722-737.
- [5] HU D, LI X, LU X. Temporal multimodal learning in audiovisual speech recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), April 19-24, 2016, Brisbane, QLD, Australia. Piscataway: IEEE Press, 2016: 3574-3582.
- [6] TAKASHIMA Y, AIHARA R, TAKIGUCHI T, et al. Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss[Z]. 2016.
- [7] CHUNG J S, SENIOR A, VINYALS O, et al. Lip reading sentences in the wild[J]. arXiv: 1611.05358, 2016.
- [8] PETRIDIS S, STAFYLAKIS T, MA P, et al. End-to-end audio-visual speech recognition[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 15-20, 2018, Calgary, AB, Canada. Piscataway: IEEE Press, 2018: 6548-6552.
- [9] WAND M, SCHMIDHUBER J, VU N T. Investigations on End-to-End Audiovisual Fusion[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 15-20, 2018, Calgary, AB, Canada. Piscataway: IEEE Press, 2018: 3041-3045.
- [10] TAO F, BUSSO C. Gating neural network for large vocabulary audiovisual speech recognition[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 2018, 26(7): 1286-1298.
- [11] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(578): 504-507.
- [12] GREZL F, KARAFIAT M, KONTAR S, et al. Probabilistic and bottle-neck features for LVCSR of meetings[C]//IEEE International Conference on Acoustics, April 15-20, 2007, Honolulu, HI, USA. Piscataway: IEEE Press, 2007.
- [13] JI N N, ZHANG J S, ZHANG C X. A sparse-response deep belief network based on rate distortion theory[J]. Pattern Recognition, 2014, 47(9): 3179-3191.
- [14] LEE H, EKANADHAM C, NG A Y. Sparse deep belief net model for visual area V2[C]//Advances in Neural Information Processing Systems, Dec 3-6, 2007, Vancouver, British Columbia, Canada.[S.l.:s.n.], 2007: 873-880.
- [15] KIAEE F, CHRISTIAN G, ABBASI M. Alternating direction method of multipliers for sparse convolutional neural networks[Z]. 2016.
- [16] KHALID F, FANANY M I. Combining normal sparse into discriminative deep belief networks[C]// International Conference on Advanced Computer Science & Information Systems, Oct 15-16, 2016, Malang, Indonesia. Piscataway: IEEE Press, 2016.
- [17] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5-6): 602-610.
- [18] LUONG M T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[J]. arXiv preprint arXiv: 1508.04025, 2015.
- [19] ANINA I, ZHOU Z, ZHAO G, et al. OuluVS2: a multi-view audiovisual database for non-rigid mouth motion analysis[C]//IEEE International Conference & Workshops on Automatic Face & Gesture Recognition, May 4-8, 2015, Ljubljana, Slovenia. Piscataway: IEEE Press, 2015.
- [20] HINTON G E. A practical guide to training restricted Boltzmann machines[Z]. 2012.
- [21] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C]//The Thirteenth International Conference on Artificial Intelligence and Statistics, May 13-15, 2010, Chia Laguna Resort, Sardinia, Italy. [S.l.:s.n.], 2010: 249-256.
- [22] PETRIDIS S, WANG Y, LI Z, et al. End-to-end audiovisual fusion with LSTM[J]. arXiv preprint arXiv: 1709.04343, 2017.
- [23] KOUMPAROULIS A, POTAMIANOS G. Deep View2View mapping for view-invariant lipreading[C]//2018 IEEE Spoken Language Technology Workshop (SLT), Dec 18-21, 2018, Athens, Greece. Piscataway: IEEE Press, 2018: 588-594.
- [24] SAITOH T, ZHOU Z, ZHAO G, et al. Concatenated frame image based CNN for visual speech recognition[C]// Asian Conference on Computer Vision, Nov 20-24, 2016, Taipei, China. Berlin: Springer, 2016: 277-289.
- [25] FUNG I, MAK B. End-to-end low-resource lip-reading with maxout CNN and LSTM[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 15-20, 2018, Calgary, AB, Canada. Piscataway: IEEE Press, 2018: 2511-2515.

- [26] LEE D, LEE J, KIM K E. Multi-view automatic lip-reading using neural network[C]//Asian Conference on Computer Vision, Nov 20-24, 2016, Taipei, China. Berlin: Springer, 2016: 290-302.
- [27] PETRIDIS S, WANG Y, LI Z, et al. End-to-end multi-view lipreading[J]. arXiv preprint arXiv: 1709.00443, 2017.



陈恩(1962-),男,宁波大学信息科学与工程学院副教授、硕士生导师,在核心期刊和重要国际会议发表论文共100余篇,参与和主持国家级、省部级、市厅级和校级科研项目共16项;获得相关科研相关奖项3项。主要研究方向为图像及视频信息处理、多媒体通信、智能控制。

[作者简介]



王一鸣(1993-),男,宁波大学信息科学与工程学院硕士生,主要研究方向为音视频信息处理、视听觉语音识别。



阿卜杜萨拉木·艾海提(1995-),男,宁波大学信息科学与工程学院硕士生,主要研究方向为机器翻译、智能语音翻译。