结合瓶颈特征的注意力声学模型

龙星延 屈 丹 张文林

(解放军信息工程大学信息系统工程学院 郑州 450001)

摘 要 目前基于注意力机制的序列到序列声学模型成为语音识别领域的研究热点。针对该模型训练耗时长和鲁棒性差等问题,提出一种结合瓶颈特征的注意力声学模型。该模型由基于深度置信网络(Deep Belief Network,DBN)的瓶颈特征提取网络和基于注意力的序列到序列模型两部分组成:DBN能够引入传统声学模型的先验信息来加快模型的收敛速度,同时增强瓶颈特征的鲁棒性和区分性;注意力模型利用语音特征序列的时序信息计算音素序列的后验概率。在基线系统的基础上,通过减少注意力模型中循环神经网络的层数来减少训练的时间,通过改变瓶颈特征提取网络的输入层单元数和瓶颈层单元数来优化识别准确率。在TIMIT数据库上的实验表明,该模型在测试集上的音素错误率降低至了17.80%,训练的平均迭代周期缩短了52%,训练迭代次数由139减少至89。

关键词 声学模型,注意力模型,瓶颈特征,深度置信网络

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.01.040

Attention Based Acoustics Model Combining Bottleneck Feature

LONG Xing-yan QU Dan ZHANG Wen-lin

(Information System Engineering College,PLA Information Engineering University,Zhengzhou 450001,China)

Abstract Currently,attention mechanism based sequence-to-sequence acoustic models has become a hotspot of speech recognition. In view of the problem of long training time and poor robustness,this paper proposed an acoustical model combining bottleneck features. The model is composed of the bottleneck feature extraction network based on deep belief network and the attention-based sequence-to-sequence model. DBN introduces the priori information of the traditional acoustic model to speed up the model convergence rate and enhance robustness and distinction of bottleneck feature. Attention model uses the time temporal information of voice feature sequence to calculate the posterior probability of phoneme sequence. On the basis of the baseline system,the training time is decreased by reducing the layer number of the recurrent neural network in the attention model,and the recognition accuracy is optimized by changing the input dims and outputs of the bottleneck feature extraction network. Experiments on TIMIT dataset show that in the core test set,the phoneme error rate decreases to 17.80%,the average time training time during an iteration decreases by 52%,and the epochs of training iterations decreases to 89 from 139.

Keywords Acoustic model,Attention model,Bottleneck feature,Deep belief network

1 引言

声学模型(Acoustic Model,AM)一直是语音识别领域的研究热点。因为隐马尔可夫模型(Hidden Markov Model,HMM)能描述语音信号的时变性和非平稳性,同时具备完备的理论体系、高效的模型参数估计与解码算法,所以它与高斯混合模型(Gaussian Mixture Model,GMM)组成的GMM-HMM模型一直是主流的声学模型并已沿用20余年。随着深度学习技术的发展,学者们在保留HMM的基础上,使用以深度神经网络(Deep Neural Network,DNN)为代表的区分性模型取代GMM用于HMM状态建模,提升了识别性能[1]。基于HMM的声学模型存在以下缺陷:假设当前状态的概率分布只受前一状态影响,不能充分学习和利用语音特征序列的时序信息;将声学建模分解成状态识别和音素识别两个过程,造成声学模型结构复杂。

Chorowski等[2]针对基于HMM声学模型的缺陷,提出将机器翻译中基于注意力机制的序列到序列模型用于声学建模,抛弃基于HMM声学模型的状态独立性假设并简化模型结构,借助循环神经网络实现特征序列到音素序列的直接转换。Bahdanau等[3]进一步提出在声学模型的基础上利用加权有限状态机引入语言模型,建立大词汇连续语音识别系统。注意力声学模型由编码网络和解码网络两部分组成,其中编码网络负责从原始声学特征中提取高层特征,解码网络根据高层特征计算音素后验概率。与传统基于HMM的声学模型不同,解码网络包含的注意力子网络在训练过程中能自动学习声学特征和音素的对应关系,使模型不依赖先验对齐信

到稿日期:2017-12-08 返修日期:2018-03-27 本文受国家自然科学基金(61673395,61403415),河南省自然科学基金(162300410331)资助。

龙星延(1992-),男,硕士生,主要研究方向为语音识别与人工智能;屈 丹(1974-),女,博士,副教授,主要研究方向为语音信号处理、智能信息处理等,E-mail:qudanqudan@sina.com(通信作者);张文林(1982-),男,博士,副教授,主要研究方向为语音识别、人工智能。

息以及强制对齐等操作。针对低资源语种，注意力模型可选择将字素(grapheme)作为建模对象，在摆脱对发音字典依赖的同时保证较高的识别率。另一方面，注意力声学模型尽管能更有效地学习和处理长时信息，但由于其编码网络使用的循环神经网络为递归结构，不能让多帧数据被同时处理，导致无法充分发挥 GPU 并行计算的优势，造成训练耗时长的问题。注意力模型彻底将语音学的先验知识摒弃，使其缺少有效的初始化参数，造成模型的收敛速度缓慢，这也是模型训练时间增加的重要原因。此外，文献[4]指出注意力模型在噪声环境中存在鲁棒性差的问题。

针对上述问题，文中提出一种结合瓶颈特征(Bottleneck Feature)的注意力声学模型。首先，以三音子(Tri-phone)状态作为建模单元训练 DBN 用于提取瓶颈特征。瓶颈特征可看作原始语音特征的非线性压缩变换，其维度不仅低于 Mel 频率的倒谱系数(Mel-scale Frequency Cepstrum Coefficient，MFCC)、线性感知预测系数(Perceptual Linear Predictive，PLP)和 Mel 滤波器组系数(Mel-scale Filter Bank，FBANK)等传统声学特征[5]，而且针对不同说话人、噪声等干扰的鲁棒性和区分性更强。此外，因为 DBN 训练的标注是基于 HMM 的声学模型生成的，所以使用该特征能够为注意力声学模型提供传统声学模型的先验信息。然后，对于引入瓶颈特征后的注意力模型，减少编码网络中循环神经网络的层数，从而在保证识别准确率的同时有效减小模型的规模并缩短训练时间。对于瓶颈特征提取网络，则通过调整输入层和瓶颈层的单元数量寻找最佳参数，进而优化系统的整体识别性能。最后，在 TIMIT 上的对比实验结果表明，结合瓶颈特征的注意力声学模型的识别性能优于原有注意力模型，并且需要的训练时间更少。

2　基于瓶颈特征的注意力声学模型

本节对提出的基于瓶颈特征的注意力声学模型进行详细阐述，包括瓶颈特征提取方法和注意力声学模型。

2.1　基于 DBN 的瓶颈特征提取

文献[6]提出在连续语音识别系统中将训练好的多层感知器作为特征提取网络，提取中间神经元数目较少的隐含层状态作为瓶颈特征。语音识别领域中的大量实验证明[7-9]，瓶颈特征能够保留和压缩原始特征中的有效信息，取得比传统声学特征更好的性能。文献[6]提出在声学模型中使用 DBN 替代多层感知器作为瓶颈特征提取网络。DBN 解决了多层感知器在训练时容易陷入局部最优的问题，并且具有更强的建模和表征能力，因而使用 DBN 提取的瓶颈特征的鲁棒性和区分性更强。

如图 1 所示，以一个 5 层 DBN 为例来介绍本文使用的瓶颈特征提取网络：网络的输入为 FBANK 特征参数，网络的输出层为 softmax 层，每个输出层单元对应绑定三音子状态的后验概率。在输入层和输出层之间，网络有 3 个隐含层，其中第 2 个隐含层为瓶颈层，它的状态单元数目相对较少。训练该网络时，首先利用基于三音子的 GMM-HMM 模型获取句子中每帧特征参数的状态标注；然后按照 DBN 的预训练方法，采用无监督训练的方式，按照从网络输入到网络输出的顺序逐层初始化相邻层单元之间的连接权重参数；最后根据 GMM-HMM 生成的标注信息，以有监督训练的方式使用反向传播算法(backpropagation)对网络参数进行微调。

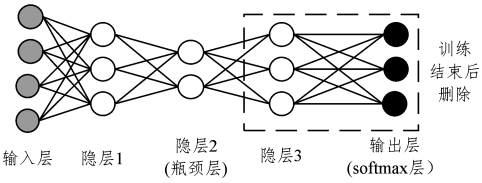

图 1　基于 DBN 的瓶颈特征提取网络

Fig. 1　Feature extraction network based on DBN

训练结束后，将瓶颈层后的网络状态单元全部删除即可得到瓶颈特征提取网络。把原始特征参数作为网络输入，通过该网络计算出瓶颈层状态值作为瓶颈特征。因为 DBN 在有监督训练过程中使用的数据包含 GMM-HMM 模型提供的对齐信息，所以与传统声学特征相比，DBN 提取的瓶颈特征对音素状态具有更强的区分性。因此，如果将该瓶颈特征应用于注意力声学模型，则能为其提供较为准确的特征与音素状态的对齐信息。换言之，这是一种将注意力声学模型与传统基于 HMM 模型相结合的方法。

2.2　结合瓶颈特征的注意力声学模型

基于注意力机制的序列到序列模型最早被应用于机器翻译领域[10]，该模型借助循环神经网络动态强大的时序建模能力，实现不同语种句子之间的直接转换。音素识别可看作原始语音特征到音素的“翻译”，因此可以将注意力模型用于声学建模，使声学模型更加有效地发掘和利用声学特征序列中的时序信息。原始注意力声学模型的结构如图 2 所示。

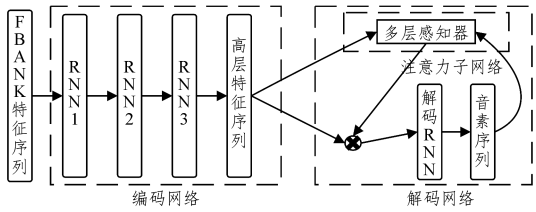

图 2　基于注意力机制的声学模型

Fig. 2　Attention mechanism based acoustic model

该模型由编码网络和解码网络两部分组成。编码网络的功能是挖掘和利用输入特征序列的前后依赖信息以及增强特征的表达能力和区分性，实现从语音特征序列到高层特征序列 $h=(h_1,h_2,\cdots,h_T)$ 的转换。解码网络根据高层特征序列 h 逐个计算输出序列位置 o 上所有音素出现的后验概率向量 y_o，最终得到输出序列 $y=(y_1,y_2,\cdots,y_O)$。

$$h=encoder(x) \tag{1}$$

$$y_o=decoder(h) \tag{2}$$

其中，$encoder$ 和 $decoder$ 分别代表注意力模型的编码网络和解码网络。给定一段长度为 T 的语音特征序列 $x=(x_1,x_2,\cdots,x_T)$ 和对应的正确标注音素序列 $p=(p_1,p_2,\cdots,p_O)$，注意力声学模型根据式(3)计算音素序列的后验概率。

$$P(p|x)=\prod_{o=1}^{O}P(p_o|x,p_{<o})=\prod_{o=1}^{O}y_o^{p_o} \tag{3}$$

其中，标量 $y_o^{p_o}$ 为输出序列位置 o 出现音素 p_o 的后验概率。令待识别的所有音素集合为 $P=\{p_1,p_2,\cdots,p_m\}$，m 为音素

数目，则注意力模型位置 o 的输出向量 $y_o=[y_o^{p_1},y_o^{p_2},\cdots,y_o^{p_m}]$。

2.2.1 编码网络

在文献[2]提出的注意力模型中，编码网络结构为 3 层基于门循环单元[11](Gate Recurrent Unit，GRU)的双向循环神经网络。它在时刻 t 的输入 x_t 和输出 h_t 的关系如式(4)所示：

$$h_t=3\text{-}layer\text{-}biRNN(x_t) \tag{4}$$

其中，$3\text{-}layer\text{-}biRNN$ 代表堆叠 3 层的双向循环神经网络。循环神经网络的双向结构赋予编码网络同时利用过去和未来时序信息的能力。堆叠多层的结构则赋予网络提取和利用高层特征信息的能力。但是循环神经网络的递归结构使得 GPU 并行计算的优势无法充分发挥，导致层叠后循环神经网络在训练过程中计算反向梯度的时间成倍增加，带来训练复杂的问题。

本文提出的结合瓶颈特征的声学模型如图 3 所示。该模型与原有编码网络的区别在于：1)在编码网络前端增加 2.1 节中训练好的基于 DBN 瓶颈特征的提取网络，并将瓶颈特征序列作为循环神经网络的输入；2)缩减循环神经网络单元的堆叠层数，只留下 1 层网络处理特征序列的时序信息。

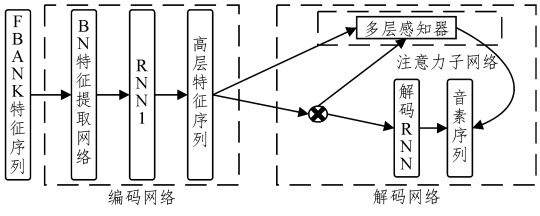

图 3 结合瓶颈特征的注意力声学模型

Fig. 3 Attention-based acoustic model combining BNF extraction network

新编码网络的输出和输入之间的关系如式(5)、式(6)所示：

$$\hat{h}_t=BNF(x_t) \tag{5}$$

$$h_t=1\text{-}layer\text{-}biRNN(\hat{h}_t) \tag{6}$$

其中，BNF 为瓶颈特征提取网络，$1\text{-}layer\text{-}biRNN$ 代表单层循环神经网络。增加基于 DBN 特征的提取网络的目的首先是取代循环神经网络的多层结构。因为 DBN 不是递归结构，所以在训练 DBN 时能够在 GPU 上并行计算多帧的梯度，从而有效提升训练效率。与此同时，通过减少循环神经网络的层数缩短反向梯度传递的距离，进一步缩短训练时间。其次，基于 HMM 声学模型的先验信息以权重矩阵的形式存贮在 DBN 中，使 DBN 网络提取的高层特征具有更好的区分性，进而使模型训练更容易收敛。此外，基于 DBN 提取的瓶颈特征对噪声有着较强的鲁棒性，使用它能弥补注意力模型在噪声环境下鲁棒性弱的缺陷。

2.2.2 解码网络

解码网络由注意力子网络和用于解码的循环神经网络组成。注意力子网络的结构主要由只含一层隐含层的多层感知器(MLP)组成，主要任务是计算输出序列位置 o 对应的目标向量 ct_o。计算输出向量 y_o 时，注意子网络首先计算输出序列前一位置的输出向量 y_{o-1} 与所有时刻 $t\in\{1,\cdots,T\}$ 的高层特征向量 h_t 的关联度；然后将对关联度进行指数归一化处理后的数值作为权重，并根据权重合并高层特征向量，最终得到目标向量 ct_o。该计算过程如式(7)－式(9)所示：

$$e_{o,t}=MLP(y_{o-1},h_t) \tag{7}$$

$$\alpha_{o,t}=\frac{\exp(e_{o,t})}{\sum_{t=1}^{T}\exp(e_{o,t})} \tag{8}$$

$$ct_o=\sum_{t=1}^{T}\alpha_{o,t}h_t \tag{9}$$

用于解码的循环神经网络根据 y_{o-1} 和目标向量 ct_o 计算状态 s_o，再经过 softmax 层计算得到输出向量 y_o，y_o 的每个分量代表对应音素出现在输出序列位置 o 的概率。计算过程如式(10)、式(11)所示：

$$s_o=RNN(y_{o-1},ct_o) \tag{10}$$

$$y_o=softmax(s_o) \tag{11}$$

2.3 模型训练和解码

模型训练过程中，采用随机梯度下降法求取如式(12)所示的负概率对数函数的最优化参数。

$$\hat{\theta}=\arg\max_{\theta}\frac{1}{N}\sum_{n=1}^{N}-\log P(p^n|x^n,\theta) \tag{12}$$

其中，N 为待训练语音样本数量，n 为样本编号，$p^n=(p_1^n,\cdots,p_o^n)$ 为语音的正确音素标注序列，$x^n=(x_1^n,\cdots,x_T^n)$ 为语音的特征序列，θ 为模型的全部参数，即编码网络、解码网络和注意力子网络中所有的权重矩阵和偏置向量。

根据式(3)，单段语音样本的后验概率如式(13)所示：

$$P(p^n|x^n,\theta)=\prod_{o=1}^{O}P(p_o^n|x^n,\theta)=\prod_{o=1}^{O}y_o^{p_o}(x^n) \tag{13}$$

其中，$y_o^{p_o}(x^n)$ 表示将编号为 n 的句子作为系统输入计算得到的音素 p_o 出现在输出序列位置 o 的后验概率。

解码时，首先将特征序列 x 作为注意力模型的输入得到输出序列 y。由于解码音素序列长度未知，因此训练前在音素集和训练语音特征序列结尾增加终止符标记〈eos〉，在解码时使用 BeamSearch[15] 算法搜索概率分数最大且以〈eos〉结尾的音素序列作为解码结果。

3 实验及分析

3.1 数据集

实验采用的 TIMIT 语料库[16]是语音识别领域最常用的标准数据库之一，它包含 6300 段英语朗读语音，从中选取 3296 条语句作为训练集，192 条语句作为测试集，400 条语句作为开发集。语音信号的采样频率为 16 kHz，采样位数为 16 bit，采用 Hamming 窗处理，帧长为 25 ms，帧移为 10 ms，预加重系数为 0.97，声学特征在 40 维 FBANK 特征的基础上拼接一阶、二阶差分，共计 120 维特征。对于提取好的特征，先在训练集范围内对每个特征分类使用标准正态分布进行归一化，并记录均值和方差，再使用训练集的均值和方差对测试集和开发集特征进行归一化。

3.2 实验配置

采用 kaldi-pdnn 工具包[12]建立和训练 DBN 模型，采用开源深度学习工具 Theano[17]建立和训练注意力模型。实验平台的硬件配置为 Intel Xeon E2670 24 核 CPU、64 GB 内存和 NVIDIA Tesla K80 显卡。

3.2.1 基于 DBN 的瓶颈特征提取网络

在提取瓶颈特征的 DBN 中，将分别对应当前帧和该帧前

后 4 帧拼接成的 120×9＝1080 维向量作为网络输入，网络输出层维度为 1236，分别对应绑定 1236 个三音子的后验概率。该网络除了输入层和输出层之外，还有 6 个隐含层，第 4 个隐含层是含有 40 个单元的瓶颈层，其余隐含层均含有 1024 个单元。DBN 隐层数和每层的单元数是参照文献[7,12]设置的，该结构的 DBN 在连续语音识别中拥有较高的准确率，且模型参数的规模较小。DBN 在阶段 1 采用基于小批量(mini-batch)随机梯度下降法的对比散度算法[18](Contrastive Divergence，CD)进行逐层预训练。批量大小(Batch Size)为 128，每层迭代周期为 5，5 个周期的动量因子(momentum)分别为 0.5，0.6，0.7，0.8，0.9。DBN 在阶段 2 采用基于小批量的随机梯度下降算法的 BP 算法进行参数微调(fine-tunning)。批量块大小为 256，动量因子固定为 0.5，初始学习率为 0.08，当开发集的帧准确率在一个周期训练结束后增幅低于 0.2%时，学习率减半，当学习率低于 0.02 时，停止迭代。训练结束后，保留 DBN 输入层到瓶颈层的网络参数，并将瓶颈层单元状态作为瓶颈特征，因此瓶颈特征的维度为 40。

3.2.2　注意力模型

基于注意力的声学模型的编码网络和解码网络均采用基于门循环单元(Gated Recurrent Unit，GRU)的循环神经网络，隐含层单元数为 256。权重矩阵通过标准正交矩阵进行初始化，偏置向量初值为 0，内部状态值采用均值为 0、方差为 0.1 的独立高斯分布初始化。解码网络的输出层单元数为 63，分别对应 61 个音素、空白符〈spc〉和序列终止符〈eos〉的后验概率。

以式(12)作为目标函数，使用 Adadelta[13] 算法对模型参数进行迭代更新。训练过程分为两个阶段：第一阶段的批量大小数为 8，目的是提高训练效率，使模型参数尽快收敛；第二阶段的批量大小数为 1，在计算梯度之前给模型的所有参数加入随机高斯噪声，以增强模型的抗噪能力和鲁棒性。在训练过程中，如果连续 5 次迭代都没有降低开发集的音素错误率，则自动进入下一阶段或者终止训练。

3.3　评价指标

为评价声学模型的识别性能，将音素错误率(Phone Error Rate，PER)作为评价指标。其计算方式如下：利用动态规划算法对模型解码输出序列与标注序列进行对比，并统计出音素的插入错误(I)、删除错误(D)和替代错误(R)。设 N 为测试集中所有序列的音素总数，则 PER 为：

$$PER=\frac{I+D+R}{N}\times 100\% \tag{14}$$

为评价和对比注意力模型的训练速度，在训练过程的第二阶段中批量样大小为 1 的条件下，将训练集的所有样本更新模型参数的平均周期(epoch)作为评价指标。

3.4　实验结果和分析

3.4.1　不同注意力模型对系统性能的影响

表 1 列出了采用不同结构编码网络的注意力声学模型对应的音素错误率和平均周期。从表 1 中的第 1－3 行可得，增加注意力模型编码网络堆叠循环神经单元的层数后，系统的识别性能得到了提升，这说明编码网络采用深层结构后特征抽象层次更高，提取的高层特征区分性更强，因此识别性能得到有效提高。但与此同时，网络层数的增加也造成了平均训练周期的快速延长。第 4－6 行是增加基于 DBN 的瓶颈特征提取网络后注意力模型的性能。循环神经网络层数相同的条件下，使用瓶颈特征后的注意力模型的音素错误率相比未使用的模型均有所下降，并且使用 1 层和 3 层循环神经网络的模型的音素错误率只相差 0.21%。这表明 DBN 提取的瓶颈特征的鲁棒性和区分性更强，可以替代基于多层循环神经网络的结构。第 7 行是直接将 DBN 与 HMM 相结合的声学模型的性能，它的音素错误率明显高于表中大部分注意力声学模型，证明了注意力模型具有比 HMM 更强的时序建模能力。

表 1　不同声学模型的音素错误率和平均周期的对比

Table 1　Performance comparison between acoustic models in PER and average epoch

模型	音素错误率/%	平均周期/min
Attention 1-RNN	21.83	36.2
Attention 2-RNN	21.41	54.8
Attention 3-RNN	19.57	76.3
BN Attention 1-RNN	19.07	37.5
BN Attention 2-RNN	18.95	56.4
BN Attention 3-RNN	18.86	78.1
DBN-HMM	21.60	—

表 1 中第 4 行对应的模型为本文提出的声学模型。按照 2.2.2 节中的方法，模型在引入瓶颈特征的同时只保留 1 层循环神经网络，不仅识别性能优于原始模型(第 3 行对应的模型)，而且有效缩短了训练时间，平均迭代周期缩短 52%。本文未统计 DBN 的训练时间，原因在于一方面可以从已经训练好的模型中迁移过来直接使用，另一方面由于 DBN 支持在 GPU 进行并行多帧运算，其训练时间相对注意力模型训练的总时间可以忽略。

图 4 是训练过程中开发集音素错误率的下降曲线。从中可以看出，本文所提模型与原始模型相比，不仅音素错误率下降速度明显加快，而且训练所需的迭代次数由 139 减少至 89。这证明了借助瓶颈特征提取网络为注意力声学模型提供先验信息的方法的有效性。

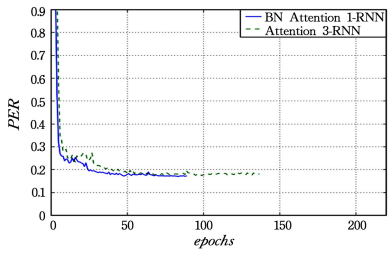

图 4　原始注意力模型和本文模型训练过程中开发集的音素错误率

Fig. 4　PER of attention model and proposed model on validation set in training process

3.4.2　不同结构瓶颈特征提取网络性能对系统性能的影响

为了进一步提升系统的识别性能，本文改变瓶颈特征提取网络的输入特征的帧数和瓶颈层的单元数后，测试系统的整体性能以寻找最优参数，实验结果如表 2、表 3 所列。

表 2　基于不同输入帧数的系统的音素错误率

Table 2　PER of systems with different input frames

输入帧数	音素错误率/%
1	17.93
3	18.49
5	18.92
7	19.31
9	19.07

表 3 基于不同瓶颈层单元数的系统的音素错误率

Table 3 PER of systems with different bottleneck units

模型	音素错误率/%
BN Attention30	18.87
BN Attention 40	17.93
BN Attention 50	17.83
BN Attention 60	17.80
BN Attention 70	17.81

表 2 列出了将 BN Attention 1-RNN 作为基线系统，基于不同帧数的连续特征作为输入的系统对应的音素错误率。从中可以看出，随着输入帧数的减少，系统识别的音素错误率下降。这是因为对连续多帧提取瓶颈特征会破坏特征序列的时序信息，从而对注意力模型计算音素和特征对齐造成干扰。只将当前 1 帧作为输入时不会破坏时序信息，此时系统的音素错误率达到最低，为 17.93%。

表 3 列出了输入帧数为 1 时，基于不同瓶颈层单元数的系统的音素错误率。从中可以看出，瓶颈层单元数从 30 增加至 50 时，音素错误率有明显降低，而瓶颈层单元数大于 50 后，音素错误率基本没有变化。这说明瓶颈层单元数目与音素种类数目相接近时，瓶颈特征对音素有较好的区分性。当瓶颈层隐含层单元数达到 60 时，系统的音素错误率下降到最低，为 17.80%。

结束语 文中提出了一种结合瓶颈特征的注意力的声学模型，先通过在编码网络中增加基于 DBN 的瓶颈特征提取网络，将基于 HMM 传统声学模型的知识迁移至注意力模型中；再通过减少循环神经网络的层数和优化瓶颈特征提取网络结构，实现简化模型结构和提升训练速度与识别率的目的。实验结果表明，所提模型能够有效提升系统识别性能并缩短训练时间，音素错误率降低至 17.80%，平均训练迭代周期和训练所需迭代次数均少于原始注意力模型。下一步的研究方向是针对注意力模型的解码网络，寻找新的算法和网络结构来优化系统识别性能和效率。

参 考 文 献

[1] HINTON G, DENG L, YU D, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[2] CHOROWSKI J, BAHDANAU D, CHO K, et al. End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results [EB/OL]. https://arxiv.org/abs/1412.1602.

[3] BAHDANAU D, CHOROWSKI J, SERDYUK D, et al. End-to-end attention-based large vocabulary speech recognition[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2016: 4945-4949.

[4] KIM S, HORI T, WATANABE S. Joint CTC-attention based end-to-end speech recognition using multi-task learning[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2017: 4835-4839.

[5] GREZL F, FOUSEK P. Optimizing bottle-neck features for lvcsr [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2008: 4729-4732.

[6] YU D, SELTZER M L. Improved Bottleneck Features Using Pretrained Deep Neural Networks[C]// 2011 Twelfth Annual Conference of the International Speech Communication Association. 2011: 237-240

[7] LI J H, YANG J A, WANG Y. New Feature Extraction Method Based on Bottleneck Deep Belief Networks and Its Application in Language Recognition[J]. Computer Science, 2014, 41(3): 263-266. (in Chinese)
李晋徽，杨俊安，王一. 一种新的基于瓶颈深度信念网络的特征提取方法及其在语种识别中的应用[J]. 计算机科学，2014，41(3)：263-266.

[8] WANG Y, YANG J A, LIU H, et al. Bottleneck Feature Extraction Method Based on Hierarchical Deep Sparse Belief Network [J]. Parttern Recognition and Artificial Intelligence, 2015, 28(2): 173-180. (in Chinese)
王一，杨俊安，刘辉，等. 基于层次稀疏 DBN 的瓶颈特征提取方法[J]. 模式识别与人工智能，2015，28(2)：173-180.

[9] CHEN L, YANG J A, WANG Y, et al. A Feature Extraction Method Based on Discriminative and Adaptive Bottleneck Deep Belief Network in Large Vocabulary Continuous Speech Recognition System[J]. Journal of Signal Processing, 2015, 31(3): 290-298. (in Chinese)
陈雷，杨俊安，王一，等. LVCSR 系统中一种基于区分性和自适应瓶颈深度置信网络的特征提取方法[J]. 信号处理，2015，31(3)：290-298.

[10] BAHDANAU D, CHO K, BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[EB/OL]. https://arxiv.org/abs/1409.0473.

[11] CHO K, MERRIENBOER B V, GULCEHRE C et, al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation [EB/OL]. https://arxiv.org/abs/1406.1078.

[12] MIAO Y. Kaldi+PDNN: Building DNN-based ASR Systems with Kaldi and PDNN[EB/OL]. https://arxiv.org/abs/1401.6984.

[13] PASCANU R, MIKOLOV T, BENGIO Y. On the difficulty of training Recurrent Neural Networks[EB/OL]. https://arxiv.org/abs/1211.5063v2.

[14] HINTON G, DENG L, YU D, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[15] SUTSKEVER I, VINYALS O. Sequence to Sequence Learning with Neural Networ-ks[EB/OL]. https://arxiv.org/abs/1409.3215.

[16] GAROFOLO J S, LAMEL L F, FISHER W M, et al. TIMIT Acoustic-Phonetic Continuous Speech (MS-WAV version)[J]. Journal of the Acoustical Society of America, 1993, 88(88): 210-221.

[17] BERGSTRA J, BREULEUX O, BASTIEN F, et al. Theano: a CPU and GPU math expression compiler[EB/OL]. http://conference.scipy.org/scipy2010/slides/james_bergstra_theano.pdf.

[18] HINTON G E, OSINDERO S, TEH Y W. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2014, 18(7): 1527-1554.