文章编号：1003-0077(2018)09-0028-07

基于 TDNN-FSMN 的蒙古语语音识别技术研究

王勇和，飞 龙，高光来

（内蒙古大学 计算机学院，内蒙古 呼和浩特 010021）

摘 要：为了提高蒙古语语音识别性能，该文首先将时延神经网络融合前馈型序列记忆网络应用于蒙古语语音识别任务中，通过对长序列语音帧建模来充分挖掘上下文相关信息；此外研究了前馈型序列记忆网络“记忆”模块中历史信息和未来信息长度对模型的影响；最后分析了融合的网络结构中隐藏层个数及隐藏层节点数对声学模型性能的影响。实验结果表明，时延神经网络融合前馈型序列记忆网络相比深度神经网络、时延神经网络和前馈型序列记忆网络具有更好的性能，单词错误率与基线深度神经网络模型相比降低 22.2%。

关键词：蒙古语；语音识别；时延神经网络；前馈型序列记忆网络

中图分类号：TP391 **文献标识码**：A

Mongolian Speech Recognition Based on TDNN-FSMN

WANG Yonghe, BAO Feilong, GAO Guanglai

(College of Computer Science, Inner Mongolia University, Hohhot, Inner Mongolia 010021, China)

Abstract: In order to improve Mongolian speech recognition, the Time Delay Neural Network (TDNN) and Feed-forward Sequential Memory Network (FSMN) are combined to model the long sequence speech frames. In addition, we investigate the influence caused by the information from the preceding and the subsequent frames in the memory block over FSMN. We compare the performance of the TDNN-LSTM using different hidden layers and nodes. The results show that the fusion of TDNN and FSMN produces better performance than DNN, TDNN and FSMN, reducing the word error rate (WER) by 22.2% compared with the DNN baseline.

Key words: Mongolian; speech recognition; Time Delay Neural Network; Feed-forward Sequential Memory Network

0 引言

语音是人类最自然、便捷的交流方式，而语音识别技术，就是让机器能够“听懂”人类的语言并将语音信号转化为对应的文本或命令。基于高斯混合模型—隐马尔可夫模型（Gaussian Mixture Model-Hidden Markov Models，GMM-HMM）的语音识别框架在很长一段时间都是语音识别系统的主导框架，其核心就是用 GMM 对语音的观察概率进行建模，而用 HMM 对语音的转移概率进行建模[1]。近年来，深度神经网络（Deep Neural Network，DNN）[2]的研究和应用极大地推动了语音识别的发展，相比传统的基于 GMM-HMM 的语音识别系统，其最大的改变是采用 DNN 替换 GMM 对语音的观察概率进行建模来计算 HMM 状态的后验概率。根据文献[3]，基于 DNN-HMM 的声学模型采用固定长度的输入窗对语音的上下文特征进行建模，而语音是一种各帧之间具有很强相关性的复杂时变信号，所以这种方法不能充分利用语音的上下文时序信息。

相比 DNN，时延神经网络（Time Delay Neural Network，TDNN）[4]同样是一种前馈网络架构，它对每个隐藏层的输出都在时域进行扩展，即每个隐藏层接收到的输入不仅是前一层在当前时刻的输出，还有前一层在之前和之后的某些时刻的输出。在文献[5]中，通过选择正确的时间步长和对隐藏层输出进行降采样，TDNN 可以从输入上下文中的所有时间步长提取足够语音特征信息。因此，TDNN

收稿日期：2017-10-20 定稿日期：2017-12-18

基金项目：国家自然科学基金(61563040, 61773224)；内蒙古自然科学基金(2016ZD06)

会参考前一层网络的历史输出，可以对更长的历史信息进行建模而不能对未来信息进行建模。Zhang 等人[6-7]提出了一种更简单的“记忆”存储神经网络结构，即前馈型序列记忆网络（Feed-forward Sequential Memory Network，FSMN），已被证明在大词汇量连续语音识别任务中具有比 DNN 和长短时记忆模块（Long-Short Term Memory，LSTM）更好的性能。FSMN 是在 DNN 隐藏层旁边引入“记忆”模块的多层前馈神经网络模型。该“记忆”模块用于临时存储固定大小的上下文信息作为短期记忆机制，能够以时间序列学习长期依赖性信息。在本文中，TDNN 融合 FSMN 的网络结构被应用于蒙古语语音识别声学模型。

目前，在中国内蒙古自治区、蒙古国及周边地区大约有 600 万人将蒙古语作为第一或第二官方语言，但是蒙古语语音识别研究仍处于初始阶段。高光来等[8]在 2006 年首次构建了蒙古语语音识别系统，在文献[9-10]中进一步对声学模型进行优化和设计。在文献[11]中，飞龙等人提出了基于词干的蒙古语语音关键词检测方法，并使用分割的方法在蒙古语大词汇量连续语音识别中取得了较好的效果[12]。在文献[13]中，张晖等人在蒙古语语音识别研究中引入了基于 DNN 的声学模型，获得了显著的性能提升。最近，基于深度神经网络的声学模型广泛应用于蒙古语语音识别中，如卷积神经网络（Convolutional Neural Network，CNN）和长短时记忆模块等，获得比 DNN 更好的识别结果[14]。然而，与其他语言如中文和英文相比，蒙古语语音识别声学模型仍有很大的优化空间。

为进一步提高蒙古语语音识别性能，本文首先将 TDNN 融合 FSMN 应用于蒙古语语音识别系统声学模型，通过对长序列语音帧进行建模来充分挖掘上下文相关信息。其次，FSMN 中“记忆”模块用于存储对判断当前语音帧有用的历史信息和未来信息，本文通过用“记忆”模块中不同的历史和未来语音帧信息长度对模型进行建模，分析其对蒙古语语音识别系统性能的影响。最后，研究了不同隐藏层数目和每个隐藏层节点数对融合的 TDNN-FSMN 模型性能的影响。

1　基于 TDNN-FSMN 的蒙古语语音识别系统

1.1　TDNN 声学模型

TDNN 是一种多层（通常三个以上）前馈神经网络模型，传统的前馈神经网络每个隐藏层的输入都是前一层网络的输出，而 TDNN 在网络传播的过程中对各个隐藏层的输出也做了扩展，它将隐藏层的当前输出与其前后若干时刻的输出拼接在一起，作为下一个隐藏层的输入。因此，TDNN 每个隐藏层的输入会参考前一层网络的历史输出，可以对更长的历史信息进行建模。

传统的 TDNN 每一个时间步长上，隐藏层的激活函数都会被计算一次。因此，在相邻时间步长中，大量的上下文相同信息被重复计算，大大增加了神经网络的训练复杂度。而 TDNN 相邻节点之间的变化可能很小，包含了大量的重复信息，因此可以每隔几帧合并计算一次结果，从而加速训练和解码过程。在文献[5]中，提出一种在 TDNN 训练中采用降采样技术来减小模型计算复杂度，通过选择合适的时间步长来大幅减少运算量，同时不能使所有的历史信息都可以被网络学习到。图 1 表示常规 TDNN（实边＋虚边）和降采样 TDNN（实边）结构图。传统 TDNN 每个隐藏层的隐藏层单元（实边＋虚边）都会被计算，而且相邻时间步长会重复计算隐藏层单元。采用降采样技术的 TDNN 在每个隐藏层只会计算一定时间间隔的隐藏层单元（实边），不仅能够对长时间依赖性的语音信号进行建模，而且模型复杂度较传统 TDNN 有大幅度降低。

1.2　FSMN 声学模型

前馈型序列记忆网络是一种含有多个隐藏层的前馈神经网络。相比传统的 DNN 结构，FSMN 在其隐藏层旁边增加了一个称为“记忆块”的模块，这些“记忆块”用于存储语音序列中与当前帧相关的历史关联信息以及未来关联信息。这些信息使得 FSMN 可以对语音序列中的长期相关性信息进行建模。图 2 表示在隐藏层中添加两个“记忆块”的 FSMN 结构图。

给定序列 $w_1=(x_{11},x_{12},\cdots,x_{1N})$，$X=\{x_1,x_2,\cdots,x_t\}$，每个 $x_t\in X$ 表示时间 t 的输入数据。相应的隐藏层输出表示为 $H=\{h_1,h_2,\cdots,h_t\}$。图 2 即为“记忆块”的结构示意图，当前语音帧 h_t 及其前 N_1 帧的输出和后 N_2 帧的输出被计算到固定大小维度，并将其与当前隐藏层的输出一起作为下一个隐藏层的输入。

图 3 即为“记忆块”结构示意图，其中 N_1 表示历史语音帧的数量，N_2 表示未来语音帧的数量，h_t^l 表示当前语音帧的输出，$h_{t-1}^l,\cdots,h_{t-N_1}^l$ 表示与当前

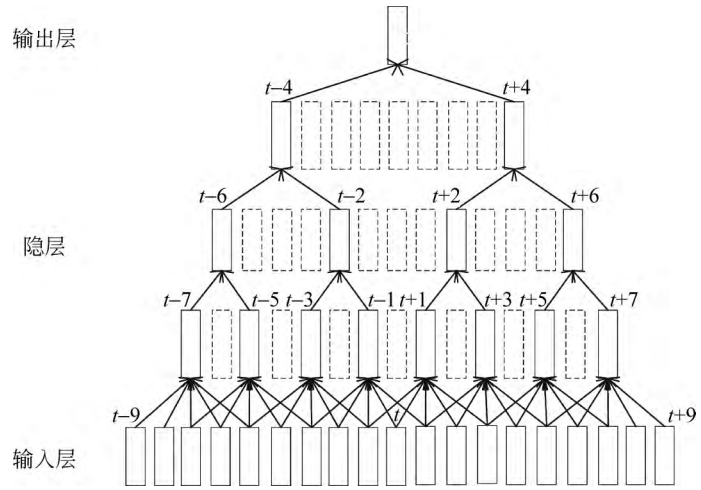

图1 TDNN结构图

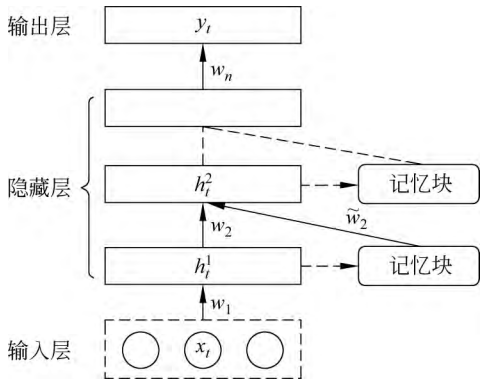

图2 FSMN模型

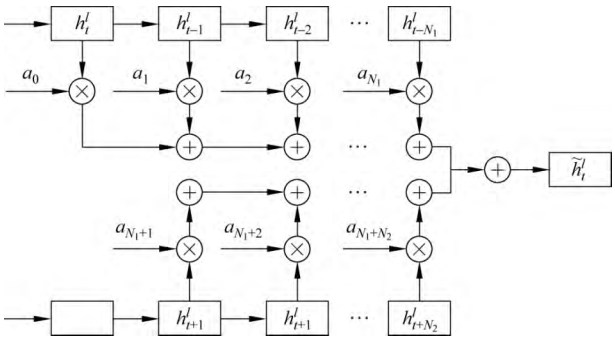

图3 “记忆块”结构图

语音帧相关的前 N_1 帧的输出，$h_{t+1}^l,\cdots,h_{t+N_2}^l$ 表示与当前语音帧相关的后 N_2 帧的输出，a 为编码系数，$\widetilde{h}_t^l$ 表示在 t 时刻 h_t^l 学习到的上下文相关的信息并与 h_t^l 一同作为下一个隐藏层的输入。

根据要使用的编码方法，编码系数 a 可以初始化为标量系数或向量系数。

(1) 如果编码系数 a 设置为标量，则FSMN称为标量FSMN(简称sFSMN)，如式(1)所示。

$$\widetilde{h}_t^l=\sum_{i=0}^{N_1}a_{t,i}^l\cdot h_{t-i}^l+\sum_{j=1}^{N_2}a_{t,M_1-1+j}^l\cdot h_{t+j}^l \tag{1}$$

(2) 如果编码系数 a 设置为向量，则FSMN称为向量FSMN(简称vFSMN)，如式(2)所示。

$$\widetilde{h}_t^l=\sum_{i=0}^{N_1}a_{t,i}^l\odot h_{t-i}^l+\sum_{j=1}^{N_2}a_{t,M_1-1+j}^l\odot h_{t+j}^l \tag{2}$$

由于vFSMN具有更好的建模能力，因此在本文中采用了vFSMN，简称为FSMN。

1.3 TDNN-FSMN声学模型

本文中，TDNN与FSMN相融合的神经网络结构被应用于蒙古语语音识别系统的声学模型。TDNN在网络传播过程中对各个隐藏层的输出做了扩展，传统前馈神经网络每个隐藏层的输入都是前一层网络的输出，TDNN则会参考前一层网络的历史输出，能对更长的历史信息进行建模，而且深层次的TDNN网络结构可以更加有效地提取训练数据中高层次信息的特征。双向FSMN神经网络结构在隐藏层旁增加了一个称为“记忆块”的模块，用于存储对判断当前语音帧有用的历史信息和未来信息。与循环网络结构一样，网络传播过程中可以学习到历史信息和未来信息。不同的是，FSMN采用非循环的前馈结构，不需要像循环网络结构那样必须等待语音输入结束才能对当前语音帧计算，其只需等待有限长度的未来语音帧输入即可。本文结合TDNN与FSMN的优点，将其融合应用于蒙古语语音识别声学模型。

如图4所示，TDNN与FSMN交替融合，包含六个隐藏层。在TDNN隐藏层中，使用$\{-n,m\}$表示将当前帧的历史第 n 帧、当前帧的未来第 m 帧和

当前帧拼接在一起作为下个网络层的输入。假设 t 表示当前帧，在 TDNN1（隐藏层 1），将帧 $\{t-2, t-1, t, t+1, t+2\}$ 拼接在一起作为下一个隐藏层的输入。在 TDNN2 和 TDNN3 处，将帧 $\{t-3, t+3\}$ 拼接在一起作为下一个隐藏层的输入。因此，在网络的最高层，至少可以学习到上下文相关的 8 帧历史信息及 8 帧未来信息。

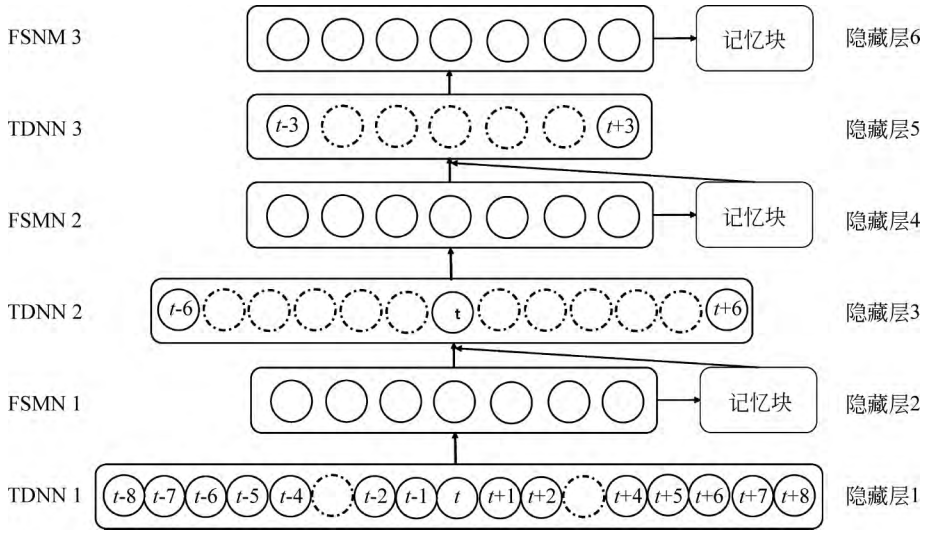

图 4　TDNN-FSMN 结构图

2　实验设置

2.1　实验语料

本文采用的蒙古语语音库是由 193 个说话人录制完成，其中采样率设为 16kHz，每采样点进行 16bit 量化，声道为单声道。语音库包含 69 781 句蒙古语朗读语音数据，总时长大约有 78h，每句话时长为 5～10s。实验中随机选择 88% 的语音数据作为训练集，12% 的语音数据作为测试集。发音词典由 38 107 个单词列表构成。对于语言模型，本文从蒙古语网站搜集大约 8 500 万单词的文本进行 3-gram 语言模型训练。

2.2　语音识别系统建立及评测

本文基于 Kaldi[15] 语音识别开发平台搭建了蒙古语语音识别系统。采用 MFCC 作为识别的特征参数。同时，对语音特征进行倒谱均值方差归一化（Cepstrum Mean Variance Normalization，CMVN）使得带噪语音特征参数的概率密度函数（Probability Density Function，PDF）更接近于纯净语音的概率密度函数，以减少训练语料与测试语料环境的不匹配度。之后使用线性判别分析与最大似然线性变换结合（Linear Discriminant Analysis-Maximum Likelihood Linear Transform，LDA-MLLT）将归一化后的上下文包含 7 帧（即 ±3）的高维特征进行区分性投影来降低特征向量维数至 40 维，保留具有分辨率的特征成分并使其集中在对角线上，以满足对声学模型在影响最小的情况下构建对角矩阵[16]。最后，使用基于特征空间最大似然线性回归（feature space Maximum Likelihood Linear Regression，fMLLR）进行说话人自适应训练，将 fMLLR 特征用于训练 DNN，TDNN，FSMN 和 TDNN-FSMN。

传统神经网络进行非线性运算时通常采用 Sigmoid，Tanh 函数作为激活函数。然而，文献[17]研究表明，修正线性单元（Rectified Linear Unit，ReLU）作为激活函数可以提高神经网络的性能。在本文中，所有神经网络的训练都使用 ReLU 非线性激活函数。

实验中采用的评价指标为国际通用的 WER 计算方式，具体如式（3）所示。

$$\mathrm{WER}=\frac{S+D+I}{T}\times 100\% \tag{3}$$

式中，S 代表替换错误词数，D 代表删除错误词数，I 代表插入错误词数，T 为句子中的总词数。WER 结果越小，表示识别性能越好。

3 实验与分析

3.1 不同神经网络的比较实验

在DNN-HMM声学模型训练中，首先对GMM-HMM训练得到的识别结果进行强制对齐，获得上下文相关的三音素状态作为声学模型训练的标签信息，共计3 762个独立的上下文相关状态，对应于DNN声学模型的输出维度。DNN的输入为15帧固定上下文窗口(即±7)，每帧提取40维MFCC特征，共计600维特征向量。实验中DNN模型包含六个隐藏层，每个隐藏层节点数为2 048个。使用基于RBM预训练方法逐层初始化DNN。小批量尺寸固定为256，初始和最终学习率参数分别设定为0.05和0.008。通过mini-batch随机梯度下降算法进行迭代更新，mini-batch大小为256，学习率在最初几次迭代中保持不变，当训练的准确率在两次迭代中没有太大的变化时，将学习率减少并进行下次迭代。

TDNN声学模型包含六个隐藏层，每个隐藏层包含512个节点。其输入为5帧固定上下文窗口(即±2)，每帧提取40维MFCC特征，共计200维特征向量。六个隐藏层的配置为{0}，{−1,1}，{−1,1}，{−3,3}，{−3,3}，{−6,3}，其中{0}表示常规的非拼接隐藏层。初始和最终学习率分别设置为0.001和0.0001。

FSMN声学模型包含六个隐藏层，每个隐藏层为512个节点，其中前三个隐藏层包含“记忆”模块，后三个隐藏层为常规隐藏层。实验中同样提取40维MFCC特征，由于FSMN的固有存储机制，不需要连续太多的语音帧序列作为输入，因此3帧固定上下文窗口(即±1)，共计120维特征向量作为FSMN的输入特征。“记忆”模块中包含5帧历史信息和5帧未来信息。FSMN在训练过程中被随机初始化，不用任何预训练方法。模型训练过程中更新策略同DNN训练参数设置保持一致。

TDNN-FSMN包含六个隐藏层。第一个隐藏层为包含512个节点的TDNN，输入特征为5帧固定上下文窗口(即±2)，共计200维特征向量。第二、四和六隐藏层为包含512个节点的FSMN，“记忆”模块中包含5帧历史信息和5帧未来信息。第三和五隐藏层是TDNN，隐藏层配置信息为{−3,3}，FSMN隐藏层输出共记1 536个输出状态作为其输入。

表1显示了在蒙古语语音数据集训练的基于DNN，TDNN，FSMN和TDNN-FSMN声学模型的识别结果。实验中调节DNN模型为最优性能，每个隐藏层包含2 048个节点，其他三种神经网络结构隐藏层节点数设置为512。从实验结果可以看出，TDNN-FSMN得到的识别性能明显优于最优性能的基线DNN模型，WER从12.90%下降到12.00%，表明基于TDNN-FSMN的声学模型在蒙古语语音识别中有显著提升。

表1 不同声学模型对比实验结果

模型	WER/%
DNN	12.90
TDNN	12.42
FSMN	12.74
TDNN-FSMN	**12.00**

3.2 FSMN隐藏层不同结构的对比实验

本文对TDNN-FSMN中FSMN隐藏层“记忆”模块中包含历史信息和未来信息的帧数对蒙古语语音识别性能的影响进行了对比实验。其中，TDNN-FSMN网络结构包含六个隐藏层，每个隐藏层为512个节点。在实验中，TDNN-FSMN_5h_5f表示“记忆”模块中包含5帧历史信息和5帧未来信息，TDNN-FSMN_5h_4f表示“记忆”模块中包含5帧历史信息和4帧未来信息。模型训练过程中更新策略与基线实验TDNN-FSMN训练参数设置保持一致。

表2 FSMN隐藏层不同结构对比实验结果

模型	WER/%
TDNN-FSMN_2h_5f	12.32
TDNN-FSMN_3h_5f	12.24
TDNN-FSMN_4h_5f	12.16
TDNN-FSMN_5h_2f	12.28
TDNN-FSMN_5h_3f	12.19
TDNN-FSMN_5h_4f	12.11
TDNN-FSMN_5h_5f	**12.00**

从表2的实验结果可以看出，“记忆”模块中包含5帧历史信息和5帧未来信息，表现出的性能最优。这是因为“记忆”模块包含历史信息帧和未来信

息帧的数量增加，将使 TDNN-FSMN 在训练过程中可以获得更多固定长度的时间上下文关联信息。而且，“记忆”模块中包含相同数量帧时，包含较多数量的历史信息帧比包含较多数量的未来信息帧表现得性能更优，表明上下文相关的历史信息对网络的性能更加有利。

3.3　TDNN-FSMN 不同结构的对比实验

在本实验中，分别对 TDNN-FSMN 中包含隐藏层的个数和隐藏层的节点数进行对比实验，其中 FSMN 隐藏层中“记忆”模块包含 5 帧历史信息和 5 帧未来信息。实验中分别设置隐藏层个数为 6、9 和 12，每个隐藏层分别包含 256、512 和 1 024 个节点。当隐藏层个数为 6 时，第 2、4 和 5 层为 FSMN 隐藏层；当隐藏层个数为 9 时，第 3、6 和 9 层为 FSMN 隐藏层；当隐藏层个数为 12 时，第 4、8 和 12 层为 FSMN 隐藏层。其余层均为 TDNN 隐藏层，其配置信息如表 3 所示，第一列表示隐藏层中使用到的降采样节点配置信息，第二列表示每个隐藏层中使用第一列的信息。例如，6-1 表示神经网络包含 6 个隐藏层，第一个隐藏层为 TDNN，降采样使用的节点数为{－2，－1，0，1，2}。使用 TDNN-FSMN-6L-256c 表示包含 6 个隐藏层，每个隐藏层包含 256 个节点。

表 3　TDNN 隐藏层配置信息

降采样信息	TDNN 隐藏层使用到的降采样信息
{－2，－1，0，1，2}	6-1，9-1，12-1
{－1，1}	6-3，9-2，12-2，12-3
{－3，3}	6-5，9-4，9-5，9-7，9-8，12-5，12-6，12-7，12-9，12-10，12-11

实验结果如图 5 所示，随着隐藏层个数增加及隐藏层节点数增加，单词错误率明显降低。这是因为随着层数和节点数的增加，将使 TDNN-FSMN 在训练过程中可以获得更多固定长度的时间上下文关联信息。最终，TDNN 融合 FSMN 的神经网络结构在蒙古语语音识别声学模型中比最优的基线 DNN 模型有很大的性能提升。其中使用 TDNN-FSMN-12L-1024c 网络结构得到的实验结果最好，单词错误率为 10.03%，与基线 DNN 模型相比相对降低 22.2%，表明 TDNN-FSMN 能有效提升蒙古语语音识别的性能。然而，TDNN-FSMN-6L-256c 网络结构识别准确率较基线 DNN 模型有所降低，由于参数规模降低，会使得 TDNN-FSMN 在训练过程中无法学习到足够的声学信息进而降低了声学模型的性能。

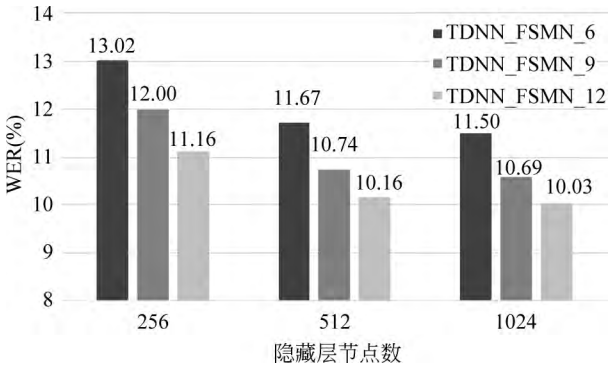

图 5　TDNN-FSMN 不同结构对比实验结果

4　总结

本文首次将融合的 TDNN-FSMN 模型应用于蒙古语语音识别中，实验结果表明，TDNN-FSMN 可以获得比 DNN 更好的性能。在不同结构 FSMN 隐藏层中，“记忆”模块包含 5 帧历史信息和 5 帧未来信息表现得性能最优，单词错误率较基线 DNN 模型相对降低 7.0%。此外，通过对 TDNN-FSMN 中包含隐藏层的个数和隐藏层的节点数进行对比实验，发现随着层数和节点数的增加，TDNN-FSMN 的性能明显提升，表明 TDNN-FSMN 在训练过程中可以获得更多固定长度的时间上下文关联信息。最终，包含 12 个隐藏层且每个隐藏层包含 1 024 个节点得到的实验结果最优，相比基线 DNN 模型，单词错误率相对降低 22.2%。最终蒙古语语音识别系统词错误率达到了 10.03%，表明基于 TDNN-FSMN 神经网络结构能有效地提升蒙古语语音识别性能。

参考文献

[1]　何珏，刘加. 汉语连续语音中 HMM 模型状态数优化方法研究[J]. 中文信息学报，2006，20(6)：83-88.

[2]　Hinton G，Deng L，Dong Y，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.

[3]　Pan J，Liu C，Wang Z，et al. Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition：Why DNN surpasses GMMS in a-

coustic modeling [C]//Proceedings of the 8th International Symposium on Chinese Spoken Language Processing,2012: 301-305.

[4] Waibel A,Hanazawa T,Hinton G,et al. Phoneme recognition using time-delay neural networks[J]. IEEE Transactions on Acoustics,Speech,and Signal Processing,1989,37(3),328-339.

[5] Peddinti V,Povey D,Khudanpur S. A time delay neural network architecture for efficient modeling of long temporal contexts[C]//Proceedings of 16th INTERSPEECH,2015: 3214-3218.

[6] Zhang S L,Jiang H,Wei S,et al. Feedforward sequential memory neural networks without recurrent feedback[J]. arXiv: 1510.02693.2015.

[7] Zhang S,Liu C,Jiang H,et al. Feedforward sequential memory networks: A new structure to learn long-term dependency [J]. arXiv: 1512.08301.2015

[8] Gao G L,Zhang S. A Mongolian speech recognition system based on HMM[C]//Proceedings of International Conference on Intelligent Computing,2006: 667-676.

[9] Qilao H,Gao G L. Researching of speech recognition oriented Mongolian acoustic model[C]//Proceedings of 2008 Chinese Conference on Pattern Recognition (CCPR),2008: 1-6.

[10] Bao F,Gao G L. Improving of acoustic model for the Mongolian speech recognition system[C]//Proceedings of 2009 Chinese Conference on Pattern Recognition(CCPR),2009: 1-5.

[11] **飞龙,高光来,王宏伟. 基于词干的蒙古语语音关键词检测方法的研究[J]. 中文信息学报**,2016,30(1): 124-128.

[12] Bao F,Gao G L,Yan X,et al. Segmentation-based Mongolian LVCSR approach [C]//Proceedings of 38th ICASSP,2013: 1-5.

[13] Zhang H,Bao F,Gao G L. Mongolian speech recognition based on deep neural networks[C]//Proceedings of 15th Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data,2015: 180-188.

[14] Zhang H W,Bao F,Gao G L,et al. Comparison on neural network based acoustic model in Mongolian speech recognition[C]//Proceedings of 20th Asian Language Processing (IALP), 2016 International Conference,2016: 1-5.

[15] Povey D,Ghoshal A,Boulianne G,et al. The Kaldi speech recognition toolkit[C]//Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding Workshop,Hawaii, USA:IEEE, 2011.

[16] **肖云鹏,叶卫平. 基于特征参数归一化的鲁棒语音识别方法综述[J]. 中文信息学报**,2010,24(5): 106-117.

[17] Maas A L,Hannun A Y,Ng A Y. Rectifier nonlinearities improve neural network acoustic models[C]//Proceedings of 30th ICML Workshop on Deep Learning for Audio, Speech and Language Processing, 2013.

王勇和(1992—),男,硕士研究生,主要研究领域为蒙古文信息处理、语音识别。
E-mail: cswyh92@163.com

高光来(1964—),硕士,教授,博士生导师,主要研究领域为人工智能、模式识别、自然语言处理。
E-mail: csggl@imu.edu.cn

飞龙(1985—),通信作者,博士,副教授,硕士生导师,主要研究领域为蒙古文信息处理、语音识别、语音合成、语音检索、语义理解、信息检索、机器翻译。
E-mail: csfeilong@imu.edu.cn