阿里开源语音识别模型DFSMN

阿里妹导读：近日，阿里巴巴达摩院机器智能实验室开源了新一代语音识别模型DFSMN，将全球语音识别准确率纪录提高至96.04%（这一数据测试基于世界最大的免费语音识别数据库LibriSpeech）。

对比目前业界使用最为广泛的LSTM模型，DFSMN模型训练速度更快、识别准确率更高。采用全新DFSMN模型的智能音响或智能家居设备，相比前代技术深度学习训练速度提到了3倍，语音识别速度提高了2倍。

开源地址：

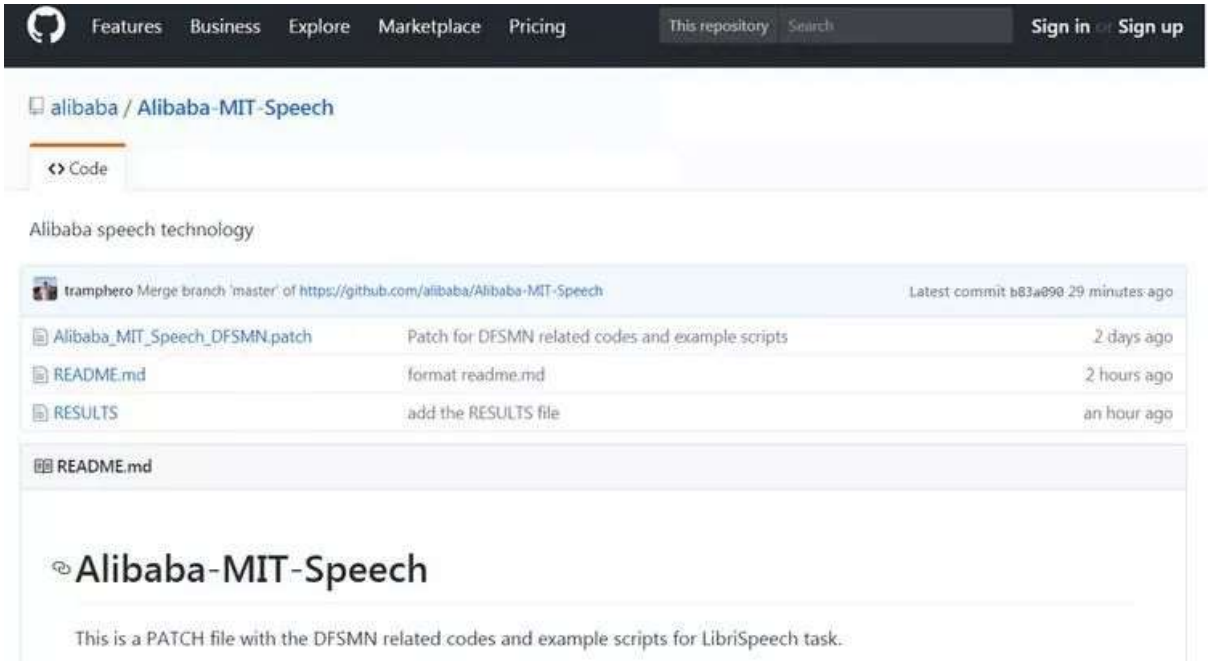
https://github.com/tramphero/kaldi

本文作者：张仕良

阿里开源语音识别模型DFSMN

在近期举行的云栖大会武汉峰会上，装有DFSMN语音识别模型的“AI收银员”在与真人店员的PK中，在嘈杂环境下准确识别了用户的语音点单，在短短49秒内点了34杯咖啡。此外，装备这一语音识别技术的自动售票机也已在上海地铁“上岗”。

著名语音识别专家，西北工业大学教授谢磊表示：“阿里此次开源的DFSMN模型，在语音识别准确率上的稳定提升是突破性的，是近年来深度学习在语音识别领域最具代表性的成果之一，对全球学术界和AI技术应用都有巨大影响。”



图：阿里在GitHub平台上开源了自主研发的DFSMN语音识别模型

语音识别声学模型

语音识别技术一直都是人机交互技术的重要组成部分。有了语音识别技术，机器就可以像人类一样听懂说话，进而能够思考、理解和反馈。

近几年随着深度学习技术的使用，基于深度神经网络的语音识别系统性能获得了极大的提升，开始走向实用化。基于语音识别的语音输入、语音转写、语音检索和语音翻译等技术得到了广泛的应用。

目前主流的语音识别系统普遍采用基于深度神经网络和隐马尔可夫（Deep Neural Networks-Hidden Markov Model，DNN-HMM）的声学模型，其模型结构如图 1所示。声学模型的输入是传统的语音波形经过加窗、分帧，然后提取出来的频谱特征，如 PLP，MFCC 和 FBK等。而模型的输出一般采用不同粒度的声学建模单元，例如单音素 (mono-phone)、单音素状态、绑定的音素状态 (tri-phonestate) 等。从输入到输出之间可以采用不同的神经网络结构，将输入的声学特征映射得到不同输出建模单元的后验概率，然后再结合HMM进行解码得到最终的识别结果。

最早采用的网络结构是前馈全连接神经网路（Feedforward Fully-connected Neural Networks, FNN）。FNN实现固定输入到固定输出的一对一映射，其存在的缺陷是没法有效利用语音信号内在的长时相关性信息。一种改进的方案是采用基于长短时记忆单元（Long-

Short Term Memory，LSTM）的循环神经网络（Recurrent Neural Networks，RNN）。LSTM-RNN通过隐层的循环反馈连接，可以将历史信息存储在隐层的节点中，从而可以有效地利用语音信号的长时相关性。

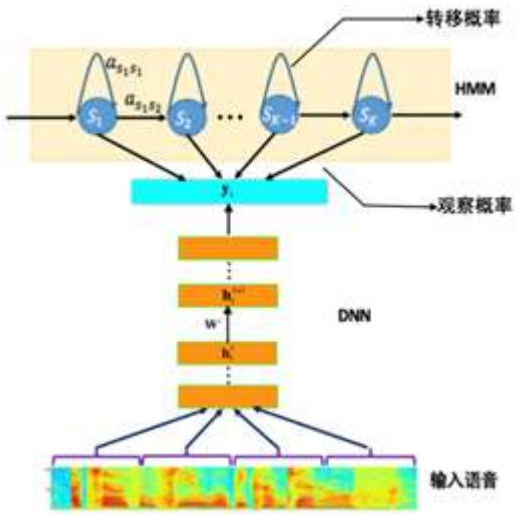


图 1. 基于DNN-HMM的语音识别系统框图

进一步地通过使用双向循环神经网络（BidirectionalRNN），可以有效地利用语音信号历史以及未来的信息，更有利于语音的声学建模。基于循环神经网络的语音声学模型相比于前馈全连接神经网络可以获得显著的性能提升。但是循环神经网络相比于前馈全连接神经网络模型更加复杂，往往包含更多的参数，这会导致模型的训练以及测试都需要更多的计算资源。

另外基于双向循环神经网络的语音声学模型，会面临很大的时延问题，对于实时的语音识别任务不适用。现有的一些改进的模型，例如，基于时延可控的双向长短时记忆单元（Latency Controlled LSTM，LCBLSTM）[1-2]，以及前馈序列记忆神经网络（Feedforward SequentialMemory Networks，FSMN）[3-5]。去年我们在工业界第一个上线了基于LCBLSTM的语音识别声学模型。配合阿里的大规模计算平台和大数据，采用多机多卡、16bit量化等训练和优化方法进行声学模型建模，取得了相比于FNN模型约17-24%的相对识别错误率下降。

FSMN模型的前世今生

1. FSMN模型

FSMN是近期被提出的一种网络结构，通过在FNN的隐层添加一些可学习的记忆模块，从而可以有效地对语音的长时相关性进行建模。FSMN相比于LCBLSTM不仅可以更加方便地控制时延，而且也能获得更好的性能，需要的计算资源也更少。但是标准的FSMN很难训练非常深的结构，会由于梯度消失问题导致训练效果不好。而深层结构的模型目前在很多领域被证明具有更强的建模能力。因而针对此我们提出了一种改进的FSMN模型，称之为深层的FSMN（DeepFSMN, DFSMN）。进一步地我们结合LFR（lowframe rate）技术构建了一种高效的实时语音识别声学模型，相比于去年我们上线的LCBLSTM声学模型可以获得超过20%的相对性能提升，同时可以获得2-3倍的训练以及解码的加速，可以显著地减少我们的系统实际应用时所需要的计算资源。

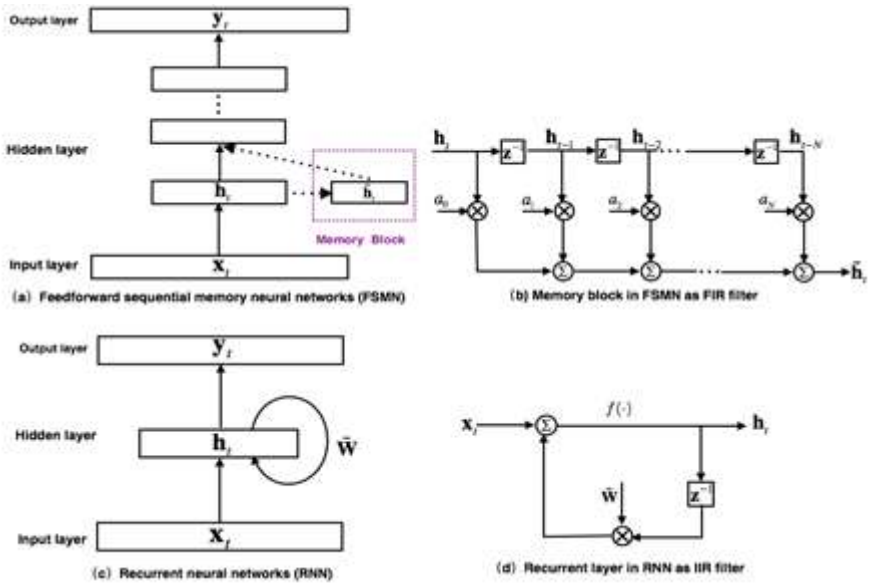


图 2. FSMN模型结构以及和RNN的对比

2. FSMN到cFSMN的发展历程

最早提出的FSMN的模型[3]结构如图 2（a）所示，其本质上是一个前馈全连接神经网络，通过在隐层旁添加一些记忆模块（memory block）来对周边的上下文信息进行建模，从而使得模型可以对时序信号的长时相关性进行建模。记忆模块采用如图 2（b）所示的抽头延迟结构将当前时刻以及之前 N 个时刻的隐层输出通过一组系数编码得到一个固定的表达。FSMN的提出是受到数字信号处理中滤波器设计理论的启发：任何无限响应冲击（Infinite Impulse Response, IIR）滤波器可以采用高阶的有限冲击响应（Finite Impulse Response, FIR）滤波器进行近似。从滤波器的角度出发，如图 2（c）所示的RNN模型的循环层就可以看作如图 2（d）的一阶IIR滤波器。而FSMN采用的采用如图 2（b）所示的记忆模块可以看作是一个高阶的FIR滤波器。从而FSMN也可以像RNN一样有效地对信号的长时相关性进行建模，同时由于FIR滤波器相比于IIR滤波器更加稳定，因而FSMN相比于RNN训练上会更加简单和稳定。

根据记忆模块编码系数的选择，可以分为：1）标量FSMN（sFSMN）；2）矢量FSMN（vFSMN）。sFSMN 和 vFSMN顾名思义就是分别使用标量和矢量作为记忆模块的编码系数。sFSMN和vFSMN记忆模块的表达分别如下公式：

$$\tilde{\mathbf{h}}_t^l = \sum_{i=0}^{N} a_i^l \cdot \mathbf{h}_{t-i}^l$$

$$\tilde{\mathbf{h}}_t^l = \sum_{i=0}^{N} \boldsymbol{a}_i^l \odot \mathbf{h}_{t-i}^l$$

以上的FSMN只考虑了历史信息对当前时刻的影响，我们可以称之为单向的FSMN。当我们同时考虑历史信息以及未来信息对当前时刻的影响时，我们可以将单向的FSMN进行扩展得到双向的FSMN。双向的sFSMN和vFSMN记忆模块的编码公式如下：

$$\tilde{\mathbf{h}}_t^l = \sum_{i=0}^{N_1} a_i^l \cdot \mathbf{h}_{t-i}^l + \sum_{j=1}^{N_2} c_j^l \cdot \mathbf{h}_{t+j}^l$$

$$\tilde{\mathbf{h}}_t^l = \sum_{i=0}^{N_1} \boldsymbol{a}_i^l \odot \mathbf{h}_{t-i}^l + \sum_{j=1}^{N_2} \boldsymbol{c}_j^l \odot \mathbf{h}_{t+j}^l$$

这里

N_1

和

N_2

分别代表回看(look-back)的阶数和向前看(look-ahead)的阶数。我们可以通过增大阶数，也可以通过在多个隐层添加记忆模块来增强FSMN对长时相关性的建模能力。

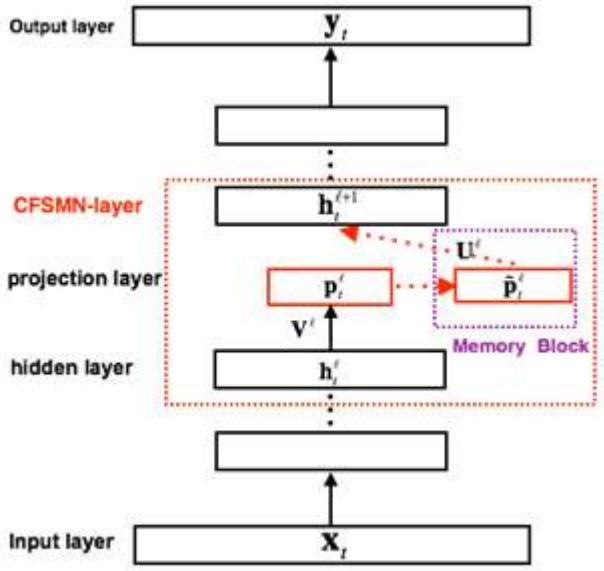


图 3. cFSMN结构框图

FSMN相比于FNN，需要将记忆模块的输出作为下一个隐层的额外输入，这样就会引入额外的模型参数。隐层包含的节点越多，则引入的参数越多。研究[4]结合矩阵低秩分解（Low-rank matrix factorization）的思路，提出了一种改进的FSMN结构，称之为简洁的FSMN（CompactFSMN，cFSMN），是一个第/个隐层包含记忆模块的cFSMN的结构框图。

对于cFSMN，通过在网络的隐层后添加一个低维度的线性投影层，并且将记忆模块添加在这些线性投影层上。进一步的，cFSMN对记忆模块的编码公式进行了一些改变，通过将当前时刻的输出显式地添加到记忆模块的表达中，从而只需要将记忆模块的表达作为下一层的输入。这样可以有效得减少模型的参数量，加快网络的训练。具体单向和双向的cFSMN记忆模块的公式表达分别如下：

$$\tilde{\mathbf{P}}_t^l = \mathbf{P}_t^l + \sum_{i=0}^{N} \boldsymbol{a}_i^l \odot \mathbf{P}_{t-i}^l$$

$$\tilde{\mathbf{P}}_t^l = \mathbf{P}_t^l + \sum_{i=0}^{N_1} \boldsymbol{a}_i^l \odot \mathbf{P}_{t-i}^l + \sum_{j=0}^{N_2} \boldsymbol{c}_j^l \odot \mathbf{P}_{t+j}^l$$

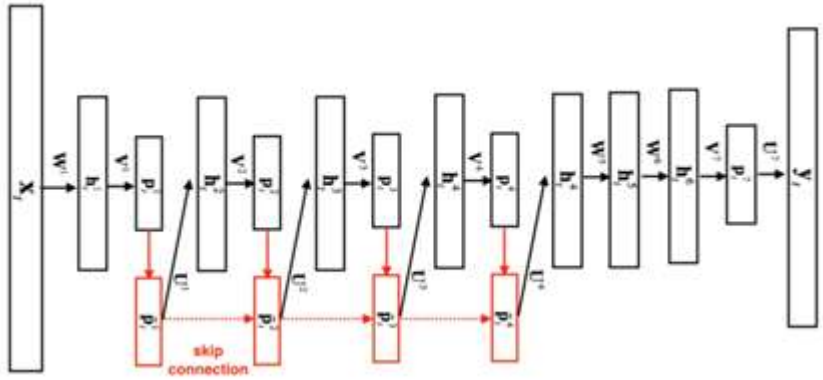


图 4. Deep-FSMN (DFSMN)模型结构框图

LFR-DFSMN声学模型

1. Deep-FSMN (DFSMN)网络结构

如图 4是我们进一步提出的Deep-FSMN（DFSMN）的网络结构框图，其中左边第一个方框代表输入层，右边最后一个方框代表输出层。我们通过在cFSMN的记忆模块（红色框框表示）之间添加跳转连接（skip connection），从而使得低层记忆模块的输出会被直接累加到高层记忆模块里。这样在训练过程中，高层记忆模块的梯度会直接赋值给低层的记忆模块，从而可以克服由于网络的深度造成的梯度消失问题，使得可以稳定地训练深层的网络。我们对记忆模块的表达也进行了一些修改，通过借鉴扩张（dilation）卷积[6]的思路，在记忆模块中引入一些步幅（stride）因子，具体的计算公式如下：

$$\tilde{\boldsymbol{P}}_t^l = \tilde{\boldsymbol{P}}_t^{l-1} + \boldsymbol{P}_t^l + \sum_{i=0}^{N_1} \boldsymbol{a}_i^l \odot \boldsymbol{P}_{t-s_1*i}^l + \sum_{j=1}^{N_2} \boldsymbol{c}_i^l \odot \boldsymbol{P}_{t+s_2*j}^l$$

其中

$\tilde{\boldsymbol{P}}_t^{l-1}$

表示第

$l-1$

层记忆模块第t个时刻的输出。S1和S2分别表示历史和未来时刻的编码步幅因子，例如S1=2则表示对历史信息进行编码时每隔一个时刻取一个值作为输入。这样在相同的阶数的情况下可以看到更远的历史，从而可以更加有效的对长时相关性进行建模。

对于实时的语音识别系统我们可以通过灵活的设置未来阶数来控制模型的时延，在极端情况下，当我们将每个记忆模块的未来阶数都设置为0，则我们可以实现无时延的一个声学模型。对于一些任务，我们可以忍受一定的时延，我们可以设置小一些的未来阶数。

相比于之前的cFSMN，我们提出的DFSMN优势在于，通过跳转连接可以训练很深的网络。对于原来的cFSMN，由于每个隐层已经通过矩阵的低秩分解拆分成了两层的结构，这样对于一个包含4层cFSMN层以及两个DNN层的网络，总共包含的层数将达到13层，从而采用更多的cFSMN层，会使得层数更多而使得训练出现梯度消失问题，导致训练的不稳定性。我们提出的DFSMN通过跳转连接避免了深层网络的梯度消失问题，使得训练深层的网络变得稳定。需要说明的是，这里的跳转连接不仅可以加到相邻层之间，也可以加到不相邻层之间。跳转连接本身可以是线性变换，也可以是非线性变换。具体的实验我们可以实现训练包含数十层的DFSMN网络，并且相比于cFSMN可以获得显著的性能提升。

从最初的FSMN到cFSMN不仅可以有效地减少模型的参数，而且可以获得更好的性能[4]。进一步的在cFSMN的基础上，我们提出的DFSMN，可以更加显著地提升模型的性能。如下表是在一个2000小时的英文任务上基于BLSTM，cFSMN，DFSMN的声学模型性能对比。

Model	BLSTM	cFSMN	DFSMN
WER%	10.9	10.8	9.4

从上表中可以看到，在2000小时这样的任务上，DFSMN模型可以获得比BLSTM声学模型相对14%的错误率降低，显著提高了声学模型的性能。

2. 基于LFR-DFSMN的语音识别声学模型

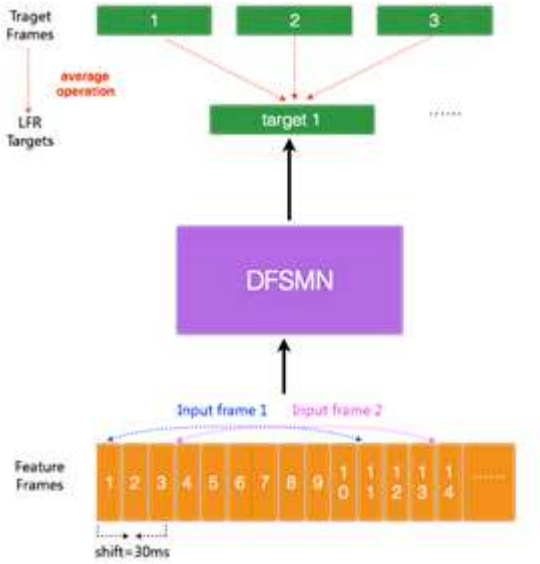


图 5. LFR-DFSMN声学模型结构框图

目前的声学模型，输入的是每帧语音信号提取的声学特征，每帧语音的时长通常为10ms，对于每个输入的语音帧信号会有相对应的一个输出目标。最近有研究提出一种低帧率（LowFrame Rate，LFR）[7]建模方案：通过将相邻时刻的语音帧进行绑定作为输入，去预测这些语音帧的目标输出得到的一个平均输出目标。具体实验中可以实现三帧（或更多帧）拼接而不损失模型的性能。从而可以将输入和输出减少到原来的三分之一甚至更多，可以极大地提升语音识别系统服务时声学得分的计算以及解码的效率。我们结合LFR和以上提出的DFSMN，构建了如图 5的基于LFR-DFSMN的语音识别声学模型，经过多组实验我们最终确定了采用一个包含10层cFSMN层+2层DNN的DFSMN作为声学模型，输入输出则采用LFR，将帧率降低到原来的三分之一。识别结果和去年我们上线的最好的LCBLSTM基线比较如下表所示。

CER%	产品线A	产品线B
LFR-LCBLSTM	18.92	10.21
LFR-DFSMN	15.00 (+20.72%)	8.04 (21.25%)

通过结合LFR技术，我们可以获得三倍的识别加速。从上表中可以看到，在实际工业规模应用上，LFR-DFSMN模型比LFR-LCBLSTM模型可以获得20%的错误率下降，展示了对大规模数据更好的建模特性。

基于多机多卡的大数据声学模型训练

实际的语音识别服务通常会面对非常复杂的语音数据，语音识别声学模型一定要尽可能地覆盖各种可能的场景，包括各种对话、各种声道、各种噪音甚至各种口音，这就意味着海量的数据。而如何应用海量数据快速训练声学模型并上线服务，就直接关系到业务相应速度。

我们利用阿里的Max-Compute计算平台和多机多卡并行训练工具，在使用8机16GPU卡、训练数据为5000小时的情况下，关于LFR-DFSMN声学模型和LFR-LCBLSTM的训练速度如下表：

	处理一个epoch需要的时间
LFR-LCBLSTM	10.8小时
LFR-DFSMN	3.4小时

相比于基线LCBLSTM模型，每个epoch DFSMN可以获得3倍的训练速度提升。在2万小时的数据量上训练LFR-DFSMN，模型收敛一般只需要3-4个epoch，因此在16GPU卡的情况下，我们可以在2天左右完成2万小时数据量的LFR-DFSMN声学模型的训练。

解码延时、识别速度和模型大小

设计更为实用化的语音识别系统，我们不仅需要尽可能地提升系统的识别性能，而且需要考虑系统的实时性，这样才能给用户提供更好的体验。此外在实际应用中我们还需要考虑服务成本，因而对于语音识别系统的功耗也有一定的要求。传统的FNN系统，需要使用拼帧技术，解码延迟通常在5-10帧，大约50-100ms。而去年上线的LCBLSTM系统，解决了BLSTM的整句延迟的问题，最终可以将延时控制在20帧左右，大约200ms。对于一些对延时有更高要求的线上任务，还可以在少量损失识别性能的情况下（0.2%-0.3%绝对值左右），将延迟控制在100ms，完全可以满足各类任务的需求。LCBLSTM相比于最好的FNN可以获得超过20%的相对性能提升，但是相同CPU上识别速度变慢（即功耗高），这主要是由模型的复杂度导致。

我们最新的LFR-DFSMN，通过LFR技术可以将识别速度加速3倍以上，进一步的DFSMN相比于LCBLSTM在模型复杂度上可以再降低3倍左右。如下表是我们在一个测试集上统计的不同的模型需要的识别时间，时间越短则表示我们所需要的计算功耗越低：

模型	整个测试集识别所需要的时间
LCBLSTM	956秒
DFSMN	377秒
LFR-LCBLSTM	339秒
LFR-DFSMN	142秒

关于LFR-DFSMN的解码时延问题，我们可以通过减小记忆模块滤波器向未来看的阶数来减小时延。具体实验中我们验证了不同的配置，当我们将LFR-DFSMN的延时控制在5-10帧时，大致只损失相对3%的性能。

此外，相对于复杂的LFR-LCBLSTM模型，LFR-DFSMN模型具有模型精简的特点，虽然有10层DFSMN，但整体模型大小只有LFR-LCBLSTM模型的一半，模型大小压缩了50%。

参考文献：

1. YuZhang, Guoguo Chen, Dong Yu, and Kaisheng Yao, ng Yao, long short term memory RNNs for distantspeech recognition,, in IEEE International Conference of Acoustics,Speech andSignal Processing (ICASSP), 2016, pp. 5755-5759.

2. XueS, Yan Z. Improving latency-controlled BLSTM acoustic models for online speech recognition[C]//Acoustics,Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on.IEEE. 2017.

3. Zhang S, Liu C, Jiang H, et al. Feedforwardsequential memory networks: A new structure to learn long-term dependency[J].arXiv preprint arXiv:1512.08301, 2015.

4. Zhang S, Jiang H, Xiong S, et al. CompactFeedforward Sequential Memory Networks for Large Vocabulary Continuous SpeechRecognition[C]//INTERSPEECH. 2016: 3389-3393.

5. Zhang S, Liu C, Jiang H, et al. Non-recurrentNeural Structure for Long-Term Dependency[J]. IEEE/ACM Transactions on Audio,Speech, and Language Processing, 2017, 25(4): 871-884.

6. Oord A, Dieleman S, Zen H, et al. Wavenet:A generative model for raw audio[J]. arXiv preprint arXiv:1609.03499, 2016.

7. Pundak G, Sainath T N. Lower Frame Rate NeuralNetwork Acoustic Models[C]//INTERSPEECH. 2016: 22-26.

End