基于深度学习的中文语音合成技术研究

**赵建民 王雨萌 王梅**

(东北石油大学计算机与信息技术学院 大庆市 163000)

摘要：语音合成技术在近些年得到较快发展，但是在相似性和自然度等方面还是存在一些问题，合成出的语音韵律差且机械音明显，使得其无法满足实际应用的要求。将深度学习技术应用于语音合成，目前已经可以表现出和基于隐马尔科夫模型以及基于语音拼接的语音合成技术相当的合成性能，达到实用的效果。由于其拥有强大的建模能力，而且有更好的灵活性、易控制性，深度学习技术在提升语音合成性能方面有很大的潜力和研究价值。其中，递归神经网络(RNN)、长短时记忆神经网络(LSTM)以及门阀递归单元神经网络(GRNN)是几项典型的深度学习模型，普遍被应用于语音识别、机器翻译时序建模方面，并且效果较为理想。通过多组对比实验，展示这些深度学习技术在语音合成方面的有效性。

关键词 深度神经网络 语音合成 长短时记忆 门阀递归单元

中图法分类号 TP311 文献标识码 A

**Research on Technology of Mandarin Speech Synthesis Based on Deep Neural Network**

Jianmin Zhao , Yumeng Wang, Mei Wang

(School of Computer&Information Technology，Northeast Petroleum University， Daqing 163000，China)

**Abstract：**Speech synthesis has been developed rapidly these years, but there are still some problems, especially in similarity and naturalness. Deep learning technologies have been used to speech synthesis, and got comparative performance as those based on HMM and waveform concatenation.Because deep learning technologies have very powerful modeling ability, and they are quite flexible and easy to control, they are very promising in improving speech synthesis and have high research value. Recurrent Neural Network(RNN), Long Short-term Memory Neural Network(LSTM) and Gated Recurrent Neural Network(GRNN) are typical deep models. And their effectiveness of speech synthesis is showed through several comparative experiments.

Keywords**：**Deep NeuralNetwork，speech synthesis，Long short-term memory,Gated recurrent unit

**引言**

语音合成技术主要分类两类：基于语音拼接和基于统计参数的语音合成方法。语音拼接语音合成需要建立较大的语音库，然后根据待合成文本的特征，从语音库中选择合适的语音片段，将语音片段进行拼接得到完整的合成语音。

基于统计参数的语音合成技术通过训练带参数的模型，再将本文的语言学特征作为输入，利用模型来预测相应的语音声学特征，最后借用声码器将声学特征合成得到语音。隐马尔科夫模型(Hidden Markov Model)就是统计参数语音合成技术的典型代表，在过去的几十年里得到广泛研究[1,2,3]。隐马尔科夫模型对时序信号有这很好的建模能力，因此，可以将其用作声学模型，用来解决语音合成中声学特征和语言学特征之间的复杂映射关系问题。并且，研究表明，隐马尔科夫模型在语音合成方面表现出了很好的性能。

但是隐马尔科夫模型也存在一些弊端。基于统计参数的语音合成技术是通过对文本特征和声学特征之间的关系进行建模，由于文本和语音都是时序信号，涉及到的文本特征必然是复杂的上下文相关的因素（[1]指出影响英语语音合成效果的上下文相关特征有50多种，包括音素学、语言学、语法学等），因此，对如此复杂的文本特征和声学特征构建其映射关系需要模型具有很强的表现能力。隐马尔科夫模型对每种上下文相关的特征独立构建模型，即上下文相关的隐马尔科夫模型。但是，由于无法在有限规模的训练数据集中包含所有可能的上下文相关的特征，因此导致上下文相关的隐马尔科夫模型不能涵盖所有特征情况。虽然，隐马尔科夫模型采用基于决策树的上下文聚类方法[3]来解决这个问题，但是，对于非常丰富而且更复杂的上下文关系，隐马尔科夫模型的表现能力还是不足。

神经网络就能更好地对复杂关系进行建模。随着深度学习技术的不断发展，将其应用于语音合成的研究工作也越来越多。深度前馈神经网络 (Deep Forward Neural Network)具有全连接特性，能直接对复杂映射关系进行建模[4,5,6]。

可以认为深度前馈神经网络替代隐马尔科夫模型中的决策树，模型表现能力比后者更强。但是，深度神经网络是一帧一帧地对文本特征和声学特征的映射关系进行建模的，完全没有考虑上下文的依赖关系，这跟语音和文本的时序特性不相符。递归神经网络(Recurrent Neural Network)能很好地对时序信号的上下文依赖关系进行建模。长短时记忆网络(LongShort-Term Memory NeuralNetwork)[7,8,10,12,13,14]和门阀递归神经网络(Gated Recurrent Neural Network)[9,11]是递归神经网络的变种。

本文会把这两种神经网络模型用于语音合成。

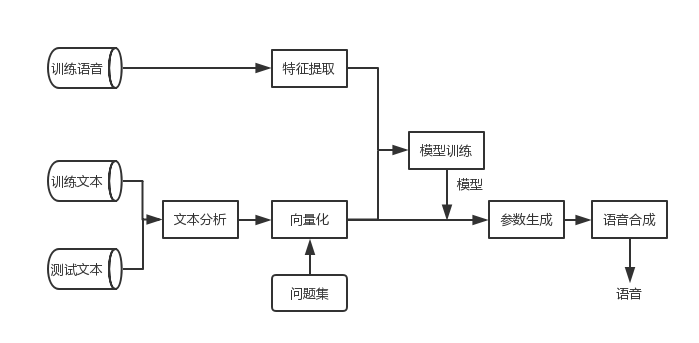
另外，在实验中，我们发现在神经网络层引入随机丢弃策略(dropout)对合成的语音效果有很大影响。

本文的组织结构如下：第二部分介绍中文语音合成的流程和相关技术；第三部分通过实验比较几种深度学习模型在中文语音合成上的表现；最后作总结。

**基于深度学习的中文语音合成技术**

2.1语音合成的基本流程

图1. 语音合成的流程



基于统计参数的语音合成，主要通过模型来描述语音的声学特征和文本的语言学特征之间的映射关系，它分为训练阶段和测试阶段（合成阶段）。在训练阶段，首先对训练语音提取声学特征，主要包括LF0、bap以及mgc等，作为模型的输出向量；对训练文本进行文本分析，得到上下文相关的标注，这将在下一部分详细介绍；借助设计好的问题集将标注转换为取值是0或1的二值型元素（比如对“该音素是否是元音”的回答结果）或者取值连续的元素（比如对“当前短语中的音素个数“的回答结果），共同构成模型的输入向量。对于深度学习模型，训练过程就是逐步迭代更新参数，以最小化预测输出和实际输出之间的误差。

2.2中文文本分析

对于中文文本，首先需要对其进行预处理，比如去除文本中的特殊标点符号（省略号、书名号等）、韵律符号的标准化处理、停顿的处理等。对经过预处理之后的文本进行分词，将整句文本拆分为中文词语并生成对应的发音。然后，统计文本的上下文相关的特征，  
在我们的实验中，文本特征主要涵盖音素、音节、词、短语、句子几个层面。另外，不仅统计当前音节、词和短语的信息，还统计前一个和后一个音节、词和短语的相关信息。统计信息包括：当前音素、前一音素、后一音素、当前音素在当前音节中的位置（包括前向和后向）、（前一音节、后一音节）当前音节的声调和所含的音素个数、词（短语、句子）中所含的音节数目等等。结合问题集，可以把统计的文本信息转化为数值向量，比如通过回答问题“当前音素是否是音素a”可以得到0（否）或者1（是），问题“当前短语中的音素个数”就可以得到连续型数值。

2.3深度学习模型

深度学习模型对复杂的映射关系有更好的建模能力。深度前馈神经网络是最普通、最简单的网络模型，它通常有多个隐含层，每一层都对其输入进行如下非线性变化：

其中，X，*ht*分别是隐含层的输入和输出，*Wx,h*是权重矩阵，*bh*是偏置，*f(\*)*是非线性函数。

经过网络输出层的输出是模型对输入的预测输出，该预测和输入对应的实际输出（标注或答案）之间存在误差，通过损失函数（比如均方误差函数）来计算该误差。通常采用随机梯度下降法来逐步调整网络参数（权值矩阵和偏置）的值来降低误差，当误差降低到一定范围内时，说明该模型对输入有了很好的预测能力。

将深度神经网络用于语音合成，就是先利用文本特征作为输入、对应的声学特征作为实际输出，训练模型参数，然后将待合成文本的特征向量输入网络，网络的输出作为声学参数用于声码器生成语音。

深度前馈网络虽然简单，而且在很多分类问题上有很好的表现，但是，它在处理语音和文本时，没有考虑其时序特性。递归神经网络就适合用来对时序信号之间的映射关系进行建模。长短时记忆网络和门阀递归神经网络是递归神经网络的典型代表。

2.3.1 长短时记忆神经网络

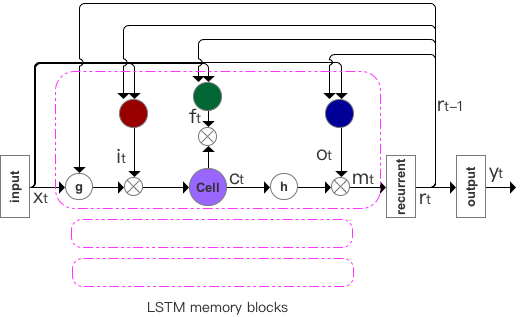
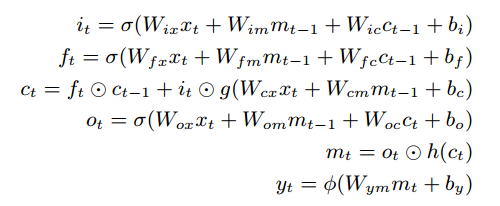


图2 长短时记忆网络记忆块

长短时记忆神经网络是一种递归神经网络，它采用带有输入门(input gate)、遗忘门(forget gate)和输出门(output gate)的记忆块作为隐含层的神经元,它能解决递归神经网络在训练网络时会出现的梯度衰减的问题，因此能学习更长时序上的依赖关系。记忆块如图2所示：其中，x是记忆块的输入，i、f、o、c分别表示输入门、输出门、遗忘门和记忆单元(cell)的激活向量；W表示连接权重（比如,W*ix*是输入层和输入门之间的权重），b表示偏置；是逻辑斯蒂回归激活函数，*g,h*分别表示单元(cell)输入和输出的激活函数，是输出层激活函数（比如softmax函数）。在长短时记忆神经网络中，输入、输出的关系如下：



可以看出，长短时记忆神经网络在预测某一时刻的输出时，不仅考虑该时刻的输入，还考虑了之前时刻的隐含层输出，即使用了前文的信息。双向长短时记忆神经网络（Bidirectional Long Short-Term Memory Neural Network）既能考虑前文的信息，又能考虑后文的信息。

2.3.3 门阀递归神经网络

门阀递归神经网络使用门阀递归单元(Gated Recurrent Unit)[11]代替长短时记忆神经网络中的记忆块，这样使得计算和实现变得更加简单。  
门阀递归单元结构如图3所示。

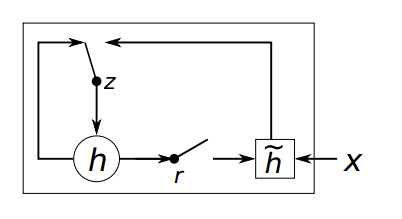
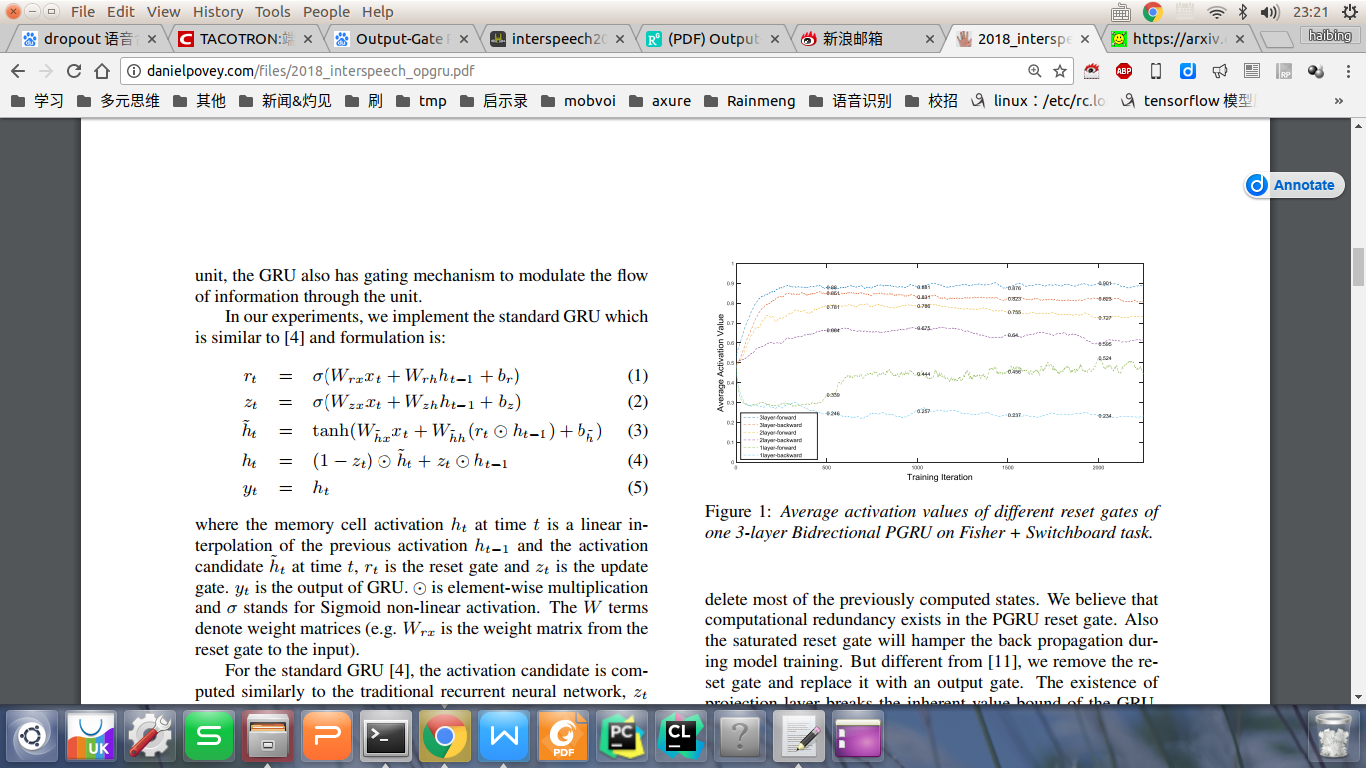


图3 GRU门阀递归单元

GRU公式如下：



其中，rt,zt,ht分别表示重置门、更新门和记忆单元激活值在第t时刻的取值，W和b表示权重和偏置。

**改进的模型--PGRU**

LSTM和GRU由于其递归特性，可以很好地长时的时序数据进行建模。但是，它们存在一个严重的缺陷，即结构复杂、速度慢。

在[15]中，Sak提出一个LSTM变种，在LSTM记忆块的输出神经元前面引入两个低维的映射层（神经元个数少），减小递归运算带来的计算量，而且在语音识别上比标准LSTM的效果要好。

受其启发，我们在GRU模型中引入映射层，称之为PGRU，如图４所示。

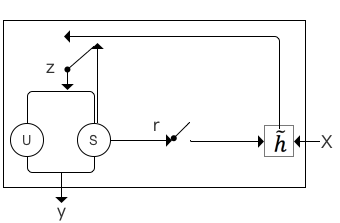
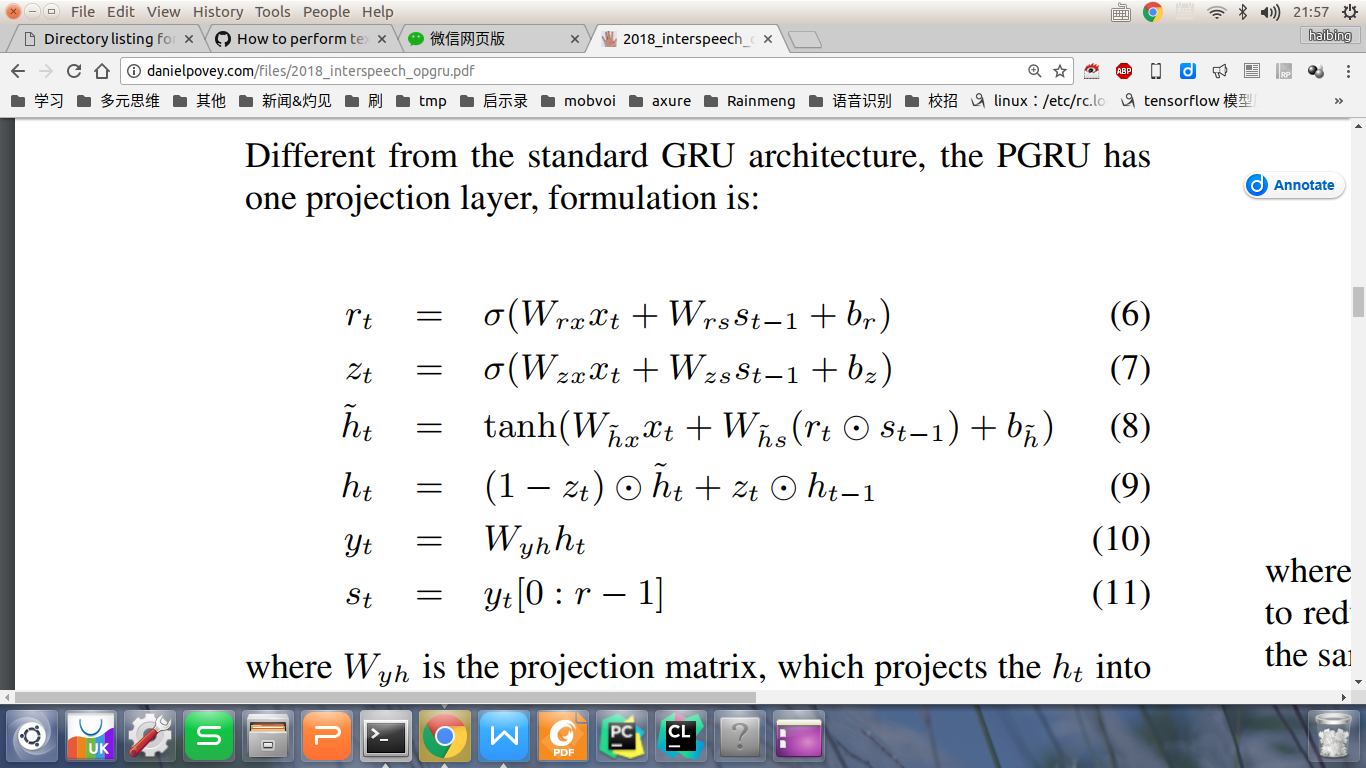


图4 PGRU门阀递归单元

公式如下：



ut = Wuh ht

st = Wsh ht

yt = concat(ut, st)

PGRU对标准GRU的记忆单元输出值做了两次降维映射，得到低维度的u和s，只有s传递给下一时刻的神经元；u和s的拼接结果作为该单元的最终输出。

我们认为，将GRU记忆单元输出值整体传递给下一时刻，不仅存在信息冗余的现象，而且会在很大程度上提高模型复杂度，降低计算速度。在PGRU模型中，由于记忆单元的输入输出维度都没有改变，只限制了递归的复杂度，模型的学习能力基本没有受到影响，但是效率会有很大提升。

**实验**

4.1数据及实验配置

实验所用数据是由一位儿童录制的共计7小时的日常口语语音，包含4242条短语音，采样率为44100Hz。将数据拆分为训练集、验证集和测试集，分别包含4000、100和142条语音。语音经过下采样，采样率被降到16000Hz。借助工具WORLD，在语音上，每5ms提取60维的梅耶尔倒谱系数、25维的频带非周期参数和1维的对数基频。实验中，将这三种特征及它们各自的一阶差分、二阶差分作为神经网络的输出。

如2.2部分所述，对文本进行分析得到471维的上下文相关的文本特征，并将其作为神经网络的输入。经过最大最小规则化，输入特征被限制在[0.01,0.99]之间；输出特征经过均值方差规则化使得其均值为0、方差为1。用于生成对齐信息的模型是一个在训练集上训练的5状态单音素HMM-GMM模型。实验基于Merlin进行的，Merlin是一款开源的语音合成工具，该工具实现了将深度学习模型用于语音合成任务。

4.2结果分析

实验对比了前馈神经网络(DNN)、RNN、LSTM、GRU在中文语音合成上的表现。网络模型配置如下：

（1）DNN：6个全连接的隐含层，每层神经元数为1024；

（2）RNN:5个全连接的隐含层，每层神经元个数为1024；第6个隐含层是一个包含256个神经元的RNN层；

（3）LSTM: 5个全连接的隐含层，每层神经元个数为1024；第6个隐含层是一个包含256个记忆块的LSTM层；

（4）(P)GRU: 5个全连接的隐含层，每层神经元个数为1024；第6个隐含层是一个包含256个门阀递归单元的GRU层；所有网络的隐含层激活函数均采用tanh函数。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型 | MCN  (dB) | BAP  (dB) | F0 RMSE  (Hz) | V/UV  (%) |
| DNN | 5.48 | 0.21 | 10.32 | 5.02 |
| RNN | 5.38 | 0.20 | 10.03 | 4.92 |
| LSTM | 5.35 | 0.20 | 9.87 | 4.89 |
| GRU | 5.36 | 0.19 | 9.80 | 4.90 |
| PGRU | 5.34 | 0.18 | 9.81 | 4.80 |

表1列举了各模型在测试集上的客观评价指标表现（包括：MCN（梅耶尔倒谱扭曲)、BAP（频带非周期性扭曲)、F0 RMSE（基频均方误差

根）和V/UE（清音、浊音错误率)）。

表1. 几种深度学习模型的客观评价

不难看出，递归神经网络的四种模型（包括RNN,LSTM,GRU,PGRU）在各项指标上比普通神经网络要表现得好，尤其在MCN方面.而且LSTM和(P)GRU，由于在递归神经网络中采用了门结构，对模型优化的能力更强，表现得也更好。

相比LSTM，GRU并非在所有客观指标上占绝对优势，但是由于其门结构比LSTM的记忆块要简单，所以训练速度要快很多。另外，PGRU比GRU的训练速度提升了将近10%。

另外，对几种模型合成的语音也进行了主观评测。评测人由98个说汉语的学生组成，每个人对同一句话的５条语音（分别由五个模型合成，顺序随机）从合成的准确性和语音自然度两个维度进行综合评分（评分为百分制，而且同一条语音的四个评分中必须有一个是100）。五种深度模型的平均得分分别是56，59，62，63，63。可以看出，在相似性和自然度方面，递归神经网络的合成表现比普通深度神经网络要好，而且长短时记忆神经网络和门阀递归单元神经网络明显优于递归神经网络。

4.3 dropout策略

Dropout是一种通常被用来防止过拟合的规则化方法。在神经网络训练阶段，实施dropout的层会随机“丢弃”一些神经元，然后在训练数据集上迭代优化模型参数。通过随机丢弃，实际上我们会训练出不同结构的网络（因为是随机丢弃，导致不同次训练，网络节点的连接情况都不一样），最终训练出的模型参数可以认为是这些不同网络结构下相应参数的平均。

在语音合成实验中，我们研究了dropout策略对合成效果的影响。发现，不采用dropout策略，得到的合成语音调很平，严重失真。

**总结**

深度学习技术在语音合成方面的应用得到印证，通过在中文语音合成上的对比实验，表明了对时序信号建模的递归神经网络比前馈神经网络有更优的性能，另外，对递归神经网络进行改进的长短时记忆神经网络和门阀递归单元神经网络也都比普通递归神经网络表现得更好。可以说，深度学习技术在语音合成上有很大的潜力。

为了提升递归神经网络的计算性能，本文对标准GRU做了改进，提出PGRU，降低递归向量的纬度，减小计算量。实验表明，在不影响建模能力的前提下，计算效率有了一些提升。

另外，端到端识别模型也是未来研究的重点。在端到端模型中，直接由文本生成语音的声学特征甚至是语音采样点，不仅减少了前端文本处理和后处理的工作，而且合成效果更理想。

**参考文献：**

[1] K. Tokuda, H. Zen, and A. Black, “An HMM-based speech synthesis system applied to English,” in Proc. IEEE Speech Synthesis Workshop, 2002, CD-ROM Proceeding.

[2] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, “Speech synthesis based on hidden Markov models,” Proceedings of the IEEE, vol. 101, no. 5, pp.1234–1252, 2013.

[3]J. Odell, The use of context in large vocabulary speech recognition, Ph.D. thesis, Cambridge University, 1995.

[4] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2013, pp. 7962–7966.

[5] H. Lu, S. King, and O. Watts, “Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis,” Proc. the 8th ISCA Speech Synthesis Workshop (SSW), pp. 281–285, 2013.

[6] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “The effect of neural networks in statistical parametric speech synthesis,” in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2015, pp. 4455–4459.

[7] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, “TTS synthesis with bidirectional LSTM based recurrent neural networks,” in Proc. Interspeech, 2014, pp. 1964–1968.

[8] H. Zen and H. Sak, “Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis,” in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2015, pp. 4470–4474.

[9] B. X. Wenfu Wang, Shuang Xu, “Gating recurrent mixture density networks for acoustic modeling in statistical parametric speech synthesis,” in Proc. IEEE Int. Conf on Acoustics, Speech, and Signal Processing (ICASSP), 2016.

[10] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997

[11] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio,”Learning phrase representations using rnn encoder-decoder for statistical machine translation,” in Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014), October. to appear.

[12]陈宙斯, and 胡文心. "简化LSTM的语音合成." 计算机工程与应用54.3(2018).

[13] 基于长短型记忆递归神经网络的英文手写识别[D]. 卫晓欣.华南理工大学 2014

[14] 基于极性转移和LSTM递归网络的情感分析[J]. 梁军,柴玉梅,原慧斌,高明磊,昝红英.  中文信息学报. 2015(05)

[15]H. Sak, A. W. Senior, and F. Beaufays, “Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition,” CoRR, vol. abs/1402.1128, pp. 157–166, 2014.

**作者联系方式：**

**联系人：王雨萌**

**E-mail：[1183650351@qq.com](mailto:1183650351@qq.com)**

手机号码：18110026749