基于深度学习的中文语音合成技术研究

**作者名**1 **作者名**1，2 **作者名**2，3

(单位名 所在市 邮编)1

**摘 要** 语音合成技术在近些年得到较快发展，但在真实性和自然度等方面还是存在一些问题。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。由于其拥有强大的建模能力，而且有更好的灵活性、可易控制性，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。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(GRU)是几项典型的深度学习模型，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 LSTM GRU**

**中图法分类号**  （细化到3位数字 ） **文献标识码** A  **DOI** （投稿时不提供DOI号）

**English Title**

NAME Name-name1 NAME Name1,2 NAME Name-name2,3

(Department of xx，xx University， City ZipCode，Country)1

(School of xx，xxUniversity，City ZipCode， Country)2

(Department of xx，xx University，City ZipCode， Country)3

**Abstract** (内容应与中文摘要内容对应，保证语法正确，符合英文表达习惯，字数一般超过200字，少于500字)

**Keywords**  Keyword 1，Keyword 2，…（与中文关键字须对应；一般不用英文缩写；若是词组，一般只大写第一个单词的首字母）

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(Long Short-Term Memory Neural Network)[7][8] [10]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。

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：  
（神经网络的表达式）  
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（W,b)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  
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)的记忆单元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。记忆单元如图所示：其中。。。。  
（记忆单元的示意图）  
在长短时记忆神经网络中，输入、输出的关系如下：  
（LSTM公式）  
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)代替长短时记忆神经网络中的记忆单元，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所示。  
（GRU示意图）  
输入X经过

* 1. 二级标题
     1. 三级标题

对不同类型文章的要求：

研究型论文：一般应包括引言（主要论述该领域面临的问题，解决问题的思路，将要提出的方法，产生的价值和意义等）、相关工作、详细的方法或技术描述、实验验证（对比验证或实例分析）、结束语等内容。

综述型论文：包括引言、问题或挑战、研究现状分析、未来研究方向、结论等内容。

正文内容的规范：

数学中使用的符号、函数名、变量或表示变化的量用斜体；使用的量符合法定计量单位标准；矢量为黑体，标量为白体；

定义、定理、公式、推论、算法、步骤等须从数字1开始按顺序编号。

图表须注明中、英文图题和表题，且在文中应明确提及或引用（例：如图1所示，如表1所列）。其中图的编号和图题应置于图下方的居中位置，表的编号和表题置于表上方的居中位置。

图形务必清晰，图宽一般在8厘米以内（如需通栏排版，宽度应在13~14厘米左右），采用Visio（2007）软件绘制：线条大小为0.5pt，线条颜色为黑色，选用04号箭头（大小为最小）；汉字为仿宋\_GB2312，西文字选用Times New Roman（图中数学变量用斜体），文字大小为6.5pt。

表格尽量采用三线表，必要时可添加辅助线。

表1 示例表

Table 1 Example table

|  |  |  |  |
| --- | --- | --- | --- |
| 算法 | 检测率/% | 准确率/% | 帧率/fps |
| PBAS | 88.40 | 85.60 | 27.4 |
| Our Method | 92.90 | 92.70 | 23.2 |

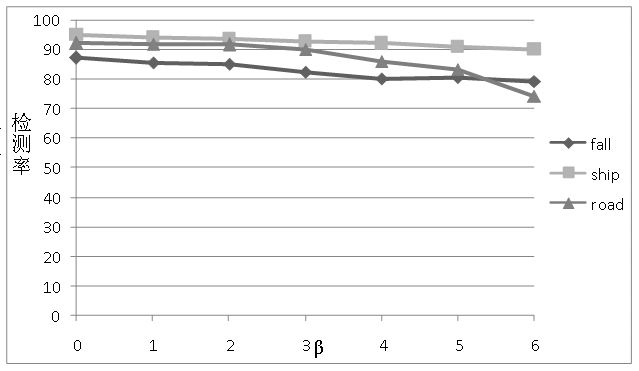


图1 示例图1

Fig.1 Example diagram 1



(a)图像1 (b)图像2



(c)图像3 (d)图像4

图2 示例图2

Fig.2 Example diagram 2

图形的具体制作要求请参加网站首页的图形规范。

**结束语**（该部分除了对本文所做工作进行总结外，还应着重指出存在的问题以及未来的研究方向）

**致 谢** （不要求每篇文章都有，若需要特别指出对文章有贡献的学者，可在此部分中说明）

**参考文献**

[1]著者.题目［J］.刊名，出版年，卷号(期号)：起止页码.

例：LI Q, WANG G, LIU J, et al. Robust Scale-Invariant Feature Matching for Remote Sensing Image Registration [J]. IEEE Geosciences & Remote Sensing Letters, 2009, 6(2):287-291.

[2]著者.书名［M］.译者，译.出版地: 出版者, 出版年：起止页码.

例：刘国钧，陈绍业，王凤翥. 图书馆目录[M]. 北京：高等教育出版社，1957：15-18.

[3]著者.析出文献题名 [C]//会议论文集.出版地：出版社，出版年：起止页码.

例：KALAL Z，MATAS J. P-N Learning: Bootstrapping Binary Classifiers by Structural Constraints[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2010:49-56.

[4] 著者.题名[D]. 所在城市：学位授予单位, 出版年.

例：王煜. 基于错误传播上下文分析的软件错误定位方法研究[D].哈尔滨：哈尔滨工业大学，2013.

[5]著者.题名：报告号[R]. 出版地 (城市名): 出版者, 出版年.

例：冯西桥. 核反应堆压力管道和压力容器的LBB分析[R]. 北京：清华大学核能技术设计研究院，1997.

[6]著者. 标准名称：准编号［S］.出版地: 出版者,出版年.

例：GB/T 16159—1996，汉语拼音正词法基本规则[S].北京：中国标准出版社，1996.

[7]著者.题名［N］.报纸名，出版日期（版次）（出版日期按YY-MM-DD格式）.

例：谢希德.创造学习的新思路[N].人民日报，1998-12-25（10）.

[8]著者.题名［文献类型标志/电子文献载体标志］.(更新日期) [引用日期].获取和访问路径（如http://www.www.arocmag.com）.

例：W3C.Web Service Choreography Interface(WSCI)(Version1.0)[EB/OL].www.w3.org/TR/wsci.

[9]专利所有者.专利题名：专利国别，专利号［P］.公告日期.获取和访问路径.

例：姜锡洲. 一种温热外敷药制备方案：中国， 881056073[P].1989-07-26.

注：

（1）参考文献中若有中文文献，请在相应英文翻译后附上中文文献；

（2）参考文献中个人著者采用姓前名后的形式。姓的每个字母均需大写，三人以上者，录入前三人姓名后加“等”，英文姓名则加“et al”。

.例：

[10] LI D, XU Z M, LI S, et al. A survey on information diffusion in online social networks [J]. Chinese Journal of Computers, 2014, 37(1):189-206 .(in Chinese)

李栋, 徐志明, 李生, 等. 在线社会网络中信息扩散[J]. 计算机学报, 2014, 37(1): 189-206.

请在文末添加作者的联系电话和邮箱！