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。

1.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]和门阀递归单元神经网络(Gated Recurrent Unit Neural Network)[9]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。

2.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

训练语音

特征提取

模型

模型训练

向量化

文本分析

训练文本

测试文本

合成语音

问题集

**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

2.3深度学习模型

2.3.1 LSTM

2.3.2 GRU

[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.