文章编号：1003-0077（2017）00-0000-00

基于神经网络的藏语语音合成

都格草 才让卓玛 南措吉
（1. 青海师范大学 青海 西宁 810008；2. 藏文智能信息处理与机器翻译重点实验室 青海 西宁 810008）

摘要：人机交互中最自然、最理想的交流方式为语音，其中主要涉及到了语音合成，即文本转换为语音的技术。随着神经网络理论的不断深入，基于神经网络的语音合成技术越来越引起人们的关注。文章通过分析藏文字结构与拼读规则，融合Sequence to Sequence模型和注意机制，研究了基于神经网络的藏语语音合成技术。实验数据表明，该文方法对藏语语音合成具有良好的性能表现。

关键词：藏语语音合成；神经网络；Sequence to Sequence；注意机制

中图分类号：TP391 **文献标识码：**A

Neural Network based Tibetan Speech Synthesis

DOU Gecao CAI Rangzhuoma NAN Cuoji

(1. Qinghai Normal University Qinghai Xining, 810008；2. Key Laboratory of Tibetan information processing,Ministry of Education Qinghai Xining 810008)

Abstract : Speech is the most natural and ideal way of communication in human-machine interaction, which mainly involves speech synthesis, that is, text to speech technology.With the continuous deepening of neural network theory, the speech synthesis technology based on neural network has attracted more and more attention. By analyzing the structure and spelling rules of Tibetan characters, combining Sequence to Sequence model and attention mechanism, this paper studies the technology of Tibetan speech synthesis based on neural network. The experimental results show that this method has good performance in the speech synthesis of Tibetan.

Key words： Tibetan speech synthesis; Neural network; Sequence to Sequence; Attention

1 引言

语音合成（Speech synthesis）是人机交互的核心技术之一，也是中文信息处理领域的一项前沿技术。语音合成的目标是将文字信息自动转换为清晰、流畅的语音，它的研究对自动控制、智能机器人和人机语音通讯系统等的研制具有重要的理论意义和实用价值。随着计算机技术和通信技术的发展，语音合成技术越来越引起社会的关注。

语音合成技术的发展按时间顺序大致经历了机械式、电子式和计算机语音合成[1]三个阶段。尽管计算机语音合成技术由于其侧重点的不同，分类方法也有所差异，但当前主流和获得较认同的分类是将语音合成方法按照设计思想分为规则驱动(Rule-based)方法和数据驱动(Data-based)方法。前者的主要思想是根据人类发音的物理过程制定一系列规则来模拟语音合成过程，例如共振峰合成和发音规则合成；数据驱动方法则是从语音库的数据用统计方法实现合成，例如波形拼接（Concatenative Synthesis）合成[2]、基于隐马尔可夫模型（Hidden Markov Model,HMM）的统计参数语音合成[3]及深度神经网络（Deep Neural Networks,DNN）的语音合成[4-8]。相对而言，基于数据驱动方法更

收稿日期：2018-07-00；**定稿日期**：2018-00-00

基金项目：国家自然科学基金资助项目（61262051,61163018），国家社科基金项目（16BYY167，15BYY132，14BYY132，13BYY141），教育部“春计划”（Z2016077），青海省基础研究项（2017-ZJ-767）

为成功，也更受研究人员的青睐。近年来，随着神经网络方法在机器翻译、文本分类、问答系统及信息抽取等领域的广泛应用，基于神经网络的语音处理技术在语音合成中也取得了显著成绩[9-12]，并从而成为当前研究语音识别及合成的主流技术。

藏语语音合成技术是中文信息处理的重要任务之一，目前藏语语音合成系统的实现主要采用波形拼接技术[13]或基于隐马尔可夫模型（HMM）统计参数语音合成技术[14,15]。考虑到波形拼接技术对存储容量要求高且系统构建周期长，而统计参数语音合成技术的合成语音效果不稳定且韵律表现不佳。本文通过分析藏文字结构与拼读规则，融合 Sequence to Sequence（简称 Seq2Seq）模型和注意机制（Attention），研究基于神经网络的藏语语音合成技术。本文主要分为以下几个部分：第一部分主要介绍语音合成技术的发展及藏语语音合成研究现状；第二部分介绍语音合成模型相关技术；第三部分给出了基于神经网络的藏语语音合成方法；第四部分进行了实验及数据分析；第五部分是结语。

2 语音合成模型

随着神经网络理论的不断深入，基于神经网络的各种模型被广泛地应用于语音合成中。例如，文献[4]中使用为 DNN 模型进行语音合成、文献[5-7]使用循环神经网络（Recurrent Neural Network, RNN）模型进行语音合成、而文献[8]使用长短记忆网络（Long Short Term Memory,LSTM）模型进行语音合成等。考虑到 Seq2Seq 模型突破了传统神经网络的固定大小输入问题框架，并在英语－法语翻译、英语-德语翻译[16,17]等的应用中有着不俗的表现。本文采用基于 Seq2Seq 模型和注意机制融合研究藏语语音合成方法。Seq2Seq 模型主要通过深度神经网络模型（常用的是 LSTM，长短记忆网络）将一个作为输入的序列映射为一个作为输出的序列。

2.1 Seq2seq

Seq2Seq 模型是 2014 年由 Google Brain 团队和 Yoshua Bengio 两个团队各自独立提出[17,18]。它是一个编码器到解码器结构的网络。根据输入序列 X 来生成输出序列 Y，当初该模型解决了机器翻译的相关问题，之后人们开始应用于智能对话与问答、自动编码与分类器训练、句法分析、文本摘要、语音合成与识别等领域被广泛应用。一般的模型而言，输入的特征通常是一个固定大小的矩阵，限制了输入的长度必须是一致。Seq2Seq 模型解决长度不固定问题。通常把 RNN 的输入称为“上下文”（Context），通过 Encoder 来产生此上下文的表示 C。C 是一个输入序列 $X=\{x_1, x_2, \ldots x_T\}$ 的向量序列。网络结构如下图 1 所示。

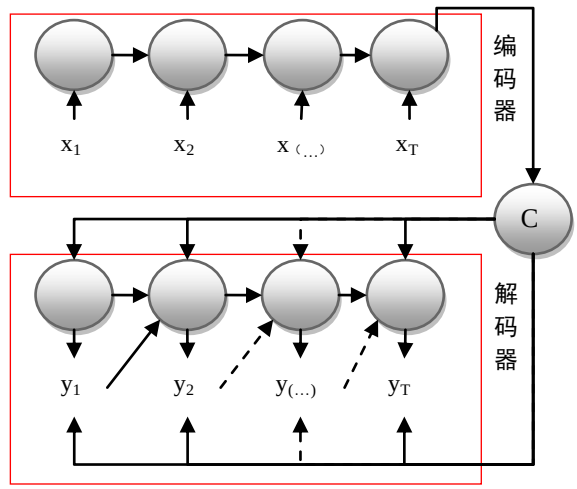

图1 Seq2seq模型

2.2 注意机制

Seq2Seq 模型中每次将 Encoder 的输出作为上下文向量 C 输入到 Decoder 中，这样每次生成音频时所使用的上下文向量是相同。也就是说 C 中必须包含原始序列中的所有信息，它的长度就限制模型性能的瓶颈。比如语音合成中当合成的句子较长时，一个 C 可能存不下那么多信息就会造成合成精度下降，因此需要引入到注意机制模型。该模型主要包括编码器、存储器以及解码器三个部分组成。上下文向量 C 的计算公式为：

$$C=\sum_{t=1}^{T} a^{(t)}h^{(t)} \qquad (1)$$

其中 $h^{(t)}$ 为编码器输出的特征向量，$a^{(t)}$ 为权重。$a^{(t)}$ 权重向量在 Decoder 每次进行预测时都不一样，计算方法如下：

$$a_{ij}=\frac{\exp(e_{ij})}{\sum_{k=1}^{T}\exp(e_{ik})} \qquad (2)$$

其中的 e_{ij} 代表 Encoder 的第 j 个输入与 Decoder 的第 i 个输出的匹配程度。

3 基于神经网络的藏语语音合成

藏语有三种主要方言，卫藏方言和康方言有声调，安多方言没有声调[19],但都遵从相同的文法与拼接规则。藏语文字是一种拼音文字，主要由 30 个辅音字母和 4 个元音组成[20]。辅音字母中有 10 个后加字，4 个下加字。后加字中 5 个字母为前加字、2 个字母为重后加字、3 个字母为上加字。藏文字以基字为中心呈横向与纵向结构。其基字的横向有前加字、后加字、重后加字拼写，而基字的纵向还有上加字、下加字和元音拼写。藏文字的基本结构如图 2 所示。藏文字的拼读按照其文字结构逐项叠加进行[21]读音，即将藏文字按其基本结构从左往右，从上到下逐项拼读便可得到相应的音节，文字中有元音的字或词时元音只出现一次。藏文字拼读顺序如图 3 所示。

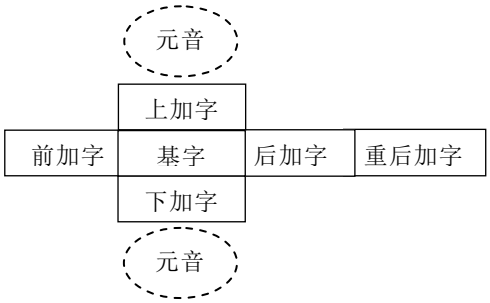

图 2 藏文字基本结构

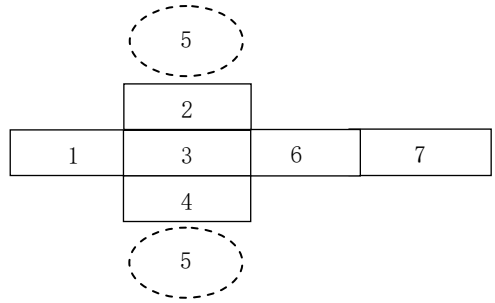

图 3 藏文字拼读顺序示意图

3.1 字符嵌入

模型的输入为一系列文本字向量，输出为声谱图，然后使用 Griffin_lim 算法[22]生成对应音频。以下将对模型进行简单介绍。

文本分析模块中纯文本数据转换为向量，向量作为神经网络的输入。因此，本文按照藏文字拼读顺序使用字典下标作为字典中每一个字对应的序列，然后每一条文本通过高斯分布函数随机生成其对应的向量。例子：“ཤ་སྲ་ལེགས་བཤད།” 句子中首先提取所有音素并给出对应序列，其次每个序列使用 8 维的二进制表示，最后生产一个矩阵 M，行大小为音素的个数 15，列大小为词向量的维度 8。比如矩阵 M 的第二行的序列号为 1,即“ཤ”对应随机生成词向量。藏文字符嵌入过程如图 4 所示。

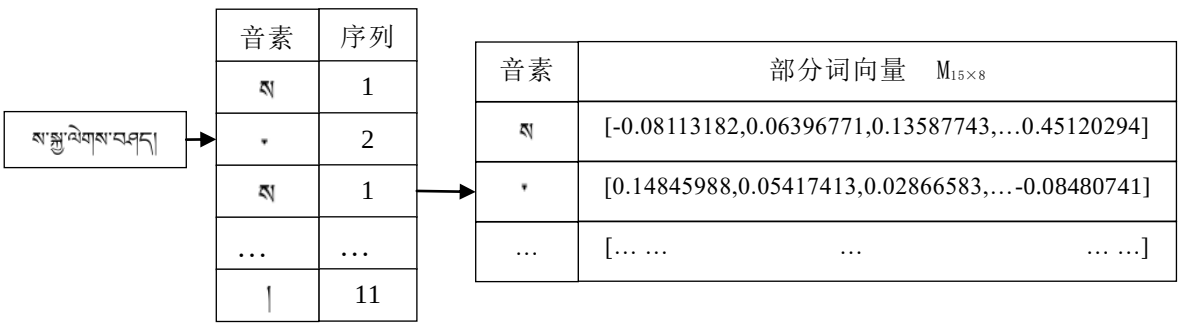

图4 藏文字符嵌入过程

3.2 提取语音特征

声音是一个一维的时域信号，直观上很难看出频域的变化规律。为了分析音频的频域，本文提取于梅尔频率倒谱系数（Mel-frequency Cepstral Coefficients,MFCC），该参数维度低且符号听觉系统作为语音特征参数。时频使用短时傅里叶（Short-time Fourier transform,STFT）。STFT 把一段长信号分帧、加窗，再对每一帧做傅里叶变换，最后把每一帧的结果沿另一个维度堆叠得到二维信号的声谱图。声谱图需要合适减缩，通过梅尔标度滤波器组（Mel-scale filter banks），变换为梅尔频谱。在梅尔频谱上做倒谱分析就得到了梅尔倒谱。

3.3 模型构建

第二节已介绍 Seq2Seq 模型与 Attention 机制。

该模型主要有编码器与解码器两大模块组成。

（1） 编码器模块

编码器模块主要包含预网（Pre-net）结构与CBHG（全称Conv Bank Highway Gru_rnn）结构。模块训练过程如图5所示。

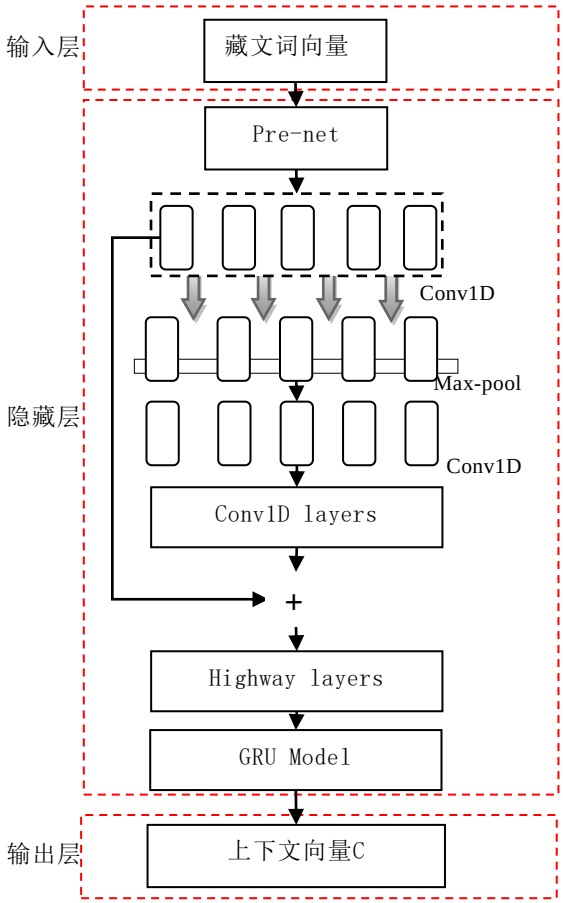

图5 编码器模块训练过程

Pre-net的主要功能是对输入进行一系列非线性的变换。它是一个3层的网络结构，有两个隐藏层，层与层之间均为全连接。第一层的神经元个数与输入藏文词向量个数一致，第二层的神经元个数为第一层的一半。两个隐藏层采用的激活函数均为ReLu，并保持0.5的dropout来提高泛化能力。CBHG主要用于提高模型的泛化能力。Pre-net的输出经过一个卷积层。它有K个大小不同的1维滤波器（Filter），其中Filter的大小为1, 2, 3…K(16)。大小不同的卷积核提取长度不同的上下文信息。然后，将经过不同大小的K个卷积核的输出堆积在一起。下一层为最大池化层（Max-pool），步长（Stride）为1。经过池化之后，会再经过两层一维的卷积层。第一个卷积层的Filter大小为3，Stride为1，采用的激活函数为ReLu；第二个卷积层的Filter大小为3，Stride为1，没有采用激活函数。然后把卷积层输出与字符的序列相加输入到Highway layers中，同时放入到两个一层的全连接网络，两个网络的激活函数分别采用ReLu和Sigmoid函数。假定输入为Input，ReLu的输出为H，Sigmoid的输出为T，Highway layer的输出为：$output = H \times T + Input \times (1 - T)$。它的输出提供GRU模型，得到上下文向量C。

（2）解码器模块

编码器模块主要分为Pre-net、Attention-RNN、Decoder-RNN三部分。Pre-net的结构与Encoder中的Pre-net相同，主要是对输入做一些非线性变换。Attention-RNN的结构为一层包含256个GRU的RNN单元，它将Pre-net的输出和Attention的输出作为输入，经过GRU单元后输出到Decoder-RNN中。Decode-RNN为两层residual GRU，它的输出为输入与经过GRU单元输出之和。每层同样包含了128个RNN单元。

在 Decoder-RNN输出之后并没有直接将输出转化为音频文件，所以添加了后处理的网络。后处理的网络在一个线性频率范围内预测幅度谱（Spectral magnitude），并且后处理网络能看到整个解码的序列，从左至右的运行。后处理网络通过反向传播来修正每一帧的错误。最后使用 Griffin-Lim 算法生成音频。

4 实验及数据分析

训练神经网络模型时语料是必不可少的主要要素，本实验使用的语料是藏族第 1 部哲理格言诗集《萨迦格言》，诗集强调知识、智慧的作用，它既是藏族学者必读著作，也在群众口头广泛流传。诗集中有 456 韵文，每个韵文有四个句子构成，每个句子 7 音节，总之 1824 个句子和 12768 个音节。这些语料作为本文的训练数据。采样率为 16KHz，采样精度为 16bits。

为了评估模型的性能和模型拟合的好坏，本文从客观和主观进行分析。客观上模型训练的误差函数来度量拟合程度。误差函数极小化，意味着拟合程度最好，对应的模型参数即为最优参数。图 6 中 loss 是 MFCC 和短时傅里叶变换的求和误差图。loss1 是 MFCC 特征训练过程的误差图。lr 是学习率，初始学习率设为 0.001。图中可见模型的

误差逐步极小化，说明模型的拟合程度最好。

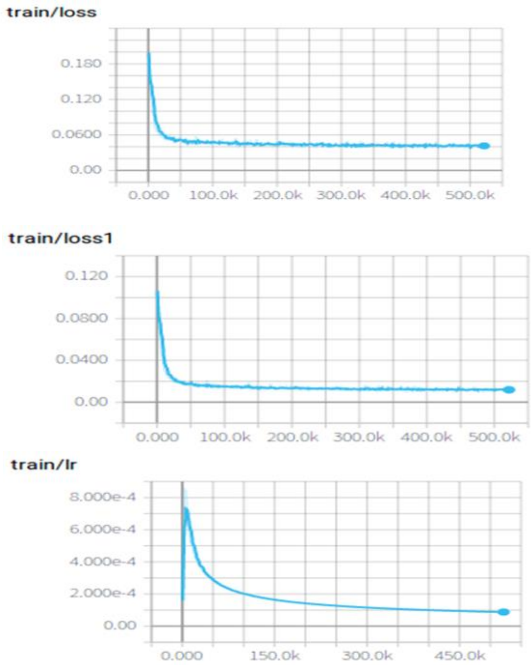

图 6 训练藏语语音合成模型的误差

主观上由 7 位测听员通过对合成语音与原始语音（设定为 5 分）进行对比打分给出了 MOS 分。如图 7 所示，合成语音的清晰度和自然度分别为 3.46 和 3.9 分，基本达到预期的目标。

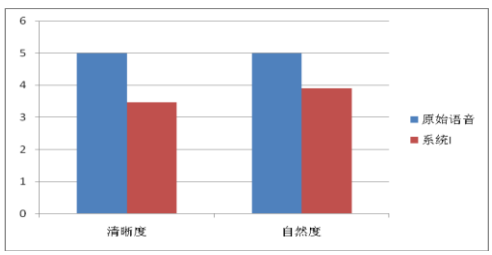

图 7 系统的 MOS 评分

评分中测听员都提出来的意见是合成语音完全能听清楚所读的内容，但语音不够清晰，有一点儿机器味儿。因此，自然度明显高于清晰度。下一步的工作是语料规模增大，并语料文体多样化，语音的清晰度进一步提高，生成任意文本对应的语音。

5 结语

语音合成技术的主要任务将文本映射为音频信号，为了提高语音合成的自然度和清晰度。本文通过分析藏文字结构与拼读规则，融合 Seq2Seq 模型和注意机制研究基于神经网络的藏语语音合成技术。为了评估模型的性能和模型拟合的好坏，本文从模型训练的误差和 MOS 评估进行客观和主观分析，合成效果均达到预期目标。

参考文献

[1] 张斌, 全昌勤, 任福继. 语音合成方法和发展综述[J]. 小型微型计算机系统, 2016.
[2] Hunt A J,Black A W.Unit selection in a speech synthesis system using a large speech database[C] Acoustics, Speech and Signal Processing,ICASSP-96,1996:373-376.
[3] Zen H,Tokuda K,Black A W.Statistical parametric speech synthesis[J].Speech Communication,2009,5 (11):1039–1064.
[4] Yao Qian,Yuchen Fan,Frank K.Song.On the training aspects of deep neural network(DNN) for parametric TTS synthesis[C].In-ternation Conference on Aconstic, Speech and Signal Processing,2014:3829-3833.
[5] P Wang,Y Qian,FK Soong,L He,H Zhao.Word embedding for recurrent neural network based TTS synthesis[J]. IEEE International Conference on Acoustics,2015:4879-4883.
[6] Lipton Z C,Berkowitz J,Elkan C.A critical review of recurrent neural networks for sequence learning[J].arXiv preprint arXiv:1506.00019,2015.
[7] Zen H,Sak H.Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis[C].IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP),2015:4470-4474.
[8] H Ming,Y Lu,Z Zhang,M Dong.A light-weight method of building an LSTM-RNN-based bilingual tts system[J]. International Conference on Asian Language Processing, 2017:201-205.
[9] VR Reddy,KS Rao.Prosody modeling for syllable based text-to-speech synthesis using feedforward neural networks[J].Neurocomputing,2016,171:1323-1334.
[10] V Rajendran,GB Kumar.Prosody prediction for Tamil text-to-speech synthesizer using sentiment analysis [J]. Asian Journal of Pharmaceutical&Clinical Research, 2017,10(13):6.
[11] T Delić,S Suzić,M Sečujski,D Pekar.Rapid Development of New TTS Voices by Neural Network Adaptation[J]. International Symposium Infoteh-jahorina,2018.
[12] Y Wang,RJ Skerry-Ryan,D Stanton.Towards End-to-End Speech Synthesis[J].interspeech,2017: 4006-4010
[13] 才让卓玛, 才智杰. 基于语料库的藏语语音合成单元选择算法[J]. 中文信报, 2017, (31)05, 59-63.
[14] 周雁, 赵栋材. 基于 HMM 模型的藏语语音合成研究[J]. 计算机应用与软件, 2015, 171-174.
[15] 高璐, 于洪志, 郑文思. 基于 HMM 的藏语拉萨话语音合成技术研究[J]. 西北民族大学学报, 2011.
[16] RJ Weiss,J Chorowski, N Jaitly,Y Wu,Z Chen. Sequece to Sequnce Models Can Directly Translate Foreign Speech[J].Interspeech,2017:2625-2629.
[17] ho et al.Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation [J].Computer Science,2014.
[18] Iiya Sutskever,Oriol Vinyals,Quoc V.Le. Sequence to sequence learning with neural networks[J].Computation and Language,2014.

[19] 羊忠旦增.藏语三大方言比较研究[D].中央民族大学,2013.
[20] 才让卓玛,李永明,才智杰.藏语语音合成单元选择[J].软件学报,2015,%(6):1409-1420.
[21] 江荻,龙从军.藏文字符研究[M]社会科学院文献出版社,2010.8.
[22] D. W. Griffin and J. S. Lim,Signal estimation from modified short-time Fourier transform[J].IEEE Trans.ASSP,vol.32,no.2,pp.236-243,Apr.1984.

都格草（1992—），女，藏族，硕士生，主要研究领域为人机语音交互，藏文信息处理。
E-mail：179418384@qq.com

才让卓玛（1970-），女，藏族，博士，教授，硕士生导师，主要研究领域为人机语音交互、藏文信息处理。
E-mail：cr-zhuoma@163.com

南措吉（1992—），女，藏族，硕士生，主要研究领为人机语音交互，藏文信息处理。
E-mail：975709705@qq.com