**Computer Science and Application** 计算机科学与应用, 2021, 11(3), 505-514
Published Online March 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.113051

Hans汉斯

# 基于BERT-CRF模型的缅甸语韵律单元边界预测

**李培英，杨　鉴***

云南大学信息学院，云南　昆明
Email: ieLiPeiYing@163.com, *jianyang@ynu.edu.cn

收稿日期：2021年2月12日；录用日期：2021年3月8日；发布日期：2021年3月12日

## 摘　要

**近年来，缅甸语语音合成引起了众多学者的关注，然而该技术的性能离推广应用还有一段距离。本文以提升缅甸语语音合成自然度作为目标，研究缅甸语韵律特征，探索通过缅甸语文本自动预测韵律单元边界的方法。本文提出并实现了一种基于BERT预训练模型和条件随机场(CRF)模型相结合的缅甸语韵律词和韵律短语边界预测方法。实验结果表明，采用BERT-CRF模型，韵律词和韵律短语的预测效果均优于CRF、BiLSTM、BiLSTM-CRF以及BERT模型。为了验证该方法的可用性，本文还将本文所提出的方法应用于语音合成前端文本分析与处理中。语音合成实验结果表明，本文所提方法能有效提高缅甸语语音合成的自然度。**

## 关键词

**缅甸语，韵律单元预测，BERT预训练模型，条件随机场模型，语音合成**

# Prosodic Unit Boundary Prediction of Myanmar Based on BERT-CRF Model

**Peiying Li, Jian Yang***

School of Information Science and Engineering, Yunnan University, Kunming Yunnan
Email: ieLiPeiYing@163.com, *jianyang@ynu.edu.cn

Received: Feb. 12th, 2021; accepted: Mar. 8th, 2021; published: Mar. 12th, 2021

## Abstract

**In recent years, Myanmar speech synthesis has attracted the attention of many scholars, but the**

*通讯作者。

文章引用: 李培英, 杨鉴. 基于 BERT-CRF 模型的缅甸语韵律单元边界预测[J]. 计算机科学与应用, 2021, 11(3): 505-514.
DOI: 10.12677/csa.2021.113051

**performance of this technology is still a long way from popularization and application. In order to improve the naturalness of Myanmar speech synthesis, this paper studies the prosodic features of Myanmar language and explores the method of automatically predicting the boundary of prosodic units through texts. In this paper, a boundary prediction method of prosodic words and phrases in Myanmar language based on BERT pre-training model and Conditional Random Field (CRF) model is proposed and implemented. The experimental results show that the prediction effect of prosodic words and phrases using BERT-CRF model is better than that of CRF, BiLSTM, BiLSTM-CRF and BERT models. In order to verify the availability of this method, the method proposed in this paper is also applied to the front-end text analysis and processing of speech synthesis. The experimental results of speech synthesis show that the proposed method can effectively improve the naturalness of Myanmar speech synthesis.**

## Keywords

**Myanmar, Prosodic Unit Prediction, BERT Model, CRF Model, Speech Synthesis**



## 1. 引言

缅甸语是缅甸联邦的官方用语。据统计，2007 年，大约有 3.3 千万人使用的第一语言是缅甸语，1 千万人使用的第二语言是缅甸语。缅甸语的电子化语言资料较为稀少，相比于英语、汉语等通用语言，缅甸语语音合成的研究速度相对缓慢，尤其在自然度方面的研究还有较大的进步空间。目前，国内外在缅甸语韵律特征方面的研究仍位于初级阶段，尤其是针对缅甸语语音合成系统的韵律研究。

语音合成系统一般可分为文本分析、韵律处理和语音合成三个模块。韵律处理模块处于文本分析与语音合成模块之间，是提高合成语音自然度的重要环节。韵律处理模块是指通过文本分析得到语言学信息，通过语言学模型设置成文本的韵律特征。最后把韵律特征传入至语音合成模块，合成声音进行输出[1]。在韵律模块中，如果计算机不能有效的识别语句的韵律结构，就会使得合成语音缺乏真实语音的抑扬顿挫。所以，准确划分连续语流中的韵律单元边界对加强语音合成自然度尤为重要。

在语音合成的声学模型训练阶段，需要构建韵律标注语料库，划分韵律单元边界。传统的标注工作一般由专业的标注人员手动标注，较为耗时，且人工标注有主观性，不同人乃至同一人在不同情况下的标注结果都有差异性。为了解决此问题，结合文本和语音实现韵律单元划分，即为韵律单元边界自动标注。而在合成阶段，应在不依赖于语音的情况下划分出韵律单元边界，只有这样，才能在合成语音中体现应有的韵律特征。在不依赖于语音的前提下，在文本中正确划分韵律单元，这即为韵律单元边界预测。

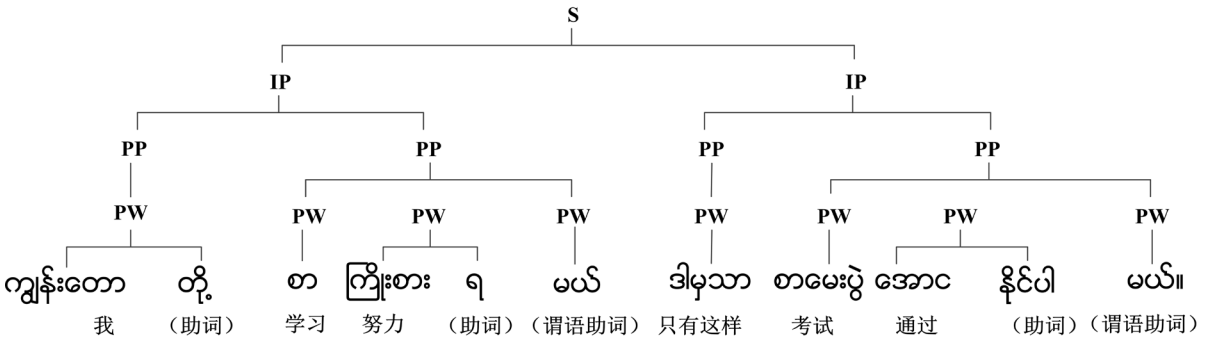
最初，基于规则的方法实现韵律结构预测[2]，这种方法要求规则制定者要具备丰富的语言学知识，需花费大量人工时间成本；后来出现了基于统计的方法，如：决策树[3]、隐马尔科夫[4]、条件随机场[5]等等方法，这些方法对模型的输入特征具有选择性，在文本语法分析计算时只利用到浅层信息，深层信息会被忽略。近几年，深度学习的发展越来越成熟，预训练语言模型普遍被用于众多领域。采用预训练模型实现任务的策略方案有：基于特征和模型微调[6]。本文在 BERT 预训练模型的基础上采用模型微调的方法实现缅甸语韵律单元边界预测，以缅甸字为建模单位，基于 BERT 模型导入自己的标注数据进一步训练得到动态字向量的特征，再利用 CRF 模型对预训练模型的输出结果进行微调。最后，将韵律单元

边界预测结果应用到 HMM-DNN 缅甸语语音合成中，对韵律单元边界的预测结果进行验证评估。

## 2. 缅甸语韵律结构

缅甸文是一种拼音文字，一个字就是一个音节。音节是缅甸语的语音层级和结构单位。缅甸语音节主要有两种构成方式：元音型和辅音元音结合型。在缅甸语中，最常见的是元音和辅音结合，一个辅音和一个元音是最基本的形式，复合型音节类型有四种：单辅音和双元音组合，双辅音和单元音或者双元音组合，三个辅音和单元音或者双元音结合，以及四个辅音和单元音组合。此外，元音可自成音节，其中单元音或双元音均可构成音节[7] [8]。缅甸文字在书写过程中没有用于标记音节边界或词边界的分隔符，在某些情况下会出现空格符号表示分隔边界，但空格符号出现情况有两种：一种是短语或换气边界，另一种是排版问题，空格位置随意添加，为避免换行引起短语分割，使得破句或造成误解，在不影响句义的位置添加空格[9]。针对这个特点，不能确切地将空格位置视为韵律短语的分隔位置。

在韵律音系学中，一般将韵律结构层级从低到高分别为：音节、词典词、韵律词、韵律短语、语调短语。根据缅甸语特点，本文着重研究缅甸语的韵律词(Prosodic Word, PW)、韵律短语(Prosodic Phrase, PP)两个层级，如图 1 所示。其中，S 表示句子，IP 表示语调短语，该例句中共有 2 个语调短语，4 个韵律短语，8 个韵律词。韵律词大多数是单个词典词，或由两个或两个以上的词典词组成，韵律词内部音节间无节奏边界，语言结构较为紧密，韵律词之后有较短的无声停顿；韵律短语由一个或若干个韵律词组成，内部的韵律词之间可能会出现节奏边界，语言结构相对较为松散，韵律短语之间有较长的停顿。




翻译：我们要努力学习，只有这样才能通过考试。

**Figure 1.** The prosody structure of Myanmar
**图 1.** 缅甸语的韵律结构

## 3. 韵律单元边界预测方法

### 3.1. BERT-CRF 模型

#### 3.1.1. BERT 模型

BERT 模型(Bidirectional Encoder Representation from Transformers)，是 Google 公司于 2018 年提出的一种基于 Transformer 模型的语言模型。我们可将 BERT 模型看作为一个通用的语言模型，用于实现不同的下游任务。BERT 模型结构[10]如图 2 所示，所使用 Transformer 的编码器部分[11]，结构如图 3 所示。

其中，图中用 Tran 表示 Transformer 模型，$E_1, E_2, \cdots, E_m$ 表示以字为单位的文本输入，$T_1, T_2, \cdots, T_m$ 表示模型的输出向量。

如图 3 所示，Transformer 模型的编码器是由多层单元组合构成，每层单元包含两个子层，每层单元内部的子层间是通过残差连接，以确保信息能完整传输：

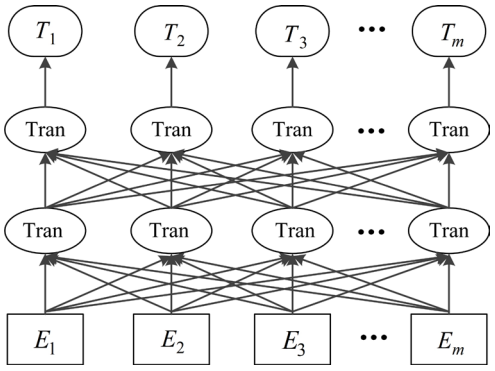



**Figure 2.** The model structure of BERT
**图 2.** BERT 模型结构

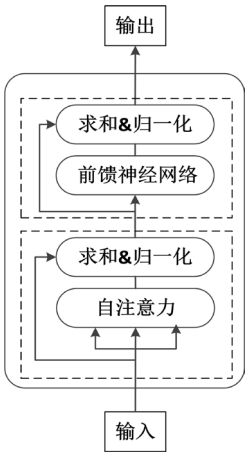



**Figure 3.** Encoder structure of transformer model
**图 3.** Transformer 模型的编码器结构

第一个子层：将以字为单位的序列输入至多注意力机制层，在编码过程中可查看该字的前后信息。在自注意力机制中，每个输入的字向量中含有三个 $d$ 维的分向量，将分别乘以查询权重、键权重，以及值权重得到查询向量 $Q$、键向量 $K$、值向量 $V$；然后为获得每个字的注意力得分，查询向量 $Q$ 和键向量 $K$ 进行内积处理；最后注意力层的输出和输入进行相加，对结果进行归一化处理得到 $A$，如式(2-1)。

$$A=\text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{2-1}$$

其中，$\sqrt{d_k}$ 为惩罚因子。

由于 BERT 模型使用的是多头注意力机制，是由多个自注意力机制组合而成的，所以，多头注意力层的输出计算方式如式(2-2)、(2-3)。

$$head_i=Attention\left(QW_i^Q,KW_i^K,VW_i^V\right) \tag{2-2}$$

$$Multihead=concat\left(head_1,head_2,\cdots,head_n\right)W^O \tag{2-3}$$

其中，$W_i^Q$、$W_i^K$、$W_i^V$ 分别表示第 $i$ 个注意力的 $W^Q$、$W^K$、$W^V$ 权重矩阵，$W^O$ 是随机初始化矩阵。

第二个子层：将第一个子层的输出传入一个全连接的前馈神经网络中，然后进行上述同样的残差处理和归一化处理。

为了训练深层的双向 Transformer 模型，本文采用了掩蔽的语言模型(Mask Language Model，MLM)作为训练任务，只需要预测被掩蔽的字，而无需重建整个句子。在每个输入句子中随机选择 15%的缅甸字作为要被遮盖的字，然后在这些被遮盖的字中，随机选择 80%的字直接用“[MASK]”替代，10%的字随机替换成其他字，10%的字保留不变[10]。

### 3.1.2. CRF 模型

条件随机场模型(Conditional Random Fields，CRF)是基于最大熵模型(Maximum Entropy Model，ME)和隐马尔科夫模型(Hidden Markov Model，HMM)构建的一种判别式概率模型，该模型使用全局优化的思想，可以更好地对文本序列实现标签预测。假设给定一组输入序列 $A=(a_1,a_2,\cdots,a_n)$，对应的预测序列为 $Y=(y_1,y_2,\cdots,y_n)$，那该预测序列的分值函数[12]如式(2-4)。

$$s(A,y)=\sum_{i=0}^{n}M_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \tag{2-4}$$

其中，$M$ 为标签之间的转移分数，$P_{i,y_i}$ 为每个字对应标签的分数。

在模型训练过程中，采用极大似然优化方法进行优化，得到预测序列的似然概率如式(2-5)。

$$\log\left(P\left(y|A\right)\right)=s(A,y)-\log\left(\sum_{\bar{y}\in Y_A}^{n}\mathrm{e}^{s(A,\bar{y})}\right) \tag{2-5}$$

其中，$Y_A$ 是指一个输入序列 $A$ 对应的所有预测的标签序列。

在模型预测阶段，采用 Viterbi 算法，得到最大得分的输出序列如式(2-6)。

$$y^{*}=\arg\max_{\bar{y}\in Y_A}s(A,\bar{y}) \tag{2-6}$$

通过公式(2-6)可以看出，$s(A,y)$ 分值越高，预测越准确。

### 3.1.3. BERT-CRF 模型

本文将韵律单元边界预测任务看作是字级别的序列标注任务，提出了基于 BERT-CRF 模型的缅甸语韵律单元边界预测模型，该模型充分将 BERT 模型和 CRF 模型的优点结合，利用 BERT 预训练模型获取上下文信息序列，自动学习序列的状态特征，将状态特征连接至一个全连接层输出一个状态分数，然后又直接传入至 CRF 网络；通过 CRF 网络对预测结果进一步加入约束条件来保证预测结果的合理性。该模型可分为三部分：输入层、BERT 层和预测层。其模型结构如图 4 所示。

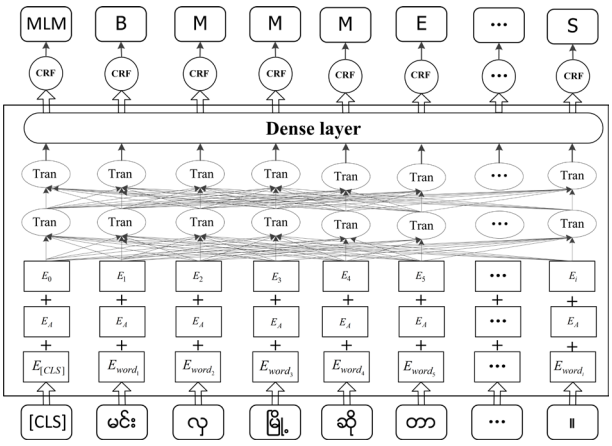



**Figure 4.** Myanmar prosodic unit prediction based on BERT-CRF model
**图 4.** 基于 BERT-CRF 模型的缅甸语韵律单元预测

BERT 模型的输入是一组将文本中的每个缅甸字经过查询字向量表转换成的向量序列。输入由 3 部分组成，由下至上分别为：字向量(token embedding)、分割向量(segment embedding)、位置向量(position embedding)。其中，字向量是 BERT 模型经过对大规模样本的无监督训练得到的词向量或字向量，本文用的是字向量，输入序列的首个字符由一个特殊的[CLS]标签填补。分割向量是用于划分文本的句子或段落，就段落而言，$E_A$ 和 $E_B$ 分别代表左句子和右句子；就句子而言，只存在 $E_A$。此次用于实验的缅甸语文本语料是单个句子，所以用 $E_A$ 表示，分割向量为 0。位置向量是表示该缅甸字在当前句中的位置信息。

BERT 模型的输出是每个缅甸字的编码向量，大小为 768，此向量中含有当前位置的语义信息。在 BERT 模型后添加一层全连接层，便可把编码向量序列转换为预测标签集合，从而为下一步实现韵律单元边界预测做准备。

在文本序列标注任务中，预测阶段往往采用 softmax 函数计算出各个标签的概率，将最大概率对应的标签作为最终的预测结果，显而易见预测过程中没有考虑标签之间的关系，而 CRF 网络可以自动学习到标签的前后信息，对预测结果加入限制条件来保证预测结果的正确性。在训练过程中，CRF 层会主动学习到所有的限制条件，从而可使得错误的预测序列极大降低。

## 3.2. 输入数据预处理和模型训练方法

### 3.2.1. 输入数据预处理方法

本文将韵律单元边界预测任务视为是序列标注任务，并且是在 BERT 预训练模型的基础上利用自己已标注数据进行训练，因此需要对文本数据进行预处理。缅甸语文本语料库是以单句为单位，需要将文本数据按照格式预处理成两列：一列缅甸字(即缅甸语音节)，一列标注标签，利用“BMES”作为标注标签。一个缅甸语句子实例：

မင်းလှမြို့ဆိုတာရှုခင်းသိပ်သာယာတဲ့မြို့ကလေးဖြစ်ငါတယ်။，其韵律标注过程如图 5 所示。过程①表示将句子分为 4 个韵律短语，过程②表示将一个韵律短语划分为相应的韵律词，过程③表示将每一个韵律词划分为对应的音节进行标注。

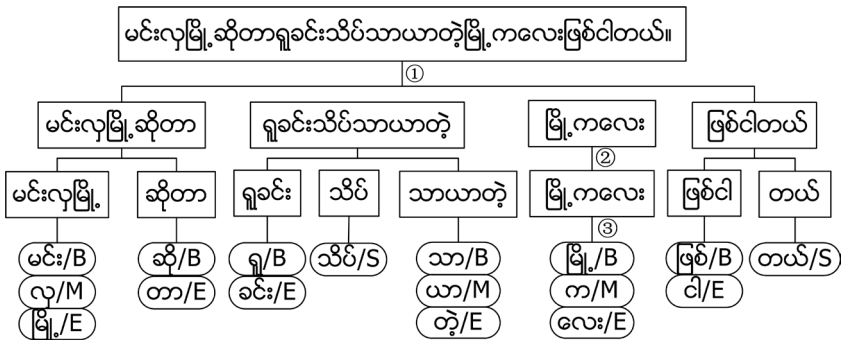



**Figure 5.** Myanmar prosodic text labeling
**图 5.** 缅甸语韵律文本标注

### 3.2.2. 模型训练方法

由于韵律词与韵律短语之间是一种递进关系，韵律词是韵律短语的构词基础，所以为了避免韵律词的预测效果影响韵律短语的预测效果，本文选择分别独立训练和预测韵律词和韵律短语。对于每个任务，首先对每个缅甸字得到其输入的特征向量，然后经过 BERT 预训练模型和全连接层得到每个字的输出结果，将输出结果直接通过 CRF 线性层，进行式(2-5)处理得到预测序列的概率向量，最后根据式(2-6)得到

最大的分值对应的预测标签。

## 4. 实验结果与分析

### 4.1. 数据库

用于本文实验的数据库，共包括 4890 个缅甸语“文本－语音”(<文本，语音>)对。文本以句子为单位，内容涵盖了新闻、科技、文体等多方面内容。语音发音人为缅甸国家级广播电台播音员。音频数据采样频率为 16 kHz，量化比特数为 16 bit。请缅甸语的专业人员通过阅读文本和听对应的音频，对上述的文本数据进行人工标注韵律词和韵律短语。本文选择 80%的数据作为训练集，10%的数据作为验证集，10%的数据作为测试集。本文的语音合成实验选择 2000 句缅甸语发音语料用于实验。

### 4.2. 实验过程

为验证本文提出的基于 BERT-CRF 模型的韵律单元边界预测方法，分别设计并实现了下列 5 种实验。

实验 1：采用传统方法 CRF 建模，使用第 3.2.1 节所述的方法进行数据预处理，选择 U-gram 作为模板特征，只考虑缅甸语音节这一特征，不考虑其他特征，选取“2 + 2”大小的窗口模板，主要对前两个音节与后两个音节的特征进行训练，由 CRF++0.58 工具进行训练。

实验 2：利用 BiLSTM 建模，在 BiLSTM 层之后加入 softmax 层，预测出最后的标签。使用 TensorFlow-1.9.0 训练模型，将训练过程中自动生成的大小为 256 维的字向量作为输入特征。batch 大小为 20，epoch 为 180，隐藏层节点数为 300，学习率为 0.001，使用 Adam 优化器更新参数。

实验 3：同样利用 TensorFlow-1.9.0 构建 BiLSTM-CRF 模型，在 BiLSTM 层之后通过 CRF 层约束标签之间的转移情况，从而预测输出标签。训练模型时，选择训练过程中自动生成的大小为 256 维的字向量作为输入特征，实验参数设置与实验 2 一样。

实验 4：在 BERT 预训练模型的基础上增加一层全连接层构建预测模型，在此基础上使用 softmax 层，预测出最后的标签。实验所用的语言模型是谷歌公司的多语言 BERT-Base 版本(下载地址：https://github.com/google-research/bert)，模型层数为 12，隐藏层单元数为 768，自注意力层个数为 12。利用 TensorFlow-1.14.0 训练预测模型，batch 大小为 32，学习率为 $2 \times 10^{-5}$，输入序列长度为 128，优化器为 Adam。

实验 5：本文提出的基于 BERT-CRF 模型是在预训练模型的基础上增加一层全连接层，之后利用 CRF 层对标签的转移情况进行限制，得到预测标签。训练模型时，实验参数设置与实验 4 一样。

### 4.3. 韵律单元预测实验的结果与分析

本文采用客观评价方法衡量预测模型的性能，客观评价方法是指利用客观测量的手段，将通过模型所标注的结果与原有的人工标注结果进行比对，利用准确率 $P$、召回率 $R$ 和 $F$ 值作为模型性能评估标准，其定义如式(3-1)、(3-2)、(3-3)。

$$P = \frac{TN}{TN + FT} \times 100\% \tag{3-1}$$

$$R = \frac{TN}{TN + TF} \times 100\% \tag{3-2}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{3-3}$$

其中，$TN$ 为预测出并符合真实韵律单元的数量，$FT$ 为被预测为韵律单元但与事实不符的数量，$TF$ 是标

记为真实韵律单元但未被预测出的数量。

将本文提出的 BERT-CRF 模型与 CRF、BiLSTM、BiLSTM-CRF 以及 BERT 模型进行比较，实验结果如表 1 所示。

**Table 1.** Experimental results of different models predicting the boundaries of prosodic units
**表 1.** 不同模型预测韵律单元边界的实验结果

| 模型 | 韵律词 | | | 韵律短语 | | |
|---|---|---|---|---|---|---|
| | *P* (%) | *R* (%) | *F* (%) | *P* (%) | *R* (%) | *F* (%) |
| CRF | 65.95 | 65.17 | 65.56 | 64.07 | 63.05 | 63.55 |
| BiLSTM | 72.21 | 70.20 | 71.19 | 67.90 | 62.27 | 64.96 |
| BiLSTM-CRF | 76.91 | 74.39 | 75.63 | 69.30 | 67.67 | 68.52 |
| BERT | 77.71 | 76.12 | 76.78 | 75.85 | 74.51 | 75.07 |
| BERT-CRF | 78.51 | 76.84 | 77.57 | 78.04 | 77.19 | 77.60 |

从表 1 可以看出以下几点结论：

(1) 在数据集相同的情况下，本文提出的基于 BERT-CRF 模型的缅甸语韵律单元边界预测效果相对更好，韵律词和韵律短语的准确率、召回率、*F* 值都明显提高了；

(2) 对比 BiLSTM 与 BiLSTM-CRF 模型、BERT 与 BERT-CRF 模型的实验结果，发现 BiLSTM-CRF 和 BERT-CRF 模型的预测效果更好，这证实了在基线模型 BiLSTM、BERT 的基础上加 CRF 层能够限制标签转移，降低错误序列，提高预测效果；

(3) 对比 BiLSTM-CRF 与 BERT-CRF 模型，BERT-CRF 模型优于 BiLSTM-CRF 模型。这是由于 BiLSTM-CRF 模型的字向量是由训练过程产生的，训练样本有限，而 BERT-CRF 模型的字向量是经过大规模样本的无监督训练获取的，另外 BERT-CRF 模型的字向量维数较高，融合到的上下文信息更丰富；

(4) 对比韵律词与韵律短语的预测结果，除 BERT-CRF 模型外的其他四种模型，韵律词的预测结果显著优于韵律短语的预测结果，这是因为韵律词单元的长度较小，无需过多考虑到的上下词和语义等深层信息，而韵律短语的内部结构具有较为复杂，如句法结构、依存关系等等，这都会影响到韵律短语的预测；对于 BERT-CRF 模型，在训练中可以获取句子内部的长时信息，从而使得在预测阶段能考虑到过多的上下文信息，而且通过 CRF 层能够提高预测的准确性，所以韵律短语的预测效果提升得较为明显；

(5) BERT 模型虽然在一定程度上能捕捉到不同字间的不同特征，但由于该模型无法分辨不同上下文中的同一个字，而上下文关系对韵律单元预测尤为重要，所以韵律单元预测效果的提升有一定的限制；

(6) 基于文本实现韵律单元预测的方法适用于语音合成的合成阶段，可用于构建大量的文本语料标注库，在一定程度上可减少人工标注的时间成本和经济成本，同时也可保证标注的一致性。

### 4.4. 语音合成实验

将基于 BERT-CRF 的韵律单元边界预测结果应用到 HMM-DNN 语音合成的前端文本分析和韵律处理中[13]，本文在 Ubuntu16.04 下基于 HTS-2.3 平台构建了缅甸语语音合成系统，其框图如图 6 所示。文本分析是语音合成的前端工作，根据词典和规则解析出后端所需的信息与文本；韵律处理处于文本分析与语音合成之间，根据韵律预测模型或韵律规则对文本进行韵律标注；语音合成是后端，结合前端的音段信息与韵律标注信息训练声学模型，基于训练好的模型实现参数预测，最后利用参数合成器合成出声音。

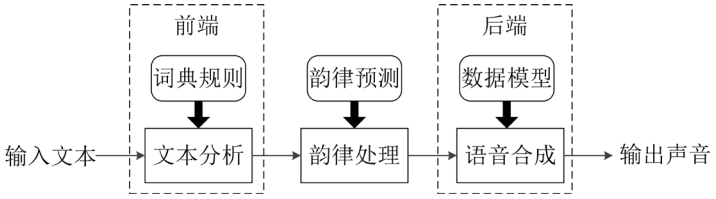



**Figure 6.** Myanmar speech synthesis system
**图 6.** 缅甸语语音合成系统

对合成语音效果采用客观评测和主观评测方法进行评估。本文采用 MCD (Mel Cepstral Distortion) [14] 评测作为语音的客观评测方法，该方法是指合成声音的 MFCC 特征与标准原始声音的 MFCC 特征之间的差距，数值越小，表示合成质量越好。本文随机挑选 500 句，将未考虑韵律信息的合成语音、加入韵律信息的合成语音分别与标准语音进行对比，实验结果如表 2 所示。

**Table 2.** Voice MCD evaluation results
**表 2.** 语音的 MCD 评测结果

| 语音 | 平均分 |
|---|---|
| 未考虑韵律信息的合成语音 | 18.13 |
| 加入韵律信息的合成语音 | 15.99 |

语音的主观评测是人听到的主观感受。本文邀请 15 位听力正常且熟悉缅甸语的听者，依照平均意见评分(MOS) [15]的标准，以双盲测试的方法对语音进行打分。听者对合成语句的自然度和连贯性进行主观评测，本文对听者的主观感受进行打分统计，实验结果如表 3 所示。

**Table 3.** Voice MOS evaluation results
**表 3.** 语音的 MOS 评测结果

| 语音 | 平均分 |
|---|---|
| 标准语音 | 4.52 |
| 未考虑韵律信息的合成语音 | 3.45 |
| 加入韵律信息的合成语音 | 3.96 |

从客观评测结果可以看出，对待输入文本进行韵律单元预测能显著缩小合成语音与标准语音之间频谱参数的差距，提高合成声音的质量，同时通过主观评测又进一步验证了加入韵律信息的合成语音更加有助于提升听者对语音质量的满意度。这表明在缅甸语语音合成中，本文所提的韵律单元预测方法可以明显提升缅甸语语音合成的自然度。

## 5. 结束语

本文以提高缅甸语语音合成自然度为目的，探讨分析了缅甸语的韵律特征，从而实现对韵律单元边界的自动预测。对于缅甸语的语言特征和韵律特征，本文提出了一种在 BERT 预训练模型的基础上微调实现韵律单元预测的模型，利用 BERT 预训练获取输入的缅甸字的特征向量，再采用 CRF 模型预测韵律单元边界。并将预测结果应用于缅甸语语音合成中来验证韵律单元预测方法的可用性。实验结果表明，与其他模型比较，基于 BERT-CRF 模型能进一步提高韵律词和韵律短语的预测结果。在语音合成的前端文本分析和处理中，对待输入文本进行韵律单元边界预测能有效的提高缅甸语语音合成自然度。在后续

工作中，我们将研究如何通过多任务学习实现韵律词和韵律短语同时预测，以便缅甸语语音合成自然度得到更大的提升。

## 基金项目

本文获得国家自然科学基金(61961043)。

## 参考文献

[1] Gu, W., Hirose, K. and Fujisaki, H. (2003) A Method for Automatic Extraction of F0 Contour Generation Process Model Parameters for Mandarin. *IEEE Workshop on Automatic Speech Recognition and Understanding*, **2003**, 682-687. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1318522

[2] 赵晟, 陶建华, 蔡莲红. 基于规则学习的韵律结构预测[J]. 中文信息学报, 2002, 16(5): 30-37.

[3] 王仁华, 胡郁, 李威, 凌震华. 基于决策树的汉语大语料库合成系统[C]//中国中文信息学会. 全国人机语音通讯学术会议, 深圳, 2001: 307-311.

[4] 熊艳娇. 基于 HMM 语音识别的韵律标记[J]. 中国新通信, 2015, 17(12): 98-99.

[5] Sun, J.W.,Yang, J., Zhang, J.P. and Yan, Y.H. (2009) Chinese Prosody Structure Prediction Based on Conditional Random Fields. 2009 5*th International Conference on Natural Computation*, Tianjian, 14-16 August 2009, 602-606. https://doi.org/10.1109/ICNC.2009.44

[6] 张鹏远, 卢春晖, 王睿敏. 基于预训练语言表示模型的汉语韵律结构预测[J]. 天津大学学报(自然科学与工程技术版), 2020, 53(3): 265-271.

[7] 钟智翔, 尹湘玲. 基础缅甸语[M]. 广州: 世界图书出版广东有限公司, 2012.

[8] Chaw, S.H. and Aye, T. (2017) Myanmar Speech Synthesis System by Using Phoneme Concatenation Method. 2017 *International Conference on Signal Processing and Communication*, Coimbatore, 28-29 July 2017, 399-404.

[9] 汪大年. 缅甸语汉语比较研究[M]. 北京: 北京大学出版社, 2012.

[10] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., *et al*. (2017) Attention Is All You Need. 31*st Annual Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[12] 田梓函, 李欣. 基于 BERT-CRF 模型的中文事件检测方法研究[J]. 计算机工程与应用, 2020: 1002-8331. http://kns.cnki.net/kcms/detail/11.2127.TP.20201027.1328.012.html

[13] Hlaing, A.M., Pa, W.P. and Thu, Y.K. (2018) DNN Based Myanmar Speech Synthesis. *The* 6*th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, Gurugram, 29-31 August 2018, 142-146. https://doi.org/10.21437/SLTU.2018-30

[14] Kubichek, R. (1993) Mel-Cepstral Distance Measure for Objective Speech Quality Assessment. *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, Victoria, 19-21 May 1993, 125-128. https://doi.org/10.1109/PACRIM.1993.407206

[15] Streijl, R.C., Winkler, S. and Hands, D.S. (2016) Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives. *Multimedia Systems*, **22**, 213-227. https://doi.org/10.1007/s00530-014-0446-1