基于深度神经网络的
韵律结构自动标注与预测

（申请清华大学工程硕士专业学位论文）

培 养 单 位：计算机科学与技术系

工 程 领 域：计算机技术

申 请 人：杜 耀

指 导 教 师：吴 志 勇 副研究员

二〇一九年五月

基于深度神经网络的韵律结构的自动标注与预测

杜　耀

Mandarin Prosodic Structure Auto-labeling and Prediction using Deep Neural Network

Thesis Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the professional degree of

Master of Engineering

by

Du Yao

(Computer Technology)

Thesis Supervisor：Associate Professor Wu Zhiyong

May, 2019

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容。

本人保证遵守上述规定。

（保密的论文在解密后遵守此规定）

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘 要

在以数据驱动的参数化语音合成中，构建高质量的合成语料库对语音合成的自然度至关重要。在构建汉语语料库的过程中，韵律层级结构标注工作往往采用人工标注。对成千上万句子进行人工标注，费时费力，而且因为不同人员对句子的理解判断存在一定的主观差异性，导致不同人员在少部分句子上的标注结果存在不一致性。因此，自动标注韵律层级结构对快速构建语音合成语料库有着重要的意义。

韵律结构预测任务是参数化语音合成系统较为关键的步骤，处在分词与词性标注模块之后。韵律结构与语音的韵律节奏、停顿密切相关，其性能关系到合成语音的自然度。

本文对语音合成系统，有关韵律层级结构的两个任务进行了深入研究，主要贡献如下：

1、提出联合深度神经网络（DNN）、带门控循环单元的双向循环神经网络（BGRU-RNN）、条件随机场（CRF）层的混合型网络结构。该方法能同时利用文本及声学信息，并且借助输出层采用条件随机在解码时考虑整句标注的上下文信息，有助于对汉语韵律层级结构进行自动标注。在特征准备方面，探索了词尾音素嵌入、词尾声学等特征应用于韵律结构标注的效果。

2、提出基于深度自注意力网络的韵律结构标注模型。该方法使用的自注意力机制，能学习到句中任意距离的词与词的特征之间的联系。相比韵律结构任务中常用的条件随机场、循环神经网络等基准模型，本文的方法能捕捉全句任意范围的依赖关系，因而取得了较基准模型更好的预测效果。

3、提出基于 Transformer 双向编码器表示模型（Bidirectional Encoder Representations from Transformers，BERT）预训练词向量的韵律结构预测模型。BERT 词向量基于自注意力机制在大语料库中预训练，富含上下文信息，在众多自然语言处理任务中，该词向量的应用效果显著。本文尝试将该词向量引入到韵律结构预测任务，并同时对比分析了各种词向量对韵律结构的影响。实验表明基于 BERT 词向量的模型能显著提升较高层级的韵律结构预测的准确性。

关键词：韵律结构自动标注；韵律结构预测；语音合成；序列标注；自注意力机制

Abstract

In a data-driven parametric speech synthesis, building a high-quality speech corpus is essential to the naturalness of synthesized speech. In the process of constructing a Chinese speech corpus, prosodic structure labeling is often accomplished by professional annotators. It's time-consuming and laborious to label tens of thousands of sentences manually. Besides, there may exist inconsistencies caused by annotators's different interpretations of sentences. Therefore, it's of significant importance to implement a prosodic structure auto-labeling system to speed up the preparation of the corpus for speech synthesis. In addition, prosodic structure prediction is also key to the text-to-speech system. The performance of the module has large impacts on the naturalness of the synthesized speech.

In this work, the tasks of automatic prosodic structure labeling and prediction are studied in depth. The main contributions are listed as follows:

1. A hybrid network structure of combined deep neural network (DNN), bidirectional recurrent neural network with gated recurrent unit (BGRU-RNN) and conditional random fields (CRF) layer is proposed. This method can exploit both text and acoustic information. The CRF layer makes the model consider the context labels of words in the whole sentence. It helps to auto-label the prosodic structures. In the preparation of features, some word-final features have been explored to improve the labeling performance.

2. A prosodic structure prediction model based on self-attention network is proposed. The self-attention mechanism can learn the dependencies of two aribitary words, regardless of their distance. Compared with the commonly used models like CRF and RNN on the task of prosodic structure prediction, the proposed method can capture the long-span dependency across the entire sentence. This method achieves better performance than the CRF and RNN baseline models.

3. A prosodic structure prediction model using word vector pre-trained by Bidirectional Encoder Representations from Transformers (BERT) is proposed. The BERT word vector is pre-trained in the large corpus based on the self-attention mechanism. It is rich in context information and has been applied to a wide range of

natural language processing tasks for its exordinary performance. In this work, the BERT is applied to the task of prosodic structure prediction. Different types of word vectors have been compared on the performance of prosodic structure prediction. The proposed BERT word vector based model can significantly improve the accuracy of higher level prosody structure prediction.

Key words: prosodic structure auto-labeling; prosodic structure prediction; speech synthesis; sequence labeling; self-attention mechanism

目 录

第 1 章 引言 1
1.1 研究背景与意义 1
1.1.1 韵律结构概念介绍 1
1.1.2 韵律结构自动标注 2
1.1.3 韵律结构预测 3
1.2 本文主要的研究内容和贡献 4
1.2.1 研究内容和各章简介 4
1.2.2 本文主要贡献 5
第 2 章 相关研究综述 1
2.1 韵律结构自动标注 1
2.1.1 基于决策树的韵律结构自动标注 2
2.1.2 基于隐马尔科夫模型的韵律结构自动标注 3
2.1.3 基于条件随机场模型的韵律结构自动标注 4
2.1.4 基于 RNN 的韵律结构自动标注 5
2.1.5 韵律结构自动标注的研究现状总结 6
2.2 韵律结构预测 7
2.2.1 基于分类回归树的韵律结构预测 8
2.2.2 基于隐马尔科夫模型的韵律结构预测 8
2.2.3 基于条件随机场模型的韵律结构预测 9
2.2.4 基于 RNN 的韵律结构预测 9
2.2.5 韵律结构预测的研究现状总结 9
第 3 章 基于 DNN-BGRU-CRF 的韵律结构自动标注 11
3.1 本章引论 11
3.2 问题定义 11
3.3 基于 DNN-BGRU-CRF 的韵律结构标注模型 11
3.3.1 深度神经网络 DNN 11
3.3.2 带门控循环单元的循环神经网络 BGRU-RNN 12
3.3.3 条件随机场 CRF 13
3.3.4 DNN-BGRU-CRF 模型结构 15

3.4 实验结果与分析 17
3.4.1 数据说明 17
3.4.2 特征提取 18
3.4.3 实验设置及对比实验 18
3.4.4 实验评估标准 19
3.4.5 实验结果及其分析 20
3.5 本章小结 22
第 4 章 基于深度自注意力网络的韵律结构预测 23
4.1 本章引论 23
4.2 问题定义 23
4.3 基于深度自注意力网络的韵律结构预测模型 23
4.3.1 自注意力子层 24
4.3.2 非线性子层 26
4.3.3 残差连接 30
4.3.4 位置编码 31
4.3.5 标签平滑 31
4.3.6 深度自注意力神经网络模型结构 32
4.4 实验结果与分析 33
4.4.1 数据说明 33
4.4.2 特征提取 33
4.4.3 实验设置及对比实验 26
4.4.4 实验评估标准 34
4.4.5 实验结果及其分析 34
4.5 本章小结 37
第 5 章 BERT 词向量及其在韵律结构预测中的应用 38
5.1 本章引论 38
5.2 问题定义 38
5.3 BERT 向量在韵律结构预测中的应用 39
5.3.1 TRANSFORMER 结构 39
5.3.2 BERT 模型结构 40
5.3.3 韵律结构预测模型 41
5.4 实验结果与分析 42

5.4.1 数据说明 42
5.4.2 特征提取 42
5.4.3 实验设置及对比实验 44
5.4.4 实验评估标准 45
5.4.5 实验结果及其分析 45
5.5 本章小结 47
第 6 章 总结与展望 49
6.1 研究工作总结 49
6.2 未来研究展望 50
参考文献 53
致 谢 58
声 明 59
个人简历、在学期间发表的学术论文与研究成果 60

插图索引

图 1.1 汉语韵律层级结构示例 2
图 1.2 语料库准备阶段的流程图 3
图 1.3 中文 TTS 合成流程图 3
图 2.1 韵律结构标注流程图 1
图 2.2 基于决策树的韵律结构自动标注系统 2
图 2.3 基于隐马尔科夫模型的韵律结构自动标注系统[9] 4
图 2.4 韵律结构预测任务流程图 7
图 3.1 DNN 深度神经网络示例 12
图 3.2 RNN 循环神经网络示例[31] 13
图 3.3 GRU 结构图[31] 13
图 3.4 CRF 链式条件随机场模型[33] 14
图 3.5 BLSTM-CRF 模型结构图[34] 15
图 3.6 基于 DNN-BGRU-CRF 的韵律结构自动标注模型 16
图 3.7 词的声学特征集的提取流程 18
图 4.1 放缩点积注意力的计算流程 24
图 4.2 多头自注意计算流程 25
图 4.3 句尾词特征经过自注意力机制后输出的计算流程图 26
图 4.4 FFN 网络结构图 27
图 4.5 GRU-RNN 结构图 29
图 4.6 BLSTM-RNN 结构图 29
图 4.7 BGRU-RNN 结构图 30
图 4.8 残差连接结构图 30
图 4.9 深度自注意力网络模型结构图 33
图 5.1 Transformer 结构图 39
图 5.2 BERT 模型结构图 40
图 5.3 韵律结构预测模型结构图 42
图 5.4 可训练词嵌入层结构图 43
图 5.5 Skip-Gram 算法模型结构图 44
图 6.1 层级 LSTM-RNN 结构图[64] 51

表格索引

表 3.1 基于 CRF 自动标注方法的混淆矩阵 20
表 3.2 基于 DNN-BGRU-CRF 自动标注方法的混淆矩阵 20
表 3.3 与其他标注方法的对比结果 21
表 4.1 不同非线性子层的实验结果 35
表 4.2 不同堆叠块数目的实验结果 35
表 4.3 CRF 模型预测结果的混淆矩阵 35
表 4.4 深度自注意力网络模型预测结果的混淆矩阵 36
表 4.5 对比实验的结果 37
表 5.1 使用 Random 词向量的韵律结构预测模型 45
表 5.2 使用 Skip-Gram 词向量的韵律结构预测模型 46
表 5.3 使用 BERT 词向量的韵律结构预测模型 46
表 5.4 对比实验结果 47

主要符号对照表

CRF	条件随机场 (Conditional Random Fields)
RNN	循环神经网络 (Recurrent Neural Network)
DNN	深度神经网络 (Deep Neural Network)
TTS	文语转换 (Text-To-Speech)
CART	分类和回归树 (Classification And Regression Tree)
GRU	门控循环单元 (Gated Recurrent Unit)
LSTM	长短时记忆单元 (Long Short-Term Memory)
BGRU	双向带门控循环单元 (Bidirectional Gated Recurrent Unit)
BLSTM	双向长短时记忆单元 (Bidirectional Long Short-Term Memory)
HMM	隐马尔科夫模型 (Hidden Markov Model)
CD-HMM	上下文相关隐马尔科夫模型 (Context Dependent HMM)
NB	非韵律结构边界 (Not a prosodic Boudnary)
PW	韵律词 (Prosodic Word)
PPH	韵律短语 (Prosodic Phrase)
IPH	语调短语 (Intonational Prosodic Phrase)
MLM	遮蔽语言模型 (Masked Language Model)
BERT	Transformer 双向编码器表示 (Bidirectional Encoder Representations from Transformers)

第1章　引言

1.1 研究背景与意义

语音合成系统（Text-to-speech，TTS）是将文本转换语音的系统。目前，语音合成在诸多领域有着广泛的应用，包括车载导航、智能音箱、游戏角色语音合成等。语音合成系统主要有以下几个大类：拼接合成、参数化合成、端到端合成。参数化语音合成与端到端语音合成均依赖大数据来训练得到模型，而参数化语音合成需要带详细标注的数据训练数据，而端到端合成系统则不需要详细标注数据。本文主要涉及的是参数化语音合成系统中的韵律结构自动标注与预测，即参数化语音合成系统中韵律结构相关的两大任务。当前，构建参数化语音合成系统主要流程有：语料库的准备阶段、语音合成阶段（包括各种模型的训练与预测）。其中韵律结构标注是处于语料库准备阶段的任务，而韵律结构预测是处于语音合成阶段的任务。接下来分别介绍韵律结构概念以及韵律结构自动标注、韵律结构预测的研究背景与意义。

1.1.1 韵律结构概念介绍

韵律信息包括语音的节奏、停顿、强调等。韵律结构是韵律节奏的表征，是一种层级的结构。一个典型的汉语韵律结构可划分为：韵律词、韵律短语、语调短语[1]。其结构如图 1.1 所示。较高层级的韵律结构单元由较低层级的韵律结构单元组合而成。下边分别介绍各个韵律层级的特点：

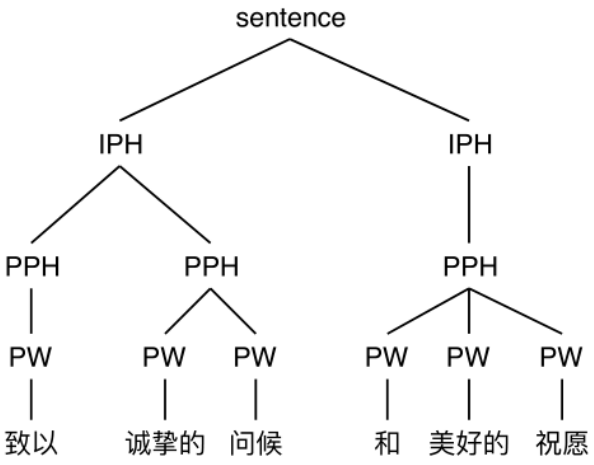
1、韵律词（Prosodic Word，PW），其所包含的音节是紧密地连在一起念出来的。比如，“美好的”、“小雨伞”。韵律词可以是由一个或者多个语法词构成。

2、韵律短语（Prosodic Phrase，PPH），由一个或者多个韵律词构成，通常在韵律短语边界会有基频重置（pitch reset）现象，有可以感知的停顿现象。所谓基频重置是指基频在韵律结构边界处出现的基频跳变现象，

3、语调短语（Intonational Prosodic Phrase，IPH），由一个或者多个韵律短语构成，可以看做是一个具有独立意义的子句，并且往往伴随着明显的、且较韵律短语更大的停顿。比如，例子中的“致以诚挚的问候”即为一个语调短语。

一个或者多个韵律词构成韵律短语，一个或者多个韵律短语构成语调短语，从而形成一种树状的层级结构。韵律结构描述的是语音中一种超音段的特征，其对语

音信号中的轻重、长短、缓急的周期性进行层级建模。这个建模对于语音合成系统至关重要，因为合成系统的停顿的插入位置也直接依赖于韵律结构边界的预测结果，其对合成语音的自然度有着重要意义。



致以 <PPH> 诚挚的 <PW> 问候 <IPH> 和 <PW> 美好的 <PW> 祝愿 <IPH>

图 1.1 汉语韵律层级结构示例

1.1.2 韵律结构自动标注

参数化合成系统依赖大量带详细标注的数据来训练得到模型，而这些带标注数据的语料库的准备是必不可少的阶段。语料库准备阶段的流程图如图 1.2 所示。构建语料库的任务通常包含音段标注和韵律标注。其中音段标注是对音素序列进行时间上的对齐，包括音素的起始、结束时间。现有的自动强制对齐工具已经能进行自动音段标注。韵律标注是对韵律信息标注，对于汉语主要是韵律层级结构的标注。现阶段韵律层级结构的标注工作主要是由人工来进行，也有不少研究人员开始尝试自动标注韵律结构的工作。在构建好语料库之后，便是根据语料库训练各模型，在利用这些训练好的模型合成语音。

由于当前韵律结构标注任务主要是依赖人工标注，而成千上万句的样本标注工作费时费力，且不同标注人员因为对部分句子的理解不同导致标注的不一致问题，往往需要多个标注人员共同协商确定有争议的标注样本，这导致构建语料库的效率不高。自动标注不仅能加快语料库的构建步骤，而且能很好地解决这种标注不一致问题，一个能精准地进行自动标注韵律结构的工具对于语音合成系统有着重要的意义。

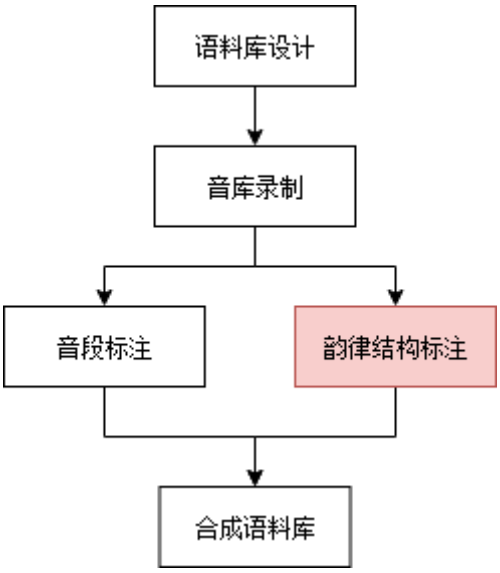


图 1.2 语料库准备阶段的流程图

1.1.3 韵律结构预测

一个典型的参数化语音合成系统，其合成步骤如图 1.3 所示，通常包括：文本正则化、分词与词性标注、韵律结构预测、字音转换、时长预测、声学特征参数预测、声码器合成语音。其中韵律结构预测模型的训练依赖大量已标注好韵律层级结构的文本数据。

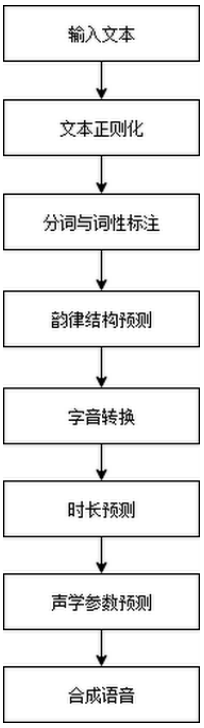


图 1.3 中文 TTS 合成流程图

参数化语音合成系统各流程的主要功能介绍如下：

1、文本正则化。将输入文本规范化为汉字序列，如将“2019”转化为“二零一九”，将“2/3”转化为“三分之二”。

2、分词与词性标注。将输入句子断为语法词序列，并标注好各词词性。

3、韵律结构预测。根据句中语法词序列，预测出韵律层级结构序列。

4、字音转换。将汉字转化为拼音音节，确定字符的发音。

5、时长预测。根据音素的文本特征预测出该音节的发音时长。

6、声学参数预测。根据音素的文本特征预测出该音节的声学参数，通常包括谱参数、清浊标记、基频、非周期信息。

7、声码器合成语音。根据声学参数信息合成语音波形信号，声码器主要有STRAIGHT[1]、WORLD[2]，以及最近提出的神经声码器 WaveNet[3]。

可以看到韵律结构预测是参数化语音合成系统中的一个必要模块，它是停顿位置预测所依赖的主要模块，也是时长模型、声学模型的前置模块，其预测效果关系到语音合成系统中的自然度。一个依据文本特征精确预测韵律层级结构的模型对提升语音合成的自然度也相当重要。

1.2 本文主要的研究内容和贡献

1.2.1 研究内容和各章简介

本文对语音合成系统里语料库准备阶段的韵律结构自动标注、合成阶段的韵律结构预测任务分别进行了研究。如下是各章简介。

第一章，文章引言，主要介绍了韵律结构、韵律结构自动标注、韵律结构预测的概念。

第二章，介绍了韵律结构自动标注与韵律结构预测任务的相关研究。

第三章，介绍了一个联合深度神经网络、双向长短时记忆网络、条件随机场的混合网络结构，联合文本与声学特征对韵律层级结构进行标注。

第四章，介绍了基于深度自注意力网络的韵律层级结构预测的方法。

第五章，介绍了基于 Transformer 双向编码器表示模型（BERT）预训练的词向量及其在韵律结构预测中应用。

第六章，对韵律层级结构自动标注与韵律结构预测工作进行了总结与展望。

1.2.2 本文主要贡献

论文主要贡献如下：

1、提出了一种基于 DNN-BGRU-CRF 的混合结构网络，联合文本、声学特征对韵律结构进行自动标注的方法，同时探究了在神经网络模型中，有利于提升自动标注性能词尾声学特征、词尾音素嵌入新型特征。

在进行韵律层级结构标注时，首先需要探究有利于提升标注性能的文本特征与声学特征，以及确定这些特征是否在神经网络模型中确有效果。通过实验发现引入词尾声学统计特征、词尾音素嵌入表达，能在神经网络模型中提升标注性能。在所提出的 DNN-BGRU-CRF 网络模型中，综合各类神经网络的优势。其中，DNN 用于学习高层的特征表达、BGRU-RNN 适用于学习上下文依赖信息，CRF 适用于解码时考虑全句整个标注序列，进行整句解码。通过该方法，可以实现对韵律词、韵律短语、语调短语进行自动标注。通过和人工标注的结果比较，能在韵律词、语调短语达到较大的一致性，韵律短语也能取得可接受的结果。通过与 CRF 自动标注结果比较，本文的方法能够达到更高的标注准确率，并且在自动标注结果与人工标注结果不一致的情况下，本文的标注结果来看存在更少的将低层级韵律结构标注为高层级的韵律结构的现象。这相对于 CRF 的标注结果是有优势的，因为将低层级韵律结构标注为高层级的韵律结构，会影响到利用这些自动标注的数据训练得到的韵律结构预测模型，接着会影响到韵律结构的预测结果，而导致错误停顿插入影响合成语音自然度的听感体验。

2、提出了基于深度注意力神经网络对韵律结构进行预测的方法。

在语音合成阶段，韵律结构预测任务至关重要，前端文本上下文特征分析模块、合成语音恰当位置的停顿插入均依赖于韵律结构预测的结果。以往的方法有使用 CRF 建模，但 CRF 需要预先指定上下文范围，并且因为该模型存在马尔科夫假设，导致难以学习到长时依赖的关系；同样也有采用词向量或者字向量，基于循环神经网络（Recurrent Neural Network，RNN）来的网络模型来进行韵律结构预测，但 RNN 也难以学习到全句范围的直接依赖关系。本文提出一个基于深度自注意力网络（Deep Self-attention Neural Network）的韵律结构预测方法，运用多头自注意力的机制，将自注意力子层与非线性子层叠加构成一个块，再堆叠多个这样相同的块形成深度网络。实验效果显示，该方法预测准确率优于 CRF、RNN 基准模型，并且不需要如同 CRF 由人工凭借经验预先指定上下文范围、设计特征模板，根据自注意力机制能自动捕捉全句中任意距离的依赖关系。

3、提出一种基于 BERT 词向量的韵律结构预测方法。

在韵律结构预测任务中，有很多方法直接采用字向量（character embedding vector）或者词向量（word vector）进行韵律结构预测。一个富含上下文信息的词向量更有利于提升各种自然语言处理任务的性能。以往的方法大都采用 Skip-Gram 方法训练一个词向量或者直接使用随机初始化的词嵌入层来获取词向量，这些词向量包含上下文的信息十分有限。最新的研究成果显示，BERT 词向量因为采用自注意力机制在大语料库上预训练得到，将预训练的词向量用于自然语言处理相关的下游任务，能取得显著效果。因为韵律结构预测也依赖于词向量特征，本文尝试将 BERT 词向量引入进韵律结构预测任务，利用其富含上下文信息的词向量作为输入特征，有助于提升韵律结构预测性能。实验结果证实，通过引入 BERT 模型在大语料库预训练的词向量到韵律结构预测任务，能提升韵律结构预测的总体正确率。

第2章 相关研究综述

在本章，分别对韵律结构自动标注、韵律结构预测两项任务进行了研究现状的综述。

2.1 韵律结构自动标注

为了构建语音合成语料库，往往需要训练有素的标注人员通过看文本、听音频的方式对文本数据进行韵律层级结构的标注。培训标注人员以及标注人员再去标注海量数据，耗时费力。一个自动标注韵律层级结构的系统需要能够根据提供的文本数据、音频数据，自动地对句子中的语法词标定韵律层级结构。

韵律结构标注任务是，在给定文本与对应的音频的情况下，标注各语法词的韵律层级结构类型，包括 NB（Not a prosodic Boundary，非韵律结构边界类型）、PW、PPH 和 IPH。其主要流程如图 2.1 所示，首先输入文本及其对应的音频文件，接着进行分词以及语音与文本的强制对齐，然后分别提取各词的文本特征、声学特征，再根据文本特征、声学特征标注标注各词的韵律类型，得到标注结果。

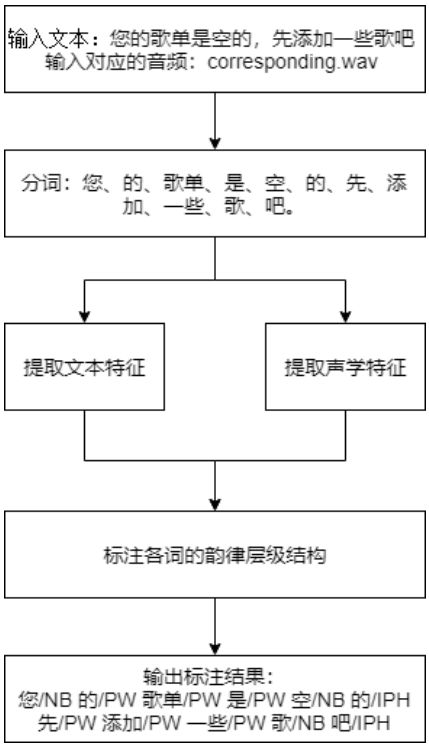


图 2.1 韵律结构标注流程图

韵律层级结构自动标注可以看做是一个分类问题，依据文本、声学特征来划分语法词对应的韵律层级结构。目前，主要是运用一些机器学习算法来进行建模，包括决策树（Decision Tree），隐马尔科夫模型（Hidden Markov Model，HMM），条件随机场（Conditional Random Fields，CRF）以及循环神经网络（Reccurent Neural Network，RNN）。接下来，分别介绍韵律结构自动标注的相关研究工作。

2.1.1 基于决策树的韵律结构自动标注

决策树模型很早就应用于韵律结构自动标注[5]，一种使用该模型进行自动标注的流程如图 2.2 所示。通过基于 HMM 的识别器进行音段标注，得到音素级的时间对齐信息。由于在韵律结构边界处，基频、能量均存在显著变化，所以也提取了语音信号的基频与能量信息。同时也提取了词尾音素时长，词后停顿这些与韵律结构密切相关的特征。在韵律结构边界处，往往存在词尾音素时长延长现象，这样的现象除了在英语中有所发现，在俄语及瑞典语中也有这类现象，该特征语音独立性[6]。提取了这些特征以后，经过决策树建模就能得到其韵律结构标签的概率，但由于仅用决策树判断忽略了其上下文信息，所以采用马尔科夫模型用于序列建模。联合决策树与隐马尔科夫模型得到最佳标注序列，从而实现韵律结构的自动标注。

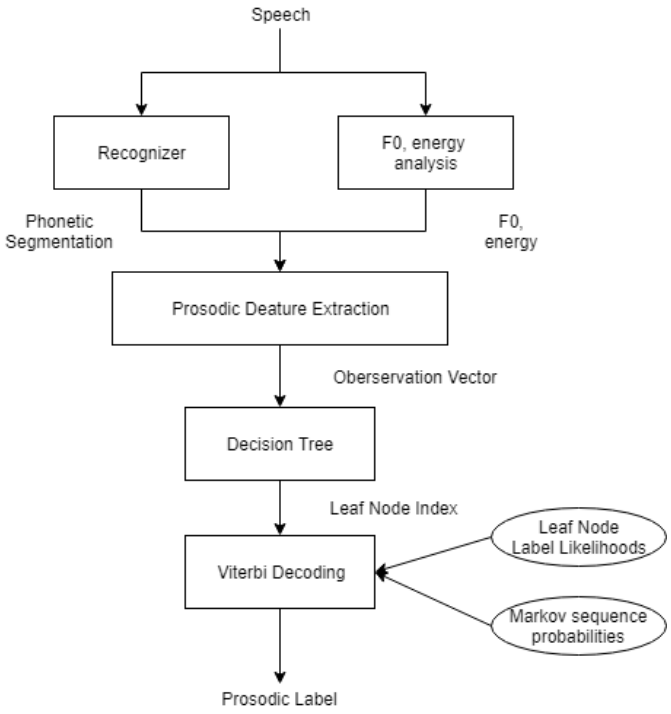


图 2.2 基于决策树的韵律结构自动标注系统

输入观察特征向量 x，决策树通过一系列的预设问题，选择树的分支直到叶子

节点，而中间节点处所询问的问题主要通过训练来进行确定。当到达的叶子节点 $T(x)$确定后，其所属类别也就确定了。根据观察特征向量 x，决策树可以得到条件概率分布$p(\alpha|T(x))$。由于上一个特征向量与当前特征向量存在一定的相关性，所以需要进行序列上的建模，一个特征序列记为$x_1^n = \{x_1, \dots, x_n\}$，标签序列记为$\alpha_1^n = \{\alpha_1, \dots, \alpha_n\}$，通过隐马尔科夫模型，可以得到：

$$p(\alpha_1, \dots, \alpha_n) = p(\alpha_1) \prod_{i=2}^{n} p(\alpha_i|\alpha_{i-1}) \tag{2-1}$$

联合决策树与隐马尔科夫模型，就能得到：

$$\begin{aligned} p(\mathrm{x}_1, \dots, \mathrm{x}_n, \alpha_1, \dots, \alpha_n) &= p(\mathrm{x}_1|\alpha_1)p(\alpha_1) \prod_{i=2}^{n} p(\alpha_i|\alpha_{i-1})p(\mathrm{x}_i|\alpha_i) \\ &= [\prod_{j=2}^{n} p(x_j)]\, L(\alpha_1|x_1) \prod_{i=2}^{n} p(\alpha_i|\alpha_{i-1})\, L(\alpha_i|x_i) \end{aligned} \tag{2-2}$$

其中

$$L(\alpha|\mathrm{x}) = \frac{p(\alpha|x)}{p(\alpha)} = \frac{p(x|\alpha)}{p(x)} \tag{2-3}$$

由决策树建模得到$p(\alpha|x)$，马尔科夫模型可以得到$p(\alpha_i|\alpha_{i-1})$。自动标注问题可以简化为以下的形式：

$$\begin{aligned} \widehat{\alpha_1^n} &= \arg\max_{\alpha_1^n} p(\alpha_1^n, x_1^n) \\ &= \arg\max_{\alpha_1^n} L(\alpha_1^n \mid x_1^n)\mathrm{p}(\alpha_1) \cdot \prod_{i=2}^{n} p(\alpha_i|\alpha_{i-1})\, L(\alpha_i|x_i) \end{aligned} \tag{2-4}$$

对于上式，可以使用动态规划的算法求解，比如 Viterbi 算法[7]，得到一个最佳标注序列，即可实现韵律结构的自动标注。

2.1.2 基于隐马尔科夫模型的韵律结构自动标注

研究人员有采用隐马尔科夫模型来实现韵律结构的自动标注[8][9][10]。杨辰雨曾采用 CD-HMM（Context-Dependent HMM，带上下文信息的隐马尔科夫模型）来进行韵律结构的自动标注[9]，其主要流程如图 2.3 所示。

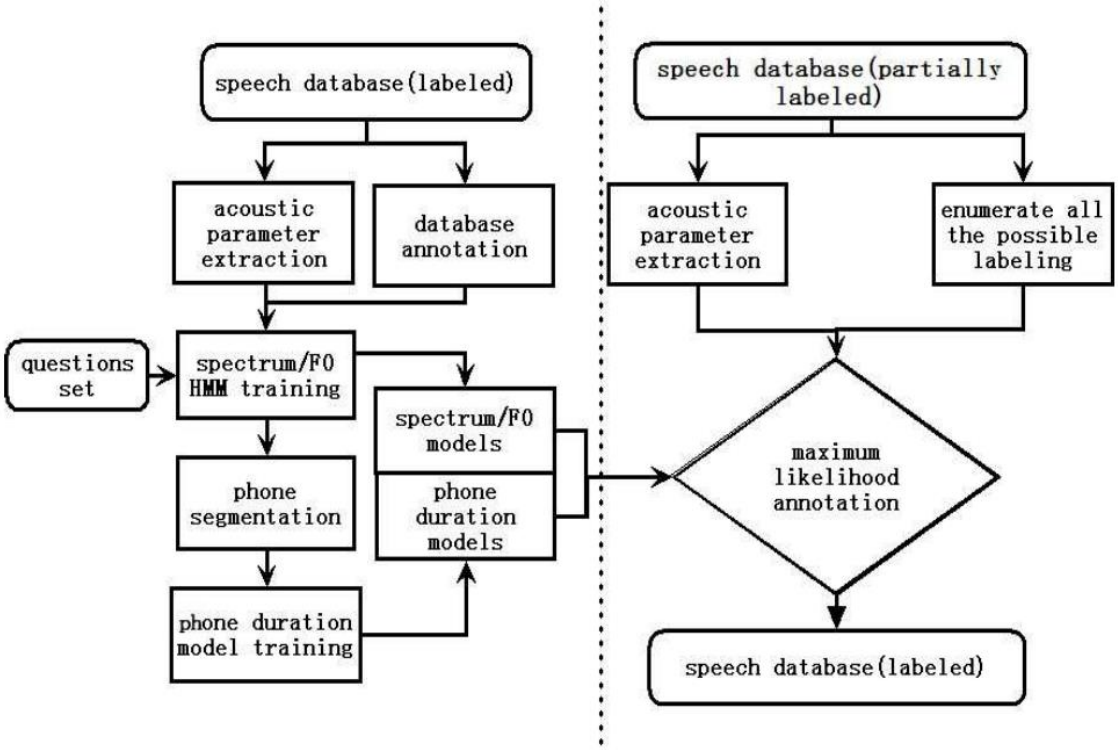


图 2.3 基于隐马尔科夫模型的韵律结构自动标注系统[9]

图中左侧为训练 HMM 阶段，右侧为标注阶段。在训练阶段，以音素与该音素后的韵律结构类型，组成带韵律结构类型的音素上下文，通过 HMM 训练得到上下文相关的 HMM 时长模型与声学模型。在标注阶段，先对句子进行分词，语法词内部的音素可以确定为非韵律结构边界类型，这样可以形成部分标注的状态。对语法词最后一个音素后接的韵律结构类型进行遍历所有可能，形成一个音素网网格。借鉴语音识别中的技术，利用已训练好的 HMM 时长、声学模型在网格中确定一条最优路径。路径中每个音素都是带韵律结构类型上下文的音素，这样也就确定了每个音素后对应的韵律层级结构边界类型，也就确定了每个语法词后对应的韵律结构类型。

2.1.3 基于条件随机场模型的韵律结构自动标注

条件随机场（Conditional Random Fields，CRF）是序列标注中常用的一个模型。在韵律结构自动标注上，该方法也得到广泛的应用[11][12][13]。有研究人员提取了较以往方法更为丰富的特征集，采用链式 CRF 来提升韵律结构的自动标注效果，在说话人相关以及说话人无关的情况下，均取得了优于之前模型的效果[11]。链式 CRF 可以用以下公式表示：

$$p_{\vec{\lambda}}(\vec{y}|\vec{x}) = \frac{1}{Z_{\vec{\lambda}}(\vec{x})} exp(\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_j f_j(y_{i-1}, y_i, \vec{x}, i)) \quad (2\text{-}5)$$

其中，i 是输入序列的位置索引，特征函数$f_j(y_{i-1}, y_i, \vec{x}, i)$可以为状态特征函数$s(y_i, \vec{x}, i)$或者转移特征函数$t(y_{i-1}, y_i, \vec{x}, i)$，$\lambda_j$为待学习的参数，$Z_{\vec{\lambda}}(\vec{x})$为规范化因子。训练好的 CRF 模型，可以根据给定的观察特征序列，输出最佳的标注序列。

根据提供的文本与音频，提取的特征分为文本特征、声学特征两大类。文本特征包括如下：

1、词面（词的字符表示）。

2、词长。词性。

3、词后标点符号。

4、是否功能词。

5、是否句首词。

6、二元语言模型。

7、短语词典。

声学特征如下：

1、词后停顿。

2、词尾音节的时长。

3、词尾音节是否重音。

以上文本与声学特征，均采用五元上下文（当前词、前两词、后两词）的拼接特征作为作为输入特征。CRF 模型根据这些输入特征，输出最佳的标注序列。模型的实现采用 CRF++工具包[23]。

2.1.4 基于 RNN 的韵律结构自动标注

近年来，深度学习在语音领域的应用得到大规模的应用，取得显著效果。研究人员也开始使用序列建模能力更强的 RNN 模型进行韵律结构的自动标注[14][15][39]。有研究人员指出传统机器学习模型 HMM、CRF 因存在的马尔科夫假设[14]，限制了其捕捉上下文依赖的能力。因此提出一个基于长短时记忆网络（Recurrent Neural Network with Long Short Term Memory unit，LSTM-RNN）的自动标注韵律结构的方法。

基于 LSTM-RNN 的自动标注方法同样采用了文本特征与声学特征。其中，文本特征包括：

1、当前词与相邻词语的耦合度。这需要训练二元语言模型，包括一个前向的二元语言模型和后向的二元语言模型。其耦合度主要有两个数值构成，分别记为$lm1_k = p(w_k|w_{k-1})$，$lm2_k = p(w_k|w_{k+1})$，其中 k 为当前词的索引值。

2、词性。

声学特征包括：

1、基频。词中基频的统计信息，统计最大、最小、平均、方差、最大值对应的标准值。

2、能量。词中能量的统计信息，统计最大、最小、平均、方差、最大值对应的标准值。

3、基频、能量的函数。韵律信息体现在基频与能量共同影响上，这里采用了对数基频与能量的乘积作为特征。

4、频谱倾斜。计算帧中频率高于 500Hz 部分的能量与该帧总能量的比值。

5、基频变化。主要分为三类：基频下倾、基频平稳、基频上升。

6、差分。计算基频、能量曲线的差分。

采用的 LSTM-RNN 的网络结构为：一层 BLSTM-RNN 隐层和一个 softmax 输出层。Softmax 输出层输出韵律层级结构类型的概率分布，取最大概率对应的韵律结构类型为标注结果。

2.1.5 韵律结构自动标注的研究现状总结

韵律结构自动标注问题，本质上是一个特征序列到标签序列的映射问题。建模方法也是采用机器学习的方法，其研究方法与机器学习方法的发展密切相关。早期，采用决策树这样的静态分类器对单个词语进行标注，这样缺乏上下文建模能力，难以学习序列映射。随着 HMM、CRF 这样有上下文建模能力的模型提出，研究人员利用这些模型的序列建模 d 能力，将其应用于韵律结构自动标注，但 HMM、CRF 这样的动态模型也有缺点，它们都存在马尔科夫假设，学习上下文依赖关系的能力不够好。近年来，RNN 在序列建模上因为其较好地序列建模能力，在韵律结构预测上得到很好的应用。

在韵律结构自动标注任务上，尽管有研究人员采用不同的机器学习方法来建模，但综合各种机器学习的优势联合进行建模的方法并没有得到充分的研究。比如，RNN 尽管能够学习上下文的依赖关系，而这些依赖关系仅限于输入特征上或者隐层特征上，在输出标签间的依赖关系没有考虑到。而这在韵律结构自动标注任务中，

相邻两词标签间的依赖关系是显而易见的。比如，上一词标注的是语调短语边界，那么当前词仍旧是语调短语边界的可能性极小。另外，在深度学习模型中，采用的词级特征也需要实验验证确定韵律结构自动标注一个合适的词级特征集，尤其是词尾部的特征，这方面适用于深度神经网络的特征并没有得到充分的研究。

2.2 韵律结构预测

韵律结构预测任务是指在语音合成阶段，给定文本预测韵律文本中各词的韵律结构，处于语音合成的文本处理前端部分。因为后续在合成语音中插入停顿以及上下文文本特征分析均依赖于韵律结构预测结果，所以一个预测精准的韵律结构对合成语音的自然度至关重要。

韵律结构预测也是一个分类问题，同样可以采用三层韵律结构。与自动标注任务不同的是，在合成阶段，能作为特征的只有文本特征。先通过分词得到语法词序列，再提取各语法词的文本特征，再通过机器学习算法根据输入的语法词文本特征预测该语法词的韵律结构。韵律结构预测任务的流程如图 2.4 所示。

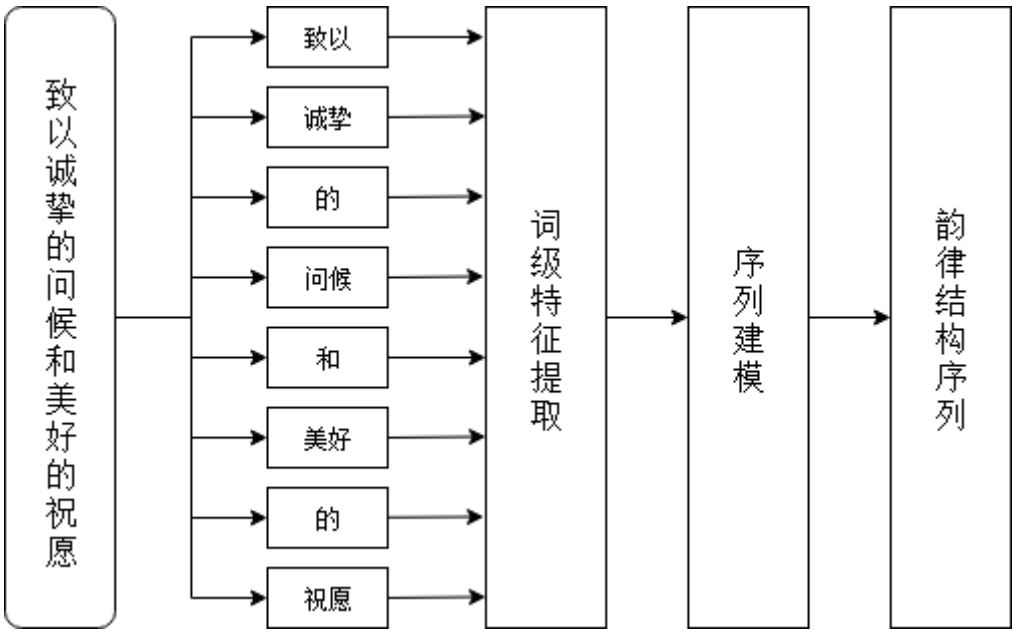


图 2.4 韵律结构预测任务流程图

韵律结构预测主要是采用机器学习的方法进行建模，如分类回归树（Classfication And Regression Tree，CART）、隐马尔科夫模型（HMM）、条件随机场（CRF）以及循环神经网络（RNN）等。接下来，分别介绍韵律结构预测的相关研究现状。

2.2.1 基于分类回归树的韵律结构预测

分类回归树是较早应用于韵律结构预测的模型[16][18]。有研究人员人分析了一系列有助于韵律结构预测任务的特征，如当前词位置到句首位置的距离，当前词位置到句尾位置的距离。因为接近句首与句尾位置的词后不太可能是语调短语。词性也与韵律结构边界密切相关。当前词距离上一韵律结构边界也可作为特征。提取这一系列特征后，作为特征变量。根据训练样本，分类回归树确定每个节点的分支问题。在测试集上，提供每个样本的特征向量，可根据问题集确定一条路径达到叶子节点，从而确定每个样本对应的韵律结构分类。

CART 树建模的缺点在于其是静态分类器，无法捕捉上下文的关联性，难以对序列进行建模。仅能从输入特征上，通过拼接前后词的特征，形成固定范围的上下文特征。随着机器学习方法的发展，能学习序列到序列间映射关系的动态模型如 HMM、CRF 在韵律结构预测任务上得到广泛应用。

2.2.2 基于隐马尔科夫模型的韵律结构预测

隐马尔科夫模型也被应用于韵律结构预测任务[19][20]。Taylor[19]等人利用 HMM 对于韵律结构进行建模，将词性作为 HMM 的观察，将韵律结构类型作为 HMM 的隐藏状态。通过隐马尔科夫模型建立词性序列到韵律结构边界序列的映射。记j_i为句子中第 i 个词的韵律结构边界类型，J_{i-1}^{N}为句子中第 i 个词前边 N 个词的韵律结构边界类型序列。主要关注$P(j_i)$，在给定的当前词之前的标注序列以及当前词的词性C_i后，有：

$$P(j_i|C_i,J_{i-1}^{N}) = P((j_i|J_{i-1}^{N})|C_i) \tag{2-6}$$

由贝叶斯概率公式，可得：

$$P(j_i|J_{i-1}^{N}|C_i) \propto P(j_i|J_{i-1}^{N}) \cdot P(C_i|(j_i|J_{i-1}^{N})) \tag{2-7}$$

假定：

$$P\left(C_i\middle|(j_i|J_{i-1}^{N})\right) = P(C_i|j_i) \tag{2-8}$$

那么有：

$$P(j_i) \propto P(j_i|J_{i-1}^{N}) \cdot P(C_i|j_i) \tag{2-9}$$

以上公式就可根据隐马尔科夫模型求解，得到当前词的韵律结构类型。进行时间步的迭代，就能得到所标注的韵律结构序列。

2.2.3 基于条件随机场模型的韵律结构预测

条件随机场模型也被广泛应用于韵律结构预测任务[11][21][22]。有研究人员将条件随机场应用于韵律结构预测[11]。与基于条件随机场模型的韵律结构自动标注不同之处，仅在于特征上的区别，韵律结构预测仅能使用文本特征，建模方法同样采用 2.1.3 的链式 CRF 模型。其韵律结构预测模型所用到的文本特征如下：

1、词面（词的字符表示）。

2、词长。词性。

3、词后标点符号。

4、是否功能词。

5、是否句首词。

6、二元语言模型。

7、短语词典。

以上特征，均采用五元上下文（当前词、前两词、后两词）的拼接特征作为作为输入特征。CRF 模型根据这些输入特征，输出最佳的韵律结构序列。模型的实现采用 CRF++工具包[23]。

2.2.4 基于 RNN 的韵律结构预测

韵律结构预测是特征序列到韵律结构序列间的建模，RNN 适用于序列建模。最近，有研究人员使用 RNN 来进行韵律结构预测[16][24][25]。比如，有研究人员使用两层 LSTM-RNN 隐层和一个 softmax 输出层的网络结构[24]，隐层大小设定为 200 来进行韵律结构预测。输入特征采用词向量，这个词向量是通过随机初始化、跟随训练数据更新的词嵌入层得到。网络输出每个词对应的韵律结构类型的概率分布，取最大概率对应的韵律结构类型为预测类型。除了词向量，也有研究人员采用字向量来进行韵律结构预测[26]，这样就不必进行预先进行分词。

2.2.5 韵律结构预测的研究现状总结

与韵律结构自动标注类似，韵律结构预测的研究和机器学习方法的发展密切相关。建模模型主要有 CART、HMM、CRF 以及 RNN。近年来，建模方法主要是采用 RNN。为了免于特征工程，模型输入是主要用字向量或者词向量。采用字向量的韵律结构预测又使得韵律结构预测工作不直接依赖于分词。也有研究人员将使用字向量的韵律结构预测任务与分词任务进行多任务学习，利用韵律结构预测与分词信息的联系，更好地进行预测[17]。最近，为了学习相邻两词的韵律结构标签

之间的依赖关系，采用 LSTM-CRF 进行建模的研究。但是，当前的研究工作仍然限于 LSTM-RNN、GRU-RNN 来学习依赖关系，而 RNN 学习依赖关系也存在局限性，只能学习到一个方向上的依赖。要学习双向依赖关系，需要两个方向的 RNN 输出进行拼接，很难学习到全句范围任意两词间的依赖关系，也难以捕捉句子的结构信息。在韵律结构预测任务上，不同的方法大致体现在模型或者特征的不同。当前，在模型上，将比 RNN 拥有更好的序列建模能力的模型应用于该任务的研究工作尚少。在特征上，研究不同方式训练的词向量对于韵律结构预测任务影响的工作，尤其是富上下文信息的词向量应用于韵律结构预测的工作，这方面的研究不多，而词向量的选择对于韵律结构预测任务非常重要。这个研究工作对后续基于深度学习韵律结构预测的特征选择上有一点的参考意义。

第3章　基于 DNN-BGRU-CRF 的韵律结构自动标注

3.1 本章引论

汉语参数化语音合成系统需要大量详细标注的数据，所以语料库的构建必不可少，而语料库中韵律结构的标注工作主要依赖于人工标注。人工标注主要有两个缺点，一是需要训练专门的标注人员来进行标注，并且成千上万句语音样本的标注工作费时费力，二是因为不同的标注人员因为对句子的理解不一样，导致少部分样本的标注存在不一致的地方。为了解决如上的确定，自动标注韵律结构的工作显得非常必要。近来也有不少研究人员采用决策树、隐马尔科夫模型、条件随机场来实现自动标注，但大多数方法均只能标注语调短语或者韵律短语等一两种韵律层级结构。本章介绍一个 DNN-BGRU-CRF 的混合网络结构，联合文本与声学特征，同时标注三种韵律层级结构，并在特征上采用了词尾声学统计特征、词尾音素嵌入等词尾特征。该标注方法不仅能替代人工标注，而且相对于传统的 CRF 方法取得更为精确的标注效果。在标注错误的情况下，本文提出的方法能较少的地将低层级的韵律结构边界预测为高层级的韵律结构边界，从而减少标注错误对后续语音合成带来的不利影响。

3.2 问题定义

给定文本与文本对应的音频，将句子文本划分为语法词序列（w_1，w_2，...，w_i，...，w_n），其中w_i为第i个语法词，根据句中各语法词的文本特征与声学特征，标定各词的韵律边界类型（y_1，y_2，...，y_i，...，y_n），其中 y_i 为 w_i 对应的韵律层级结构类型，包括非韵律边界（NB）、韵律词（PW）、韵律短语（PPH）和语调短语边界（IPH）。

3.3 基于 DNN-BGRU-CRF 的韵律结构标注模型

3.3.1 深度神经网络 DNN

深度神经网络(DNN)是近年来机器学习众多算法中的一个热点研究方向，在语音信号处理中，DNN 也取得了较好的效果。DNN 是一个多层神经网络的结构，相较于浅层的模型如 SVM、单层感知机等，DNN 能够模拟更为复杂的函数，具有较

强的建模能力。神经网络的神经元之间的运算通常是由线性运算和非线性的激活函数构成。全连接为当前层的神经元与上一层的每个神经元都有连接。通过堆叠全连接层，DNN 可以将底层的特征形成更为抽象的高层特征表示。如图 3.1 所示，一个有着一个输入层、三个隐层、一个输出层结构的神经网络。

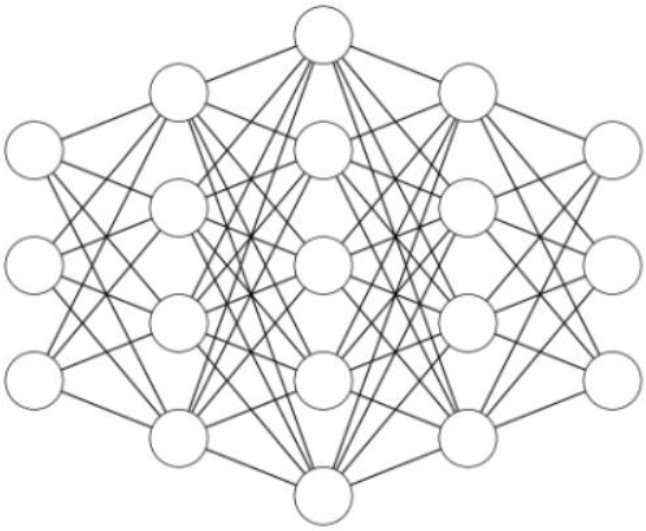


图 3.1 DNN 深度神经网络示例

神经网络的训练过程，就是一个根据训练数据调整模型参数的过程。反向传播是学习模型参数的算法。它基于梯度下降的策略对参数沿负梯度方向更新参数。本文采用 DNN，一是对词性、词长、词后标点等文本特征进行特征融合，二是输出这些特征的高层表示。

3.3.2 带门控循环单元的循环神经网络 BGRU-RNN

韵律层级结构标注问题可以看作是一个序列标注问题，RNN 是适合用于序列建模的神经网络，但因为存在梯度消散或者爆炸的问题，难以有效地学习到长时依赖关系，带门控制的长短时记忆单元（Long Short Term Memory，LSTM）[28]是为了解决梯度消散问题，用以学习长时依赖的结构，门控循环单元（Gated Recurrent Unit，GRU）[29]是 LSTM 的一个变种，它有着类似于 LSTM 的门控结构，但更为精简，训练参数较少。双向 RNN[30]是一种能捕捉双向上下文信息的网络结构。本文采用 BGRU-RNN 的结构来学习双向的上下文依赖关系。

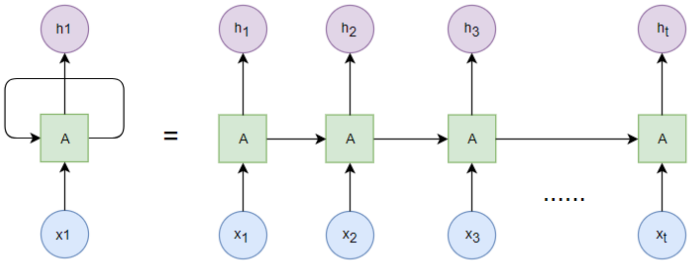


图 3.2 RNN 循环神经网络示例[31]

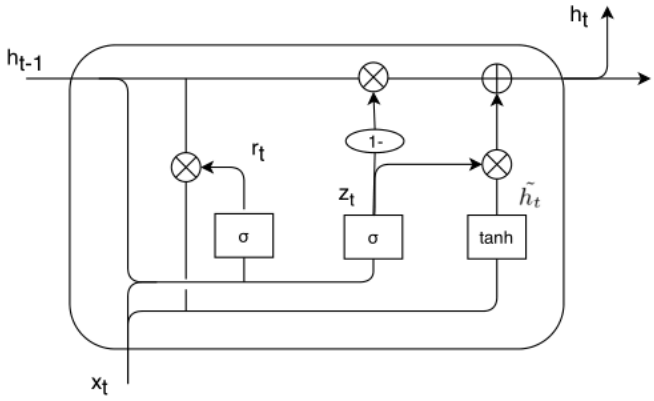


图 3.3 GRU 结构图[31]

GRU 的具体计算公式如下所示：

$$z_t = \sigma(W_z[h_{t-1}, x_t]) \tag{3-1}$$

$$r_t = \sigma(W_r[h_{t-1}, x_t]) \tag{3-2}$$

$$\widehat{h_t} = tanh(W[r_t \odot h_{t-1}, x_t]) \tag{3-3}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widehat{h_t} \tag{3-4}$$

其中，⊙表示按元素相乘，h_{t-1}和$\widehat{h_t}$分别为前一记忆单元传来的内容以及当前记忆单元更新内容。W_z，W_r以及 W 分别是要学习的权重参数。时间步 t 时刻的输入为x_t。更新门z_t控制前一时刻信息被保留的程度，重置门r_t用于控制忽略前一时刻的状态信息的程度。

3.3.3 条件随机场 CRF

条件随机场（CRF）是一种基于概率图模型的统计学习算法，它被广泛应用于词性标注、韵律结构标注任务中[32][12]。条件随机场对条件分布进行建模，是一种判别式模型。具体来看，给定观测序列$x = (x_1, x_2, \dots, x_n)$，以及对应的标记序列$y = (y_1, y_2, \dots, y_n)$，条件随机场则是要建模得到条件概率$P(y|x)$，其条件概率通常定义

为状态特征函数与转移特征函数的和，也称之为点函数与边函数，可用以下公式表示：

$$P(y|x) = \frac{1}{Z}exp(\sum_{j}\sum_{i=1}^{n-1}\lambda_j t_j(y_{i+1}, y_i, x, i) + \sum_{k}\sum_{i=1}^{n-1}\mu_k s_k(y_i, x, i)) \qquad (3\text{-}5)$$

其中，$t_j(y_{i+1}, y_i, x, i)$为转移特征函数，描述了两个相邻时间步的标签转移以及观察状态间的关系，$s_k(y_i, x, i)$为状态特征函数，描述了观察状态与标签之间的关系，Z为规范化因子，λ_j与μ_k为学习的参数。链式条件随机场模型结构图如 3.4 所示。

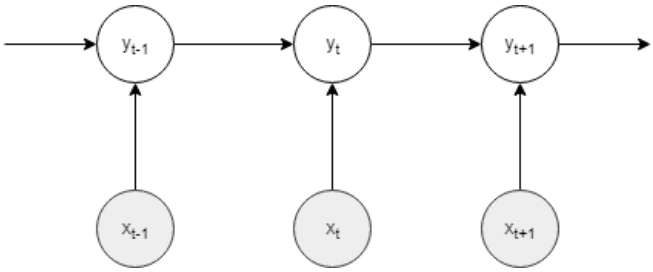


图 3.4 CRF 链式条件随机场模型[33]

条件随机场需要定义特征函数，而这些特征函数往往依赖于经验，需要专家知识来设计特征函数。在工程实现上，特征函数的取舍、调试均会耗费大量的时间精力以确定一个合适的特征模板。该模型在序列标注，包括在词性标注、韵律结构自动标注、韵律结构预测等任务上，有着广泛的应用。

隐层为 BLSTM、输出层为 CRF 的混合网络（BLSTM-CRF）成功应用于词性标注任务[35]。联合 BLSTM 与 CRF 建模，CRF 可以做到解码时考虑到整个序列的标注，并得到较好的词性标注效果，其模型结构如图 3.5 所示。BLSTM 后的 softmax 层输出各时间步所有标签的概率分布。此时，CRF 主要对每两相邻标签转移概率进行建模，且转移概率与输入不相关。该方法经实验结果证实，在词性标注的任务上优于单纯使用 BLSTM 的模型。

在韵律结构标注中，上一语法词的标注结果与当前词的标注结果存在密切的联系，比如上一语法词的标注结果为语调短语边界，那么当前词再标注为语调短语边界的可能性极小。CRF 因为能学习到句子中各词的不同韵律结构之间的转移关系，所以在韵律结构标注时使用 CRF 作为输出层能考虑到这种规律，从整个句子上统筹考虑，得到更好的结果。以往逐时间步采用简单的选择最大概率作为标注结果，将各语法词的标注看做相互独立的。采用 BLSTM-CRF 这样的混合结构就能将两相邻时间步的标注结果建立起联系。本文尝试将这种混合结构引入到韵律结

构标注任务。

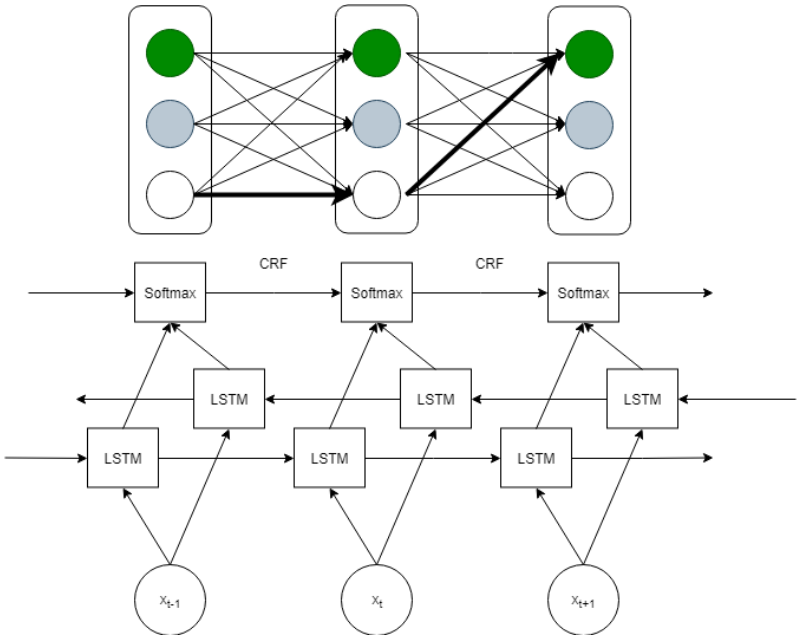


图 3.5 BLSTM-CRF 模型结构图[34]

3.3.4　DNN-BGRU-CRF 模型结构

本文提出的模型整体框架如图 3.6 所示。一个句子经过分词后得到词序列，再通过特征提取得到词的特征集，形成词级特征序列。每个时间步，输入当前词的词级特征，输出该词对应的韵律层级结构类型。模型采用 DNN 来学习文本特征的高层表示，采用 BGRU-RNN 学习句中各词上下文之间的依赖，采用 CRF 学习标签之间的转移概率，通过 DNN、BGRU-RNN、CRF 混叠的形式，综合各个模型的优势，将其应用于韵律结构自动标注任务。

在输入特征上，本文研究了对韵律结构自动标注有用的一系列特征，总体上包括文本特征与声学特征。用词向量[36]表示当前输入词的语义特征，词性也与韵律结构密切相关[37][38]，本文也把词性作为输入特征的一种。词后若存在标点符号例如逗号、句号，往往也意味着词后是一个较高的韵律层级结构边界[39][40]。一个较长的语法词有较大可能就是一个韵律词[41]。在声学特征方面，有研究表明词后存在停顿往往意味着该词后存在韵律结构边界[42]；在韵律短语边界处存在着词尾音节的延长现象[43]；在韵律短语边界以及语调短语边界处，存在有明显的基频重置现象[44]。有研究人员对韵律短语边界处的基频曲线进行统计分析，发现基频曲线变化程度与韵律结构类型相关[45]。

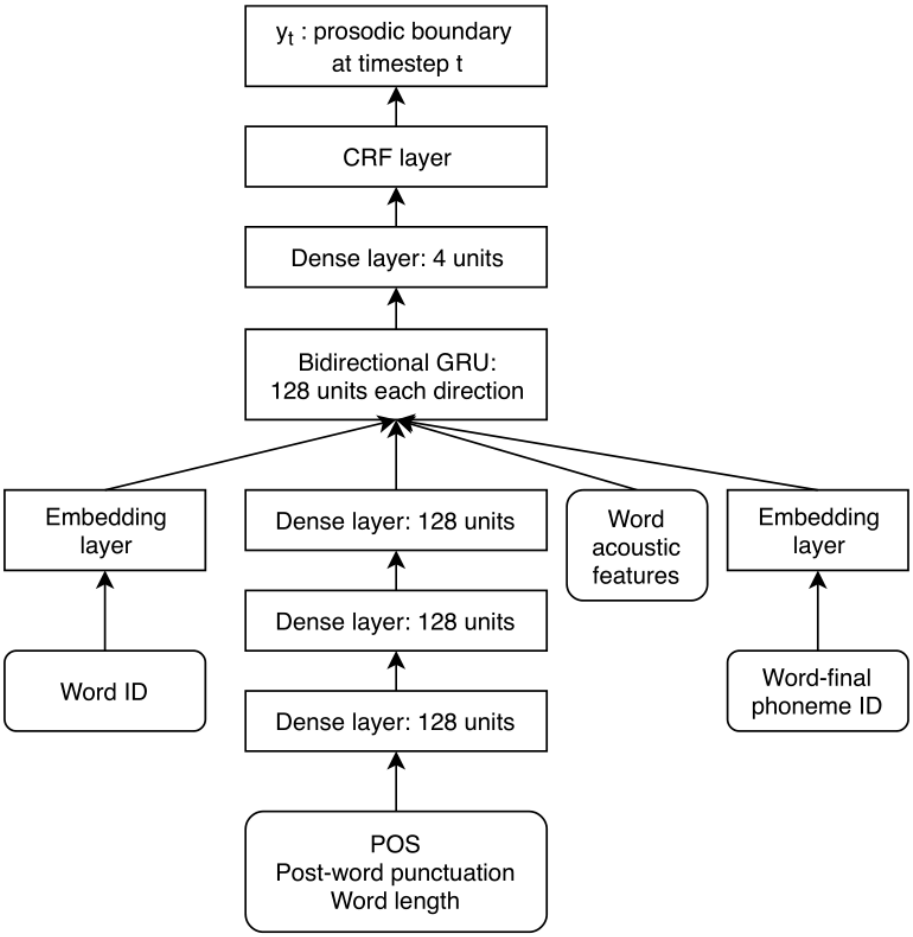


图 3.6 基于 DNN-BGRU-CRF 的韵律结构自动标注模型

本文模型的输入特征总体包括两大类，文本特征和语音特征。其中，文本特征包括：

1、词 ID，表示输入词的索引。

2、词尾音素 ID，表示输入词词尾音素的索引。

3、词性，表示词的语法特征，如名词、形容词等。

词长，语法词中字的个数。

词 ID 输入到一个可训练的词嵌入层，得到 200 维的词向量。词尾音素 ID 输入到一个可训练的词嵌入层，得到 64 维的音素向量。词性用 45 维独热编码（one-hot encoding）表示，词长采用 10 维独热编码表示，词后标点采用 15 维独热编码表示。

本文所用到的声学特征包括：

1、词后停顿等级，为了将停顿特征表示为向量作为神经网络的输入，本文将连续的停顿时长值离散化为停顿等级。依据停顿的时长本文划分了五个停顿等级：Pause-0 （$0 \le p < 50$ ms），Pause-1 （$50 \le p < 150$ ms），Pause-2（$150 \le p < 350$ ms），Pause-3（$350 \le p < 450$ ms），Pause-4 （$p \ge 450$ ms），这样可以用五维的独热编码表示停顿等级。

2、词尾音节时长等级，类似于词后停顿，本文将词尾音节的时长划分为九个等级，可以用九维的独热编码表示。

3、词尾音节的声学统计特征，包括词尾音节的基频、能量统计特征，统计量包括：最大值、最小值、范围、均值、方差。

4、词边界的基频、能量对数差值，当前词词尾浊音与下一词词首浊音帧的基频对数差、能量对数差。

以上声学特征中，词后停顿等级采用五维独热向量表示，词尾音节时长等级采用九维独热编码表示。词尾音节基频声学统计特征包含五个统计量，每个统计量采用一维的标量表示。词尾音节能量声学统计特征包含五个统计量，每个统计量采用以为的标量表示。当前词尾浊音帧、下一词词首浊音帧基频对数差分别采用一维白标量表示。当前词尾浊音帧、下一词词首浊音帧能量对数差分别采用一维标量表示，这些特征拼接，形成输入的词声学统计特征。

将表示词性、词长、词后标点的特征向量拼接形成 DNN 的输入向量。DNN 由三层全连接层，每层 128 个神经元，最后一层输出特征的高层表示。再与词向量、音素向量、词的声学特征拼接，作为 BGRU-RNN 的输入。BGRU-RNN 学习到下文长时关系，其输出通过一个全连接层输出一个 4 维的向量，此时采用 CRF 层学习到的状态转移矩阵，综合各时间步的输出以及状态转移概率，采用 Viterbi 序列解码的方式，输出一个最佳的标注序列。

3.4 实验结果与分析

3.4.1 数据说明

本文使用的数据集一共有 41483 句样本，包括文本以及由女声采用普通话朗读的音频，音频文件采样率 16kHz。为了便于与自动标注比较效果，这些句子都由人工标注了韵律层级结构。随机选定 37483 句作为训练集，另外 2000 句作为验证集，剩下的 2000 句作为测试集。

3.4.2 特征提取

对中文句子使用前端预处理工具进行分词与词性标注，通过文本分析可以获取到每个语法词后是否有标点、标点的类型，这样就能得到词的文本特征。

通过基于 HMM 的强制对齐工具，可以得到音段切分信息的数据。以 5ms 为一帧，那么根据强制对齐的结果，得到词尾音节的帧边界，参考各帧的基频信息，也能得到词尾浊音帧与下一词词首浊音帧的帧序号。分帧提取音频文件的基频和能量特征，得到每个句子的基频、能量曲线，再根据强制对齐结果的帧序号信息，可以得到词尾音节浊音帧的对数基频值、对数能量值，再计算得到词尾音节的对数基频、对数能量的统计值，词尾浊音帧与下一词词首浊音帧的对数基频差值、对数能量差值。如上计算的这些声学特征构成该词的声学特征集合。

经过以上步骤的处理，即可得到神经网络的各时间步的输入的词级文本特征与词级声学特征。

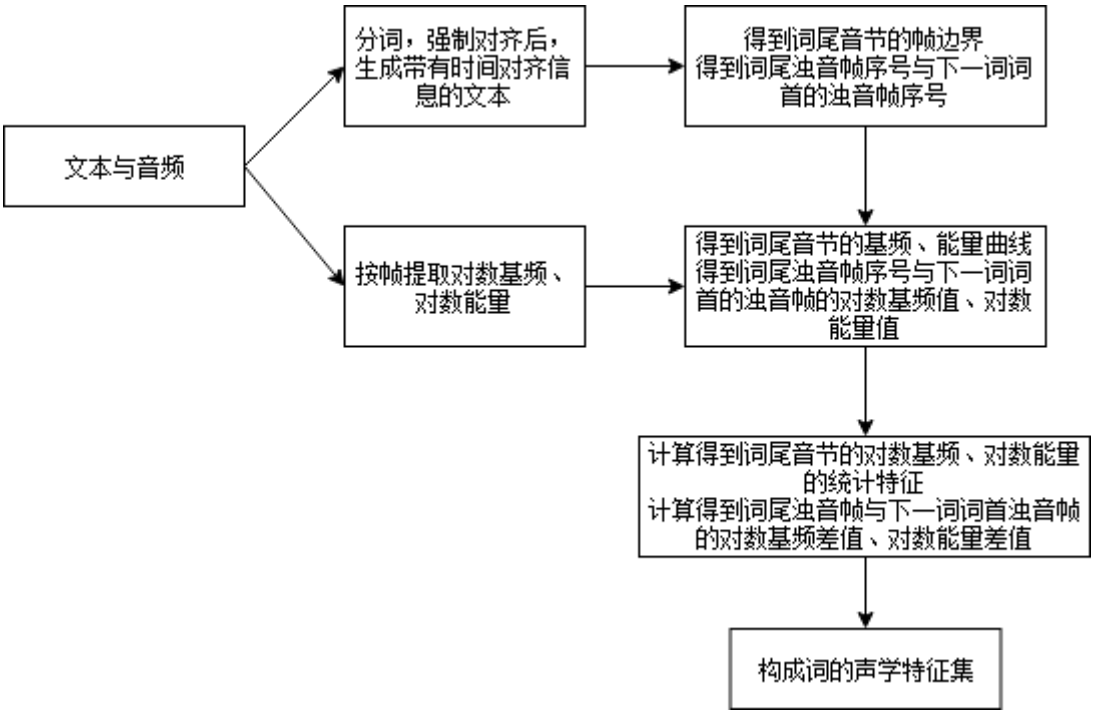


图 3.7 词的声学特征集的提取流程

3.4.3 实验设置及对比实验

本文采用了 Adam 优化方法，初始化学习率设置为 0.0001，用以训练神经网络。激活函数采用 ReLU，且每层后接 dropout 层，dropout 概率设置为 0.5，用以

防止过拟合。词嵌入层输出词向量的维度设定为 200 维，词尾音素向量输出维度设定为 64 维。

除此以外，本文还设置了几个对比实验，用以验证模型的效果：

CRF，传统的条件随机场模型。以词面、词性、词长、词后标点、词后停顿时长等级、词后音节时长等级作为输入，同时利用了文本特征与声学特征。

D-BLSTM-CRF，采用带 LSTM 单位的 RNN，网络其他配置与所提模型相同，用以对比，以确定在韵律结构自动标注任务上本文的混合结构采用的 RNN 种类。

D-BGRU-S，输出层采用 softmax 层，网络其他配置与所提模型相同，用以对比，以确定输出层采用 softmax 层还是能学习到一定上下文标签关联的 CRF 层。

D-BGRU-CRF⋆，模型结构总体与本文所提出的模型相同，但没有词尾音素 ID 以及相应的嵌入层作为输入，作为参照用于衡量词尾音素 ID 的输入对于自动标注效果的影响。

D-BGRU-CRF，本文所提出的模型，综合 DNN、BGRU、CRF 各自的优势，形成的一个混合结构模型。

3.4.4 实验评估标准

自动标注实验结果的对比评估主要有两个方面，一方面需要和人工标注的结果进行对比评估，另一方面需要和前人做自动标注的模型作对比。

与人工标注结果比较，本文采用混淆矩阵的形式进行对比。以人工标注的结构为真实值，自动标注的结果为预测值，就可以得到相应的混淆矩阵。与其他自动标注模型的比较，本文采用总体正确率 T-ACC、韵律词 F1 值、韵律短语 F1 值、语调短语 F1 值这四个指标来对比各个模型，其中以 T-ACC 为主要评估指标。

T-ACC 计算方式如下：

$$T-ACC=\frac{N_{自动标注正确的样本}}{N_{全部样本}} \tag{3-6}$$

因为不同韵律结构类型的 F1 值计算方法类似，F1 同时权衡了精确率与召回率，此处以计算语调短语 F1 值为例：

$$Precision=\frac{N_{自动标注正确的IPH样本}}{N_{自动标注为IPH的样本}} \tag{3-7}$$

$$Recall=\frac{N_{自动标注正确的IPH样本}}{N_{人工标注为IPH的样本}} \tag{3-8}$$

$$F1 = 2 * Precision * Recall \tag{3-9}$$

3.4.5 实验结果及其分析

3.4.5.1 与人工标注结果对比

以人工标注的结果为真实值，自动标注的结果为预测值，可以得到混淆矩阵。表 3.1 为基准模型 CRF 的自动标注结果的混淆矩阵，表 3.2 为本文提出的 DNN-BGRU-CRF 自动标注方法的混淆矩阵：

表 3.1 基于 CRF 自动标注方法的混淆矩阵

Manual \ Auto	NB	PW	PPH	IPH
NB	4173 (84.01%)	706 (14.21%)	86 (1.73%)	2 (0.04%)
PW	527 (7.42%)	5914 (83.31%)	644 (9.07%)	14 (0.20%)
PPH	101 (3.37%)	1093 (36.42%)	1678 (55.91%)	129 (4.30%)
IPH	4 (0.11%)	14 (0.38%)	183 (4.98%)	3471 (94.53%)

表 3.2 基于 DNN-BGRU-CRF 自动标注方法的混淆矩阵

Manual \ Auto	NB	PW	PPH	IPH
NB	4227 (85.10%)	691 (13.91%)	47 (0.95%)	2 (0.04%)
PW	505 (7.11%)	6044 (85.14%)	542 (7.63%)	8 (0.11%)
PPH	104 (3.47%)	1138 (37.92%)	1661 (55.35%)	98 (3.27%)
IPH	14 (0.38%)	14 (0.38%)	177 (4.82%)	3467 (94.42%)

与人工标注的结果相比，可以看到标注 PW、IPH 与人工标注的结果大致相当，但是自动标注容易将 PPH 与 PW 混淆，混淆并不一定代表错误，事实上不同人在标注 PPH 边界时也可能会存在分歧。在自动标注结果与人工标注结果不一致的情

况下，将较高层级韵律结构判断为较低层级的情形是要优于将底层韵律结构判定为高层级的情形。观察混淆矩阵对角线的上半部分，本文的方法相比 CRF 方法，在与人工标注不一致的样本里，本文的标注方法会较少地将低层级的韵律结构预测为高层级韵律结构。比如，本文的方法将人工标注的 PPH 判定为 IPH 的样本数是 98，而 CRF 将人工标注的 PPH 判定为 IPH 的样本数是 129。在语音合成中错误的插入停顿会导致语音合成质量的大幅下降，本文的方法较 CRF 出现错误停顿的次数更少。

3.4.5.2 与其他自动标注方法的结果对比

本文实现了四个韵律结构自动标注的模型作为对比。根据四个评价指标 T-ACC、 PW F1、PPH F1、IPH F1 来对比模型的性能。对比结果如下表所示：

表 3.3 与其他标注方法的对比结果

Model	PW F1	PPH F1	IPH F1	T-ACC
CRF	0.7978	0.6001	0.9525	0.8131
D-BLSTM-CRF	0.8018	0.5984	0.9584	0.8169
D-BGRU-S	0.8026	0.5964	0.9599	0.8199
D-BGRU-CRF⋆	0.7993	0.5965	0.9586	0.8163
D-BGRU-CRF	0.8066	0.6120	0.9646	0.8218

与 CRF 相比，所提出的模型在各指标均取得优于 CRF 基准标注系统的效果，具体地，将总体正确率从 0.8131 提升到 0.8218，PW F1 值从 0.7978 提升到 0.8066，PPH F1 值从 0.6001 提升到 0.6120，IPH F1 值从 0.9525 提升到 0.9646。与 D-BLSTM-CRF 比，可见在混合网络结构中，使用 GRU 单元是要优于使用 LSTM 单元的。与 D-BGRU-S 比，可见所提出的模型使用 CRF 层进行整句解码是要优于使用 Softmax 后取最大概率韵律结构这样的判定方法的，这说明考虑到韵律结构序列中相邻两词的韵律结构标签之间的转移关系，以整句解码韵律结构序列，能有效提升韵律结构自动标注的准确率。与 D-BGRU-CRF⋆相比，所提出的的模型因为加入了词尾音素向量，取得了更好的标注效果，这说明除了词尾声学特征外，词尾音素向量联合词尾声学特征可以作为提升韵律结构自动标注的效果，因为词尾的声学特征也与词尾音素密切相关。

3.5 本章小结

本章描述了一种基于 DNN-BGRU-CRF 的混合网络结构实现韵律结构自动标注的方法。该方法联合文本与声学特征作为输入，探究了在神经网络的模型下，词尾音素嵌入、词尾声学统计特征等特征用于自动标注。实验结果显示，与人工标注相比，能得到大体一致的标注结果，与其他自动方法 CRF 相比，取得更优的标注效果，并验证了 DNN-BGRU-CRF 中 GRU 单元、CRF 层的采用的合理有效性。实验结果表明使用 GRU 效果优于采用 LSTM 单元，输出层采用 CRF 优于 Softmax 层。这样的混合结构综合了 DNN、BGRU-RNN、CRF 各种方法的优势，取得较 CRF 更优的效果，使得自动标注结果跟人工标注的结果能达到大部分一致，为快速构建带韵律结构标注的语料库提供了一个有效的方案。

第4章　基于深度自注意力网络的韵律结构预测

4.1 本章引论

韵律结构预测是在语音合成阶段，根据文本特征标定韵律层级结构的任务，它与合成语音的自然度密切相关。不少语音合成系统的停顿位置的确定直接依赖于韵律结构预测的结果。

在韵律结构预测任务上，研究人员有采用条件随机场模型 CRF 进行预测[11]，采用词嵌入、字嵌入等特征基于 RNN 来进行韵律结构的预测，但 CRF 与 RNN 存在难以学习到整句范围上下文关系的问题，并且 CRF 需要一定的专家知识来进行特征模板的设计。

在本章，采用基于深度自注意力网络的方法。该方法能够学习句中词语之间任意距离依赖关系。比如一个句子长度为 n，若要计算句末词与句首词间的依赖关系，RNN 需进行$n-1$次循环计算得到固定维度的向量，经过多次运算得到的该向量是无法完全保留句首词的特征信息的。此向量与句末词特征作为 RNN 最后一个时间步的输入计算才有可能学到两词的依赖关系，其计算复杂度为$O(n)$。自注意力机制，对两词距离不敏感，例如可以直接对句首词特征、句末词特征两者进行计算，计算复杂度为$O(1)$。自注意力机制还有着更好的捕捉全句各词间的上下文依赖关系的能力。实验结果证明，该方法有利于提升预测的准确率。

4.2 问题定义

给定待合成语音的文本，将句子文本经过分词后得到语法词序列（w_1，w_2，...，w_i，...，w_n）。其中w_i为第 i 个语法词，然后提取各语法词的文本特征，再根据句中各语法词的文本特征，预测各词的韵律边界类型（y_1，y_2，...，y_i，...，y_n），其中 y_i 为 w_i 对应的韵律层级结构类型，包括非韵律边界（NB）、韵律词（PW）、韵律短语（PPH）和语调短语边界（IPH）。

4.3 基于深度自注意力网络的韵律结构预测模型

本节先介绍构成深度自注意力网络用的子层结构如自注意力子层、非线性子层的概念。再介绍连接子层输入与输出的残差连接，用以对序列中位置信息进行位置编码的技术，对数据集标签进行预处理的标签平滑技术，最后则描述由基本的子层构成的深度神经网络的整体结构。

4.3.1 自注意力子层

注意力机制可以看成是根据一个查询（query）和一系列的键（key）值（value）对得到该查询的一个表示。其步骤可描述为，将该查询与每个键进行相似度计算得到一系列的权重，然后对相应的值进行权重求和得到该查询的表示。其中，相似度计算函数有多种，加性注意力和点积注意力为常用的计算相似度的方法。放缩点积注意力（Scaled dot-product attention）即为一种点积注意力[46]，其计算流程如图所示。其中 Q 为查询序列，K 为一系列的键，V 为键所对应的值。

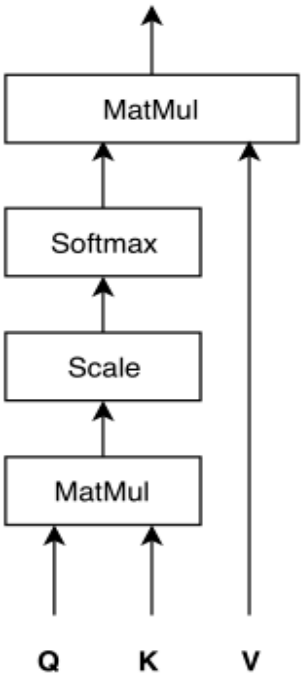


图 4.1 放缩点积注意力的计算流程

首先 Q 与 K 进行矩阵乘法运算，然后通过一个放缩因子进行放缩变换，接着通过 softmax 进行规一化操作，最后通过与 V 进行矩阵乘法得到输出。

用公式可表示为如下形式：

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4-1}$$

其中 d 为放缩因子。在本文的模型中，指的是 Q 中向量的维度。

自注意力机制[47]只需要一个序列就可以计算这个序列的表示，也即其中 Q、K、V 取值相同，均为该序列。多头注意力则是通过对查询、键、值进行 h 个线性变换，再并行地进行放缩点积，每一个放缩点积都会得到一个d_v维的表示，通过将

h 个d_v维的值拼接，形成 $h * d_v$的向量，得到一个输出。多头自注意计算流程，如图 4.2 所示。

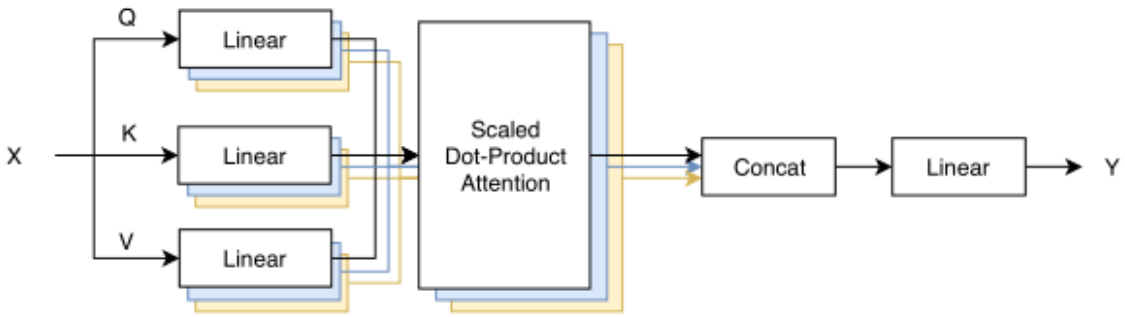


图 4.2 多头自注意计算流程

其计算公式可以用如下等式表示：

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{4-2}$$

$$MultiHead(Q,K,V) = Concat(head_1, \ldots\ldots, head_h)W \tag{4-3}$$

其中，$W_i^Q \in R^{\,d\times d_k}, W_i^K \in R^{d\times d_k}, W_i^V \in R^{d\times d_v}$分别是查询、键、值的线性变换矩阵，$W \in R^{hd_v\times d}$为拼接各放缩点积输出值后所做的最后一次线性变换矩阵。在本文的任务中，多头的数目设置为$h = 8$。对于每一个头，参数设定为$d = 256$，$d_k = d_v = \frac{d}{h} = 32$。

自注意力机制被成功应用于很多自然语言处理任务，例如自然语言推断[48]、神经机器翻译[47]和序列标注[49]等。本文同样对自注意力机制进行应用探索，将自注意力机制应用于韵律结构预测任务。

自注意力子层在本文的模型中，主要用于捕捉全句各词上下文间的依赖关系，形成富含上下文信息的词级特征表示。较高层级韵律结构有可能依赖于相距较远的词，运用自注意力机制主要为了捕捉相距较远词间的依赖关系，从而有助于提升预测效果。如图 3.3 所示，假设一个句子有 n 个词，那么计算句中最后一个词的特征表示是基于该句中所有词的输入特征，通过求他们的相似度得到对应各词的权重值，再以权重求和的形式得到该词的特征表示。自注意力机制直接计算就能直接捕捉两者的依赖关系，计算复杂度为$O(1)$。RNN 计算句首位置与句尾位置的两词依赖关系，需要通过循环神经网络进行 n 次循环计算，其计算复杂度为$O(n)$。自注意力机制能直接建立两词间的联系，它对于两词间的距离不敏感，相比于 RNN 更有利于学习相距较远的词间的依赖关系，而韵律结构预测中的语调短语边界的预

测往往依赖于相距较远的上一个语调短语边界。RNN 仅靠上一词的输出信息和当前词的输入信息来计算当前词的输出，这样的计算方式难以学习到句子的结构信息，而自注意力机制直接依赖于全句范围内各词的输入特征，其计算方式更有利学习到句子的结构信息。

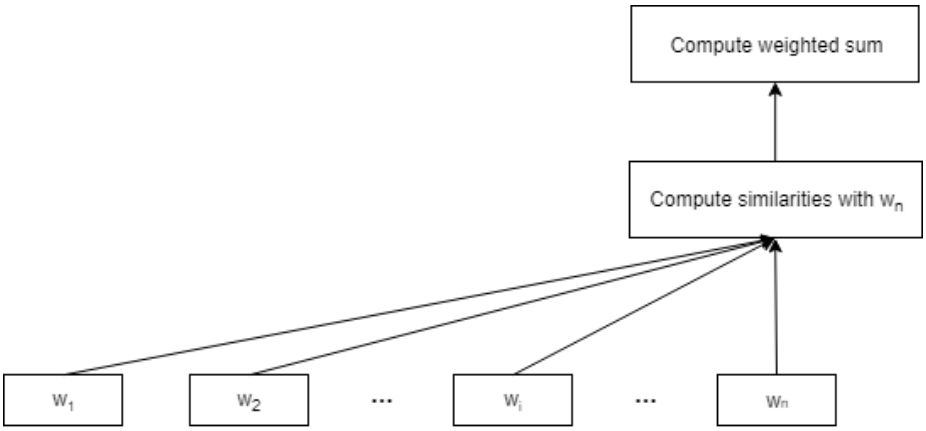


图 4.3 句尾词特征经过自注意力机制后输出的计算流程图

4.3.2 实验设置及对比实验

4.3.2.1 模型初始化及参数设置

对非线性子层、自注意力子层的参数进行正交初始化，词嵌入层随机初始化并且设定设定为可训练的。词嵌入层输出的词向量维度设定为 200 维，隐层神经元数目设置为 256。模型采用 Adam 优化方法[54]，初始学习率设定为 0.0001。Dropout 层[55]设置的保存率为 0.8，每个子层前都采用了 Dropout 层。采用标签平滑使得模型学到一些不确定性，防止过拟合，有助于提升模型泛化能力。对数据集中标签进行预处理的平滑值设定为 0.1。

4.3.2.2 对比实验

为了衡量所提出的模型的性能，实现并比较了如下模型：

CRF，特征上采用词面、词性、词长、词后标点的文本特征，实现了一个基准系统。

BLSTM-EMB，输入特征采用词向量，模型结构采用两层 BLSTM，每层神经元数目设定为 256，最后一层 softmax 输出韵律结构类型的概率分布。

BGRU-RICH，输入特征采用词向量、词性、词长、词后标点，一层神经元数目为 256 全连接层，四层双向带门控循环单位的 RNN 层（BGRU-RNN），最后一

层 softmax 输出韵律结构类型的概率分布。

SLEF-ATT，即为本文所提出的模型。

4.3.3 非线性子层

4.3.3.1 全连接网络子层

全连接网络子层可以与自注意力网络子层联合使用，用以将输入进行非线性变换。其主要包括两个线性变换层以及在中间的 ReLU 激活函数[50] ，如图 4.4 所示。

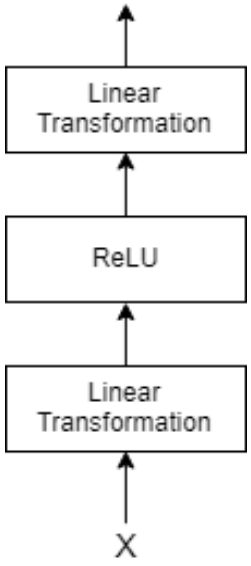


图 4.4 FFN 网络结构图

可以用如下公式表示这个计算流程：

$$FFN(X) = ReLU(XW_1)W_2 \tag{4-4}$$

其中，X为全连接网络子层的输入，$W_1 \in \mathrm{R}^{d\times d}$，$W_2 \in \mathrm{R}^{d\times d}$为全连接网络子层所需要学习的参数。

4.3.3.2 循环神经网络子层

因为韵律层级结构预测，是输入一个词特征序列，输出相应韵律结构序列，是序列到序列间的映射，RNN 适合序列建模。但是当序列较长时，RNN 存在因为梯度爆炸或者梯度消散，导致训练困难问题。带门控机制的 RNN 是一个比较好的解决以上训练问题的方法，目前主要有 LSTM 单元以及其变体 GRU 单元。相比于 LSTM，GRU 有着更为简洁的门结构，且参数少，模型收敛更快，目前被广泛应用

于序列建模。由于 RNN 仅有单向的上下文信息，双向 RNN 则用来获取双向的上下文信息。

LSTM 单元的公式可表示如下：

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{4-5}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{4-6}$$

$$\widehat{C_t} = tanh(W_C[h_{t-1}, x_t] \ + \ b_C) \tag{4-7}$$

$$C_t = f_t \odot \ h_{t-1} + i_t \odot \widehat{C_t} \tag{4-8}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{4-9}$$

$$h_t = o_t \ \odot \tanh C_t \tag{4-10}$$

其中，C_t表示当前 LSTM 单元的记忆内容，f_t为遗忘门，控制遗忘之前时间步信息的程度，i_t和o_t分别为输入门与输出门，用以控制输入和输出信息，⊙表示按元素相乘。

LSTM 单元的变体 GRU，其单元的计算过程在 3.3.2 节有详细介绍，此处不作展开。

尽管带门结构的 RNN（LSTM-RNN、GRU-RNN）能学习到过去时间步的依赖关系，但由于只能学到一个方向上的信息使得它的性能受到了限制。双向 RNN 能够使得网络学到两个方向上的上下文依赖关系，所以双向 RNN 的结构也应用在本文的模型中。

为了分析不同的循环神经网络子层作为非线性子层，对韵律结构预测效果的影响，我们分别尝试了不同循环神经单元的设置（LSTM、GRU）以及不同方向的设置（单向、双向），实现了如下三种韵律结构子层：

1、单向 GRU-RNN 子层。记子层各时间步的输入为（x_1，x_2，...，x_i，...，x_n），循环神经单元采用 GRU，仅学习单向的依赖关系。用以和配置为 BGRU-RNN 子层的预测结果对比，以确定学习双向的依赖关系更有助于提升韵律结构预测效果。其网络结构如图 4.5 所示。

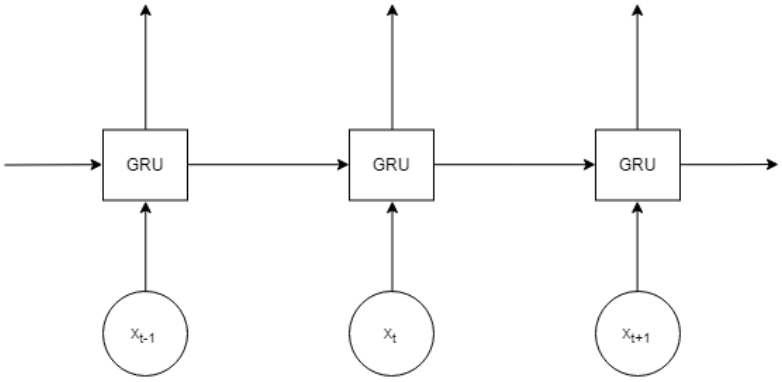


图 4.5 GRU-RNN 结构图

2、双向 LSTM-RNN 子层（BLSTM）。记子层各时间步的输入为（x_1，x_2，...，x_t，...，x_n），循环神经单元采用 LSTM，通过拼接正向 LSTM 的输出与反向 LSTM 的输出，形成子层的输出。用以和配置为 BGRU-RNN 子层对比，选取 LSTM 单位与 GRU 单元中效果更好的计算单元作为循环神经网络的配置。BLSTM-RNN 网络结构如图 4.6 所示：

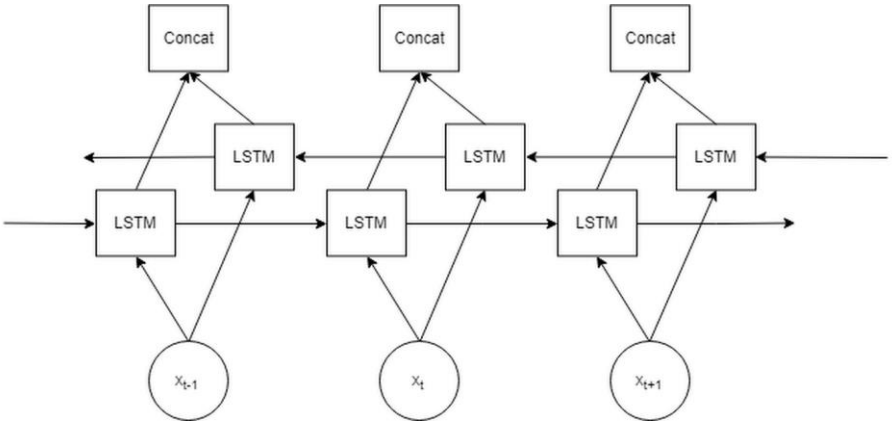


图 4.6 BLSTM-RNN 结构图

3、双向 GRU-RNN 子层（BGRU）。记子层各时间步的输入为（x_1，x_2，...，x_t，...，x_n），循环神经单元采用 GRU。通过拼接正向 GRU 的输出与反向 GRU 的输出，形成子层的输出。其网络结构如图 4.7 所示。

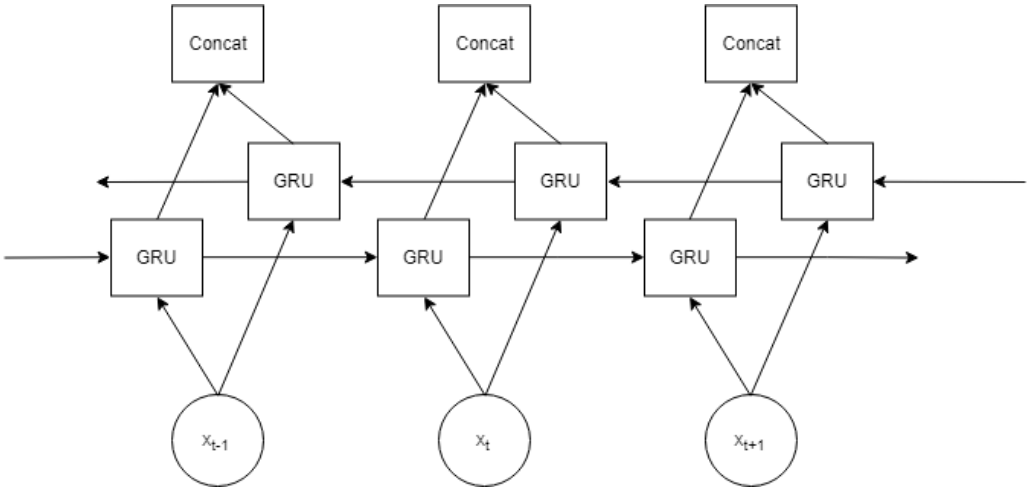


图 4.7 BGRU-RNN 结构图

因为子层的输入与输出存在残差连接，需要保持数据输入维度与输出维度相同，采用单向 GRU-RNN 子层时，设置神经元数目为 256 维。采用 BLSTM 或者 BGRU 时，每个方向设置的神经元数目为 128 维，双向的输出拼接形成 256 维。

4.3.4 残差连接

深度神经网络随着层数的增加，在训练集的准确率会存在饱和甚至下降的现象，这就是神经网络的退化问题。残差连接是一个训练深度神经网络的有效方法，其实现方式是除了主路的层与层之间存在直接连接之外，还有旁路的跳过中间若干层的直接连接，并在连接处对各维度进行加法操作。如图 4.8 所示：

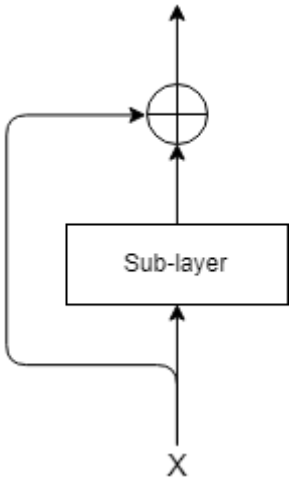


图 4.8 残差连接结构图

将其用在所提出的网络模型各子层中，可以用如下公式表示：

$$Y = X + SubLayer(X) \tag{4-11}$$

其中 X，Y 分别是各子层的输入与输出，$SubLayer$表示非线性子层或者自注意力子层对输入特征的变换。在残差连接后，有个层规范化[52]操作，主要用以控制子层输出数据的分布。

该方法最早用于解决图像识别的深度神经网络随着层数的加深，出现难以优化的问题[51]。随着层数的增加，识别错误率不降低，反而升高。简单堆叠的神经网络由于学习的是恒等映射，那么堆叠的深度模型会退化成浅层模型。通过拟合残差，模型学习的变成了学习一个残差函数，较学习恒等映射更为简单，残差连接就是一个恒等映射，不引入额外的参数，可以直接使用反向传播训练。

本文所提出的的模型需要堆叠多个相同的层，随着堆叠的层数增多，模型的深度增加，也会带来难以训练的问题，通过采用残差连接，有助于模型的训练，也利于尝试更深的网络结构配置。

4.3.5 位置编码

尽管自注意力机制能够学习到句中词语任意距离的依赖关系，但是词语间的相对位置距离却因为注意力机制而忽略了。为了能够让自注意力网络能够利用相对位置信息，采用了时间信号（timing signal）[47]的机制来利用公式生成位置编码，不用学习任何参数。可以使用如下公式对位置信息进行编码：

$$timing(t, 2i) = sin(\frac{t}{10000^{\frac{2i}{d}}}) \tag{4-12}$$

$$timing(t, 2i+1) = cos(\frac{t}{10000^{\frac{2i}{d}}}) \tag{4-13}$$

其中 t 为时间步的索引值，$2i$ 及 $2i$+1 为编码的维度索引，d 为位置编码的维度。

4.3.6 标签平滑

标签平滑[53]是机器学习在学习分类问题时，对标签进行预处理一种技术。它能让模型学到一定的不确定性，但是有助于提升模型的泛化能力，从而提升模型预测的效果。其具体实施方式如下，本文所描述的韵律结构预测任务是一个四分类问题，预测的四类为：非韵律结构边界（NB）、韵律词边界（PW）、韵律短语边界（PPH）、语调短语边界（IPH）。假如一个语法词的标签为语调短语边界（IPH），那么其标签用独热编码（one-hot encoding）表示如下：

$$TAG_{IPH} = (0, 0, 0, 1) \tag{4-14}$$

使用标签平滑，相当于加入了噪声，引入一定程度的不确定性。假设平滑值设定为 0.1，那么平滑后的标签向量可表示为：

$$SMOOTH_{IPH} = (0.03, 0.03, 0.03, 0.9) \tag{4-15}$$

在模型训练之前，本文对所有的训练数据的标签均进行了标签平滑的处理方式，所设定的平滑值为 0.1，平滑后的标签作为训练集的标签。

4.3.7 深度自注意力神经网络模型结构

深度自注意力网络结构如图 3.9 所示，图示的输入是每个时间步的输入，一个时间步输入一个词的词级特征。

将语法词序列表示为：（w_1，w_2，...，w_i，...，w_n），其中w_i为句中第 i 个词的 ID，输入到词嵌入层得到得到该词的词向量e_i，词嵌入层采用随机初始化，其中参数配置为可训练，在网络训练过程中，跟随整个网络更新其参数，这样的词向量往往是跟当前目标任务密切相关，不必在大语料库上进行预训练。

由于自注意力机制忽略了各词间相对的位置信息。所以，在词嵌入层输出词向量后，与位置编码求和，从而输出带位置编码信息的特征。除了词向量，另外一系列的文本特征也为韵律结构预测提供了丰富的特征，文本特征集合r_i由第 i 个词的词性、词长、词后标点类型独热编码拼接构成。

网络前端的全连接层用以将特征进行混合以学到特征的高层表达，随后采用 N 个相同的块堆叠形成深度网络，这个 N 可以根据需要配置不同的值。网络中每个相同的块均由一个非线性子层、一个自注意力子层构成。其中，非线性子层可以有多种灵活的配置，如配置成全连接网络子层、GRU-RNN 子层、BLSTM-RNN 子层以及 BGRU-RNN 子层。这些配置可以满足不同的需求，如对训练、预测速度的需求以及对预测准确率的需求。

子层的输入、输出采用残差连接的结构，这样可以使得深度自注意力网络可以通过设置更大的 N，尝试较深的网络结构。在残差连接后，还有规范化层，用于控制子层输出数据的分布。

最后一层采用 softmax 层用以输出当前输入词对应的韵律结构类型的概率分布。采用分布中最大概率所对应的韵律结构类型作为当前词的韵律结构预测结果。

深度自注意力网络在韵律结构预测任务上引入自注意力机制，使得该网络能捕捉到全句范围内任意两次间的依赖关系。并且网络具备一定的灵活性，可以配置不同堆叠块的数目 N 以及非线性子层。

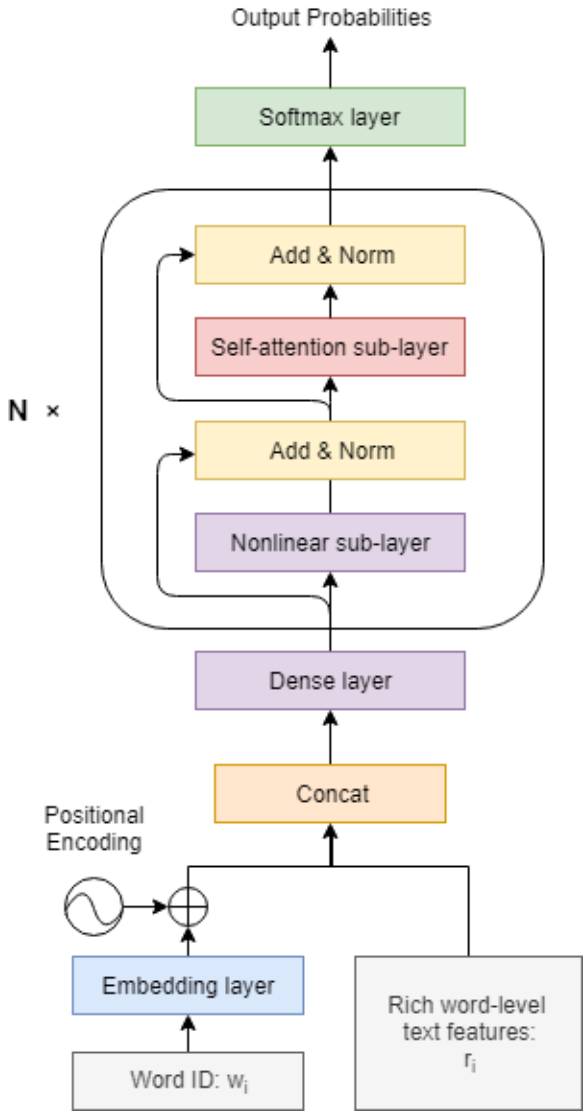


图 4.9 深度自注意力网络模型结构图

4.4 实验结果与分析

4.4.1 数据说明

数据集一共有 41483 句文本，句子都由专业标注人员标注了韵律层级结构。从中随机选定 37483 句作为训练集，另外 2000 句作为验证集，剩下的 2000 句作为测试集。

4.4.2 特征提取

在深度神经网络中，表示词语可以采用一个低维度的词向量，本实验采用了一个可训练的词嵌入层用以学习 200 维的词向量。词性与韵律结构存在一定的相关

性；一个词长较长的语法词往往也是一个韵律词；词后的标点往往也意味着存在韵律结构边界，如词后存在逗号或者句号，那么该词后有较大可能是语调短语边界。

综上，采用了如下特征：

1、词 ID，该词在词汇表的位置索引。

2、词性，表示语法词的语法属性，如形容词、副词、名词等。

3、词长，表示语法词中字的个数。

4、词后标点符号类型，表示语法词后的标点类型。

词 ID 表示该词在词汇表中的索引值。根据数据集中的统计显示，词性有 45 种，因此将词性用 45 维独热编码表示。最长的词长为 10，词长则采用 10 维独热向量表示。词后标点有 14 种，加上无标点的类型，共计 15 种，将词后标点信息表示为 15 维的独热编码。

4.4.3 实验评估标准

以测试集里人工标定韵律结构为真实值，韵律结构预测模型的结果为预测值，可以得到一个混淆矩阵。为了和其他韵律结构预测方法进行比较，采用总体正确率 T-ACC（预测正确的样本数与总体样本数目的比值）、韵律词 F1 值、韵律短语 F1 值、语调短语 F1 值这四个指标来对比各个模型，其中以 T-ACC 为主要评估指标。这些指标的计算方法与 3.4.4 介绍的一致，此处不再赘述。

4.4.4 实验结果及其分析

4.4.4.1 模型配置参数的实验结果

本文实验模型结构配置主要有两个参数，一是模型的非线性子层，二是模型堆叠相同神经网络块（每块含一个非线性子层与自注意力子层）的数目 N。不同的非线性子层设置以及不同的 N 设置，均能影响到模型的效果。本文尝试了不同的参数设置，以确定最佳的参数配置。

针对非线性子层，本文实现了四种不同的非线性子层：全连接子层（FFN）、单向带门控循环单元的循环神经网络（GRU-RNN），双向长短时记忆网络（BLSTM），双向带门控循环神经单元的循环神经网络（BGRU-RNN）。评估指标采用 T-ACC，实验结果如表 4.1 所示，可以看到非线性子层设置为 BGRU 效果最佳，总体正确率达到 0.7991。设置为全连接子层 FFN 虽然总体准确率略低于 RNN 子层，但因为其能方便并行实现，在训练速度上有相应的优势，如果对模型训练、预测的速度有需求，可以采用设置 FFN 非线性子层的模型。

表 4.1 不同非线性子层的实验结果

非线性子层	FFN	GRU	BLSTM	BGRU
T-ACC	0.7829	0.7941	0.7970	0.7991

针对模型堆叠相同块的数目 N，分别实验了四个不同的设置值。实验结果如表 3.2 所示。从下表可以看到，N 设置为 4 的效果最佳。

表 4.2 不同堆叠块数目的实验结果

N	2	3	4	5
T-ACC	0.7961	0.7959	0.7991	0.7984

4.4.4.2 混淆矩阵

以专业标注人员标注的韵律结构为真实值，所提出的模型预测的结果可以采用混淆矩阵来衡量。表 3.3 与表 3.4，分别是 CRF 与本文所提模型预测结果的混淆矩阵。与 CRF 模型相比，本文所提出的的模型在 NP，PW，IPH 这三类上取得更好的预测效果，但是在 PPH 这类边界的预测上效果不佳。本文所提出的模型容易将 PPH 与 PW 混淆，可能存在的原因是在数据集中相对于 IPH、PW 的样本数，PPH 的样本数目最少，从而导致模型未能充分学习到该类区别与其他类的特征。总体来说，和 CRF 模型相比，本文所提出的的模型在绝大多数韵律结构边界处取得更优的效果。

表 4.3 CRF 模型预测结果的混淆矩阵

Predict / Actual	NB	PW	PPH	IPH
NB	4142 (83.39%)	702 (14.13%)	123 (2.46%)	0 (0.00%)
PW	526 (7.41%)	5714 (80.49%)	843 (11.87%)	16 (0.23%)
PPH	122 (4.07%)	1240 (41.32%)	1553 (51.75%)	86 (2.87%)
IPH	17 (0.46%)	111 (3.02%)	147 (4.00%)	3397 (92.51%)

表 4.4 深度自注意力网络模型预测结果的混淆矩阵

Actual \ Predict	NB	PW	PPH	IPH
NB	4224 (85.04%)	676 (13.61%)	64 (1.29%)	3 (0.06%)
PW	473 (6.66%)	5866 (82.63%)	747 (10.52%)	13 (0.18%)
PPH	97 (3.23%)	1341 (44.69%)	1483 (49.42%)	80 (2.67%)
IPH	18 (0.49%)	101 (2.75%)	151 (4.11%)	3402 (92.65%)

4.4.4.3 对比实验结果

四个模型的对比实验结果如表 4.5 所示。与 CRF 相比，本文所提出的的模型将 T-ACC 从 0.7901 提升到了 0.7991。与 BLSTM-EMB 相比，所提出的的模型有着显著的提高，这同时也说明，仅仅采用词向量作为特征划分三种韵律层级结构是不够的。BGRU-RICH，采用了更为丰富的特征集合包括：词向量、词性、词长、词后标点，可以看到取得了不错的效果。比较 BGRU-RICH 与 BLSTM-EMB，用了更深的网络模型，更丰富的输入特征后，总体正确率从 0.7320 提升到了 0.7912，IPH F1 值以及 PPH F1 值均得到了显著提升，分别从 0.8151 提升至 0.9493、从 0.4564 提升至 0.5366，PW F1 值略有提升，从 0.7301 提升至 0.7657。

与 BGRU-RICH 相比，本文所提出的模型 SELF-ATT，T-ACC 有进一步的提高，因为自注意力机制能够捕捉句子中任意距离的词特征之间的相互依赖关系。具体地，所提出的模型将 T-ACC 从 0.7912 提升至 0.7991，将 PPH F1 值从 0.5366 提升至 0.5446，将 PW F1 值从 0.7657 提升到 0.7778，IPH F1 未见显著差别。

通过比较实验，可以得到，本文采用的深度自注意力网络 SELF-ATT 在韵律结构预测任务上不仅优于条件随机场模型 CRF，也优于仅使用带门控单元的循环神经网络 BGRU-RNN。本文所提出的的模型采用自注意力网络与 BGRU-RNN 网络的堆叠，联合了两者的优势。自注意力机制有。实验结果显示，该方法在所有对比的模型中取得最优的效果。

表 4.5 对比实验的结果

Model	PW F1	PPH F1	IPH F1	T-ACC
CRF	0.7687	0.5481	0.9474	0.7901
BLSTM-EMB	0.7301	0.4564	0.8151	0.7320
BGRU-RICH	0.7657	0.5366	0.9493	0.7912
SELF-ATT	0.7778	0.5446	0.9490	0.7991

4.5 本章小结

本章描述了一种基于自注意力机制神经网络的韵律结构预测方法，该方法利用自注意力机制从而能够学到到句中任意距离的词之间的依赖关系，并且根据全句范围内的词特征计算当前词的特征表达，能捕捉句子的结构信息。实验结果显示，该方法能取得优于 CRF 模型、双向带门控循环单元的循环神经网络（BGRU-RNN）的效果。与 CRF 相比，不必依赖专家知识预先进行特征模板设计，能够学到长时依赖。与 BGRU-RNN 相比，能够对任意距离的词语直接计算依赖关系以及捕捉全句范围的内的依赖关系。本方法是自注意力机制在韵律结构预测上的应用，联合了 BGRU-RNN 与自注意机制，能更好地捕捉上下文依赖关系，从而提升韵律结构预测的效果。

第5章　BERT 词向量及其在韵律结构预测中的应用

5.1 本章引论

韵律结构预测是在语音合成阶段，根据文本特征标定韵律层级结构。一种广泛应用的方法是采用 RNN 来进行建模，输入特征采用字向量或者词向量的方式[24][56][57]，来进行韵律结构预测。但是，词向量存在多种训练形式，一种是随机初始化，利用跟随训练数据更新的词嵌入层得到的词向量，一种是利用大语料库预先训练好的词向量，并且预先训练采用的模型有多种，不同的模型得到的词向量也存在差异。研究不同词向量对韵律结构预测的影响显得十分必要。最近的研究成果显示，使用 BERT 模型[58]在大语料库上预训练的词向量对于自然语言处理的下游任务效果显著。

在本章，采用 BERT 模型训练得到的词向量来进行韵律结构预测，并分析了不同词向量在神经网络模型中对韵律结构预测任务的影响。

5.2 问题定义

对句子进行分词得到句中的词序列（w_1，w_2，…，w_i，…，w_n），其中 W 为一个句子的语法词序列，w_i为第 i 个语法词，再经过文本分析得到w_i的文本特征如词向量、词性、词长、词后标点，在这里将它们分别表示为e_i、s_i、l_i、p_i。通过各语法词的文本特征预测各词的韵律边界类型（y_1，y_2，…，y_i，…，y_n），其中 y_i 为 w_i 对应的韵律层级结构类型，包括非韵律结构边界（NB）、韵律词（PW）、韵律短语（PPH）和语调短语边界（IPH）。

其中词向量特征的获取方式，如下三种：

1、使用随机初始化且可训练的嵌入层，嵌入层的参数跟随训练数据更新。将词 ID 输入到嵌入层，输出词向量。此方法的好处是少了预训练词向量模型的步骤。

2、使用 Skip-Gram 模型[59]在维基百科、百度百科大语料库上进行预训练得到词向量。

3、使用 BERT 模型在大语料库上预训练得到词向量。

保持网络主体结构不变，以输入词向量特征e_i为变量，分别将以上三种词向量分别作为模型的输入词向量e_i，得到实验结果用以分析 BERT 词向量对韵律结构预测任务的影响。

5.3 BERT 向量在韵律结构预测中的应用

5.3.1 Transformer 结构

Transformer[47]是一个运用了多头自注意力机制的神经网络结构，也是 BERT 模型的基础组成部分，其主要结构如图 4.1 所示：

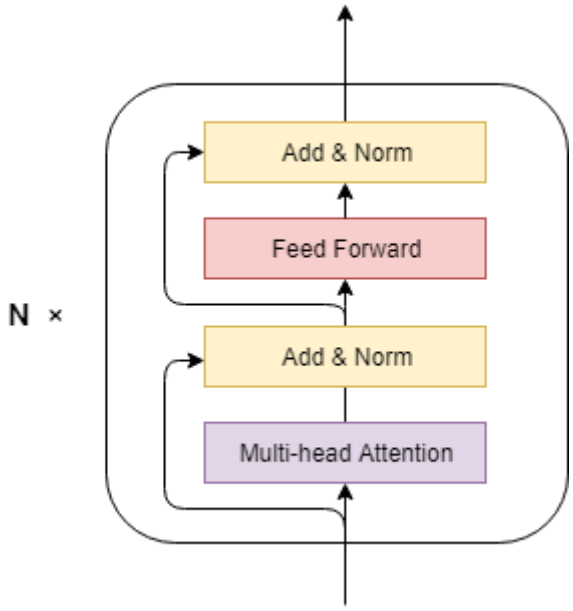


图 5.1 Transformer 结构图

Transformer 可以看作是一个块，每块又含有多个层，首先输入进入一个多头自注意层，其次用一个残差连接该层的输入与输出，再对残差结果进行规范化，输入到下一个前馈神经网络，该前馈神经网络通常是由两个全连接层和中间一个非线性激活函数 ReLU 构成。前馈神经网络的输入与输出也有一个残差连接，残差连接后同样也会进行一个层规范化操作，最后输出结果。在使用 Transformer 时，通常采用多个 Transformer 块堆叠而成深度网络，图中 N 即表示堆叠 N 个 Transformer 块的网络。

Transformer 因为多头自注意机制以及前馈神经网络均能进行并行计算，从而有着较好的并行性，相比 RNN 需要依赖前一时刻的输入，其训练速度上存在优势。联合位置编码，Transformer 结构适用于对序列进行建模，且能学习到序列中任意距离的依赖关系。最近，不少研究人员将其应用于序列建模的任务中，例如机器翻译[60]等任务。

5.3.2 BERT 模型结构

BERT 模型是一个基于 Transformer 的用以学习语言模型的深度神经网络模型。其主要结构如图 5.2 所示：

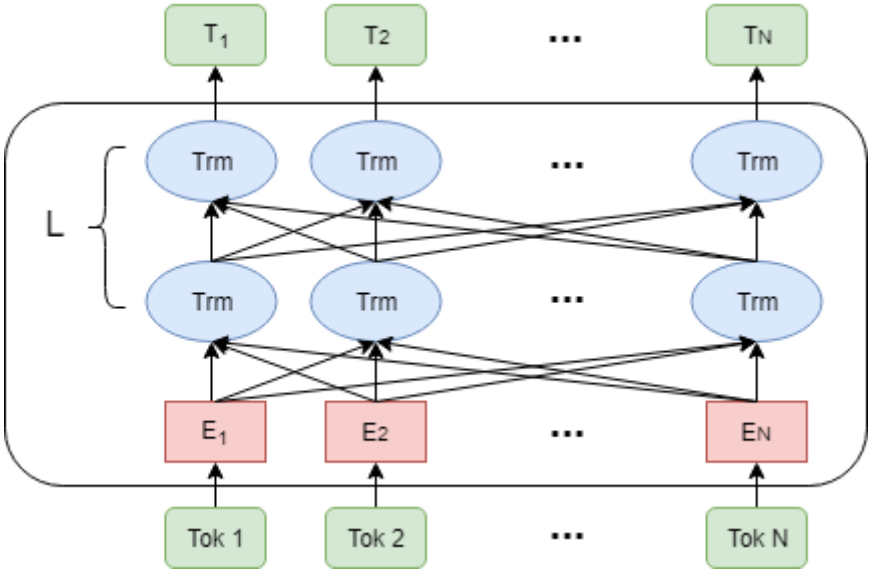


图 5.2 BERT 模型结构图

用 L 表示 Transformer 的块数，H 表示隐层节点数，A 表示注意力的头数。BERT 根据这几种参数的不同配置，可分为 BERT 基础版与加强版。本文使用 BERT 基础版的配置：L=12，H=768，A=12。

BERT 模型在预训练中有两个目标：

1、采用遮蔽语言模型 MLM[61]。随机选中句子中的 15%的词语按概率遮蔽，预测的时候，仅预测被遮蔽的词语，而不是预测整句话，近似于完形填空。这样的学习目标，能使得模型学习到一个词语的上下文表示。但遮蔽语言模型也存在相应的缺点，遮蔽的词汇引入了[MASK]这样的特殊字符，这在下游任务中进行微调的时候，句中其实是没有[MASK]这样的字符的。为了缓解这样的问题，遮蔽语言模型选择了以下按概率处理的方式：80%的情况下将选中词语替换为[MASK]。10%的情况下，将其替换为随机词语。10%的情况下，对词语保持原状，不作任何处理。由于预测的是随机掩蔽句中词语，使得模型不得不保持学习各词语的上下文。尽管存在随机替换词语的情况，但总体比率不高，不会影响到模型的语言理解能力。另外，遮蔽语言模型由于只预测 15%的词语，会导致收敛速度较预测整句词语的模型更慢。

2、下一句预测，是一个二分类任务，有助于学习到句子级的特征表达。在数据集中选择句子对构成样本（A，B）：其中 50%的情况 B 是 A 的下一句，可在数据集中选择段落中的连续两句来形成这样的句子对；另外 50%的情况，则 B 不是 A 的下一句。在 B 不是 A 的下一句的情况下，B 句子是随机在语料库中选定其他句子与 A 拼接成句子对。B 是 A 的下一句的正样本例子如下：

A：他今天上午八点起床了。B：他就去参加晨练。

负样本的例子如下：

A：他今天上午八点起床了。B：他刚刚踢完球回来。

训练时，同时优化以上两个目标，最终得到模型的训练参数，可以用来获取 BERT 词向量。其好处在于，通过遮蔽语言模型，能够获取到双向的依赖关系，并且基于自注意力机制，学习到长时依赖。另外，通过下一句预测，能学习到句子与句子间的关系，得到句子级的信息表达。通过 BERT 模型结构，在大语料库上预训练后，就能根据所提供的词语获取到该词的向量表示。类似于迁移学习的方法，在大语料库上训练得到预训练好的模型，在下游任务中通过监督式学习进行模型参数的微调[62][63]。最近的研究成果显示，BERT 词向量在下游的众多自然语言处理任务上获得最佳性能。

本文任务的输入特征也包括词向量。预测韵律结构序列也需要利用上下文信息。BERT 模型预训练的词向量富含上下文信息，将其引入到韵律结构预测任务，以分析其对该任务的影响。

5.3.3 韵律结构预测模型

本文的预测模型主要由 DNN 与 BGRU-RNN 两部分叠加而成，如图 5.3 所示，输入特征有两部分，一部分是词向量，一部分是词性、词长、词后标点拼接的文本特征向量。模型网络每时间步读取一个词向量和文本特征向量，文本特征向量采用三层全连接层的 DNN 用以学习高层特征表达，DNN 的输出与词向量进行拼接形成 BGRU-RNN 的输入，再通过一层全连接层与 softmax 层预测出当前时间步的词语对应的韵律结构类型的概率分布，取对应最大概率的韵律结构类型作为预测值。

图中所示的词向量为 BERT 模型预训练的词向量。通过固定模型结构，主要变换输入特征中的词向量，以分析不同的词向量对于韵律结构预测任务的影响，并确定 BERT 词向量在韵律结构预测任务上的效果是否优于传统方法训练得到的词向量。

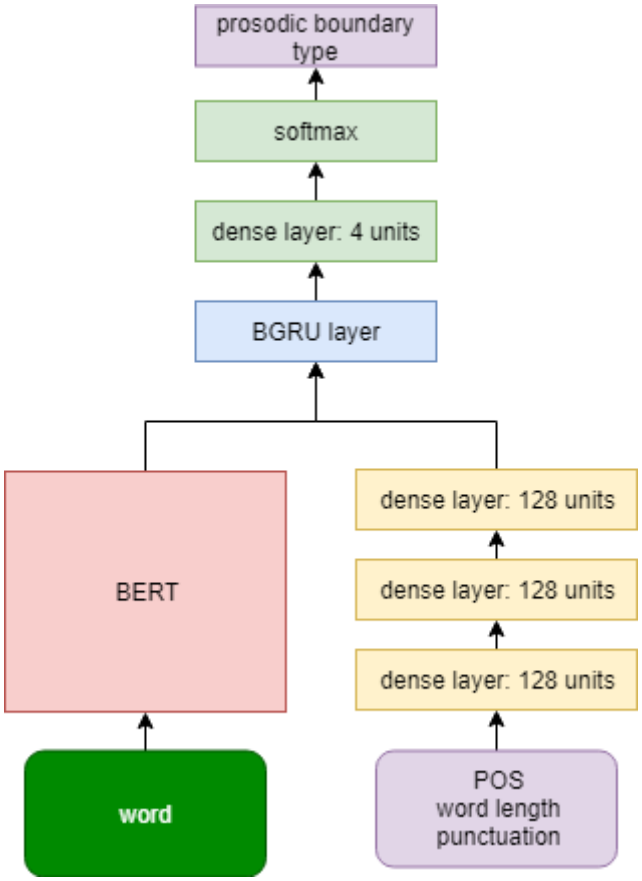


图 5.3 韵律结构预测模型结构图

5.4 实验结果与分析

5.4.1 数据说明

数据集一共有 41483 句文本，句子都由专业标注人员标注了韵律层级结构。从中随机选定 37483 句作为训练集，另外 2000 句作为验证集，剩下的 2000 句作为测试集。

5.4.2 特征提取

针对词级特征输入，采用了词性、词长、词后标点特征以及词向量特征。在分词与词性标注后，就能得到词性、词长的信息，词后标点则通过简单的文本分析就能得到。

在深度神经网络中，表示词语可以采用一个低维度的词向量，本实验采用了一个可训练的词嵌入层用以学习 200 维的词向量。词性与韵律结构存在一定得相关

性，一个词长较长的语法词往往也单独构成一个韵律词，词后的标点往往也意味着存在韵律结构边界，如词后存在逗号或者句号，那么该词后有较大可能是语调短语边界。

综上，采用了如下特征：

1、词 ID，该词在词汇表的索引。

2、词性，表示语法词的语法属性，如形容词、副词、名词等。

3、词长，表示语法词中字的个数。

4、词后标点符号类型，表示语法词后的标点类型。

词 ID 表示该词在词汇表中的索引值。根据数据集中的统计显示，词性有 45 种，将词性用 45 维独热编码表示。最长的词长为 10，词长则采用 10 维独热编码表示。词后标点有 14 种，加上无标点的类型，共计 15 种，将词后标点用 15 维独热编码表示。

通过词面来获取词向量，词向量获取方式主要是以下三种：

1、随机初始化、可训练的词嵌入层

2、用 Skip-Gram 算法训练的词向量

3、用 BERT 模型训练的词向量

下面分别介绍这三种词向量的获取方式：

可训练词嵌入层：通过随机初始化、可训练的词嵌入层获取词向量的方式可以用图 5.4 表示。

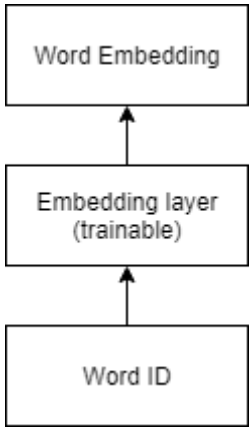


图 5.4 可训练词嵌入层结构图

Skip-Gram 训练得到词向量：Skip-Gram 是训练词向量的一种常用模型，不同

的词可能存在相同的上下文，那么他们具有一定的相关性，词向量能让词义相关的词语在向量空间中距离相近。Skip-Gram 通过一个滑动窗口的机制，假设滑动窗口大小设置为 2，则选取当前词作为模型的输入词，选取当前词的前边两词、后边两词作为模型的输出词，输入词与输出词之间形成词对，作为模型的训练数据。以滑动窗口内的词作为样本的标签，通过当前词和滑动窗口内的各词组合，形成以词对为表示形式的训练样本。

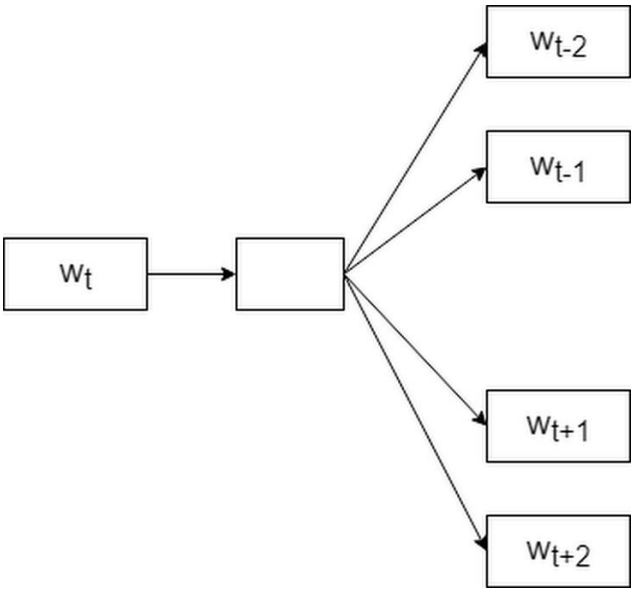


图 5.5 Skip-Gram 算法模型结构图

BERT 模型训练得到词向量：BERT 模型是一个能学习到双向上下文的、使用自注意力机制的模型结构，主要通过预训练来得到一个词的向量表示，再将该向量用于下游任务。这个向量表示，较 Skip-Gram 词向量更富含上下文信息，由于 Skip-Gram 需要指定滑动窗口的大小，其能捕捉的上下文信息限制在窗口内，而 BERT 模型因为自注意力机制能获取全句的上下文信息，其在众多自然语言处理任务上获得当前最优的效果，比如在问答、自然语言推断等任务。

5.4.3 实验设置及对比实验

本文所提出的模型的全连接层以及 BGRU-RNN 层参数数据化采用正交初始化，每个全连接层采用 128 个神经元，全连接层后接 dropout 层，dropout 层的保持概率设置为 0.8。

作为对比实验所采用的模型结构基本保持不变，以输入特征中的词向量为变量，做了三个实验：分别以随机初始化、可训练的词嵌入层得到的词向量的模型表示为 Random；以 Skip-Gram 算法训练的词向量简记为 Skip-Gram；以 BERT 模型

训练的词向量简记为 BERT。

5.4.4 实验评估标准

以测试集里人工标定韵律结构为真实值，韵律结构预测模型的结果为预测值，可以得到混淆矩阵来评估模型的预测效果。

为了和其他韵律结构预测方法进行比较，本文采用总体正确率 T-ACC（预测正确的样本数与总体样本数目的比值）、韵律词 F1 值、韵律短语 F1 值、语调短语 F1 值这四个指标来对比各个模型，其中以 T-ACC 为主要评估指标。这些指标的计算方法与 3.4.4 介绍的一致，此处不再赘述。

5.4.5 实验结果及其分析

5.4.5.1 混淆矩阵

以测试集中人工标注的韵律结构为真实值，韵律结构预测模块预测结果为预测值，可以采用混淆矩阵的形式来评估模型性能。表 5.1、表 5.2、表 5.3 分别为使用随机词向量、Skip-Gram 预训练的词向量、BERT 预训练的词向量作为输入特征韵律结构预测实验结果的混淆矩阵。

对比表 5.1 与表 5.2，可知使用预训练的词向量在预测正确的 PW 样本数、PPH 样本数指标上有显著提升，预测正确的 IPH 样本数无明显变化。对比表 5.1 与表 5.3，使用预训练的 BERT 词向量预测正确的 PW 样本数指标上略有下降，预测正确的 PPH 样本数指标上显著提升，预测正确的 IPH 样本数指标上无显著变化。

表 5.1 使用 Random 词向量的韵律结构预测模型

Actual \ Predict	NB	PW	PPH	IPH
NB	4252 (85.60%)	654 (13.17%)	60 (1.21%)	1 (0.02%)
PW	536 (7.55%)	5858 (82.52%)	696 (9.8%)	9 (0.13%)
PPH	116 (3.87%)	1536 (51.18%)	1272 (42.39%)	77 (2.57%)
IPH	23 (0.63%)	127 (3.46%)	119 (3.24%)	3403 (92.67%)

表 5.2 使用 Skip-Gram 词向量的韵律结构预测模型

Actual \ Predict	NB	PW	PPH	IPH
NB	4306 (86.69%)	609 (12.26%)	52 (1.05%)	0 (0.00%)
PW	476 (6.71%)	6024 (84.86%)	592 (8.34%)	7 (0.10%)
PPH	84 (2.80%)	1417 (47.22%)	1419 (47.28%)	81 (2.70%)
IPH	9 (0.25%)	106 (2.89%)	154 (4.19%)	3403 (92.67%)

表 5.3 使用 BERT 词向量的韵律结构预测模型

Actual \ Predict	NB	PW	PPH	IPH
NB	4296 (86.49%)	588 (11.84%)	82 (1.65%)	1 (0.02%)
PW	487 (6.86%)	5831 (82.14%)	775 (10.92%)	6 (0.08%)
PPH	76 (2.53%)	1224 (40.79%)	1619 (53.95%)	82 (2.73%)
IPH	17 (0.46%)	81 (2.21%)	170 (4.63%)	3404 (92.70%)

5.4.5.2 对比实验结果

以四项评估指标衡量模型的性能：总体正确率、预测语调短语的 F1 值、预测韵律短语的 F1 值、预测韵律词的 F1 值。分别以 Random、Skip-Gram、BERT 分别表示使用随机初始化向量、Skip-Gram 向量、BERT 向量的韵律结构预测模型。从表 5.4 中可以看出，使用随机初始化的词向量的结果略低于基准模型 CRF。当使用

Skip-Gram 预训练的词向量后，能看到显著的提升，总体正确率优于基准模型 CRF。使用 BERT 预训练的词向量也可以看到总体正确率的提高，同样优于基准模型 CRF。对比 Skip-Gram 与 BERT，可以发现两者的总体正确率无明显差异，IPH F1 值也没有差异，但是在 PPH F1 上，使用 BERT 词向量能显著提升 PPH 的预测效果，一个重要的原因在于相比于使用 Skip-Gram 词向量，BERT 词向量富含上下文信息，这些上下文信息对于预测 PPH 有一定的帮助，但是在 PW 上的预测效果略有下降。如果在 T-ACC 指标无显著差异性的情况下，依次从 IPH F1、PPH F1、PW F1 来评估模型，因为这种自顶向下的方法，符合高层级韵律结构的预测性能对语音合成的性能影响大于低层级韵律结构的预测性能带来的影响，因为后续的停顿插入位置就是依赖高层级的韵律结构预测结果。总结而言，使用 BERT 词向量优于使用 Skip-Gram 向量的预测结果，使用这两个预训练的词向量均显著优于随机初始化嵌入层得到的词向量，也优于基准 CRF 模型。

表 5.4 对比实验结果

Model	PW F1	PPH F1	IPH F1	T-ACC
CRF	0.7687	0.5481	0.9474	0.7901
Random	0.7671	0.4942	0.9503	0.7890
Skip-Gram	0.7898	0.5439	0.9502	0.8085
BERT	0.7868	0.5734	0.9502	0.8084

5.5 本章小结

本章介绍了 BERT 模型训练上下文相关词向量的原理，并首次在韵律层级结构预测任务中应用该向量提升韵律层级结构预测性能。使用了一个深度神经网络与带门控循环单元神经网络的混合结构，以不同的词向量作为输入，衡量词向量类型的不同对于韵律结构预测任务的影响。通过实验显示，采用在大语料库中预训练的词向量有助于提升韵律结构预测效果，并且 BERT 预训练的词向量相较于 Skip-Gram 预训练的词向量，在韵律结构预测的总体正确率上大致相当，但 BERT 词向量的优势在于在较高层级的韵律结构的预测效果上显著优于 Skip-Gram 预训练的

词向量。该方法对韵律层级结构选定词向量作为深度神经网络的特征有一定的参考价值，并且也证实了 BERT 词向量富含上下文信息的特性，对预测韵律结构的效果有一定的提升。

第6章　总结与展望

6.1 研究工作总结

本文主要研究了语音合成系统里关于韵律结构的两大任务。一是在音库准备阶段，为了快速构建语料库，需要自动标注韵律结构的任务。另一个是语音合成阶段，在前端文本处理流程中，对句子进行分词以及词性标注后，需要预测韵律结构类型的任务。

为了快速构建音库，本文提出了基于 DNN-BGRU-CRF 的韵律结构自动标注模型，使用深度神经网络，联合文本、声学特征对韵律结构进行自动标注任务。所提出的 DNN-BGRU-CRF 网络模型综合了各类神经网络的优势，DNN 用于学习高层的特征表达和特征融合、BGRU-RNN 适用于学习上下文依赖信息，CRF 适用于解码时考虑全句整个标注序列，进行整句解码。取得了较 CRF 自动标注更高的标注准确率。自动标注结果与人工标注结果能达到大体的一致性。在不一致的情况下，本文的标注结果将低层级韵律结构标注为高层级韵律结构的样本数显著少于 CRF，这说明本文的标注方法优于 CRF，因为较少的错误标注为高层级韵律结构的样本有助于得到更精确韵律结构预测模型，从而错误插入的停顿更少。这对快速构建带精确标注的语音合成海量语料库有着重要意义。

为了能在合成阶段更精确地预测韵律结构，本文提出了一个基于深度自注意力网络的预测模型，用以学习句中词语任意距离之间的依赖关系。该方法与 CRF 相比，不必需要专家知识人工确定特征模板，不受马尔科夫假设限制导致难以捕捉长时依赖关系；与 RNN 相比，能利用全句范围上任意距离的两词间的依赖关系，自注意机制还能捕捉句子结构信息。实验结果显示，本文提出的基于深度自注意力网络的韵律结构预测模型取得了较传统 CRF、RNN 更高的预测准确率。

为了研究不同词向量对于韵律结构预测任务的影响，本文分析了三种词向量在韵律结构预测任务上的应用。词向量包含一定的语义信息，这些信息对于韵律结构预测是有帮助的。目前缺少相应的研究工作探讨不同方法训练的词向量在深度神经网络方法中，对韵律结构预测的影响。本文尝试了三种词向量，包括随机确定、跟随训练数据更新的词向量，用 Skip-Gram 算法在大语料库上预训练的词向量，以及采用 BERT 模型在大语料库上预训练的词向量。通过实验测定不同词向量对于韵律结构预测任务的效果，对于将来以词向量作为输入、以深度神经网络方法建模的韵律结构预测任务，在选取词向量的方案时，有一定的参考意义。

6.2 未来研究展望

对于韵律结构自动标注任务，除了利用文本、音频两类特征外，可以考虑利用传感器捕捉说话人下巴的运动数据。研究人员报道了不同韵律结构边界下巴运动存在特别之处。另外，如果有视频数据的话，可以捕捉头部动作、手部动作，找到与韵律结构密切相关的联系。比如在演讲的情况下，头部视线转移的时候，有较大可能是出在韵律结构边界的情况。不同人也在韵律结构边界也可能存在一些习惯性的动作，比如有部分人演讲会有手部动作的突然停顿。这些动作均与个人特性密切相关，所以仅限于自动标注单人数据的韵律层级结构。这些带个人特性的动作规律与韵律结构边界的联系均值得探索。

韵律结构标注的过程中，标注人员会遇到在少部分样本上难以确定韵律结构的情况，但能确定大致的分类，比如可以确定一个语法词后边存在 B 集合里的韵律结构类型，但无法确定是 B 集合的具体一种韵律结构类型，假定集合$A = \{NB, PW\}$，集合$B = \{PPH, IPH\}$，举例说明，标注人员认为 50%的可能标注韵律短语边界（PPH），50%的可能标注语调短语边界（IPH），标注人员能确定标注类型不属于集合 A 中的韵律结构类型。标注人员能确定大致分类，更精细的分类无法确定，这是在实际标注过程中存在的，而当前的标注的方法都让标注人员必须确定一个具体分类，不同标注人员的不一致性问题的产生，部分原因也在于此。为了模拟标注人员不确定的情况，做自动标注的时候设计一些弱分类器，如该分类器只能将韵律结构类型确定在所有韵律结构类型构成的集合的子集中，这样可以允许模糊分类，这是与实际情况相符的。并且，根据以往机器学习算法的经验来看，若干个弱分类器构成的强分类器，比单纯的多分类器拥有更好的性能。针对韵律层级结构自动标注研究模糊分类或是弱分类器是一个值得深入探索的领域。

对于韵律结构预测任务，因为语音合成后续的停顿插入的位置往往依赖于这个韵律结构预测的结果。在预测错误的情况下，高层级韵律结构预测成低层级的情况要比低层级预测成高层级的情况更让人接受，因为后者出现频次较多的话，会导致停顿插入较多，极大地降低语音合成的体验效果。设计一个损失函数控制错误的倾向，向低层级韵律结构方向标注，或者将韵律结构分类问题看作是一个带偏好的分类问题，均是一个值得深入探索的地方。

把韵律结构预测看做是一个扁平分类问题，是一个相对简化的处理，不同韵律层级结构仅仅是不同的分类，这与实际情况不符合。不同韵律结构之间存在着层级的关系，一个句子按韵律结构划分成树状的层级结构，这种层级关系，并没有得到有效的建模。如果能设计一种神经网络结构能够学习到层级的这种依赖关系，也就

更符合实际问题的建模，这也是一个值得探索的地方。层级结构的 RNN 可能是一种解决方案，有研究人员曾提出过一种层级的 LSTM-RNN 来进行多任务学习的网络结构[64]，且表明该方式能学习到层级间关系。其网络结构如图 6.1 所示，（x_1，x_2，x_3）为输入词序列，（y_1，y_2，y_3）为任务一的标签序列，（z_1，z_2，z_3）为任务二的标签序列，这样的结构能使得模型能够捕捉任务一与任务二之间的层级依赖关系。这种层级结构能很好地与韵律结构层级相对应，因为韵律层级结构也存在层级关系。把不同韵律结构的分类问题，看成是多个任务，每个任务分别预测一种韵律结构层级，这样形成一个多任务问题，再通过这样的层级 LSTM-RNN 结构进行建模，这样的建模方式更符合实际情况。尽管这种建模方法更符合韵律结构间的实际关系，但有效性仍需实验数据进行评估。此外，自注意力机制因为计算当前词与句中所有词的相似度，这种机制能捕捉到句子的结构信息，将层级 LSTM-RNN 网络结构与自注意力机制进行融合，能否更好地捕捉不同韵律结构间的层级依赖关系，有待深入研究。总体来讲，将韵律结构预测看成扁平分类是目前普遍的处理方式，能设计一种学习层级依赖关系的网络能很好地推进韵律结构预测技术发展，对有层级依赖关系的的输入序列的类似问题也能起到推动作用。

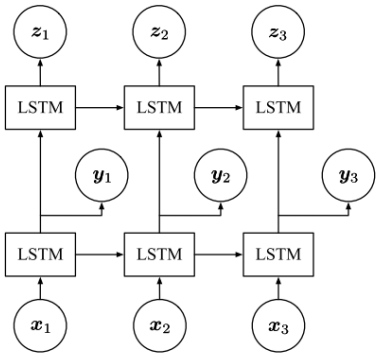


图 6.1 层级 LSTM-RNN 结构图[64]

韵律结构预测当前是依赖于大数据驱动的技术，而训练数据往往采用单说话人的数据。个性化语音合成是当前语音合成的一个热点研究领域。不同人的韵律结构也存在着细微不同之处。研究多说话人的韵律结构的不同之处，也有助于个性化语音合成的效果。另外，儿童与成人说话的韵律结构以及停顿方式也有着明显不同之处，现在针对孩童的语音合成的需求也越来越多，比如儿童陪聊机器人等应用。研究儿童的韵律结构不同的地方，分析各年龄段人们的韵律的变化，既有一定的科学研究价值，也有现实应用价值。

语音合成系统中，关于停顿位置的确定方式，并不存在统一的标准。有的合成

系统将以韵律结构预测出较高韵律结构边界，比如韵律短语边界、语调短语边界，再在这些边界处插入停顿；有的专门训练一个停顿模型，用停顿模型的预测结果，确定停顿位置。在专门训练停顿模型的合成系统里，停顿模型的训练与韵律结构预测模型的训练是分开进行的，然而两个任务之间实际上存在着联系，如果以多任务的形式，其中一任务预测韵律结构，另一个任务进行停顿位置预测，这样能充分利用两者的关联性，但又不是在韵律结构预测结果的基础上以简单的规则来预测停顿位置。

参考文献

[1] Kawahara H. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds [J]. Acoustical science and technology, 2006, 27(6): 349-353.

[2] Morise M, Yokomori F, Ozawa K. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications [J]. IEICE TRANSACTIONS on Information and Systems, 2016, 99(7): 1877-1884.

[3] Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K. WaveNet: A generative model for raw audio [J]. SSW, 2016.

[4] Ma J, Zhang W, Shi Q, Zhu W, Shen L. Automatic prosody labeling using both text and acoustic information [C]. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[5] Wightman C W. Automatic labeling of prosodic patterns [J]. IEEE Transactions on speech and audio processing, 1994: 469-481.

[6] Vaissière J. Language-independent prosodic features [J]. Prosody: Models and measurements, 1983, 61(2): 53-66.

[7] Forney G D. The Viterbi algorithm [J]. Proceedings of the IEEE, 1973, 61(3): 268-278.

[8] Ananthakrishnan S, Narayanan, S S. An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model [C]. International Conference on Acoustics, Speech, and Signal Processing, 2005, 1: I-269.

[9] Yang C, Ling Z, Lu H, W Guo, Dai L. Automatic phrase boundary labeling for Mandarin TTS corpus using context-dependent HMM [C]. International Symposium on Chinese Spoken Language Processing (ISCSLP), 2010: 374-377.

[10] Gallwitz F, Niemann H, Nöth E, Warnke V. Integrated recognition of words and prosodic phrase boundaries [J]. Speech Communication, 2002, 36(1-2): 81-95.

[11] Qian Y, Wu Z, Ma X, Soong F. Automatic prosody prediction and detection with Conditional Random Field (CRF) models [C]. International Symposium on Chinese Spoken Language Processing (ISCSLP), 2010: 135-138.

[12] Levow G A. Automatic prosodic labeling with conditional random fields and rich acoustic features [C]. Proceedings of the Third International Joint Conference on Natural Language Processing, 2008.

[13] Wang X, Xie L, Ma B, Chng E S, Li H. Modeling broadcast news prosody using conditional random fields for story segmentation [C]. APSIPA ASC, 2010: 253-256.

[14] Rosenberg A, Fernandez R, Ramabhadran B. Modeling phrasing and prominence using deep recurrent learning [C]. In Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[15] Tao J. Acoustic and Linguistic Information Based Chinese Prosodic Boundary Labelling [J]. Lecture Notes in Computer Science, 2004: 489-496.

[16] Wang M Q, Hirschberg J. Predicting intonational boundaries automatically from text: the ATIS domain [C]. Speech and Natural Language, Proceedings of a Workshop, 1991.

[17] Huang Y, Wu Z, Li R, Meng H, Cai L. (). Multi-Task Learning for Prosodic Structure Generation Using BLSTM RNN with Structured Output Layer [C]. INTERSPEECH, 2017: 779-783.

[18] Shen X, Xu B. A CART-based hierarchical stochastic model for prosodic phrasing in Chinese [C]. Proc. of ISCSLP, 2000: 105-108.

[19] Taylor P, Black A W. Assigning phrase breaks from part-of-speech sequences [J]. Computer Speech & Language, 1998, 12(2): 99-117.

[20] Yu Y, Li D, Wu X. Prosodic modeling with rich syntactic context in HMM-based Mandarin speech synthesis [C]. IEEE China Summit and International Conference on Signal and Information Processing, 2013: 132-136.

[21] Kang H, Liu W. Prosodic words prediction from lexicon words with CRF and TBL joint method [C]. International Symposium on Chinese Spoken Language Processing. 2006: 161-168.

[22] Chiang C, Wang Y, Chen S. Punctuation generation inspired linguistic features for mandarin prosodic boundary prediction [C]. International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012: 4597-4600.

[23] Kudo T. CRF++: Yet another CRF toolkit [OL]. 2005. http://crfpp. sourceforge. net/.

[24] Vadapalli A, Gangashetty S V. An investigation of recurrent neural network architectures using word embeddings for phrase break prediction [C]. Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH), 2016: 2308-2312.

[25] Shao Y Q, Han J Q, Liu T, Zhao Y Z. Prosodic word boundaries prediction for Mandarin text-to-speech [C]. In International Symposium on Tonal Aspects of Languages: with Emphasis on Tone Languages, 2004.

[26] Ding C ,Xie L, Yan J, Zhang W, Liu Y. Automatic Prosody Prediction for Chinese Speech Synthesis using BLSTM-RNN and Embedding Features [J]. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015: 98-102.

[27] Sun J, Ynag J, Zhang J, Yan Y. Chinese Prosody Structure Prediction Based on Conditional Random Fields [C]. International Conference on Natural Computation, 2009, 3: 602-606.

[28] Hochreiter S and Schmidhuber J. Long short-term memory [J]. Neural computation, 1997: 1735-1780.

[29] Cho K, Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. arXiv preprint arXiv:1406.1078, 2014.

[30] Schuster M, Paliwal K K. Bidirectional recurrent neural networks [J]. IEEE Transactions on Signal Processing, 1997, 45(11):2673-2681.

[31] Christopher Olah. Understanding LSTM Networks [OL]. 2005. http://colah.github.io/posts/2015-08-Understanding-LSTMs.

[32] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C]. International Conference on Machine Learning (ICML), 2001.

[33] Cheng H, Griss M. Learning and Recognizing the Hierarchical and Sequential Structure of Human Activities [D]. Research Showcase, CMU, 2013.

[34] Li F, Zhang M, Tian B, Chen B, Fu G, Ji D. Recognizing irregular entities in biomedical text via deep neural networks [J]. Pattern Recognition Letters, 2017: S0167865517302155.

[35] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging [J]. arXiv preprint arXiv:1508.01991, 2015.

[36] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space [J]. arXiv preprint arXiv:1301.3781, 2013.

[37] Read I, Cox S. Using part-of-speech for predicting phrase breaks [C]. International Conference on Spoken Language Processing, 2004.

[38] Taylor P, Black A W. Assigning phrase breaks from part-of-speech sequences [J]. Computer Speech & Language, 1998, 12(2): 99-117.

[39] An RNN-based algorithm to detect prosodic phrase for Chinese TTS [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing. 2001, 2:809-812.

[40] Sanders E, Taylor P. Using statistical models to predict phrase boundaries for speech synthesis [C]. In Proceedings of Eurospeech, 1995, 2: 1811–1814.

[41] Vaissière J. Rhythm, accentuation and final lengthening in French. Music, language, speech and brain. 1991: 108-120.

[42] Liu Y, Li A. Cues of prosodic boundaries in Chinese spontaneous speech [C]. Proc. ICPHS2003, 2003.

[43] Wightman C W, Ostendorf M. Automatic labeling of prosodic patterns [J]. IEEE Transactions on speech and audio processing 2.4 (1994): 469-481.

[44] Yang Y, Wang B. Acoustic correlates of hierarchical prosodic boundary in Mandarin [C]. Speech Prosody, 2002.

[45] Lai L, Gooden S. Acoustic cues to prosodic boundaries in Yami: A first look [C]. Proceedings of Speech Prosody, 2016.

[46] Luong M, Pham H, Manning C. Effective approaches to attention-based neural machine translation [J]. arXiv preprint arXiv:1508.04025, 2015.

[47] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L. Attention is all you need [C]. Advances in Neural Information Processing Systems (NIPS), 2017: 5998-6008.

[48] A decomposable attention model for natural language inference [J]. arXiv preprint arXiv:1606.01933, 2016

[49] Tan Z, Wang M, Xie J, Chen Y, Shi X. Deep semantic role labeling with self-attention [C]. AAAI Conference on Artificial Intelligence, 2018.

[50] Nair V, Hinton G. Rectified linear units improve restricted boltzmann machines [C]. Proceedings of the 27th international conference on machine learning (ICML), 2010.

[51] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition [C]. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2016: 770-778.

[52] Ba L J, Kiros J R, Hinton G E. Layer normalization [J]. arXiv preprint arXiv:1607.06450, 2016.

[53] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision [C]. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2016.

[54] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv preprint arXiv:1412.6980, 2014.

[55] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting [J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[56] Huang Y, Wu Z, Li R, Meng H, Cai L. Multi-Task Learning for Prosodic Structure Generation using BLSTM RNN with Structured Output Layer [C]. Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH), 2017.

[57] Zheng Y, Li Y, Wen Z, Ding X, Tao J. Improving Prosodic Boundaries Prediction for Mandarin Speech Synthesis by Using Enhanced Embedding Feature and Model Fusion Approaches [C]. INTERSPEECH, 2016.

[58] Devlin J, Chang M W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. arXiv preprint arXiv:1810.04805, 2018.

[59] Mikolov T, Sutskever I, Chen K. Distributed Representations of Words and Phrases and their Compositionality [C]. Advances in Neural Information Processing Systems (NIPS), 2013: 3111-3119.

[60] Ahmed K, Keskar N S, Socher R. Weighted transformer network for machine translation [J]. arXiv preprint arXiv:1711.02132, 2017.

[61] Taylor W L. Cloze procedure: A new tool for measuring readability [J]. Journalism Bulletin, 1953, 30(4): 415-433.

[62] Dai A M, Le Q V. Semi-supervised sequence learning [C]. Advances in Neural Information Processing Systems (NIPS), 2015.

[63] Howard J, Ruder S. Universal language model fine-tuning for text classification [J]. arXiv preprint arXiv:1801.06146, 2018.

[64] Zhou Q, Wen L, Wang X, Ma L, Y Wang. A Hierarchical LSTM Model for Joint Tasks [C]. China National Conference on Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, 2016: 324-335.

致　谢

衷心感谢导师吴志勇副研究员三年来的悉心指导。吴老师在课题研究上给予的帮助与支持，让我受益匪浅。

感谢腾讯科技有限公司 AI Lab 语音研究组康世胤提供的研究、实践平台以及在专业实践上给予的指导。

感谢实验室老师们三年来的培养。感谢李旭师兄在其毕业之际、百忙之中抽出时间帮忙解答语音合成的技术细节。感谢宁义双师兄在我所遇到的技术问题上给予的热心指导。感谢黄雨晨在韵律结构相关课题上提供的帮助，让我有机会接触并理解该领域的研究现状，并确定了自己的研究课题进行深入探索。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1990 年 04 月 23 日出生于湖北省襄阳市。

2009 年 9 月考入重庆大学大学软件学院软件工程专业，2013 年 7 月本科毕业并获得工学学士学位。

2013 年 9 月到 2014 年 7 月，就职于北京神州普惠科技股份有限公司。

2015 年 8 月到 2016 年 8 月，就职于北京海鑫科金高科技股份有限公司。

2016 年 9 月考入清华大学计算机科学与技术系攻读计算机技术工程硕士至今。

发表的学术论文

[1] Yao Du, Zhiyong Wu, Dan Su, Dong Yu, Helen Meng. Automatic Prosodic Structure Labeling using DNN-BGRU-CRF Hybrid Neural Network. (INTERSPEECH 2019, submitted)

[2] Yao Du, Zhiyong Wu, Dan Su, Dong Yu, Helen Meng. Prosodic Structure Generation using Deep Self-attention Neural Network. (INTERSPEECH 2019, submitted)

研究成果

[1] 吴志勇，杜耀，康世胤，苏丹，俞栋，蒙美玲. 一种韵律层级自动标注、模型训练的方法及装置: 中国, CN109697973A (中国专利公开号)

[2] 吴志勇，杜耀，康世胤，苏丹，俞栋，蒙美玲. 一种基于自注意力机制的韵律层级结构预测方法 (在申状态)

参与的科研工作

[1] 2017 年 8 月-2018 年 11 月，在腾讯科技有限公司 AI Lab 实习。参与犀牛鸟专

项研究项目(JR201803，JR201942)