基于内容和声学特征层级融合的自动韵律边界标注

傅睿博　陶建华　温正棋

摘　要　韵律边界标注对于语料库建设和语音合成有着至关重要的作用，而自动韵律标注可以克服人工标注中不一致、耗时的缺点。仿照人工标注流程，本文对文本特征和声学特征采用层级融合的方式运用循环神经网络训练自动韵律边界标注模型。本文以词为单位提取了静音时长，与传统以帧为单位的声学特征相比更具有明确的物理意义，与韵律边界的联系更加紧密。实验结果表明，相比于传统声学特征，本文所采用的静音时长特征使自动韵律标注的性能有所提高；相比于直接特征层面融合和决策层融合的方法，特征层级融合方法更好地结合了声学和文本的特征，进一步提高了标注的性能。

关键词　韵律边界标注，特征层级融合，语料库构建，语音合成

Automatic Prosodic Boundaries Labeling Based On Hierarchically Fusing the Context and Acoustic Features

FU Ruibo　TAO Jianhua　WEN Zhengqi

Abstract Automatic prosodic boundaries labeling plays an important role in the construction of speech corpus. Compared with the manual annotation of prosodic boundaries, which is time consuming and inconsistent, the automatic annotation of prosodic boundaries can overcome the above drawbacks. In this paper, inspired by the manual labeling procedure, a method that hierarchically fuses the context features and acoustic features by using LSTMs is proposed to imitate the human annotators. To handle the discrepancy of the basic unit of text and audio, a subsystem is designed to detect the duration of silence. Experiments show that the effectiveness of the duration of silence being extracted as the acoustic features and the hierarchically fused method in improving the F1-score of the boundaries prediction compared with previous features fusion strategies.

Key words Prosodic boundaries labeling, Hierarchically fusion, Corpus construction, Speech synthesis

1. 引言

语料库建设，特别是语音合成音库的构建，在语音相关技术中占有重要的地位。在目前主流的语音合成方法中，参数语音合成方法和波形拼接合成方法都需要精细的语料库标注工作。这些标注工作将直接影响到最后合成语音的音质、表现力等。

语音合成中，语料库标注主要包括音段标注和韵律标注：音段标注针对的是音素序列标注对应的起始和结束时间；韵律标注是对和语言相关的韵律信息进行标注，对于中文合成音库，韵律信息的标注主要是指对韵律层级进行标注。考虑到目前音段的自动标注技术已经相对成熟，基本可以满足目前系统构建的需求，本文将以语料库建设中韵律标注为研究的入手点。韵律信息在语音合成中被用于模型的上下文文本信息，其精度直接影响到语音合成的质量，该工作通常需要专业标注人员进行标注。然而，随着目前语料库的加大，人工进行韵律边界的标注比较耗时，而且人工标注存在一定主观性，不同人乃至同一人在不同状况下的标注结果都会存在不一致性，通常需要多人重复标注并采用投票的方式来保证一致性。因此，如何精确自动地对语料库进行韵律边界标注已经成为目前一个亟需解决的问题。

汉语的韵律边界通常被分为三类：韵律词、韵律短语和语调短语［1］。在已有的韵律边界标注研究中，可以大致分为三类：第一类采用文本特征进行韵律边界自动标注，该方法包括采用分类回归树［2］、条件随机场［3］、深度回归学习［4］、基于记忆学习［5］等方式，主要以词性、字词的位置、数量等信息为特征进行分类。该类方法主要依靠自然语言处理技术，适用于仅有文本的语料库标注。对于相同的文本，有可能存在不同的表达方式，其所对应的韵律的标注也不唯一，若存在对应的音频语料，该方法不能保证给出最切合音频发音节奏的标注。第二类采用声学特征进行韵律边界自动标注，如Wightman等人提取了每个音节的时长、基频、能量相关特征，利用决策树和隐Markov模型（hidden Markovmodel，HMM）对英文语料库采用ToBI体系进行标注［6］。该类方法需借助语音识别或大量人工音频处理［7］，其准确率很大程度上取决于切分精度，同时不同的音节由于时长不同，声学参数的提取及其归一化也存在一定困难。第三类结合文本和声学特征进行韵律边界自动标注，如Hasegawa-Johnson等人采用（multilayer perceptron，MLP）分类器对基频和时长特征建模，采用支持向量机（support vector，machine，SVM）对文本和句法特征建模［8］［Chen等人采用文本相关的HMM（CD-HMM）和N-gram语言模型联合文本和声学特征对韵律边界进行建模［9］［此类方法可以综合文本和音频两个通道的特征［10］，其难点在于文本和声学特征的提取单元存在不一致性，同时每类特征各自所对应的数据类型不同。前两类研究分别聚焦于文本或音频单个通道，基于文本特征的韵律标注适用于语音合成前端的韵律预测模块；基于声学特征的韵律标注适用于语音识别中的韵律停顿识别。然而对于语音合成语料库构建而言，音频是由专业录音人在录音室录制，比语音识别中的实际应用场景噪声小，同时有较为精准的文本。因此，第三类结合文本和声学参数对韵律边界建模的方法更适用于语料库的构建，同时结合文本和声学特征不仅仅是简单地对前两类方法的综合，文本特征和声学特征存在一定相关性，因此本文的研究核心是如何更好地将文本和声学特征融合并用于对韵律边界的建模。

在研究文本和声学特征融合之前，首先要考虑的是文本和声学特征的选取。随着自然语言技术的发展，以往的研究对韵律边界建模所采用的文本特征已经较为丰富，词向量等被神经网络所采用的特征已经被用于韵律建模研究中［11］，然而声学特征的选取，大部分研究还采用语音识别技术中所采用的声学参数（如基频、能量、谱参数等）［12］，由于语音识别的声学前端首要目的是识别发音的基元，在后端语言模型才会对韵律等较高层面信息进行分析识别，其在识别过程中所采用的声学特征更倾向于刻画发音等浅层次信息，这点与韵律边界的检测标注是不一致的。因此，本文的出发点是挖掘深层次且与韵律标注目的一致的声学特征，并探究所选取的声学特征与文本特征的融合方式。

本文主要探索了将文本特征和声学特征层级融合用于韵律边界标注模型建模。重点研究了静音时长作为深层次抽象特征对韵律边界自动标注精度提升的贡献，并采用无标注数据判别静音模型来提高模型的鲁棒性。

2. 自动韵律边界标注系统

本文构建的模拟人工标注的自动标注系统整体框架如图1所示，分为文本和音频两个通道，在文本分析的基础之上，通过提取音频的声学参数来实现对静音时长特征的提取，采用基于长短时记忆模型（long short-term memory，LSTM）的循环神经网络，运用层级融合的方式将文本特征和声学特征输入预测模型网络，得到最终的韵律边界预测结果。

本文系统构建仿照人工标注的流程，如图2所示。人工韵律标注的方法大致可归纳为在机器自动分词对文本预处理的基

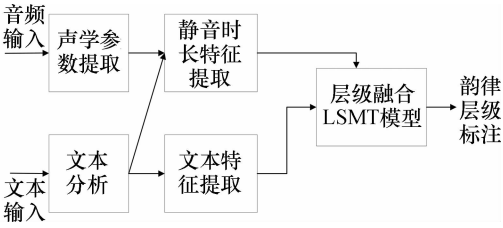

图 1　整体系统框架

础上，标注人员对照文本和音频的频谱图，在听取音频过程中，根据语法词所在的位置结构和实际发音情况微调生成韵律词，如发现频谱中有较大的“空隙”，则参考边界两侧音高和边界调，此外在对应的文本中结合自身经验和频谱“空隙”大小对韵律短语和语调短语进行标注。

在静音时长特征提取过程中，传统方法中采用基频、能量等参数以帧为单位建模，刻画短时发音能力较强，在以字或词为单位的归一化过程中会受噪声和音段切分精度不高的影响，削弱其刻画词间、短语间韵律关系的能力，属于较浅层的信息。本文所采用的静音判别模型，在结合自动音段切分与文本自动语法词分析的基础上，提取静音时长，将浅层以帧为单位的梅尔频率倒谱系数（Mel frequency cepstrum coefficient，MFCC）转化为较深层次以词为单位的静音时长。在时长的提取及归一化过程中，将静音判别模型的输出概率和词的时长结合到对静音时长不同层面的描述过程中，共提取 74 维度的静音时长信息。由于经过人工校对过的音段标注较少，且标注精度和一致性不高，本文采用无标注数据判别静音模型来提高模型的鲁棒性：先用小语料预训练，根据预训练的模型得到大的未标注的语料标签，对该大语料进行再训练。

在韵律标注模型训练过程中，采用层级预测的方式，运用基于 LSTM 的循环神经网络层级融合文本特征和声学特征训练标注模型。由于声学参数静音时长已经过一系列处理，其特征的维度较低，而文本特征采用独热编码（one-hot）归一化的形式较为稀疏。因此，本文采用了将文本特征和声学特征分别在神经网络的不同层级输入来训练韵律边界标注模型，相比直接将文本和声学层面在特征层面上直接融合和在决策层融合的方式，韵律边界标注的效果有所提高。

3. 静音时长特征提取

静音时长特征的提取涉及图 1 的声学参数提取、文本分析和静音时长提取三部分，具体流程如图 3 所示，将提取好的 39 阶 MFCC 和经过预校对的文本使用音素自动切分工具得到音素的时间边界。文本通过语法词分析工具得到以语法词为单位的文本。但由此难以得到每个语法词间的静音段时长，因此在该环节加入一个静音判别模型以得到以帧为单位的静音段位置信息，结合音素时间边界和以语法词为单位的文本，得到最终的语法词间的静音段时长。

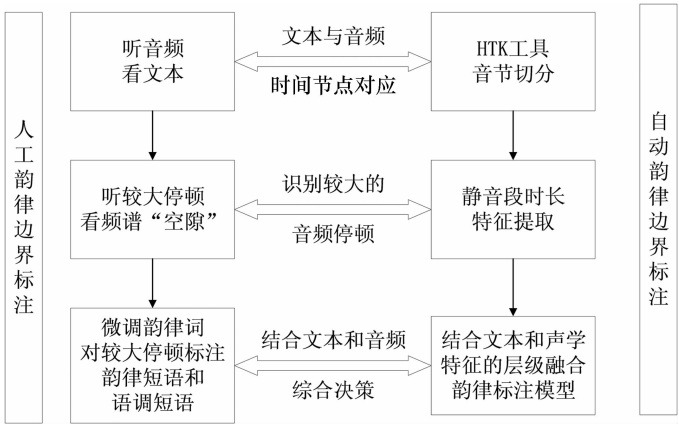

图 2　韵律标注各环节类比

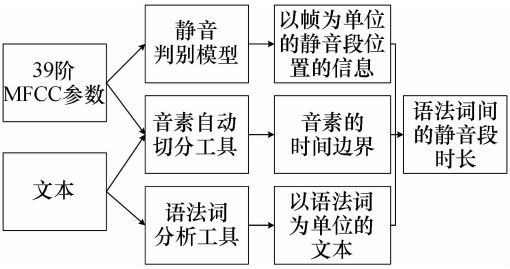

图3　整体静音时长特征提取流程

3.1 静音判别模型

静音判别模型的主要目的是将语料中各字之间的静音识别出来，然而本文面临的一个问题就是精确标注的每个字后带有静音标注的语料较少。同时少量经过人工标注的语料也存在标注不准确的现象，其原因是部分字间的静音段长度较短和人工观察音频频谱存在一定的误差。用该数据训练出来的判别模型极易出现“过拟合”现象。

为了解决有标注数据较少的问题，本文采用了两轮训练的方式来扩充训练数据。第一轮运用已有的少量有标注数据训练一个静音标注模型，用该模型对大量无标注数据进行标注。第二轮训练使用第一轮扩充后的语料进行训练。此方法具有可拓展性，对某个特定需要标注的语料，该方法可以更加有针对性地训练静音判别模型。

3.2 静音时长提取

本文对语法词间静音时长提取如下：

假定一句话由 m 个语法词组成，即：

$$[w_1, w_2, \cdots, w_m]$$

设定第 i 个词：

第一个字的起始时刻为 f_i^S，终止时刻为 f_i^E。

最后一个字的起始时刻为 l_i^S，终止时刻为 l_i^E。

则对于第 i 个词，其静音段搜索区间为：

$$\left[\frac{l_i^S + l_i^E}{2}, \frac{f_{i+1}^S + f_{i+1}^E}{2}\right]$$

设在该区间内共有 t 帧，对于第 j 帧音频，静音段判别模型的输出为 o_j^α，则：

$$o_j^\alpha = \begin{cases} 1, p_j \geq \alpha; \\ 0, p_j < \alpha. \end{cases} \tag{1}$$

其中：p_j 为第 j 帧为静音段的概率，α 为置信度系数，$0 < \alpha < 1$，默认情况为 $\alpha = 0.5$。

则对于第 i 个词与第（$i+1$）个词之间的绝对静音段时长为：

$$H_{ab}^\alpha(i) = \sum_{j=1}^{t} o_j^\alpha. \tag{2}$$

将模型输出概率信息融入时长信息，定义加权绝对时长为：

$$H_{wab}^\alpha(i) = \sum_{j=1}^{t} p_j o_j^\alpha. \tag{3}$$

考虑到长句和短句在实际发音时的静音时长有所差别，以句子为单位进行归一化处理，归一化后的时长为：

$$H_{nab}^\alpha(i) = \frac{H_{ab}^\alpha(i)}{\overline{H_{ab}^\alpha}}, \tag{4}$$

$$H_{nwab}^\alpha(i) = \frac{H_{wab}^\alpha(i)}{\overline{H_{wab}^\alpha}}. \tag{5}$$

其中：$\overline{H_{ab}^\alpha}$ 和 $\overline{H_{wab}^\alpha}$ 分别代表一句话的每个静音段的平均绝对时长和平均加权绝对时长。

考虑到静音时长与当前字整体时长比例，定义相对时长1为：

$$H_{r1ab}^\alpha(i) = \frac{H_{ab}^\alpha(i)}{D_E(i)}, \tag{6}$$

$$H_{r1wab}^\alpha(i) = \frac{H_{wab}^\alpha(i)}{D_E(i)}. \tag{7}$$

考虑到静音时长会受到静音段前后相邻两个字的时长影响，定义相对时长2为：

$$H_{r2ab}^\alpha(i) = \frac{H_{ab}^\alpha(i)}{D_E(i) + D_S(i+1)}, \tag{8}$$

$$H_{r2wab}^\alpha(i) = \frac{H_{wab}^\alpha(i)}{D_E(i) + D_S(i+1)}. \tag{9}$$

其中：$D_E(i)$ 代表第 i 个词最后一个字的时长，$D_S(i+1)$ 代表第（$i+1$）个词开头第一个字的时长。

静音段特征应该对不同韵律层级的标

注更有针对性：在比较严格的判别条件下，只有语调短语等较大的静音段才会被识别，此时的时长特征更易于区分语调短语；在判别条件比较宽松的情况下，词之间的静音时长中更加细微的差异得到体现，更加易于区分韵律词与非韵律词。因此，本文将 α 取从 0.1 到 0.9，以 0.1 为间隔递增，得到 9 组特征。

4. 层级融合模型

我们采用文本特征和声学特征层级融合的 LSTM 模型训练韵律边界预测模型。我们采用有 3 层隐含层的 LSTM 神经网络，首先文本特征先输入含有 1 层隐含层的 LSTM-A，其输出与声学特征结合输入 2 层 LSTM-C，最终输出预测结果。在网络中我们优先输入文本特征的是由于文本特征提取所经过的处理较少，维度较高，而声学特征经过多步骤提取，维度低，已经能较好地反映其物理信息，能较好地反映最终需预测的韵律边界信息。

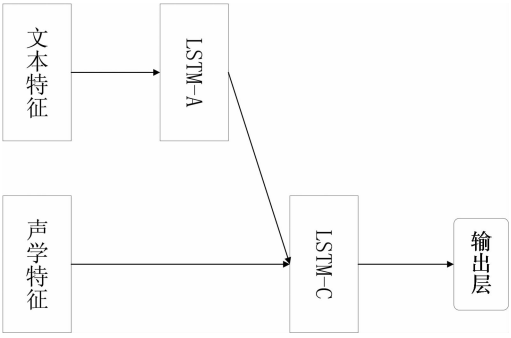

图 4　层级融合网络结构

5. 实验及结果分析

5.1 实验语料

本文采用有专业女性录音人录制的用于语音合成的语料。该语料总共 20000 句，总字数大于 40000，韵律边界（如韵律词、韵律短语、语调短语边界）由两位专业标注人员对照音频和文本标注，有较高的一致性。语料使用语音合成前端语法词和词性分析工具完成，语法词切分正确率为 96.6%，词性分析正确率为 96.4%。语料被分为训练集、验证集和测试集三部分，其比例分别为 80%、10% 和 10%。用于静音段判别模型训练的、人工标注好时间边界的语料占大语料的 10%，共 2000 句。

5.2 静音时长特征提取及对比

5.2.1 系统构建

静音段判别模型依次使用小语料和扩充后的大语料训练，均采用 DNN 进行训练，网络输入均为 39 阶 MFCC，小语料训练网络为 2 层 128 个隐含层节点数，大语料训练网络为 2 层 256 个隐含层节点，输出均含有 2 个 softmax 节点单元。

网络使用 RMSprop 算法［13］训练 10 代，batch-size 为 128，dropout［14］系数为 0.2。

5.2.2 静音切分结果评价

小语料和扩充标注后的大语料训练的静音判别模型准确率分别为 97.4% 和 98.7%，原因是第一轮小语料训练相当于对原始人工标注的结果进行了修正，修正后重新标注的语料更加易于判别。

因此，静音判别模型的准确率不足以说明通过该模型提取的特征对韵律标注的贡献。本文对提取的每一维特征采用一维线性分类器进行训练，计算单个特征进行韵律边界标注的准确率、召回率和 F1 值，将第一和第二轮训练的静音判别模型所提取的参数进行对比实验，本文以韵律短语这一层级为例，实验结果如表 1 所示。

表 1 中所列的数据斜杠右侧为第二轮静音判别模型训练后所提取的静音时长特征各自在一维线性分类器中的评价效果，以 F1 值作为评价指标；斜杠左侧代表第二轮相比第一轮提取该特征后分类器评价指标 F1 的变化值，正数代表提高，负数代表下降。

表1　　各特征在一维线性分类器中评价结果（韵律短语）

α	$H^{\alpha}_{ab}(i)$	$H^{\alpha}_{wab}(i)$	$H^{\alpha}_{r1ab}(i)$	$H^{\alpha}_{r1wab}(i)$	$H^{\alpha}_{r2ab}(i)$	$H^{\alpha}_{r2wab}(i)$	$H^{\alpha}_{nab}(i)$	$H^{\alpha}_{nwab}(i)$
0.1	+0.2/76.1	+0.1/75.9	+0.3/72.3	+0.2/74.1	+0.2/75.0	+0.2/75.6	+0.3/74.5	+0.3/74.4
0.2	+0.3/75.8	+0.3/75.8	+0.4/73.4	+0.1/74.1	+0.1/74.9	+0.1/75.6	+0.2/74.8	+0.2/74.3
0.3	+0.3/76.4	+0.2/76.6	+0.3/73.8	+0.3/74.1	+0.3/75.5	+0.3/75.4	+0.3/74.5	+0.2/74.2
0.4	+0.6/76.2	+0.1/76.1	+0.5/73.6	+0.2/74.1	+0.4/75.2	+0.4/75.2	+0.1/73.9	+0.3/74.4
0.5	+0.4/75.6	+0.0/76.1	+0.3/73.8	+0.3/74.3	+0.2/75.3	+0.2/75.2	+0.3/74.4	+0.2/74.1
0.6	+0.6/75.8	+0.2/76.0	+0.2/74.0	+0.2/74.2	+0.2/75.2	+0.2/75.0	+0.3/74.7	+0.2/74.8
0.7	+0.2/75.8	+0.2/75.8	+0.3/73.9	+0.3/74.1	+0.3/74.9	+0.3/74.9	+0.2/75.8	+0.1/75.6
0.8	+0.3/75.2	-0.1/75.3	+0.3/74.2	+0.2/74.3	+0.3/74.5	+0.4/74.7	+0.3/75.9	+0.3/75.5
0.9	+0.2/75.5	+0.1/75.2	+0.2/74.6	+0.3/74.6	+0.2/75.0	+0.3/74.8	+0.2/75.9	+0.2/75.6
均值	+0.34/75.8	+0.12/75.9	+0.31/73.7	+0.31/74.2	+0.24/75.1	+0.24/75.2	+0.24/74.9	+0.22/74.8

第二轮采用扩充标注数据训练后提高了静音段识别精度，各个静音时长特征单独用于韵律边界检测的F1值有所提高。

通过该实验结果，还可以观察到将提取到的时长进行归一化处理，结合输出概率及考量静音段占发音音节时长的方式，可以更加全面地反映音节间的静音段在全句中声学时长的分布。

5.3 韵律边界标注系统构建及评价

本节所有实验中，韵律词、韵律短语和语调短语边界的标注都采用层级标注的方式，低一级的韵律标注结果会作为高一级韵律标注模型来输入。

为了将文本特征和声学特征更好地结合以提高韵律层级自动标注的精度，本文尝试了两种方式：一是将文本和声学特征直接在特征层面融合，二是分别训练基于文本特征和声学特征的标注模型在决策层面融合。

5.3.1 基于传统声学特征的系统构建

将传统方法中所采用的包括时长（音节时长，静音时长，相邻音节时长比例）、基能量（极值、均值、比例）共20维特征，用基于LSTM的循环神经网络训练基于声学特征的韵律边界标注模型。

5.3.2 基于静音时长声学特征的系统构建

将音节时长和节2.2所介绍的一系列静音时长特征，共74维，用基于LSTM的循环神经网络训练基于声学特征的韵律边界标注模型。

5.3.3 基于文本特征的系统构建

文本特征包括了词性、字和词层面的数量和位置信息、音调，还有根据5度标音法给出的字之间基频的差异。所有文本特征均采用独热编码（one-hot）的方式进行归一化，文本特征共214维。

5.3.4 基于特征直接融合的系统构建

将文本和静音时长声学特征直接组成一个288维的向量作为网络的输入，采用两层基于LSTM的循环神经网络进行训练。

5.3.5 决策融合的系统构建

将基于静音时长声学特征和文本特征分别训练得到模型，用决策层融合策略进行融合。即分别使用文本特征和声学特征训练韵律边界标注模型，在决策层面根据两个模型的输出判别结果概率和模型整体正确率作为决策依据，其决策过程具体为：

$$f(x_a, x_t) = \begin{cases} f_a(x_a), P_a(C) > P_t(C) \\ f_t(x_t), P_t(C) > P_a(C) \end{cases} \quad (10)$$

其中：

$$P_a(C) = P_a(C|f_a(x_a))P_a(f_a(x_a)) \quad (11)$$

$$P_t(C) = P_t(C|f_t(x_t))P_t(f_t(x_t)) \quad (12)$$

$f_a(x_a)$ 和 $f_t(x_t)$ 分别是基于声学和文本特征的分类器，$P_a(C)$ 和 $P_t(C)$ 分别表示分类器模型整体正确率，$P_a(C \mid f_a(x_a))$ 和 $P_t(C \mid f_t(x_t))$ 分别表示分类器的先验概率，即输出判别结果概率。

5.3.5 特征层级融合的系统构建

将静音时长声学特征和文本特征同时送入如第4节所述的层级融合模型。

5.3.6 实验中的超参数设置

本文实验所采用的超参数如表2所示。

表2　　实验超参数设置

系统	隐含层节点数	隐含层层数	学习率	输出层节点类型
传统声学特征	256	1	0.001	softmax
本文声学特征	256	2	0.001	softmax
文本特征	256	2	0.001	softmax
直接特征融合	512	2	0.001	softmax
特征层级融合	256512	3	0.001	softmax

表3　　韵律自动标注评价结果

韵律层级	传统声学特征	本文声学特征	文本特征	直接特征融合	决策融合	特征层级融合
韵律词	83.72	85.98	95.10	95.08	96.35	96.68
韵律短语	72.31	85.64	69.43	87.70	87.85	88.53
语调短语	74.62	84.81	84.24	85.66	85.83	86.64

表3为6个系统的评价效果，本文使用F1值作为对各个韵律层级预测的评价指标。

6. 分析与讨论

相比传统声学特征，本文所采用的静音时长等一系列声学特征对韵律标注的准确度有明显提高。原因是本文所采用的是边界自动切分，其切分精度的误差会传递给后续基频、能量等参数的提取，会给传统方法所用到的声学参数带来较大误差。本文采用静音时长特征的出发点就是考虑到自动音节切分精度不高的问题，因此所提取的静音时长特征受切分不准的影响较小。从另一个角度看，传统声学特征中的所涉及的关于基频的参数已经在文本特征中的声调信息有所体现，传统声学特征更多体现了某个音节的发音状况，而不是音节间的停顿情况。

声学特征对韵律短语这一层级的预测效果提高明显。在人工韵律标注中，这一层级也是最难标注的，有较大的不一致性，文本所采用的静音时长对各音节间的停顿进行了较为细致的度量，弥补了文本所采用的文本特征对句内各音节间停顿关系描述的不足。

对于直接在特征层面对文本和声学特征进行融合的方法，本文发现，就韵律短语和语调短语这两个层级而言，特征融合提高了自动标注的性能。但是对于韵律词这一层级，融合未能提高该层级标注的性能，其原因是静音时长对于韵律短语和语调短语这两个层级的判别区分度较大，而静音时长对韵律词这一层级的区分度较小，只要依靠文本特征就可以达到较好的识别效果。

特征层级融合的方式相比直接特征融合和决策层融合的方式在各个韵律层级的标注效果都有所提高。原因在于采用的直接特征融合方式对于静音时长和文本两类不同类别的特征不能进行有效的区分训练：静音时长特征相比文本特征有着更大的贡献。特征层级融合的网络结构，该网络可以凸显较高韵律层级的标注，从而获得较好的效果。

7. 结论

本文探究了静音时长特征的提取及其

相比传统声学特征在韵律边界自动标注上性能的提升。实验结果表明，在音频采用自动音节切分未经过人工校对的情况下，本文所采用的静音时长特征对韵律边界的检测性能有明显提升。同时，特征直接融合、决策融合和特征层级融合的对比试验表明，特征层级融合更加适用于两类不同数据类型、不同抽象程度的特征，特征层级融合可以兼顾特征直接融合和决策层融合的优点，可以进一步提高中文韵律边界标注的精度。

下一步，将探索并丰富文本特征，如加入词向量等语义、语法结构特征等；同时，探索其他网络拓扑结构，让文本特征和声学特征在模型中融合得更加合理。

8. 致谢

本文研究得到了国家自然科学基金面上项目（编号：61425017，61773379，61603390，61771472）；国家重点研发计划（编号：2018YFB1005003）的经费支持。

参考文献

[1] Chu M, Qian Y. 2001. Locating boundaries for prosodic constituents in unrestricted Mandarin texts. *Computational linguistics and Chinese language processing*. 6(1), 61-82.

[2] Wang M Q, Hirschberg J. 1992. Automatic classification of intonational phrase boundaries. *Computer Speech & Language*. 6(2), 175-196.

[3] Levow G A. 2008. Automatic Prosodic Labeling with Conditional Random Fields and Rich Acoustic Features. *Proc. International Joint Conference on Natural Language Processing*. Hyderabad, India, 217-224.

[4] Rosenberg A, Fernandez R, Ramabhadran B. 2015. Modeling phrasing and prominence using deep recurrent learning. *Proc. the Annual Conference of the International Speech Communication Association*. Dresden, Germany, 136-141.

[5] Busser B, Daelemans W, Bosch A. 2001. Predicting phrase breaks with memory-based learning. *Proc. 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*. Edinburgh, United Kingdom. The University of Edinburgh, 29-34.

[6] Wightman C W, Ostendorf M. Automatic labeling of prosodic patterns . 1994. *IEEE Transactions on speech and audio processing*, 2(4): 469-481.

[7] 胡伟湘、徐波、黄泰翼（2002）汉语韵律边界的声学实验研究。《中文信息学报》，16(1)：44—49。

[8] Hasegawa J M, Chen K, Cole J. 2005. Simultaneous recognition of words and prosody in the boston university radio speech corpus. *Speech Communication*, 46(3): 418-439.

[9] Chen Q, Ling Z H, Yang C Y. 2015. Automatic phrase boundary labeling of speech synthesis database using context-dependent HMMs and N-Gram Prior Distributions. *Proc. the Annual Conference of the International Speech Communication Association*. Dresden, Germany, 227-234.

[10] 吴晓如、王仁华、刘庆峰（2003）基于韵律特征和语法信息的韵律边界检测模型。《中文信息学报》，17(5):49—55。

[11] Ding C, Xie L, Yan J. Automatic prosody prediction for Chinese speech synthesis using BLSTM-RNN and embedding features . 2015. *Proc. Automatic Speech Recognition and Understanding*. Arizona, USA, 98-102.

[12] Lin C K, Lee L S. 2005. Improved spontaneous Mandarin speech recognition by disfluency interruption point (IP) detection using prosodic features. *Proc. Ninth European Conference on Speech Communication and Technology. Lisbon, Portuguese*, 78-85.

[13] Tieleman T, Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. https://www.coursera.org/learn/neural-networks. visited 5-Jan-172017.

[14] Hinton G E, Srivastava N, Krizhevsky A, et al. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *Computer Science*, 3(4): 212-223.

傅睿博 中国科学院自动化研究所,博士研究生,主要研究领域为语音合成。
E-mail: ruibo.fu@nlpr.ia.ac.cn

陶建华 中国科学院自动化研究所,博士,研究员,博士生导师,主要研究领域为情感识别、语音识别与合成、人机交互、模式识别、网络音视频信息处理。
E-mail: jhtao@nlpr.ia.ac.cn

温正棋 中国科学院自动化研究所,博士,副研究员,主要研究领域为语音识别与合成。
E-mail: zqwen@nlpr.ia.ac.cn