# 基于静音时长和文本特征融合的韵律边界自动标注*

傅睿博[1,3]，李雅[1]，温正棋[1]，陶建华[1,2,3]
(1. 中国科学院自动化研究所 模式识别国家重点实验室，北京 100190；
2. 中国科学院自动化研究所 中国科学院脑科学与智能技术研究中心，北京 100190；
3. 中国科学院大学 人工智能技术学院，北京 100190)

**文　摘：**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的不一致、耗时的缺点。仿照人工标注流程，本文运用循环神经网络分别对文本和音频两个通道训练子模型，对子模型的输出采用模型融合，从而获得最优标注。我们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加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，本文所采用的静音时长特征相比于传统声学特征对自动韵律标注的性能有所提高，决策融合方法相比于直接特征层面融合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；决策融合；静音时长；语料库构建；语音合成

**中图分类号：** H116.4；TP181

语料库建设在语音相关技术中占有重要的地位，特别是语音合成音库的构建。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：音段标注是对音素序列标注对应的起始和结束时间，考虑到目前音段的自动标注技术已经相对成熟，可以基本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。韵律信息在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急需解决的问题。

在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：第一类即为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]等方法，主要以词性、字词的位置、数量信息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；第二类即为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，利用决策树和隐马尔科夫模型(Hidden Markov Model,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，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；第三类采用结合文本和声学特征的韵律边界自动标注，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[7]；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[8]；此类方法可以综合文本和音频两个通道的特征，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与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

***基金项目：**国家高技术研究发展计划（863 计划）（No.2015AA016305），国家自然科学基金项目(NSFC)（No.61425017,No.61403386），中国科学院战略性先导科技专项（GrantXDB02080006），中国社会科学基金重大项目（13&ZD189）。

**作者简介：**傅睿博（1993-），男（汉），浙江，博士研究生。
**通讯联系人：**陶建华，研究员，E-mail：jhtao@nlpr.ia.ac

中的韵律停顿识别。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，较语音识别中的实际应用场景噪声较小，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是存在一定相关性的，因此文本的研究的核心是如何更好地将文本和声学特征融合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，词向量等基于神经网络所采用的特征已经被用于韵律建模研究中[9]，然而在声学特征的选取大部分研究还采用的语音识别技术中所采用的声学参数（如基频、能量、谱参数）[10]，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基于文本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。

本文主要探索了静音时长特征的引入和音频和文本两个通道的韵律边界标注模型决策融合方法。主要贡献有如下三点：（1）探究了静音时长作为深层次抽象特征对韵律边界自动标注的贡献。（2）决策融合的方法被用于结合声学特征和文本的韵律边界标注模型，相比在特征层面的直接融合有所提高。（3）采用无标注数据判别静音模型来提高模型的鲁棒性。

文本的行文组织结构如下：在第一部分，对韵律边界自动标注系统的构建思路和整体框架进行阐述；在第二部分，介绍了静音时长特征提取流程；第三部分，介绍了所采用的决策融合策略；第四部分，介绍实验系统的构建、结果及评价分析。第五部分，对本文所做的工作进行总结与展望。

## 1. 模拟人工标注的自动标注系统

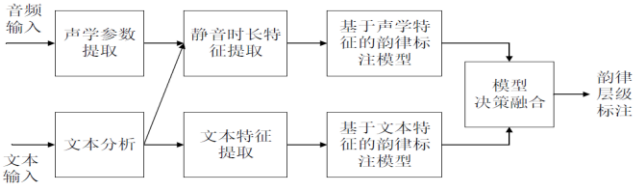


图 1 整体系统框架

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）的循环神经网络的分别训练基于文本特征和基于静音声学特征的韵律层级标注模型。

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”，并参考边界两侧音高重置、边界调和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。

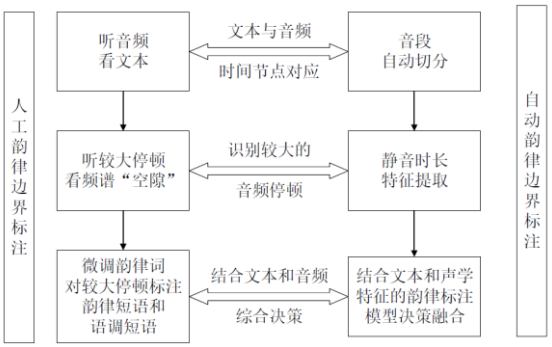


图 2 韵律标注各环节类比

在静音时长特征提取过程中，传统方法中采用的基频、能量等参数的建模单位为帧，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、短语间韵律的关系的能力，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，对的静音时长提取，将浅层以帧为单位的梅尔频率倒谱系数（Mel Frequency Cepstrum Coefficient，MFCC）转化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的标签，对该大语料进行再训练的方式。

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 的循环神经网络分别训练基于文本和声学特征的标注模型。由于声学参数静音时长已经过一系列处理，处于较高层次的抽象特征，而文本特征采用 one-hot 归一化的形式较为稀疏。本文因此采用了对两通道分别训练模型的输出层采用决策融合的方式，相比于直接将文本和声学层

面在特征层面上融合的方式，韵律边界标注的效果有所提高。

## 2 静音时长特征提取

静音时长特征提取涉及图 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和语音合成前端的语法词分词工具得到以词为单位的时间边界信息，但由于该切分边界精度不高，自动切分难以保证在每一个词后面切分出较为精准的静音，若采用人工对静音进行切分，也会出现一些人工判别标准不一致导致的误差，故在该环节加入一个静音判别模型。

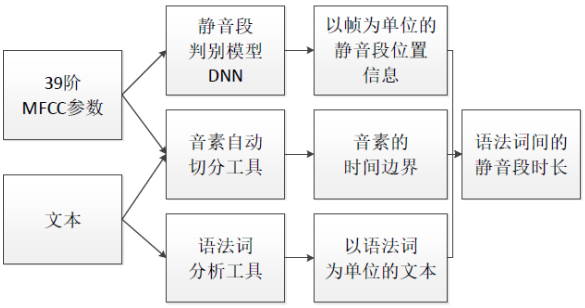


图 3 整体静音时长特征提取流程

### 2.1 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，然而我们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的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，我们采用了两轮训练的方式来扩充训练数据。第一轮运用已有少量有标注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。第二轮训练将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训练静音判别模型。

### 2.2 静音时长提取

本文对语法词间静音时长提取如下：

假定一句话的由 m 个语法词组成，即

$$[w_1, w_2, w_3, \dots, w_m]$$

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$o_j^\alpha$，则

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

其中$p_j$为第 j 帧为静音段的概率，$\alpha$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

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

考虑到在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

考虑到静音时长会受到静音段前后相邻两个字的时长所影响，定义相对时长

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

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

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

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

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。

为了让静音段特征对于不同韵律层级的标注更加有针对性：在比较严格判别条件下，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语调短语的区分；在判别条件比较宽松的情况下，词之间的静音时长更加细微的差异得到体现，更加易于韵律词与非韵律词的区分。因此，我们将$\alpha$取从 0.1 到 0.9，以 0.1 为间隔递增，得到 9 组特征。

## 3. 模型决策融合

由于声学特征是连续特征，文本特征是离散特征，且静音时长这一组声学特征已经过一系列处理，具

有明显的物理意义，而文本特征属于浅层特征，直接将文本和声学特征在特征层面融合有一些不合理。本文采用的是模型决策融合的方式，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的概率和模型整体正确率作为决策依据，其决策过程具体为：

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

其中

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

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

$f_a(x_a)$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$P_t(C)$分别表示分类器判别正确的概率，$P_a\big(C|f_a(x_a)\big)$和$P_t\big(C|f_t(x_t)\big)$分别表示分类器的先验概率，即模型对该类别判别的正确率，$P_a(f_a(x_a))$和$P_t(f_t(x_t))$分别表示模型输出的类别判别概率

## 4. 实验及结果分析

### 4.1 实验语料

我们采用有专业女性录音人录制的用于语音合成的语料。该语料总共含有 20000 句，总字数大于 40000，韵律边界（韵律词，韵律短语，语调短语）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%，词性分析的正确率为 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为大语料的 10%，共 2000 句。

### 4.2 静音时长特征提取及对比

**(1)系统构建**

静音段判别模型分为小语料训练，扩充后的大语料模型训练，均采用 DNN 模型进行训练，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[9]训练 10 代，batch-size 为 128, dropout[10]系数为 0.2。

**(2)静音切分结果评价**

小语料训练的静音判别模型的准确率为 97.4%，扩充标注后的大语料训练的静音判别模型的准确率为 98.7%，原因为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。

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。我们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，将第一轮和第二轮两组训练静音判别模型所提取的参数进行对比实验，我们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。

采用二轮训练扩充无标注数据的方法提高静音段识别精度，各个静音时长特征单独用于韵律边界检测的 F1 值平均提高了，

通过该实验结果我们还可以观察到将提取到的时长进行归一化、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的反应音节间的静音段在全句中声学时长分布。

### 4.3 韵律边界标注系统构建及评价

本节所有实验，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的输入。

为了探究如何更好地将文本特征和声学特征

**表 1 各特征在一维线性分类器中评价结果（韵律短语）**

| | $H_{ab}^{\alpha}(i)$ | $H_{wab}^{\alpha}(i)$ | $H_{r1ab}^{\alpha}(i)$ | $H_{r1wab}^{\alpha}(i)$ | $H_{r2ab}^{\alpha}(i)$ | $H_{r2wab}^{\alpha}(i)$ | $H_{nab}^{\alpha}(i)$ | $H_{nwab}^{\alpha}(i)$ |
|---|---|---|---|---|---|---|---|---|
| 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 |

注：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。

更好地结合以提高韵律层级自动标注的精度，我们尝试了将文本和声学特征直接在特征层面融合，和分别训练基于文本特征和声学特征的标注模型，在决策层面融合的方式。

**(1)基于传统声学特征的系统构建**

我们将传统方法中所采用的处理较少的包括时长（音节时长，静音时长，相邻音节时长比例）、基频（拟合参数、极值、区间、均值、梯度）、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。

**(2)基于静音时长声学特征的系统构建**

我们将音节时长和第 2.2 节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。

**(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、音调和根据 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 one-hot 的方式进行归一化，文本特征共 214 维。

**(4)基于特征融合的系统构建**

我们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)决策融合的系统构建**

我们将基于静音时长声学特征和文本特征分别训练得到最好的模型，用第 3 部分所介绍的决策策略进行融合。

**(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

| | 隐含节点数 | 隐含层层数 | 学习率 | 输出层节点 |
|---|---|---|---|---|
| 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 |
| 特征融合 | 512 | 2 | 0.001 | softmax |

**(7)实验结果**

表 3 展示了上文介绍的 5 个系统的评价效果，我们对各个韵律层级使用 F1 值作为评价指标。

表 3 韵律自动标注评价结果

| | 传统声学特征 | 本文声学特征 | 文本特征 | 特征融合 | 决策融合 |
|---|---|---|---|---|---|
| 韵律词 | 83.72 | 85.98 | 95.10 | 95.08 | **96.35** |
| 韵律短语 | 72.31 | 85.64 | 69.43 | 87.70 | **87.85** |
| 语调短语 | 74.62 | 84.81 | 84.24 | 85.66 | **85.83** |

**(8)分析与讨论**

本文所采用的静音时长等一系列声学特征相比传统声学特征取得更好效果，原因为传统的声学特征提取参数的所采用的语料是建立在各音节边界十分准确的情况下的，而本文的目的是构建自动韵律边界标注，所以采用的是边界自动切分，切分精度的误差会传递给后续基频、能量等参数的提取。本文所采用的静音时长参数的出发点就是考虑到自动音节切分精度不高的问题。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，传统声学特征更多是体现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韵律特征提高明显。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成分关系的不足。

对于直接在特征层面对文本和声学特征进行融合的方法，我们发现对于韵律短语和语调短语这两个层级，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，我们发现融合未能提高该层级标注的性能，其原因为静音时长对于是否是韵律词的判别区分度较小，两个语法词之间的停顿时长对韵律词的区分度较小，依靠文本特征就可以达到较好的识别效果。

决策融合的方式相比于特征融合的方式在各个韵律层级的标注效果都有所提高。分析原因在于我们目前采用的特征融合方式对于两类不同类别的特征在模型训练阶段不能得到有效的区分训练。目前的网路结构不能凸显对于较高韵律层级的标注，静音时长特征相比文本特征有着更大的贡献。决策融合的方法由于考虑了各子模型的先验概率，更加易于“取长补短”，在每个标注过程中都综合给出最优的判别结果，从而获得较好的效果。

## 5. 总结

本文探究了静音时长特征的提取及其相比于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于韵律边界的检测有明显提升。同时，特征直接融合和模型决策融合的对比试验表明，模型决策融合更加适合对两类不同数据类型、不同抽象程度的特征，决策融合的方式可以进一步提高中文韵律边界标注的精度。

未来，我们将探索丰富文本特征，如加入词向量等语义、语法结构特征等。同时，我们将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更加合理。

## 参 考 文 献


[1] Locating boundaries for prosodic constituents in unrestricted Mandarin texts[J]. Computational linguistics and Chinese language processing, 2001, 6(1): 61-82.

[2] Automatic classification of intonational phrase boundaries[J]. Computer Speech & Language, 1992, 6(2): 175-196.

[3] Automatic Prosodic Labeling with Conditional Random Fields and Rich Acoustic Features[C]//IJCNLP. 2008: 217-224.

[4] Modeling phrasing and prominence using deep recurrent learning[C]//INTERSPEECH. 2015: 3066-3070.

[5] Predicting phrase breaks with memory-based learning[C]//4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis. 2001.

[6] Automatic labeling of prosodic patterns[J]. IEEE Transactions on speech and audio processing, 1994, 2(4): 469-481.

[7] Simultaneous recognition of words and prosody in the boston university radio speech corpus[J]. Speech Communication, 2005, 46(3): 418-439.

[8] Automatic phrase boundary labeling of speech synthesis database using context-dependent HMMs and n-gram prior distributions[C]//INTERSPEECH. 2015: 1581-1585.

[9] Automatic prosody prediction for Chinese speech synthesis using BLSTM-RNN and embedding features[C]//Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on. IEEE, 2015: 98-102.

[10] Improved Spontaneous Mandarin Speech Recognition by Disfluency Interruption Point (IP) Detection Using Prosodic[J]. 2005.

[11] Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude[J]. COURSERA: Neural networks for machine learning, 2012, 4(2).

[12] Improving neural networks by preventing co-adaptation of feature detectors[J]. arXiv preprint arXiv:1207.0580, 2012


# Automatic Prosodic Boundaries Labeling based on Fusing the Duration of Silence and the Lexical Features


Ruibo Fu[1, 3], Ya Li[1] ,Zhengqi Wen[1] ,Jianhua Tao[1,2,3]

1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

2. CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

3. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China



**Abstract:**

Automatic prosodic boundaries labeling plays an important role on 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 by reducing human factors. In this paper, inspired by the manual labeling method, an ensemble strategy that combines the lexical 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 sub-system is designed to detect the duration of silence. Considering that the duration of silence has the intuitionist physics meaning, the contribution of the duration of silence and its normalization methods in the automatic prosodic boundaries labeling are investigated.

The acoustic features that are extracted from the sub-system are being more processed and more abstractive and have more intuitive physic meaning. On the other hand, the lexical features extracted from the text is less intuitive and more sparse, which may weaken the contribution of the acoustic features. Therefore, an ensemble strategy that combines the two sub-systems trained separately is proposed instead of fusing the acoustic features and the lexical features. Experiments show that the effectiveness of the duration of silence being extracted as the acoustic features and the ensemble strategy in improving the F1-score of the prosodic boundaries labeling compared with previous features fusion strategies.



**Key words:** prosodic boundaries labeling; the ensemble strategy; the duration of silence; the construction of corpus; speech synthesis