基于决策树的英语焦点语音转换

孟凡博[1,2]，吴志勇[1,2,3]，蒙美玲[2,3]，贾珈[1,2]，蔡莲红[1,2]

（1. 清华大学 计算机科学与技术系，普适计算教育部重点实验室，清华信息科学与技术国家实验室，北京 100084；

2. 清华大学 深圳研究生院，清华大学-香港中文大学媒体科学、技术与系统联合研究中心，深圳 518055；

3. 香港中文大学 系统工程与工程管理学系，香港）

摘　要： 焦点是语言表达的重要方式，焦点重音是重要的韵律特征，实现中性语音到焦点语音的转换可以提高语音的表现力。该文提出了声学特征局部凸显度的表示方法，分析了由中性到焦点语音，焦点单词所属音节声学特征变化与中性语音相应音节声学特征局部凸显度的相关性，提出了一种基于决策树的英语焦点语音的转换模型。该模型采用决策树对训练语料进行聚类，所用上下文包括音节与焦点单词的相对位置以及音节在韵律结构中（韵律短语、韵律词）的位置。在此基础上，提出了一种基于局部凸显度的中性到焦点语音声学特征变化的预测算法。采用该算法后，客观实验中声学特征变化平均绝对值误差降低到 0.08，主观实验表明，本文提出的模型的转换语音具有更好的焦点表达效果和自然度。

关键词：语音转换；焦点语音；声学特征；韵律结构；局部凸显

中图分类号：TP391　　**文献标志码：**A

English emphatic speech conversion based on decision tree

MENG Fanbo[1,2], WU Zhiyong[1,2,3], Helen Meng[2,3], JIA Jia[1,2] and CAI Lianhong[1,2]

(1.Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China; 2. Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China; 3. Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, China)

Abstract: Emphasis is an important feature of prosody. The technology of emphatic speech conversion can improve the expressiveness of the converted speech. This paper firstly defines the local prominences (LPs) of acoustic features, and then analyzes the correlations between the feature changes of the syllables of the emphasized words from neutral to emphatic speech and the LPs of the corresponding syllables of the neutral speech. Based on the analysis, an English emphatic speech conversion model based on decision tree is proposed. This model firstly clusters the training data with decision tree. After data clustering, a prediction algorithm of the feature changes from neutral to emphatic speech based on LPs is proposed. Experiments demonstrate that the mean average error of the proposed model decreases to 0.08. The naturalness and the emphasis intensity of the converted speeches are also improved.

key words: speech conversion; emphatic speech; acoustic feature; prosodic structure; local prominence

语音是人们在日常生活中重要的交流手段，它不仅能够表达文字所含的语义信息，还可以通过说话者的说话方式如语气、焦点等表达出其他含义。其中，焦点是语言表达所必要的。根据广泛认可的焦点-重音（focus-to-accent）理论，在音高重音语言（如英语）中，成为焦点的词或成分会以音高重音的形式在语音中表现出来，即形成焦点重音[1]。

焦点重音具有局部凸显性，声学特征高于临近音节的音节更容易被感知为重音[2]。一般的，与焦点重音感知相关的声学特征主要有基频、时长和能量[3]。研究发现，焦点语音的声学表现受多种因素影响。语句中焦点的声学表现与其在句中的位置是相关的。文[4]发现焦点的基频和时长变化随着所处韵律层级（如音节、韵律词、韵律短语和语调短语）的增大而增大。焦点语音中音节的声学特征也与焦点的相对位置有关。文[5]指出焦点会增大所在单词重读音节的基频范围，同时减小焦点之后音节的基频范围。焦点的声学特征（如基频、时长和能量）之间也有一定的相关性。文[6]发现，在焦点重音的声学表现上，基频和时长具有互补的作用，从中性到焦点语音，焦点的基频变化和时长变化呈负相关性。此外，焦点重音的基频与能量呈正相关性[7]。

焦点重音是重要的韵律特征，它使得语音的基频起伏更为明显，听起来更具有表现力。文[8-9]分别采用线性预测模型和隐 Markov 模型对中性到焦点语音声学特征的变化进行建模。文[10]将重音等级分为轻、中、重三类，统计了不同重音等级发音单元处于不同韵律边界时声学特征变化，并建立了相应的统计模型。已有的焦点语音转换模型，没有很好地利用焦点声学特征分析的结果（如焦点后的基频抑制现象），降低了转换语音的焦点感知，并且会有修改幅度过大的情况，降低了转换语音的自然度。

本文提出了声学特征局部凸显度的表示方法，分析了焦点单词所属音节声学特征变化与中性语音中相应音节声学特征局部凸显度的相关性，建立了基于决策树的中性到焦点语音转换模型。转换模型

收稿日期：2012-09-25

基金项目：国家自然科学基金面上项目（60805008，60928005，61003094）；香港特区政府研究资助局基金项目（N-CUHK414/09）；国家“九七三”重点基础项目（2013CB329304）；国家“八六三”高技术项目（2012AA011602）

作者简介：孟凡博（1984—），男（汉），黑龙江，博士研究生。

通信作者：蔡莲红，教授，E-mail: clh-dcs@tsinghua.edu.cn

首先根据焦点的位置对训练语料进行聚类；由于焦点数据远少于非焦点数据，为了避免过度分类导致的数据稀疏，采用决策树选择对数据区分度最大的问题进行聚类。在聚类之后，针对模型预测幅值过大问题，提出了一种基于中性语音声学特征局部凸显度的中性到焦点语音声学特征变化预测算法；为了进一步提高算法预测精度，该算法还兼顾了不同声学特征变化之间的相关性。

1 语料库

本文研究采用的语音是由 350 句英文文本录制的数据。平均句长为 16 个音节。在每个语句中设置一个或多个焦点词，且处于句中不同位置，例如：

“*I have met Peterson on one occasion.*”和“*Fighting thirst is the first thing to be done in the country.*”

其中“Peterson”、“occasion”、“thirst”和“first”为焦点词。

本文邀请了一位母语为英语的加拿大女性发音人录音，每句文本分别录制了中性语音和对应的突出焦点词的语音，要求在录制中性语音时发音平白不带焦点，在录制带有焦点的语音时发音突出焦点。一共得到 700 句语音，以 Microsoft Windows Wav 格式保存（16 b，单声道，16 kHz 采样率）。语料库基于 Festival[11]进行了音节切分和文本分析，标注了音节是否重读，音节与焦点的相对位置以及音节所在韵律词、韵律短语的位置信息等，对错误的切分进行了人工修正，并提取了每个音节的基频、时长和短时能量。

从 350 句文本中随机选取 20 句文本（以及相应 40 句录音）作为测试语料，其余语料作为训练语料。

2 英语焦点声学特征变化及其与中性语音声学特征局部凸显度的相关性分析

英语是非声调语言，然而在连续语流中，焦点的声学表现不仅仅是基频均值的提高，还有基频最大值、基频范围的增大，时长的增长以及能量的增强等。文[12]统计分析了从中性到焦点语音的声学特征变化及其方差，数据表明：由于焦点的基频最大值变化大于基频最小值变化，并且基频最大值与基频最小值的差并不稳定，导致从中性到焦点语音基频范围变化的方差较大(大约为其他特征的 100 倍)，因此基频范围不够稳定不适合用于建模。而基频最大值的增大也是焦点的重要表现，并且文[12]的统计表明，基频最大值和最小值的变化较为稳定。因此，本文选择对音节的基频最大值（pitch maximum，记为 P_{Max}）和基频最小值（pitch minimum，记为 P_{Min}）进行分析、建模，此外，建模特征还包括音节的平均时长（duration，记为 D）和短时能量（short time energy，记为 E）。

2.1 从中性语音到焦点语音的声学特征相对表示

当中性语音变为焦点语音时，声学特征差异的相对值体现了这种转换。设中性语料库中第 i 个音节的基频最大值为 $P_{\mathrm{Max}}^{i,\mathrm{neu}}$，相对应的焦点语音的基频最大值为 $P_{\mathrm{Max}}^{i,\mathrm{foc}}$，则由中性语音到焦点语音第 i 个音节基频最大值的相对变化 $\Delta P_{\mathrm{Max}}^{i}$ 表示为

$$\Delta P_{\mathrm{Max}}^{i}=\frac{P_{\mathrm{Max}}^{i,\mathrm{foc}}}{P_{\mathrm{Max}}^{i,\mathrm{neu}}}. \tag{1}$$

由中性语音到焦点语音第 i 个音节其他声学特征变化（包括基频最小值变化 $\Delta P_{\mathrm{Min}}^{i}$、时长变化 ΔD^{i} 和能量变化 ΔE^{i}）的计算方法与 $\Delta P_{\mathrm{Max}}^{i}$ 相同。

2.2 声学特征局部凸显性的相对表示

焦点的声学特征具有局部凸显性，基频、时长和能量超过临近音节的音节更容易被感知为焦点重音[2]，即焦点的感知与其所处的上下文是有关系的。本文采用当前音节声学特征与所在韵律短语的所有音节声学特征平均值的比值来表示当前音节该声学特征在所处上下文中的突显程度，即局部凸显度（local prominence），设语料库中第 i 个音节所在韵律短语的音节编号为 $i_1,i_2,\ldots,i_k$，则音节 i 基频最大值局部凸显度 $\hat{P}_{\mathrm{Max}}^{i}$ 表示为

$$\hat{P}_{\mathrm{Max}}^{i}=\frac{P_{\mathrm{Max}}^{i}}{\frac{1}{k}\sum_{j=i_1}^{i_k}P_{\mathrm{Max}}^{j}}. \tag{2}$$

其中 k 为第 i 个音节所在韵律短语的音节个数。第 i 个音节其他声学特征局部凸显度（包括基频最小值局部凸显度 $\hat{P}_{\mathrm{Min}}^{i}$、时长局部凸显度 $\hat{D}^{i}$ 和能量局部凸显度 $\hat{E}^{i}$）的计算方法与 $\hat{P}_{\mathrm{Max}}^{i}$ 相同。

2.3 焦点单词所含音节声学特征变化与中性语音相应音节局部凸显度的相关性分析

本文统计了中性和焦点语音中焦点单词所含音节声学特征变化与中性语音相应音节声学特征局部凸显度的相关系数，如图 1 所示。在声学特征变化之间的相关性方面，基频（包括最大值和最小值）变化与能量变化有正相关性，而与时长变化有明显的有负相关性，即音节基频提高幅度较大时，音节的能量提高的幅度也较大，而时长的提升的幅度反而较小。在声学特征变化与局部凸显度的相关性方面，焦点单词所含音节各声学特征的变化与中性语音中该音节相应声学特征的局部凸显度有明显的负相关性（如 $\Delta P_{\mathrm{Max}}^{i}$ 与 $\hat{P}_{\mathrm{Max}}^{i}$、$\Delta P_{\mathrm{Min}}^{i}$ 与 $\hat{P}_{\mathrm{Min}}^{i}$）；此外，各声学特征的变化还与其他声学特征的局部凸显度有相关性，基频变化与时长局部凸显度以及时长变化与基频局部凸显度有明显的正相关性。这说明，由于生理等条件约束，基频不能无限制地增大。因此当原始中性语音中音节的基频局部凸显度较高时，说话人会较大幅度增长时长、较小幅度增加基频来突出焦点，而当原始中性语音中音节的基频局部凸显度较低时，说话人会较大幅度提高基频、较小幅度增长时长来突出焦点。

	ΔP_{Max}	ΔP_{Min}	ΔD	ΔE	$\hat{P}_{\mathrm{Max}}$	$\hat{P}_{\mathrm{Min}}$	$\hat{D}$	$\hat{E}$
ΔP_{Max}	1.00	0.86	−0.80	0.58	**−0.80**	−0.60	0.32	−0.20
ΔP_{Min}	0.86	1.00	−0.82	0.38	−0.89	**−0.83**	0.48	−0.50
ΔD	−0.80	−0.82	1.00	−0.40	0.66	0.52	**−0.50**	0.35
ΔE	0.58	0.38	−0.40	1.00	−0.27	−0.05	−0.19	**−0.05**

图 1 中性和焦点语音中焦点单词所含音节声学特征变化与中

性语音相应音节声学特征局部凸显度的相关系数

在目标为焦点语音的转换过程中，由于焦点语音中各音节的声学特征及其局部凸显度是未知的，但是其可以通过中性语音到焦点语音各音节声学特征变化以及中性语音相应音节声学特征局部凸显度计算。因此本文选择基于中性语音声学特征的局部凸显度对中性语音到焦点语音声学特征变化进行预测。

3 基于决策树的焦点语音转换模型

焦点语音的声学特征与焦点的位置是相关的[4,5]，当焦点在句中的位置不同时，由中性语音到焦点语音，焦点的声学特征变化是不同的；此外，焦点对临近音节的声学特征也有影响（如焦点后的基频抑制现象）。因此为了提高模型的预测精度，需要根据焦点在句中位置的上下文以及音节与焦点的相对位置的上下文对训练数据进行聚类，然而由于一句话仅有少数几个词为焦点，语料中焦点数据要远少于非焦点数据。为了避免过度分类而导致数据稀疏，本文提出了基于决策树的焦点语音转换模型，采用决策树对训练数据进行聚类。在构建决策树时，每次选择区分性最大的问题对训练语料进行聚类，决策树分裂停止条件有两个：1）待选问题集里没有能够降低距离的问题；2）当前节点中数据个数少于一定阈值。在聚类之后对每个叶节点中的数据分别建立声学特征变化的预测模型。

3.1 基于位置上下文的决策树聚类

根据焦点语音中各音节的声学表现与焦点位置的关系，本文一共引入了 3 个音节与焦点的相对位置问题和 9 个音节在韵律结构中的位置问题，如表 1 所示。由于焦点数据在总数据中的比例较低，导致焦点位置相关问题的区分性弱于其他问题。为了避免将焦点数据与非焦点数据聚类到一起，引起较大的预测误差，本文在决策树聚类过程中首先使用焦点相关问题进行分裂，再使用句中位置问题进行分裂。

由中性语音到焦点语音焦点单词所含音节的声学特征变化之间是相关的，并且声学特征变化也与中性语音相应音节声学特征局部凸显度有一定的相关性，因此本文针对每个音节的从中性语音到焦点语音的各声学特征变化和中性语音相应音节各声学特征的局部凸显度组成的 8 维特征向量进行聚类。令 $\boldsymbol{V}_i$ 为第 i 个音节对应的特征向量：

$$\boldsymbol{V}_i=[\Delta P^i_{\mathrm{Max}}\quad \Delta P^i_{\mathrm{Min}}\quad \Delta D^i\quad \Delta E^i\quad \hat{P}^i_{\mathrm{Max}}\quad \hat{P}^i_{\mathrm{Min}}\quad \hat{D}^i\quad \hat{E}^i]. \tag{3}$$

表 1 决策树问题集

问题类型	问题
音节与焦点相对位置问题	音节所在单词是焦点/在焦点之前/在焦点之后
音节所在韵律短语在句中位置问题	音节所在韵律短语位于句首/句中/句末
音节所在韵律词在韵律短语中位置问题	音节所在韵律词位于韵律短语首/中/末
音节与重读音节相对位置问题	音节 是重读音节/在重读音节之前/在重读音节之后

为了将声学特征相近的样本聚为一类以提高模型的预测精度，决策树训练时根据节点内样本的集内距离进行聚类。集内距离为节点中各样本到节点中心的平均 Euclidean 距离。设 L 为当前节点，该节点内的音节序号为 $l_1,l_2,\ldots,l_n$，则节点 L 的集内距离为

$$d(L)=\frac{1}{n}\sum_{i=l_1}^{l_n}\mathrm{f}\left(\boldsymbol{V}_i,\frac{1}{n}\sum_{j=l_1}^{l_n}\boldsymbol{V}_j\right). \tag{4}$$

其中 n 为 L 内的音节个数，$f(\cdot)$ 表示欧氏距离计算方法。设待选问题集为 Q，L 被问题 q 分裂为子节点 L_{ql} 和 L_{qr}，则选择令集内距离**下降最大**的问题对当前节点进行分裂，如下所示：

$$\begin{cases}\Delta d=d\left(L_{ql}\right)+d\left(L_{qr}\right)-2d(L)\\ q_0=\arg\min\limits_{q\in Q}\left(\Delta d\right)\end{cases}. \tag{5}$$

只有当 Δd 为负时，节点才会分裂。

3.2 基于局部凸显度的焦点语音声学特征变化预测算法

当采用决策树对数据进行聚类之后，由于聚类时选择的问题是使得集内距离下降最大的问题，因此属于同一个叶节点的特征向量具有相似的特性，所以可以在决策树聚类时采用由中性语音到焦点语音各音节声学特征变化组成的 4 维向量 $\boldsymbol{V}_i^{'}=[\Delta P^i_{\mathrm{Max}}\ \Delta P^i_{\mathrm{Min}}\ \Delta D^i\ \Delta E^i]$ 进行聚类，然后采用叶节点中所有特征向量的平均值（average of the feature vectors，AFV）作为这个叶节点所对应上下文的声学特征变化预测值。然而由于预测时没有考虑中性语音音节声学特征的局部凸显度，该预测方法经常会出现预测值过大或过小的情况。针对这个问题，本文提出了基于中性语音音节声学特征局部凸显度的焦点语音声学特征变化预测算法。假设声学特征的变化与中性语音音节声学特征的局部凸显度具有线性关系，即有：

$$\begin{cases}\Delta P_{\mathrm{Max}}=a_1\hat{P}_{\mathrm{Max}}+b_1,\\ \Delta P_{\mathrm{Min}}=a_2\hat{P}_{\mathrm{Min}}+b_2,\\ \Delta D=a_3\hat{D}+b_3,\\ \Delta E=a_4\hat{E}+b_4.\end{cases} \tag{6}$$

考虑到声学特征之间的相关性（correlations between the changes of acoustic features，CCAF），式（6）扩展为式（7）：

$$\boldsymbol{R}=\boldsymbol{A}\boldsymbol{T}+\boldsymbol{B} \tag{7}$$

其中：

$$\boldsymbol{R}=\begin{bmatrix}\Delta P_{\mathrm{Max}}\\ \Delta P_{\mathrm{Min}}\\ \Delta D\\ \Delta E\end{bmatrix},\ \boldsymbol{A}=\begin{bmatrix}a_{11}&a_{12}&a_{13}&a_{14}\\ a_{21}&a_{22}&a_{23}&a_{24}\\ a_{31}&a_{32}&a_{33}&a_{34}\\ a_{41}&a_{42}&a_{43}&a_{44}\end{bmatrix},\ \boldsymbol{T}=\begin{bmatrix}\hat{P}_{\mathrm{Max}}\\ \hat{P}_{\mathrm{Min}}\\ \hat{D}\\ \hat{E}\end{bmatrix},\ \boldsymbol{B}=\begin{bmatrix}b_1\\ b_2\\ b_3\\ b_4\end{bmatrix}.$$

$a_{ii}(i=1,2,3,4)$ 表示相应声学特征变化与中性语音该声学特征局部凸显度的相关性，而 $a_{ij}(i\neq j)$ 表示该声学特征变化与其他声学特征局部凸显度的相关性，当 $\boldsymbol{A}$ 为 $\boldsymbol{0}$，$\boldsymbol{B}$ 为特征向量的平均值时，该算法退化为采用 AFV 进行预测。

实验中，本文针对决策树每个叶节点的数据，分别采用非线性最小方差回归方法计算式（7）中的

系数矩阵 $\boldsymbol{A}$、$\boldsymbol{B}$，即不同的叶节点（也即不同的焦点位置上下文）对应不同的系数矩阵。

3.3 焦点语音转换模型预测及实现

本文首先提取中性语音的声学参数，进而计算每个音节声学特征的局部凸显度，根据焦点位置上下文取得局部凸显度与声学特征变化间的系数矩阵，进而计算得到每个音节的声学特征变化。再通过 STRAIGHT 算法[13]修改语音的基频、时长，在 Hamming 窗平滑下修改能量，最后拼接得到焦点语音。算法具体步骤如下：

设 $S(n)$为中性语音的波形，n 为离散的时间索引，由 b 开始到 e 结束，中性语音共有 N 个音节，其中 $S_i(n)$为第 i 个音节的波形，由 b_i 开始到 e_i 结束。

1）提取参数：采用 STRAIGHT 算法提取 $S_i(n)$的声学参数，设 $W_i(n)$、$P_i(n)$和 $D_i(n)$分别为第 i 个音节的频谱、基频序列和相应的时间序列，其起始时间为 b_i，结束时间为 e_i，P^i_{Max} 和 P^i_{Min} 分别为该音节的基频最大值和基频最小值，E_i 为该音节的短时能量，由于本文模型不修改频谱，因此有

$$W'_i(n) = W_i(n). \tag{8}$$

2）根据式（2）计算当前音节声学特征的局部凸显度 $\hat{P}^i_{\text{Max}}$、$\hat{P}^i_{\text{Min}}$、$\hat{D}^i$ 和 $\hat{E}^i$。根据位置上下文取得决策树中对应叶节点的 $\boldsymbol{A}$、$\boldsymbol{B}$。根据式（7）计算得到当前音节声学特征变化的预测值 ΔP^i_{Max}、ΔP^i_{Min}、ΔD^i 和 ΔE^i。

3）修改基频、时长：

$$P'^i_{\text{Min}} = P^i_{\text{Min}} \times \Delta P^i_{\text{Min}}, \tag{9}$$

$$P'^i_{\text{Max}} = P^i_{\text{Max}} \times \Delta P^i_{\text{Max}}, \tag{10}$$

$$P'_i(n) = P'^i_{\text{Min}} + \frac{P'^i_{\text{Max}} - P'^i_{\text{Min}}}{P^i_{\text{Max}} - P^i_{\text{Min}}} \times \left(P_i(n) - P^i_{\text{Min}}\right), n \in [b_i, e_i], \tag{11}$$

$$D'_i(n) = b_i + \left(D_i(n) - b_i\right) \times \Delta D^i, n \in [b_i, e_i]. \tag{12}$$

4）生成波形：令 $S'^i(n)$ 为目标语音第 i 个音节的波形，在得到目标语音的基频和时间序列之后，采用 STRAIGHT 算法生成波形：

$$S'_i(n') = g\left(W'_i(n), P'_i(n), D'_i(n)\right), n \in [b_i, e_i], n' \in [b'_i, e'_i]. \tag{13}$$

其中 $g(\cdot)$ 表示 STRAIGHT 语音生成算法。

5）修改能量：音节 i 的能量在窗长为 M、窗移为 $M/2$ 的 Hamming 窗 $H_{i,k}(n)$的平滑下，根据 ΔE_i 进行调整：

$$S''_{i,k}(n) = S'_i(n) H_{i,k}(n) \Delta E_i, k \in \left[0, 2\left\lfloor \frac{e'_i - b'_i}{M} \right\rfloor\right]; \tag{14}$$

$$H_{i,k}(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi(n - b'_i - kM/2)}{e'_i - b'_i}\right), n \in \left[b'_i + kM/2, b'_i + (k/2+1)M\right]; \\ 0, n \notin \left[b'_i + kM/2, b'_i + (k/2+1)M\right]; \end{cases} \tag{15}$$

$$S''_i(n) = \sum_{k=0}^{2\left\lfloor \frac{e'_i - b'_i}{M} \right\rfloor} S''_{i,k}(n), n \in [b'_i, e'_i]. \tag{16}$$

含有焦点的转换语音由此拼接得到：

$$S''(n) = \left\{S''_1(n), S''_2(n), ..., S''_i(n), ..., S''_N(n)\right\}. \tag{17}$$

4 实验及结果分析

4.1 预测误差的客观评价实验

为了测试模型的预测精度，本文分别计算采用不同建模参数（AFV、LP、LP+CCAF）时模型预测参数的平均绝对值误差（mean absolute error，MAE）和均方误差（root of the mean square error，RMSE）。设测试语料中共有 N 个音节，F^i_{Max}、F^i_{Min}、F^i_{D} 和 F^i_{E} 分别表示音节 i 的基频最大值、最小值、时长和能量，F'^i_{Max}、F'^i_{Min}、F'^i_{D} 和 F'^i_{E} 为相应的模型预测值，则 MAE、RMSE 计算方法如下：

$$\text{MAE} = \frac{\sum_{j \in \{\text{Max, Min, D, E}\}} \sum_{i=1}^{N} \left|F'^i_j - F^i_j\right|}{4N}, \tag{18}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{j \in \{\text{Max, Min, D, E}\}} \sum_{i=1}^{N} \left(F'^i_j - F^i_j\right)^2}{4N}}. \tag{19}$$

实验结果见表 2，相对于采用特征向量的平均值进行预测，引入局部凸显度之后模型的预测精度有了较大的提高，减小了当中性语音局部凸显度较高或较低时模型的预测误差，降低了均方误差。建模时引入声学特征变化之间的相关性后，模型的预测精度略有提高，均方误差变化不大。

表 2 预测误差客观评价实验结果

测试集	MAE			RMSE		
	AFV	LP	LP+CCAF	AFV	LP	LP+CCAF
集内	0.16	0.10	0.08	0.14	0.09	0.09
集外	0.19	0.14	0.12	0.17	0.13	0.12

4.2 转换语音的焦点表达效果感知实验

本实验用来测试模型是否正确转换焦点。本文对分别采用 AFV 进行预测的模型与采用 LP+CCAF 进行预测的模型的转换语音以及相应的原始焦点录音进行评测。从测试集中取出 10 句中性语音，分别采用 2 个模型转换生成相应的 10 句转换语音，每句转换语音含有 1 个或多个焦点单词。将转换语音以及相应的焦点录音（共 30 句）混合随机顺序播放给听音人，听音的同时提供不含焦点标注的文本；听音人在听音之后，识别句中哪个或哪些单词为焦点，并根据焦点的重读强度进行 5 分制 MOS 评分。

共有 10 位听音人参加了实验，实验结果如表 3 所示。其中准确率指正确识别出的焦点单词的百分比，误识率指非焦点被识别为焦点的百分比，漏识率指未被识别的焦点的百分比，评分指听音人对所识别焦点强度的平均评分。仅采用 AFV 进行预测的模型，其转换语音的焦点识别准确率仅为 85%，而在采用 LP+CCAF 进行预测后，转换语音的焦点识别准确率提高到 97%，达到了焦点录音的水平。此外，采用 LP+CCAF 的转换语音的误识率为 5%，略高于焦点录音 2%，但错误识别的焦点强度评分要明显小于焦点录音。实验结果说明采用声学特征局部凸显度、以及声学特征相关系数可以更有效地预测焦点的声学特征，提高转换语音的焦点表达效果。

4.3 转换语音的自然度评价实验

本实验用来测试模型转换语音的自然度。本文对焦点录音以及在使用相同决策树条件下分别采用 AFV 进行预测的模型和 LP+CCAF 进行预测的模型

的转换语音进行评测。从测试集中取出 10 句中性语音，分别使用 2 个模型转换生成相应的 10 句转换语音将转换语音和相应的焦点录音（共 30 句）混合随机顺序播放给听音人，听音人在听音之后，根据听到的语音的自然度进行 5 分制 MOS 评分。

表 3 转换语音的焦点表达效果评价实验结果

实验语音	准确率/%	误识率/%	评分		漏识率/%
			准确识别	错误识别	
焦点录音	97	2	4.7	4.5	3
AFV 转换语音	85	6	3.5	2.5	15
LP+CCAF 转换语音	97	5	4.3	3.6	3

同样 10 位听音人参加了转换语音的自然度评价实验，对焦点录音、采用 AFV 进行预测的模型的转换语音和采用 LP+CCAF 进行预测的模型的转换语音的自然度 MOS 评分分别为 4.8、3.8 和 4.4。从实验结果来看，采用 LP+CCAF 进行预测很少出现预测参数过大或过小的情况，与采用 LFV 进行预测相比，改进了转换语音的自然度。

5 结论

本文根据焦点的韵律特点，提出了一种基于决策树的焦点语音转换模型。引入了音节与焦点的相对位置以及音节在韵律结构中的位置上下文，采用决策树对训练数据进行聚类；根据焦点的局部凸显性，定义了声学特征局部凸显度，提出了基于声学特征局部凸显度的焦点语音声学特征预测算法。实验证明，该模型可以有效进行焦点语音转换。

参考文献 (References)

[1] Ladd R. Intonational phonology [M]. Cambridge: Cambridge University Press, 1996.

[2] LI Kun, ZHANG Shuang, LI Mingxing, et al. Prominence model for prosodic features in automatic lexical stress and pitch accent detection [C]// INTERSPEECH 2011. Grenoble, France: ISCA, 2011: 2009-2012.

[3] Tamburini F. Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system [C]// Eurospeech 2003. Grenoble, France: ISCA, 2003: 129-132.

[4] 李雅, 卢颖超, 许小颖, 等. 连续语流中韵律层级和调型组合对重音感知的影响[J]. 清华大学学报(自然科学版), 2011, **51**(9): 1239-1243.

LI Ya, LU Yingchao, XU Xiaoying, et al. Influence of rhythm and tone pattern on Mandarin stress perception in continuous speech [J]. *J Tsinghua Univ (Sci and Tech)*. 2011, **51**(9): 1239-1243. （in chinese）

[5] Xu Y, Xu C X. Phonetic realization of focus in English declarative intonation [J]. *Journal of Phonetics*, 2005, **33**(2): 159-197.

[6] 许洁萍, 初敏, 贺琳, 等. 汉语语句重音对音高和音长的影响[J]. 声学学报, 2000, **25**(4): 335-339.

XU Jieping, CHU Min, HE Lin, et al. The influence of Chinese sentence stress on pitch and duration [J]. *Chinese Journal of Acoustics*, 2004, **25**(4): 335-339. （in chinese）

[7] Plag I. The variability of compound stress in English: structural, semantic and analogical factors [J]. *English Language and Linguistics*, 2006, **10**(1): 143–172.

[8] Bou-Ghazale S E, Hansen J H L. Generating stressed speech from neutral speech using a modified CELP vocoder [J]. *Speech Communication*, 1996, **20**(1-2): 93-110.

[9] Bou-Ghazale S E, Hansen J H L. HMM-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress [J]. *IEEE Transactions on Speech and Audio Processing*, 1998, **6**(3): 201-216.

[10] 李雅，潘诗峰，陶建华. 采用重音调整模型的 HMM 语音合成系统[J]. 清华大学学报(自然科学版), 2011, **51**(9): 1171-1175.

LI Ya, PAN Shifeng, TAO Jianhua. HMM-based expressive speech synthesis with a flexible Mandarin stress adaptation model [J]. *J Tsinghua Univ (Sci and Tech)*, 2011, **51**(9): 1171-1175. （in chinese）

[11] Black A W, Taylor P, Caley R. The festival speech synthesis system [Z/OL]. [2012-07-15]. http://www.festvox.org/festival/.

[12] MENG Fanbo, WU Zhiyong, MENG Helen, et al. Generating emphasis from neutral speech using hierarchical perturbation model by decision tree and support vector machine [C]// ICALIP 2012. Shanghai, China: IEEE Press, 2012: 442-448.

[13] Kawahara H, Masuda-Katsuse I, Cheveigné A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds [J]. *Speech Communication*, 1999: **27**(3–4), 187–207.