基于融合特征的短语音汉语声调自动识别方法

沈凌洁，王 蔚
(南京师范大学教育科学学院，江苏南京 210097)

摘要：提出一种基于韵律特征(基频、时长)和梅尔倒谱系数(Mel-Frequency Cepstral Coefficient，MFCC)特征的融合特征进行短语音汉语声调识别的方法，旨在利用两种特征的优势提高短语音汉语声调识别率。该融合特征包括 7 个根据不同模型得到的韵律特征和统计参数以及 4 个从每个音段的梅尔倒谱系数计算得来的对数化后验概率，使用高斯混合模型表示 4 个声调的倒谱特征的分布。实验分两步：第一步，将基于韵律特征和倒谱特征的分类器在决策阶段混合起来进行声调分类，分别赋予两个分类器权重，计算倒谱特征和韵律特征在声调分类任务中的权重；第二步，将基于字的韵律特征和基于帧的倒谱特征结合起来生成融合特征的超向量，使用融合特征进行汉语声调识别，根据准确率、未加权平均召回率(Unweigted Average Recall，UAR)和科恩卡帕(Cohen's Kappa)系数 3 个指标，比较并评估 5 种分类器(两种设置的高斯混合模型，后向传播神经网络，支持向量机和卷积神经网络(Convolutional Neural Network，CNN))在不平衡数据集上的分类效果。实验结果表明：(1) 倒谱特征方法能够提高汉语声调的识别率，该特征在总体分类任务中的权重为 0.11；(2) 基于融合特征的深度学习(CNN)方法对声调的识别率最高，为 87.6%，与高斯混合模型的基线系统相比，提高了 5.87%。该研究证明了倒谱特征法能够提供与韵律特征法互补的信息，从而提高短语音汉语声调识别率；同时，该方法可以运用到韵律检测和副语言信息检测等相关研究中。

关键词：韵律特征；倒谱特征；梅尔倒谱系数；短语音声调；声调分类；融合；卷积神经网络
中图分类号：H107 **文献标识码**：A **文章编号**：1000-3630(2018)-02-0167-08
DOI 编码：10.16300/j.cnki.1000-3630.2018.02.013

Fusion feature based automatic Chinese short tone classification

SHEN Ling-jie, WANG Wei
(*College of Education Science*, *Nanjing Normal University*, *Nanjing* 210097, *Jiangsu*, *China*)

Abstract: This study proposes an approach to automatically recognizing short Chinese tone based on the fusion of prosodic and cepstral features to improve the recognition rate of Chinese tone. The fused features include seven prosodic features and their statistic parameters based on different models as well as four MFCC log posterior probabilities calculated from four Gaussian mixture models (GMM). Experiments have two steps: First, the classifiers based on prosodic features and cepstral features are combined to classify tone, and both of the two classifiers are given weights to examine the functions of prosodic features and cepstral features on tone classification; Second, seven reduced prosodic features based on different models and four log posterior probabilities obtained from frame-level MFCC which are modeled by Gaussian mixture model are concatenated into a fusion feature. Then, the tone classification performances of five classifiers, namely GMM with two configurations, back propagating neural network (BPNN), support vector machine (SVM) and convolutional neural network (CNN), are compared and evaluated with three indicators of accuracy, unweighted average recall (UAR) and Cohen's Kappa coefficient. Results show that: (1) Cepstral feature method can improve the recognition rate of Chinese tone classification and the weight of the features in the overall tone classification is 0.11; (2) Deep learning method of CNN using fused features outperforms other classifiers with a recognition rate of 87.6%, which is improved by 5.87% compared with the GMM baseline system. This study indicates that cepstral features provide complementary information to tone classification and hence improve the recognition rate. This new method could also be applied to other relevant researches on prosody detection and paralinguistic information detection.

Key words: prosodic feature; cepstral feature; Mel-Frequency Cepstral Coefficient (MFCC); short tone; tone classification; fusion; convolutional neural network

0 引 言

在汉语中，有四种声调，分别是阴平、阳平、上升和去声。这四种声调与元音、辅音结合起来成

收稿日期: 2017-05-06; **修回日期**: 2017-08-12
基金项目: 国家社会科学基金教育学一般项目(BCA150054)资助。
作者简介: 沈凌洁(1992－), 女, 江苏常州人, 硕士研究生, 研究方向为人机交互、机器学习等。
通讯作者: 王蔚, E-mail: wangwei5@njnu.edu.cn

为汉语的三个必要成分。汉语是一种单音节结构的语言，每个字由一个音节和一个声调表示，代表不同的语义。因此，声调对于汉字的区分起着重要的作用。

由于汉语是声调语言，因此发音的准确性不仅与每个音节相关，还与声调相关。在噪声环境中，语言的声调信息可以帮助提高汉语语音识别的准确性[1-2]。在 0 dB 信噪比环境下，给予正确声调信息的语音其识别率非常高，但当声调信息去除后，其识别率降低到 70%以下[1]。在汉语计算机辅助语言学习(Computer-Assisted Language Learning，CALL)领域中，声调的识别和评价是系统的重要组成部分，Qu 等[3]提出一种声调测评的混合方法，文献[4-6]表明，利用声调和重音等信息可以检测抑郁症及相关的疾病。

然而，声调识别是语音识别的子问题，仍有问题没有解决。例如，在连续的语音中，相邻字的声调会互相作用从而影响声调的识别率，较短的字词的声调识别也具有挑战性[3]。

传统的声学特征主要有韵律特征(基频、能量、时长)、音质特征(基频微扰 jitter、振幅微扰 shimmer)和时频特征(梅尔倒谱系数、线性预测倒谱系数)。韵律特征最能体现语音的副语言信息，因此是最常用的声调识别特征[7-8]。不同的声调通常由不同的基频曲线表示。图 1 展示了 4 个不同声调的频谱以及基频曲线，基频通常用 F0 来表示。图 1 中 4 种声调的频谱图在窄频带下绘制出来，不同的灰度代表相应的频率的能量值，颜色越深，能量越大。黑色线代表 F0 曲线，由自相关算法得出，该图来源于文献[9]。除了声调的基频特征，其他的声学特征如持续时间、声强等同样可以辅助进行声调识别[10]。

由于基频能够体现声调的变化，因此它成为研究声调识别与分类的主流特征。在有关语音识别等的任务中，倒谱特征被认为是一种鲁棒性较强的特征，尤其是梅尔倒谱系数(Mel-Frequency Cepstral Coefficient，MFCC)[11]，它能够很好地模拟人耳的听力特性，因而成为语音识别中广泛使用的特征。相关研究表明[12]，声调不仅与人的声带相关(通常由基频表示)，还与声道的振动相关。声道信息通常由频谱特征表示，它与人的生理特性相关，代表了声道的大小和长度，因而能影响不同的发音。因此，将频谱特征与韵律特征结合起来能更好地进行声调识别，提高声调识别率。然而很少有将两种特征结合起来进行汉语声调识别的研究。

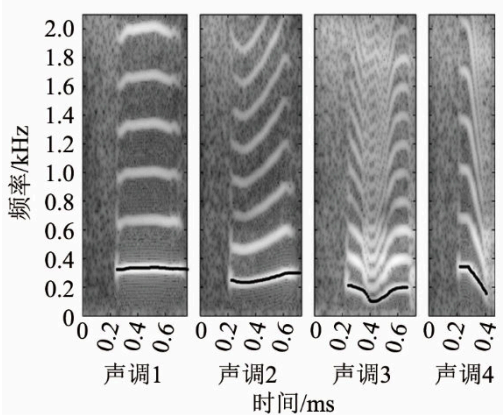

图 1 由一个女性表达的“shi”汉语音节的 4 个声调的频谱图和 F0 曲线[9]

Fig.1 Spectrum and F0 curves of the four tones of the Chinese syllable "shi" expressed by a woman[9]

该研究的目的是将韵律特征(基频、时间)和倒谱特征结合起来提高短语音汉语声调的识别率。研究分为两部分：第一部分，通过实验验证了倒谱特征(MFCC)能够提高汉语声调的识别率，并且计算该特征在声调识别中的权重；第二部分，分析比较基于融合特征的 5 个分类器在不平衡数据上的分类效果。该研究使用了两种设置的高斯混合模型、神经网络，支持向量机和卷积神经网络，比较准确率、未加权平均召回率((Unweighted Average Recall，UAR)和科恩卡帕系数 3 个指标。

该研究提出了一个将基于超音段的韵律特征和基于帧的倒谱特征结合起来的方法来提高短语音汉语声调识别率。首先在特征选择上，将韵律特征和倒谱特征结合起来提高汉语声调识别率。其次，基于融合特征，选择不同的算法提高在不平衡数据上声调的识别率。该研究的创新之处有以下几个方面：

(1) 特征选择的方法：将基于不同统计特性的、不同模型的以及其他文献提出的有效的韵律特征结合起来，使用顺序前向特征选择(Sequence Forward Feature Selecture，SFFS)方法，提取出对该数据库有效的特征，减少数据的冗余度，从而简化算法。同时，针对不同的问题和数据，该方法对其他相关研究如副语言信息挖掘、情感识别等相关问题研究有较好的泛化能力。

(2) 特征融合：该研究通过早期融合的方法将基于字的韵律特征和基于帧的频谱特征融合起来形成超向量。该超向量融合了语音的韵律信息和生理信息，被应用于不同的分类器，从而提高汉语声调识别率。

(3) 解决问题：汉语声调识别的相关研究目前已经相当成熟，但大多数研究仅关注发音清晰、时间较长、音质较高的语音信息，而短时语音的汉语

声调识别仍然具有一定的挑战。该研究聚焦短语音的汉语声调识别，从特征和分类器的角度进一步提高声调识别率。

1　相关研究

对于近年来汉语声调识别的研究情况，相关文献详见表 1。该表体现了近 10 年汉语声调识别在算法和特征提取上的变化，人们不仅仅只使用基频特征等超音段特征，还关注倒谱特征在声调识别中的作用。虽然已有研究表明倒谱特征能够很好地进行汉语声调识别，但是相关研究基于不同的数据库，识别率不能进行绝对的比较，同时，相关研究没有指出频谱特征和韵律特征对汉语声调识别的贡献率。关于声调模型的研究，目前已有三种基频曲线模型，分别为 Tilt 模型，Bézier 模型，量化轮廓模型(Quantized Contour Model，QCM)等等[13]。本研究的启发来自于文献[14]，其创新性地使用基频 F0，MFCC 和 Frequency Modulation 特征进行越南语声调分类，研究表明，与只使用韵律特征的分类方法相比，将倒谱特征和韵律特征结合起来进行分类的方法，准确率提高 7.5%，并指出声调语言如汉语、粤语等都可以使用相似的方法。基于前人的研究，本文尝试使用韵律特征和倒谱特征相结合的方法进行短语音汉语声调识别，验证该方法在汉语声调识别中的可行性。

表 1　汉语声调的相关研究

Table 1　Recent researches on Chinese tone classification

作者	算法	特征	准确率/%
C. Chen, Bunescu, Xu, & Liu (2016)[9]	卷积神经网络	梅尔倒谱系数	95.5
Ryant, Yuan, & Liberman (2014)[15]	深度神经网络	梅尔倒谱系数，基频	82.27
Hu et al. (2014) Wu et al. (2013)[16,17]	隐马尔可夫模型	离散余弦变换系数，离散余弦级数系数	60.3
Qu et al. (2011)[3]	隐马尔可夫模型	梅尔倒谱系数	86.6
Zhou (2008)[18]	人工神经网络	基频	85.0
Zhaojie et al. (2007)[19]	高斯混合模型	基频，时长	65.9
Xin et al. (2006)[20]	神经网络	基频，时长	76.2
Cao et al. (2004)[21]	决策树/隐马尔可夫模型	基频	69.1

由于汉语声调的数据不平衡，因此解决不平衡数据对分类结果产生的影响是声调分类任务不得不面临的一个问题。文献[18]列举了不平衡数据对最终结果带来的消极影响，指出在不平衡数据下分类器倾于向将样本分为最多样本数所属的那个类别。为了减少不平衡的声调数据带来的消极影响，相关研究进行了多种实验，如过采样实验、欠采样实验或整体采样实验等[19]。解决不平衡数据的方法主要分为两类，分别为基于算法和基于数据的两个层级[20]。第一种方法采用新的算法或者对已有算法进行改进来解决问题。第二种方法对较少数据的类别进行多次采样、过采样，或对较多数据的类别进行欠采样。该文章采用不同算法来解决不平衡数据带来的影响，获得了较好的总体分类效果。

2　本文提出的方法

2.1　特　征

2.1.1　韵律特征

使用 praat[22]软件提取每个短时汉字语音段的基频特征，该软件默认提取基频值的方法为自相关法[23]。每个语音段的基频特征将从该语音段的基频以及它的一阶、二阶差分中提取，使用 z-score 进行标准化。这些特征如下：

(1) 基本统计量：最大值 M_1，最小值 M_2，最大值对应的时间 T_1，最小值对应的时间 T_2，$|T_1-T_2|$，$(M_1-M_2)/|T_1-T_2|$，平均值 m，标准差 S_1，偏度 S_2，峰度 S_3，上四分位数 Q_1，中位数 m_1，下四分位数 Q_3，四分位差 I (interquartile range)，四分位差与标准差之差的绝对值 $|I\text{-}S_1|$，开始时刻的值 f_1，中间时刻的值 f_2，结束时刻的值 f_3，f_2 与 f_1 的差的绝对值，f_3 与 f_1 的差的绝对值，f_3 与 f_2 的差的绝对值 (22 个特征)；

(2) 基于 Tilt 模型[24]的特征：基频上升值，基频上升时间，基频下降值，基频下降时间，基频上升时间和下降时间的总和，基频上升值和下降值的总和，Tilt 值，一共 7 个特征；

(3) 文献[25-26]提出的特征：基频上升的平均量 f_4，下降的平均量 f_5，上升次数的百分比 f_6，下降次数的百分比 f_7，一次拟合的一次项系数和常数项系数 C_1, c_1，2～7 次拟合的最高次数项的系数 C_2～C_7(12 个特征)。

由于这些特征代表和区分声调的能力各不相同，因此该研究使用过滤式特征选择方法(RELIEFF)算法[27]评估不同特征在分类任务中的区分性和代表性，按这些特征对分类任务的贡献率从高到低进行排序并产生相应的权重。然后使用顺序

前向算法(Sequence Forward Feature Selection，SFFS)进行特征的筛选，将这些排序好的特征依次投入相应的分类器，并只保留能提高分类效果的特征，分类器采用 KNN (k-Nearest Neighbor) 算法。最后，根据以上方法，保留 7 个特征(基频曲线一次拟合的一次项系数 C_1，上升次数的百分比 f_6，下降次数的百分比 f_7，最大值与结束时刻值之差的绝对值 $|M_1-f_3|$，最大值与开始时刻值之差的绝对值 $|M_1-f_1|$，结束时刻值 f_3，最大时刻与结束时刻之差 $|T_1-f_3|$。

2.1.2 倒谱特征

在提取 MFCC 时，首先进行语音信号的预处理，设置帧长为 20 ms，帧移为 10 ms，窗函数为汉明窗。然后进行声音活动检测(Voice Activity Detection，VAD)，去除无声段。每帧提取 24 维的特征向量，包括 12 个 MFCC 和它的一阶差分(ΔMFCC)。每段语音的倒谱使用倒谱均值相减法(Cepstral Mean Substraction，CMS)进行标准化。

2.1.3 融合特征

利用每帧的 MFCC 特征训练分别代表 4 个声调的高斯混合模型(Gaussian Mixture Model，GMM)，计算每段语音的 MFCC 在这 4 个 GMM 上的对数化后验概率，将 7 个韵律特征和这 4 个 MFCC 的对数化后验概率结合起来形成 11 维的融合特征[28]，如图 2 所示。该研究使用 10 折交叉验证的方法计算每段语音段的 4 个 MFCC 对数化后验概率。

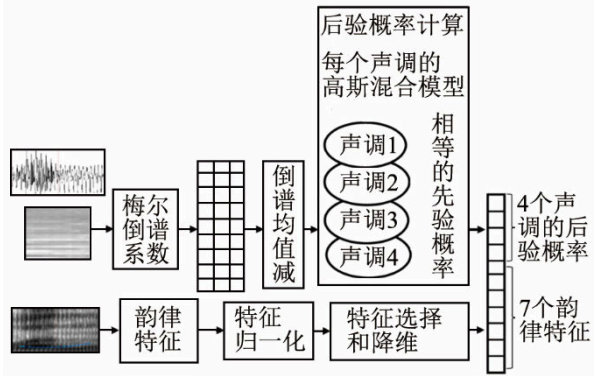

图 2 融合特征的生成流程

Fig.2 Diagram of fusion features' generation

2.2 实验设计

2.2.1 实验一：将基于韵律特征和基于倒谱特征的分类器混合，计算两种特征的权重

为了探究韵律特征和倒谱特征对声调分类任务的贡献率，证明倒谱特征能提高声调识别率，分别使用韵律特征和倒谱特征进行分类，并赋予两个分类器权重，探究在该权重变化的情况下，基于韵律特征和基于倒谱特征的混合分类器的声调识别率的变化。两个分类器为高斯混合模型(GMM)。训练韵律特征的 GMM 使用 8 个成分，训练倒谱特征的 GMM 选择 32 个成分。测试语音根据韵律特征和倒谱特征识别出的声调分别为 T^*_{pitch}，T^*_{MFCC}，它们分别由两个分类器计算得到的后验概率 $P_{\text{pitch}}(T_n|\boldsymbol{S}_i)$ 和 $P_{\text{MFCC}}(T_n|\boldsymbol{X}_i)$ 中得到，$T_n(n=1,2,3,4)$ 分别表示声调 1、声调 2、声调 3、声调 4。

$$T^*_{\text{pitch}}=\arg\max_{1\leqslant n\leqslant 4}\left[P_{\text{pitch}}(T_n|\boldsymbol{S}_i)\right] \tag{1}$$

$$T^*_{\text{MFCC}}=\arg\max_{1\leqslant n\leqslant 4}\left[P_{\text{MFCC}}(T_n|\boldsymbol{X}_i)\right] \tag{2}$$

其中：$\boldsymbol{S}_i$ 为样本 i 的 7 个韵律特征构成的特征向量，$1\leqslant i\leqslant N$，N 为样本总数，$\boldsymbol{X}_i=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_K\}$，$K$ 为样本 i 的帧数，$\boldsymbol{x}_1$ 为第一帧语音的 MFCC 特征向量，该研究假设每个声调的先验概率相等($P(T_n)=\frac{1}{4}$)，并且对于每个语音段，$P(\boldsymbol{S})$和 $P(\boldsymbol{X})$都是相等的。因此，可以将 T^*_{pitch} 和 T^*_{MFCC} 近似表示为

$$\hat{T}^*_{\text{pitch}}=\operatorname{argmax}_{1\leqslant n\leqslant 4}\left[P(\mathbf{S}_i|T_n)\right] \tag{3}$$

$$\hat{T}^*_{\text{MFCC}}=\operatorname{argmax}_{1\leqslant n\leqslant 4}\left[P(\boldsymbol{S}_i|T_n)\right]=\operatorname{argmax}_{1\leqslant n\leqslant 4}\left[\prod_{k=1}^{K}P(\boldsymbol{x}_k|T_n)\right]\cong \operatorname{argmax}_{1\leqslant n\leqslant 4}\left[\sum_{k=1}^{K}\lg P(\boldsymbol{x}_k|T_n)\right] \tag{4}$$

在汉语声调分类任务中，韵律特征和倒谱特征是两种性质完全不同的特征，对于声调分类的贡献程度也不相同，因此该研究将这两种不同的分类器混合起来，探究两种不同的特征是否能改善声调分类的准确率。尽管有许多混合不同分类器的方法[29-31]，研究使用两个分类器的后验概率加权和的方法[29]：

$$T^*=\arg\max_{1\leqslant n\leqslant 4}\left[\alpha\times P(T_n|\boldsymbol{S}_i)+(1-\alpha)\times P_{\text{MFCC}}(T_n|\boldsymbol{X}_i)\right] \tag{5}$$

该方法能够体现不同分类器对整体声调识别的贡献程度。该研究使用两种特征，即韵律特征和倒谱特征，韵律特征对整体分类效果的贡献程度为 α ($0\leqslant\alpha\leqslant 1$)，因此，倒谱特征对整体分类效果的贡献程度为 $1-\alpha$。为了检验这两个分类器的相似程度，计算了 Q 统计量[32]：

$$Q_{\text{pitch,MFCC}}=\frac{N^{00}N^{11}-N^{01}N^{10}}{N^{00}N^{11}+N^{01}N^{10}} \tag{6}$$

式中：N^{00} 表示两个分类器都识别错误的个数；N^{11} 表示两个分类器都识别正确的个数；N^{10} 表示第一个分类器分类正确的同时第二个分类器分类错误的个数；N^{01} 表示第一个分类器分类错误的同时第二个分类器分类正确的个数。Q 统计量介于[–1,1]之

间，Q 值越接近 0，两个分类器的分类效果越相近，反之，Q 值越接近 1 或–1，两个分类器的分类效果越不同。

2.2.2　实验二：将韵律特征和倒谱特征混合，比较 4 个分类器的识别结果

在验证了倒谱特征能提高汉语声调的识别率之后，使用融合特征，从算法的水平上提高汉语声调识别率。将 7 个韵律特征和这 4 个 MFCC 的对数化后验概率结合起来形成 11 维的融合特征，如图 2 所示。

由于实验来自 4 个声调的数据量相差较大，使用不同的分类器来比较它们在不平衡数据下的分类表现。使用如下 4 种分类器，分别为两种设置的 GMM、后向传播神经网络(Back Propagating Neural Network，BPNN)、支持向量机、卷积神经网络。

(1) 高斯混合模型(GMM)：使用该融合特征分别训练 4 个高斯混合模型，对应 4 个不同的声调。GMM 分类器有两种设置，一个称为小 GMM，即训练 MFCC 时使用 16 个成分，训练该 11 维融合特征时使用 4 个成分，训练仅基于韵律特征的模型使用 4 个成分，另一个称为大 GMM，即训练 MFCC 时使用 32 个成分，训练该 11 维融合特征时使用 8 个成分，训练仅基于韵律特征的模型使用 8 个成分。

(2) 后向传播神经网络(BPNN)：该网络的拓补结构为 11*10*4，有 10 个隐藏节点，隐藏节点的激活函数为 sigmoid，输出层的激活函数为 softmax，调整 BP 网络参数的方式为自适应、有动量的梯度下降法。选择概率最大的那个节点对应的声调作为该语音的类别。

(3) 支持向量机：支持向量机(Support Vector Machine，SVM)[33]能够将具有高维特征的两类数据进行较好的分类和判别，是一种判别性分类器。本研究需要解决的是多类别问题，因此分别设计 6 个 SVM，测试时将测试数据分别投入 6 个样本，分类器分别对该样本进行类别投票，投票结果最多的那一类即为该测试样本的类别。当出现一个以上相同的最大票数时，选取第一个出现的最大值的那一类作为该测试样本的类别。SVM 选取高斯径向基核函数。

(4) 卷积神经网络(Convolutional Neural Network，CNN)：卷积神经网络的拓扑结构为 11*8*4，使用一维卷积层，卷积核大小(kernal size)为 3，滤波器个数为 32，使用最大池化层(max-pooling)，激活函数为修正线性单元(Rectified Linear Unit，ReLU)[34]，优化器为 Adam[35]，学习率为 10^{-4}，迭代次数为 10。该实验在 keras 平台上实现。

该实验的基线系统为基于韵律特征并仅使用 4 个成分的高斯混合模型(GMM)。

3　实验与结论

实验分为两步。第一步，探究韵律特征和倒谱特征对声调分类的贡献程度，验证倒谱能提高汉语声调的识别率。第二步，将两种特征混合起来，利用 5 个分类器进行声调识别，比较不同分类器在不平衡数据上的表现。

3.1　数据描述

语音数据来自中国科学院自动化研究所疑问句语料库，该语料库中语料的采样频率为 16 kHz，精度为 16 bit。该语料库由两男两女朗读，每人朗读相同的 590 句。使用隐马尔可夫工具 (HMM toolkit，HTK)进行字词切分，请 5 个本科生对标注好的数据进行筛选，挑选时长较短的语音。收集到的 4 个声调的数据分布见表 2。

该数据集被分为训练集和测试集两部分，为了避免局部最优的试验结果，使用 10 折交叉验证进行训练和测试。

表 2　数据分布
Table 2　Data distribution

	声调 1	声调 2	声调 3	声调 4	踪迹
男	534	414	75	969	1 992
女	643	804	282	1 632	3 361
总计	1 177	1 218	357	2 601	5 353

3.2　评价指标

用来评价算法分类效果的指标有三个，分别为准确率、未加权平均召回率(UAR)和科恩卡帕系数(κ)[36]。准确率用来评估总的分类准确率；未加权平均召回率用来评估每一类的准确率的均值，它对待每一类的错误率给予相同的权重，因而能更客观地评价基于不平衡的数据集下算法的分类表现；科思卡帕系数 κ 用来评估人和机器对声调的识别的一致程度。未加权平均召回率(UAR)指标用于体现在不平衡数据上的表现，其定义为

$$P_{\mathrm{UAR}}=\frac{1}{N}\sum_{i=1}^{N}\frac{c_i}{n_i} \tag{7}$$

式中：c_i 表示被正确划分为类别 i 的个数；n_i 表示类别为 i 的样本数；N 表示类别数。

3.3　实验结果

3.3.1　实验一：将基于韵律特征和基于倒谱特征的分类器混合，计算两种特征的权重

为了探究基频特征和 MFCC 特征对声调识别的贡献率，将两种特征的 GMM 分类器混合起来，识别结果见图 3。

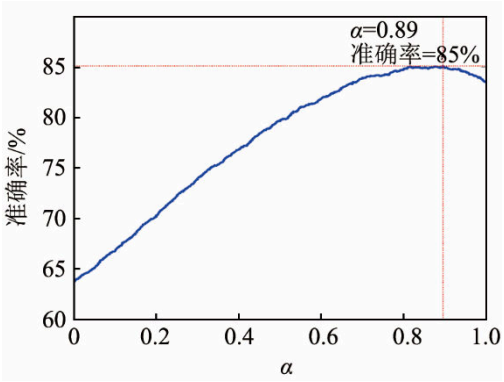

图 3 混合分类器的声调识别率

Fig.3 Fusion recognition rate as a function of weight α attributed to both prosodic (α=1) and spectral (α=0) classifiers

图 3 中，基于韵律特征的分类器权重为 α，基于倒谱特征分类器的权重为 1–α。由图 3 可知，当基于韵律特征的 GMM 分类器的权重 α 为 0.89，基于倒谱特征的 GMM 分类器权重 1–α 为 0.11 时，声调分类的准确率最高，为 85%。在该研究中，两个分类器的 Q 值等于 0.229 5，表明韵律特征和倒谱特征在声调分类任务中能够提供互补的信息。由此证实，韵律特征(主要是基频)仍然是声调分类的主要特征，但倒谱特征可以在一定程度上提高声调分类的识别率。

3.3.2 实验二：将韵律特征和倒谱特征混合，比较 4 个分类器的识别结果

实验结果见表 3。图 4 是基于基线系统和卷积神经网络算法在声调识别上的混合矩阵，灰度值表示正确识别每一种声调的百分比，图 4 中，基线系统(a)和卷积神经网络(b)的声调识别率。j 行 k 列的值表示本属于声调 j 的样本却被误分为声调 k 的比例。(j=1,2,3,4；k=1,2,3,4)。

表 3 基线系统与基于融合特征的 5 个分类器的分类结果

Table 3 Classification results of baseline system and 5 classifiers with fusion features

分类器	识别率/%	P_{UAR}/%	κ
高斯混合模型(基线系统)	81.73	76.47	0.800
高斯混合模型(成分为 32 和 8)	84.55	76.63	0.832
高斯混合模型(成分为 16 和 4)	83.07	76.02	0.815
后向传播神经网络	86.28	76.22	0.851
支持向量机	85.50	77.60	0.842
卷积神经网络	87.60	81.54	0.869

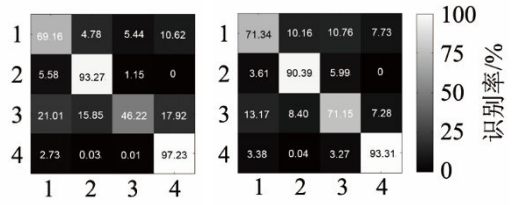

图 4 声调识别混合矩阵

Fig.4 Tone classification confusion matrices: the tone recognition error patterns of GMM baseline system (left) and CNN (right)

从表 3 和图 4 可以得到如下结果：

(1) 与基线系统相比较起来，基于融合特征的卷积神经网络分类器的准确率提高了 5.87%；

(2) 卷积神经网络的准确率最高，为 87.6%。除了用卷积神经网络的方法之外，实验结果表明，神经网络的准确率最高，其次是支持向量机。并且，在不平衡数据上，支持向量机的表现仅次于卷积神经网络。支持向量机的优势是能够利用有限的数据生成较好的决策面，从而获得较优的识别率[33]；

(3) 在该实验中，判别性模型(SVM)比生成性模型(GMM)表现好[8]，这是因为判别性模型能够对潜在的变量进行分类并生成较好的决策面，从而判别数据的类别。

4 讨论和总结

本研究验证了倒谱特征对短时声调识别的作用。实验结果表明，虽然韵律特征在声调识别中仍然起到重要的作用，但是由于倒谱特征能够获取韵律所不能表达的特征，能提供与韵律信息互补的代表生理特性的频谱特征，因此能够提高汉语声调的识别率，文献[13, 37]研究的结果证实了这一点。

根据上述实验结果，进一步将韵律特征和倒谱特征融合起来进行短语音汉语声调分类。分别比较在不平衡数据上基于融合特征的高斯混合模型、神经网络、支持向量机和卷积神经网络的分类效果，结果表明卷积神经网络能够获得最高的识别率。

在进行韵律特征的筛选与降维时，采取与文献[38]类似的方法，即针对每个特征的分类能力从高到低进行排序，虽然得到的特征不完全一致，但大致都是描述基频曲线走势的特征。

然而，为了能够充分证明本文提出的短语音汉语声调分类方法的泛化能力，今后还需要在其他数据库上进行实验。

本研究从特征提取的角度来提高短语音汉语声调的识别率。随着近年来深度学习的快速发展和

其显著的分类能力[39]，该研究未来可以进一步从算法角度提高汉语声调的识别率，将深度神经网络(Deep Neural Network，DNN)、循环神经网络(Recurrent Neural Network，RNN)，短长时记忆(Long Short Term Memory，LSTM)神经网络等深度神经网络同样可以应用到相关的研究中[14,15]。此外，由于该方法同时涉及到音段特征和超音段特征，因此还可以将类似的方法泛化到汉语重音检测与评价、汉语韵律的检测与评价、副语言信息的检测与分类等相关的研究中，扩大该方法的应用范围。

5　结 论

该研究通过将韵律特征和倒谱特征结合起来进行汉语声调识别，使用深度学习方法(CNN)和传统机器学习方法进行分类，实验结果表明将韵律特征和倒谱特征结合起来能显著提高传统基于韵律特征的声调识别率，基于深度学习(CNN)的声调识别效果最好。该研究方法和研究思路可以进一步扩展到语音情感识别、副语言信息检测与识别等相关研究中，今后将进一步探究相关深度学习方法来提高语音声调识别。

参 考 文 献

[1] CHEN F, WONG L L N, HU Y. Effects of lexical tone contour on Mandarin sentence intelligibility[J]. Journal of Speech, Language, and Hearing Research, 2014, **57**(1): 338-345.

[2] WANG J, SHU H, ZHANG L, et al. The roles of fundamental frequency contours and sentence context in mandarin chinese speech intelligibility.[J]. Journal of the Acoustical Society of America, 2013, **134**(1): EL91-97.

[3] QU Y, HE X, LU Y, et al. A hybrid method of tone Assessment for mandarin CALL system[M]. Pattern Recognition, Machine Intelligence and Biometrics. Springer Berlin Heidelberg, 2011: 61-80.

[4] PAUL R, AUGUSTYN A, KLIN A, et al. Perception and production of prosody by speakers with autism spectrum disorders[J]. Journal of Autism & Developmental Disorders, 2005, **35**(2): 205-220.

[5] RINGEVAL F, DEMOUY J, SZASZAK G, et al. Automatic intonation recognition for the prosodic assessment of language-impaired children[J]. IEEE Transactions on Audio Speech & Language Processing, 2011, **19**(5): 1328-1342.

[6] DIEHL J J, PAUL R. Acoustic differences in the imitation of prosodic patterns in children with autism spectrum disorders[J]. Research in Autism Spectrum Disorders, 2012, **6**(1): 123-134.

[7] GONZALEZ-FERRERAS C, ESCUDERO-MANCEBO D, VIVARACHO-PASCUAL C, et al. Improving automatic classification of prosodic events by pairwise coupling[J]. IEEE Transactions on Audio Speech & Language Processing, 2012, **20**(7): 2045-2058.

[8] SRIDHAR V K R, BANGALORE S, NARAYANAN S S. Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework[J]. IEEE Transactions on Audio Speech & Language Processing, 2008, **16**(4): 797-811.

[9] CHEN C, BUNESCU R, XU L, et al. Tone classification in mandarin chinese using convolutional neural networks [C]//Conference of the International Speech Communication Association. 2016.

[10] SURENDRAN D R. Analysis and automatic recognition of tones in mandarin chinese[D]. The University of Chicago, 2007.

[11] DAVIS S, MERMELSTEIN P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. IEEE transactions on acoustics, speech, and signal processing, 1980, **28**(4): 357-366.

[12] ERICKSON D, IWATA R, ENDO M, et al. Effect of tone height on jaw and tongue articulation in Mandarin Chinese[C]//International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages. 2004.

[13] JOHNSON D O, KANG O. Automatic prosodic tone choice classification with Brazil's intonation model[J]. International Journal of Speech Technology, 2016, **19**(1): 95-109.

[14] LE P N, AMBIKAIRAJAH E, CHOI E H C. Improvement of Vietnamese Tone Classification using FM and MFCC Features [C]//International Conference on Computing and Communication Technologies. IEEE, 2009: 1-4.

[15] RYANT N, YUAN J, LIBERMAN M. Mandarin tone classification without pitch tracking[C]//ICASSP 2014-2014 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2014: 4868-4872.

[16] WU J, ZAHORIAN S A, HU H. Tone recognition for continuous accented Mandarin Chinese[C]//Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 7180-7183.

[17] HU H, ZAHORIAN S A, GUZEWICH P, et al. Acoustic features for robust classification of Mandarin tones[C]//INTERSPEECH. 2014: 1352-1356.

[18] ZHOU N, ZHANG W, LEE C Y, et al. Lexical tone recognition with an artificial neural network[J]. Ear & Hearing, 2008, **29**(3): 326-335.

[19] LIU Z J, SHAO J, ZHANG P Y, et al. Research on tone recognition in Chinese spontaneous speech[J]. Acta Physica Sinica, 2007, **56**(12): 7064-7069.

[20] XIN L, SIU M H, HWANG M Y, et al. Improved tone modeling for Mandarin broadcast news speech recognition.[C]//INTERSPEECH 2006-Icslp, Ninth International Conference on Spoken Language Processing, Pittsburgh, Pa, Usa, September. DBLP, 2006.

[21] 曹阳，黄泰翼，徐波，等. 基于统计方法的汉语连续语音中声调模式的研究[J]. 自动化学报, 2004, **30**(2):191-198.
CAO Yang, HUANG Taiyi, XU Bo, et al. A stochastically-based study on Chinese tone patterns in continuous speech[J]. Acta Automatica Sinica, 2004, **30**(2): 191-198.

[22] Boersma P, Weenink D. Praat: Doing phonetics by computer[J]. Ear & Hearing, 2011, **32**(2): 266.

[23] MEI X D, PAN J, SUN S H. Efficient algorithms for speech pitch estimation[C]//Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on. IEEE, 2001: 421-424.

[24] TAYLOR P. Analysis and synthesis of intonation using the tilt model[J]. The Journal of the acoustical society of America, 2000, **107**(3): 1697-1714.

[25] VU M Q, BESACIER L, CASTELLI E. Automatic question detection: prosodic-lexical features and crosslingual experiments [C]//INTERSPEECH 2007, Conference of the International Speech Communication Association, Antwerp, Belgium, August. DBLP, 2007: 2257-2260.

[26] MA M, EVANINI K, LOUKINA A, et al. Using f0 contours to assess nativeness in a sentence repeat task[C]//Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[27] ROBNIK-Šikonja M, KONONENKO I. Theoretical and empirical analysis of relieff and rreliefF[J]. Machine Learning, 2003, **53**(1): 23-69.

[28] FERRER L, BRATT H, RICHEY C, et al. Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems[J]. Speech Communication, 2015, **69**: 31-45.

[29] KUNCHEVA L I. Combining pattern classifiers: methods and algorithms[J]. Technometrics, 2005, **47**(4): 517-518.

[30] MONTE-MORENO E, CHETOUANI M, FAUNDEZ-ZANUY M, et al. Maximum likelihood linear programming data fusion for speaker recognition[J]. Speech Communication, 2009, **51**(9): 820-830.

[31] JAIN A, NANDAKUMAR K, ROSS A. Score normalization in multimodal biometric systems[J]. Pattern recognition, 2005, **38**(12): 2270-2285.

[32] YILDIRIM S, NARAYANAN S. Automatic detection of disfluency boundaries in spontaneous speech of children using audio-visual information[J]. IEEE Transactions on Audio Speech & Language Processing, 2009, **17**(1): 2-12.

[33] BURGES C J C. A tutorial on support vector machines for pattern recognition[J]. Data mining and knowledge discovery, 1998, **2**(2): 121-167.

[34] GLOROT X, BORDES A, BENGIO Y. Deep sparse rectifier neural networks[C]//International Conference on Artificial Intelligence and Statistics. 2012.

[35] KINGA D, ADAM J B. A method for stochastic optimization[C]//International Conference on Learning Representations (ICLR). 2015.

[36] COHEN J. A coefficient of agreement for nominal scales[J]. Educational & Psychological Measurement, 1960, **20**(1): 37-46.

[37] BAO W, LI Y, GU M, et al. Combining prosodic and spectral features for Mandarin intonation recognition[C]//International Symposium on Chinese Spoken Language Processing. IEEE, 2014: 497-500.

[38] HAN R, CHOI J Y. Prosodic boundary tone classifickation with voice quality features[J]. J. Acoust. Soc. Am., 2013, **133**(4): 1862-1866.

[39] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, **29**(6): 82-97.