文章编号:1000-582X(2009)07-0770-05

基于 MFCC 和 SVM 的说话人性别识别

肖汉光[1,2],何　为[1]
(1.重庆大学 输配电装备及系统安全与新技术国家重点实验室,重庆 400030;
2.重庆工学院 数理学院,重庆 400054)

摘　要:建立了普通话语音性别数据库,提出联合梅尔频率频谱系数(Mel-frequency Cepstrum Coefficients, MFCC)的特征提取方法和支持向量机(Support Vector Machine, SVM)的分类方法进行说话人性别识别,并与其它分类方法进行比较,实验结果表明该方法的说话人性别识别准确率达到 98.7%,明显优于其它分类器。

关键词:模式识别;分类器;性别识别;支持向量机;梅尔频率频谱系数
中图分类号: TP391.4　　文献标志码:A

Gender recognition of speakers based on MFCC and SVM

XIAO Han-guang[1,2], *HE Wei*[2]
(1. State Key Laboratory of Power Transmission Equipment & System Security and New Technology, Chongqing University, Chongqing 400030, P. R. China; 2. School of Mathematics and Physics, Chongqing Institute of Technology, Chongqing 400054, P. R. China)

Abstract: A Chinese speech (mandarin) database was established for speakers gender recognition. A combination method is proposed for gender recognition of speakers based on support vector machine and Mel-frequency cepstrum coefficients (MFCC) for classification and feature extraction respectively. The comparative result shows that the accuracy of SVM is 98.7%, which is better than other methods.

Key words: pattern recognition; classifiers; gender recognition; mel-frequency cepstrum coefficients; support vector machine

说话人性别识别是语音识别研究中的一个重要分支,即通过说话人的语音识别说话人的性别。性别识别作为说话人识别的预分类技术可以降低研究问题的复杂度,提高系统的准确率[1]。在人机交互系统中,可以根据说话人性别选择不同性别的语音应答,使系统服务更加友好和人性化。计算机可对短时(如几十毫秒)语音信号进行性别自动识别,是人耳无法相比的。

在生理和心理学方面,男性女性说话有明显的差异,如声带产生的基音、口腔结构(喉咽、舌、腭、唇、齿等)产生的共振峰频率、呼出气流的大小和强弱等。语音信号中包含说话人性别特征,这使得利用语音信号实现说话人的性别识别成为可能。目前常用方法是分析说话人语音的基音,以基音判断标准进行性别识别,得了较好的结果[2]。但是,在说话人的语音多样性增大时,单一特征和判据进行性别识别不易取得高准确率,而梅尔频率频谱系数(Mel-frequency Cepstrum Coefficients, MFCC)能体现说话人时域和频域空间中的差异,并被有效地应用于语音识别领域[3-4]。

收稿日期:2009-02-26
基金项目:国家自然科学基金资助项目(50877082);重庆工学院青年教师科研基金资助项目(20062D39)
作者简介:肖汉光(1980-),男,重庆大学博士研究生,主要从事机器学习、模式识别等研究。
何为(联系人),男,重庆大学研究员,博士生导师,(E-mail) hewei@cqu.edu.cn。

目前，常用于模式识别的方法包括：向量量化(Vector Quality，VQ)、决策树(Decision Tree，DT) C4.5、K 近邻分类器(K-nearest Neighbor，KNN)、神经网络(Neural Network，NN)和支持向量机(Support Vector Machine，SVM)。由于 VQ 方法原理简单、易于实现和识别准确率良好，所以常被应用于语音识别领域。在线性可分情况下，VQ 能得到较好的识别效果；在非线性可分情况下，特别是同性别说话人语音特征较复杂时，VQ 的准确率会有较大影响。为克服该缺点，笔者利用在非线性可分情况下仍有良好分类效果的支持向量机进行复杂语音情况下的性别识别。支持向量机是由 Vapnik 及其合作者基于结构风险最小化原理提出的一种有监督的统计学习方法，被公认为小样本情况下统计及其学习的经典[5-6]。由于其不需要确定各类的条件概率密度和先验概率就能找到全局最优解，并且具有较好的泛化能力，所以被广泛应用于诸多领域，如文本分类，手写体数字识别，图像识别与目标探测，水文预报，空气质量预报，地球空间物理和实验高能物理数据分析与处理，肿瘤及癌症诊断，基因微阵列表达数据分析，药物设计，蛋白质-蛋白质相互作用预测以及蛋白质结构与功能预测等[6-10]。

笔者利用代表说话人性别差异的梅尔频率频谱系数和支持向量机进行说话人性别识别，并将获得的结果和其它分类器进行比较。

1　分类原理

1.1　KNN 的原理

KNN 和其他分类方法相比是最简单但准确率较高的分类器。该方法遵从的假设为：同类样本在特征空间中距离相近，而异类的样本距离较远。若给定一待分类的 L 维样本 $\boldsymbol{x}'=(x_1',x_2',\cdots,x_L')$，计算其与训练样本$\{\boldsymbol{x}_i\}$(即已知类别的样本)的相似度或距离，如式(1)为待测样本与训练集中第 i 个样本欧氏距离：

$$S_i = \| \boldsymbol{x}' - \boldsymbol{x}_i \| 。 \tag{1}$$

由 K 个最相似或最接近的样本根据自身类别进行少数服从多数的投票决定待识别样本的类别。一般 K 取 1 到 N(N 为训练样本的样本数)。

1.2　PNN(probability neural network)的原理

PNN 是根据贝叶斯最优决策规则而设计的分类方法，由输入层、径向基层、比较层和输出层组成[11]。当待测样本输入到输入层，和径向基层的所有神经元进行运算，计算其与神经元的距离，神经元一般设定为训练集中的各样本。在比较层中进行距离比较，计算待测样本与所有正和负样本神经元的平均距离，若与正样本神经元的平均距离小于负样本神经元的平均距离，则输出为正类别，反之为负类别。实际 PNN 相当 K 为 N(N 为训练集的样本数)时的 KNN，但计算距离的表达式略有不同。式(2)为径向基层中计算待测样本与神经元的距离公式：

$$S_i = \mathrm{e}^{-\| x'-x_i \|/2\sigma^2}, \tag{2}$$

其中 $-1/2\sigma^2$ 为伽玛参数 g，在训练 PNN 时，需进行该参数优化，一般采用网格搜索法。

1.3　SVM 的原理

以两类(正样本和负样本)分类问题为例，在线性可分的情况下，SVM 构建一个超平面 H

$$\boldsymbol{W} \cdot \boldsymbol{P} + b = 0, \tag{3}$$

式中：$\boldsymbol{W}$ 为权重向量；$\boldsymbol{P}$ 为特征向量；b 为一参数。该超平面以最大边界的形式将正负样本区分开。该超平面的构建是通过寻找向量 $\boldsymbol{W}$ 和参数，使其在满足条件

$$\boldsymbol{W} \cdot \boldsymbol{P}_i + b \geqslant 0,(\text{对正样本}, y = +1), \tag{4}$$

$$\boldsymbol{W} \cdot \boldsymbol{P}_i + b < 0,(\text{对负样本}, y = -1) \tag{5}$$

时 $\| \boldsymbol{W} \|$ 达到最小。式中 $\boldsymbol{P}_i$ 代表第 i 个训练样本的特征向量；$\| \boldsymbol{W} \|^2$ 代表权重向量 $\boldsymbol{W}$ 的欧几里德范数；y 为样本类别标记。在求出 $\boldsymbol{W}$ 和 b 后，通过决策函数

$$y_i = \mathrm{sign}[\boldsymbol{W} \cdot \boldsymbol{P}_i + b] \tag{6}$$

判断向量 $\boldsymbol{P}_i$ 所对应测试样本的类别。若决策函数值为 $+1$，该样本属于正样本；否则，属于负样本。

在线性不可分的情况下，SVM 利用核函数 $K(\boldsymbol{P}_i,\boldsymbol{P}_j)$ 将特征向量映射到一个高维空间。在此高维空间中，线性不可分问题被转化为线性可分问题，其决策函数为

$$y_j = \mathrm{sign}\Big[\sum_{i=1}^{l} \alpha_i y_i K(\boldsymbol{P}_i,\boldsymbol{P}_j) + b\Big], \tag{7}$$

式中 l 为训练样本数，系数 α_i 和 b 应使拉格朗日表达式

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\boldsymbol{P}_i,\boldsymbol{P}_j) \tag{8}$$

达到最大值，且应满足

$$C > \alpha_i \geqslant 0 \text{ 和} \sum_{i=1}^{l} \alpha_i y_j = 0, \tag{9}$$

其中，C 为错误惩罚参数，它控制对错误分类样本的惩罚程度，C 越大支持向量的个数越多，最优超平面越复杂。

核函数 $K(\boldsymbol{P}_i,\boldsymbol{P}_j)$ 一般取径向基函数

$$K(\boldsymbol{P}_i,\boldsymbol{P}_j) = \mathrm{e}^{-\| \boldsymbol{P}_i-\boldsymbol{P}_j \|^2/2\sigma^2}。 \tag{10}$$

一般训练过程中需要对径向基函数中的伽玛参数 $g=-\dfrac{1}{2\sigma^2}$ 进行优化，大多采用的方法为网格搜

索法[5]。

2 结果及讨论

2.1 数据采集

采用专业录音笔录音，环境为噪音相对较小的一般办公场所。男女说话人各选择43名和35名。说话人以一般语速和语调叙述一段内容，说话内容不限。所有语音数据均在采样频率为32 kHz时采集，每人录音时间为30 s。录音完毕得到78个语音文件，为降低语音数据的繁冗度，采用11 024 Hz重新采样，即下采样(Down-sampling)，并将语音文件保存为波形文件格式，然后进行预处理。

2.2 预处理

由于录制的语音数据中存在非语音段，需要将非语音段剔除。一般采用方法为平均能量和平均过零率检测[12]。平均能量和平均过零率是最基本的语音信号时域特征，其定义如式(11)、(12)所示。

平均能量

$$\mathrm{RMS}=\frac{1}{L}\sum_{n=1}^{L}s^2(n)。\tag{11}$$

平均过零率

$$\mathrm{ZCS}=\frac{1}{2L-1}\sum_{n=1}^{L-1}\left|\operatorname{sign}(s(n+1))-\operatorname{sign}(s(n))\right|,\tag{12}$$

其中L为一段语音的采样点数，n为L个采样点中的任意一点。根据不同的采样频率可以选择不同大小的L，一般选择加窗处理中的窗口大小为2^N，N为正整数。

在处理过程中，首先计算待处理语音的总平均能量和总平均过零率，然后用大小为L的窗口采用重叠式连续截取语音，并计算各段语音数据的平均能量和平均过零率，最后根据全语音段的平均能量和平均过零率设定门限值，将平均能量和平均过零率低于或高于该门限的语音段判断为非语音段，并将其剔除，如图1所示。图1中，(a)图为待处理的一段语音，(b)图为各窗口的平均能量，(c)图为各窗口的平均过零率，通过比较(a)和(b)图，发现平均能量较大时对应语音段，平均能量较小时对应非语音段，并且对应关系较为良好。而(c)和(a)图的对应关系不明显，这可能是背景噪音的不同引起的。所以笔者只采用能量进行非语音段剔除，门限值设为待处理全语音段平均能量的1/10，如果截取语音段的平均能量大于该值，即判断该段为语音段。图1(d)为非语音段剔除后的语音序列。通过比较(a)和(d)图，可以看出新的语音序列保存了原始语音信号的语音段。

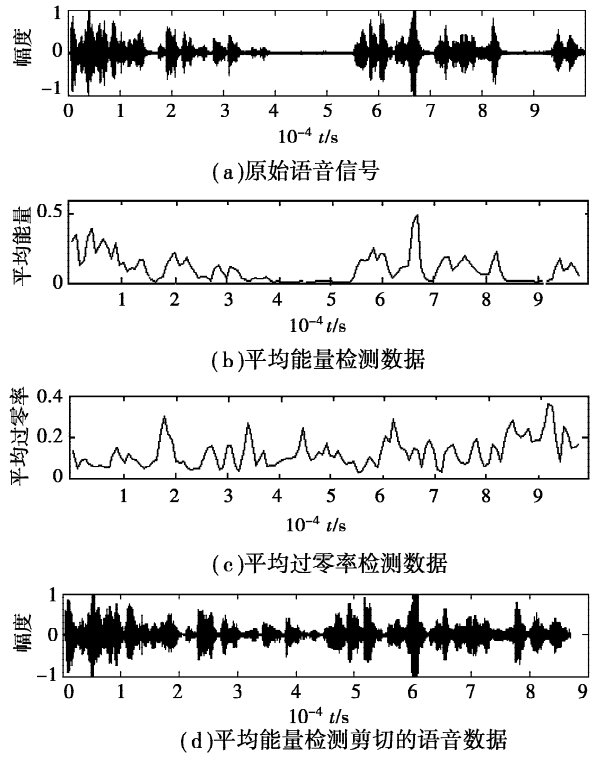

图1 语音检测和剪切预处理结果

语音信号是一种长时非平稳信号，但短时间内可视为平稳信号，语音特性主要体现在短时特性上，所以需对语音信号其分帧提取短时性[13]。采用帧长约为45 ms，帧移约为30 ms，即重叠式截取短时语音序列，如图2所示。

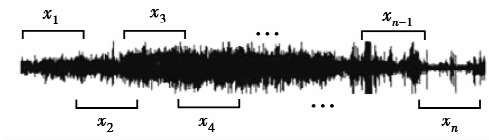

图2 短时时间序列$\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$的重叠式截取

为了避免进行快速傅里叶变换产生吉布斯效应，必须对$\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$进行平滑过滤[14]。平滑过滤器一般选择汉明窗口。其表达式为

$$W_i=0.54-0.46\cos\left(2\pi\frac{i}{L}\right),i=0,1,\cdots,L-1,\tag{13}$$

$$\boldsymbol{X}_{ji}=\boldsymbol{X}_{ji}W_i,j=1,2,\cdots,N,i=0,1,\cdots,L-1。\tag{14}$$

2.3 MFCC的特征提取

设$\boldsymbol{x}_j$代表某一语音信号经预处理后的其中一帧，对$\boldsymbol{x}_j$进行快速傅里叶变换，获取频谱系数并求出能谱系数$\boldsymbol{f}_j$，然后利用梅尔频谱特征标度的三角形滤波器组进行滤波处理。三角形滤波器组如图3所示。每个三角形可视为一个中心频率和一个上下截止频

率的带通滤波器。中心频率为人耳对某频段的感知中心，上下截止频率为人耳在该频段的感知范围。滤波器在低频段的个数比较均匀，随着频率的增加，滤波器的个数呈指数衰减[15]。滤波器的形状可供选择，如三角形、汉明形和汉宁形，但使用最多的是三角形。滤波器的个数一般选择为 24。

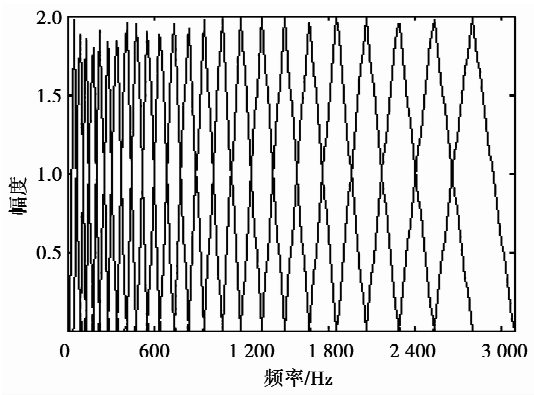

图 3　三角形滤波器组及其中心频率

利用滤波器组对频谱系数 $\boldsymbol{f}_j$ 进行加和滤波，即在同一三角形下，频谱与对应的三角形幅值相乘然后求和，得到 24 个系数。对这 24 个系数 $\boldsymbol{m}_j$ 取对数，并作离散余弦变换，得到 MFCC 系数，即

$$C_i = \sqrt{\frac{2}{k}} \sum_{j=1}^{k} \ln(m_j) \cos \mid \pi i/k(j-0.5) \mid , \tag{15}$$

式中：k 是三角形滤波器个数；m_j 为第 j 个滤波器的输出；C_i 为 MFCC 的第 i 个分量，其中，$i=1,2,\cdots,k$。

2.4　实验及结果

利用 MFCC 特征提取方法提取 78 个语音文件前 10 s 语音数据的特征向量构成训练向量集。提取 78 个语音文件第 2 个 10 s 语音数据的特征向量构成测试向量集。利用训练样本集对二类 SVM 进行训练构建一个 SVM 预测模型。SVM 的核函数为径向基函数，$C=10\ 000$，g 经过优化为 0.001。

将性别为男和女的特征向量分别视为正样本和负样本。测试过程中，将待识别性别说话人的全部测试样本输入 SVM 预测模型，得到各测试样本的正负类别，并与实际正负类别进行比较，若该说话人的特征向量类别一半以上被判断正确，则设该说话人的性别判断正确。利用相同的方法分别对 78 个说话人对应的测试样本集进行说话人性别的判断，计算性别判断正确的说话人的个数，得出总的准确率。

设 TP(True Positive)代表在测试集中被判断正确的男性说话人个数；FN (false negative)代表在测试集中被错判为女性说话人的个数；TN(True Negative)代表在测试集中被判断正确的女性说话人个数；FP (False Positive)代表在测试集中被错判为男性说话人的个数。

测试准确率公式为

$$Q_p = TP/(TP+FN), \tag{16}$$

$$Q_n = TN/(TN+FP), \tag{17}$$

$$Q = (TP+TN)/(TP+FN+TN+FP)。 \tag{18}$$

为比较 SVM 与不同分类方法对该数据的识别效果，分别选择了 KNN 和 PNN 方法进行识别。KNN 的最近邻数选择为 3。PNN 采用高斯函数度量神经元与样本之间距离，伽玛参数优化为 0.001。

表 1　KNN、PNN 和 SVM 对测试样本集的识别准确率

方法	TP	FN	TN	FP	Q_p/%	Q_n/%	Q/%
KNN	40	3	31	4	93.0	88.6	91.0
PNN	41	2	31	4	95.4	88.6	92.3
SVM	43	0	34	1	100	97.1	98.7

表 1 中数据表明，男(正)样本测试准确率 Q_p 明显比女(负)样本测试准确率 Q_n 高，这说明男性语音比女性语音更容易识别。从总的准确率 Q 可以看出，SVM 的准确率明显高于 KNN 和 PNN，是一种高识别率的分类方法。

3　结　语

在自建的语音性别数据库的基础上，提出联合梅尔频率频谱系数的特征提取方法和支持向量机的分类方法进行说话人性别识别，结果表明 SVM 的识别准确率高于其它几种常用的分类器，达98.7%。这表明用于说话人识别的 MFCC 特征能有效地用于说话人性别识别，且与 SVM 联合得到最佳效果。由于该数据库说话人数量较小，为达到实际应用的要求，下一步将扩大说话人数量。另外，由于录音环境人为控制得较为理想，可加入一定的噪声，研究分类器的抗噪能力，使之更符合实际应用的要求。

参考文献：

[1] 张捍东，李金炜. 基于性别识别的分类 CHMM 语音识别[J]. 计算机工程与应用，2007，43(21)：187-189.
ZHANG HAN-DONG，LI JIN-WEI. Speech recognition based on CHMM classified by gender identification [J]. Computer Engineering and Applications，2007，43(21)：187-189.

[2] 李娟娟，俞一彪，薛广荣. 说话人性别识别系统的 DSP 实现[J]. 现代电子技术，2005，215(24)：37-39.
LI JUAN-JUAN，YU YI-BIAO，XUE GUANG-RONG. Speaker gender identification using DSPs[J]. Modern Electronic Technique，2005，215(24)：37-39.

[3] 邓英，欧贵文. 基于 HMM 的性别识别[J]. 计算机工程与应用，2004,40(15)：74-75.

DENG YING, OU GUI-WEN. Gender identification using HMM [J]. Computer Engineering and Applications, 2004, 40(15): 74-75.

[4] 王伟，邓辉文. 基于MFCC参数和VQ的说话人识别系统[J]. 仪器仪表学报，2006，27(6)：2253-2255.
WANG WEI, DENG YUI-WEN. Speaker recognition system using MFCC features and vector quantization[J]. Chinese Journal of Scientific Instrument, 2006, 27(6): 2253-2255.

[5] VAPNIK V. The nature of statistical learning theory[M]. New York: Springer, 1995.

[6] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报，2000，26(1)：32-42.
ZHANG XUE-GONG. Introduction to statistical learning theory and support vector machines[J]. Acta Automatica Sinica, 2000, 26(1): 32-42.

[7] 肖汉光，蔡从中，廖克俊. 利用声波和地震波识别军事车辆类型[J]. 系统工程理论与实践，2006，26(4)：108-113.
XIAO HAN-GUANG, CAI CONG-ZHONG, LIAO KE-JUN. Recognition of military vehicles by using acoustic and seismic signals[J]. Systems Engineering-Theory & Practice, 2006, 26(4): 108-113.

[8] CAI C Z, HAN L Y, JI Z L, et al. SVM2Prot: Web2based support vector machine software for functional classification of a protein from its primary sequence[J]. Nucleic Acids Research, 2003, 31(13): 3692-3697.

[9] CAI C Z, HAN L Y, JI Z L, et al. Enzyme family classification by support vector machines[J]. Proteins, 2004, 55(1): 66-76.

[10] 蔡从中，袁前飞，肖汉光，等. 中药组方的计算机辅助分类与识别[J]. 重庆大学学报：自然科学版，2006，29(10)：42-46.
CAI CONG-ZHONG, YUAN QIAN-FEI, XIAO HAN-GUANG, et al. Computer-aided classification and identification of traditional Chinese medicine herbal formula[J]. Journal of Chongqing University: Natural Science Edition, 2006, 29(10): 42-46.

[11] SPECHT D F. Probabilistic neural networks[J]. Neural Networks, 1990, 3(5): 109-118.

[12] AVCI E. A new optimum feature extraction and classification method for speaker recognition: GWPNN[J]. Expert Systems with Applications, 2007, 32(2): 485-498.

[13] GIULIANI D, GEROSA M, BRUGNARA F. Improved automatic speech recognition through speaker normalization[J]. Computer Speech and Language, 2006, 20(1): 107-123.

[14] LI L. Ground vehicle acoustic signal processing based on biological hearing models[D]. Maryland: University of Maryland College Park, 1999.

[15] 张小玫，张雪英，梁五洲. 基于小波Mel倒谱系数的抗噪语音识别[J]. 中国电子科学研究院学报，2008，3(2)：187-189.
ZHANG XIAO-MEI, ZHANG XUE-YING, LIANG WU-ZHOU. A noise robust speech recognition based on wavelet MFCC[J]. Journal of China Academy of Electronics and Information Technology, 2008, 3(2): 187-189.

（编辑 张 苹）

(上接第765页)

[7] BOSSANYI E A. Bladed for windows-theory manual[M]. England: Garrrad Hassan and Parteners Limited, 1999.

[8] RAZAVI H, ABOLMAALI A, GHASSEMIEH M. Invisible elastic bolt model concept for finite element analysis of bolted connections [J]. Journal of Constructional Steel Research, 2007, 63(5): 647-657.

[9] 方栋，陈继志. 高强度螺栓螺纹根部应力集中的有限元分析[J]. 材料开发与应用，2007(2)：37-39.
FANG DONG, CHEN JI-ZHI. Finite element analysis of stress concentration at the root of screw thread of high strength bolt[J]. Development and Application of Materials, 2007(2): 37-39.

[10] 许岚，龚曙光，陈艳萍，等. 基于有限元风机轮毂结构形状优化与模态分析[J]. 现代制造工程，2005(12)：3-6.
XU LAN, GONG SHU-GUANG, CHEN YAN-PING, et al. Structural shape optimization and modal analysis for axial fan hub by finite element method[J]. Modern Manufacturing Engineering, 2005(12): 3-6.

[11] 孙传宗，姚兴佳，单光坤. MW级风力发电机轮毂强度分析[J]. 沈阳工业大学学报，2008，30(1)：46-49.
SUN CHUAN-ZONG, YAO XING-JIA, SHAN GUANG-KUN. Strength analysis of MW wind turbine hub[J]. Journal of Shenyang University of Technology, 2008, 30(1): 46-49.

[12] 何婧，何玉林，金鑫，等. 失速型风力发电机系统振动仿真分析[J]. 重庆大学学报：自然科学板，2007，30(5)：91-95.
HE JING, HE YU-LIN, JIN XIN, et al. Vibration analysis and system simulation for a stalled wind turbine [J]. Journal of Chongqing University: Natural Science Edition, 2007, 30(5): 91-95.

[13] 蒋丽. 拉力作用下高强螺栓联接的有限元模拟[J]. 山西建筑，2006，32(21)：56-57.
JIANG LI. Finite element simulation of the high strength bolted connection [J]. Shanxi Architecture, 2006, 32(21): 56-57.

（编辑 张 苹）