软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(Suppl.(2)):24−31

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种基于背景声音识别的社会情境感知方法*

杨 曜[1], 郭 斌[2], 於志文[2]

[1](西北工业大学 软件与微电子学院,陕西 西安 710072)

[2](西北工业大学 计算机学院,陕西 西安 710072)

通讯作者: 於志文, E-mail: zhiwenyu@nwpu.edu.cn

摘 要: 随着社会需求的不断扩大及技术的不断发展,人与人之间的社会交互也越来越多.理解社会交互特征并能感知用户所处的社会情境语义(如在开会、在上课),对于促进和辅助用户社会活动具有重要意义.从背景声音的角度对社会交互进行理解,目的是通过对背景声音差异性特征的提取,识别用户所处的社会情境.提出了一种基于背景声音识别的社会情境感知方法,该方法采用 Mel frequency cepstral coefficients (MFCCs,即 Mel 频率倒谱系数)分析声音信号,将路径搜索限制和搜索过滤的改进 Dynamic Time Warping(DTW)算法作为识别器.通过对 11 种社会情境背景声音的采集和识别,表明该算法能够有效地识别用户所处的社会情境,且其运算效率与识别率比传统 DTW 算法有所提高.

关键词: 社会情境感知;背景声音识别;Mel 频率倒谱系数;DTW 算法

中文引用格式: 杨曜,郭斌,於志文.一种基于背景声音识别的社会情境感知方法.软件学报,2013,24(Suppl.(2)):24−31. http://www.jos.org.cn/1000-9825/13020.htm

英文引用格式: Yang Y, Guo B, Yu ZW. Approach of social context awareness based on background sound recognition. Ruan Jian Xue Bao/Journal of Software, 2013,24(Suppl.(2)):24−31 (in Chinese). http://www.jos.org.cn/1000-9825/13020.htm

Approach of Social Context Awareness Based on Background Sound Recognition

YANG Yao[1], GUO Bin[2], YU Zhi-Wen[2]

[1](School of Software and Microelectronics, Northwestern Polytechnical University, Xi'an 710072, China)

[2](School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China)

Corresponding author: YU Zhi-Wen, E-mail: zhiwenyu@nwpu.edu.cn

Abstract: With the spread of social needs and development of techniques, social interaction is more and more frequent among people. To promote and assist human social interaction, it's important to understand the social context the user situates. The paper mainly studies the understanding of social contexts based on background sounds, the goal of which is to recognize the social context in which users reside through analyzing the differences of background sounds. It uses the Mel frequency cepstral coefficients to analyze sound features and classify the sounds based on an improved Dynamic Time Warping (DTW) algorithm. Experimental results show that the proposed algorithm is more effective than traditional methods.

Key words: social context awareness; background sound recognition; Mel frequency cepstral coefficients (MFCC); DTW algorithm

情境感知技术最早由 Schilit 于 1994 年提出[1].其目标是通过传感器及其相关技术使计算机设备(特别是移动计算设备)能够"感知"用户当前的情境.情境包括多个方面,如个体情境、环境情境、社会情境等.本文主要从社会情境角度出发进行分析,它一般是指用户所处的社会场景或正在参加的群体活动[2],比如开会、派对、上课

* 基金项目: 国家自然科学基金(61373119, 61222209, 61103063); 国家重点基础研究发展计划(973)(2012CB316400); 新世纪优秀人才支持计划(NCET-12-0466); 高等学校博士学科点专项科研基金(20126102110043); 陕西省自然科学基础研究计划(2012JQ8028); 西北工业大学基础研究基金(JC20110267)

收稿时间: 2012-06-15; 定稿时间: 2013-07-22

等.通过准确感知当前情境,可以了解所处环境情况,辅助人与人之间的交互.

获取人类活动信息主要依靠多种传感器(如视频传感器,音频传感器,加速度传感器等),而背景声音识别技术是近年来发展起来的一种普适技术.它通过音频传感器(如智能手机等)获取背景声音,可以实时、准确地识别人类活动中的个体和群体行为,并做出正确的理解.尽管目前已有很多对语音特征提取和识别的技术,但是运用在背景声音识别方面还很少见.此外,已有的基于背景声音识别的研究主要面向个体情境[3]或环境动态[4],在基于背景声音的社会情境识别方面则未见探索.

社会情境的识别对于辅助人类交互、提供基于情境的服务具有重要作用.鉴于此,本文提出基于背景声音识别社会情境的方法,采用 Mel 频率倒谱系数提取声音特征,识别过程采用搜索路径限制和结果过滤的 DTW 算法,在传统的 DTW 算法的基础上提高了识别率与计算效率,通过对 11 类不同的社会场景声音的识别证明了该方法的有效性.

1 相关工作

背景声音识别活动是近年来出现的一个研究领域,作为情境感知的一部分,具有背景声音识别功能的计算终端设备更具人性化.目前,相关的声音特征提取一般采用 LPCC,MFCC 和 HCC 技术[5,6],LPCC 参数是线性预测系数(linear prediction coefficient,简称 LPC)在倒谱中的表示,该特征是基于语音信号为自回归的假设,利用线性预测分析获得倒谱参数.MFCC 参数将频谱转化为基于 Mel 频率的非线性频谱,利用了人耳听觉特性.而 HCC 是在 MFCC 基础上发展而来的[5].在识别算法方面,传统的 DTW 算法和基于高斯混合模型的方法均可作为识别算法.

在背景识别中,识别器安装在终端移动设备中,识别算法的计算效率是决定用户体验的一个重要因素[7],在传统的 DTW 算法上,通过路径约束条件改进其计算能力,并在原有基础上降低了存储空间,但是,其计算效率的提高和存储空间的降低将略微导致其识别率的下降,因此,在识别结果中采取一种筛选策略,过滤掉不符合要求的结果[8],在此基础上重新进行识别,可以有效地提高识别率.本文综合考虑识别率和算法效率之间的取舍,结合搜索路径限制和结果过滤,既能成功地提高识别率,也使算法效率得以改进.

2 系统架构设计

本系统是基于背景声音识别的社会情境感知系统框架,可以通过背景声音识别感知当前所处环境[3],系统架构如图 1 所示.实验中,背景声音采集使用 RX98M 便携式录音工具.

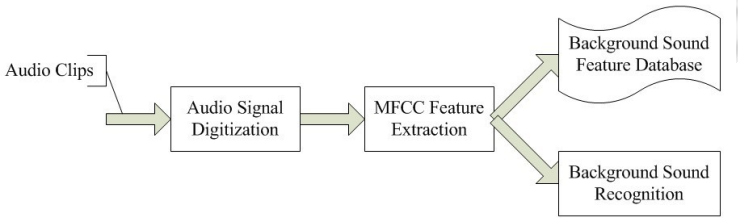

Fig.1 System architecture

图 1 系统架构

(1) 声音信号数字化

声音信号经过采样后,由 A/D 变换器变换为二进制数字码,方便后面特征计算.

(2) MFCC 特征提取

对每类背景声音采集多个该类声音文件,通过标准的 MFCC 特征提取,形成一个 22 维的 MFCC 特征参数.每个声音文件经过 MFCC 过程形成一个 MFCC 特征参数,对不同的 MFCC 特征参数求和再取平均值,该平均值为最终的 MFCC 特征参数.

(3) 背景声音特征模板库

每一类背景声音通过特征提取形成 MFCC 特征参数,保存在背景声音特征模板库中.

(4) 背景声音识别

当一个待识别的背景声音信号通过移动计算设备采集后,经过与建立背景声音特征模板库时相同的 MFCC 特征提取,采用改进后的 DTW 算法计算出与特征模板库中所有背景声音特征的距离,找到最接近的一个模板声音.

3 基于背景声音的情境识别算法

背景声音识别采用动态时间规整(dynamic time warping,简称 DTW)算法,该算法基于动态规划(dynamic programming,简称 DP)的思想.普通的 DTW 算法由于要计算每个帧距离,并且保存所有帧距离,付出了很大的时间与空间代价,在此基础上,提出了一种提高计算效率和识别率的方法.

3.1 特征提取——MFCC

在语音特征提取过程中,我们对特征参数的要求是能够有效地代表语音特征,具有更好的区分性;特征语音参数之间更易于区分,具备良好的独立性.本文采取 MFCC 方法作为声音特征提取,提取过程如图 2 所示.

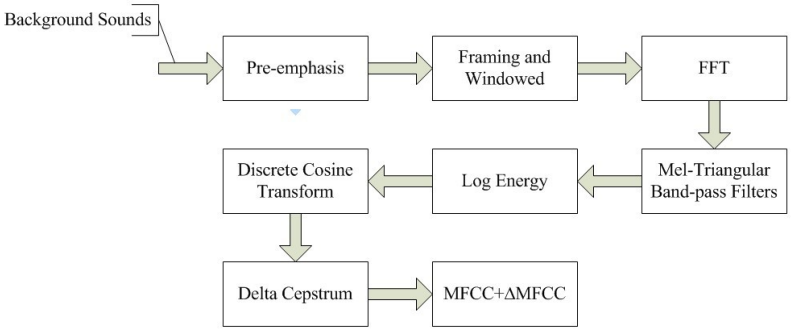

Fig.2 MFCC feature extraction process

图 2 MFCC 特征提取流程

MFCC 提取过程如下:

(1) 对输入的声音信号进行预强调(pre-emphasis),令声音信号通过一个高通滤波器:

$$H(z)=1-\alpha\times z^{-1} \tag{1}$$

公式(1)中的α介于 0.9~1.0 之间,作用是为了突出高频部分的共振峰以及使信号变得平坦.实验中设置α=0.9375.

(2) 在识别过程中,每次计算以帧为单位,所以要对声音信号进行分帧(frame blocking)处理,实验中选取的帧长为 16ms,帧叠 8ms.然后对每帧信号进行加窗处理,增加每帧之间的连续性.窗形状选择为汉明窗(Hamming window).

(3) 由于信号在时域上的变化很难区分,所以对分帧加窗处理后的信号进行快速傅里叶转换(fast Fourier transform,简称 FFT),目的是将信号在时域上的变化转化到频域上的分布,再求频谱幅度的平方,得到能量谱.

(4) 将能量通过一组 Mel 尺度的三角带通滤波器组,然后求取每一个滤波器输出的对数能量,Mel 频率和一般频率f的关系如公式(2)所示:

$$mel(f)=2595\times\log_{10}(1+f/700) \tag{2}$$

然后对对数能量进行离散余弦变换(discrete Cosine transformation,简称 DCT),去除各维之间的相关性,将信号映射到 12 维的低维空间,形成 12 维的 Mel 倒频谱参数.离散余弦变换如公式(3)所示:

$$DCT_m=\sum_{k=0}^{n-1}\cos[(2k+1)\times m\pi/2n],\ \text{其中},m=1,2,3,...,12 \tag{3}$$

这里,n 是三角滤波器的个数,实验中取 n=24.最后对映射到 12 维空间的 Mel 特征参数求取一阶差分,形成

一个 24 维的特征参数,由于特征参数中第 1 个参数对结果误差影响太大,所以删除 MFCC 和一阶差分参数中第 1 个参数[9],差分参数主要是为了显示倒谱参数对时间的变化,这就是 22 维 MFCC 参数.

3.2 识别算法——DTW

(1) 传统的 DTW

传统的 DTW 算法基于动态规划思想,具有运算量较大、但技术上容易实现、正识率也较高等特点.为了描述传统的 DTW 算法,我们用 R 和 T 分别代表参考模板和测试模板.其中,

$$R=\{R(1),R(2),...,R(m),...,R(M)\},$$
$$T=\{T(1),T(2),...,T(n),...,T(N)\}.$$

m 为训练声音帧的时序标号,n 为测试声音帧的时序标号,这样可以形成一个 $m\times n$ 的网格,DP 算法就是寻找一条从(1,1)到(M,N)的最短路径.

由于 DTW 算法基于动态规划思想,其最优解为累积距离 $D(M,N)$,由 R 和 T 对应的帧距离 $d(m,n)$计算得出.累积距离的递归求解如公式(4)所示.其中,初始条件 $D(1,1)=d(1,1)$.

$$D(m,n)=d(m,n)+\min\{D(m,n-1),D(m-1,n-1),D(m-1,n)\} \tag{4}$$

(2) 改进的 DTW

在动态规整过程中,计算量大且需要存储每一个帧距离 $d(n,m)$,为了减少计算量和存储空间,采取约束搜索路径方法,该方法将搜索路径限制在两边斜率分别为 0.5 和 2 的平行四边形区域内,缩小了搜索范围从而提高了计算效率[8].如果搜索路径通过当前坐标(n,m),则下一个坐标只可能是(n+1,m+2),(n+1,m+1),(n+1,m).约束路径如图 3 所示.

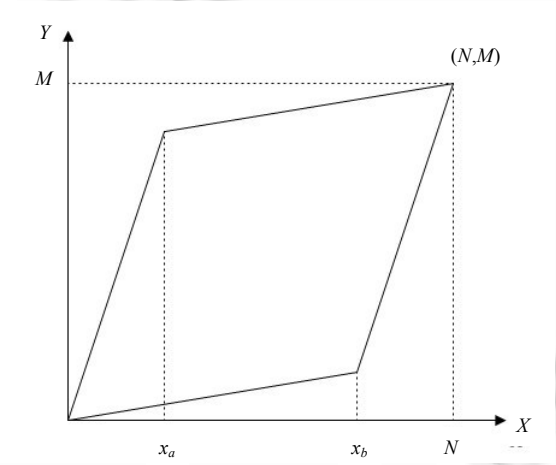

Fig.3 DP algorithm

图 3 DP 算法示意图

如图 3 所示,由于减少了搜索路径,最终结果不一定是最优解,识别率会有所下降.由于平行四边形的两边斜率分别为 0.5 和 2,y 的上界和下界随 x 的变化而变化,为了计算 y 的边界值,必须把搜索路径计算过程动态地分为 3 个阶段:(1,x_a),(x_a+1,x_b),(x_b+1,N),其中每个阶段对应不同的 y 的上边界和下边界,x 的划分必须满足公式(5):

$$\begin{cases} x_a=\dfrac{1}{3}(2M-N) \\ x_b=\dfrac{2}{3}(2N-M) \end{cases} \tag{5}$$

由于搜索路径限制为一个平行四边形,对于平行四边形以外的帧距离不需要计算,那么,只需要计算平行四边形的上界和下界之间的帧距离.平行四边形边界计算如式(6)、式(7)所示.

$$y_{\min}=\begin{cases}\frac{1}{2}x, & 0\leqslant x\leqslant x_b\\ 2x+(M-2N), & x_b<x\leqslant N\end{cases} \tag{6}$$

$$y_{\max}=\begin{cases}2x, & 0\leqslant x\leqslant x_a\\ \frac{1}{2}x+\left(M-\frac{1}{2}N\right), & x_a<x\leqslant N\end{cases} \tag{7}$$

由于 DTW 算法把搜索路径限制在如图 3 所示的平行四边形中,在搜索路径中用来计算最短路径的格点都落在平行四边形内,那么平行四边形外的所有点将不会对结果产生影响,计算量会降低很多.其次,累积距离在 x 轴每前进一格,都只和前一格的 3 个点的累积距离有关,所以不需要一次性存储所有平行四边形内的格点,这样可以大大缩短算法的运行时间并提高其效率.但是由于减少了搜索路径,导致识别率会略微下降.

其次,基于背景声音识别的声音模板不同于语音识别,前者具有复杂性、多变性等特点,往往不同类的背景声音,由于存在部分相同的元素,导致识别率会有所下降.基于上述原因,提出一种改进 DTW 算法的思想:在进行第 1 次 DTW 识别的过程中,将识别率低于某个阈值的模板删除掉,然后将测试环境声音在剩下的训练模板中再经过一次 DTW 识别过程,这样,在识别率上会所提高,特别是当训练模板种类很多时,识别率将有明显的提高.

但是,由于经过 2 次 DTW 识别过程,算法效率将明显下降.所以,采取另外一种方法来达到相同的结果:在第 1 次 DTW 算法过程中,计算测试声音模板在每个训练模板下的识别率并保存与每个模板的距离,将识别率低于某个阈值(实验中阈值选取为 10%)的距离设为最大值,最后再计算测试声音模板在每个训练模板下的识别率.这样可以避免经过 2 次 DTW 算法带来的效率降低.表 1 是 4 种不同阈值在 2 次 DTW 识别过程中的识别率.

Table 1 Recognition rate under different thresholds

表 1 不同阈值下的识别率

阈值(%)	识别率(%)
10	84.76
15	87.54
20	91.36
25	93.18

由表 1 所示,虽然阈值在 15%,20%,25%时识别率依次递增,但是不能简单地认为阈值越大越好.在实际应用中,背景声音种类模板远远大于实验中的 11 类,并且多种不同性质的声音干扰,少量的误识在所难免,为了避免阈值过大而把正确的识别结果过滤掉以及提高识别率,最终,实验选取的阈值为 10%.

4 实验结果与分析

本次实验中,选取了 11 种不同的背景声音,这些背景声音通过用 RX98M 便携式录音工具录制,声音文件格式为 wav 格式,采样频率为 16KHz,采样率为 16bit,每个声音文件长度从 40s 到 10min 不等.表 2 列出实验中的背景声音采集地点.

Table 2 Data collection places of background sounds

表 2 背景声音采集地点

序号	背景声音采集地点
1	公路
2	公交车站
3	超市
4	图书馆
5	食堂
6	公交车
7	地铁站
8	露天篮球场
9	餐馆
10	自习室
11	室内体育馆

由于背景声音具有复杂性和多样性,在采集背景声音时,尽量选取具有该背景声音特点的声音模板,使得MFCC特征更容易加以区别,在识别时降低误识率.其次,为了使背景声音更具普遍性,模板背景声音也在不同的时间采集.

如表 3 所示,虽然部分场景识别率都有所提高,但是也存在部分背景声音(如超市)的识别率低于 50%的情况,其主要原因是多种不同性质的声音混合和相似背景下的误识,超市中存在音乐、谈话等影响,说明该方法在鲁棒性上存在一定缺陷.而如表 4 和表 5 所示,分别对长度为 1min 和 3min 的公路背景声音进行识别,识别结果没有明显差异,在相似背景声音中,由于具有相似的声音特征,与其他背景声音相比,都具有较高的误识率.在图书馆和自习室的比较中,由于都具有比较安静的声音特征,可以看到,在图书馆背景声音识别中,自习室的误识率达到10.15%,而其他类的背景声音误识率总和只有8.49%.

如图 4 所示,在传统的 DTW 算法的基础上利用路径搜索限制后的算法,识别背景场所花费的平均时间从7.99s 降低到了 3.64s.但如表 3 所示,这样处理的代价是略微降低了识别率,平均识别率从 79.03%降低到了77.65%.采用结果过滤的 DTW 算法后,识别率得到了一定程度的提升,从传统 DTW 算法的平均识别率 79.03%提高到了 84.76%.如图 4 所示,在结果过滤的基础上采用路径限制的 DTW 算法,那么在时间效率上也得到了提高,从传统 DTW 算法的 7.95s 降低到 3.56s,而识别率只是略微有所降低.实验结果表明,改进后的 DTW 算法在时间效率和识别率上都要高于传统的 DTW 算法,为今后在移动设备终端上得以实现奠定了基础.

Table 3 Recognition rate of the 4 algorithms

表 3 4 种算法的识别率

背景声音采集地点	传统 DTW 算法(%)	搜索路径限制 DTW 算法(%)	结果过滤 DTW 算法(%)	搜索路径限制和结果过滤 DTW 算法(%)
公路	76.19	77.78	79.37	80.95
公交车站	81.82	80.00	100.00	87.27
超市	31.67	41.67	43.33	56.67
图书馆	81.36	83.05	81.36	83.05
食堂	80.00	80.00	80.00	80.00
公交车	100.00	100.00	100.00	100.00
地铁站	71.67	63.33	71.67	65.00
露天篮球场	88.33	78.33	100.00	100.00
餐馆	95.00	91.67	100.00	100.00
自习室	91.67	81.67	100.00	83.33
室内体育馆	71.67	76.67	76.67	78.33
平均识别率	79.03	77.65	84.76	83.15

Table 4 Recognition rate of the same scene with different time length

表 4 不同时间长度的相同场景下的识别率

	传统 DTW 算法(%)	搜索路径限制 DTW 算法(%)	结果过滤 DTW 算法(%)	搜索路径限制和结果过滤 DTW 算法(%)
公路(1 分钟)	76.19	77.78	79.37	80.95
公路(3 分钟)	77.42	76.90	80.32	80.09

Table 5 Correct and error recognition rate in similar scenes (%)

表 5 相似背景声音下正识率和误识率(%)

	图书馆	自习室	其他
图书馆	81.36	10.15	8.49
自习室	1.50	91.67	6.83
	食堂	餐馆	其他
食堂	80.00	3.17	16.83
餐馆	0.00	95.00	5.00
	公交车站	地铁站	其他
公交车站	81.82	5.79	12.39
地铁站	8.17	71.67	20.16

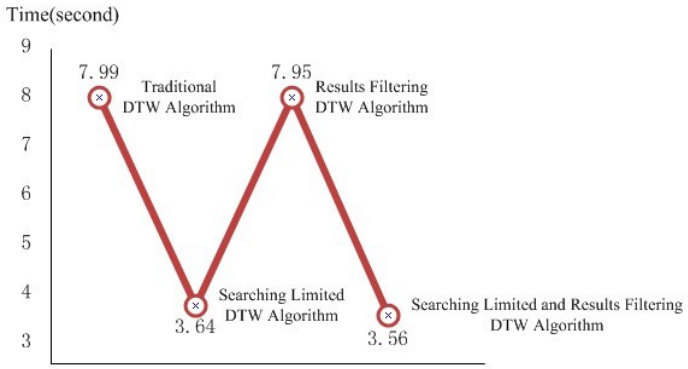

Fig.4 Efficiency of the 4 algorithms (Unit: s)

图 4 4 种算法的时间效率(单位:秒)

5 结束语

本文提出一种基于环境声音识别的社会情境感知方法,针对背景声音识别过程中的特征提取和识别算法进行研究.实验结果表明,本方法能够有效地识别用户所处情境.

在今后的工作中,主要将系统原型应用在移动终端设备,例如手机中,从实验中使用的计算设备到实际中的计算设备,其处理器的性能会有所下降,不可避免地会带来计算效率的降低,除了在移动设备中实现系统模型以外,在识别算法效率和能耗上也须加以改进.

致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其是日本庆应义塾大学詹毅老师表示感谢.

References:

[1] Schilit B, Adams N, Want R. Context-Aware computing applications. In: Mobile Computing Systems and Applications. Santa Cruz: IEEE Press, 1994. 85−90.

[2] Guo B, Zhang D, Yu Z, Liang Y, Wang Z, Zhou X. From the Internet of things to embedded intelligence. In: World Wild Web, Vol.16. Springer-Verlag, 2012. 399−420. [doi: 10.1007/s11280-012-0188-y]

[3] Zhan Y, Miura S, Nishimura J, Kuroda T. Human activity recognition from environment background sounds for wireless sensor network. In: Proc. of the IEEE Int'l Conf. on Networking, Sensing and Control. London: IEEE Press, 2007. 307−312. [doi: 10.1109/ICNSC.2007.372796]

[4] Rana R, Chou C, Kanhere S, Bulusu N, Hu W. Ear-Phone: An end-to-end participatory urban noise mapping system. In: Proc. of the 9th ACM/IEEE Int'l Conf. on Information Processing in Sensor Networks (IPSN 2010). New York: ACM Press, 2010. 105−116. [doi: 10.1145/1791212.1791226]

[5] Cowling M, Sitte R. Comparison of techniques for environmental sound recognition. Pattern Recognition Letters, 2003,24(15): 2895−2907.

[6] Davis SB, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. on Acoustics, Speech and Signal Processing, 2003,28(4):357−366. [doi: 10.1109/TASSP.1980.1163420]

[7] Yu Z, Yu Z, Zhou X. Social awareness computing: Concepts, issues and research progress. Chinese Journal of Computers, 2012,35(1):16−26 (in Chiense with English abstract). [doi: 10.3724/SP.J.1016.2012.00016]

[8] Zhang J. Research of improved DTW algorithm in embedded speech recognition system. In: Proc. of the 2010 Int'l Conf. on Intelligent Control and Information Processing (ICICIP). Dalian: IEEE Press, 2010. 73−75. [doi: 10.1109/ICICIP.2010.5564195]

[9] Zhen F, Zhang GL, Song ZJ. Comparisons of different implementations of MFCC. Journal of Computer Science and Technology, 2001,16(6):582−589. [doi: 10.1007/BF02943243]

附中文参考文献:

[7] 於志文,於志勇,周兴社.社会感知计算:概念、问题及其研究进展.计算机学报,2012,35(1):16−26. [doi: 10.3724/SP.J.1016.2012.00016]

杨曜(1988－),男,四川人,硕士生,CCF 学生会员,主要研究领域为移动计算,普适计算.

E-mail: yangyao308@gmail.com

郭斌(1980－),男,博士,副教授,CCF 高级会员,主要研究领域为普适计算,社会智能,移动社会网络.

E-mail: guob@nwpu.edu.cn

於志文(1977－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为普适计算,移动互联网,人机交互,智能信息技术.

E-mail: zhiwenyu@nwpu.edu.cn