基于 Elman 神经网络的语音情感识别应用研究*

余伶俐[1]，周开军[1,2]，邱爱兵[3]

(1. 中南大学 信息科学与工程学院，长沙 410083；2. 湖南商学院 计算机与电子工程学院，长沙 410205；3. 湖南信息职业技术学院 机电工程系，长沙 410200)

摘　要： 针对语音情感的动态特性，利用动态递归 Elman 神经网络实现语音情感识别系统。通过连接记忆上时刻状态与当前网络一并输入，实现 Elman 网络模型的状态反馈。基于此设计了语音情感识别系统，该系统能在后台修改网络类型，并实现单语句与批量语句识别模式。针对系统进行语音情感识别实验表明，基于 Elman 神经网络的语音情感识别在同等参数模型设置前提下优于 BP 神经网络识别效果，且 BP 神经网络参数设置较 Elman 网络敏感。

关键词： 语音情感识别；Elman 网络；BP 网络；MFCC

中图分类号： TP391　　**文献标志码：** A　　**文章编号：** 1001-3695(2012)05-1809-06

doi:10.3969/j.issn.1001-3695.2012.05.055

Application research of speech emotion recognition based on Elman neural network

YU Ling-li[1], ZHOU Kai-jun[1,2], QIU Ai-bing[3]

(1. *School of Information Science & Technology, Central South University, Changsha* 410083, *China*; 2. *School of Computer & Electronic Engineering, Hunan University of Commerce, Changsha* 410205, *China*; 3. *Dept. of Mechanical & Electrical Engineering, Hunan College of Information, Changsha* 410200, *China*)

Abstract: This paper utilized the Elman dynamic recurrent neutral network to realize speech emotion recognition system for dynamic characteristics of speech. Meanwhile, it realized state feedback of Elman neutral network through input both of connection memory from last state and current state together. Finally, it designed speech emotion recognition system which could not only modify network structure types in backstage, but also practiced two mode of signal speech recognition and batch proces-sing recognition. Based on this platform, the experiments show that the recognition effect of Elman is superior to BP network un-der the same model parameters. Furthermore, parameters setting of BP are more sensitive than Elman network.

Key words: speech emotion recognition; Elman network; back propagation network; mel frequency cepstrum coefficient

0　引言

近年来，语音情感识别已成为人工智能领域的研究热点，受到国内外研究者的广泛关注，目前已涌现了众多的国内外研究机构，著名的有 MIT 的 Affective Computing 研究小组、慕尼黑工业大学的 Institute for Human-Machine Communication，在国内有模式识别国家重点实验室、清华大学人机交互与媒体集成研究所[1]、中国科学院语言研究所和浙江大学人工智能研究所等。根据语音识别系统提取的客户情感，服务机器可以了解到客户的精神状态，进而作出相应的友好反应；自动语音翻译系统可以让翻译更恰当地表达说话者的实际情感；电话客户服务中心语音识别系统可以更准确地掌握客户的心理和需求等[2]。

有效的语音特征集合是保证语音情感识别率的先决条件。人的语音含有丰富的信息量，能表征语音情感特征的参数也很多样化。目前并没有统一的标准认定哪种语音特征更适合语音情感识别[3]，受国内外广大语音情感识别研究者欢迎的语音特征有韵律特征[4]（代表性特征有基音周期、短时能量、语速、持续时间等及其统计特征值）和音质特征[5]（代表性特征有共振峰、MFCC、LPC 等及其统计特征值）等。

分类器的设计是语音情感识别系统中非常关键的一环，由于研究的条件和背景不同，目前也没有一致认可的最好的分类器，国内外研究者常用的分类器各有各的优缺点。具体而言，主成分分析法（PCA）[6]简单易行，但是识别效果不是很理想；隐马尔科夫模型（HMM）[7,8]虽有坚实的数学基础，且能捕获瞬时状态，记录状态转移，但在建立和训练时间上较长，实际应用时还需要解决计算复杂度过高的问题；高斯混合模型（GMM）[9]具有很强的数据分布表示能力，可得到原始观测特征参数的概率密度函数的最大似然估计，但是不能对数据作变换，而且需要足够多的训练样本；而神经网络（ANN）[10,11]以训练误差作为优化问题的约束条件，具有非线性和极强的分类能力，其推广能力明显优于一些传统的学习方法。近年来常用静态前向工作模式的神经网络对语音情感进行识别，网络的自学

收稿日期： 2011-09-30；**修回日期：** 2011-11-01　**基金项目：** 中央高校基本科研业务费专项资金资助项目(2012QNZT060)；第 49 批中国博士后科学基金面上资助项目(20110491272)；中南大学博士后基金资助项目(2010-2012)；湖南省教育厅青年基金项目(11B070)

作者简介： 余伶俐（1983-），女，江西景德镇人，讲师，博士（后），主要研究方向为语音情感识别、机器人感知与规划（llyu@csu.edu.cn）；周开军（1979-），男，湖南常德人，讲师，博士（后），主要研究方向为语音与图像处理技术；邱爱兵（1975-），男，湖南娄底人，工程师，硕士，主要研究方向为工业自动控制及应用.

习、自适应能力使之总能达到可接受的识别结果,但其缺乏处理时序特征能力。为了能更生动、直接地反映语音情感的动态特性,本文提出利用 Elman 反馈神经网络进行情感分类。

1 普通话情感分类

情感可以分为主要情感(原始情感)和次要情感(派生情感)[12]。大部分研究人员认为主要情感包括害怕(fear)、愤怒(anger)、高兴(joy)、悲伤(sadness)和厌恶(disgust)。次要情感由主要情感变化或混合得到。这类情感的生成理论也叫情感的调色板理论[13]。次要情感包括自豪(高兴的一种变化形式)、感激(高兴的一种派生形式)、惊奇等。除此情感分类外还有其他分类方法,如 Fox 提出的三级情感模型[13],它是按照情感中表现的主动和被动的程度将情感分成不同的等级,如表 1 所示,等级越低,分类越粗糙,等级越高,分类越精细。

表 1 Fox 的情感三级分类模型

rate		approach			withdrawal	
1st	joy	interest	anger	distress	disgust	fear
2nd	pride	concern	hostility	misery	contempt	horror
3rd	bliss	responsibility	jealousy	agony	resentment	anxiety

除了以上将情感在离散类别上的表示外,连续空间表示情感类别的方式——维度论逐渐受到广大语音情感学者的青睐。维度论把不同的情感看做是逐渐的、平滑的转变,不同情感间的相似性和差异性根据彼此在维度空间中的距离来显示。最广为接受的维度模式是激活度或唤醒度(activation or arousal)与评估度或愉悦度(evaluation or pleasure)组成的 Activation-Evaluation二维空间[14],其理论基础是正负情感的分离激活。本文研究人类语音中的情感分类方法,根据主要情感分类,分别对生气(anger)、高兴(joy)、悲伤(sadness)、中性(normal)语句进行识别分类。

2 语音信号预处理及情感特征提取

声音是声波通过空气的传播而产生的,是模拟信号,要进行计算机处理,就必须转换为数字信号,通过采样使输入的声音变为数字信号。语音信号前端处理流程如图 1 所示。

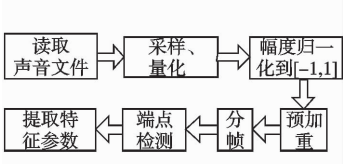


图1 语音信号前端处理流程

2.1 语音信号的幅度归一化

幅度归一化主要是为了方便数据处理而在预加重之前把数据映射到[-1,1]范围之内处理。本文以 fs = 16 kbit 的采样频率、以 bit 占 16 位字节,对模拟信号文件进行采样,将采样转换为双精度数存入数组而后归一化处理。

2.2 预加重

由于语音信号的平均功率频谱受声门激励和口鼻辐射的影响,高频端大约在 800 Hz 以上,按 6 dB/倍频程跌落。为了消除声带和嘴唇的效应,以提升高频共振峰的振幅部分,提高高频分辨率,使信号的频谱变得平坦,以便频谱分析或声道参数分析,为此,在预处理中先进行预加重。预加重数字滤波器如式(1)所示。提取语音信号特征前通过预加重滤波器,可达到消除滞留漂移、抑制随机噪声和提升清音部分能量的效果。

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

其中:μ 是预加重系数,在 0.9 ~ 1 之间。本文中的所有实验均采用 0.937 5。

2.3 分帧

浊音的基音周期、清浊音信号幅度和声道参数等均随时间而缓慢变化。为了减少语音帧的截断效应,对语音进行加窗分帧处理。由于发声器官的惯性运动,在小段时间内(如 10 ~ 30 ms)语音信号近似不变,即语音信号具有短时平稳性。短时语音片段的分帧采用移动有限长度窗口进行加权的方法来实现,一般每秒帧数为 33 ~ 100 帧。分帧一般采用如图 2 所示的交叠分段法,目的是使帧与帧之间过渡平滑,保证其连续性。本文实验取帧长 FrameLen = 256,帧移 FrameInc = 80。

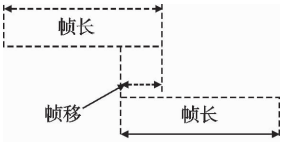


图2 语音信号的分帧

当窗口较宽时,平滑作用大,能量变化不大,故反映不出能量的变化;当窗口较窄时,没有平滑作用,反映了能量的快变细节,而看不出包络的变化。语音信号的分帧处理,实际上就是对各帧进行某种变换或运算。设这种变换用 $T[\ ]$ 表示,$x(n)$ 为输入语音信号,$w(n)$ 为窗序列,$h(n)$ 是与 $w(n)$ 有关的滤波器,则各帧经处理后输出表示为式(2)。

$$Q_n = \sum_{m=-\infty}^{\infty} T[x(m)]h(n-m) \tag{2}$$

2.4 端点检测算法

语音端点检测是指从包含语音的信号中判断语音的起始点和结束点。判别语音段的起始点和终止点的问题主要归结为区别语音和噪声的问题[15]。如果能保证系统的输入信噪比很高,通过计算短时能量就基本能把语音片段和背景噪声区分开。但实际应用中很难保证高信噪比,为此仅根据能量来判断较为粗糙,需进一步利用短时平均过零率判断,因为清音和噪声的短时平均过零率比背景噪声的平均过零率高。

短时能量代表音量高低,为每帧采样点的加权平方和。短时能量 E_n 的表达式如式(3)所示。其中,汉明窗函数式(4)平方的物理含义是一个冲激响应为 $w(n)^2$ 的滤波器;$x(n)$ 为离散语音信号时间序列,N 为窗长。$w(n)$ 的选择与短时能量计算直接相关。如果窗长 N 过长,对信号的平滑作用太强,使短时能量几乎没有什么变化,无法反映语音信号的时变特性;反之,如果窗长 N 过小,不能达到足够的平滑作用,语音振幅瞬间变化的细节被保留,难以分辨振幅包络的变化规律。本文采用的窗长,在满足对语音振幅瞬间变化的有效平滑前提下,保证了短时能量的明显变化。

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 = \sum_{m=n-n+1}^{n} [x(m)w(n-m)]^2 \tag{3}$$

$$w(n) = \begin{cases} 0.5 - 0.5\cos(2\pi n/(N-1)) & 0 \leqslant n \leqslant N-1 \\ 0 & \text{其他} \end{cases} \tag{4}$$

相邻的两个离散信号取样值具有不同符号时,便会出现“过零”现象,单位时间内轨迹经过零的次数叫做过零率。语音信号不仅是宽带信号,而且还是时变信号,频谱特性随时间变化。对一帧语音信号而言,只要前一个采样点的值与后一个采样点的值符号相反,便可视为一次过零。遍历整个帧,总的过零次数就是这帧语音信号的过零率。整个语音信号的平均过零率可由每帧的过零率相加再除以总帧数得到。语音信号的短时过零率定义为

$$\text{zcr} = \sum_{m=-\infty}^{+\infty} |\text{sgn}[x(m)] - \text{sgn}[x(m-1)]|w(n-m) = |\text{sgn}[x(n)] - \text{sgn}[x(n-1)]| \times w(n) \tag{5}$$

其中:$s_i(n)$为第 i 帧语音信号;N 为帧长;sgn(x)为 sgn[$x(n)$] = $\begin{cases}1 & x(n)\geqslant 0\\ -1 & x(n)<0\end{cases}$。

利用短时能量与短时平均过零率相结合的判定方式改进端点检测算法,具体步骤如下:

a)将输入的情感语音信号加窗分帧,帧长选为 21.77 ms(240 点),帧移为 7.25 ms(80 点),窗函数选为汉明窗。

b)设定较高能量门限 $amp_1=10$,较低能量门限 $amp_2=2$,以及过零率较高门限 $zcr_1=10$,较低门限 $zcr_2=5$;根据采用的语音片段设置最大静音段 $max_{silence}=8$(80 ms),最小语音长度 $min_{len}=15$(50 ms);设置初始参数状态 status = 0,计数器 count = 0,静音片段 silence = 0,$x_1=0$,$x_2=0$。

c)计算短时过零率、短时能量。

d)根据短时过零率与短时能量中计算得到的 amp 重新调整 amp_1 和 amp_2。

e)进入循环。设定 n 的范围是从 1 到短时过零率采集到的数组总维数。如果 status = 0 或 1,进入 f);否则进入 i)。

f)如果 $amp(n)>amp_1$,则证明进入了语音段,$x_1=\max(n-count-1,1)$,并令状态 status = 2,silence = 0,计数器每次自动加 1,进入 i)。

g)如果 $amp(n)>amp_2$ 或者 $zcr(n)>zcr_2$,则可能进入了语音段,令状态 status = 1,计数器每次自动加 1。

h)如果不能满足 g)的条件,则一定处于静音段,令状态 status = 0,计数器为 0。

i)status = 2,证明已经进入有声段。如果 $amp(n)>amp_2$ 或者 $zcr(n)>zcr_2$,表明有音段继续,进入 l),否则进入 j)。

j)进入无音段,语音可能会结束。如果无音段 silence 小于设定的门限 $max_{silence}$,则可能是有音段中出现的短暂无音,不用处理,进入 l);否则进入 k)。

k)无音段长度大于门限,则判断有音段是否足够长。如果足够长,则有音段结束 status = 3,由 count 向前找到最后一个满足 $amp(n)>amp_2$ 或者 $zcr(n)>zcr_2$ 的帧,作为有音段的终止帧;如果有音段不够长,则认为该段有音段为噪声干扰,不予记录,此前确定的起始帧也无效,令 status = 0,silence = 0,count = 0,进入 l)。

l)如果待处理信号的所有帧已经处理完,则算法结束 count = count − silence/2,$x_2=x_1+count-1$,有效语音起点帧 $n_1=((x_1+2)/3)\times 240$,有效语音结束帧 $n_2=((x_2+2)/3)\times 240$;否则退回到 e)。

图 3 是利用本文端点检测算法对“台球很有意思”这句不同情感的语音样本进行端点检测的结果,其中竖直线表示语音帧开始和结束的位置。

同一语句在不同的情感状态下,生气时,短时能量的最大值最大;平静时,短时能量的最大值最小;难过时,短时过零率的最大值最小。在四种情感状态下,短时能量和短时过零率的变化曲线有差别。

2.5　基于 MFCC 的情感特征提取

MFCC 参数具有良好的识别性能和抗噪声能力,但计算量较大、计算精度要求较高。MFCC 计算过程如图 4 所示。

采用 MFCC 作为情感特征的情感识别参数,其提取过程如下:

a)预滤波。前端带宽为 300 ~ 3 400 Hz 的抗混叠滤波器;A/D 变换并归一化倒频提升窗口,16 kHz 的采样频率,16 bit 的线性量化精度;预加重,通过一阶有限激励响应高通滤波器,使信号的频谱变得平坦,不易受到有限字长效应的影响;分帧,根据语音的短时平稳特性,以帧为单位进行处理。

b)加窗 hamming(256)。采用汉明窗对一帧语音加窗,以减小吉布斯效应的影响,增加音框左端和右端的连续性。假设分帧后的语音信号为 $S(n)$,$n=0,1,2,\cdots,N-1$,那么乘上汉明窗后为 $S'(n)=S(n)\times w(n)$,$w(n)$形式为

$$w(n,\alpha)=(1-\alpha)-\alpha\cos(2\pi n/(N-1))\quad 0\leqslant n\leqslant N-1 \qquad (6)$$

c)快速傅里叶变换(fast Fourier transformation, FFT)。将时域信号变换成为信号的功率谱。

d)三角窗滤波。用一组 Mel 频率上线性分布的三角窗滤波器(共 24 个三角窗滤波器),对信号的功率谱滤波,每个三角窗滤波器覆盖的范围都近似于人耳的一个临界带宽,以此来模拟人耳的掩蔽效应。

e)求对数能量。对三角窗滤波器组的输出求对数,可得近似于同态变换结果。

f)离散余弦变换(discrete cosine transformation, DCT)。去除各维信号之间的相关性,将信号映射到低维空间。

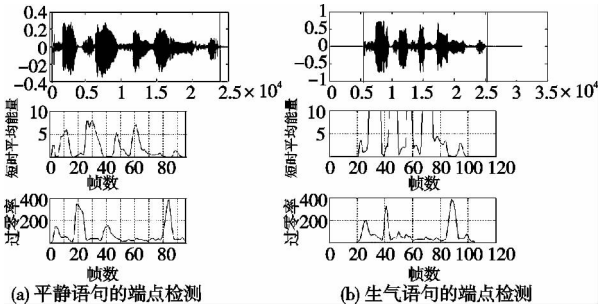


图3　各类情感端点检测结果

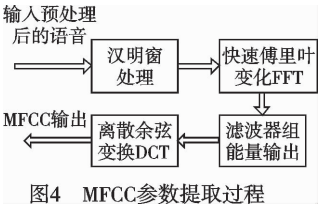


图4　MFCC参数提取过程

3　Elman 神经网络

Elman 网络是具有局部记忆单元和局部反馈连接的前向神经网络。反馈连接由一组结构单元构成,用来记忆前一时刻的输出值,其连接权值是固定的。在这种网络中,除了普通的隐含层外,还有一个特别的隐含层,称为关联层(或联系单元层)。该层从隐含层接收反馈信号,每一个隐含层节点都有一个对应的关联层节点与之连接。通过连接记忆上个时刻的隐层状态同当前时刻网络输入一起,作为隐层输入,相当于状态反馈。隐层的传递函数仍为某种非线性函数,输出层为线性函数,关联层也为线性函数,如图 5 所示。

设反馈层的输出为 $y_c(t)$,网络在外部输入时间序列 $u(t)$ 作用下的输出序列为 $y(t)$,则网络输出 $y(t)$ 如式(7) ~ (9)所示。其中:${}^1W$、${}^2W$ 和 ${}^fW$ 为输入层至隐含层、隐含层至输出层

和隐含层节点之间的连接权矩阵;$f_1(\cdot)$和$f_2(\cdot)$为非线性作用函数。

$$x_o(t+1) = {}^{H}\boldsymbol{W}y_c(t+1) + {}^{1}\boldsymbol{W}u(t+1) - {}^{1}\theta \tag{7}$$

$$y_c(t) = o(t-1) = f_1(x_0(t-1)) \tag{8}$$

$$y(t) = f_2({}^{2}\boldsymbol{W}o(t) - {}^{2}\theta) \tag{9}$$

网络采用动态BP学习算法,步骤如下:

a)设网络的目标函数为式(10),其中y_d和y分别为样本输出与网络输出。

$$J(t) = \sum_{\tau=t_0+1}^{t} E(\tau) = \frac{1}{2}\sum(y_d(\tau) - y(\tau))^2 = \frac{1}{2}\sum e^2(t) \tag{10}$$

b)在Elman网络中采用动态BP算法对权值进行调整。由于${}^{1}\boldsymbol{W}$与${}^{2}\boldsymbol{W}$的调整不受反馈回路的影响,故采用动态BP算法与采用标准BP算法相似,其主要区别在连接层到隐含层的权矩阵${}^{H}\boldsymbol{W}$的调整上。由式(7)~(10)可得

$$\frac{\partial E(\tau)}{\partial {}^{H}\boldsymbol{\omega}_{ij}} = \frac{\partial E(\tau)}{\partial o_i(\tau)} \cdot \frac{\partial o_i(\tau)}{\partial x_{o,i}(\tau)} \cdot \frac{\partial x_{o,i}(\tau)}{\partial {}^{H}\boldsymbol{\omega}_{ij}} \tag{11}$$

其中:

$$\frac{\partial E(\tau)}{\partial o_i(\tau)} = \frac{\partial E(\tau)}{\partial x(\tau)} \cdot \frac{\partial x(\tau)}{\partial o_i(\tau)} = \delta(\tau)\frac{\partial}{\partial o_i(\tau)}[\sum_i^2 \boldsymbol{\omega}_i o_i(\tau)] = \delta(\tau)^2\boldsymbol{\omega}_i(\tau)$$

$$\frac{\partial o_i(\tau)}{\partial x_{o,i}(\tau)} = f'_1(x_{o,i}(\tau))$$

$$\frac{\partial o_i(\tau)}{\partial {}^{H}\boldsymbol{\omega}_{ij}} = \frac{\partial o_i(\tau)}{\partial x_{o,i}(\tau)} \cdot \frac{\partial x_{o,i}(\tau)}{\partial {}^{H}\boldsymbol{\omega}_{ij}} =$$

$$f'_i(x_{o,i}(\tau))\frac{\partial}{\partial {}^{H}\boldsymbol{\omega}_{ij}}[\sum_j {}^{H}\boldsymbol{\omega}_{ij}o_j(\tau-1) + \sum_q {}^{1}\boldsymbol{\omega}_{iq}u_q(\tau-1)] =$$

$$f'_1(x_{o,i}(\tau))\left[o_j(\tau-1) + \sum_j {}^{H}\boldsymbol{\omega}_{ij}\frac{\partial o_j(\tau-1)}{\partial {}^{H}\boldsymbol{\omega}_{ij}}\right] \tag{12}$$

式(12)表示动态递推过程,即动态BP学习算法。

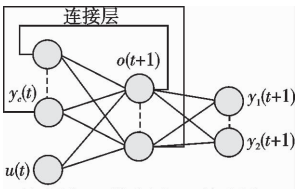


图5 Elman网络结构

3.1 输入层、输出层的设计

利用Elman神经网络模型作为识别分类器时,其输入节点数为特征向量的维数。本文对语音进行预处理后,提取了包括短时能量最大值、短时过零率最大值与MFCC参数在内的12个语音情感的特征值,组成了12维特征向量。神经元网络的输出层节点数目由类别数确定。本文对四种语音情感进行识别,所以Elman神经网络输出层节点数为4。对网络进行训练时,如果网络的输入特征向量属于第K类语音,则在网络输出单元的第K个节点输出1,其余节点输出0。

3.2 隐含层、连接层的设计

隐含层节点数的确定是神经网络设计中非常重要的一个环节。隐含层单元数太少,不能准确提取样本特征,容错性差;隐含层单元数太多,会导致网络规模过于庞大,结构复杂,有时甚至不收敛;另外会使特征空间划分过细,网络的判决曲面将只包含训练样本,导致网络没有判断能力,可能降低训练样本以外样本的识别率,导致过训练。本文根据经验公式$n_1 = \sqrt{n+m} + \alpha$($n$为输入层节点数,$m$为输出层节点数,$\alpha$为1~10的常数)来确定隐含层中节点数目的范围,通过误差对比,确定隐含层神经元的个数为15。隐含层的输出在重新反馈到隐含层之前,在连接层被延时,隐含层节点与连接层节点一一对应,连接权值固定为单位矩阵,其中连接层的节点数等于隐层节点数。

3.3 初始权值的选取

Elman网络在学习之前,必须进行网络初始化的工作。由于系统是非线性的,初始值对于学习是否达到局部最小、训练时间的长短以及是否能够收敛的关系很大,一个重要的条件是希望初始加权后的每个神经元的输出值都接近于零,这样可以保证每个神经元的权值都能够在它们的S型激活函数变化最大处进行调节。初始化主要是将Elman网络模型中所有神经元之间的连接权值均赋予绝对值小于1的随机值。本文将网络的初始权值设置在[-1,1]之间,这样避免了权值过大使训练难以收敛。

3.4 学习速率

学习速率决定了每一次循环训练中所产生的权值变化量。学习速率如果设置过大,会导致系统不稳定;如果设置过小,又会导致网络收敛速度很慢,训练时间大大增加。但学习速率偏小点能保证网络的误差值不跳出误差表面低谷,最终趋于误差最小值。因此,通常在应用中倾向于选取较小的学习速率以保证系统的稳定性。

3.5 期望误差的选择

期望误差设定过小,网络收敛速度慢,甚至无法收敛;期望误差设定过大,网络收敛速度快,但是错判率会增加。在神经网络的训练过程中,期望误差值通常可同时选择几个不同的期望误差进行训练,最后对比这些网络的效果,以确定实际应该采用的期望误差。

3.6 语音情感识别系统设计

语音情感识别系统分为训练阶段和识别阶段。在训练阶段,首先对训练样本的情感语音进行预处理操作,主要是对语音进行预加重、加窗、分帧等;然后提取语音情感特征参数,并对Elman神经网络进行目标训练。在识别阶段,测试语音通过预处理、特征提取和特征分析后,输入到训练好的网络中进行识别判决。情感语音识别系统如图6所示。

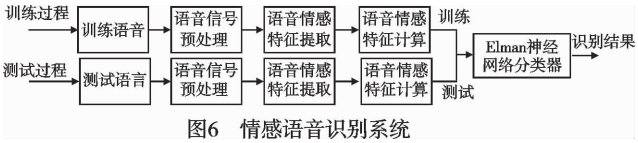


图6 情感语音识别系统

4 语音情感识别系统实验分析

4.1 情感识别系统功能模块设计

识别系统可在后台更改网络类型,并设计了两种识别模式:批处理模式和单个处理模式。其中批处理模式针对统计识别率设计,将对保存在文件夹路径下的所有同一情感样本文件进行识别,输出识别情感状态、总识别率、误识别数、识别率等相关数据用于验证系统性能状态。先选择“批处理”复选框,然后在“请输入测试样本路径”地址栏输入批处理语音文件夹,单击“识别”按钮开始识别,如图7所示。

单个处理模式给出单个文件的波形谱、平均短时能量、过零率、识别结果等相关数据。先选择“单个处理”复选框,然后单击“选择待测试语音”按钮,通过打开的路径选择器选择要测试的文件,点击“识别”按钮开始识别。文件路径将在“单个测试样本路径”框内显示,如图8所示。

图7　训练angry进行批处理识别结果

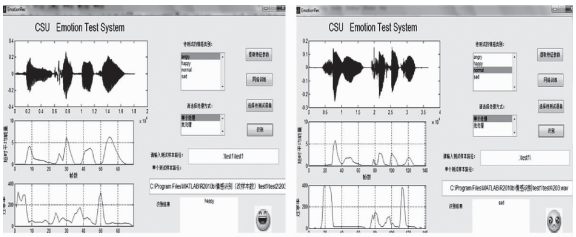


图8　单个处理模式识别结果状态

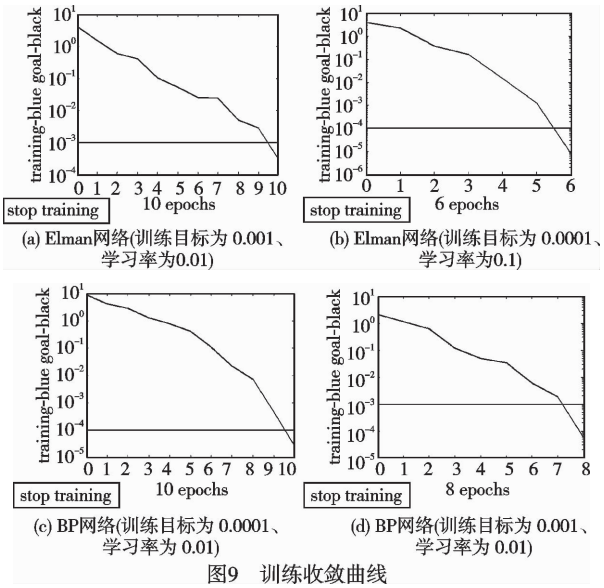
4.2　实验结果分析

对每种情感的 75 句不同文本语音文件进行识别测试,其中前 25 句用于训练,后 50 句用于测试,统计 angry、happy、normal、sad 情感的识别率。为了与 BP 神经网络模型进行比较,文中构建了一个与 Elman 神经网络具有相同输入/输出的 BP 神经网络识别模型,并使用相同的语音情感特征参数训练 BP 神经网络,利用训练后的 Elman 神经网络和 BP 神经网络模型对测试情感语音进行识别率分析。Elman 网络中分别设置训练目标为 0.000 1、学习率为 0.01,训练目标为 0.001、学习率为 0.01,训练目标为 0.000 1、学习率为 0.1 三种不同参数进行实验分析对比,如表 2 所示。

表 2　语音情感识别实验分析与对比

	识别情感	实际情感(训练目标为 0.000 1、学习率为 0.01)			
		angry	happy	normal	sad
Elman 网络语音情感识别结果	angry	**0.76(38)**	0.1(5)	0.12(6)	0.02(1)
	happy	0.08(4)	**0.86(43)**	0.02(1)	0.04(2)
	normal	0.08(4)	0.06(3)	**0.78(39)**	0.08(4)
	sad	0.04(2)	0.12(6)	0.08(4)	**0.76(38)**
	识别情感	实际情感(训练目标为 0.001、学习率为 0.01)			
		angry	happy	normal	sad
	angry	**0.74(37)**	0.1(5)	0.12(6)	0.04(2)
	happy	0.10(5)	**0.84(42)**	0.02(1)	0.04(2)
	normal	0.08(4)	0.06(3)	**0.78(39)**	0.08(4)
	sad	0.04(2)	0.16(8)	0.08(4)	**0.72(36)**
	识别情感	实际情感(训练目标为 0.000 1、学习率为 0.1)			
		angry	happy	normal	sad
	angry	**0.70(35)**	0.14(7)	0.14(7)	0.02(1)
	happy	0.06(3)	**0.80(40)**	0.08(4)	0.06(3)
	normal	0.02(1)	0.10(5)	**0.78(39)**	0.10(5)
	sad	0.08(4)	0.04(2)	0.18(9)	**0.70(35)**
BP 网络语音情感识别结果	识别情感	实际情感(训练目标为 0.000 1、学习率为 0.01)			
		angry	happy	normal	sad
	angry	**0.78(39)**	0.08(4)	0.10(5)	0.04(2)
	happy	0.08(4)	**0.76(38)**	0.06(3)	0.10(5)
	normal	0.06(3)	0.10(5)	**0.76(38)**	0.08(4)
	sad	0.08(4)	0.08(4)	0.10(5)	**0.74(37)**
	识别情感	实际情感(训练目标为 0.001、学习率为 0.01)			
		angry	happy	normal	sad
	angry	**0.74(37)**	0.08(4)	0.16(8)	0.02(1)
	happy	0.06(3)	**0.84(42)**	0.06(3)	0.04(2)
	normal	0.02(1)	0.16(8)	**0.76(38)**	0.06(3)
	sad	0.10(5)	0.06(3)	0.14(7)	**0.70(35)**

从横向对比表 2 可得,对于 angry 情感误识别为 happy 和 normal 的概率达到一定比例;对于 happy 情感误识别而言,Elman 识别率达到 86%;而对于 normal 情感整体上与 sad 的相似度误判程度较高,同时 sad 情感被误判成 normal 的概率更高。Happy 和 angry 虽然是对立的情感,但是都属于激烈的语气,在发音及各种生理特征上有着很多相似之处,两者在振幅、时间和基音构造上的差异也较小。纵向对比实验结果,选定不同的训练目标会对实验结果产生不同影响,对于 BP 网络而言,happy 情感降低了训练目标反而识别率上升,且对被误判为 sad 情感的概念缩小;但对于 angry 和 sad 情感,随着训练目标值的降低而识别率降低,且识别成 normal 的概率上升。其原因可能是随着训练目标设定值的降低,训练精度的下降,导致 BP 神经网络训练过程变得相对粗糙,而 normal 情感是一种相对中性的情感,于是导致各种情感和 normal 情感的相似度变大。对于动态 Elman 网络而言,训练目标降低,对识别情感的影响不大,但增大学习率后,导致系统动态变化略大,影响 Elman 网络的识别效果。但总体而言,动态 Elman 较 BP 网络更适合于语音情感识别系统。图 9 为训练收敛曲线。



(a) Elman网络(训练目标为 0.001、学习率为0.01)　(b) Elman网络(训练目标为 0.0001、学习率为0.1)

(c) BP网络(训练目标为 0.0001、学习率为 0.01)　(d) BP网络(训练目标为 0.001、学习率为 0.01)

图9　训练收敛曲线

4.3　其他相关方法比较

高斯混合模型 GMM (Gaussian mixture model)可以平滑地逼近任意形状的概率度函数,具有容易处理的特点。GMM 通过最大似然度估算来建立说话人概率模型,该模型能较好地反映说话人语音特征在特征空间上的分布,从而获得语音情感识别的能力。该方法对平静、高兴、难过、生气的识别率[16]分别为 58%、73%、72%、72%,各种情感识别率有较大波动,如表 3 所示。总体而言,情感识别效果不如 Elman 神经网络模型。

表 3　GMM 语音情感识别方法识别率

识别情感	实际情感			
	angry	happy	normal	sad
angry	**0.72**	0.12	0.08	0.05
happy	0.13	**0.73**	0.12	0.03
normal	0.12	0.12	**0.58**	0.20
sad	0.03	0.03	0.22	**0.72**

5　结束语

本文以普通话语音情感为研究对象,学习并理解了各类语音情感的特征,在 MATLAB 平台上构建了一个完整的语音情

感识别系统,实现了语音情感识别。本文设计了语音情感识别操作界面,可直观使用该系统来进行语音识别。本文采用了短时能量、短时过零率与 MFCC 情感特征,实现了基于 Elman 神经网络的情感识别交互系统。

实践证明,语音情感识别仍然面临着巨大的挑战:a)语音特征对于各种复杂情感识别的甄别效力问题的研究有待进一步深入,同时需要保持一定特征维数,降低计算复杂度;b)同一句语音中可能包含多种情感成分,对于复杂情感的分离并不存在很明显的界限,即使由人来判断也可能是不准确的;c)说话人的情感受环境与文化背景影响而呈现出地域特性,存在差异。因此,目前国内外语音情感识别研究之路依旧任重而道远。

参考文献:

[1] 韩文静,李海峰. 基于韵律语段的语音情感识别方法研究[J]. 清华大学学报:自然科学版,2009, 49(S1):1363-1368.

[2] AYADI M E, KAMEL M S, KARRAY F. Survey on speech emotion recognition: features, classification schemes and databases [J]. Pattern Recognition, 2011, 44(3): 572-587.

[3] YANG Bin, LUGGER M. Emotion recognition from speech signals using new harmony features [J]. Signal Processing, 2010, 90(5): 1415-1423.

[4] HOZJAN V, KACIC Z. Context-independent multilingual emotion recognition from speech signals[J]. International Journal of Speech Technology, 2003, 6(3): 311-320.

[5] RONG Jia, LI Gang, CHEN Y P P. Acoustic feature selection for automatic emotion recognition from speech[J]. Information Processing & Management, 2009, 45(3): 315-328.

[6] WEI Xiao-peng, ZHAO La-sheng, ZHANG Qiang. Speech emotion recognition using PCA-based spectral features and HMM [J]. Journal of Information and Computational Science, 2009, 6(2): 741-747.

[7] MPORAS I, GANCHEV T, FAKOTAKIS N. Phonetic segmentation of emotional speech with HMM-based methods [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2010, 24(7): 1159-1179.

[8] YUSUKE I, TAKASHI N, MAKOTO T, *et al*. A rapid model adaptation technique for emotional speech recognition with style estimation based on multiple-regression HMM [J]. IEICE Trans on Information and Systems, 2010, E93. D(1): 107-115.

[9] MATROUF D, VERDET F, ROUVIER M. Modeling nuisance variabilities with factor analysis for GMM-based audio pattern classification [J]. Computer Speech and Language, 2011, 25(3): 481-498.

[10] 颜才柄. 基于 BP 神经网络的语音情感识别算法的研究[D]. 武汉:武汉理工大学,2009.

[11] HUANG K C, KUO Y H. A novel objective function to optimize neural networks for emotion recognition[C]//Proc of the 2nd World Congress on Nature and Biologically Inspired Computing. 2010: 413-417.

[12] MURRAY I R, ARNOTT J L. Towards the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion [J]. Journal of the Acoustic Society of America, 1993, 93(2): 1097-1108.

[13] COWIE R, DOUGLAS C E, TSAPATSOULIS N, *et al*. Emotion recognition in human-computer interaction [J]. IEEE Signal Processing Magazine, 2001, 18(1): 32-80.

[14] COWIE R. Describing the emotional states expressed in speech [EB/OL]. (2000). http://www. cs. columbia. edu/ julia/papers/cowie00. pdf.

[15] 张雪英. 数字语音信号处理及 MATLAB 仿真[M]. 北京:电子工业出版社,2010: 24-26.

[16] WU C H, LIANG Wei-bin. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels [J]. IEEE Trans on Affective Computing, 2011, 2(1): 10-21.