文章编号:1006-9348(2009)02-0183-04

归一化振幅商在语音情感识别中的应用

白 洁[1,2],蒋冬梅[1]

(1. 西北工业大学计算机学院,陕西 西安 710072;
2. 海军兵种指挥学院作战指挥系,广东 广州 510430)

摘要:提出了一种新的连续语音情感识别特征:语音元音段声门激励的时域参数归一化振幅商(the normalized amplitude quotient,NAQ)。该方法首先运用迭代自适应逆滤波器(Iterative Adaptive Inverse Filtering,IAIF)估计声门波,然后采用NAQ值来描述声门开启和闭合的特性。采用eNERFACE'05听视觉情感语音数据库中六种不同情感的语音为实验数据,以情感语音元音段的归一化振幅商值为特征,使用直方图和盒形图分析其特征的分布和对情感的区分能力;以情感语句元音段的NAQ值的均值、方差、最大值、最小值作为特征,用高斯混合模型(Gaussian Mixture Models,GMM)和k-近邻法进行了语音情感识别实验,结果表明NAQ特征对语音情感具有较强的区别能力。

关键词:归一化振幅商;迭代自适应逆滤波;高斯混合模型;近邻法

中图分类号:TP391.42 **文献标识码**:A

Normalized Amplitude Quotient Feature in Emotion Recognition

BAI Jie[1,2],JIANG Dong-mei[1]

(1. School of Computer Science,Northwestern Polytechnical University,Xi'an Shanxi 710072,China;
2. Tactical Command Department,Naval Arms-Commanding Academy, Guangzhou Guangdong 510430, China)

ABSTRACT: A time-domain parameter of the glottal flow, the normalized amplitude quotient (NAQ) is presented as a new emotion feature in this paper. Six emotional speeches from the eNTERFACE'05 audio-visual emotion database are inversely filtered using Iterative Adaptive Inverse Filtering (IAIF) to estimate the glottal flow and parameterized using NAQ. To evaluate the properties of the emotion features based on NAQ values, firstly, the histogram and boxplot of NAQ features are plotted to see their ability of distinguishing different emotions. Then, the mean, variance, maximum value and minimum value of NAQ features are used in speech emotion classification using Gaussian Mixture Models andk-nearest neighbor classifier. Experimental results show that NAQ value of vowel segments can be used as an effective emotion feature in emotion recognition from speech.

KEYWORDS: Normalized amplitude quotient(NAQ); Iterative adaptive inverse filtering(IAIF); Gaussian mixture models(GMM); Nearest neighbor algorithm

1 引言

语音是人类交流的重要手段,它作为语言的声音表现形式其中不仅包含了语言学信息,还包含了人们的情感和情绪等非语言信息。随着计算机技术的高速发展,对语音信号中情感信息的处理也越来越重要。基于心理学和韵律学研究的结果表明,在语音中说话者情感最直观的表现就是韵律和语音质量的变化,如音强、音调、音质等的变化。因此语音情感识别研究中很重要的一部分就是语音的韵律特征和音质特征。

在语音情感认知的层面,常见的音质为以下几种形式:呼吸声(breathy voice),叽叽嘎嘎声(creaky voice),粗糙声(harsh voice),松懈声(lax-creaky voice),正常声(modal voice),绷紧声(tense voice),轻声(whispery voice)。有研究表明,在进行语音情感识别时,加入音质特征对于区分那些韵律特征比较相近的情感具有明显改善[1]。文献[2]总结了一些讲英语的人说话时情感的表达和音质间的关系:表示亲密的情感的发音伴有breathy,表示机密情感的发声伴有whispery,生气的情感伴有harsh,厌倦的情感伴有creaky。文献[3]中提出tense对应着生气,高兴,害怕;lax对应悲伤。不同的音质对应不同的声门波,因此文献[4]用声门波参数

基金项目:国家自然科学基金项目(60703104)
收稿日期:2008-01-02 修回日期:2008-01-04

来定量分析音质。逆滤波是估计声门波的一种很好的方法。文献[5]中提出了一个新的声门时域参数 NAQ,它由声门波最大振幅和对应一阶导数的最大负峰值两个振幅值来度量,避免了准确得到声门开启和闭合时刻的难题。文献[6]中使用 NAQ 对三种不同的音质进行分析,实验表明不同的音质对应不同的 NAQ 值。文献[7]中在连续语音中提取 anger, joy, neutral, sadness, tenderness 五种情感语音中元音/a:/的 40 毫秒片段,统计分析的结果表明 NAQ 参数对不同情感具有区分能力。

但文献[6,7]中所做的 NAQ 参数分析实验仅是针对情感语句中单一元音/a:/的 40 毫秒片段,对于声门时域参数 NAQ 在整句连续语音中的应用并没有做探讨。

本文对声门时域参数 NAQ 作为整句连续语音的情感识别特征做了初步研究,提出将情感语音中所有元音段的 NAQ 值作为语音情感特征。采用 eNERFACE'05[8] 的听视觉情感语音数据库,对六种情感 anger, disgust, fear, happiness, sadness, surprise 语音中所有元音段的 NAQ 值,首先通过画直方图包络,盒形图比较其分布和对情感的判别能力,结果表明不同情感语音中元音段的 NAQ 值分布不同,也就是说元音段 NAQ 值可以分辨出不同语音情感。最后以情感语音中所有元音段的 NAQ 值的均值,方差,最大值,最小值为特征,分别使用 k - 近邻法和 GMM 方法进行了语音情感识别实验。

2 基于 NAQ 的语音情感特征

2.1 迭代自适应逆滤波器[9]

本文采用 IAIF 对语音信号进行逆滤波。基本原理是:如果能从原始语音中消除声门激励和口鼻辐射的影响,离散全集点模型(Discrete All - pole Modelling, DAP)分析就可以相当精确地估计出声道特性,并产生可信的声门波。

整个算法包括两部分迭代。如图 1 所示,第一个部分为模块 2 ~ 6,产生声门激励的初步估计,第二部分 7 ~ 12,以第一部分的初步估计为输入更准确地估计声门激励。其中:$s(n)$为声压波,即原始语音信号;$g(n)$为输出,即估计的声门波;$H_{g1}(z)$,$H_{vt1}(z)$,$H_{g2}(z)$,$H_{vt2}(z)$是转移函数。

实验中调节声道共振峰的数量和唇辐射的系数以获得最佳的声门波估计。共振峰的数量一般为 8 ~ 14,唇辐射系数为 0.97 ~ 1.0。

2.2 归一化振幅商[5]

文献[2]给出一个新的声源时域参数 NAQ:

$$NAQ = \frac{AQ}{T} = \frac{f_{ac}}{d_{peak} \cdot T} \qquad (1)$$

式中 T 为基音周期,其中振幅商(amplitude quotient, AQ)被定义为声门波最大振幅和其对应一阶导数的最大负峰值之比,是描述声源特征最有效的参数之一。

$$AQ = \frac{f_{ac}}{d_{peak}} \qquad (2)$$

f_{ac}:是声门脉冲的最大波峰值;

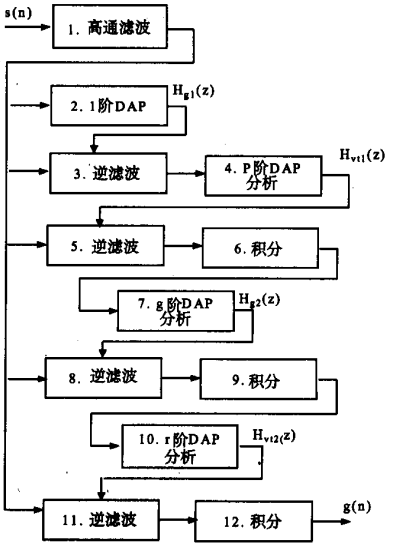

图 1 IAIF 流程图

d_{peak}:声门脉冲对应一阶导数的最大负峰值;因为不需要测量声门波开启或闭合的瞬间时刻,AQ 值比较容易得到,但是 AQ 的值依赖于信号的基频(F0),因此式(1)中将 AQ 用基音周期归一化得到 NAQ,去除了这种对基频的依赖性[5]。

图 2 给出了元音/a:/的一段经 IAIF 处理得到的声门激励与其对应的一阶导数的波形(T:基音周期;f_{ac}:一个周期内声门波最大峰值;d_{peak}:相对应的一阶导数的最大负峰值)。

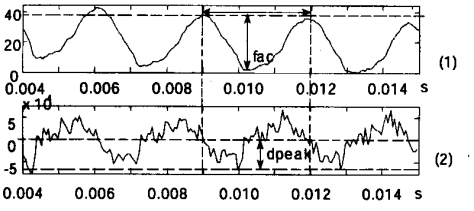

图 2 上图为元音/a:/经 IAIF 逆滤波后得到的声门波,下图为其对应的一阶导数

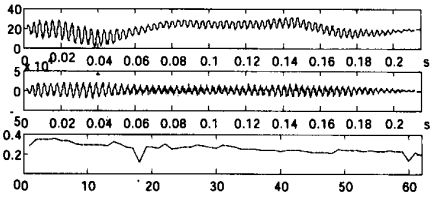

图 3 元音 o 经 IAIF 处理后声门波形,对应一阶导数波形及 NAQ 值

图 3,4,5,6 分别是元音 o、e,爆破音 p,清辅音 s 经 IAIF 逆滤波后的声门波形、对应的一阶导数波形,及其 NAQ 值,由图中可以看出元音段 NAQ 值的变化比较平稳,而且不同元音段的 NAQ 值比较接近,爆破音 p 只求出了两个 NAQ

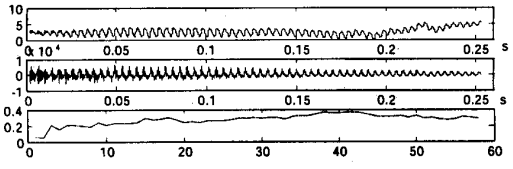

图4 元音 e 经 IAIF 处理后声门波形，对应一阶导数波形及 NAQ 值

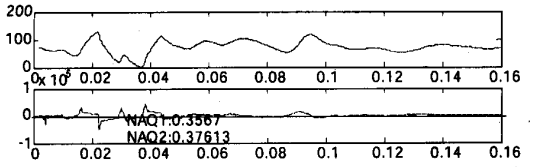

图5 爆破音/p/经 IAIF 处理后声门波形及其对应一阶导数波形

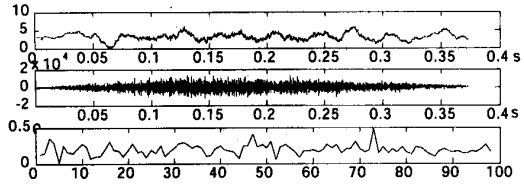

图6 清辅音/s/经 IAIF 处理后声门波形，对应一阶导数波形及 NAQ 值

值，而清辅音 s 的激励类似于白噪声，其求出的 NAQ 值也具有很大的随机性。因此，如果采用整个语句中的所有辅音和元音段的 NAQ 值作为情感特征，这种特征的分布将会比较发散，由语音单元不同引起的 NAQ 值变化，将会超出由情感引起的变化，由此可见语音情感特征不宜采用整个语句的 NAQ 值，我们只采用元音段的 NAQ 值作为语音情感特征。

3 特征判别力分析

3.1 数据样本

采用 eNERFACE'05 听视觉情感语音数据库[9] 中的语音作为实验数据，共六种情感：anger，disgust，fear，happiness，sadness，surprise；由来自 14 个不同国家的 42 个说话人录制，使用英语，每种情感由每个人的 5 句话来表达。本文用 cooledit 从视频文件中提取 16kHz，16 位，单声道的音频用于实验。为了提高实验的可靠性，首先由三名同学通过交叉主观试听，从每种情感中挑出表达效果好的 130 句作为实验数据，其中 100 句作为训练数据，30 句作为识别数据，并将识别数据分给另外三名同学进行情感感知实验。

对所有实验用语句，采用语音处理工具包 HTK[11]，在用 TIMIT 标准语音语料库训练的三音素模型的基础上，进行音素的强迫对准，并对元音段进行切分。因为存在元音与辅音的过渡段，为了保证提取的元音段的可靠性，对每段元音仅取其四分之一至四分之三部分。

3.2 直方图

对六种情感的所有语音元音段的 NAQ 值分布画出其对应直方图的包络图，如图 7 所示。可见 happy，disgust，anger，fear，surprise，sadness 的分布近似于正态分布，其 NAQ 值的方差明显不同，从图中可以看出分布大概分为了 happy，disgust，anger 和 surprise，fear 和 sadness 这四层，即这四层的情感之间可以明显区别开，但是 anger 和 surprise，及 fear 和 sadness 之间较容易混淆，总体来说以情感语音元音段的 NAQ 值为特征对情感语音还是有分辨力的。

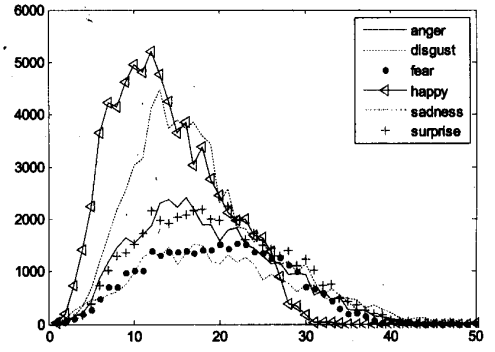

图7 六种情感所有元音段 NAQ 直方图比较

3.3 盒形图

图 8 是用六种情感所有元音段的 NAQ 值画出的盒形图。从图中可以看出六种情感元音段的 NAQ 值除了 happy 和 surprise 的中位数较接近，其它情感元音段的 NAQ 值中位数基本上都不相同，另外六种情感的盒子长度位置都不一样，也就是有不一样的四分位数，说明各种情感元音段的 NAQ 值分布不一样，有不同的均值和标准差。

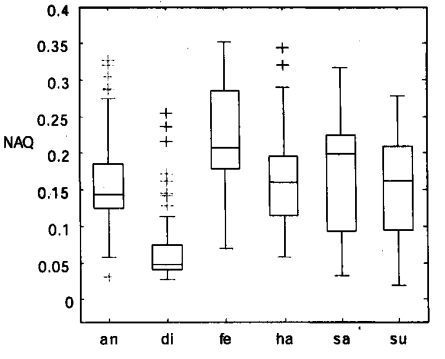

图8 六种情感所有元音段 NAQ 盒形图比较

其中：an，di，fe，ha，sa，su 分别代表 anger，disgust，fear，happy，sadness，surprise

4 语音情感识别

4.1 基于 GMM 的语音情感识别

GMM 是单一高斯概率密度函数的延伸，GMM 可以平滑地近似任意形状的密度分布。一个 M 阶 GMM 的概率密度函数由 M 个高斯概率密度函数的加权和表示：

$$p(x \mid \lambda) = \sum_{i=1} c_i b_i i = 1,2,\cdots,M \qquad (3)$$

其中 x 是 D 维随机向量，混合权重为 $c_i, i = 1,2,\cdots,M$，需满足 $\sum_{i=1}^{M} c_i = 1$，第 i 个高斯分布概率密度函数 $b_i(x)$ 为：

$$b_i(x) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-u_i)\sum_i^{-1}(x-u_i)\right), \quad i = 1,2,\cdots,M \tag{4}$$

式中 u_i 为均值向量，Σ_i 为协方差矩阵。

假设每种情感的模型为 $\lambda = [u, \Sigma, c]$。使用期望最大化（expectation - maximization，EM）算法进行训练，首先初始化用 k - means 聚类算法进行聚类，得到中心向量 $u = (u_1, \cdots, u_M)$ 作为均值 u 的初始值，并计算其协方差 Σ_i，作为 $\Sigma = (\Sigma_1, \cdots, \Sigma_M)$ 的初始值，权重定为 $c_i = 1/M, i = 1,2,\cdots,M$。

识别时，计算输入特征序列在每种模型下的概率，最大者对应情感为识别结果。文中实验取 3 个高斯。

4.2 K—近邻分类器

k - 近邻方法是一种基于统计的分类方法，是比较基础的分类器。其基本思想是：找到和待分类语音文件最相似的 k 个已分类情感语音，根据这 k 个情感语音的类别来判断待分类情感语音的类别值。k - 近邻的错误率上下界在一到两倍贝叶斯决策方法的错误率范围内。从错误率的角度看，k - 近邻法还是优越的。

k - 近邻一般取 k 为奇数，经实验验证 k 从 7 逐渐取大至 51 时结果基本不再变化。

4.3 识别实验与分析

表 1 是以情感语音元音段的 NAQ 值的均值，方差，最大值，最小值作为特征，分别使用 k - 近邻方法，GMM 方法对情感语句进行识别的结果以及情感感知实验结果。可见 happy 的识别率为最低，在 GMM 方法下识别率 6.7%，在 k - 近邻方法下为 10%，其它情感识别率都大于 10%，尤其是使用 GMM 方法时，情感 fear，disgust 的识别率分别达到 60%，50%，fear 情感的识别率已经很接近感知实验结果。说明情感语句元音段的 NAQ 值可以作为语音情感识别的特征之一。

情感语音数据库采用国外的，由于中西方文化的差异，感知实验的识别率也不是很高，fear，surprise 只达到 63%，63.3%，happy 的识别率最高达到 90%，但在使用两种分类方法的识别结果中 happy 的识别率都是最低，其它相对识别率较高的情感反而感知实验识别率较低，这其中有语义的影响，因此在感知实验就需要去除语义的影响。

表 1 k - 近邻及 GMM 情感识别结果 %

识别结果	anger	disgust	fear	happy	sadness	surprise
感知实验	73.3	76.7	63	90	86.7	63.3
GMM	13.3	50	60	6.7	26.7	23.3
KNN	33.3	40	30	10	26.7	23.3

5 结论

本文尝试将声门时域参数 NAQ 用在连续语音的情感识别中，作为语音情感识别的特征之一。对连续情感语音，以元音段的 NAQ 值的均值，方差，最大值，最小值为特征，分别用 GMM 方法和 k - 近邻法对六种情感：anger，disgust，fear，happiness，sadness，surprise 进行识别实验，结果表明连续语音元音段的 NAQ 值可以作为语音情感识别的有力特征之一。在今后的工作中，将进一步将 NAQ 相关的特征和其他语音情感特征，如基频、语调特征等相结合，以提高语音情感识别的识别率。

参考文献：

[1] Tato Requel, Santos Rocio, Kompe Ralf, J M Pardo. Emotion space improves emotion recognition[C]. Proc. ICSLP. Denver, Colorado. 2002,3: 2029 - 2032.

[2] Laver John. The Phonetic Description of Voice Quality[M]. Cambridge University Press, 1980.

[3] Klaus R Scherer. Vocal affect expression: A review and a model for future research[J]. Psychological Bulletin, 1986, 99(2): 143 - 165.

[4] Gobl Christer, Chasaide Ailbhe Ni. The role of voice quality in communicating emotion, mood and attitude[J]. Speech Communication, 2003, 40: 189 - 212.

[5] Alku Paavo, Bäckström Tom, Vilkman Erhhi. Normalized amplitude quotient for parameterization of the glottal flow[J]. Journal of the Acoustical Society of America, 2002, 112(2): 701 - 710.

[6] Lehto Laura, et al. Comparison of two inverse filtering methods in parameterization of the glottal closing phase characteristics in different phonation types[J]. Journal Voice, 2007, 21(2): 138 - 150.

[7] Airas Matti, Alku Paavo. Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalized amplitude quotient[J]. Phonetica, 2006, 63(1): 26 - 46.

[8] O Martin, et al. The eNTERFACE'05 audio - visual emotion database[C]. Proceedings of the 22nd International Conference on Data Engineering Workshops, 2006.

[9] Alku Paavo. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering[J]. Speech Communication, 1992, 11(2 - 3): 109 - 118.

[10] S J Young. The HTK Hidden Markov Model Toolkit: Design and Philosophy[R]. Technical Report, CUED, Cambridge University, 1994.

[作者简介]

白 洁（1977 - ），女（汉族），宁夏银川人，讲师，硕士研究生，主要研究方向为数字语音处理。

蒋冬梅（1973 - ），女（汉族），河南商丘人，副教授，博士，主要研究方向为音频信号处理、语音处理、听视觉融合的语音识别和说话人头部动画。