情感语音的嗓音参数提取与分析*

李向伟[1]，方强[2]，李爱军[2]，王红[1]
1. 山东师范大学 信息科学与工程学院，济南 250014；
2. 中国社会科学院语言研究所，北京 100732；

文 摘：本文主要寻找嗓音音质中能够区分情感的因素，为下一步的情感语音合成作准备。我们基于同一发音人的七种不同情感（七种情感分别为：悲伤、高兴、害怕、厌恶、生气、惊讶、中性）语音样本提取了基频抖动 jitter、振幅抖动 shimmer、谐波噪声率 HNR、基频均值 meanF0、声门波震动幅度 Pulse Amp、声门波形最大下降率 MFDR 等与嗓音声源密切相关的 8 个声学参数并进行统计分析。结果表明在不同情感下一些参数如 NAQ,MFDR 具有显著性差异，而其他参数如 shimmer,h1-h2 差异较小。在两种具体情感对组合的分析过程中，各个参数表现出的差异性也有所不同。

关键词：情感语音；嗓音音质参数

语音是人类交流最自然快捷的方式，人们在交谈中不仅传递语言文字本身携带的信息，同时也包含说话人自身主观的情感状态，透露出说话人的情绪，让我们的交流丰富多彩。在传统的语音信号处理中，往往只关注语音信号要表达的文字意义，忽略了对情感因素分析和建模。现有的语音合成技术得到的声音大多也不带感情色彩。试想如果计算机能够通过说话声音判断人们的情感状态，对不同情感的声音输入信号作不同的响应，将会使得很多应用比如语音导游系统、儿童玩具等更加人性化，更能满足人们的需求。在现代的语音学研究领域中，提取和处理说话人的语音情感特征，并以此作为依据区分不同的情感是一项具有重大现实意义的研究课题。

在我们的说话过程中，气流由肺部通过声门，使声带振动产生声带音，声带音经过声腔共振和唇辐射形成我们听到的声音，在同一个音的不同情感表达中，这一过程任何部分都可能是影响情感听觉的重要方面。在情感语音的声学特征分析中，过去的研究主要集中在基频、音强、音长等特征上[1]，因为这些特征值便于使用计算机软件进行提取，并且在区分某些情感上也具有一定的作用，但还不足以作为准确预测说话人情感的依据。近来的研究表明嗓音音质在情感表达中起着重要作用，国外有 Sona Patel [2] 等对法语发音人不同情感下的单音节/a/进行参数提取和分析，一共得到 12 个参数,其中 11 个在情感语音区分中作用显著，并将这 11 个参数归为三类主要影响因素。国内的研究有王磊[3]等以 jitter 为目标，研究不同情感下 jitter 的分布，将 jitter 分为确定部分和随机部分，所谓确定部分就是由情感决定的部分，不同的情感有特定的 jitter 模式，随机部分则与情感无关。本文通过量化元音嗓音音质，以提取和分析声门波参数为主，对汉语普通话中的阴平元音/a/进行分析，寻找嗓音音质特征中能够区分不同情感的成分以及这些成分的作用。

提取和分析过程如下，我们首先从发音人在七种不同情感（七种情感分别为悲伤、高兴、害怕、生气、厌恶、惊讶、中性）下朗读语料中提取得到要分析的音节，本文选取声调为阴平的元音/a/，得到七组样本，接下来我们假设 8 个参数（分别为 jitter, shimmer, HNR, meanF0, Pulse Amp, MFDR, NAQ 和 h1-h2，在后续研究不同声调的元音中，音高的变化范围将作为一个参数加入）可能在不同情感中具有显著性差异，从样本中分别提取这 8 个参数，作统计分析得到不同情感下每个参数的差异显著性。

1 情感语音数据库

研究语料来自中国社会科学院语音与言语科学重点实验室情感语音数据库[4]。本文选用其中一位女性发音人，在七种不同情感状态下朗读相同文本得到的样本。每种情感得到的样本数目相同，从中分离出研究使用的阴平元音/a/进行特征提取和

*基金项目： 国家自然科学基金项目（No.60975081），国家自然科学基金重点项目(No.61233009)，中国社会科学院创新工程，山东省科技发展计划（No.2012GGB01058）；山东省研究生创新基金（No.SDYY10059）
作者简介：李向伟（1989），男（汉），山东省潍坊市，在读硕士研究生。
通讯联系人：方强，助研，fangqiang@cass.org.cn

分析，每种情感得到 20 个有效样本。语音材料在相同的环境下录制，初始采样率为 44 kHz，为了便于分析在研究过程中重采样为 16 kHz。

2 参数提取

本文主要考察不同情感对嗓音特征的影响，以及各嗓音特征对情感分类的贡献。影响情感发声的因素有很多，本文选取了 8 个参数，主要与嗓音音质相关。8 个参数如下：

2.1 Jitter（基频抖动）

基频抖动描述了基频从一个区域到另一个区域之间快速和反复的变化。考虑到在不同的情感下说话人的声带震动频率变化可能存在显著性差异。计算公式如下：

$$\text{js} = \sum_{i=2}^{N} |T_i - T_{i-1}|/(N-1),$$

$$\text{meanP} = \sum_{i=1}^{N} T_i/N,$$

$$\text{jitter} = js/meanP$$

其中T_i是第 i 个基音周期的时长（单位为 ms），N 是所有基音周期的数量，js 计算出所有相邻的两个基音周期差值的绝对值均值，meanP 计算得到所有基音周期时长的平均值，然后两值相比，得到发音期间的基频抖动，jitter 的值分布在 0 到 2 之间。

2.2 shimmer（振幅抖动）

与 jitter 相似，shimmer 描述了振幅的快速反复变化,公式如下：

$$\text{shimmer} = \left[\sum_{i=2}^{N} |A_i - A_{i-1}|/(N-1)\right] \Big/ \left[\sum_{i=1}^{N} A_i/N\right]$$

其中A_i是提取的振幅峰值，N 为提取的脉冲个数。

2.3 HNR（harmonics-to-noise ratios，谐波噪声率）

谐波噪声率体现了信号中周期部分与噪声部分占的比重，很大程度上反映出声音的嘶哑程度。我们定义延迟信号τ的自相关函数$r_x(\tau)$为

$$r_x(\tau) = \int x(t)x(t+\tau)dt \qquad (1)$$

其中$x(t)$是一个稳定的时间信号，这个函数在τ =0 时有一个全局最大值。如果函数除在τ =0 的时，其他时间也有最大值，那么存在一个时间 T_0 是时间信号的周期。对于任何整数 n，有

$$r_x(nT_0) = r_x(0) \qquad (2)$$

如果除了τ =0 外没有最大值点，也可能存在一个局部最大值点，定义

$$r_x'(\tau) = \frac{r_x(\tau)}{r_x(0)} \qquad (3)$$

我们定义一个周期为T_0的周期信号 H(t)和一个噪声信号 N(t)，即有$r_x(0) = r_H(0) + r_N(0)$,我们定义

$$r_x'(\tau_{max}) = \frac{r_H(0)}{r_x(0)}; 1 - r_x'(\tau_{max}) = \frac{r_N(0)}{r_x(0)} \qquad (4)$$

$r_x'(\tau_{max})$表示信号中周期部分的相对能量，它的补$1 - r_x'(\tau_{max})$表示信号中噪声部分的相对能量。进一步可定义 HNR。[5]

$$\text{HNR(in dB)} = 10 * log_{10} \frac{r_x'(\tau_{max})}{1 - r_x'(\tau_{max})}$$

2.4 mean F0（基频均值）

meanF0 是元音基频的平均值，即声音发声期间声带震动的平均频率，它的大小反映了声带震动的快慢。

2.5 Pulse Amp（振动幅度）

声门波形图中表现为周期内振幅的极大值与极小值之差，如图 1 所示，它的值越大，表示声门震动的幅度越大。我们首先使用线性预测编码逆滤波得到声门波，然后从声门波形中提取这个参数，线性预测模型最常见的表示为

$$\hat{x}(n) = -\sum_{i=1}^{p} a_i x(n-i)$$

其中 $\hat{x}(n)$是预测的信号值，$x(n-i)$ 是前面观测到的值，a_i 是预测系数。这种预测产生的误差是

$$\text{e(n)} = \text{x(n)} - \hat{x}(n)$$

x(n)是真正的信号值，这个等式对于所有类型的一维线性预测都是有效的，它们的不同之处是参数a_i的选择方式不同。

2.6 MFDR（maximum flow declination rate，最大下降率）

MFDR 即一个周期内声门波导数最小值的绝对值，如图 1 中所示。MFDR 表示声带关闭的快慢。

2.7 NAQ（标准化振幅商）

标准化振幅商是一种量化声门关闭相的方法，它处理的对象是两个振幅域中经过逆滤波后的声门波。[6]

计算公式：

$$\text{NAQ} = Pulse\ Amp/(T_0 * MFDR)$$

其中 Pulse Amp 即振动幅度，T_0为周期，MFDR 为声门波最大下降率。

2.8 h1-h2（第一谐波与第二谐波的差值）

对声音的时域信号作快速傅里叶转换，得到声音的频域信号，然后计算第一谐波与第二谐波的能量差 h1-h2。

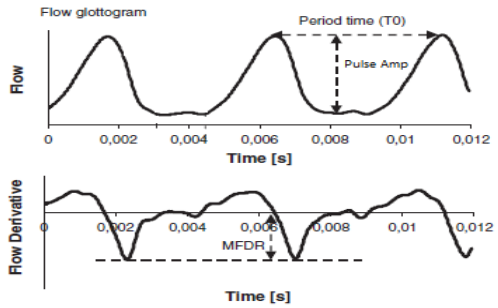

图 1 声门波参数图示[7]

3 分析结果

为了研究参数在不同情感下的差异显著性，需要对每个参数作单因素方差分析，在作单因素方差分析前为了确保数据可用，我们首先对数据样本作单样本 k-s 检验。

单样本 k-s 检验结果显示除高兴情感下的 jitter 参数(精确双尾显著性 p=0.037<0.05)、害怕情感下的 shimmer 参数(精确双尾显著性 p=0.016<0.05)以外，其余各类样本都符合标准正态分布。Jitter 和 shimmer 的检验结果如下表所示：

表 1 单样本 k-s 检测结果(p 值)

情感	Jitter	shimmer
Sad	0.478	0.828
Joy	***0.037***	0.247
Fear	0.307	***0.016***
Anger	0.585	0.4
Neutral	0.147	0.563
Disgust	0.75	0.57
Surprise	0.494	0.223

注：表中斜体加粗表示 p<0.05 的值

从表中可以看出高兴情感下的 jitter 参数与害怕情感下的 shimmer 参数显著性小于 0.05，拒绝原假设 H_0，样本不符合标准正态分布。剩余其他参数的样本双尾显著性都大于 0.05，没有理由拒绝原假设，图形在这里不再画出。

接下来依次对每个参数作单因素方差分析，检验不同情感下参数值之间是否存在显著性差异。从结果来看，有些参数在不同情感中差异显著，比如 NAQ,MFDR 等，有些参数则差异相对较小，比如 shimmer, h1-h2 等，图 2 为七种情感下 NAQ,MFDR,shimmer,h1-h2 的分布，可以形象的看出它们之间区分作用的大小，表 2 是 8 个参数的单因素 ANOVA 分析结果(F 值和 p 值)，用统计分析的方法量化区分显著性大小，参数对应的 F 值越大，p 值越小，表明这个参数对不同情感的区分性越好。

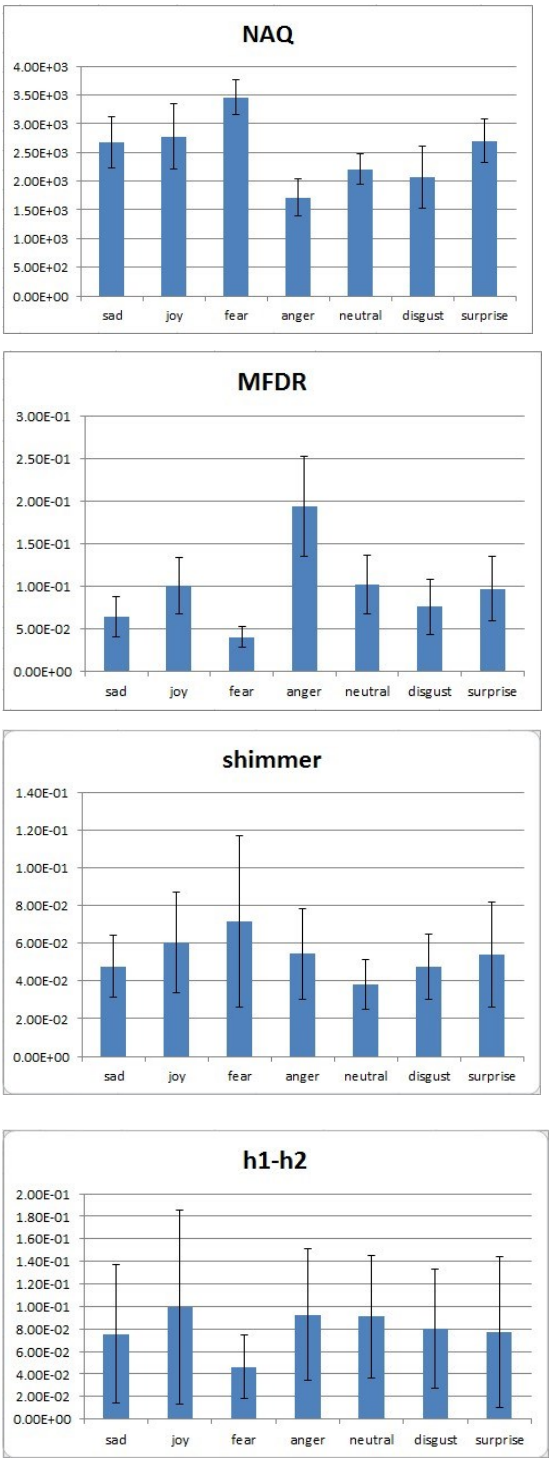

图 2 不同情感下 NAQ,MFDR,shimmer,h1h2 的分布

表 2 单因素 ANOVA 分析结果

参数	F 值(组间 df=6，组内 df=133)	p 值
NAQ	37.764	0
MFDR	36.708	0
meanF0	22	0
Pulse Amp	18.616	0
HNR	7.024	0
Jitter	5.789	0
Shimmer	3.255	0.005
h1-h2	1.677	0.131

注：表中参数按照 F 值递减排列

由表 2 可以看出除 h1-h2 明显不显著外(p=0.131>0.05)，其余的七个参数情感都有显著性。为了进一步观察参数在具体哪些情感对作用显著，我们做了两两比较单因素方差分析，结果说明所有的参数都可以区分至少一种情感对，但区分能力有强有弱，有的参数几乎可以区分所有的情感对，比如 NAQ，有的则只能区分很少的情感对，如 h1-h2 只有在高兴-害怕、害怕-生气、生气-中性三种情感对中区分作用显著。同样有些情感对可以被几乎所有的参数区分，比如害怕-中性，而有的情感对只有少数参数可以区分，如厌恶-惊讶。表 3 列出两两比较单因素 ANOVA 分析结果(p<0.05 说明参数对情感对区分作用显著)。

4 总结和前景

本文研究内容是嗓音音质成分中 8 个参数在不同情感下的差异，结果表明不同情感下一些参数如 NAQ,MFDR 具有显著性差异，而其他参数如 shimmer,h1-h2 差异较小。其中由表 2 看出 h1-h2 差异最小（p=0.131>0.05）。在分析两种具体情感组合的过程中，各个参数表现出的差异性也有所不同，NAQ 与 MFDR 在情感对组合中依然是差异最显著的两个参数，在大多数情感对中具有显著性差异，shimmer 与 h1-h2 差异较小，其中 h1-h2 只有在高兴-害怕、害怕-生气、害怕-中性三对情感组合中具有显著性差异。分析结果表明，NAQ，MFDR，meanF0，Pulse Amp，HNR，Jitter，Shimmer 七个参数在不同情感下具有显著性差异，h1-h2 差异不显著。基于以上结果，我们有可能在语音合成过程中通过调节上述相关嗓音声源参数，合成情感语音。

本文只研究了阴平下的元音/a/，接下来将扩大样本量，分析所有元音在不同声调下和不同情感状态下的嗓音声源参数，找出共同的参数。此外，我们还将利用主成分分析法分析不同参数之间变化的关联性，减小参数调整的复杂度。

5 致谢

本工作得到国家自然科学基金项目（No.60975081），国家自然科学基金重点项目(No.61233009)，中国社会科学院创新工程，山东省科技发展计划（No.2012GGB01058）；山东省研究生创新基金（No.SDYY10059）的支持。

表 3 两两比较单因素 ANOVA 分析结果(p 值)

情感对	NAQ	MFDR	shimmer	h1h2	Jitter	Pulse Amp	HNR	meanF0
Sad-joy	0.457	***0.002***	0.133	0.213	0.752	***0.007***	0.578	***0***
Sad-fear	***0***	***0.036***	***0.005***	0.125	***0***	***0.041***	***0***	0.06
Sad-anger	***0***	***0***	0.429	0.375	0.729	***0***	***0.009***	0.223
Sad-neutral	***0***	***0.001***	0.246	0.419	0.121	***0.029***	0.074	0.225
Sad-disgust	***0***	0.315	0.973	0.801	0.592	0.71	***0.042***	***0.005***
Sad-surprise	0.868	***0.004***	0.472	0.943	0.99	0.224	0.141	***0***
Joy-fear	***0***	***0***	0.186	***0.006***	***0***	***0***	***0.001***	***0.001***
Joy-anger	***0***	***0***	0.473	0.719	0.975	***0***	***0.038***	***0***
Joy-neutral	***0***	0.917	***0.008***	0.66	0.063	0.577	***0.02***	***0***
Joy-disgust	***0***	***0.03***	0.124	0.32	0.825	***0.018***	0.136	***0***
Joy-surprise	0.563	0.767	0.43	0.241	0.743	0.125	0.358	0.059
Fear-anger	***0***	***0***	***0.042***	***0.016***	***0***	***0***	0.213	0.501

情感对	NAQ	MFDR	shimmer	h1h2	Jitter	Pulse Amp	HNR	meanF0
Fear-neutral	***0***	***0***	***0***	***0.02***	***0***	***0***	***0***	0.498
Fear-disgust	***0***	***0.002***	***0.005***	0.075	***0.001***	***0.016***	0.066	***0***
Fear-surprise	***0***	***0***	***0.036***	0.109	***0***	***0.001***	***0.017***	***0***
Anger-neutral	***0***	***0***	0.052	0.936	0.059	***0***	***0***	0.997
Anger-disgust	***0.01***	***0***	0.41	0.525	0.85	***0***	0.55	***0***
Anger-surprise	***0***	***0***	0.943	0.415	0.72	***0***	0.242	***0***
Neutral-disgust	0.271	***0.023***	0.26	0.578	***0.038***	0.069	***0***	***0***
Neutral-surprise	***0***	0.689	0.061	0.462	0.124	0.327	***0.001***	***0***
Disgust-surprise	***0***	0.06	0.451	0.857	0.584	0.398	0.566	***0***

注：斜体加粗表示区分作用显著(p<0.05)

参 考 文 献

1. Juslin, P.N. and Scherer, K.R., *Vocal expression of affect.* The new handbook of methods in nonverbal behavior research, 2005: p. 65-135.
2. Patel, S., et al., *Mapping emotions into acoustic space: The role of voice production.* Biological psychology, 2011. **87**(1): p. 93-98.
3. WANG, L., LI, A., and FANG, Q., *A Method for Decomposing and Modeling Jitter in Expressive Speech in Chinese.* Proc. Speech Prosod, 2006.
4. Li, A., Fang, Q., and Dang, J., *Emotional Intonation in a Tone Language: Experimental Evidence From Chinese.* ICPhS XVII, Hong Kong, 2011: p. 17-21.
5. Boersma, P. *Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound*. in *Proceedings of the institute of phonetic sciences*. 1993: Amsterdam.
6. Alku, P., Bäckström, T., and Vilkman, E., *Normalized amplitude quotient for parametrization of the glottal flow.* the Journal of the Acoustical Society of America, 2002. **112**: p. 701.
7. Björkner, E., et al., *Voice source differences between registers in female musical theater singers.* Journal of Voice, 2006. **20**(2): p. 187-197.

The extraction and analysis on voice parameters of emotional speech

Li Xiangwei[1] , Fang Qiang[2],Li Aijun[2], Wang Hong[1]

1. College of Information Science and Technology, Shandong Normal University, Jinan 250014, China

2. Institute of Linguistics, Chinese Academy of Social Sciences, Beijing 100732, China

Abstract: In this study, we mainly looked for the distinguishable factors in voice quality, to prepare for emotional speech synthesis. Eight parameters which are based on one speaker's voice samples in seven different kinds of emotions (They are sad, joy, fear, disgust, surprise, anger and neutral) were extracted and analyzed. They are jitter, shimmer, HNR, mean F0, Pulse Amp, MFDR， NAQ,h1-h2, which have close relationship with glottal voice. A series of statistic analysis showed that some of these parameters such as NAQ and MFDR are more significant in distinguishing emotions and in pair-wise comparisons between each emotion pair than others.

Key words: emotional speech; voice quality parameter