基于统计方法的普通话情感语调模型

苏庄銮 [1] 汪增福 [1]

摘 要 提出了一种基于数据驱动的语调建模方法. 该方法采用主成分分析 (Principal component analysis, PCA) 技术, 给出了特征语调, 统计了语音情感模式在特征语调空间中的分布规律, 经过分析得出了普通话中情感模式所对应的情感语调. 针对语音产生的机理复杂、语音语调受众多因素影响的特点, 为了避免这些干扰因素的影响, 设计了相应的情感语音库. 利用所设计的语音库, 进行了相关实验. 实验结果表明, 利用所提出的特征语调模型不仅能够非常完美地重构出语调样本的语调, 而且具有相当的情感表达能力.

关键词 主成分分析, 语音合成, 情感计算, 特征语调
中图分类号 TP18

Affective Intonation-modeling for Mandarin by Statistical Method

SU Zhuang-Luan[1] WANG Zeng-Fu[1]

Abstract This paper proposes a data-driven intonation modeling method. With the principal component analysis (PCA) technique, the concept of eigenintonation is presented. The distribution of emotional states in the eigenintonation space is then studied and the corresponding emotional intonations in Mandarin are given out. In order to avoid the influences caused by the unwanted factors, the affective corpora for the purpose of evaluation are designed. With the corpora, the related experiments have been performed. The experimental results show that the eigenintonation model has a quite good ability of expressing emotions, and all intonation samples in our corpora can be well recovered with the eigenintonations.

Key words Principal component analysis, speech synthesis, affective computing, eigenintonation

1 引言

语音是最重要的人际交流工具之一. 人类的话语不仅起着表字达意的作用, 而且还包含了说话人的状态和情感等信息. 语音中所包含的情感信息对提高说话人识别和语音内容识别的效率, 改善合成语音的自然度等具有重要的意义. 基于语音的情感研究要解决的基本问题, 是要找到情感和语音模式之间的对应关系. 其中, 如何抽取有效的语音特征并运用恰当的模型来表达语音特征和情感之间的关联性是亟待解决的一个关键问题. 通常, 特征的抽取通过观察语音随说话人感情的变化而得到. 研究表明, 在普通话中语音基频特性对情感的表达起着非常关键的作用[1]. 反过来, 在普通话中说话人情感的表达方式也极大地影响着语音的基频包络[2].

普通话是典型的有调语言, 共有四个字调和一个轻声. 赵元任认为, 普通话句调是字调和语调的"代数和"[3]. 其中, 字调在连续语流中会受前后的影响而变调, 同时字调的调域受语气的影响. 字调叠加在语调上就像 "小浪加大浪". 赵元任给汉语中字调与语调之间的关系明确地下了定义[4]: 汉语的字调是表义的, 而语调是表情的. "语调只能表达语气、情调、用途, 等等". 这种将基音包络分解为全局趋势和局部变化这样两个元素的方法在其他语言中也很常见[5, 6]. 因此, 实现字调和语调的分离对于进行语音情感研究具有重要的意义, 引起了众多研究者的关注. 例如, Tian 和 Nurminen 使用统计的方法对字调进行训练建模, 取得了较好的效果[7]. 此外, 笔者也曾提出了一个面向普通话的情感声调模型, 并利用该模型实现了普通话中字调和语调的分离[8]. 但是, 从总体上来看目前国内有关普通话情感语调的建模研究还相对较少.

针对上述研究现状, 本文借助主成分分析 (Principal component analysis, PCA) 技术提出了一种基于数据驱动的语调模型. 实验结果表明, 该特征语调模型不仅能够非常完美地重构出语调, 而且具有相当的情感表达能力.

2 语音库的设计及其基频统计

为了进行语音情感方面的研究, 首先需要根据某些特性标准对情感做一个有效合理的分类, 然后

收稿日期 2006-1-9 收修改稿日期 2006-7-14
Received January 9, 2006; in revised form July 14, 2006
模式识别国家重点实验室开放基金, 中国科学院自动化研究所和中国科学技术大学智能科学与技术联合实验室开放基金 (JL0602) 资助
Supported by Open Foundation of National Laboratory of Pattern Recognition, P. R. China, and Open Foundation of Joint Laboratory of Intelligent Science & Technology, Institute of Automation, Chinese Academy of Sciences and University of Science and Technology of China (JL0602)
1. 中国科学技术大学自动化系 合肥 230027
1. Department of Automation, University of Science and Technology of China, Hefei 230027
DOI: 10.1360/aas-007-0673

再在分类的基础上研究语音特征参数的性质. 目前使用较多的情感模型主要有 Plutchik 的 “情感轮” 模型和 Fox 的三级情感模型[9]. 具体在普通话情感研究中, 国内使用较多的则是四情感模型. 它们将情感分类成 “欢快”、“愤怒”、“惊奇”、“悲伤”[9, 10], 或者是 “欢快”、“愤怒”、“恐惧”、“悲伤”[1, 2] 四类. 综合语音情感研究的现状和上述情感模型的特点, 本文采用五情感分类法, 将情感分类成 “欢快”、“愤怒”、“惊奇”、“恐惧” 和 “悲伤” 五种.

本文研究基于基频曲线全局趋势变化的语调建模问题. 考虑到韵律在结构上具有分层的特点, 句法结构复杂、字数较长的语句可以按照韵律边界划分为多个结构简单、长度短小的韵律单元. 因此, 以韵律单元为对象建立语调模型能够把复杂问题分为多个简单问题进行求解. 在韵律单元中, 又以韵律短语具有相对稳定的语调模式[11]. 所以, 本文把语调模型建立在韵律短语层次上, 具有一般意义.

影响语调的因素很多, 诸如语法句式、重音、说话人情感以及说话人个性特征等都可能是其中的一个方面. 本文关注语调受情感影响的部分, 因此, 需要尽量消除与情感关联不大的成分. 考虑到目前还没有可靠的技术手段可以通过语音信号处理消除情感之外的因素对语调的影响, 所以, 本文设法通过语音库的设计来消除这些干扰因素.

为了设计符合研究需求的语音库, 对所使用语句的内容和长度做了一定的限制, 以避免干扰因素和简化算法. 语音库所采用的语料来自具有不同内容的 40 个句子. 语料内容本身没有特定的情感倾向, 并能够由表演者根据自己的意愿或语料录制者的要求在其上附着上任意指定的情感. 为消除语法结构对语调的影响, 实验中所采用的语料均由双音节名词主语、双音节动词谓语和双音节名词宾语六个音节所构成. 例如,“北京召开奥运”. 将每句语料设计成这样的长度, 是因为韵律短语的长度一般约为六个音节[11]. 这种设计一方面可以简化后续处理步骤, 另一方面也可以突出关注语调受情感影响的研究主旨. 为了使最终的实验结果不受说话人个性特征的影响, 所有语句的发声均由同一名专业女演员完成. 每个句子被要求分别按照 “欢快”、“愤怒”、“惊奇”、“恐惧”、“悲伤” 以及 “不带感情” 的六个方式朗读. 最终获得的语音库包含 240 个句子, 共 1440 个音节. 所有语音均在 16 KHz 的采样率下采用 16 位精度数字化.

为了评估所设计的语音库的情感模式的有效性, 使用改进的自相关算法提取基频, 并对基频的均值、最大值和最小值进行了统计, 其结果如图 1 所示.

由图 1 可见, “惊奇”、“欢快” 和 “愤怒” 的基频相对很高, 而 “悲伤” 的基频则比自然状态的要低些. 统计结果还显示 “悲伤” 的基频动态变化范围比其他情感状态的要小, “恐惧” 的基频几乎与 “悲伤” 具有一样的特点. 另一方面, “愤怒”, “欢快” 和 “惊奇” 则表现得很类似. 所有这些统计结果都与已报道的情感语音研究结果[1, 2, 10] 相吻合, 说明本文所设计的语音情感数据库对情感的表达具有代表性和一般性.

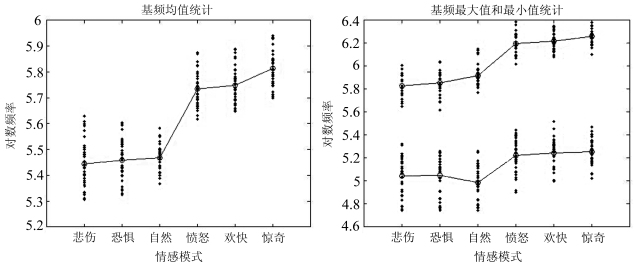

图 1 语音库基频统计结果

Fig. 1 Statistic results for F_0 of the speech corpora

3 特征语调分析

特征语调的概念是从 PCA 技术衍生而来的. PCA 是一种多变量分析方法 [12], 它能给出原始数据集合的低维压缩表达形式. PCA 可以将高维空间中的一组具有相关性的向量转变成一组按照能量降序排列的不相关的向量.

在本文设计的语音库中, 由于语法结构、说话人个性等因素的影响都是一致的, 所以所有语调的形状应该非常相似, 即具有一定的相关性, 它们应该能够被一组 “基本语调” 所描述. 我们知道 PCA 的作用之一就是可以从原始相关变量集合中提取出新的不相关特征. 因此, 一个自然的想法就是, 使用 PCA 对所有语调进行分析, 获取这样一组 “基本语调”. 在误差的允许范围内, 使用这些 “基本语调” 能逼近所有原始语调. 本文把这组作为主成分的 “基本语调” 称为 “特征语调”.

在数学上, PCA 是针对一个方差矩阵求解其特征值和特征向量, 将所得到的特征值进行降序排列, 并选取排在前面的特征值所对应的特征向量作为主成分.

3.1 语调提取算法

从原始语音信号中提取语调的步骤如下:

1) 使用改进自相关法从语音中提取浊音段的基频值, 并对无声段和清音段进行插值;

2) 使用分段三次多项式对整段基频曲线进行迭代拟合, 获得连续光滑基频曲线;

3) 计算基频曲线的自然对数, 然后对其进行截止频率为 f_c 的高通滤波.

由 [8] 知, 步骤 3) 中滤波残留分量就是句子基

音包络中的全局趋势部分. 我们知道, 由于实际语音的复杂性, 严格分离其基音包络中的声调和语调并不是一件容易的事. 所以, 本文在分离语调时采取了一种保守策略, 即在保证结果不含局部变化的前提下, 尽可能提取出基音包络的全局趋势. 结合本文语音库特点, 在 f_c 取 0.5 Hz 时, 把算法结果作为语调来研究. 实验中将语调长度标准化为 N 个样点.

3.2　语调的主成分分析

设语调在原始语调空间 O 中用 N 维向量 I 表示, 其中, N 为语调长度的标准化样点数. 则语调样本集的总体方差矩阵 C 为

$$C=\frac{1}{M}\sum_{i=1}^{M}(I_i-m)(I_i-m)^{\mathrm{T}} \tag{1}$$

其中, I_i 为第 i 个语调样本向量, M 为所有语调样本的总数, m 为语调样本集的平均值, C 为 $N\times N$ 的矩阵. 求解 C 的特征值, 并按降序排列, $\{\lambda_1,\lambda_2,\cdots,\lambda_N\}$. 选取排在前面的 L $(L<N)$ 个特征值所对应的特征向量作为主成分, 由它们构成特征向量集合, 记为 U. 以 U 中特征向量作为基向量构造特征子空间 P, 由于这些特征向量反映了语调样本的语调特征, 为方便起见, 将特征子空间 P 称为语调子空间.

3.3　重构分析

在主成分分析的基础上, 将原始空间 O 中所有的语调样本 I_k $(k=1,2,\cdots,M)$ 向由上述 U 构成的语调子空间 P 投影, 有

$$\Omega_k=U^{\mathrm{T}}(I_k-m),\quad k=1,2,\cdots,M \tag{2}$$

借助于上述投影 $\Omega_k(k=1,2,\cdots,M)$ 以及特征向量集合 U, 可以得到原始语调样本 I_k 的估计 J_k, 从而实现对原始语调样本 I_k 的逼近

$$J_k=U\Omega_k+m,\quad k=1,2,\cdots,M \tag{3}$$

为了衡量重构的准确性, 引入语调样本重构率的概念: 若第 k 个语调样本的重构率记为 R_k, 则对所有语调样本进行重构的重构率 r 定义为

$$r=\sum_{k=1}^{M}\frac{R_k}{M}\times 100\%=\sum_{k=1}^{M}\frac{1-\|I_k-J_k\|/\|I_k\|}{M}\times 100\% \tag{4}$$

4　情感语调

4.1　特点

情感语调是指拥有该语调的语音能够表达某一类情感. 已有的韵律研究有许多关于情感语调的定性分析, 本文则尝试借助特征语调方法给出定量的情感语调曲线.

为了研究情感问题, 对情感进行分类是一种可行的方法. 但是由于人类情感的复杂, 不论是 Plutchik 的 “情感轮” 模型还是 Fox 的三级情感模型[9], 都认为大多数时候人类情感并不是简单地属于哪一类, 而是几类的混合结果. 其中, 所包含的每个类别的强度都会随情感的变化有所不同. 为了适应情感的这种表达需要, 情感语调向量需要能够被控制. 由 3.3 节的分析可以知道, 原始语调可以被投影到特征语调 U 所张成的子空间 P 中. 由于子空间 P 的维数比原始空间 O 低许多, 使得控制目标向量成为可能. 而如果直接对语调向量在原始空间 O 内求平均或者聚类来定量情感语调, 会由于语调向量维数太高而无法直接对其控制, 所以, 借助特征语调建立情感语调模型的方法实际中更合用一些.

4.2　计算方法

设情感类别为用 α, $\alpha=1,2,\cdots,6$ 分别指代 “欢快”、“愤怒”、“惊奇”、“恐惧”、“悲伤” 和 “自然” 等情感状态, 用 N 维向量 I^α 表示从情感状态为 α 的语音中提取出的语调在原始语调空间 O 中的取值. 根据式 (2), 将 I^α 投影到子空间 P 中, 记为 Ω^α. Ω^α 在空间 P 内分布于不同的区域, 计算 Ω^α 的聚集核心向量

$$\bar{\Omega}^\alpha=\sum_{k=1}^{K_\alpha}\frac{\Omega_k^\alpha}{K_\alpha},\quad \alpha=1,2,\cdots,6 \tag{5}$$

其中, Ω_k^α 表示情感类别为 α 的第 k 个语调在空间 P 内的投影向量, K_α 为具有情感类别 α 这一类内的所有语调样本的总数.

于是, $\{\bar{\Omega}^\alpha,\ \alpha=1,2,\cdots,6\}$ 就构成了本文给出的情感语调基于特征语调 U 的 6 维表示. 在合成中生成情感语调曲线时, 按式 (3) 把 $\bar{\Omega}^\alpha$ 转换回原始语调空间 O 中

$$T_\alpha=U\bar{\Omega}^\alpha+m,\quad \alpha=1,2,\cdots,6 \tag{6}$$

其中, $\{T_\alpha,\ \alpha=1,2,\cdots,6\}$ 即为各类情感类别对应的情感语调在实际语调空间内的表达.

5　实验与讨论

5.1　特征语调分析

为了说明特征语调的特点以及据此重构原始语调样本的有效性, 在本文设计的语音库的基础上进行了相关实验. 实验中, 语调长度标准化样点数 N 取为 100, 并选择前 6 个 (即取 $L=6$) 特征向量构造特征语调 U, 其分析结果如图 2 所示. 从图中可以看出, 在重构语调时, 前面的分量占据主导地位, 它们确定了被拟合对象的基本调型, 而后面的分量在

这方面则贡献相对较小, 后述的实验结果也说明这一点.

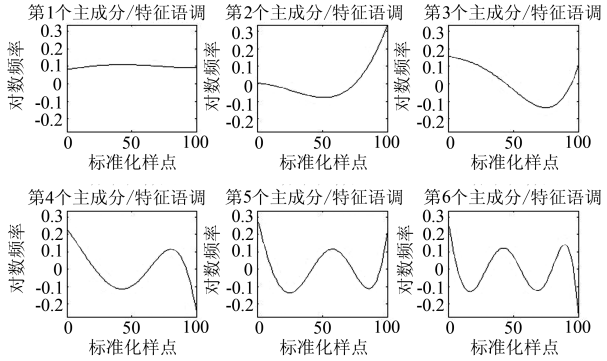

图 2 特征语调

Fig. 2 Eigenintonations of the affective speech

在前述特征语调实验的基础上, 进行了使用这些特征语调重构所有原始语调样本的实验, 实验结果如表 1 所示.

表 1 使用前 L 个特征语调的重构率 r 对比

Table 1 The restoring rate r with L components selected

L	3	4	5	6
r	81.61%	95.71%	99.46%	99.89%

表中, L 为重构中所使用的特征语调的个数, r 为重构率, 由式 (4) 得出. 从表 1 可以看出, 选择低于 5 个特征语调时的重构效果并不十分理想, 采用 6 个特征语调时其重构效果已非常好. 一个拟合的结果示例如图 3 所示. 可以看出, 本文语音库所有语调都可以使用 6 个特征语调来完全拟合, 证明文中提出的特征语调是非常有效的.

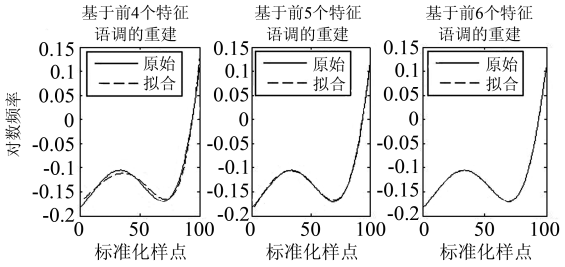

图 3 语调重构拟合示例

Fig. 3 Illustration for reconstructing with eigenintonations

5.2 情感语调分析

本文设计的语音库中的所有语调的情感模式都是已知的, 每个情感模式含 40 句, 共 6 个情感模式. 按照 4.2 节的方法计算, 得到不同情感类别语调在子空间 P 中的投影 Ω^{α}, 其前三个分量分布如图 4 所示.

根据图 4 分析前三个分量的分布可知, “惊奇”、“欢快”、“愤怒” 这三类情感远离其它几类情感模式分布, 其中 “惊奇” 离 “自然” 最远, 而 “悲伤” 和 “恐惧” 则分布得不是很开. 由特征语调的分析可知, 后几个分量所包含的能量很小, 所以其分布差异不如前三个分量突出.

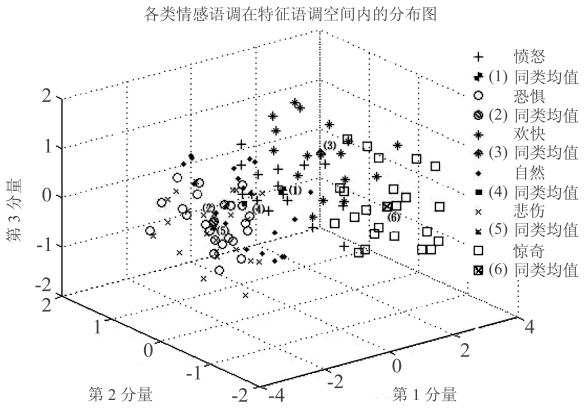

图 4 情感语调投影 Ω^{α} 前三个分量的分布

Fig. 4 Distribution of first 3 weights of affective intonations in eigen sub-space

由情感语调的计算方法, 按式 (5) 和式 (6), 计算得出不同模式情感对应的语调 $\{T_{\alpha},\ \alpha = 1, 2, \cdots, 6\}$, 实际表示如图 5 所示. 可以看出, “愤怒”、“欢快”、“惊奇” 的调值都比较高, 其中 “惊奇” 的调域最大, 而 “恐惧”、“悲伤” 的差别则不大. 这些定性的结果与已有的研究结果[9] 吻合, 说明本文情感语调结果具有代表性.

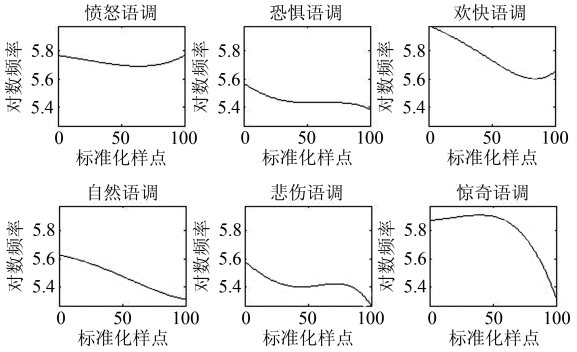

图 5 情感语调 T_{α}

Fig. 5 Affective rule-intonations T_{α}

5.3 合成结果

语音合成实验中, 采用线性预测方法分析、合成语音. 按照情感语调修改自然语音的语调, 再重新合成. 比如, 一个把 “自然” 语调语音修改为 “惊奇” 语调的变化过程如图 6 所示. 图中所示例句内容为 “团结产生力量”, 依次为原始波形图和频率曲线图. 其中, 中间的频率曲线图包括原始基频曲线即句调、原始语调以及修改以后的句调和语调, 还有来自同样文字内容的原始情感语音的语调作为对比. 由图可见, 修改后的合成语调与相应原始情感语音的语

调基本吻合.

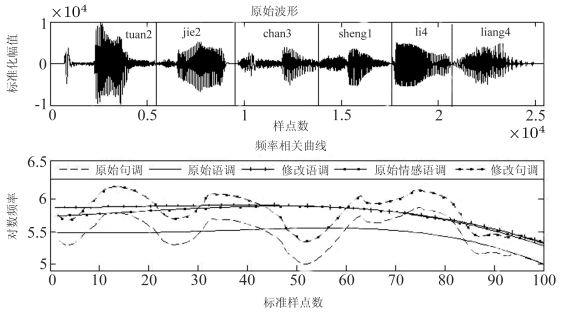

图 6　修改语调合成语音

Fig. 6　Illustration for modifying intonation with surprise rule-intonation

在听觉实验中, 要求测试者分辨出语音听起来最像 "欢快"、"愤怒"、"惊奇"、"恐惧"、"悲伤" 以及 "自然" 状态中的哪一种情感状态. 实验结果显示, 尽管清楚区分开 "愤怒" 和 "欢快", "恐惧" 和 "悲伤" 有点困难, 但是 "欢快、"惊奇"、"恐惧" 的情感状态区分得相对很清晰. 可以看出, 使用特征语调分析出的情感语调具有相当的情感表达能力, 这也说明基于特征语调的分析方法是行之有效的.

6　结束语

语音语调中蕴含着说话人的语气、情感和状态. 本文提出了一种数据驱动的语调建模方法, 导出了特征语调的概念, 并将该方法用于分析普通话中情感模式对应的情感语调. 为了避免干扰因素对情感语调的影响, 设计了相应的情感语音库. 基于本文设计的语音库, 进行了相关实验. 结果表明, 特征语调在一定误差范围内可以拟合出所有语调, 该结果与理论分析完全一致. 实验结果还表明, 使用特征语调分析出的情感语调具有相当的情感表达能力, 说明本文提出的情感语调分析方法是行之有效的. 另外, 本文方法同基于直接平均和直接聚类的方法相比, 具有能够直接控制情感语调变化的优点.

本文提出的方法还需要在其它语法结构的韵律短语上进行更多分析并给出情感语调, 以实现普通话情感语调的完备性, 并需要运用模型的低维表示归纳情感语调的变化控制规律, 这些将是下一步的工作重点.

References

1 Tao J H, Kang Y G. Features importance analysis for emotional speech classification. *Affective Computing and Intelligent Interaction (Lecture Notes in Computer Science)*. Berlin: Springer-Verlag, 2005. 449∼457

2 Yuan J H, Shen L Q, Chen F X. The acoustic realization of anger, fear, joy and sadness in Chinese. In: Proceedings of IEEE International Conferences on Spoken Language. IEEE, 2002. 2025∼2028

3 Zhao Yuan-Ren. *Problems of Language*. Beijing: Commercial Press of China, 1980
(赵元任. 语言问题. 北京: 商务印书馆, 1980)

4 Wu Zong-Ji. Zhao Yuanren′s contribution to the research on mandarin intonation. *Journal of Tsinghua University*, 1996, **11**(3): 58∼63
(吴宗济. 赵元任先生在汉语声调研究上的贡献. 清华大学学报, 1996, **11**(3): 58∼63)

5 Abe M, Sato H. Two-stage *F*0 control model using syllable based *F*0 units. In: Processings of IEEE International Conference on Acoustics, Speech, and Signal. IEEE, 1992. **2**: 53∼56

6 Bellegarda J R, Silverman K E A, Lenzo K, Anderson V. Statistical prosodic modeling: from corpus design to parameter estimation. *IEEE Transactions on Speech and Audio Processing*, 2001, **9**(1): 52∼66

7 Tian J L, Nurminen J. On analysis of eigenpitch in Mandarin Chinese. In: Proceedings of 2004 International Symposium on Chinese Spoken Language Processing. IEEE, 2004. 89∼92

8 Su Z L, Wang Z F. An approach to affective-tone modeling for Mandarin. *Affective Computing and Intelligent Interaction (Lecture Notes in Computer Science)*. Berlin: Springer-Verlag, 2005. 390∼396

9 Zhao Li. *Speech Signal Processing*. Beijing: Mechanics Industry Press, 2003. 259∼268
(赵力. 语音信号处理. 北京: 机械工业出版社, 2003. 259∼268)

10 Zhao Li, Wang Zhi-Ping, Lu Wei, Zou Cai-Rong, Wu Zhen-Yang. Speech emotional recognition using global and time sequence structure feature. *Acta Automatica Sinica*, 2004, **30**(3): 423∼429
(赵力, 王治平, 卢韦, 邹采荣, 吴镇扬. 全局和时序结构特征并用的语音信号情感特征识别方法. 自动化学报, 2004, **30**(3): 423∼429)

11 Cai Lian-Hong, Huang De-Zhi, Cai Rui. *The Theory and Application of Modern Speech Technology*. Beijing: Tsinghua University Press, 2003. 206∼222
(蔡莲红, 黄德智, 蔡锐. 现代语音技术基础与应用. 北京: 清华大学出版社, 2003. 206∼222)

12 Fukunaga K. *Introduction to Statistical Pattern Recognition*. Dordrecht: Academic Press, 2000

苏庄銮　中国科学技术大学自动化系博士研究生. 主要研究方向为信号处理和情感语音合成.
(**SU Zhuang-Luan**　Ph. D. candidate at Department of Automation, University of Science and Technology of China. His research interest covers signal processing and affective speech synthesis.)

汪增福　中国科学技术大学教授. 主要研究方向为视听觉信息处理, 智能机器人和模式识别. 本文通信作者.
E-mail: zfwang@ustc.edu.cn
(**WANG Zeng-Fu**　Professor at Department of Automation, University of Science and Technology of China. His research interest covers audio-video information processing, intelligence robots, and pattern recognition. Corresponding author of this paper.)