

希尔伯特边际谱在语音情感识别中的应用

谢 珊, 曾以成, 蒋阳波

(湘潭大学光电工程系, 湖南湘潭 411105)

摘要: 利用希尔伯特-黄变换(Hilbert-Huang Transform, HHT)对情感语音进行处理, 得到其边际谱, 然后对比分析四种情感即高兴、生气、厌恶、无情感语音信号边际谱的特征, 提出四个特征量: 子带能量(SE)、子带能量的一阶差分(DSE)、子带能量倒谱系数(SECC)、子带能量倒谱系数的一阶差分(DSECC)用于情感识别。用它们作说话人无关, 文本无关的语音情感识别, 得到最高 90%的识别率, 比基于傅立叶变换的梅尔频率倒谱系数(MFCC)高 22 个百分点。实验结果表明, 基于 HHT 边际谱的特征能够较好地反映语音信号中的情感信息。

关键词: 情感识别; 边际谱; HHT

中图分类号: TN912

文献标识码: A

文章编号: 1000-3630(2009)-02-0148-05

DOI 编码: 10.3969/j.issn1000-3630.2009.02.012

Application of Hilbert marginal spectrum in speech emotion recognition

XIE Shan, ZENG Yi-cheng, JIANG Yang-bo

(Photoelectronics Department, Xiangtan University, Xiangtan, 411105, Hunan, China)

Abstract: Marginal spectrum of the emotional speech is obtained through Hilbert-Huang Transform. Speech signals of four different emotions, namely happy, angry, boring and natural, are analyzed contrastively focusing on the characteristics of the marginal spectrum. Then four features: SE, DSE, SECC and DSECC are extracted for emotion recognition. Finally speaker-independent and text-independent emotion recognitions are simulated by using these features respectively, which gains the best recognition rate of 90%, which is 22 percentage higher than Fourier Transform based feature MFCC. Thus, conclusion is drawn that HHT marginal spectrum can well reflect the emotional information in speech.

Key words: emotion recognition; marginal spectrum; HHT

1 引 言

语音情感识别具有广阔的应用前景。在人机交流方面^[1], 能赋予计算机识别情感和表达情感的能力, 使交流更加人性化; 在自动电话服务系统中^[2], 可使合成的语音更加动听, 不再单调; 在娱乐方面, 互动电影^[3]通过情感识别让观众参与电影的创作, 是未来电影发展的方向; 智能玩具和电子宠物^[4]通过情感识别可与人进行更加人性化的情感交流。此外, 语音情感识别还可以应用于辅助教学、精神诊断和测谎等方面^[1]。

语音情感识别包括三个步骤: 样本采集, 特征提取和特征识别^[5], 其中特征提取最重要, 它直接决定了识别的效果。到目前为止, 用于情感识别的主要有基音频率、共振峰、频谱、能量和时长等几

个方面的特征^[6]。陶建华分析了不同情感语音在基音、时长、能量等方面的差异, 并对基音均值等 18 个特征根据其各自的识别效果进行了排序^[7]。Nicholson 等人提取基音频率、功率谱和线性预测系数(LPC), 对 9 种情感进行识别, 但识别率只有 51%^[4]。Vladmir 将短时特征和长时特征结合起来, 分别对不同的语种进行情感识别, 最高识别率为 67%^[8]。

希尔伯特-黄变换(HHT)是一种适用于非线性、非平稳信号的处理方法^[9]。它通过经验模式分解(Empirical Mode Decomposition, EMD)将信号的内部振荡模式逐级分离出来^[10], 并通过希尔伯特变换得到其瞬时频率和幅度。本文运用 HHT 对语音信号进行处理, 比较分析不同情感语音的边际谱, 并在此基础上提取一组特征用于情感识别, 实验结果证明基于 HHT 边际谱的特征能够较好地反映语音的情感信息。

2 基于 HHT 的边际谱

根据 HHT 的原理, 首先对语音信号进行 EMD,

收稿日期: 2008-03-01; 修回日期: 2008-04-13

作者简介: 谢珊(1983-), 女, 湖南岳阳人, 硕士研究生, 研究方向为语音情感识别。

通讯作者: 曾以成, E-mail: yichengz@xtu.edu.cn

得到该语音的 n 个内在模式函数(IMFs), 然后对每个 IMF 作希尔伯特变换, 分别得其希尔伯特幅度谱 $H_i(\omega, t)$, 它是关于时间和瞬时频率的函数。将 $H_i(\omega, t)$ 在时域积分, 可得到能量在频域的分布。即用有限个点将整个频域划分为若干小的区间如 $[\omega_1, \omega_2, \dots, \omega_N, \omega_{N+1}]$, 对任意 $0 < k < N+1$, 找出在整个时域范围内瞬时频率落在第 k 个小区间的点, 将其幅度相加, 即:

$$h_i(\Omega_k) = \sum_{\omega_k < \omega \leq \omega_{k+1}} H_i(\omega, t) \quad (1)$$

其中, $\Omega_k = (\omega_k + \omega_{k+1})$ 为第 k 个区间的频率代表值。由此可得 $h_i(\Omega)$, 即 IMF 边际谱。

对所有 IMF(Intrinsic Mode Function)边际谱求和, 可得语音信号的总边际谱 $h(\Omega)$:

$$h(\Omega) = \sum_{i=1}^n h_i(\Omega) \quad (2)$$

3 语音信号情感分析

实验使用的情感语音样本在专门的语音实验室录制, 由 5 男 5 女 10 个专业演员录音, 包括 10 个文本不同的长句。各说话人分别用高兴、生气、悲伤、厌恶、恐惧和平静等六种情感状态说一遍这 10 个句子。根据情感空间理论, 高兴、生气、厌恶和平静分别在激活-评估二维坐标系的第一、二、三和四象限^[11], 因此本文选取此四种典型情感(以下分别用字母 H, A, B, N 表示)进行研究。

选取一组说话人和文本都相同的此四种情感语音, 分别用 $x_H(t)$ 、 $x_A(t)$ 、 $x_B(t)$ 、 $x_N(t)$ 表示, 先进进行如图 1 所示的处理, 然后比较分析不同情感语音的 IMF 边际谱和总边际谱。

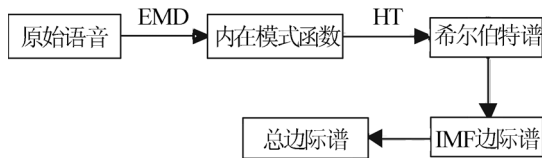


图1 求语音信号总边际谱框图

Fig.1 Extraction diagram of speech signal general marginal spectrum

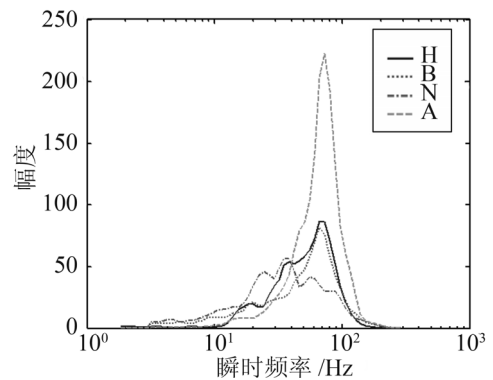
3.1 IMF 边际谱比较

EMD 实质上是自适应滤波的过程, 分解得到的一系列 IMF 是按照频率从大到小排列的, 且每种情感语音经 EMD 分解后得到的 IMF 数目大致为 8 到 10 个, 基本相同。因此, 可以很方便地比较同一情感语音的不同 IMF 边际谱, 也可以比较不同情感语音的同一级别 IMF 边际谱。

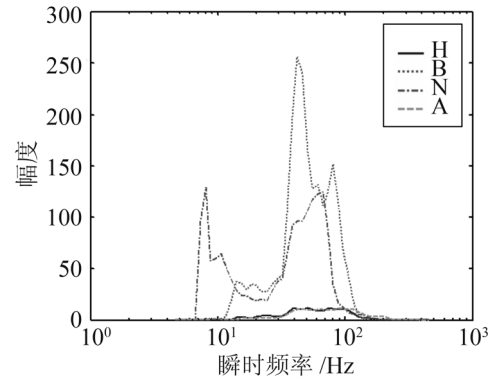
从语音库中选取的高兴、生气、厌恶和无情感

状态语音样本, 进行图 1 所示的处理后, 可得它们各自的所有 IMF 边际谱。为表述简便, 选择适当级数的三个 IMF 分别代表语音信号的高频、中频和低频部分, 如选取 IMF₂、IMF₅ 和 IMF₈。上述四种情感语音的 IMF₂、IMF₅ 和 IMF₈ 边际谱曲线分别如图 2(a)、2(b)、2(c)所示, 其中实线代表情感“H”, 虚线代表情感“A”, 点线代表情感“B”, 点划线代表情感“N”。

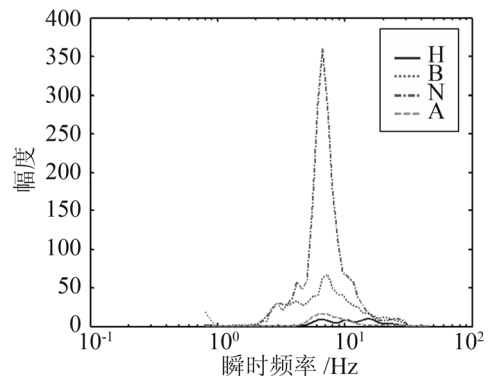
从图 2 可以看出, 在不同的频率范围内, 四种情感语音能量的分布规律各有不同。在图 2(a)中, 由于情感 A 具有非常丰富的高频能量, 其边际谱曲线出现明显的尖峰, 该峰值是情感 N 峰值的 3.9 倍,



(a) Marginal spectrum of IMF2



(b) Marginal spectrum of IMF5



(c) Marginal spectrum of IMF8

图2 情感语音 IMF 边际谱比较: (a) IMF2 (b) IMF5 (c) IMF8
Fig.2 Comparison of IMF marginal spectrum of emotional speech: (a) IMF2, (b) IMF5, (c) IMF8

B 的 2.7 倍。且在此频率范围内, A 和 H 的边际谱有非常明显的差别, 前者能量更大, 分布更集中。而 B 和 N 差别较小。从图 2(b)可以明显地看出, 在中频段, A 和 H 的能量非常小, 并且它们的边际谱曲线几乎重合, 说明此两种情感的中频能量小且其分布趋向一致。而情感 B 和 N 这部分的能量非常大, 且由于 EMD 自适应滤波的结果, 边际谱表现出明显的带通特性, 从通带的位置可以看出, B 的平均频率比 N 高。在低频部分, 如图 2(c)所示, A 和 H 仍表现了能量小且分布相似的特点。而情感 B 和 N 的能量很大, 尤其是 N, 有非常明显的尖峰。

3.2 总边际谱比较

图 3 所示为四种情感语音的总边际谱。可以大致看出这四条曲线在各个频段的差异。即 A 和 H 高频能量大, 两者高频部分边际谱差异也大, 中、低频部分能量小, 两者差异也小。而 B 和 N 的情况恰好相反。总体上与前面的结论一致。

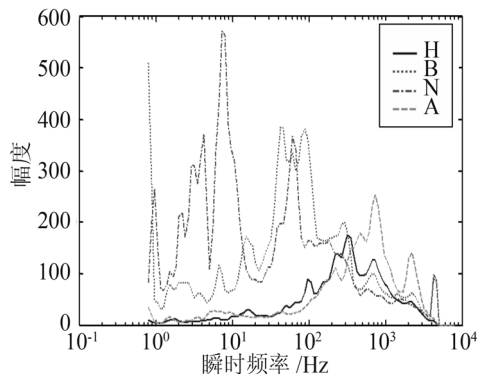


图 3 情感语音总边际谱比较

Fig.3 Comparison of general marginal spectrum of emotional speech

以上分析表明, 边际谱能够较好地体现语音中的情感信息, 因此可以作为语音情感识别的一个重要特征。但也可以看出, 并不能直接将边际谱作为特征进行情感识别。因此下一节中将在情感语音的总边际谱基础上提取出几个特征量用于情感识别。

4 情感识别

提出四个特征量, 即边际谱子带能量(SE)及其一阶差分(DSE), 边际谱子带能量倒谱系数(SECC)及其一阶差分(DSECC)用于情感识别。特征提取过程如下:

- (1) 对信号进行 HHT, 得到 $H(\omega, t)$;
- (2) 求出信号的总边际谱 $h(\omega)$, 其中 ω 用对数坐标;
- (3) 将 ω 划分为 Y 个子频率带, 各子带的带宽

BW 在 ω 对数域呈等差递减, 即第 i 个子带的带宽 BW_i 满足 $BW_i = BW_{i-1} - 1$;

(4) 分别计算每个子频率带的能量 E_i , 得到特征矢量 $SE = (E_1, E_2, \dots, E_Y)$, 即子带能量;

(5) 对 E_i ($1 \leq i \leq Y$) 求对数, 并作离散余弦变换, 得到特征矢量 $SECC$, 即子带能量倒谱系数;

(6) 对 SE 求一阶差分, 得到特征矢量 DSE ;

(7) 对 $SECC$ 求一阶差分, 得到特征矢量 $DSECC$;

特征提取框图如图 4 所示。

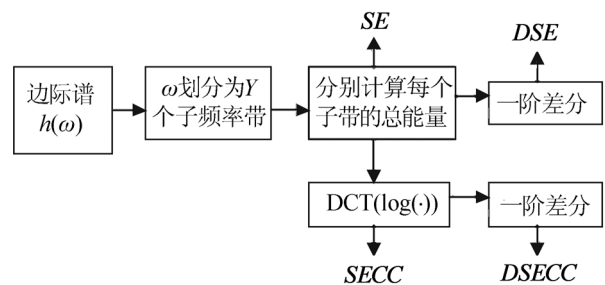


图 4 特征提取过程

Fig.4 Procedure of feature extraction

用于识别的语音共 300 句, 其中 255 句用于训练, 45 句用于识别。语音数据各种情感的分布如表 1 所示:

表 1 语音数据各种情感的分布
Table 1 Distribution of the utterances among emotions

	H	B	N	A
训练	50	55	60	90
识别	7	8	13	17
总数	57	63	73	107

识别采用矢量量化(VQ)方法, 每种情感训练一个码本^[11]。训练阶段, 先对每个用于训练的语句提取一个 Y 维的特征矢量。然后用 LBG 算法训练码本, 码本大小为 $M=2^k$ ($1 \leq k \leq 5$)。识别阶段, 对每个识别语句提取同样的特征矢量, 分别计算该特征矢量与 4 个码本的距离, 使距离最小的码本所对应的情感即为识别情感^[12]。

特征维数 Y 和码本长度 M 的取值均会影响识别结果。首先考察特征维数对识别率的影响。取码本长度 M 为 2, 使特征维数 Y 从 9 逐渐增至 24, 分别得出每一维时各特征的识别率。作为比较, 还用 MFCC 作了识别, 取码本长度为 256, 得出其在不同特征维数下的识别率。(在这里, 情感识别率取四种情感识别率的平均值)。五种特征各自的识别率随特征维数变化的曲线如图 5 所示。其中星号代表特征“SE”, 加号代表特征“DSE”, 圆圈代表特征“SECC”, 菱形代表特征“DSECC”, 点代表特征 MFCC。从图 5 可以看出, SE 的识别效果最好, 在不同特征维数下, 其识别率均高于 MFCC, 且最高

达到 73.3%。SECC 在特征维数较低时, 识别效果和 MFCC 相当, 而当维数大于 20 时, 其识别率明显比 MFCC 高。DSECC 在不同特征维数下的平均识别率比 MFCC 高 0.62 个百分点。而 DSE 识别效果不够理想, 平均识别率仅为 62.35%。

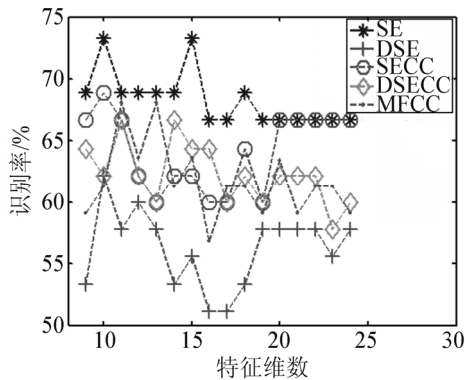


图 5 五种特征的识别结果
Fig.5 Recognition results of the five features

然后考察码本大小对识别率的影响。取特征维数 Y 为 12, 分别得出 5 个特征在不同码本下的识别率, 结果如表 2 所示。从表 2 可看出, Y 等于 12 时, SE 和 SECC 的识别率均比较高, 且其随码本大小 M 的变化波动较小。另外, 基于边际谱的 4 个特征识别率随码本大小的变化没有明显的变化规律, 而基于傅立叶变换的 MFCC 识别率随码本的增加而增大。这是因为前者用于训练的特征矢量很少, 为 50~90 个。后者的特征矢量通过分帧提取, 有 20000~30000 个, 是码本大小的几百倍甚至几千倍, 增加码本的长度相当于增加了码本包含的信息, 从而使识别更准确^[13]。

表 2 各种特征在不同码本下的识别率
Table 2 Recognition results of the five features in different code book size

特征量	识别率/%				
	码本大小=2	4	8	16	32
SE	68.9	64.4	62.2	62.2	64.0
DSE	60	62.2	73.3	51.1	66.7
SECC	62.2	60.0	68.9	64.4	66.7
DSECC	62.2	62.2	48.9	53.3	53.3

特征量	识别率/%				
	码本大小=216	32	64	128	256
MFCC	38.6	52.3	56.8	59.1	63.6

另外, 表 3 列出了各个特征能得到的最高识别率及此时的特征维数和码本大小。从表 3 可以看出, 基于边际谱的 4 种特征的最高识别率相对于 MFCC 的提高了 5%~10%。

值得注意的是, 不同类型的情感, 其识别率是不同的, 而同一种情感因说话人性别不同, 识别率

表 3 各种特征的最高识别率及所用特征维和码本大小
Table 3 Best result of each feature and its feature dimension and codebook size

特征量	特征维数	码本	识别率/%
SE	15	2	73
DSE	23	8	73
SECC	18	16	78
DSECC	24	16	76
MFCC	13	256	68

也存在差异。因此在下面的实验中, 将所有语音按说话人性别分为两部分, 分别对其进行训练和识别, 并给出这两种情况下各种情感的识别率。为简化问题, 特征维数和码本尺寸取固定值, Y 为 12, M 为 8。表 4 和表 5 分别给出了采用不同特征时男性和女性各种情感的识别率。

表 4 采用不同特征时男性各种情感的识别率
Table 4 Recognition rate of different emotions for male adopting different features

	SE	DSE	SECC	DSECC	MFCC
H	63%	63%	75%	50%	50%
B	50%	58%	75%	66%	62%
N	68%	50%	60%	53%	50%
A	90%	85%	90%	82%	82%

表 5 采用不同特征时女性各种情感的识别率
Table 5 Recognition rate of different emotions for female adopting different features

	SE	DSE	SECC	DSECC	MFCC
H	70%	75%	90%	60%	68%
B	86%	70%	75%	66%	75%
N	62%	53%	68%	62%	60%
A	82%	50%	75%	75%	70%

从表 4、5 可看出, 此实验的识别效果比上文的好, 这是因为男性和女性表达情感的方式存在差异, 因此将其分开训练和识别能够提高识别率。另外, 女性的平均语音情感识别率高于男性, 特别是情感 H, 最高识别率达 90%。而男性情感 A 的识别率相对较高, 在 80% 以上。比较不同特征下的识别率, 可以看出, 本文提出的特征量整体上优于传统特征 MFCC, 男性情感 A 和女性情感 H 的识别率可达 90%, 明显高于相同条件下 MFCC 的识别率。这是因为 HHT 先对信号进行平稳化处理, 然后计算其频谱, 能够准确提取信号的频域信息。

可见, 基于边际谱的特征具有更好的情感识别效果。这是由于与傅立叶变换相比, 希尔伯特变换更适合处理语音信号, 它不必对信号分帧, 既保留了信号的动态信息, 又大大减小了码本的容量。

5 结 论

本文以 HHT 为原理对语音信号的四种不同情

感(高兴、生气、厌恶和自然)进行了分析。通过比较不同情感语音的边际谱得出:情感 A 和 H 高频能量相对较大,而 B 和 N 的中频和低频能量较大。并且前两种情感的区别主要在高频部分,后两种情感的区别主要在中、低频部分。还提取了 SE 等特征对这四种情感进行了文本无关和说话人无关的识别,结果表明,SE 和 SECC 具有较好的识别效果,当码本大小为 2 时,它们在不同特征维数时的平均识别率分别为 68.4 和 64.2。SECC 在码本大小为 16、特征维数为 18 时,可得到最高识别率 78%,比基于傅立叶变换的特征 MFCC 高 10 个百分点。而对不同性别的语音,分开训练和识别具有更好的识别效果,能使识别率达到 90%,比相同条件下 MFCC 的识别率高 8%~22%。

此外,基于边际谱的特征用于情感识别时,能大大减小码本的长度,从而提高识别效率,具有一定的实际意义。

参 考 文 献

- [1] Yi Linlin, Wei Gang. Speech emotion recognition based on HMM and SVM[A]. Proceedings of the Fourth International Conference on Machine Learning and Cybernetics[C]. Guangzhou. 2005: 18-21.
- [2] Chul Min Lee, Shrikanth S.Narayanan. Toward detecting emotions in spoken dialogs[J]. IEEE Transactions on Speech and Audio Processing, 2005, 13(1): 230-235.
- [3] Laurence Vidrascu, Laurence Devillers. Detection of real-life emotions in call centers[A]. INTERSPEECH 2005[C]. 2005, 10: 1841-1844.
- [4] Nicholson J, Takahashi K, Nakatsu R. Emotion recognition in speech using neural networks[J]. Neural Computer & Applic, 2000, 9: 290-296.
- [5] 詹永照, 曹鹏. 语音情感特征提取和识别的研究与实现[J]. 江苏大学学报(自然科学版), 2005, 26(1): 20-25.
ZHAN Yongzhao, CAO peng. The analysis and research of the emotional speech feature extraction and emotion recognition[J]. Transaction of Jiangsu University, 2005, 26(1): 20-25.
- [6] 陈建厦, 李翠华. 语音情感识别的研究进展[J]. 计算机工程, 2005, 31(13): 50-55.
CHEN Jianxia, LI Cuihua. Evolution of emotional speech research[J]. Computer Engineering, 2005, 31(13): 50-55.
- [7] TAO Jianhua, KANG Yongguo. Features importance analysis for emotional speech classification[J]. ACII 2005, 37: 449-457.
- [8] Vladimir, Hozjan Zdravko Kacic. Context-independent multilingual emotion recognition from speech signals[J]. International Journal of Speech Technology, 2003, 6: 311-320.
- [9] 樊长博, 张来斌, 王朝晖, 等. 基于 EMD 与功率谱分析的滚动轴承故障诊断方法研究[J]. 机械强度, 2006, 28(4): 628-631.
FAN Changbo, ZHANG Laibin, WANG Zhaohui, et al. Research of ball bearing fault detection based on EMD and power spectrum analysis[J]. Machine Intensity, 2006, 28(4): 628-631.
- [10] 李凌, 曾以成, 雷雄国. EMD 在说话人辨认中的应用[J]. 湘潭大学自然科学学报, 2006, 28(3): 108-111.
LI Ling, ZENG Yicheng, LEI Xiongguo. Application of EMD in speaker recognition[J]. Natural Science Journal of XiangTan University, 2006, 28(3): 108-111.
- [11] 张石清, 赵知劲. 支持向量机应用于语音情感识别的研究[J]. 声学技术, 2008, 27(1): 87-95.
ZHANG Shiqing, ZHAO Zhijing. A study of support vector machine for speech emotion recognition[J]. Technical Acoustics, 2008, 27(1): 87-95.
- [12] 邱作春, 曾庆宁. 独立分量分析在说话人识别技术中的应用[J]. 声学技术, 2008, 27(6): 863-866.
QIU Zuochun, ZENG Qingling. A application of indepdent componet analysis for speaker recognition[J]. Technical Acoustics, 2008, 27(6): 863-866.
- [13] Kyung Hak Hyun, Eun Ho Kim, Yoon Keun Kwak. Improvement of emotion recognition by bayesian classifier using non-zero-pitch concept[J]. IEEE International Workshop on Robots and Human Interactive Communication, 2005, 13(15): 312-316.