文章编号:1007-5321(2018)02-0056-06 **DOI**:10.13190/j.jbupt.2017-157

基于谱信息熵和互补模型的声效模式检测

晁 浩, 鲁保云, 刘永利, 刘志中, 宋 成
(河南理工大学 计算机科学与技术学院, 河南 焦作 454000)

摘要: 提出了一种基于模型融合的声效检测方法. 首先提取对所有声效模式都具有良好辨识能力的谱信息熵特征,并进行声效辨识度分析;然后引入互补模型进行声效识别,从模型层面实现了整体谱特征、梅尔频率倒谱系数和谱信息熵的融合. 对孤立词测试集进行了声效检测实验,识别精度为 81.6%,实验结果表明,谱信息熵在 3 类特征中具有最好的分类能力,而互补模型能够有效集成 3 种特征蕴含的显著性信息.

关 键 词: 声效; 谱信息熵; 支持向量机; 高斯混合模型; 多层感知器
中图分类号: TN391.42 **文献标志码**: A

Vocal Effort Detection Based on Spectral Information Entropy and Complementary Models

CHAO Hao, LU Bao-yun, LIU Yong-li, LIU Zhi-zhong, SONG Cheng
(School of Computer Science and Technology, Henan Polytechnic University, Henan Jiaozuo 454000, China)

Abstract: A new vocal effort detection method based on model fusion was presented. By analyzing the ability to discriminate the vocal effort modes, the spectral information entropy feature which contains more salient information regarding the vocal effort level was proposed. Then, the complementary models were presented to achieve the fusion of the spectrum features, Mel-frequency cepstral coefficients and spectral information entropy feature. Experiments are conducted on isolated words test set, and the proposed method achieves 81.6% average recognition accuracy. The results show the spectral information entropy has the best performance among the three kinds of features and the complementary models can integrate the three kinds of features effectively.

Key words: vocal effort; spectral information entropy; support vector machine; Gaussian mixture model; multilayer perceptron

声音效果被定义为正常人的一种发音变化的衡量[1]. 准确地识别出语音信号的声效模式对于提高语音处理技术,如语音识别、语音合成及说话人识别的鲁棒性有着重要的作用[2-4]. 5 种声效模式中,耳语相关的研究较多[5-6]. 剩余 4 种声效模式的发音方式没有明显的差异,因此检测难度较大,相关的研究较少. 声效检测常用的 2 种特征为整体谱特征[7]和梅尔频率倒谱系数(MFCC, mel-frequency cepstral coefficients)[8-9]. 与整体谱特征相比,MFCC 对于耳语外的 4 种声效模式有着更强的辨识能力. 然而该特征毕竟是为了语音识别而提出的,更多地蕴含了与语音内容相关的显著性信息,用于声效检测时潜

收稿日期: 2017-08-10
基金项目: 国家自然科学基金项目(61502150, 61403128); 河南省高等学校青年骨干教师资助项目(2015GGJS-068); 河南省科技攻关项目(172102210279)
作者简介: 晁 浩(1981—), 男, 讲师, E-mail: chaohao1981@163.com.

力有限.

为了提高所有声效模式的检测精度,首先提出了一种基于元音帧的谱信息熵特征(SIE, spectral information entropy). 与 MFCC 特征和整体谱特征相比,谱信息熵表现出了更强的辨识能力. 考虑到整体谱特征、MFCC 和谱信息熵特征分别是从不同角度描述语音信号,蕴含的声效量级相关的显著性信息不会完全重叠,上述 3 种特征在声效检测时应该具有互补的作用. 因此,引入了基于互补模型的特征融合方法.

1　谱信息熵

声效检测的关键是获取声效量级相关的显著性信息,进而提取声效特征. 与频谱倾斜等整体特性相比,基于帧的特征更能捕获由于声效改变造成的语音信号谱的细微变化. 为此,提出了基于语音帧的谱信息熵特征.

1.1　谱信息熵特征提取

对语音信号进行分帧、加窗、预加重处理后,再进行快速傅里叶变换得到时频域上的能量分布:每 1 帧的能量分布可以视为 1 个标准正交基上的系数向量. 首先对各频带能量进行平滑处理,然后对该向量内所有频率成分进行归一化,在此基础上估计概率密度函数,并依据概率密度函数计算谱信息熵参数.

计算某一个频带的谱信息熵:假设 $X(k)$ 是当前语音帧 $x(n)$ 在该频带的能量谱,k 取值从 $k_1 \sim k_M$,那么第 k 个频率成分的频率内容占整个频带的比例为

$$p(k) = \frac{|X(k)|^2}{\sum_{j=k_1}^{k_M} |X(j)|^2}, k = k_1, \cdots, k_M \qquad (1)$$

可以得出 $\sum_{k=k_1}^{k_M} p(k) = 1$,那么 $p(k)$ 就可以认为是描述该频带内能量分布的概率分布函数. 而该频带的谱信息熵为

$$H = -\sum_{k=k_1}^{k_M} p(k) \log p(k) \qquad (2)$$

计算谱信息熵时语音信号的频带具体划分如表 1 所示.

由于元音信号的能量分布基本在 0 ~ 5.0 kHz,且能量主要集中在中低频带,所以在确定各个频带的频域范围时有如下规则:低频带的带宽较窄,粒度较小,能够获取中低频域的频谱能量的细微变化;高频带的带宽较宽,主要是刻画中高频的能量分布. 这样,谱信息熵重点反映的是语音信号在频域的能量分布情况,而不是整个频域的能量总量. 各个频带具体的频域范围主要通过多次实验来调整. 对于每 1 帧,根据式(2)分别计算其在表 1 所示的 12 个频带的谱信息熵,得到维数为 12 的信息熵特征.

表 1　12 个频带及其频域范围

频带	f/kHz	频带	f/kHz
1	0.0 ~ 0.3	7	1.7 ~ 2.1
2	0.2 ~ 0.5	8	2.2 ~ 2.6
3	0.4 ~ 0.7	9	2.7 ~ 3.2
4	0.6 ~ 1.0	10	3.3 ~ 3.8
5	0.8 ~ 1.2	11	3.9 ~ 4.4
6	1.2 ~ 1.6	12	4.5 ~ 5.0

1.2　声效辨识度分析

语音信号中声效模式敏感信息分布是不均匀的,在之前的研究工作中,已经分析了元音与辅音在不同声效模式下的语谱变化[9]. 与绝大部分辅音相比,元音在不同声效模式下其语谱变化更为明显,意味着蕴含了更多的声效模式显著性信息. 这里将进一步分析从元音中提取的谱信息熵和 MFCC 在声效检测中哪个具有更好的辨识能力.

为了进行声效辨识度对比,定义了基于欧氏距离的谱距离计算公式:

$$D_c = \sqrt{\sum_{i=1}^{N} (c^{E(1)}(i) - c^{E(2)}(i))^2} \qquad (3)$$

以描述语谱的变化程度. 式(3)计算同一音素分别在 2 种声效模式 $E(1)$、$E(2)$ 下发音时,产生的 2 个语音信号的谱距离 D_c. $\boldsymbol{c}^{E(\cdot)}$ 为音素在声效模式 $E(\cdot)$ 下对应的语音信号的谱特征序列求均值后形成的向量,$c^{E(j)}(i)$ 为该向量的第 i 个参数.

谱距离描述了音素在不同声效模式下的语谱变化程度. 首先将 MFCC 作为谱特征计算各个音素的谱距离均值;然后将信息熵作为谱特征再计算各个音素的谱距离均值;再统一进行线性函数归一化处理. 图 1 显示了单元音 a e o i u 分别使用 2 种特征时的谱距离均值对比.

从图 1 可以看出,当使用 SIE 时,5 种单元音的平均谱距离均明显要高于使用 MFCC 时得到的平均谱距离. 这就表明谱信息熵特征在检测声效模式时能提供更多的区分性信息.

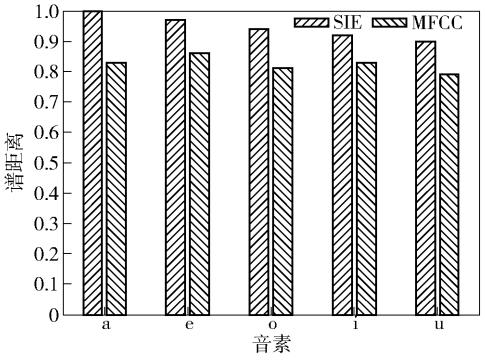

图1 SIE和MFCC的谱距离均值对比

2 基于互补模型的声效检测

互补模型最初被提出来用于汉语韵律间断检测[10]和汉语重音检测[11],相比传统的集成机器学习方法,其在模型层面进行特征融合能够取得更好的效果.

鉴于整体谱特征、MFCC和谱信息熵分别是从不同角度描述语音信号的,笔者考虑结合3种特征来实现声效模式检测. 存在的问题是:1)建模时同时使用3种特征是否比单独使用1种或2种特征的效果要好? 2)如何同时使用3种特征?

对于第1个问题笔者借助于信息熵理论来说明:假设$H(X)$是随机变量X的熵,随机变量X和随机变量Y是不独立的,$H(X|Y)$是条件熵. 利用信息熵的结论,可以得到

$$H(X|Y)\leqslant H(X) \tag{4}$$

式(4)表明,条件可以减少熵,也就是说,有用的信息Y可以减少X的不确定性. 因此,声效建模中在确定使用1种特征的情况下,引入其余的特征也可以减少声效事件的不确定性,如

$$H(E|I,M,F)\leqslant H(E|I,M)\leqslant H(E|I)\leqslant H(E) \tag{5}$$

其中:E为声效模式事件随机变量,F为整体谱特征随机变量,M为MFCC特征随机变量,I为谱信息熵特征随机变量.

对于第2个问题,常用的解决方案是同时使用3种特征来构建1个分类模型p:

$$E^* = \arg\max\ p(E|F,M,I) \tag{6}$$

为了简化计算,式(6)可以转变为

$$E^* = \arg\max\ p(E|F,M,I) = \arg\max\ p(E|F)p(E|M)p(E|I) \tag{7}$$

式(7)成立的条件是F、M和I之间是相互独立的,但是在现实情况下这是不可能实现的,如整体谱特征F中的帧能量分布与谱信息熵特征I明显具有相关性. 因此,式(7)尽管简化了计算,但在一定程度上牺牲了检测精度. 考虑到这一点,这里引入了一种不依靠F、M和I独立假设的建模方法.

通过对第1个问题的分析可以得出以下结论:有用信息的引入可以减少条件熵,提高模型的精度,而将整体谱特征、MFCC和谱信息熵联合建模则是这一认识的重要体现,在这里使用互补模型来进行3种特征的联合建模.

首先,将式(6)转换为

$$E^* = \operatorname{argmax}\ p(E|F,M,I) = \operatorname{argmax}\ (\lambda p(E|F,M,I) + (1-\lambda)p(E|F,M,I)) = \operatorname{argmax}\ (\lambda p_1(E|F,M,I) + (1-\lambda)p_2(E|F,M,I)) \tag{8}$$

式(7)仅仅是式(5)的一个变形,给$\lambda p(E|F,M,I)$一个新符号$\lambda p_1(E|F,M,I)$,同时给$(1-\lambda)p(E|F,M,I)$一个新符号$(1-\lambda)p_2(E|F,M,I)$. 如果用同样的方法建模$p_1(\ )$和$p_2(\ )$,则式(8)可以看作是一个传统的模式分类方法,如式(6);如果不用同样的方法建模$p_1(\ )$和$p_2(\ )$,并假设F、M和I之间相互独立,则式(8)就可以写成式(7);如果不用同样的方法建模$p_1(\ )$和$p_2(\ )$,并放弃F、M和I之间独立性假设,则是基于互补模型的模型融合方法.

从式(8)可以看到:1) 对于每个分类器$p_1(\ )$和$p_2(\ )$,所有的特征,即整体谱特征、MFCC和谱信息熵都用于建模,而$p_1(\ )$和$p_2(\ )$则是1对互补的模型;2) 在利用所有的特征建模之后,2个不同的分类器$p_1(\ )$和$p_2(\ )$是线性结合的;3) 从本质上讲,互补模型方法是一种分类器结合的方法,而互补的含义是一种模型,能够检测声效模式,而另外一种方法也可以检测声效模式,但是它们的分布不同. 虽然检测的结果有重叠,但是不完全重叠,它们之间存在互补,通过这种方式来综合利用3种特征蕴含的声效显著性信息. 该思想来源于系统融合,不同的系统可以进行融合,以提高精度.

由于元音特别是单元音相比辅音蕴含了更多的声效模式显著性信息[9],所以语音信号中的单元音被提取出来用于声效检测. 单元音可以通过手工切分或元音端点检测来获取. 基于互补模型的声效检测具体过程如下.

1) 检测语音信号中的元音,得到元音序列. 对于元音序列的声效模式,确立如下规则:如果语句的

声效模式是 E，那么该语句中所有单元音的声效模式都是 E，反之亦然.

2）提取每 1 个元音的整体谱特征 F_t、MFCC 特征 M_t 和谱信息熵特征 I_t. 其中，F_t 表示元音序列中第 t 个元音的整体谱特征，M_t 表示元音序列中第 t 个元音的 MFCC 特征，I_t 表示元音序列中第 t 个元音的谱信息熵特征. 整体谱特征包括该元音信号的频谱倾斜、声强级、时长以及帧能量分布；MFCC 特征和谱信息熵为基于帧的特征序列.

3）根据 1）中确立的规则，将式(8)进一步展开得到式(9)，根据式(9)来判定语音信号最优的声效模式.

$$\begin{aligned} E^* = & \\ \operatorname{argmax}\ & (\lambda p_1(E|F,M,I) + (1-\lambda)p_2(E|F,M,I)) = \\ \operatorname{argmax}\ & \left(\frac{\lambda}{(1-\lambda)}p_1(E|F,M,I) + p_2(E|F,M,I)\right) = \\ \operatorname{argmax}\ & (\gamma p_1(E|F,M,I) + p_2(E|F,M,I)) = \\ \operatorname{argmax}\ & \left(\gamma \prod_{t=1}^{n} p_1(E|F_t,M_t,I_t) + \right. \\ & \left. \prod_{t=1}^{n} p_2(E|F_t,M_t,I_t)\right) \end{aligned} \tag{9}$$

其中：$p_1(E|F_t,M_t,I_t)$ 为利用模型 $p_1(\,)$ 估计第 t 个元音的声效模式为 E 的概率，$p_2(E|F_t,M_t,I_t)$ 为利用模型 $p_2(\,)$ 估计第 t 个元音的声效模式为 E 的概率，n 为元音序列中元音的总数. 互补模型 $p_1(\,)$ 和 $p_2(\,)$ 通过统计机器学习的方法获得，例如，决策树、神经网络、支持向量机等. 参数 γ 为

$$\gamma = \frac{\lambda}{1-\lambda} \tag{10}$$

2 种模型 $p_1(\,)$ 和 $p_2(\,)$ 检测的结果会部分重叠. 因此在选择互补模型时，2 个模型分布的 K-L 距离越大，则分布的差异性就越大，分类的效果就越好. 1 个极端的例子就是 K-L 距离为 0，说明两个分布相同. 因此，可以选择 K-L 距离大的哪些分布作为互补模型的候选.

此时，式(7)可以展开为

$$\begin{aligned} E^* &= \operatorname{argmax} p(E|F)p(E|M)p(E|I) = \\ \operatorname{argmax} & \prod_{t=1}^{n} p\,(E|F_t)^{\alpha} p\,(E|M_t)^{\beta} p(E|I_t) \end{aligned} \tag{11}$$

其中 α 和 β 为调整权重系数.

3　实验仿真及结果分析

3.1　实验语料及模型

实验语料库包含了由 20 个男性说话人录制的 2.5 万个汉语数字(0 ~ 9). 在训练集中，每种声效模式对应 4 000 个汉语数字；测试集中，每种声效模式对应 1 000 个汉语数字. 所有语料均为安静实验室环境录制，信噪比小于 50 dB，采样率为 16 kHz，16 bit 量化，窗长为 25.6 ms，帧移为 10 ms. 整体特征为声强级、谱倾斜、元音时长及帧能量分布[7]，MFCC 特征包含 12 维梅尔频率倒谱系数.

3.2　结果与分析

首先分析 3 种特征哪种对声效模式的辨识能力更好，为此，轮流使用整体谱特征、MFCC 和 SIE 单独构建分类模型，这意味着式(11)中的 $p(E|F_t)$、$p(E|M_t)$ 和 $p(E|I_t)$ 分别单独使用. 高斯混合模型(GMM，gaussian mixture model)、支持向量机(SVM，support vector machine)和多层感知器(MLP，multi-layer perceptron)3 种模型分别用于声效检测. 其中，MLP 拥有 1 个隐含层，隐含层节点数为 $2N+1$，N 为输入层节点数. SVM 采用 RBF(radial basis function)核，利用 LibSVM 工具训练[12]. GMM 采用对角线协方差矩阵，拥有 128 个分量. 声效识别结果如表 2 ~ 4 所示.

表 2　基于 GMM 的声效识别结果　%

特征类型	耳语	轻声	正常	大声	呼喊
谱特征	92.4	56.7	51.7	57.2	62.4
MFCC	90.7	72.9	64.6	68.2	74.7
SIE	92.6	75.1	68.5	71.6	77.5

表 3　基于 MLP 的声效识别结果　%

特征类型	耳语	轻声	正常	大声	呼喊
谱特征	93.4	58.4	53.2	59.0	63.9
MFCC	91.6	73.7	65.8	70.0	76.1
SIE	93.5	75.8	69.2	72.9	79.3

表 4　基于 SVM 的声效识别结果　%

特征类型	耳语	轻声	正常	大声	呼喊
谱特征	94.2	59.7	54.4	59.3	64.7
MFCC	93.3	75.5	67.2	71.7	77.5
SIE	94.2	77.2	70.6	74.3	80.4

从上述 3 个表中可以看出，不管使用哪种分类模型，SIE 特征的表现都是最好，意味着谱信息熵对于声效变化更加敏感. 整体谱特征在识别耳语时的精度与 SIE 相近，在识别其余 4 种声效模式时与 MFCC 特征及 SIE 特征的差距较大.

表 5 所示为利用式(11)定义的模型融合方法

来识别声效模式的结果. 其中,整体谱特征用于构建 GMM 模型,MFCC 用于训练 MLP,SIE 用于训练 SVM. 当权重系数 α 的值位于区间[0.15, 0.3],β 的值位于区间[0.2, 0.4]时识别效果最优. 可以看出,当利用模型融合的方式来利用 3 种特征识别声效时,其精度要高于表 2 ~ 表 4 中单独使用 1 种特征时的精度.

表 5 基于模型融合的声效识别结果 %

模型组合	耳语	轻声	正常	大声	呼喊
GMM/MLP/SVM	94.6	78.6	72.3	76.0	81.5

基于互补模型方法用于声效检测结果如表 6 所示. 与模型融合的方法不同,表 6 中所有的模型 GMM*/MLP*/SVM* 都是同时使用 3 种特征来训练获得.

表 6 基于互补模型的声效识别结果 %

模型组合	耳语	轻声	正常	大声	呼喊
GMM*/MLP*	94.2	78.4	71.9	75.7	81.4
MLP*/SVM*	94.9	79.0	72.8	76.5	81.9
GMM*/SVM*	95.5	79.7	73.1	77.4	82.3

从表 6 中可以看出,互补模型 MLP*/SVM* 和 GMM*/SVM* 的识别结果要高于表 5 中的结果,这意味着互补模型在集成多类特征进行声效检测方面似乎要比式(7)中定义的模型融合方法效果更好. 互补模型 GMM*/MLP* 的识别结果略微低于表 5 中的结果.

为了分析 3 种特征在互补模型下的表现,在采用互补模型 GMM*/SVM* 的情况下,分别利用不同的特征组合来训练上述 2 种模型,检测结果如表 7 所示. 可以看出,在识别耳语时 3 种特征组合的表现比较接近,原因可能是 3 种特征对于耳语均表现出了很强的辨识能力. 在识别其余 4 种声效模式时,特征组合 MFCC/SIE 表现最优,其识别精度与使用全部 3 种特征时的精度比较接近(见表 6 最后 1 行).

表 7 基于 GMM*/SVM* 互补模型的声效检测结果

特征组合	耳语	轻声	正常	大声	呼喊
谱特征/MFCC	95.1	76.9	69.5	74.2	79.4
谱特征/SIE	95.4	78.8	71.6	76.5	81.6
MFCC/SIE	95.2	79.6	72.9	77.1	82.1

通过分析上述实验结果,可以有以下结论:1)相比整体谱特征和 MFCC 特征,谱信息熵由于精确描述了元音信号在不同频域的能量分布信息,对于声效模式具有更强的辨识能力;2)3 种特征均蕴含了声效模式相关的显著性信息,且在确定 1 种特征的情况下引入其余 2 种特征,可以减少条件熵,提高模型的精度,即 3 种特征具有互补性. 在 2 类特征的组合中,MFCC/SIE 组合的互补性更强;3)采用不同的方法建模所有的特征,通过不同的模型在不同的侧面刻画声效的属性,然后再融合不同模型的模型互补方法可以很好地刻画了声效的这种属性. 3 种互补模型组合中,GMM*/SVM* 的效果明显更好,原因可能是这 2 个模型分布的 K-L 距离较大.

4 结束语

针对语音声效检测问题,提出了描述元音信号在各频域能量分布的谱信息熵特征,并分析了该特征与 MFCC 特征在元音域对声效变化的敏感性;最后引入了基于互补模型的特征融合方法进行声效模式识别. 实验结果表明,与整体谱特征和 MFCC 特征相比,谱信息熵特征蕴含了更丰富的声效模型显著性信息. 而互补模型能够更好地集成 3 种特征的声效显著性信息,进而提高了检测精度.

参考文献:

[1] Hartmut T, Anders E. Acoustic effects of variation in vocal effort by men, women, and children [J]. Journal of the Acoustical Society of America, 2000, 107(6): 3438-3451.

[2] Ghaffarzadegan S, Bořil H, Hansen JHL. UT-rocal effort Ⅱ: analysis and constrained-lexicon recognition of whispered speech [C]// ICASSP. Florence: IEEE, 2014: 2563-2567.

[3] Saeidi R, Alku P, Backstrom T. Feature extraction using power-law adjusted linear prediction with application to speaker recognition under severe vocal effort mismatch [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(1): 42-53.

[4] Tumomo R, Antti Suni, Jouni P, et al. Analysis and synthesis of shouted speech [C]// Interspeech. Lyon : ISCA, 2013: 1544-1548.

[5] Zhang Chi, Hansen, John H L. Advancements in whisper-island detection within normally phonated audio streams[C]// Interspeech. Brighton: ISCA, 2009: 860-863.

[6] Carlin M A, Smolenski B Y, Wenndt S J. Unsupervised

detection of whispered speech in the presence of normal phonation[C]// Interspeech. Pittsburgh: ISCA, 2006: 685-688.

[7] Zhang Chi, Hansen, John H L. Analysis and classification of speech mode: whispered through shouted[C]// Interspeech. Antwerp: ISCA, 2007: 2289-2292.

[8] Petr Z, Milan S, Jiri S. Impact of vocal effort variability on automatic speech recognition[J]. Speech Communication, 2012, 54(6): 732-742.

[9] 晁浩, 宋成, 刘志中. 基于元音模板匹配的声效多级检测[J]. 北京邮电大学学报, 2016, 39(4): 98-102.
Chao Hao, Song Cheng, Liu Zhizhong. Multi-level detection of vocal effort based on vowel template matching[J]. Journal of Beijing University of Posts and Telecommunications, 2016, 39(4): 98-102.

[10] Ni Chongjia, Liu Wenju, Xu Bo. Mandarin prosodic break detection based on complementary model [C]// ISCSLP. Tainan: IEEE, 2011: 353-357.

[11] Ni Chongjia, Liu Wenju, Xu Bo. From English pitch accent detection to mandarin stress detection, where is the difference? [J]. Computer Speech & Language, 2012, 26(3): 127-148.

[12] Chang Chihchung and Lin Chihjen. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-27.