文章编号:1003-0530(2016)09-1117-09

听觉注意模型的语谱图语音情感识别方法

张昕然[1] 查 诚[1] 宋 鹏[2] 陶华伟[1] 赵 力[1]
(1. 东南大学水声信号处理教育部重点实验室, 江苏南京 210096; 2. 烟台大学计算机与控制工程学院, 山东烟台 264005)

摘 要: 在语音情感识别技术中, 由于噪声环境、说话方式和说话人特质原因, 造成特征向量空间分布不匹配的情况。从语音学上分析, 该问题多存在于跨数据库情感识别实验。训练的声学模型和用于测试的语句样本之间的错位, 会使语音情感识别性能剧烈下降。语谱图的特征能从图像的角度对现有情感特征进行有效的补充。本文据此所研究的听觉选择性注意模型, 模拟人耳听觉特性, 能有效探测语谱图上变化的情感特征。同时, 利用时频原子对模型进行改进, 取得频率特性信号匹配的优势, 从时域上提取情感信息。选择注意机制使模型能提取跨语音数据库中的显著性特征, 提高语音情感识别系统的情感辨识能力。实验结果表明, 利用文章所提方法在跨库情感样本上进行特征提取, 再通过典型的分类器, 识别性能提高了约 9 个百分点, 从而验证了该方法对不同数据库具有更好的鲁棒性。

关键词: 语音情感识别; 跨数据库; 语谱图特征; 听觉注意机制; 时频原子
中图分类号: TN912.34 **文献标识码:** A **DOI:** 10.16798/j.issn.1003-0530.2016.09.15

Spectrogram Speech Emotion Recognition Method Based on Auditory Attention Model

ZHANG Xin-ran[1] ZHA Cheng[1] SONG Peng[2] TAO Hua-wei[1] ZHAO Li[1]
(1. Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing, Jiangsu 210096, China; 2. School of Computer and Control Engineering, Yantai University, Yantai, Shandong 264005, China)

Abstract: When there exists mismatch between the trained acoustic models and the test utterances due to noise conditions, speaking styles and speaker traits, unmatched features may appear in cross-corpus. The resulting is the drastic degression in the performance of speech emotion recognition. Hence, the auditory attention model is found to be very effective for variational emotion features detection in our work. The auditory model of selective attention simulating to the human ear hearing characteristics, can effectively detect the changes of emotional features in spectrogram. Meanwhile, Chirplet has been adopted to obtain the advantages of frequency characteristic matching signals and extract emotional information from the time domain. Selective attention mechanism model can extract the salient gist features which show their relation to the expected performance in cross-corpus testing. In our experimental results, the prototypical classifier with the proposed feature extraction approach can deliver a gain of up to about 9% accuracy in cross-corpus speech recognition, which is observed insensitive to different databases.

Key words: speech emotion recognition; cross-corpus; spectrogram feature; auditory attention cues; chirplet

1 引言

在人工智能和模式识别领域,语音情感识别(SER)能够为人机交互提供自然而基本的媒介。随着实用计算机性能的爆炸式进步和语音技术的显著提高,在目前 SER 技术研究中,如何得到大量实用的语音情感数据,即跨数据库问题成为关注的热点[1]。相比语音识别成百上千小时的语音库,SER

收稿日期: 2015-12-23; 修回日期: 2016-03-21
基金项目: 国家自然科学基金项目(61273266, 61375028); 教育部博士点专项基金(20110092130004); 山东省自然科学基金(ZR2014FQ016)

的数据库依然稀少,尤其是公开的可用的数据库。此外,不同的研究人员因研究需要,录制了各自不同的情感语音库,它们具有自己的录制设备、录制环境、录制人员、录制内容、录制语言等等,库与库之间各不相同。因此在进行 SER 时,一般只选用某一个特定的语音情感数据库进行训练和识别,其余的数据库则被排除在外。然而,人耳对不同情感的感知却不受这些外在条件的影响。比如,对于不同的语言,虽然可能无法听懂每条语句的具体含义,但是却可以判断每条语句的情感,即情感的感知不依赖于语义。如果把不同的库整合在一起,一方面扩充了数据库,另一方面也可以找到真正的不依赖于语言和语义的情感特征,对 SER 将会有积极的促进作用。

事实上,情感数据库的泛化性扩展技术是跨库 SER 的关键。情感特征提取在所有 SER 系统中都起到关键性的作用,这部分是本文着重研究讨论的内容。在跨库条件下,如果提取的特征在各个数据库上具有鲁棒性,那么它们必须包含有足够表征说话人情感状态的信息。近年来,在关于跨数据库 SER 的研究工作中已经取得了一些进展,提出了一些解决方案。例如:隐因子分析[2-3],唤醒和效价维度映射[4],基于稀疏自动编码的特征迁移学习[5],以及针对独立说话人的多核学习策略[6]等。以上提到的技术无一例外都使用传统的语言学特征构建 SER 系统,这些特征属于底层描述符(low-level descriptor (LLD)),比如:语音强度,梅尔倒谱系数,基频 F0 参数,过零率等。虽然跨库的 SER 取得一定的研究成果,但是还存在很多的问题。首先从情感类别看,现有的研究基本上处于两类问题的研究。有些文献把情感往唤醒度和效价维映射,把多类情感转换为两个维度的识别[7];有的考虑某一种情感作为目标情感,其余作为非目标情感[8];有的研究只识别单一的情感。通过这样的方式把复杂的多类情感问题转换为简单的两类问题。此外,现阶段的研究,对来自不同库的训练样本和测试样本,只是简单的采用了归一化的方法,并没有从二者的分布特性来分析如何解决其差异问题,所以识别效果并不明显。

因此,现有的传统情感特征遇到瓶颈,从语音频谱图角度寻找更加有效的情感特征成为本文的思路。听觉注意模型来源于生物学原理,它模拟了人类听觉系统运行的进程[9]。在模型研究中,获得输入语音信号的语谱图进行分析,这是模拟人类听觉系统的初级阶段:由耳蜗滤波器和内耳细胞组成的感知系统对语音信息进行预处理[10]。之后,利用听觉显著原理分析语谱图的中心环绕区分度信息,这是模拟听觉系统中从头部基底膜到耳蜗神经核的处理过程。基于听觉的声学频谱图的特征能够描述语音的渐变时间演化过程。再有,这些语谱图特征模仿了人类听觉系统的感知情感信息的能力。关于这点之前的一些研究报告称,语谱图的视觉显著性成分包含重要的语言学信息[11-12]。这些特别的观测报告形成了一类来源于视觉语谱图的语音特征,它们能够探测说话人的情感状态并且与现有语音特征形成互补。听觉注意机制从语谱图中提取如多尺度梯度、能量变化、时间及频率变化等特征,考虑相邻频谱间的互相关信息,这是传统的方向特征所忽略的,并在已有的研究中表明其具有重要的情感能力[13]。

本文安排如下:第一部分关于跨库 SER 进行简要说明,讨论有效的情感特征并引入基于视觉显著性的新特征类型;第二部分提出基于时频原子和语谱图特征的声学注意模型,并将其用于 SER 系统;据此,第三部分在跨数据的语音情感库上进行仿真实验并针对结果分析;最后第四部分对提出的听觉注意 SER 系统进行了讨论和总结。

2 基于语谱图的听觉注意 SER 方法

2.1 基于听觉注意语谱图特征的 SER 模型

跨库情感识别的应用是由于目前的技术已达到瓶颈,需要多数据的支持。然而,增加的不同来源样本会在训练和测试过程中造成差异,使用语谱图特征减少训练测试样本差异,可以对传统情感特征进行补充。本文提出一种基于语谱图特征的听觉注意模型,并用于跨库 SER 系统。模型将听觉注意思想应用于多语音情感数据库上,提取出有效的情感特征。选择性注意机制是人类的高级认知能力,有助于关注目标信息,抑制干扰。目前,该技术在语音情感方面的研究工作主要集中在视觉显著性注意机制上,听觉方面的研究主要集中在生理和认知的研究上,情感计算相关算法类的应用性研究

较少[14]。

考虑到选择性注意机制的特点,并结合听觉理论的研究,本文提出的特征提取模型如图 1 所示。本文提出听觉注意 SER 方法的动机如下:在一个语音频谱图中,我们通常能够辨别出灰度级图像的特质和基音边界周围的不连续点,特别是在元音周围,因为它们往往显示出高能量和清晰的共振峰结构。利用这种“中心-周围”差异的特性[15-16],在语音频谱图中,关注能包含较多情感信息的关键帧,并根据相关性获得周边点的方向、时域的信息。本文利用时频原子对语谱图“中心-周围”进行时频分解,提出的四类特征使 SER 系统能够判断不同情感数据库中多说话人的情绪状态。

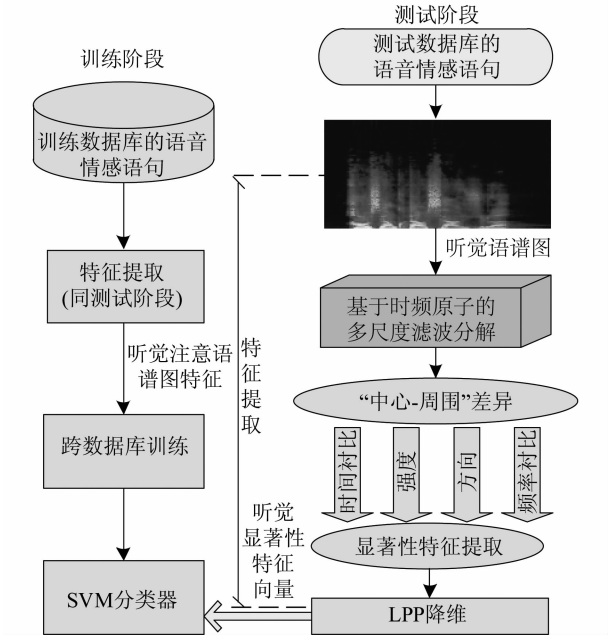


图 1　基于听觉注意的 SER 系统框图

Fig. 1　Diagram of the proposed auditory attention-based SER system

基于听觉注意机制的语音情感识别系统主要分为三个部分:基于听觉注意机制的特征提取、LPP 特征降维和分类识别模型,模型如图 1 所示。

在特征提取部分,首先计算训练集和测试集语音样本的语谱图,在听觉注意模型中,语谱图类似于一个场景的一帧视觉图像。然后仿照听觉注意机制对语谱进行预处理通过多尺度时频原子滤波器将对语谱图信号进行频谱滤波和信号分解。再计算滤波频谱的局部取向(方向特征),时域信息(时间衬比特征),强度和频率衬比作为情感特征向量。将这些特征级联,最终得到局部听觉显著向量。该特征可以用以探测语音的共振峰以及捕捉它们的变化。

在特征降维部分,通过得到的底层声学 Gist 特征,应用局部保留映射(LPP)策略[17]对特征进行降维。LPP 可以在对高维数据进行降维映射后有效地保留数据内部的非性结构,包括语谱图的内在几何性质和局部信息,提高了语音情感特征降维的效率。首先,计算训练集 LPP 投影矩阵,然后采用投影矩阵对训练集和测试集显著特征进行降维处理,得到最终用于分类的特征向量。为了使得 SER 系统对不同的数据库保持不敏感性,还需要同时对情感特征数据进行跨库训练。通过 LPP 降维过程筛选出相关方向的通道,再利用相关通道建立声学显著特征,使来自不同数据库情感类别的特征产生映射。

在分类识别部分,采用基于改进蛙跳算法的支撑向量机算法(Improved Shuffled Frog Leaping Algorithm-Support Vector Machine, Im-SFLA-SVM)[18]实现。Im-SFLA-SVM 采用改进的蛙跳算法来选择 SVM 参数,解决了 SER 中数据规模大情况下训练困难的问题。通过跨库样本训练建立识别模型,从而实现不同情感的区分。

2.2　基于时频原子改进的语谱图特征提取表示

首先利用情感数据库中的语音信号,获得语谱图并表示为能量对数卷积的形式:

$$\mathbf{Im}_{\mu,\nu} = \mathrm{conv}(\mathbf{E}, \psi_{\mu,\nu}) \tag{1}$$

其中,$\mathbf{E}$ 是 Mel 谱的对数能量[7],ψ 为所使用的核函数,μ 表示所使用核的方向,ν 表示核尺度。

由于时频原子具有频率特性信号匹配的优势,将其引入语谱图特征提取模型,对语谱图进行滤波分解,可获得更丰富的情感信息。通过对窗函数 $g(t)$ 的拉伸、调制和平移等边变化,可以得到用于建立 SER 数据库的语谱图情感特征。这些特征构成的向量使得窗函数在时域上具有良好的局部辨识性[19]。当窗函数满足以下条件时,包含语谱图特征的窗函数可以表示为高斯函数 $g(t)=2^{1/4}e^{-\pi t^2}$:

①$\|g\|=1$ 且连续可微;

②窗函数 $g(t)$ 为实函数且 $g(t)\in O((1/t^2+1))$;

③$\int g(t)\mathrm{d}t \neq 0$ 且 $g(0) \neq 0$。

为增加语谱图特征的区分度,需要引入时频原子 Chirplet 以使窗函数维度增加 1,并表示如下:

$$g_{\phi}(t) = \frac{1}{\sqrt{s}} \cdot g\left(\frac{t-u}{s}\right) e^{j\xi t} \tag{2}$$

这里 u 为时间域能量聚集中心,ξ 是波形的相位偏移。满足条件①、②、③的时频原子特征数据库可形成式(2)函数的集合,其中包含了对高斯窗函数的位移、延展和调制。Chirplet 时频原子具有良好的时频特性[20],被广泛应用于信号分解中。本文中基于 Chirplet 时频原子改进的的函数定义如下:

$$\psi_{\mu,\nu}(z) = \frac{\| \boldsymbol{k}_{\mu,\nu} \|^2}{\sigma^2} e^{-\frac{\| k_{\mu,\nu} \|^2 \| z \|^2}{2\sigma^2}} \left[e^{j\boldsymbol{k}_{\mu,\nu} \cdot z} - e^{-\frac{\sigma^2}{2}} \right] \tag{3}$$

$$\boldsymbol{k}_{\mu,\nu} = \begin{pmatrix} k_{\nu}\cos\phi_{\mu} \\ k_{\nu}\sin\phi_{\mu} \end{pmatrix} \tag{4}$$

其中,$z=(i,j)$ 是像素点的空间位置,σ 表示高斯函数的半径。$\phi_{\mu}=\pi\mu/8$。

特征分解图通过 Chirplet 滤波器实现,通常情况下,Gabor 过完备时频原子库可由 Gabor 时频原子经过伸缩、平移、调制而成。由于通过以上方法得到的过完备原子库为一无穷空间,在实际情况中不可用。考虑到语音情感数据库中,样本信号持续时间短、波形变化剧烈、时频局部化信息量较广等特点,本文拟在 Gabor 原子基础上,增加相位偏移 ξ。因此,Chirplet 是对传统三参数的 Gabor 函数,进行调制转换,并增加调频参数而构成的[19]。其形式如下:

$$g_{\phi}(t) = \frac{1}{\sqrt{s}} g\left(\frac{\lambda}{s}\right) \exp[i(\xi(\lambda) + 0.5c\lambda^2)] \tag{5}$$

这里 $\lambda=(t-u)$,同时 $\phi=(s,u,\xi,c)$ 为时频参数,其中包括 u、ξ、脉宽价差 s 和频率调制斜率 c。由于时频原子具有多尺度特征性,本文的方法优越于依赖单一尺度特征提取的传统稀疏信号分解法[21]。这样,多尺度 Chirplet 通过频率参数,提取语谱图中的时间衬比特征,对比传统的 Gabor 小波,取得了频率特性信号匹配的优势。

2.3 语谱图的 Chirplet 时频原子图谱

利用改进的时频原子表示方法对语谱图进行图像分解。在接下来的阶段中,从图谱提取的多尺度特征集由四类特征组成:强度特征(I)、频率衬比特征(F)、时间衬比特征(T)和方向特征(O)。这里,强度特征对应颜色特征,代表语谱图能量特征;频率衬比对应亮度特征;而时间衬比由公式(5)的时频参数来定义;方向特征对应语谱图相邻频率间的相关梯度特征。

本文设置的 Chirplet 为 2 维方向滤波器,根据式(5)从分解的语谱图中提取情感特征。通过在相应尺度中滤波器和图像的卷积,可以获得各方向通道上的灰度级分解图像。不同方向上的多尺度多角度图像特征可以表示为:

$$P_{\theta}(\sigma) = |P_I(\sigma) * G_0(\theta)| + |P_I(\sigma) * G_{\pi/2}(\theta)| \tag{6}$$

在本文的研究中,将时频原子参数设置为:4 尺度和 6 方向。然后应用生成的 Chirplet-Gauss 核函数,与语谱图的灰度图像进行卷积运算。进而可得到语谱图的 24 个多尺度多角度时频原子图谱,这里方向特征我们取 $\theta=\{0°,30°,45°,90°,120°,135°\}$,如图 2 所示。根据图谱提取的图像特征能反映所代表语音样本的情感状态,并用于情感识别。

颜色通道特征图像包含两组。根据德国生理学家赫林提出的拮抗色学说,用 R-G 和 B-Y 的拮抗作用来表示颜色信息对最终显著图的贡献[22],这两对颜色通道特征图像由如下公式算得:

$$P_{\text{R-G}}(\sigma) = (r-g)/\max(r,g,b) \tag{7}$$

$$P_{\text{B-Y}}(\sigma) = (b - \min(r,g))/\max(r,g,b) \tag{8}$$

式中,$P_{\text{R-G}}(\sigma)$ 和 $P_{\text{B-Y}}(\sigma)$ 分别表示 R-G 和 B-Y 颜色对在尺度 σ 上的分解图像,r、g、b 分别表示一幅彩色图像中红、绿、蓝分量值。

亮度通道特征图像由图像的 r、g、b 分量的平均值来表示:

$$P_{\text{I}}(\sigma) = (r+g+b)/3 \tag{9}$$

式中,$P_{\text{I}}(\sigma)$ 表示在相应尺度 σ 上的亮度通道分解图像。

这里,I 特征滤波器相当于脑干听觉投射区的接收区域,具有刺激兴奋阶段和同步对称的禁止边频带[10]。模型中的每个 Chirplet 滤波器,都可以探测

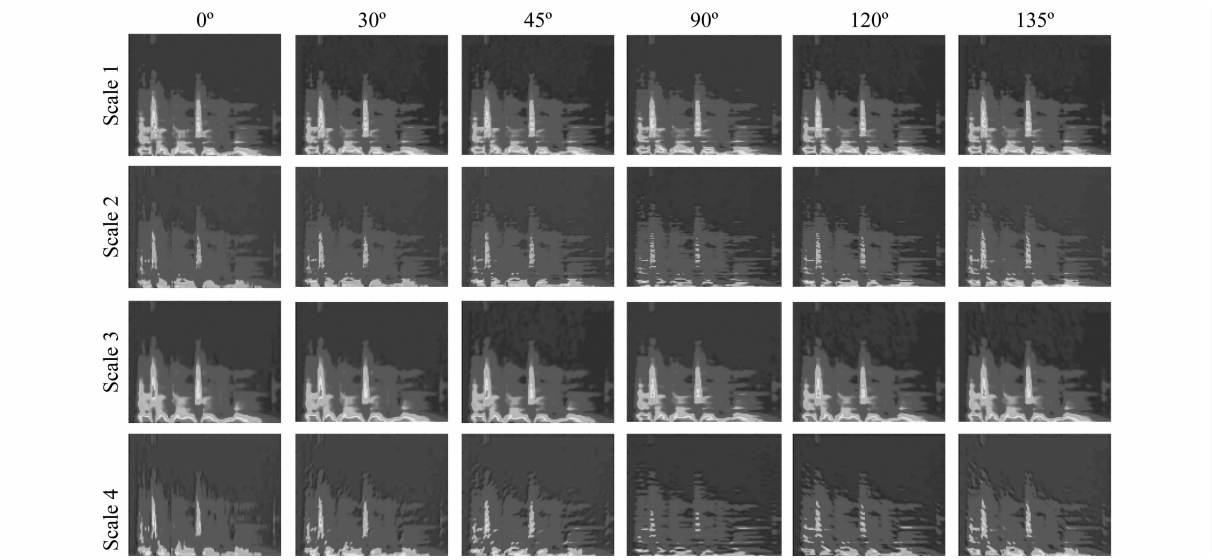


图 2　4 尺度 6 角度的 Chirplet 时频原子语谱图

Fig. 2　Chirplet spectrograms with 4 scales and 6 directions

和获取对应图谱区域语音信号的变化,并最终将其形成情感特征数据。再有,F 和 T 特征滤波器能够分别在频率和时间坐标轴上检测和捕捉信号的变化。同时,O 特征滤波器会探测显著灰度尺度图像中波纹的起伏变化(上扬或下降曲线)并记录[8]。值得一提的是,模型中探测异变点阶段对高区别性特征的提取和数据库的泛化至关重要。进一步,提取出的四类特征 I、F、T、O 的映射构成听觉显著向量,再通过对每个特征映射的增广得到累加的显著向量。最后,应用 LPP 对增广听觉显著向量进行降维和去除冗余,再将不同方向和尺度下得到的四类特征级联,可以得到用于分类识别的语谱图特征。

事实上,根据中心听觉系统的处理阶段[23-24],利用本文提出的基于时频原子改进的听觉选择注意模型,从语谱图中提取出这些特征。提取过程使用 Chirplet 滤波器模拟人类脑干听觉皮层的分析阶段,而多尺度情感特征反映了跨数据库情况下,不同来源的语音样本在语言学上的特质,提高了语音情感辨识度。

3　仿真实验

3.1　语音情感识别实验设置

本文使用三种语音情感数据库进行 SER 特征评估实验,包括:柏林库(EMO-DB)[25],eNTERFACE[26]库和一个中文语音情感库(CNDB)。关于 CNDB 数据库,它是由东南大学语音情感实验室录制并建立的。库中包含的语音语句来自于本实验室师生,并使用影像技术、噪声刺激、视频剪辑观看和电脑游戏等方法诱发录制的。然后为保证录制声音的质量,对实验者进行听力筛选以剔除明显的质量问题语句。经过滤波后,数据库中保留了 4617 条不同情感的语句样本。

为验证跨数据库 SER 性能,训练和测试样本中包含情感类别需要保持一致性。在三个语音情感数据库里,选取了五种情感作为对比实验数据:“生气”,“厌恶”,“恐惧”,“喜悦”和“悲伤”,而其他的情感类别被剔除。本文评价情感特征性能的数据库平台,都遵从“leave-one-corpus (LOCO)”的跨库分析策略,也就是:使用一个数据库作为测试样本集,另外则两个用于监督或者非监督的训练。这种方法对比决策融合策略,在结合 SVM 分类器用于跨数据库 LOCO 实验中具有优越的分类识别性能[27]。实验中 SVM 分类器核函数采用高斯核函数,参数根据支持向量个数为 43 的蛙跳算法[18],设置为:惩罚系数 $C'=100$,核方差 $\sigma=50$。关于实验识别率的定义标准,本文采用情感类别正确判别数与总数据量的比值,即:“识别正确率=1-误报率”。这里,错误率由漏验率和虚警率组成。

为了与听觉注意语谱图特征进行对比分析,本文使用了 Interspeech 2010[28] 中定义的标准特征提取方法获得传统特征集作为对照实验。实验中利

用 openSMILE 工具包[29]进行特征提取,设置特征构成成分参数为 1582,包括 34 个底层描述符和它们相应的一阶 delta 回归系数。本文通过实验对比了基于传统特征和基于语谱特征的语音情感识别算法。所选用传统特征包括音质特征、韵律特征和混沌特征[26],结合 LPP 分析,选取 11 维特征向量作为主特征。本文方法特征包含:语谱图像特征(Spectral Pattern features, SPs)和谐波能量特征(Harmonic Energy features, HEs)[28,30]。其中,语谱图像特征包含每个子带的每帧语谱图的平均值和每个子带的每幅语谱图的相对值,一共 188 维;谐波能量特征指的是谐波能量包络的统计特征,一共 234 维。考虑目前情感识别特征都是传统特征叠加上新特征的做法,因此本文进行特征分析,都是在传统特征的基础上叠加不同语谱特征构成。情感分类算法都采用 SVM 算法并用蛙跳算法进行参数设置。

3.2 实验结果和分析

表 1 给出了跨数据库和单一数据库情况下,使用两种特征提取方法的识别结果。其中三类情况对比实验包括:"实验 1"、"实验 2"和"实验 3"。在"实验 1"中,eNTERFACE 和 CNDB 作为训练数据库,而 EMO-DB 的情感样本被用于测试;在"实验 2"中,eNTERFACE 单独作为测试样本集,并仍使用 EMO-DB 柏林库进行测试实验。在"实验 3"中,在每个单一语音情感数据库上应用两种特征提取方法,并进行 SER 实验的测试和训练。

表 1 两种特征提取方法在不同数据库类型中的 SER 识别率(%)

Tab. 1 Overall recognition rates of different database types with 2 feature extraction methods(%)

特征提取方法	数据库类型				
	跨数据库实验		单一数据库实验(实验 3)		
	实验 1	实验 2	EMO-DB	eNTER-FACE	CNDB
本文方法	**59.2**	**51.9**	87.6	**78.3**	**72.1**
标准方法	49.6	43.8	**88.4**	73.5	70.5

从表 1 中可以看出,本文提出的基于语谱图的听觉注意特征识别方法相比标准特征提取方法,在两种跨库实验中都取得了更好的识别效果,分别达到了 59.2% 和 51.9%。同时可以发现,对比"实验 2"中单一训练库的方法,"实验 1"在情感识别率上取得了明显提升。这是由于在使用两种数据库进行作为训练集的情况下,所得特征判决函数更具有泛化性,使得该方法更适合于跨库识别。

进一步分析表 1 的实验结果,5 组语音情感识别对比实验中,本文所提的语谱图特征方法在 4 组中的识别性能都优于标准特征提取法,这得益于时频原子在图像特征提取中的全局优势,能较好的提取声谱图中的情感差异信息。同时也说明了听觉注意语谱图特征使 SER 模型在跨库任务中更具有鲁棒性。在"实验 3"的 EMO-DB 库实验中,标准方法取得了略微的优势,这是由于柏林库的语音数据量较少,且研究技术较成熟,针对它而建立的标准特征提取模型性能提高空间小,因此语谱图特征在该单一数据库上并未取得更好的识别效果。

语谱图特征适合于语音情感识别的证据可以由图 3 表现出来。音高特征是已知经典语音特征,在单一数据库的 SER 系统中具有良好的类别区分度。因此,在语谱图特征分析实验中,采用了基于语谱图可视化的强度特征和传统音高特征进行对比。图 3 所示是选取实验 1 和 2 中"悲伤"和"厌恶"两种情感的特征向量,将其中传统音高特征和语谱图强度特征进行对比。图中归一化的散点分布图展示了两种特征在识别"悲伤"和"厌恶"情感时的表现情况。可以看出,语谱图强度特征(横轴)能够较为准确的区分出两类情感数据。相比之下,音高特征(纵轴)对两种情感的分类结果出现了巨大的重叠区域。利用散点(坐标点)的分布形态反映两种特征变量统计关系。直观表现出语谱图特征相比传统音高特征具有更好的分类性能。

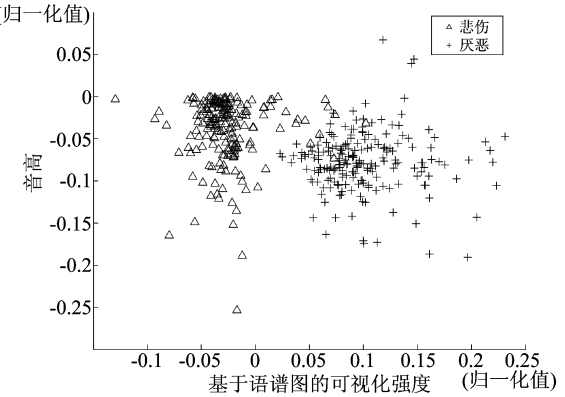


图 3 语谱图强度特征和音高特征 SER 归一化散点图

Fig. 3 Normalization scatter plot of visual-based intensity VS. pitch

由于语音信号是非平稳的自然信号,而像音高和能量这类特征在语言学使通常被称为韵律,它们是基于帧级别的框架内的局部特征,只能描述语音信号特质的单个或少数几个方面[31]。相比之下,语谱图是语音信号在时域和频域上的联合表示。那么,语谱图的分析和整合则是为一种全局化的语音信号处理思想。因此,全局特征(语句级别)可以看作是从语句中提取的所有局部特征的统计,比如图 3 的基于可视化的语谱图强度特征。另外,听觉注意思想是一种基于显著性理论的生理感知机制。鉴此,在本文研究中提出基于该思想的特征提取方法,并应用于 SER 系统建模,通过去除无效的频谱成分来提高情感特征的时效性。

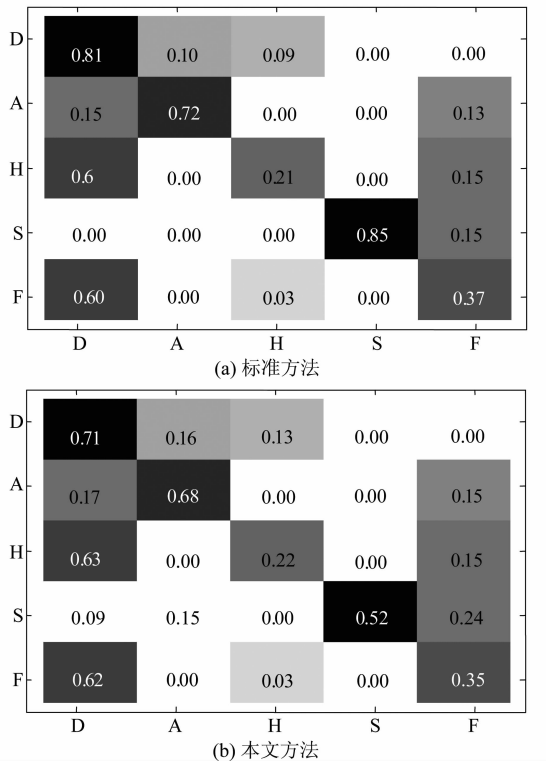


图 4　“实验 1”中跨数据库 SER 模型的混淆矩阵

Fig. 4　Confusion matrix of cross-corpus SER in case1

图 4 展示了“实验 1”两类特征提取方法在跨语料库仿真实验的混淆矩阵。图中给出了所有语音情感类别的识别准确率,包括:厌恶(D)、生气(A)、喜悦(H)、悲伤(S)和恐惧(F)。从混淆矩阵可以看出,“厌恶”和“悲伤”情绪取得了最好的识别效果。同时,“喜悦”和“恐惧”两类情感的模糊度最高,并可以定义为所使用特征提取方法的误判情况。事实上,引起这种混淆的原因主要是本文 SER 实验中所使用的大部分声学特征都与唤醒度相关[32],它们对于与效价度联系紧密的情感类型(如“喜悦”和“恐惧”)的区分度较低。相比之下,在更适合语谱图特征识别的唤醒度情感类别之间,优越的平均识别率显示出其性能的较大提高,也说明本文提出的方法在跨数据库的 SER 中具有优势,从不同尺度语谱图特征间相关性的角度,为跨库 SER 提供了新思路。

4　结论

本文中我们基于听觉注意思想提出了一种新的跨数据库语音情感识别方法。该方法能够利用对语谱图特征变化的捕捉,有效的探测语音场景中的声学情感活动。而这些提取到的特征变化正是使样本点区别于其相邻点之间的显著性感知信息。然后,通过加入带有多尺度的 Chirplet 时频原子对特征提取模型进行改进,通过形成的过完备原子库提高语谱图特征提取包含的信息量。来自多个数据库的样本具有多成分的特点,其频率特性具有时变性。本文所提出的听觉注意模型相比传统特征方法更适用于跨数据库中的非平稳语音情感特征,使模型更加适合不同数据来源语音情感的分类。在跨数据库上的仿真实验结果表明,所提听觉注意方法相比传统标准特征提取方法具有更优越的 SER 性能。在对较旧的柏林库的实验显示出,本方法对数据量的较低的单一情感类别性能提升有限。因此,分析语谱图特征与现有标准特征的联系,形成融合特征进行情感识别是下一步研究的方向。

参考文献

[1] Huang R S. Information Technology in An Improved Supervised Locally Linear Embedding for Recognizing Speech Emotion[C]// Advanced Materials Research. Trans Tech Publ, 2014: 375-378.

[2] Song P, Jin Y, Zha C, et al. Speech emotion recognition method based on hidden factor analysis[J]. Electronics Letters, 2014, 51(1): 112-114.

[3] Zhang X, Tao H, Zha C, et al. A Robust Method for Speech Emotion Recognition Based on Infinite Student's-Mixture Model[J]. Mathematical Problems in Engineering, 2015, 2015: 58-70.

[4] Schuller B, Zhang Z, Weninger F, et al. Synthesized speech for model training in cross-corpus recognition of human emotion[J]. International Journal of Speech Tech-

nology, 2012, 15(3): 313-323.

[5] Deng J, Zhang Z, Marchi E, et al. Sparse autoencoder-based feature transfer learning for speech emotion recognition[C]// Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on. IEEE, 2013: 511-516.

[6] Yun J, Peng S, Zheng W, et al. Speaker-independent speech emotion recognition based on two-layer multiple kernel learning[J]. IEICE Transactions on Information and Systems, 2013, 96(10): 2286-2289.

[7] Jin Y, Song P, Zheng W, et al. A feature selection and feature fusion combination method for speaker-independent speech emotion recognition[C]// Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014: 4808-4812.

[8] Ivanov A, Riccardi G. Kolmogorov-Smirnov test for feature selection in emotion recognition from speech[C]// Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012: 5125-5128.

[9] Alho K, Salmi J, Koistinen S, et al. Top-down controlled and bottom-up triggered orienting of auditory attention to pitch activate overlapping brain networks[J]. Brain research, 2015, 1626: 136-145.

[10] Kong L, Michalka S W, Rosen M L, et al. Auditory spatial attention representations in the human cerebral cortex [J]. Cerebral Cortex, 2014, 24(3): 773-784.

[11] Ajmera P K, Jadhav D V, Holambe R S. Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram[J]. Pattern Recognition, 2011, 44(10-11): 2749-2759.

[12] Kalinli O, Narayanan S. Prominence detection using auditory attention cues and task-dependent high level information[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2009, 17(5): 1009-1024.

[13] Mao Q, Dong M, Huang Z, et al. Learning salient features for speech emotion recognition using convolutional neural networks[J]. Multimedia, IEEE Transactions on, 2014, 16(8): 2203-2213.

[14] Shinn-Cunningham B, Best V. Auditory Selective Attention[J]. The Handbook of Attention, 2015: 99.

[15] Stevens C, Harn B, Chard D J, et al. Examining the role of attention and instruction in at-risk kindergarteners electrophysiological measures of selective auditory attention before and after an early literacy intervention[J]. Journal of Learning Disabilities, 2013, 46(1): 73-86.

[16] Kayser C, Petkov C I, Lippert M, et al. Mechanisms for allocating auditory attention: an auditory saliency map [J]. Current Biology. 2005, 15(21): 1943-1947.

[17] Wong W K, Zhao H T. Supervised optimal locality preserving projection[J]. Pattern Recognition, 2012, 45(1): 186-197.

[18] 张潇丹,胡峰,赵力. 基于改进的蛙跳算法与支持向量机的实用语音情感识别[J]. 信号处理, 2011,27(5): 678-689.
Zhang Xiaodan, Hu Feng, Zhao Li. Recognition of Practical Speech Emotion based on Improved Shuffled Frog Leaping Algorithm and Support Vector Machine[J]. Signal Processing, 2011,27(5):678-689. (in Chinese)

[19] Yin Q, Qian S, Feng A. A fast refinement for adaptive Gaussian chirplet decomposition[J]. Signal Processing, IEEE Transactions on, 2002, 50(6): 1298-1306.

[20] Bayram I. An analytic wavelet transform with a flexible time-frequency covering[J]. Signal Processing, IEEE Transactions on, 2013, 61(5): 1131-1142.

[21] 范海宁,郭英,艾宇. 基于原子分解的跳频信号盲检测和参数盲估计算法[J]. 信号处理, 2010,26(5): 695-702.
Fan Haining, Guo Ying, Ai Yu. Blind detection and parameter estimation algorithm based on atomic decomposition[J]. Signal Processing, 2010,26(5):695-702. (in Chinese)

[22] Eagleman D M. Visual illusions and neurobiology[J]. Nature Reviews Neuroscience, 2001, 2(12): 920-926.

[23] Noriega G. A Neural Model to Study Sensory Abnormalities and Multisensory Effects in Autism[J]. Neural Systems and Rehabilitation Engineering, IEEE Transactions on, 2015, 23(2): 199-209.

[24] Khoubrouy S, Panahi I, Hansen J H. Howling Detection in Hearing Aids Based on Generalized Teager-Kaiser Operator[J]. Audio, Speech, and Language Processing, IEEE/ACM Transactions on, 2015, 23(1): 154-161.

[25] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech[C]// Interspeech, 2005: 1517-1520.

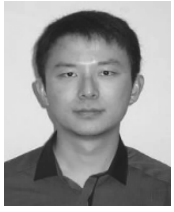
[26] Martin O, Kotsia I, Macq B, et al. The eNTERFACE'05 audio-visual emotion database[C]// Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on. IEEE, 2006: 8.

[27] Moustakidis S, Mallinis G, Koutsias N, et al. SVM-based fuzzy decision trees for classification of high spatial resolution remote sensing images[J]. Geoscience and Remote Sensing, IEEE Transactions on, 2012, 50(1): 149-169.

[28] Schuller B, Steidl S, Batliner A, et al. The INTERSPEECH 2010 paralinguistic challenge[C]// INTER-

SPEECH. 2010: 2794-2797.
[29] Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor [C]//Proceedings of the international conference on Multimedia. ACM, 2010: 1459-1462.
[30] Kalinli O. Syllable Segmentation of Continuous Speech Using Auditory Attention Cues [C] // INTERSPEECH, 2011: 425-428.
[31] Ali S A, Khan A, Bashir N. Analyzing the Impact of Prosodic Feature (Pitch) on Learning Classifiers for Speech Emotion Corpus[J]. International Journal of Information Technology and Computer Science (IJITCS), 2015, 7(2): 54.
[32] Kim E H, Hyun K H, Kim S H, et al. Improved emotion recognition with a novel speaker-independent feature[J]. Mechatronics, IEEE/ASME Transactions on, 2009, 14 (3): 317-325.

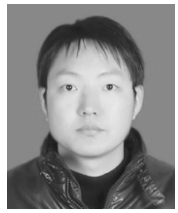
作者简介



张昕然　男,1987 年生,河南开封人,东南大学信息科学与工程学院博士生。主要研究方向为语音情感识别。
E-mail: zxrzxr87324@126.com

查　诚　男,1979 年生,东南大学信息科学与工程学院博士生。主要研究方向为语音情感识别。
E-mail: chengzha@seu.edu.cn

宋　鹏　男,1983 年生,烟台大学助理教授,主要研究方向为语音信号处理、语音情感识别。
E-mail: pengsongseu@gmail.com



陶华伟　男,1987 年生,东南大学信息科学与工程学院博士生。主要研究方向为语音情感识别。
E-mail: ttkltao@163.com

赵　力(通讯作者)　男,1958 年生,东南大学信息科学与工程学院教授,博士生导师。主要研究方向为语音信号处理、图像信号处理。
E-mail: zhaoli@seu.edu.cn