第 38 卷 第 2 期
2019 年 3 月

应用声学
Journal of Applied Acoustics

Vol. 38, No. 2
March, 2019

◇ 研究报告 ◇

基于变分模态分解的语音情感识别方法*

王玮蔚[1] 张秀再[1,2†]

(1 南京信息工程大学电子与信息工程学院 南京 210044)

(2 江苏省大气环境与装备技术协同创新中心 南京 210044)

摘要 针对传统语音情感特征参数在进行情感分类时性能不佳的问题，该文提出了一种基于变分模态分解的语音情感识别方法。情感语音信号首先由变分模态分解提取固有模态函数，然后对所选主导固有模态函数进行重新聚合，再提取梅尔倒谱系数和各固有模态函数的希尔伯特边际谱。为了验证该文提出的特征性能，选用两种语音数据库 (EMODB、RAVDESS) 进行实验，按该文方法提取特征后使用极限学习机进行语音情感分类识别。实验结果表明：相比基于经验模态分解和集合经验模态分解的语音情感特征，该文提出的特征有更好的识别性能，验证了该方法的实用性。

关键词 变分模态分解，Mel 倒谱系数，希尔伯特谱，极限学习机

中图法分类号: TN912.34 **文献标识码:** A **文章编号:** 1000-310X(2019)02-0237-08

DOI: 10.11684/j.issn.1000-310X.2019.02.013

Speech emotion recognition based on variational mode decomposition

WANG Weiwei[1] ZHANG Xiuzai[1,2]

(1 *Nanjing University of Information Science and Technology, Nanjing* 210044, *China*)

(2 *Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology CICAEET, Nanjing* 210044, *China*)

Abstract In view of the problem of poor performance of traditional speech emotion feature parameters in emotion classification, this paper proposes a speech emotion recognition method based on variational mode decomposition (VMD). The emotion speech signal is first extracted by the VMD into the intrinsic mode functions (IMF), then the selected dominant IMFs are re-aggregated, after that the Mel frequency cepstral coefficents (MFCC) and the Hilbert marginal spectrum of each IMF are extracted. In order to verify the performance of the features proposed in this paper, two speech databases(EMODB、RAVDESS) are selected for the experiment. After extracting features according to the method of this paper, the extreme learning machine (ELM) is used for speech emotion classification and recognition. The experimental results show that compared with the emotion features based on empirical mode decomposition (EMD) and ensemble empirical mode decomposition (EEMD), the features proposed in this paper have better recognition performance, and the practicability of the method is verified.

Key words Variational modal decomposition, Mel frequency cepstral coefficents, Hilbert marginal spectrum, Extreme learning machine

2018-07-26 收稿; 2018-10-15 定稿

*江苏省自然科学青年基金项目 (BK20141004), 国家自然科学青年基金项目 (11504176，61601230), 江苏高校优势学科建设工程资助项目

作者简介: 王玮蔚 (1993-), 男, 江苏扬州人, 硕士研究生, 研究方向: 语音情感分析。

†通讯作者 E-mail: xz_zhang@nuist.edu.cn

0 引言

在多种通信方式中，语音信号是人与人、人与机器通信最快的自然方法。人类甚至可以从语音交流中感觉到说话人的情绪状态。语音情感是分析声音行为的一种方法，是指各种影响(如情绪、情绪和压力)的指针，侧重于语音的非言语方面。在这种情况下，语音情感识别的主要挑战是提取一些客观的、可测量的语音特征参数，这些参数可以反映说话人的情绪状态。近年来，语音情感识别在人机通信、机器人通信、多媒体检索等领域得到了广泛关注。语音情感识别研究主要是利用语音中的情感和语音特征的统计特性，进行一般定性的声学关联[1−2]。

语音情感识别的主要工作为语音情感特征提取和分类网络模型选择。当前国内外的研究方向多为分类网络模型选择，而情感特征提取方向研究内容较为匮乏，因此，提取有效的语音情感特征也是当前语音情感识别的关键任务。2004 年，Ververidis 等[3] 从能量、基音和语音频谱的动态行为中提取出 87 个静态特征，并提出了谱平坦度测度与谱中心的比值作为说话人独立的特征，利用帧级特征、基音周期、能量和 Mel 倒谱系数(Mel frequency cepstral coefficents, MFCC)对性别和情感进行了层次分类。2011 年，Sun 等[4] 将 Teager 能量中提取的小波系数引入到语音情感识别中。2008 年，韩一等[5] 将 MFCC 参数作为特征对语音情感进行识别，也取得了较好的结果。

2011 年，He 等[6] 首先将经验模态分解(Empirical mode decomposition, EMD)引入到语音情感识别中。2015 年，Sethu 等[7] 利用 EMD 将语音进行分解，以分解得到的固有模态函数(Intrinsic mode functions, IMF)分量进行语音分类。Shahnaz 等[8] 将 EMD 和小波分析相结合，通过选取主导 IMF 分量，不仅减少了计算负担，而且避免包含冗余或信息量较少的数据，得到了 80.55% 的语音情感识别准确率。向磊[9] 将集合固有模态函数(Ensemble empirical mode decomposition, EEMD)和希尔伯特(Hilbert)边际谱相结合，有效地解决了传统 EMD 分解带来的模态混叠问题。

为了提高语音情感特征识别性能，解决基于 EMD 和 EEMD 算法的语音情感特征模态混叠和计算量过大的缺点，本文将变分模态分解(Variational modal decomposition, VMD)方法引入到语音情感特征提取中[10]，提出基于 VMD 分解的语音情感特征，采用极限学习机(Extreme learning machine, ELM)将本文特征与语音基音特征、谱特征作为分类特征进行实验。结果表明，相较于传统语音特征以及基于 EMD、EEMD 的语音情感特征，本文提出的特征能更好地表示语音的情感特征，提高了语音情感的识别准确率。

1 特征提取

1.1 VMD 分解

VMD 方法与反复循环剥离进行模态函数分解的 EMD 方法不同，VMD 通过对变分模型的最优极值求解，实现自适应地获取 IMF，在迭代过程中不断更新每个 IMF 分量的中心频率和带宽[10−11]。

IMF 分量表达式为

$$u_k(t) = A_k(t)\cos(\varphi_k(t)), \tag{1}$$

其中，$u_k(t)$ 为第 k 个 IMF 分量，$0 < k < K+1$，$A_k(t)$ 为第 k 个 IMF 分量的幅值，$\varphi_k(t)$ 为第 k 个 IMF 分量的相角，t 为时间。

约束条件为

$$\min_{\{\boldsymbol{u}_k\}\{\boldsymbol{w}_k\}} \left\{ \sum_k \left\| \partial_t \left[\left(\delta(t)+\frac{\mathrm{j}}{\pi t}\right) \cdot u_k(t) \right] \mathrm{e}^{-\mathrm{j}\omega_k t} \right\|_2^2 \right\}$$
$$\text{s.t.} \sum_k u_k = f, \tag{2}$$

式(2)中，$\{\boldsymbol{u}_k\} := \{u_1, \cdots, u_K\}$，$u_k(t)$ 记为 u_k，$\{\boldsymbol{u}_k\}$ 为分解到的 K 个有限带宽的 IMF 分量的集合，u_k 表示分解到的第 k 个有限带宽的 IMF 分量，∂_t 为微分算子，$\delta(t)$ 为狄利克来函数，j 为虚数符号，e 为自然常数，$f(t)$ 为约束函数，$\{\boldsymbol{\omega}_k\} := \{\omega_1, \cdots, \omega_K\}$，$\{\boldsymbol{\omega}_k\}$ 为 K 个 IMF 分量所对应的中心频率的集合，ω_k 表示第 k 个 IMF 分量所对应的中心频率，$\|\cdot\|_2^2$ 表示范数；通过拉格朗日函数求该约束条件下的最优解，生成的拉格朗日表达式为

$$L(\{\boldsymbol{u}_k\},\{\boldsymbol{\omega}_k\},\lambda) = \alpha \sum_k \left\| \partial_t \left[\left(\delta(t)+\frac{\mathrm{j}}{\pi t}\right) u_k(t) \right] \mathrm{e}^{-\mathrm{j}\omega_k t} \right\|_2^2 + \left\| f(t) - \sum_k u_k(t) \right\|_2^2 + \left\langle \lambda(t), f(t) - \sum_k u_k(t) \right\rangle, \tag{3}$$

式 (3) 中，$L(\{\boldsymbol{u}_k\},\{\boldsymbol{\omega}_k\},\lambda)$ 为拉格朗日函数，α 为惩罚系数，$\lambda(t)$ 为拉格朗日乘子，$\langle\cdot\rangle$ 表示内积。

采用乘法算子交替的方法求式 (3) 的鞍点，就得到IMF 分量，求解过程中 u_k^{n+1} 的值会不断更新。公式 (4) 取得最小值时，u_k^{n+1} 与 u_k^n 的误差小于预设值，u_k^{n+1} 为第 $n+1$ 次迭代的第 k 个IMF 分量，其表达式为

$$u_k^{n+1}=\underset{u_k\in\mathbf{X}}{\arg\min}\left\{\alpha\left\|\partial_t\left[\left(\delta(t)+\frac{\mathrm{j}}{\pi t}\right)\cdot u_k(t)\right]\mathrm{e}^{-\mathrm{j}\omega_k t}\right\|_2^2+\left\|f(t)-\sum_i u_i(t)+\frac{\lambda(t)}{2}\right\|_2^2\right\},\tag{4}$$

式 (4) 中，$\mathbf{X}$ 为 u_k 的集合，w_k^{n+1} 为第 $n+1$ 次迭代的第 k 个IMF 分量的中心频率，$\sum_{i\neq k}u_i(t)^{n+1}$ 表示将第 $n+1$ 次迭代的除了第 k 个IMF 分量之外的分量进行求和。

利用Parseval/Plancherel傅里叶等距变换可将式 (4) 转换到频域进行计算，可得到各模态的频域更新，就可将中心频率的取值问题转换到频域，得到中心频率的更新方法；同时更新 λ，表达式如下：

$$\hat{u}_k^{n+1}(\omega)=\frac{\hat{f}(\omega)-\sum_{i\neq k}\hat{u}_i(\omega)+\dfrac{\hat{\lambda}(\omega)}{2}}{1+2\alpha(\omega-\omega_k)^2},\tag{5}$$

$$\omega_k^{n+1}=\frac{\int_0^\infty\omega\left|u_k(\omega)\right|^2\mathrm{d}\omega}{\int_0^\infty\left|\hat{u}_k(\omega)\right|^2\mathrm{d}\omega},\tag{6}$$

$$\hat{\lambda}^{n+1}(\omega)\leftarrow\hat{\lambda}^n(\omega)+\tau\left(\hat{f}(\omega)-\sum_k\hat{u}_k^{n+1}(\omega)\right).\tag{7}$$

每个IMF 分量的频率中心及带宽在模型求解过程中，随着迭代次数不断更新，直到满足迭代条件 $\sum_k\left\|\hat{u}_k^{n+1}-\hat{u}_k^n\right\|_2^2/\left\|\hat{u}_k^{n+1}\right\|_2^2<\mathrm{e}$，即可根据相应的频域特征得到 K 个IMF 分量。该分解模式可以自适应地对信号频带进行切割，有效避免模态混叠，且IMF 分量被固定划分为 K 个，消除了EMD 算法大量的无效分解分量，使得计算量大幅下降 [10]。

1.2　基于VMD-HT 的语音情感特征

对语音信号进行VMD 分解得到IMF 分量后，为了得到能对语音情感分析的特征，利用IMF 分量为平稳信号的特点 [6]，对VMD 各分量进行Hilbert 变换，得到IMF 的瞬时频率和幅值 [12]，特征提取流程如图1所示。

$$H_k(t)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{u_k(t')}{t-t'}\mathrm{d}t,\tag{8}$$

式 (8) 中，$H_k(t)$ 为IMF 分量的Hilbert 变换函数，$u_k(t')$ 为基于时间常数 t' 的第 k 个IMF 分量。

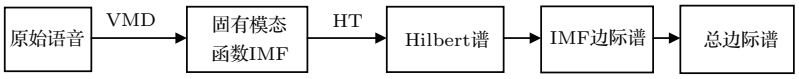


图1　VMD-HT 特征提取流程图

Fig. 1　VMD-HT feature extraction flow chart

$$Z_k(t)=u_k(t)+\mathrm{j}H_k(t)=a_k(t)\mathrm{e}^{\mathrm{j}\theta_k(t)},\tag{9}$$

式 (9) 中，$Z_k(t)$ 为解析函数，$a_k(t)=\sqrt{u_t^2(t)+H_k^2(t)}$ 为第 k 个IMF 分量的瞬时幅值，$\theta_k=\arctan\dfrac{H_k(t)}{u_k(t)}$ 为相位，$u_k(t)$ 为第 k 个IMF 分量，$H_k(t)$ 为第 k 个分量的Hilbert 变换。

式 (9) 中，$Z_k(t)$ 的相位表达方式突出了Hilbert 变换的物理意义，是基于时间序列形成的一个振幅和相位调制的三角函数。则Hilbert 谱的瞬时频率定义为 [8]

$$W_k(t)=\frac{\mathrm{d}\theta_k}{\mathrm{d}t},\tag{10}$$

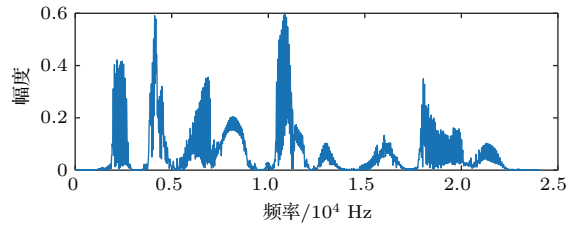
其中，θ_k 表示第 k 个IMF 分量的相位。

然后，对于语音信号第 k 个IMF 分量 $u_k(t)$ 的幅值 $a_k(t)$ 和瞬时频率 $W_k(t)$，计算 $u_k(t)$ 的平均瞬时频率 (Mean instantaneous frequency, MIF)。根据获得的各IMF 分量的MIF 及幅值，计算原始信号的MIF 表示为 [10]

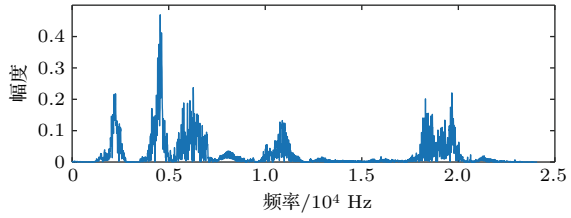
$$\mathrm{MIF}=\frac{\sum_{k=1}^{K}\|a_k\|\mathrm{MIF}_k}{\sum_{k=1}^{K}\|a_k\|}.\tag{11}$$

将各IMF 分量的平均瞬时频率、幅值以及原始信号的瞬时频率作为该语音信号的VMD-HT 特征。

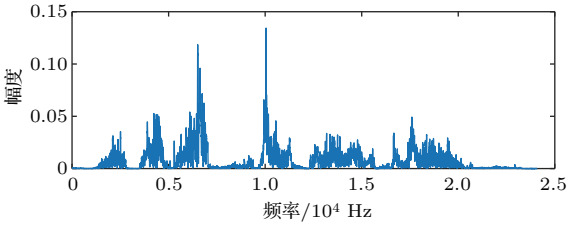
模态 K 通过人为方式进行调整，根据测试结果，K 设置为4时，提取的特征效果最好，以害怕 (FEAR) 语音为例，得到的4个IMF 边际谱如图2所示。



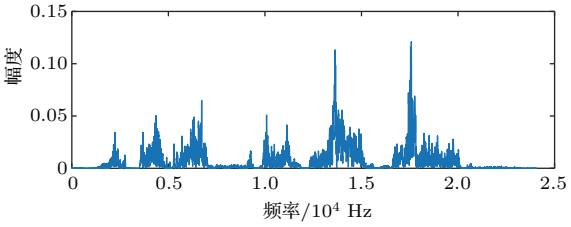
(a) IMF1



(b) IMF2



(c) IMF3



(d) IMF4

图 2　各 IMF 信号的边际谱图

Fig. 2　The marginal spectrum of each IMF signal

1.3　基于 VMD-MFCC 的语音情感特征

MFCC 由 Stevens 在 1937 年提出[11]，MFCC 参数是基于人耳对不同频率声音有不同敏感度的特点提出的，揭示了人耳对高频信号的敏感度低于低频信号的特点。语音信号由频率 f 转换到 Mel 尺度的表达式为[12−13]

$$f_{\mathrm{Mel}}(f)=2595\times\lg(1+f/700). \tag{12}$$

语音信号通过 VMD 分解后，剔除余波分量，再重新聚合，对聚合信号提取 MFCC 参数，即得到 VMD-MFCC 特征。在将信号进行 VMD 分解之后，提取 MFCC 参数的过程分为数步，流程如图 3 所示。

MFCC 参数提取采用一组基于 Mel 尺度的三角带通滤波器，将语音信号转换到频域后，对语音信号进行滤波处理，使语音信号遵循 Mel 尺度的衰减特性。滤波器组对频域信号进行切分，每个频段产生一个对应的能量值。本实验中滤波器个数取 24，因此可得到 24 个能量值。

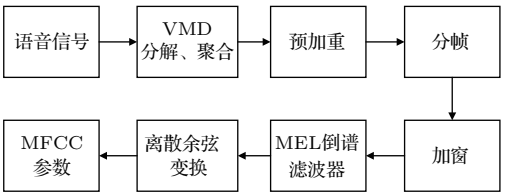


图 3　MFCC 参数提取流程图

Fig. 3　MFCC parameter extraction flow chart

由于人耳对声音的感知程度具有非线性特性，用对数形式描述更好。因此，对能量值进行对数处理，再倒谱分析。

根据 MFCC 定义，对对数能量进行反傅里叶变换，再通过低通滤波器获得低频信号。使用离散余弦变换 (Discrete cosine transform, DCT) 可以直接获取低频信息，DCT 与离散傅里叶变换相似，但只有实数部分，该过程可表示为

$$C_m=\sum_{k=1}^{Q}E_k\times\cos\left[m\left(k-\frac{1}{2}\right)\frac{\pi}{Q}\right],$$
$$m=1,\cdots,L, \tag{13}$$

式 (13) 中，E_k 为第 k 个滤波器的对数能量值；Q 为三角滤波器个数，一般取 22~26；m 为当前计算的 MFCC 特征参数的维数，L 取 12，12 维 MFCC 特征参数足以代表一帧语音特征[14]。

以 EMODB 中害怕情感语句为例，以 256 个点为一帧，帧移为 64，Mel 倒谱滤波器取 24 个，预加重系数为 0.95，计算 12 阶 MFCC 参数如图 4 所示。采用本文方法对语音进行分解后提取的 MFCC 参数如图 5 所示。由图 4 可知，直接提取的 MFCC 特征参数每一帧之间差别较大，经过处理后的语音信号的 MFCC 特征参数每帧之间差别明显降低，可以使 MFCC 特征更易于通过分类器进行识别。

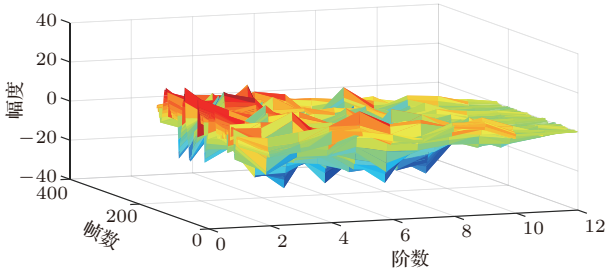


图 4　FEAR 语句 12 阶 MFCC 参数

Fig. 4　FEAR statement 12th order MFCC parameters

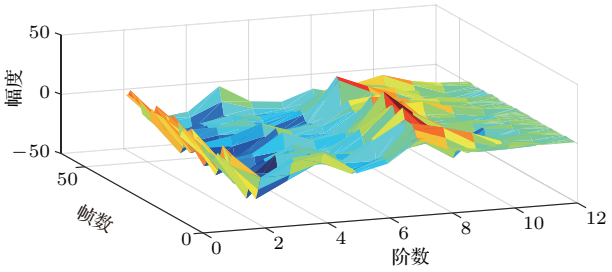


图 5　FEAR 语句 12 阶 VMD-MFCC 参数

Fig. 5　FEAR statement 12th order VMD-MFCC parameters

2　分类算法

2.1　分类算法简介

语音情感识别中最常用的分类器是支持向量机[15−16] (Support vector machine, SVM)、人工神经网络[11,17−18] (Artificial neural network, ANN)、K 最近邻算法[12] (K-nearest neighbor, KNN)、Elman 神经网络[12]、高斯混合模型[19] (Gaussian mixture model, GMM) 长短时神经网络[20](Long short-term memory, LSTM) 和隐马尔可夫模型[10] (Hidden Markov model, HMM)。在众多人工神经网络中，将快速模型学习与准确预测能力相结合的极限学习机，应用于多模式情感识别和计算语言学，以适度的计算资源获得了最好的结果[21−23]。

2.2　ELM 简介

最初，ELM 作为单隐层前馈网络的一种快速学习方法——反向传播的另一种方法提出[21]。与传统的神经网络和机器学习算法相比，ELM 方法学习速度快、泛化性能好。因此，本实验采用 ELM 方法进行情感特征分类，基本 ELM 的体系结构如图 6 所示。

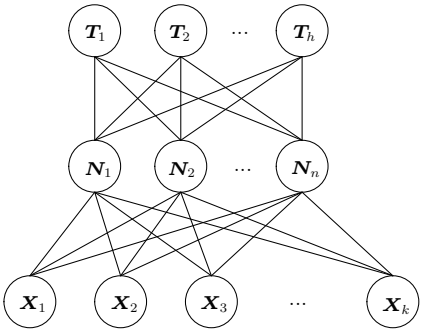


图 6　ELM 基本结构图

Fig. 6　ELM basic structure

$$\sum_{i=1}^{L}\boldsymbol{\beta}_i g(\boldsymbol{W}_i\cdot\boldsymbol{X}_j+\boldsymbol{b}_i)=\boldsymbol{o}_j,\quad j=1,\cdots,N,\tag{14}$$

式 (14) 为 ELM 神经网络处理输入数据的公式，式中 $g(x)$ 为激活函数，$\boldsymbol{W}_i=[w_{i,1},w_{i,2},\cdots,w_{i,n}]^{\mathrm{T}}$ 为输入权重，$\boldsymbol{\beta}_i$ 为输出权重，$\boldsymbol{b}_i$ 为第 i 个隐藏单元的偏置，$\boldsymbol{X}_j$ 是输入的数据，$\cdot$ 表示内积。

单隐层神经网络学习目标是使输出误差最小，表示为

$$\sum_{j=1}^{N}\|\boldsymbol{o}_j-\boldsymbol{t}_j\|=0,\tag{15}$$

即存在 $\boldsymbol{\beta}_i$、$\boldsymbol{W}_i$ 和 $\boldsymbol{b}_i$，使得

$$\sum_{i=1}^{L}\boldsymbol{\beta}_i g(\boldsymbol{W}_i\cdot\boldsymbol{X}_j+\boldsymbol{b}_i)=\boldsymbol{t}_j,\quad j=1,\cdots,N.\tag{16}$$

以矩阵的形式表示为

$$\begin{aligned}&\boldsymbol{N}\boldsymbol{\beta}=\boldsymbol{T},\\&\boldsymbol{N}(\boldsymbol{W}_1,\cdots,\boldsymbol{W}_L,\boldsymbol{b}_1,\cdots,\boldsymbol{b}_L,\boldsymbol{X}_1,\cdots,\boldsymbol{X}_L)\\&=\begin{bmatrix}g(\boldsymbol{W}_1\cdot\boldsymbol{X}_1+\boldsymbol{b}_1)&\cdots&g(\boldsymbol{W}_L\cdot\boldsymbol{X}_N+\boldsymbol{b}_L)\\\vdots&\cdots&\vdots\\g(\boldsymbol{W}_1\cdot\boldsymbol{X}_N+\boldsymbol{b}_1)&\cdots&g(\boldsymbol{W}_L\cdot\boldsymbol{X}_N+\boldsymbol{b}_L)\end{bmatrix}_{N\times L},\\&\boldsymbol{\beta}=\begin{bmatrix}\boldsymbol{\beta}_1^{\mathrm{T}}\\\vdots\\\boldsymbol{\beta}_L^{\mathrm{T}}\end{bmatrix}_{L\times m},\quad\boldsymbol{T}=\begin{bmatrix}\boldsymbol{T}_1^{\mathrm{T}}\\\vdots\\\boldsymbol{T}_N^{\mathrm{T}}\end{bmatrix}_{N\times m},\end{aligned}\tag{17}$$

式 (17) 中，$\boldsymbol{N}$ 为隐含层节点输出，$\boldsymbol{\beta}$ 为隐含层到输出层的权重系数，$\boldsymbol{T}$ 为训练所需要得到的期望结果。为了对隐含层神经元进行训练，得到 $\boldsymbol{\beta}_i$、$\boldsymbol{W}_i$ 和 $\boldsymbol{b}_i$ 的解为

$$\|\boldsymbol{N}(\hat{\boldsymbol{W}}_i,\hat{\boldsymbol{b}}_i)\hat{\boldsymbol{\beta}}_i-\boldsymbol{T}\|=\min_{\boldsymbol{W},\boldsymbol{b},\boldsymbol{\beta}}\|\boldsymbol{N}(\boldsymbol{W}_i,\boldsymbol{b}_i)\boldsymbol{\beta}_i-\boldsymbol{T}\|,\tag{18}$$

式 (18) 中，$i=1,\cdots,L$，该式用最小化损失函数表示为

$$E=\sum_{j=1}^{N}\left(\sum_{i=1}^{L}\boldsymbol{\beta}_i g(\boldsymbol{W}_i\cdot\boldsymbol{X}_j+\boldsymbol{b}_i)-\boldsymbol{t}_j\right)^2.\tag{19}$$

传统的一些基于梯度下降法算法 (如反向传播 (Back propagation, BP)、多层感知器 (Multi-layer perception, MLP)) 可以用来求解这样的问题，但这些学习算法需要在迭代过程中调整所有参数。而 ELM 算法的输入层权重 $\boldsymbol{W}_i$ 和隐含层 $\boldsymbol{b}_i$ 在初始化时已被随机产生且唯一，因此隐含层的输出矩阵

$\boldsymbol{N}$ 就被确定，只需要调整隐含层到输出层的权重系数 $\boldsymbol{\beta}_i$，对该系数的训练可转化为求解一个线性系统 $\boldsymbol{N}\beta = \boldsymbol{T}$。输出权重可由式 (20) 确定，

$$\hat{\boldsymbol{\beta}} = \boldsymbol{N}^{\dagger}, \tag{20}$$

式 (20) 中，$\boldsymbol{N}^{\dagger}$ 是矩阵的 Moore-Penrose 广义逆。可证明求得解的范数最小且唯一，且 ELM 的计算速度较基本梯度下降算法快数倍 [21]。

3 实验验证

3.1 数据集选取

本实验基于德国 BerlinEMODB 语音情感数据库和美国 RAVDESS 视听情感数据库，下面对两种数据库进行简单的介绍。

德国 BerlinEMODB 语音情感数据库是最为常用的公开语音情感数据库之一，它是由德国柏林工业大学录制的德语情感数据库，由 10 位专业演员 (5 男 5 女) 参与录制，得到包含生气、无聊、厌恶、害怕、高兴、中性和悲伤等 7 类基本情感的 800 条语句。对于文本语料的选择遵从选择语义中性、无明显情感倾向的日常语句，且语音在专业录音室中录制而成。经过 20 个说话人的听辨测试，最终得到 494 条情感语句用于实验评价 [11]。

美国 RAVDESS 视听情感数据库是为北美英语的科学家和治疗师提供一个可自由使用的动态视听语音录音库，由 24 名演员 (12 男，12 女) 参与录制，他们用北美英文口音说话和唱歌，语音中包含各种情绪。包含 7356 个情感中性陈述的高品质视频录音，用一系列情绪说出和唱出。演讲集包括 8 个情绪表达：中性、冷静、快乐、悲伤、愤怒、恐惧、惊讶和厌恶。歌曲集包括 6 种情绪表达：中性、冷静、快乐、悲伤、愤怒和恐惧。除了中性以外的所有情绪都表现为两种情绪强度：正常和强烈。有 2452 个独特的发声，所有这些都有三种模式格式：完整的音频-视频 (720p，H.264)、纯视频和纯音频 (波形)。该数据库已经在涉及 297 名参与者的感知实验中得到验证 [24]。

3.2 特征选取

传统语音情感特征为基频特征、韵律谱特征以及部分非线性特征 [10]，本文将 VMD-MFCC、VMD-HT 和传统语音情感特征相结合作为实验选取的特征，称为底层特征，底层特征描述见表 1。

表 1 底层特征描述

Table 1 Description of the underlying features

底层特征描述	维度
短时能量最大值、最小值、均值、方差、帧差、自相关系数、均方误差、能量比	1∼8
基音频率最大值、最小值、均值、方差、第一、二阶抖动	9∼14
浊音频率最大值、最小值、均值、方差	15∼18
第一共振峰的最大值、最小值、均值、方差、一阶抖动	19∼24
VMD-MFCC(0-12 阶) 最大值、最小值、均值、方差	25∼76
DB3 小波高、低频能量、总能量	77∼79
VMD-HT 边际谱 (K=4) 高频、低频、总能量	80∼91

3.3 仿真结果

为了验证 VMD-HT 和 VMD-MFCC 特征在语音情感识别中的应用效果，取两种语音情感数据集中共有的生气、伤心、害怕、开心、中性五种情感，取 10 名说话人的情感语句各 50 句。其中，随机抽取 40 句用来做训练，10 句用来测试，进行 10 次实验，实验结果以 10 次实验识别率的平均值作为评估指标，整个实验与说话人无关。采用 KNN(K=5)、SVM(核函数设置为 sigmoid)、ELM 作为分类方法，输入为 91 维底层情感特征，并采用 Sethu V 的 EMD 特征和向磊的 EEMD 特征进行对比实验，对比实验中的输入特征中 25∼76 和 80∼91 维分别替换为基于 EMD 和 EEMD 的特征。实验结果见表 2、表 3。

由表 2、表 3 可知，ELM 分类准确度要高于 KNN 和 SVM；在两个数据集中，加入 VMD 特征的 ELM 方法分别在中性和害怕情绪的识别率达到最高，而开心情感识别率在两个数据集中都为最低。相较于传统语音情感特征，基于 EMD 的特征通过选取主导 IMF 分量，不仅减少了计算负担，而且避免包含冗余或信息量较少的数据，有效地提升了语音情感识别性能；基于 EEMD 的特征，由于避免了 EMD 分量的模态混叠问题，识别率在 EMD 特征的基础上有所提升；在加入 VMD 特征之后，由于 VMD 分解方法不仅解决了 EMD 方法模态混叠的问题，还提升了 IMF 信号的分解完整性，因此，基于 VMD 的特征在三种分类方式上的识别度都高于基于 EMD 和 EEMD 的特征。以 EMODB 为例，害怕的识别率提高了 2%，中性的识别率提高了 5%，生气的识别率提高了 2%。因此，将 VMD 特征用于语音情感识别，可以有效提高识别准确率，且将 VMD 特征和 ELM 分类器结合，有更好的识别效果。

表 2　EMODB 数据集分类实验结果 (识别率)

Table 2　EMODB data set classification experiment results

	KNN	SVM	ELM	EMD +KNN	EMD +SVM	EMD +ELM	EEMD +KNN	EEMD +SVM	EEMD +ELM	VMD +KNN	VMD +SVM	VMD +ELM
害怕	0.65	0.66	0.77	0.6	0.75	0.83	0.56	0.78	0.88	0.6	0.8	0.9
开心	0.58	0.72	0.71	0.8	0.85	0.87	0.88	0.81	0.85	0.95	0.75	0.88
中性	0.51	0.65	0.68	0.6	0.7	0.82	0.68	0.67	0.89	0.65	0.9	0.94
伤心	0.67	0.81	0.82	0.88	0.75	0.85	0.84	0.73	0.81	0.9	0.8	0.86
生气	0.43	0.67	0.85	0.68	0.75	0.86	0.72	0.85	0.89	0.65	0.8	0.91
平均	0.568	0.702	0.766	0.712	0.76	0.846	0.736	0.768	0.864	0.75	0.81	0.898

表 3　RAVDESS 数据集分类实验结果 (识别率)

Table 3　RAVDESS data set classification experiment results

	KNN	SVM	ELM	EMD +KNN	EMD +SVM	EMD +ELM	EEMD +KNN	EEMD +SVM	EEMD +ELM	VMD +KNN	VMD +SVM	VMD +ELM
害怕	0.71	0.68	0.77	0.85	0.75	0.95	0.79	0.80	0.94	0.8	1	1
开心	0.39	0.77	0.69	0.45	0.85	0.92	0.66	0.85	0.91	0.65	0.9	0.93
中性	0.44	0.69	0.72	0.3	0.9	0.87	0.58	0.86	0.93	0.55	0.95	0.95
伤心	0.57	0.55	0.66	0.95	0.75	0.75	0.77	0.79	0.89	0.9	0.95	0.95
生气	0.61	0.72	0.78	0.85	0.8	0.89	0.81	0.88	0.91	0.9	0.9	0.94
平均	0.544	0.682	0.724	0.68	0.81	0.876	0.722	0.836	0.916	0.76	0.94	0.954

4　结论

根据语音信号非平稳、非线性特点，本文将变分模态分解 (VMD) 引入到语音情感特征识别中，通过 Hilbert 变换和提取 MFCC 参数，组成新的语音情感非线性联合特征。将该特征应用于语音情感识别，实验将基于 VMD 提取的 VMD-MFCC 特征和 VMD-HT 特征与传统语音情感特征相结合，采用极限学习机进行语音情感分类。实验结果表明，相较于基于 EMD 和 EEMD 的情感特征，基于 VMD 的语音特征结合极限学习机进行语音情感分类的方法，具有更高的识别率。

参　考　文　献

[1] Lin Y L, Wei G, Yang K C. Research progress of speech emotion recognition[J]. Journal of Circuits and Systems, 2007, 12(1): 90–98.

[2] Schuller B, Rigoll G, Lang M. Hidden Markov model-based speech emotion recognition[C]//2003 International Conference on Multimedia and Expo. ICME '03. Proceedings, 2003: 401–404.

[3] Ververidis D, Kotropoulos C, Pitas I. Automatic emotional speech classification[C]//2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004, 1: 593–596.

[4] Sun R, Moore E. Investigating glottal parameters and teager energy operators in emotion recognition[M]//Affective computing and intelligent interaction. Berlin, Heidelberg: Springer, 2011: 425–434.

[5] 韩一, 王国胤, 杨勇. 基于 MFCC 的语音情感识别 [J]. 重庆邮电大学学报: 自然科学版, 2008, 20(5): 597–602.
Han Yi, Wang Kuangyin, Yang Yong. Speech emotion recognition based on MFCC[J]. Journal of Chongqing University of Posts and Telecommunications: Natural Science, 2008, 20(5): 597–602.

[6] He L, Lech M, Maddage N C, et al. Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech[J]. Biomedical Signal Processing and Control, 2011, 6(2): 139–146.

[7] Sethu V, Ambikairajah E, Epps J. Empirical mode decomposition based weighted frequency feature for speech-based emotion classification[C]//Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE Interna-

tional Conference on. IEEE, 2008: 5017–5020.

[8] Shahnaz C, Sultana S, Fattah S A, et al. Emotion recognition based on EMD-wavelet analysis of speech signals[C]//Digital Signal Processing (DSP), 2015 IEEE International Conference on. IEEE, 2015: 307–310.

[9] 向磊. 语音情感特征提取与识别的研究 [D]. 杭州: 浙江理工大学, 2013.

[10] Dragomiretskiy K, Zosso D. Variational mode decomposition[J]. IEEE Transactions on Signal Processing, 2014, 62(3): 531–544.

[11] Zhao H, Li L. Fault diagnosis of wind turbine bearing based on variational mode decomposition and Teager energy operator[J]. IET Renewable Power Generation, 2016, 11(4): 453–460.

[12] Grimm M, Kroschel K, Narayanan S. Support vector regression for automatic recognition of spontaneous emotions in speech[C]//Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. IEEE, 2007, 4: IV-1085-IV-1088.

[13] Hu H, Xu M X, Wu W. GMM supervector based SVM with spectral features for speech emotion recognition[C]. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, 2007: 413–416.

[14] Neumann M, Vu N T. Attentive convolutional neural network based speech emotion recognition: a study on the impact of input features, signal length, and acted speech[J]. arXiv preprint arXiv: 1706.00612, 2017.

[15] 朱菊霞, 吴小培, 吕钊. 基于 SVM 的语音情感识别算法 [J]. 计算机系统应用, 2011, 20(5): 87–91.
Zhu Juxia, Wu Xiaopei, Lyu Zhao. SVM-based speech emotion recognition algorithm[J]. Computer System Application, 2011, 20(5): 87–91.

[16] Lin Y L, Wei G. Speech emotion recognition based on HMM and SVM[C]//Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on. IEEE, 2005, 8: 4898–4901.

[17] Pao T L, Chen Y, Yeh J H. Emotion recognition and evaluation from mandarin speech signals[J]. International Journal of Innovative Computing, Information and Control, 2008, 4(7): 1695–1709.

[18] Yüncü E, Hacihabiboglu H, Bozsahin C. Automatic speech emotion recognition using auditory models with binary decision tree and svm[C]//Pattern Recognition (ICPR), 2014 22nd International Conference on. IEEE, 2014: 773–778.

[19] Pan Y, Shen P, Shen L. Speech emotion recognition using support vector machine[J]. International Journal of Smart Home, 2012, 6(2): 101–108.

[20] Wöllmer M, Kaiser M, Eyben F, et al. LSTM-modeling of continuous emotions in an audiovisual affect recognition framework[J]. Image and Vision Computing, 2013, 31(2): 153–163.

[21] Huang G B. An insight into extreme learning machines: random neurons, random features and kernels[J]. Cognitive Computation, 2014, 6(3): 376–390.

[22] Han K, Yu D, Tashev I. Speech emotion recognition using deep neural network and extreme learning machine[C]//Fifteenth Annual Conference of the International Speech Communication Association, 2014: 223–227.

[23] Huang G B, Wang D H, Lan Y. Extreme learning machines: a survey[J]. International Journal of Machine Learning and Cybernetics, 2011, 2(2): 107–122.

[24] Livingstone S R, Peck K, Russo F A. Ravdess: the ryerson audio-visual database of emotional speech and song[C]//Annual meeting of the Canadian Society for Brain, Behaviour and Cognitive Science, 2012: 205–211.