运用高斯混合模型识别动物声音情绪*

刘 恒 吴 迪 苏家仪 杨春勇 侯 金
(中南民族大学电子信息工程学院智能无线通信湖北省重点实验室 武汉 430074)

摘 要:针对动物的情绪识别问题,提出高斯混合模型在动物声音情绪识别上的应用方法。利用语音信号处理与机器学习技术,提取动物声音信号的过零率、共振峰、梅尔-频率倒谱系数3种描述动物情绪的特征参数。采用高斯混合模型对采集到的动物声音信号训练样本进行聚类分析,计算测试样本后验概率,实现动物情绪的自动识别。通过分析特征参数的权重系数组合、高斯混合数目对识别率的影响来择选最优参数。实验结果表明,经参数优化后的高斯混合模型可将动物声音情绪的识别率由84.25%提高至96.67%。

关键词:动物情绪识别;高斯混合模型;权重系数;提取特征
中图分类号: TN912.34 B842.3 **文献标识码**:A **国家标准学科分类代码**: 520.20

Recognition of animal sound's emotion based on Gaussian mixture model

Liu Heng Wu Di Su Jiayi Yang Chunyong Hou Jin
(Hubei Key Laboratory of Intelligent Wireless Communications, College of Electronic Information Engineering, South-Central University for Nationalities, Wuhan 430074, China)

Abstract: For the problem of animal emotion recognition, this paper proposes an approach of applying Gaussian Mixture Model algorithm in animal sounds emotion recognition by combining speech processing and machine learning technique. The automatic recognition approach of animal emotion includes three key steps: three feature parameters extraction (Zero-crossing rate, Formant and Mel-Frequency Cepstral Coefficients), cluster analysis of training samples by using Gaussian Mixture Model, and computation of posterior probability of testing samples. Combination of feature weight coefficients and the number of Gaussian mixture components are analyzed to find the influence to the recognition rate. After that, choosing the optimal parameter, the experiment result shows that the Gaussian Mixture Model algorithm with optimal parameter effectively improves the recognition rate of animal emotion from 84.25% to 96.67%.

Keywords: animal emotion recognition; Gaussian mixture model; weight coefficients; feature extraction

1 引 言

随着动物行为学研究的发展,动物的情绪表达日益受到关注。动物通过声音和动作等特定行为表达情绪,其中,动物声音与人类语言类似,可实现同物种之间的交流。长期以来,人们对动物情绪的认知局限于长期的经验总结和直觉判断。如何实时有效地感知动物情绪成为新兴的研究方向。

近年来,人工智能领域的机器学习和机器翻译技术新进展使得人类识别动物声音情绪成为可能。GMM成为说话人识别中常用的模型[1]。但在情绪语音识别上,GMM的应用鲜有所见。美国南加州大学的LEE C C等人应用决策二叉树对人类的五种情绪进行分析[2],促进了机器学习在情绪语音识别上的应用。那不勒斯第二大学的ESPOSITO A等人在对情绪语音识别方法分析中提到利用GMM进行识别[3]。而在动物声音情绪的研究中GMM的应用更是鲜有所见。英国格拉斯哥大学的BELIN P团队对猫和猴子叫声进行分析,判识了积极和消极两种情绪[4],但情绪维度较少;匈牙利罗兰大学的MOLNAR C等人应用机器学习的方法对狗的6种不同行为相对应的叫声进行了分析[5],但识别率为64%左右。国内学者在人类语音方面,主要研究了说话人识别[6]和人类

收稿日期:2016-08
* 基金项目:国家自然科学基金项目(61002013)、国家林业局野生动植物保护与自然保护管理项目(BZY13002)资助

情感识别[7]，在动物声音方面也有基于动物叫声的物种识别[8]和个体辨认技术[9]的相关报道，但研究进展相对缓慢且识别率低。总体看，GMM在动物情绪语音识别上少有应用报道，基于单一特征的GMM在应用上识别率低。

本文研究一种GMM在动物声音情绪识别上的应用方法，以解决目前动物声音情绪识别存在的维度不足、识别率低的问题。通过提取动物声音信号特征参数[10]，为其典型情绪特征建立特定模型，并对情绪特征的权重系数组合[11]、高斯混合数目进行情绪辨识分析，以获得多维情绪特征与识别率的优化关系，从而提升识别率。

2 基本原理

动物声音情绪识别分为训练和测试两个阶段，工作流程如图1所示。在训练阶段，对动物声音情绪训练样本进行特征参数提取，建立高斯混合模型；在测试阶段，先对未知动物的声音测试样本进行预处理，再提取特征参数，最后计算特征向量在每个GMM的后验概率，最大后验概率对应的模型所代表的情绪即为识别结果。

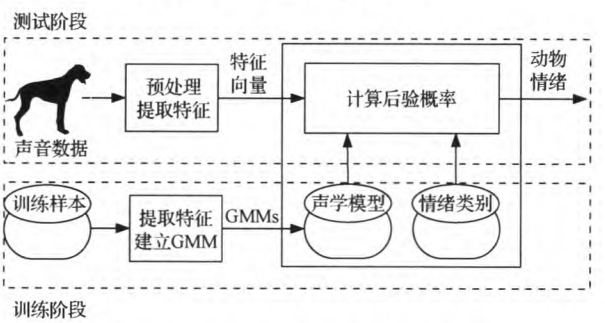


图1 动物声音情绪识别基本原理

3 高斯混合模型

3.1 模型描述

高斯混合模型是一种常用的概率模型，其本质上是一种多维概率密度函数，可用于对若干个高斯概率密度函数进行加权和，以逼近任意分布函数，因此可用来描述各种形式的语音特征参数的统计分布。

将GMM作为动物声音情绪模型，则每一种情绪对应一个GMM，如：

$$P(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i G_i(\vec{x}) \tag{1}$$

式中：M为高斯混合数目，$\vec{x}$为动物声音的D维观察向量，维度D取决于选取的特征参数，$p_i(i=1,\ldots,M)$为高斯成分的混合权重，满足$\sum_{i=1}^{M} p_i = 1$，$G_i(\vec{x})$，$i=1,\ldots,M$为高斯成分密度，每个成分密度是一个D维变量高斯函数的形式，则：

$$G_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \cdot \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\vec{x}-\vec{\mu}_i)\right\} \tag{2}$$

式中：$\vec{\mu}_i$为第i个高斯分布的D维均值向量，$\boldsymbol{\Sigma}_i$为第i个高斯分布的$D\times D$对角协方差矩阵。

M个单高斯模型(single gaussian model，SGM)的混合权重p、均值向量$\vec{\mu}$与协方差矩阵$\boldsymbol{\Sigma}$共同组成一个GMM，则SGM参数集合$\boldsymbol{\lambda}$可写为：

$$\boldsymbol{\lambda} = \{p_i, \vec{\mu}_i, \boldsymbol{\Sigma}_i\} \quad i=1,\ldots,M \tag{3}$$

式中，$\boldsymbol{\lambda}$对应动物某一种声音的情绪状态，则该情绪状态的M维混合结构[12]如图2所示。

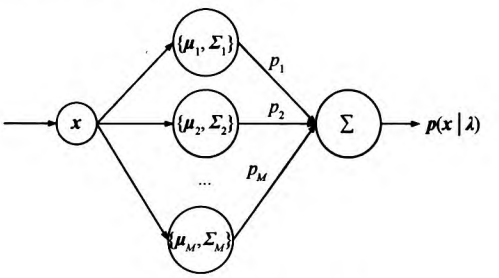


图2 GMM混合结构

3.2 模型训练

给定动物声音情绪的训练样本，进行情绪模型的训练估计可得到GMM的参数集合$\boldsymbol{\lambda}$。当参数集合$\boldsymbol{\lambda}$在一定程度上对训练样本的特征向量分布达到最佳匹配时，最能表征训练样本的情绪信息。有效估计$\boldsymbol{\lambda}$的方法有若干种，其中最大似然估计(maximum likelihood estimation，MLE)是最常用的方法，MLE的目标是在给定训练数据的条件下找到使GMM似然函数最大的模型参数集合$\boldsymbol{\lambda}$。

假设某一种动物声音情绪模型的训练向量序列为$\boldsymbol{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}$，则GMM似然函数为：

$$P(\boldsymbol{X} \mid \boldsymbol{\lambda}) = \prod_{t=1}^{T} p(\vec{x}_t \mid \boldsymbol{\lambda}) \tag{4}$$

式中：$P(\cdot)$为概率函数。由于$P(\boldsymbol{X}|\boldsymbol{\lambda})$是关于参数$\boldsymbol{\lambda}$的一个非线性函数，难以计算出精确的最大值，因此直接计算MLE是极其困难的。根据文献[13]，可用期望最大化(expectation maximization，EM)迭代估计最大似然参数。

4 特征提取

4.1 特征分析

由于动物声音数据中存在大量冗余信息，为了减少数据量和提升计算速度，需要对动物声音数据提取特征参数，这实际上是一个降维的过程，是动物声音情绪识别的关键步骤。接下来，根据文献[14]以狗声音为例，用Praat软件，依次分别从时域、频域、倒谱域3个角度讨论能够准确描述相应情绪的声学特征参数。

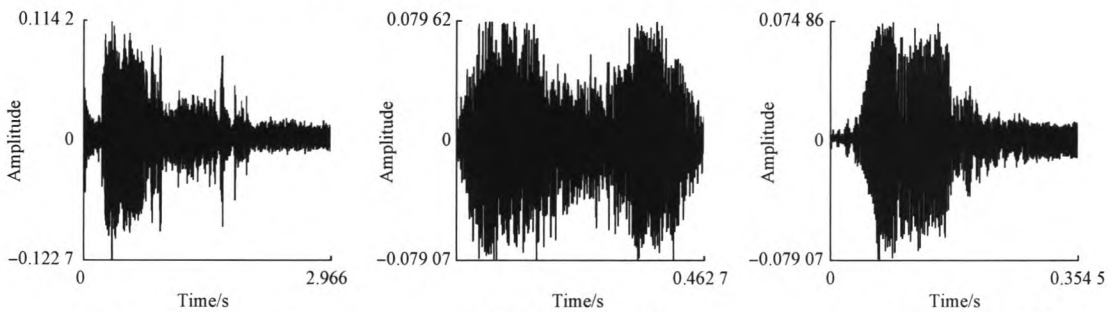
在时域上，一般是利用过零率(zero-crossing rate，ZCR)描述动物声音波动的激烈程度。过零率体现动物声音信号中相邻两个采样点幅度符号反转的比率。因不同的声音情绪下波动激烈程度对应着不同的ZCR，故可提取ZCR特征来识别动物情绪。ZCR定义为：

$$R=\frac{1}{T-1}\sum_{t=1}^{T-1}I\{s_t s_{t-1}<0\} \tag{5}$$

式中：R为过零率，T为采样点数，s_t为动物声音信号中的第t个样本点，$I\{\cdot\}$为指示函数，若此括号内的表达式逻辑为真，则指示函数结果为1，否则为0。

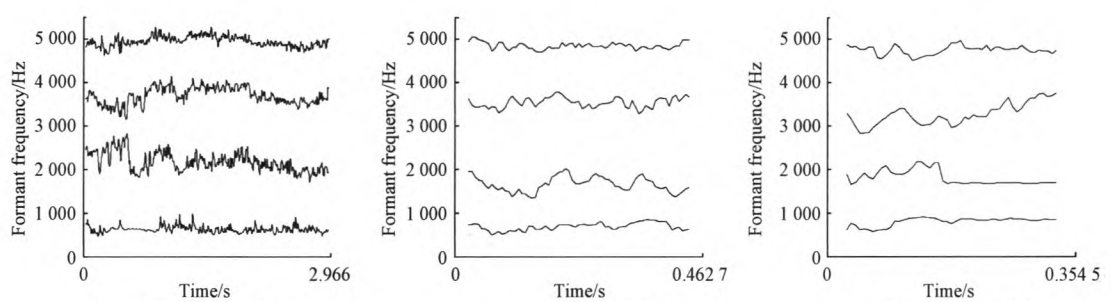
如图3所示为狗3种情绪“悲伤”、“愤怒”、“高兴”的声音信号时域波形。按式(5)计算获得这3种情绪声音信号的过零率，结果依次对应为10.44%、7.41%、13.17%。通过对10组实验数据进行统计，可得动物3种情绪声音信号的ZCR，其表现出$R_{悲伤}>R_{愤怒}>R_{高兴}$这一规律，因此验证了ZCR可作为动物声音情绪识别特征的设想。

在频域上，利用共振峰(Formant)[16]来描述动物通过声音表达不同情绪的声道(共振腔)位置。因动物发出的不同声音对应着声音信号中不同形态的共振峰，故可提取共振峰特征来识别动物情绪，其中，第一共振峰(定义为F1)、第二共振峰(定义为F2)常被作为发声特征的考察对象[17]。如图4所示为狗声音信号中的“悲伤”、“愤怒”和“高兴”3种情绪共振峰。



(a) 为“悲伤”声音时域波形　(b) 为“愤怒”声音时域波形　(c) 为“高兴”声音时域波形

图3　狗3种情绪“悲伤”、“愤怒”、“高兴”的声音信号时域波形



(a) 为“悲伤”共振峰信号　(b) 为“愤怒”共振峰信号　(c) 为“高兴”共振峰信号

图4　狗声音信号中“悲伤”、“愤怒”和“高兴”3种情绪共振峰

通过对10组实验所获得的狗叫声样本进行计算统计，F1、F2分布情况如表1所示。

表1　狗叫声情绪共振峰分布

情绪	样本数/个	F1/Hz	F2/Hz
高兴	10	513～517	1 024～1 028
悲伤	10	848～851	2 110～2 115
愤怒	10	691～695	1 308～1 310

从表1中可见，每种情绪的共振峰差异特征明显，证明其共振峰可作为动物声音情绪识别特征。

在倒谱域上，从人耳听觉感知的角度利用梅尔-频率倒谱系数(mel-frequency cepstral coefficients，MFCC)作为有效的特征参数[18]。MFCC特征提取流程如图5所示。

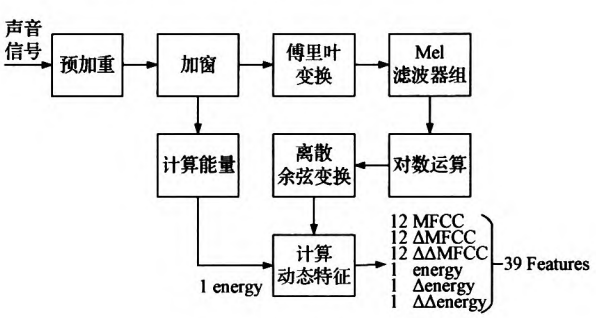
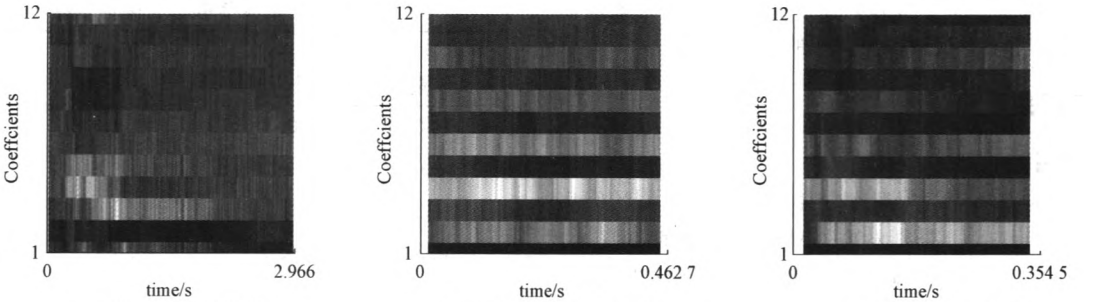


图5　MFCC特征提取流程

如图5所示，最终提取到的MFCC共有39维特征，其中包括36维MFCC特征(12维MFCC原始特征，12维一阶MFCC动态特征，12维二阶MFCC动态特征)，以及3维能量特征(1维原始能量特征，1维一阶能量动态特征，1维二阶能量动态特征)。MFCC通过听觉感知、能量指示以及相应的动态特性描述动物情绪的变化过程。本文通过如图5所示流程对“悲伤”、“愤怒”和“高兴”3种情绪提取出39维特征中的前12维MFCC原始特征，如图6所示。



(a) 为"悲伤"MFCC灰度图　(b) 为"愤怒"MFCC灰度图　(c) 为"高兴"MFCC灰度图

图 6　狗声音信号中"悲伤"、"愤怒"和"高兴"3 种情绪 MFCC 灰度图

图 6 中的纵坐标 Coefficients 表示 MFCC 特征值的维数；灰度值表示 MFCC 的大小，如灰度越深表示参数值越大。由图 6 可见，3 种情绪的灰度值在每一维度特征分布上差异特征均较明显，验证了 MFCC 可作为动物声音情绪识别特征。

4.2　特征加权

通过前文的实验结果分析，可以认为以上过零率、共振峰和 MFCC 3 种特征均可用于描述动物声音的情绪。为了弥补单一特征描述在实际环境的应用中存在的识别率较低的缺陷，本文提出 ZCR、共振峰和 MFCC 的三特征加权描述法，即将这 3 种声音信号的情绪特征按照某种权重组合，通过反复迭代识别率测试来确定最优加权参数。具体优化过程如图 7 所示。

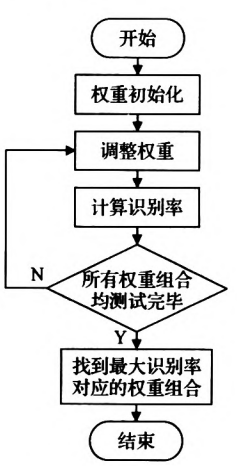
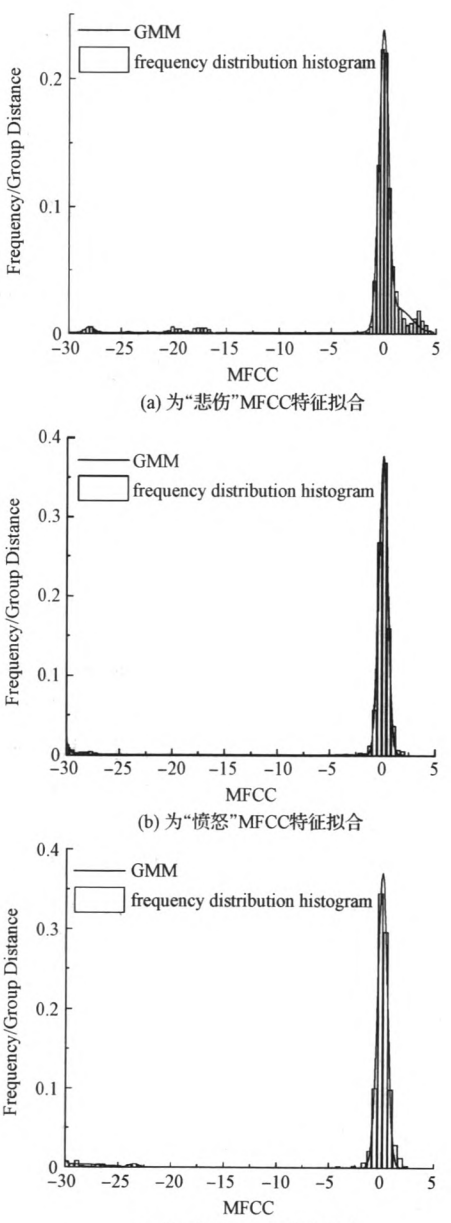


图 7　特征加权优化过程

4.3　模型训练

利用 GMM 可对 MFCC 特征参数的统计特性（频率分布直方图）进行拟合，"高兴""悲伤""愤怒"3 种情绪的拟合效果依次分别如图 8 所示。

图 8 纵轴表示频率-组距比，横轴表示 MFCC 的大小。图中 frequency distribution histogram 为 MFCC 频率分布直方图，体现了 MFCC 的分布特征。曲线 GMM 为 GMM 对 MFCC 的拟合结果。可以看出，GMM 拟合曲线基本包括频率直方图中所有频率分量（即为图中的柱状图），因而



(a) 为"悲伤"MFCC特征拟合
(b) 为"愤怒"MFCC特征拟合
(c) 为"高兴"MFCC特征拟合

图 8　狗声音信号中"悲伤"、"愤怒"、"高兴"3 种情绪 MFCC 特征拟合

图中所示拟合图即是最终拟合结果。由以上各图可知，GMM 可对不同的情绪特征进行拟合。当拟合曲线基本

涵盖频率分布直方图中所有频率分量时即可定为最优拟合结果，这可为情绪识别提供相应的统计模型。

5 情绪识别

5.1 最大后验概率

假设已训练的声音模型库中共有 τ 种情绪，则情绪集合 $T=\{1,2,\ldots,\tau\}$ 分别对应 GMM 模型中的参数 λ_1，$\lambda_2,\ldots,\lambda_\tau$。下面的工作即对测试样本进行情绪识别，分两步完成，第一步是提取测试样本的特征参数，第二步是计算特征参数在情绪模型库中每个模型下的后验概率，最大后验概率所对应的模型所代表的情绪即为识别结果。根据贝叶斯定理，可利用下式求得最大后验概率所对应的模型代表的情绪：

$$\hat{T}=\arg\max_{1\leqslant k\leqslant\tau}\Pr(\lambda_k\mid X)=\arg\max_{1\leqslant k\leqslant\tau}\frac{p(X\mid\lambda_k)\Pr(\lambda_k)}{p(X)} \tag{6}$$

式中：$\arg\max_{1\leqslant k\leqslant\tau}(\cdot)$ 表示使圆括号内表达式取得最大值时对应的变量 τ，即找到最有可能的模型所代表的情绪。$\Pr(\lambda_k)=1/\tau$ 为常数，$p(X)$ 表示动物声音提取特征序列 X（即训练样本的特征参数）的可能性，即 $p(X)=1$，因而这两项的数值对于每一种情绪模型均相等，故可省略。另根据观察向量（即测试样本的特征参数）之间的独立性，式(6)可变换为：

$$\hat{T}=\arg\max_{1\leqslant k\leqslant\tau}\sum_{t=1}^{T}\log p(\vec{x_t}\mid\lambda_k) \tag{7}$$

值得说明的是，此处使用对数概率可防止数值计算中出现的数值下溢。

5.2 识别实验

动物声音情绪识别实验在实验室环境下进行，由于现实动物情绪声音数据的缺乏，训练与测试样本均来自狗的录音。实验中，共采用 3 组训练样本进行模型的训练，对应 3 种情绪，每种情绪有 10 个训练数据；测试样本共 3 组，每组 10 个测试数据。

目前已知的动物声音的情绪特征分析通常是针对多个独立特征，因此本文提出将不同特征权重系数组合的方法，对比分析不同权重系数组合与识别率的关系，如表 2 所示。

从表 2 可知，未加权的系数组合{0.0,1.0,0.0}、{1.0,0.0,0.0}、{0.0,0.0,1.0}对应的识别率并非最优识别率。当权重系数组合为{0.2,0.3,0.5}时达到最高识别率，而且各特征权重的重要程度满足：MFCC＞共振峰＞过零率，因此不同权重系数组合一定程度上影响识别率，最优权重系数组合为{0.2,0.3,0.5}。

进一步，在保持{0.2,0.3,0.5}权重系数组合不变的条件下，通过改变 GMM 的混合个数 M，可得到 GMM 数目 M 与识别率的关系，如图 9 所示。图中，T-GMM 表示对从狗叫声中所提取的多情绪特征进行组合应用 GMM，S-GMM 表示对从狗叫声中所提取的单一情绪特征应用 GMM。

表 2 权重系数组合与识别率

过零率权重	共振峰权重	MFCC权重	总识别率（%）
0.0	0.1	0.9	93.33
…	…	…	…
0.2	0.2	0.6	90.00
0.2	0.3	0.5	96.67
0.2	0.4	0.4	93.33
…	…	…	…
0.0	1.0	0.0	80.00
…	…	…	…
1.0	0.0	0.0	76.67
0.0	0.0	1.0	86.67

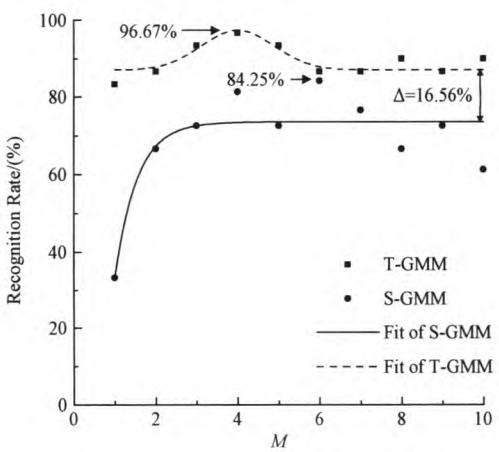


图 9 高斯混合数目与识别率的关系

由图 9 可见，相较于对单一特征 MFCC 应用 GMM（即 S-GMM），本文提出的多特征组合应用 GMM（即 T-GMM）在动物声音情绪识别上效果更佳，其最佳识别率可由 84.25%提高至 96.67%，平均识别率提升相较 S-GMM 可达 $\Delta=16.56\%$。同时还可发现，GMM 数目影响识别率。图中随着 M 的变化，因 GMM 的拟合过程中存在欠拟合（under-fitting）和过拟合（over-fitting），从而导致了识别率波动。但从趋势来看，随着 GMM 的混合个数 M 增加，T-GMM 与 S-GMM 的识别率趋于分别稳定在 89.33%和 72.77%左右。

6 结 论

动物声音情绪识别技术在诸多领域具有较重要的应用。以狗叫声为原型，针对“高兴”、“悲伤”、“愤怒”3 种情绪，利用过零率、共振峰及 MFCC3 种特征进行加权，基于高斯混合模型实现了动物声音情绪识别，且通过对这三种特征权重系数组合、高斯混合数目与识别率关系的分析，进一步提高了系统的识别率，最高可达到 96.67%。目前，国内外对动物情绪研究较少，声音数据资源不完善，样

本数目有限，且至今尚未发现能100%准确描述动物情绪的特征，故针对现实中的动物声音实验、样本库完善及特征分析方面仍需进一步开展研究。本文的研究结果可为动物声音情绪识别在动物行为学研究、宠物叫声翻译、动物园应急报警、珍惜动物生态保护等领域提供应用上的理论参考。

参 考 文 献

[1] STURIM D E, CAMPBELL W M, REYNOLDS D A. Classification methods for speaker recognition[J]. Speaker Classification I Fundamentals Features & Methods Springer Berlin, 2007(4343):278-297.

[2] LEE C C, MOWER E, BUSSO C, et al. Emotion recognition using a hierarchical binary decision tree approach [J]. Speech Communication, 2011, 53(910):1162-1171.

[3] ESPOSITO A, ESPOSITO A M. On the recognition of emotional vocal expressions: motivations for a holistic approach[J]. Cognitive Processing, 2012, 2(增刊2):541-50.

[4] BELIN P, FECTEAU S, CHAREST I, et al. Human cerebral response to animal affective vocalizations[J]. Proceedings of the Royal Society B: Biological Sciences, 2008, 275(1634): 473-481.

[5] MOLNáR C, KAPLAN F, ROY P, et al. Classification of dog barks: a machine learning approach[J]. Animal Cognition, 2008, 11(3): 389-400.

[6] 王明合，唐振民，张二华. 基于i-vector局部加权线性判别分析的说话人识别[J]. 仪器仪表学报，2015，36(12):2842-2848.

[7] 黄程韦，赵艳，金赟，等. 实用语音情感的特征分析与识别的研究[J]. 电子与信息学报，2011，33(1):112-116.

[8] 王岩. 基于动物叫声的物种识别技术的研究[D]. 哈尔滨：东北林业大学，2008.

[9] 李娜. 基于动物鸣声的物种识别研究[D]. 北京：中国科学院大学，2013.

[10] 韩文静，李海峰，阮华斌，等. 语音情感识别研究进展综述[J]. 软件学报，2014，25(1):37-50.

[11] 庞程，王秀玲，张结，等. 基于多特征融合的GMM汉语普通话口音识别[J]. 华中科技大学学报：自然科学版，2015(增刊1): 381- 384, 388.

[12] DONG Y, LI D. Gaussian mixture models[M]. Automatic Speech Recognition, 2015:13-21.

[13] JEZIERSKA A, CHAUX C, PESQUET J C, et al. An EM approach for time-variant poisson-gaussian model parameter estimation[J]. IEEE Transactions on Signal Processing, 2014, 62(1):17-30.

[14] YEO C Y, AL-HADDAD S A R, NG C K. Animal voice recognition for identification (ID) detection system[C]. IEEE 7th International Colloquium on Signal Processing and its Applications (CSPA), 2011: 198-201.

[15] BACHU R G, KOPPARTHI S, ADAPA B, et al. Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy[M]. Springer Netherlands Advanced Techniques in Computing Sciences and Software Engineering, 2010: 279-282.

[16] RABINER L R, SCHAFER R W. Digital Speech Processing[J]. The Froehlich/Kent Encyclopedia of Telecommunications, 2011(6):237-258.

[17] FRANCO-PEDROSO J, GONZALEZ-RODRIGUEZ J. Linguistically-constrained formant-based i-vectors for automatic speaker recognition[J]. Speech Communication, 2015(76):61-81.

[18] BORDE P, VARPE A, MANZA R, et al. Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition[J]. International Journal of Speech Technology, 2015, 18(2): 167-175.

作 者 简 介

刘恒，现为中南民族大学本科生，主要研究方向为语音识别与语音分离。
E-mail:lhmachine@foxmail.com

吴迪，2015年于文华学院获得学士学位，现为中南民族大学硕士生，主要研究方向为语音识别与语音分离增强。
E-mail:478208146@qq.com

苏家仪，2015年于中南民族大学获得学士学位，主要研究方向为语音识别。
E-mail:joey_su@qq.com

杨春勇(通讯作者)，1998年于华中师范大学获得学士学位，2001年于中国地震局地震研究所获得硕士学位，2005年于华中科技大学获得博士学位，现为中南民族大学教授，主要研究方向为光传感与光通信技术。
E-mail:cyyang@mail.scuec.edu.cn

侯金，2003年于哈尔滨理工大学获得学士学位，2006年于哈尔滨工业大学获得硕士学位，2011年于华中科技大学获得博士学位，现为中南民族大学副教授，主要研究方向为光通信与光器件。
E-mail:houjin@mail.scuec.edu.cn