# 基于调制频谱特征的自动语音情感识别*

张鼎天，徐明星
普适计算教育部重点实验室 清华信息科学与技术国家实验室（筹）
清华大学 计算机科学与技术系，北京 100084

**文　摘：** 本文采用调制频谱特征来自动识别人的语音中的情感信息。受人耳听觉系统启发，语音信号通过听觉滤波器组以及调制滤波器组得到长时频域-时域表示，从而获得声学频率和时域调制频率的信息，进而提取出调制频谱特征。通过将该特征在演员表演的德语 Berlin 语音库和采集自真实生活的中文情感语音库上进行语音情感分类实验，发现该特征与传统的短时频谱特征，如梅尔频率倒谱系数和感知线性预测系数相比，具有良好的性能和应用前景。

**关键词：**情感识别；语音调频；频域-时域表示；情感计算；语音分析
**中图分类号：**TP 391

情感计算是目前活跃的跨学科研究领域。该领域中语音情感识别(SER)的目标是从说话人的语音信号识别潜在的情感状态。识别的结果可以广泛应用到包括人机交互等各个方面。

频谱特征（包括倒谱特征）在 SER 中发挥了显著的作用。它们传达语音信号的频率含量，并为韵律特征提供补充信息。然而，设计更有效的情感识别频谱特征的工作还嫌不足。传统频谱特征，例如著名的梅尔频率倒谱系数(MFCC)，只考虑信号的短时频域属性，而忽略了重要的长时演化趋势。这种局限性会进一步影响 SER 的性能。另一方面，神经科学的研究成果显示哺乳动物听觉皮层的频域时域(ST)感受野可以长达数百毫秒，并对时间-频率域的调制产生反应。语音调制频谱的重要性也在很多领域得以验证，包括听觉生理，心理声学，语音感知，信号分析与合成。

因此，人们提出了情感识别的长时的调制频谱特征(MSFs)[1]。对语音信号的多个声学频率窗进行时域取包络操作，然后进行频率分析得到的特征便同时包含频域和时域的属性。这些特征被应用于对已标注分类好的语音信号进行情感分类，实验结果显示其性能优于传统特征。

由于该特征仅被用于德语的 Berlin 情感语音库上，而且该语音库为演员表演的结果，并非真实情感语音。因此，我们将该特征用于采集自真实生活的中文情感语音库上，通过对比试验来验证其在不同语种和真实情感语音上的性能。

本文的组织结构如下：第一部分介绍提取语音信号长时 ST 表示的算法；第二部分详细介绍本文采用的 MSFs 以及用于对比的短时频谱特征；第三部分介绍使用的数据库；第四部分分析讨论实验结果；第五部分为总结。

## 1　语音信号的 ST 表示

受动物的听觉系统启发，语音信号的频域时域(ST)表示可用图 1 所示方法提取。首先进行预处理，将语音信号以 8kHz 重采样，并用 P.56 语音电压计将其活跃语音强度归一化为-26dBov。由于情感信息可通过频带有限的电话会话即可可靠传达，我们认为 8kHz 的采样频率即可胜任 SER。不含重叠部分的语音帧被 G.729 语音活跃检测算法鉴别为活跃或非活跃，只有活跃的语音帧予以保留。

预处理后的语音信号 $s(n)$ 进行加窗，采用 256ms 窗长、64ms 窗移的汉明窗，得到 $s_k(n)$，$k$ 代表帧号。相对较长的窗长对于最低滤波器中心频率为 4kHz 的低调制频率而言是必须的。

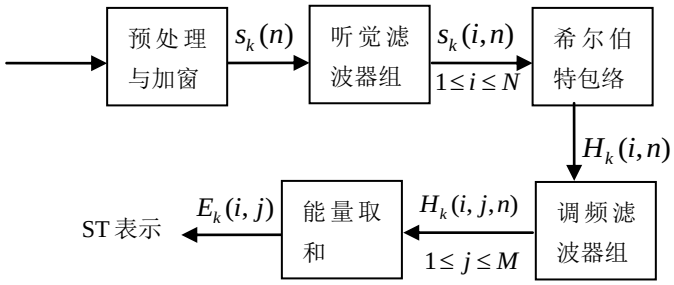


图 1　ST 表示的提取流程图

我们都知道人类听觉系统可以用一系列相交的带通频率滤波器组模型来表示。这些滤波器中心频率依次上升。第 $i$ 个滤波器在第 $k$ 帧的输出信号为：

*基金项目：本项目工作受到国家自然科学基金面上项目（61171116）支持。

$$s_k(i,n) = s_k(n) * h(i,n) \tag{1}$$

其中 $h(i,n)$ 表示第 $i$ 个通道的冲击响应。这里我们采用包含 $N$ 个子带滤波器的 gammatone 滤波器组。这些滤波器的中心频率与带宽成正比，可以用以下式子进行描述：

$$ERB_i = \frac{F_i}{Q_{ear}} + B_{\min} \tag{2}$$

其中 $F_i$ 表示第 $i$ 个滤波器的中心频率，$Q_{ear}$ 和 $B_{\min}$ 为常数 9.26449 和 24.7。我们采用 19 个滤波器的 gammatone 滤波器组，第一个和最后一个滤波器的中心分别为 125Hz 和 3.5kHz，带宽 38 和 400Hz。该滤波器组的幅值响应如图 2 所示。

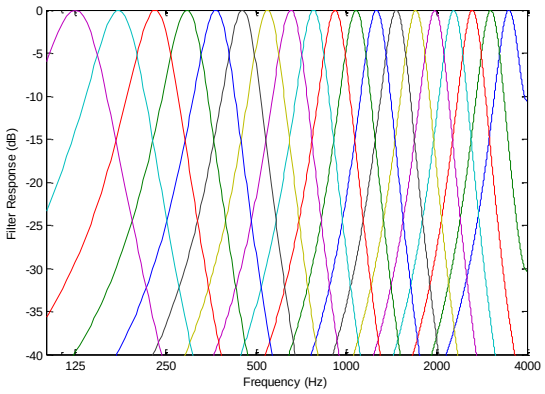


图 2　听觉滤波器组的幅值响应

由 $s_k(i,n)$ 计算希尔伯特包络，取 $\hat{s}_k(i,n) = s_k(i,n) + jH\{s_k(i,n)\}$ 的幅值，其中 $H\{\cdot\}$ 代表希尔伯特变换。因此：

$$H_k(i,n) = \left|\hat{s}_k(i,n)\right| = \sqrt{\hat{s}_k^2(i,n) + H^2\{\hat{s}_k(i,n)\}} \tag{3}$$

滤波器组的听觉频谱分解只是人类听觉系统处理声音信号的第一阶段模型。处理之后的输出进一步在听觉皮层中提取频域-时域调制模式。在这里，一个 $M$ 带调制滤波器被用来模仿这种功能，它被应用在 $H_k(i,n)$ 上产生输出 $M$ 个输出 $H_k(i,j,n)$，$1 \le j \le M$。这里的调制滤波器为 2 阶带通滤波器，品质因子为 2。本文我们令 $M=5$，滤波器中心频率在对数坐标上等距分布从 4 至 64Hz。其幅值响应见图 3。

最后，第 $k$ 帧的 ST 表示 $E_k(i,j)$ 由计算 $H_k(i,j,n)$ 的能量得到：

$$E_k(i,j) = \sum_{n=1}^{L} \left|H_k(i,j,n)\right|^2 \tag{4}$$

其中 $1 \le k \le T$，$L$ 和 $T$ 表示一帧中的样本数和帧总数。通过结合听觉滤波器组和调频滤波器组，我们能够得到一个二维频率矢量以便在不同声学频率带上分析调制频率内容。

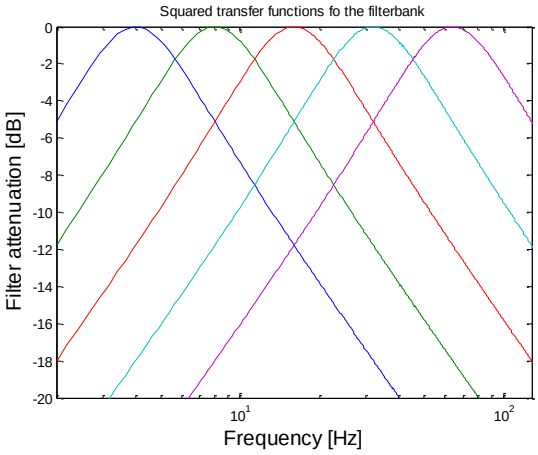


图 3　调制滤波器组的幅值响应

## 2　特征提取

在这一部分，我们详细阐述从 ST 表示中如何提取 MSF 和短时频谱特征。

### 2.1　调制频谱特征

有两种 MSF，分别通过 ST 表示的频域测量和线性预测来计算得到。对于每一帧 $k$，ST 表示 $E_k(i,j)$ 首先归一化为单位能量值。六种频域测量值 $\Phi_1 \sim \Phi_6$ 随后在每一帧上计算得出。对于第 $k$ 帧，$\Phi_{1,k}(j)$ 被定义为第 $j$ 个调频带上能量的均值（$1 \le j \le 5$）：

$$\Phi_{1,k}(j) = \frac{\sum_{i=1}^{N} E_k(i,j)}{N} \tag{5}$$

$\Phi_1$ 描绘了沿调制频谱上语音能量的分布情况。第二种测量方式是频谱平缓度，即频域能量测量值的几何平均与算数平均的比值：

$$\Phi_{2,k}(j) = \frac{\sqrt[N]{\prod_{i=1}^{N} E_k(i,j)}}{\Phi_{1,k}(j)} \tag{6}$$

该值接近于 1 说明频谱平缓，接近 0 说明频谱较为分散。第三种测量手段是频谱重心：

$$\Phi_{3,k}(j) = \frac{\sum_{i=1}^{N} f(i) E_k(i,j)}{\sum_{i=1}^{N} E_k(i,j)} \tag{7}$$

这里令 $f(i) = i$。由于临近调制带会有很大的相关性，这里的 $\Phi_{2,k}(j)$ 和 $\Phi_{3,k}(j)$ 取 $j \in \{1,3,5\}$ 以减少信息冗余度。

以上三种特征测量的都是单个调制带的频域信息，此外还需测量不同调制带间的关系。首先，19 个声学带被分为四部分：1-4，5-10，11-15，16-19，即 $D_l(1 \le l \le 4)$，粗略涵盖了频率<300，300-1000，1000-2000，和>2000 四个部分。同一部分下的能量求和：$\tilde{E}_k(l,j) = \sum_{i \in D_l} E_k(i,j)$。于是调制频域重心：

$$\Phi_{4,k}(l) = \frac{\sum_{j=1}^{M} j\tilde{E}_k(l,j)}{\sum_{j=1}^{M} \tilde{E}_k(l,j)} \tag{8}$$

$\Phi_{5,k}(j)$ 和 $\Phi_{6,k}(j)$ 为对 $\tilde{E}_k(l,j)$，$1 \le j \le M$ 进行一阶多项式拟合的线性回归系数（斜率）和相应的回归误差（方均根误差，RMSE）。这三种特征体现的是声学频率区域的变化速率，很好地捕捉了时域动态信息。

除频域测量外，线性预测分析（LPA）也被应用到第 1、3、5 个调制带上以抽取第二类特征并减少冗余信息。这里使用了自回归（AR）模型的自相关方法，采用 5 阶全极点模型，计算线性预测倒谱系数（LPCC）$C_k(n,j)(0 \le n \le 5)$。综上，每一帧可计算一共 41 个 MSF 特征。

通常的 SER 文献都会用统计的手段将帧级别（FL）的特征提取为发音级别（UL）的特征，从而体现全局的属性，消除与说话内容的关联。这里我们计算 FL 的 MSF 的平均值和标准偏差，产生 82 个 UL 的 MSF。

### 2.2 MFCC 特征

梅尔频率倒谱系数（MFCC）是当下普遍采用的短时频谱特征。我们用 MFCC 来做对比试验。语音信号经过预加重系数 0.97 的高通滤波器，从窗长 25ms、窗移 10ms 的汉明窗中提取前 13 个 MFCC（包括零阶系数），计算其 delta 值以及双重 delta 值，得到 39 维 FL 特征向量。计算这些向量的平均值、标准偏差、第 3~5 阶中心距，我们得到 195 个 MFCC 特征。

### 2.3 PLP 特征

除 MFCC 以外，短时频域特征中的感知线性预测（PLP）也被提取出来作为对照。这里使用一个 5 阶全极点，提取出的 PLP 系数被转换为倒谱系数 $c(n)(0 \le n \le 5)$。同 MFCC 一样，计算它自身、delta 值以及双重 delta 值的平均值、标准偏差、第 3~5 阶中心距，我们得到 90 个 PLP 特征。

## 3 情感语音数据

### 3.1 Berlin 情感语音数据库

Berlin 情感语音数据库是公开的用于情感识别的最流行的数据库之一。10 个播音员（5 男 5 女）用德语说出 10 个的日常句子（五短五长，典型的在 1.5s 和 4s 间），可归类为 7 种情感。原始数据大约有 800 句。通过 20 名听众的主观感受测试，情感识别正确率高于 80%，超过 60%的听众认为自然的话语被包含在最终的数据库。Berlin 情感语音数据库包含 7 种情感，其数目分别为：愤怒（127），无聊（81）厌恶（46），恐惧（69），快乐（71），中性（79），和悲伤（62）。

### 3.2 中文情感语音数据库

为了验证该特征在不同语种和不同情感分类上的性能，我们从网上采集了流媒体（主要是视频）中的中文情感语音数据，其中包括采访、播音等，绝大部分是自然状态下的情感流露。经过进一步筛选和主观感受测试，选取 5 种情感：厌恶、快乐、中性、愤怒、悲伤，每种情感男、女声各 50 句，一共 500 句，大部分在 2s~6s 之间。

## 4 实验

我们采用支持向量机（SVM）进行情感识别。SVM 的内核采用径向基函数（RBF），RBF 的好处在于它可以很好地处理特性和目标间的非线性关系，含有较少的超参数，数值计算的难度较小。

我们采用 10 折交叉验证。我们将说话人混合，从中随机分成 10 个不相交的子集，将其中 9 组用于训练，剩下 1 组用于测试。

### 4.1 特征选取

使用所有的特征用于机器学习会由于维数灾难导致识别性能的下降。这里我们采用两步特征选取方法。第一步计算特征的 Fisher 区分度（FDR）并排序，去除不相关的“噪声”特征。FDR 定义为：

$$FDR(u) = \frac{2}{C(C-1)} \sum_{c_1} \sum_{c_2} \frac{(\mu_{c_1,u} - \mu_{c_2,u})^2}{\sigma^2_{c_1,u} + \sigma^2_{c_2,u}} \tag{9}$$

其中 $1 \le c_1 < c_2 \le C$，$\mu_{c_1,u}$ 和 $\sigma^2_{c_1,u}$ 分别为第 $u$ 个特征在第 $c_1$ 类的中的均值和方差，通过实验，将阈值设定为 0.15，去除低于此值者，可以保留那些类间区分度高而类内区分度低的特征，这一步可以减少大约 10%的 MSF 和 40%的 MFCC 和 PLP 特征。

在第二阶段，通过有名的多类线性判别分析（LDA）可以进一步提高类间区分度，减小类内区分度。LDA 解下式求特征值：

$$S_b w = \lambda S_w w \tag{10}$$

其中 $S_b$ 和 $S_w$ 为类间和类内的散度矩阵。本征矢量 $w$ 用以建立变换矩阵 $W$，将 $x$ 转为 $y = W^T x$。由于 $C$ 类划分问题 $S_b$ 的最大秩为 $C-1$，LDA 特征

表 1 使用 MSF，MFCC 和 PLP 特征的识别效果（Berlin）

| 特征 | 识别率（%） | | | | | | | 平均 |
|---|---|---|---|---|---|---|---|---|
| | 愤怒 | 无聊 | 厌恶 | 恐惧 | 快乐 | 中性 | 悲伤 | |
| MSF | **90.8** | **84.6** | 78.1 | 72.2 | **60.5** | **82.9** | **86.7** | **79.4** |
| MFCC | 83.3 | 82.6 | **78.8** | **75.6** | 51.9 | 78.8 | 82.7 | 76.2 |
| PLP | 83.2 | 72.4 | 55.3 | 51.9 | 47.8 | 75.6 | 79.8 | 66.6 |

表 2 使用 MSF，MFCC 和 PLP 特征的识别效果（中文）

| 特征 | 识别率（%） | | | | | 平均 |
|---|---|---|---|---|---|---|
| | 厌恶 | 快乐 | 中性 | 气愤 | 悲伤 | |
| MSF | 76.3 | **64.6** | 70.6 | **77.3** | **82.6** | **74.3** |
| MFCC | **79.9** | 59.9 | **73.2** | 76.7 | 76.7 | 73.3 |
| PLP | 66.3 | 48.7 | 54.7 | 68.8 | 59.6 | 59.6 |

的最大数目也是 $C-1$。

### 4.2 实验结果

从表 1 中可以看出，MSF 应用在在 Berlin 语音库大多数情感识别中取得了最好的准确度，平均识别率胜过了 MFCC 和 PLP。

表 2 是 MSF 在中文语音库中的识别效果。可以看出，MSF 依然超过了传统的短时频谱特征。由于从网上获取的语音片段录音质量参差不齐，情感划分方法有待改进等原因，整体识别率不如 Berlin 语音库的好，但也因此显示出 MSF 对于不同语种的普适性。

## 5 结论

本文将 MSF 用于不同的语种和情感分类上的情感识别。使用听觉滤波器组和调频滤波器组得到 ST 表示，从中用频域测量和线性预测提取本文所述的特征，在演员表演的 Berlin 语音库中实验来区分愤怒，无聊，厌恶等 7 种情感，在情感真实流露的中文语音库中区分厌恶、中性、悲伤等 5 种情感。通过和短时频谱特征 MFCC 和 PLP 的对比可以看出，MSF 具有更好的性能；同时，MSF 具有普适性，在不同的语种和真实情感语音上都显示出优越的性能。

## 参 考 文 献

[1] Siqing Wu, Tiago H. Falk, Wai-Yip Chan. Automatic speech emotion recognition using modulation spectral features [J]. Speech Communication, 2011. 768-785.

[2] Young S, Evermann G, Gales M , et al . The HTK Book(HTK version 3.4) [M] . Cambridge: Cambridge University, 2006.

[3] Zissman M A. Comparison of four approaches to automatic language identification of telephone speech [J]. IEEE Transactions on Speech and Audio Processing , 1996, 4(1) : 31 - 44.

# Automatic speech emotion recognition based on modulation spectral features

Zhang Dingtian，Xu Mingxing

Key Laboratory of Pervasive Computing, Ministry of Education

Tsinghua National Laboratory for Information Science and Technology (TNList）

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

**Abstract:** In this study, modulation spectral features (MSFs) are used for the automatic recognition of human affective information from speech. The features are extracted from an auditory-inspired long-term spectro-temporal representation. Obtained using an auditory filterbank and a modulation filterbank for speech analysis, the representation captures both acoustic frequency and temporal modulation frequency components, thereby conveying information that is important for human speech perception but missing from conventional short-term spectral features. On experiments using performed Berlin speech database and natural Chinese emotional speech database assessing classification of different discrete emotion categories, the MSFs show promising performance in

comparison with features that are based on mel-frequency cepstral coefficients and perceptual linear prediction coefficients, two commonly used short-term spectral representations.

**Key words:** Emotion recognition; Speech modulation; Spectro-temporal representation; Affective computing; Speech analysis