



中国科学院自动化研究所  
INSTITUTE OF AUTOMATION  
CHINESE ACADEMY OF SCIENCES

# 情感计算 —音频情感识别



授课教师：陶建华

# 目录

---

- 背景及意义
- 研究现状与进展
- 音频情感数据库
- 语音情感识别
- 音乐情感识别
- 展望

# 目录

---

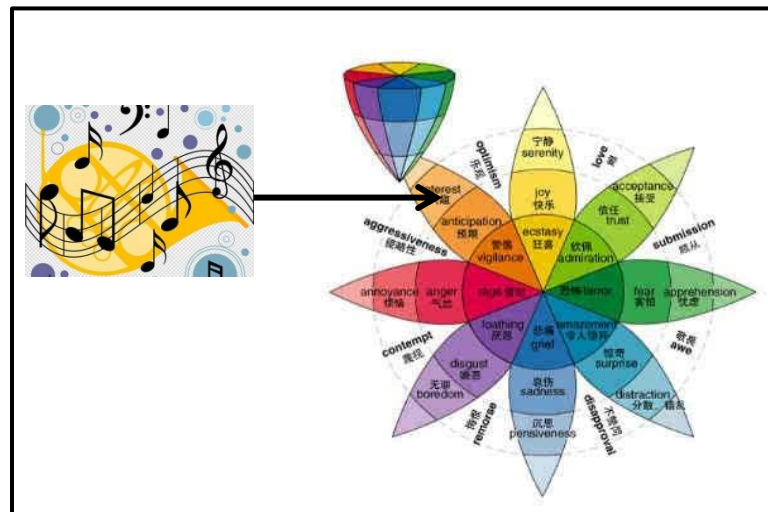
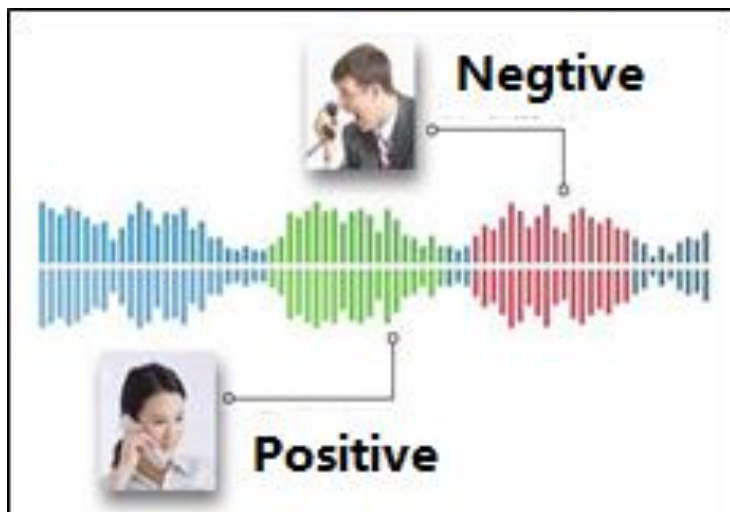
- 背景及意义
- 研究现状与进展
- 音频情感数据库
- 语音情感识别
- 音乐情感识别
- 展望

# 背景及意义

## ■ 音频：人类能够听到的所有声音

- 人声
- 音乐
- 环境音效

## ■ 音频中包含丰富的情感



# 没有情感会是什么样的？

## ■ 情感对语意理解的影响（言不尽意）



女：我从火车站怎么到你那？

男：我到火车站接你。（正常，Level 0）。

女：不，谢谢。告诉我去的路就行。

男：我到火车站接你。（有点不高兴，Level 1）。

女：只要告诉我去的路，我自己能去。

男：我到火车站接你！（有点急躁，Level 2）。

女：我自己去。

男：我到火车站接你！！（生气，Level 3）。

女：你真要来接我呀？

男：我到火车站接你！！！（愤怒，Level 4）。



# 意义

- 在**电话服务**中，系统可以检测谈话的语气和情感，从而提高服务质量。
- **医学研究**中，烦躁、焦虑、抑郁等不良情绪对治疗有很大的阻碍作用，如果能够更早发现病人情绪波动并及时稳定，对病人的康复也有着积极作用。
- 互动电影、情感翻译、机器人、电子游戏等等。



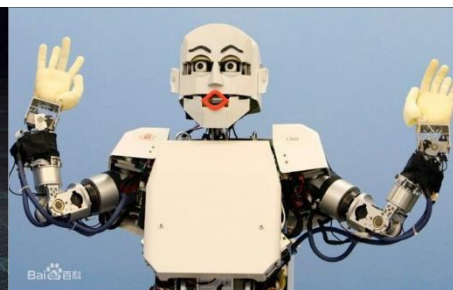
情感电话客服



精神疾病分析



谎言分析



机器人

# 目录

---

- 背景及意义
- 研究现状与进展
- 音频情感数据库
- 语音情感识别
- 音乐情感识别
- 展望

# 研究进展与现状

---

- 80年代末90年代，根据**韵律**控制人机对话的过程
- 90年代后期应用**模式识别**方式开始快速发展
- 目前应用**深度学习方法**性能获得了极大的提升，并得到了一些应用



# 研究进展与现状

---

## ■ 国外

### ■ 英国贝尔法斯特女王大学的情感语音组

收集并创建了第一个大规模的高自然度情感数据库, Roddy Cowie和 Ellen Douglas-Cowie 教授。重点研究心理学和语音分析。

### ■ 美国MIT媒体实验室情感计算研究所

Rosalind Picard 教授90年代初开始进行情感语音研究  
(<http://affect.media.mit.edu>)

### ■ 美国南加州大学语音情感组

Shri Naranya教授, 情感语音的声学分析、合成和识别, 以及有关笑声合成研究(<http://sail.usc.edu/emotion/index.php>)

### ■ 爱尔兰都柏林大学

Nick Campbell教授, 从事自然情感语音合成

# 研究进展与现状

---

## ■ 国外

### ■ 新加坡科技设计大学

**Soujanya Poria**教授，从事多模态和对话情感识别 (<https://declare-lab.net>)

### ■ 新加坡南洋理工大学

**Erik Cambria**教授，从事文本情感分析以及多模态和对话情感识别 (<https://sentic.net>)

### ■ 美国卡内基梅隆大学

**Louis-Philippe Morency**教授，从事多模态情感识别 (<http://multicomp.cs.cmu.edu>)

### ■ 英国帝国理工学院&德国奥斯堡大学

**Björn Schuller**教授，从事语音、视觉以及多模态情感识别 (<http://www.schuller.one>)

# 研究进展与现状

---

## ■ 国外

- 以色列Nemesysco公司实际应用以分层声音分析技术（LVA）在安全、商业和个人娱乐领域为客户提供解决方案。创业公司 Beyond Verbal以通过识别音域变化，从而分析出愤怒、焦虑、幸福或满足等情绪，其中包括11个类别， 400个复杂情绪的变量。
- 英国的初创企业EI Technologies可以分析人声的音调，识别高兴、悲伤、害怕、愤怒及无感情等5种用户的基本情绪。识别的准确率约为70-80%左右，这个数字要高于人类60%的平均水平，而受过训练的心理学家的判断准确率约为70%。
- 日本SGI研究院能感知人类情感：KOTOHANA  
(<http://www/.sgi/co/jp/solutions/bbu/ST/index/html>)

# 研究进展与现状

---

## ■ 国内

- 中科院自动化所模式识别国家重点实验室
- 东南大学无线电工程系
- 清华大学计算机科学与技术系
- 台湾大同大学资讯工程学系
- 中国公司Emotibot竹间智能科技、清帆科技EduBrain（专注教育领域技术创新）
- 其他：中国社科院语言研究所，西北工业大学、中国人民大学、哈工大，浙大，华南理工、中科大，南京师范大学、江苏大学等

# 目录

---

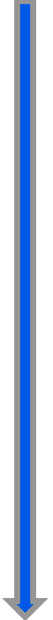
- 背景及意义
- 研究现状与进展
- 音频情感数据库
- 语音情感识别
- 音乐情感识别
- 展望

## ■ Cowie 提出了情感数据库建立必须依据的四个原则

- **真实性**，数据库中的素材应是人们所经历过的真实的情感体验
- **交互性**，数据库中的情感素材应是人们在人与人交互过程中产生的，这样更接近与语音情感人机交互的目的
- **连续性**，情感素材应该连续的情感场景中发生，存在着多种情感状态的转移
- **丰富性**，数据库中的情感素材应尽可能地包含多媒体信息，如声音，表情等

# 语音情感数据库

## ■ 获取语音情感数据库方法

- 
- **演员表演**：职业演员以模仿的方式表现出相应的情感状态，虽然说话人被要求尽量表达出自然的情感，但刻意模仿的情感还是显得更加夸大
  - **引导情感**：跟测试者交互谈话，引导情感的表达，比如让讲一个高兴的事或者看恐怖电影，或者测试者观看一段电影
  - **取自媒体**：从影视或者广播节目中截取的片段
  - **现实生活**：自然场景下的情感表达，一般较难获得，有些数据库基于电影电视剧片段进行类似

表演性 (acted)

引导性 (elicited)

自发性 (naturalistic)

# 语音情感数据库

---

## ■ Belfast 英语情绪语料库

- **引导情感：**由50位说话人根据引导文本，表达愤怒、恐惧、高兴、悲伤和中性五种情感
- **自然语料：**从电视访谈节目中选取剪辑的125位说话人的多种情感

## ■ 柏林（EMO-DB）情感数据库

- **引导情感：**日常交流中常用的十个德语语句，共800语句，含七种情感：中性、愤怒、恐惧、高兴、悲伤、厌恶和惊奇

## ■ FAU AIBO儿童德语情感语音库

- **自然情感：**录制51儿童（10-13岁，21男30女）与索尼公司生产的电子宠物AIBO游戏过程中的自然语音，保留情感信息明显的语料，共9.2小时，包括48401单词



# 语音情感数据库

---

## ■ 汉语普通话语音数据库

- **演员表演：**中科院自动化所研制，由演员模仿情感进行录制，共9600条语音，包括6中情感：高兴、生气、惊奇、控制、悲伤、平静
- **演员表演：**东南大学研制，表演性情感语音，由10名男性话者对4个语句分别用喜、怒、惊、悲四种情感录制480句

## ■ CREST情绪语料库

- **自然情感：**日本的国际电气通信基础研究所（ATR）录制，包含完全自然状态下的1000小时情感语音，60%是日语，汉语和英语各占20%

## ■ 丹麦情感语音库

- **演员表演：**含5种情感：高兴、生气、惊奇、悲伤、平静。共260条

# 语音情感数据库

---

## ■ CHEAVD 数据库

- **取自媒体：**由中科院自动化所从32部中文电影、79集电视剧、20期综艺节目中剪辑出2629个音视频情感片段，共计时长141分钟，共包括238位发音人。

## ■ Semaine数据库

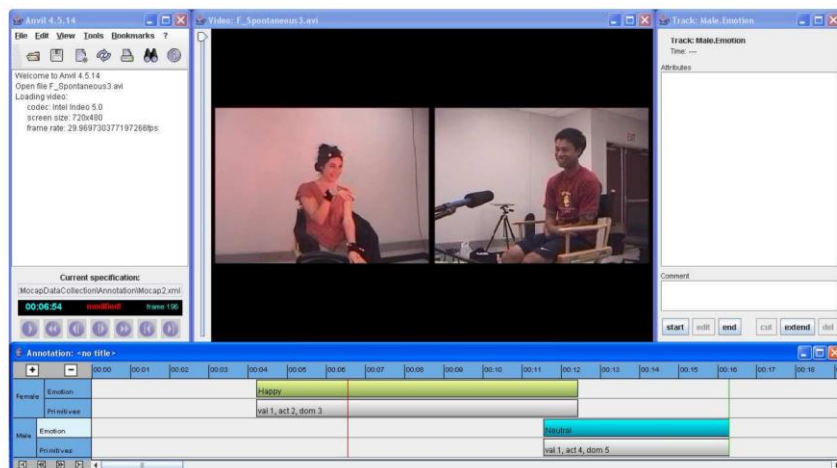
- **引导情感：**面向自然人机交互和人工智能研究的数据库，20 个用户（22 岁-60 岁, 8 男12 女）被要求与性格迥异的4个机器角色进行交谈. 这4 个角色分别是: 1) 温和而智慧的Prudence; 2) 快乐而外向的Poppy; 3) 怒气冲冲的Spike 和4) 悲伤而抑郁的Obadiah

# 语音情感数据库

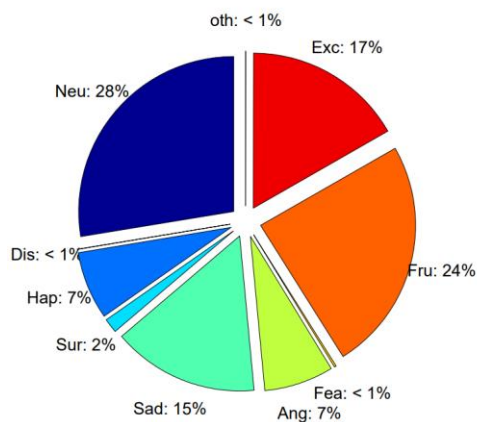
## ■ 南加州大学交互式情感对话数据库：IEMOCAP

- **演员表演：**含10种情感：如中性、高兴、生气、悲伤等，且提供PAD维度情感。10名专业演员两两进行对话（脚本/即兴表演），共10039条数据。

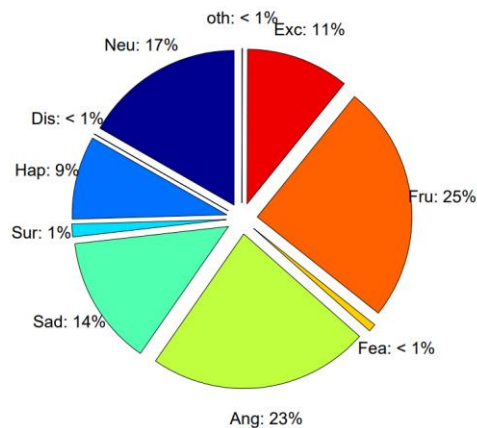
标注界面



情感分布



(a) Scripted session



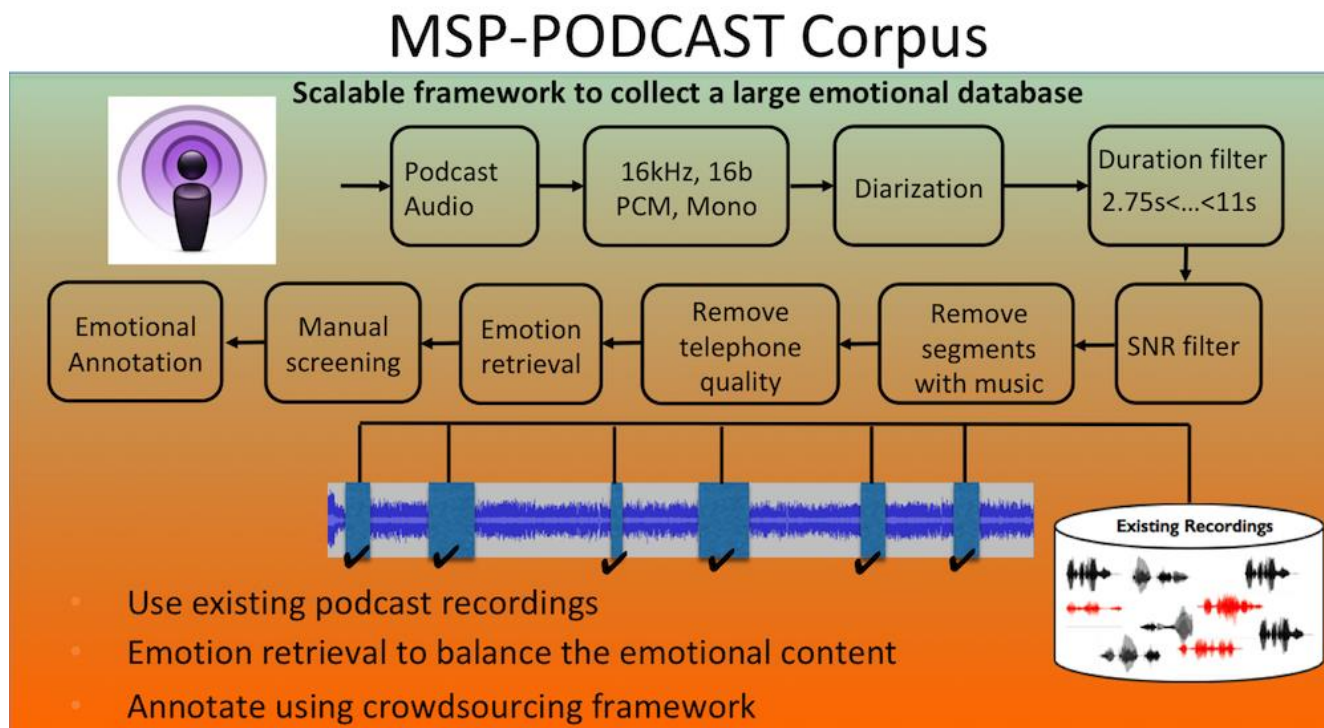
(b) Spontaneous sessions

# 语音情感数据库

## ■ 南加州大学自然语音情感数据库：MSP-Podcast

- **自然情感：** 提供PAD维度标签，以及8种离散情感：如生气、伤心、高兴、惊讶等。该数据库样本从音频共享网站获取，目前版本大小已超过100h。

数据收集流程



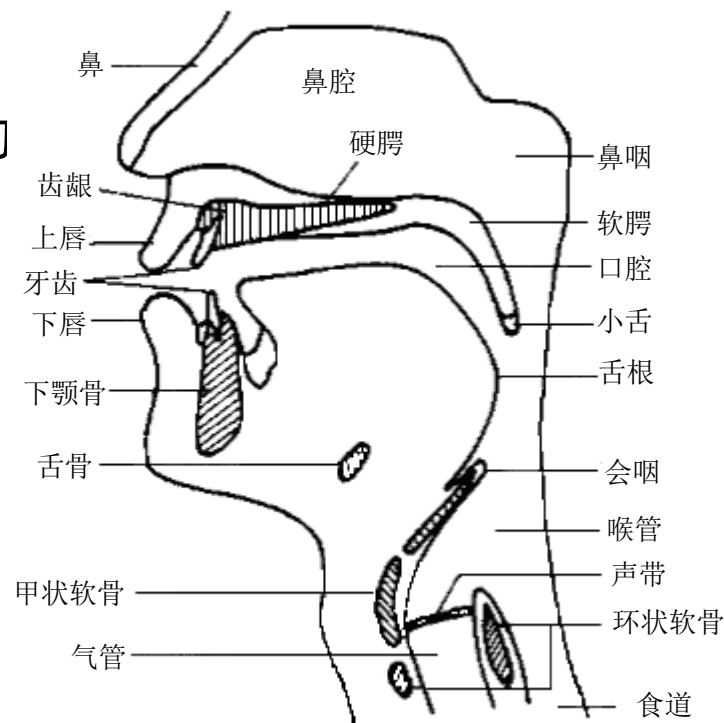
# 目录

---

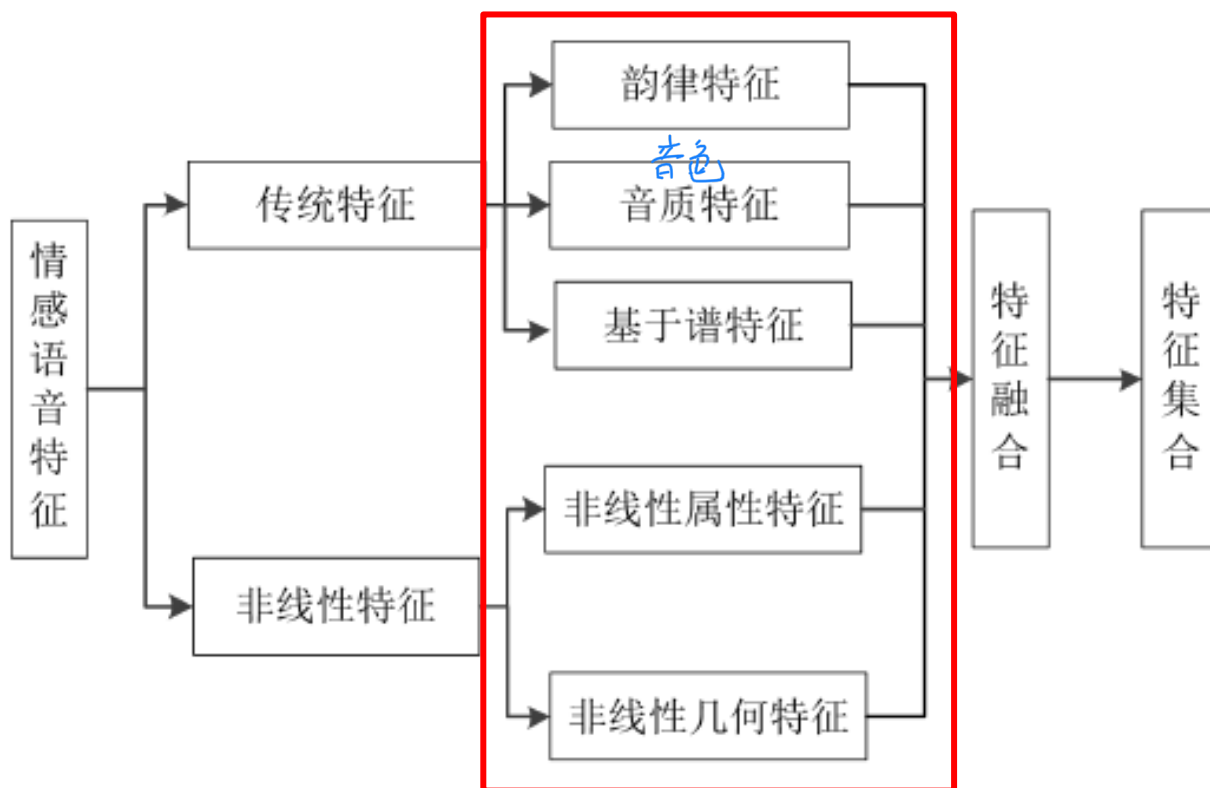
- 背景及意义
- 研究现状与进展
- 音频情感数据库
- 语音情感识别
- 音乐情感识别
- 展望

# 语音特征

- 人类的发声器官由肺、气管、声带、鼻、口和唇等组成。
- 肺部的气流经气管呼出时，呈一定张力的声带振动发声。
- 声带的长短和张力决定了声音的基频。
- 声音的强度取决于气流的大小和强度。
- 声音经过形状变化的口腔、鼻腔等共振，最后经唇部辐射传出。
- 其中声音的共振过程形成了共振峰，各共振峰的频率由共振腔的大小和形状决定。



# 语音情感特征



# 语音情感特征

---

## ■ 语音情感特征种类

- 韵律特征：最主要的语音情感特征，如语速、音量和音调等，例如发怒时都会增加；振幅、基音频率，持续时间；
- 音质特征：音频抖动（Jitter）和振幅抖动（shimmer），谐波噪声率，共振峰；
- 频谱特征：MFCC, LPCC。



# 语音情感特征

---

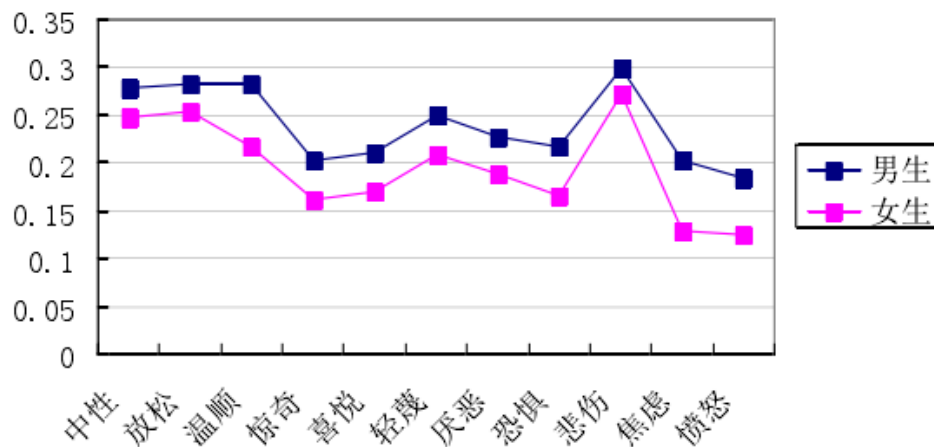
## ■ 语音情感特征种类

- 韵律特征：最主要的语音情感特征，如语速、音量和音调等，例如发怒时都会增加；振幅、基音频率，持续时间；
- 音质特征：音频抖动（Jitter）和振幅抖动（shimmer），谐波噪声率，共振峰；
- 频谱特征：MFCC, LPCC。

# 语音情感特征

## ■ 语速

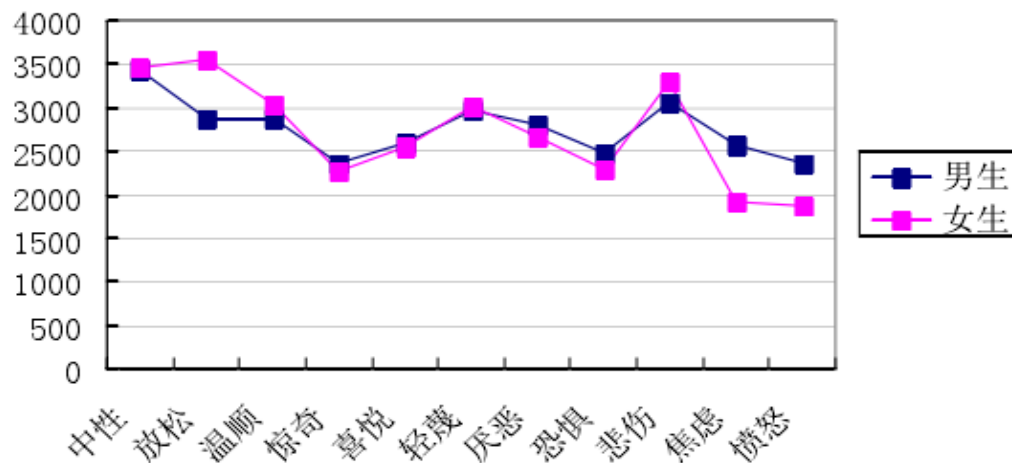
- 语音可以反应出说话者的情绪状态：当人的情绪比较激动的时候，比如处于愤怒状态，语言的表达速度明显加快，相反在人的情绪比较低落时，比如处于悲伤状态，语言的表达速度则明显较慢。



# 语音情感特征

## ■ 时长

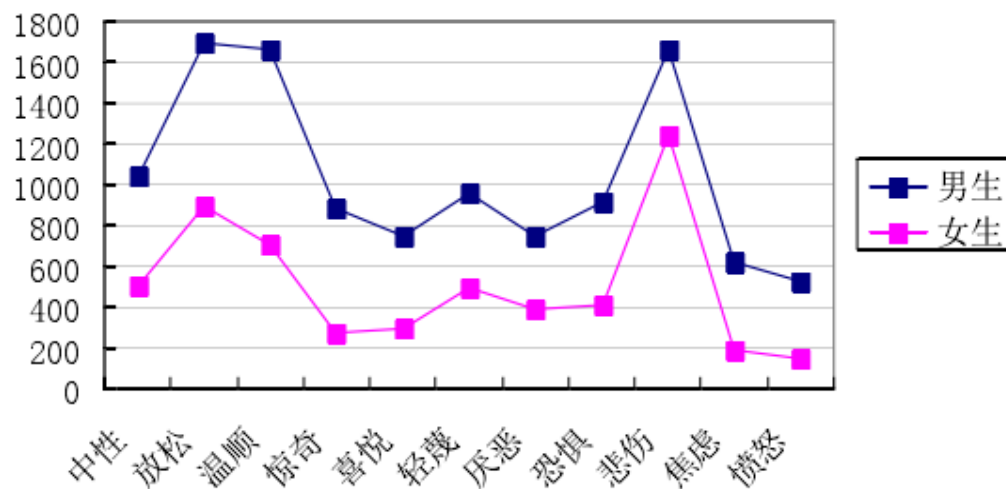
- 语句的发音持续时间指每一情感语句从开始到结束的持续时间，与感知的语速相对应，情感语音的时长构造主要着眼于不同情感语音发话时间构造的差别，时长分析常采用音节、句子为单元来测量。



# 语音情感特征

## ■ 停顿

- 停顿也反映了情感信息，停顿指的是前一个音节与下一个音节之间无声的时间。



# 语音情感特征

---

## ■ 分析

- 语速：语速的变化是表达情感的一个重要手段。它反映一个人在不同环境，不同情感下说话时的心情急切度。人在焦虑和愤怒状态下，说话速度很快；惊奇和喜悦次之；而悲伤情感下说话速度最慢。
- 停顿和时长：
  - 不管是男性还是女性，不同情感下，其停顿和时长对于基本类型的情感变化有一定的一致性，对于微妙复杂的情感两者的变化有一定的差异。
  - 性别不同，也会引起一些情感之间特征的变化差异性。比如男性说话人在放松和温顺情感下的时长变化略不同于女性说话人，停顿分析中惊奇与喜悦情感在不同性别中的变化稍有不同。

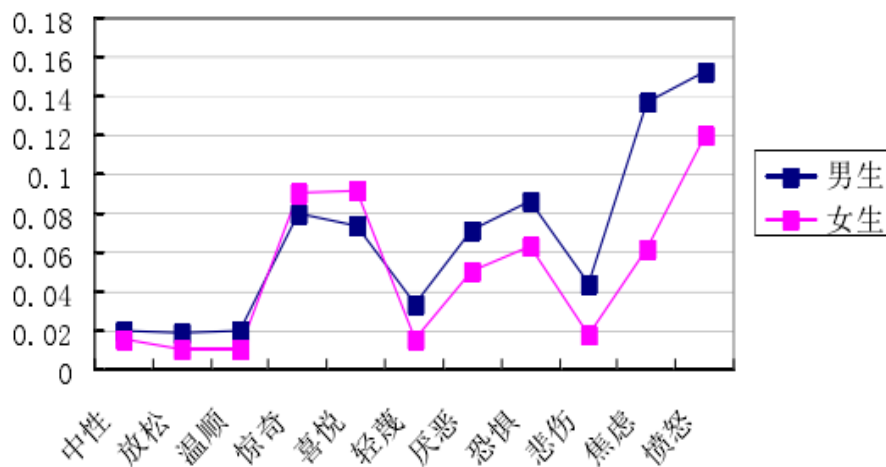
# 语音情感特征

## ■ 音强（能量）

- 能量表现为语音的音量的高低, 而音量的高低又是通过声音的响度大小来反映。

$$E_n = \sum_{m=-\infty}^{\infty} [s(m) w(n-m)]^2 = \sum_{m=n-N+1}^n [s(m) w(n-m)]^2$$

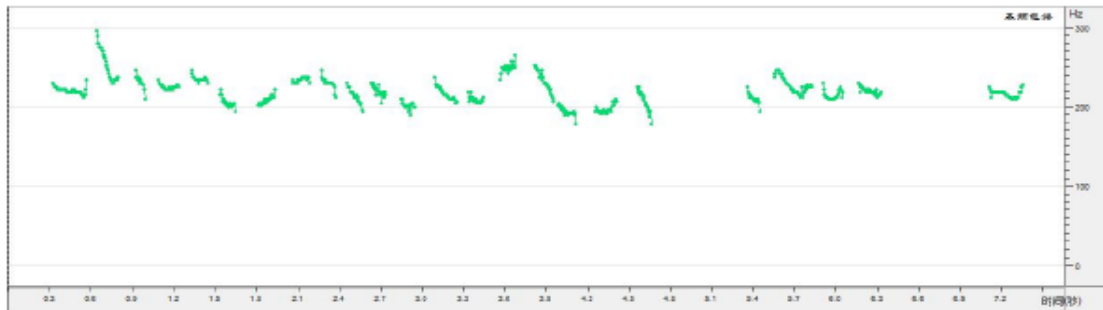
- 不管是男性还是女性说话人, 其中中性、放松、温顺情感的能量基本在同一水平, 总体能量较低, 其次是轻蔑和悲伤能量相近, 惊奇、喜悦和恐惧能量处于一个水平, 能量最强的属于愤怒情感。



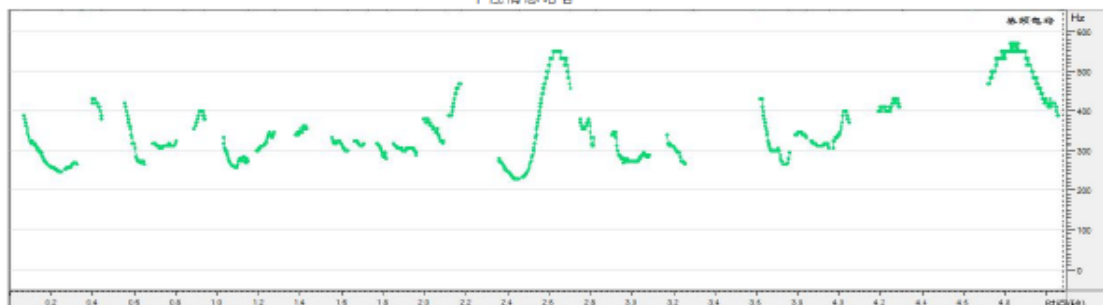
# 语音情感特征

## ■ 基频

- 在发出浊音时, 声门波形成的周期性脉冲, 即声带的振动周期被称作是浊音的基音周期, 基音频率即为其倒数, 简称基频, 通常用 $F_0$ 表示。基频值取决于声带大小、厚薄、松紧程度以及声门上下之间的气压差效应等。



中性情感语音



喜悦情感语音

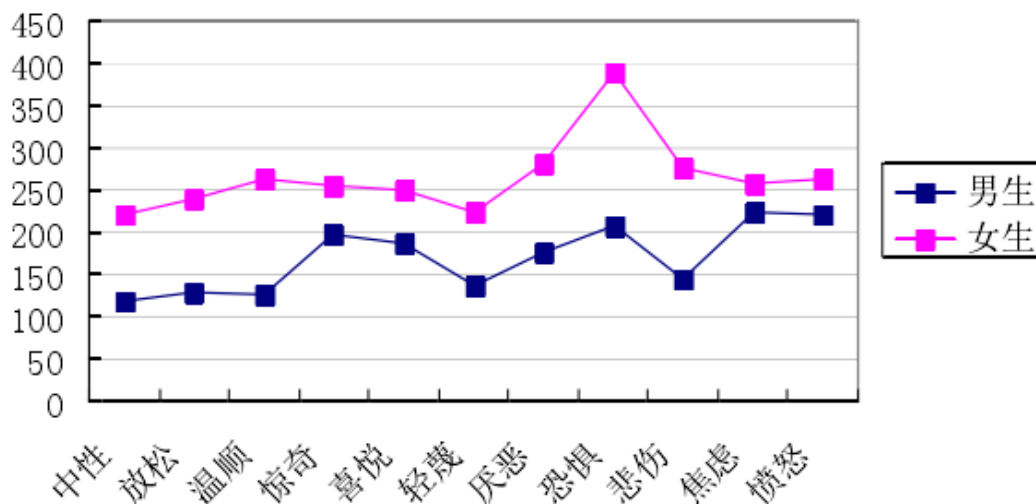


典型的声门脉冲串波形

# 语音情感特征

## ■ 基频

- 基音频率体现出以下规律：处在**激动情绪**下如愤怒的人所表达出的语音的**基频较高，变化范围较大**；处于低落情绪如悲伤的人所表达的语音的基频较低，变化范围较小，处于平静情绪下的人所表达出的语音的基频则相对稳定。

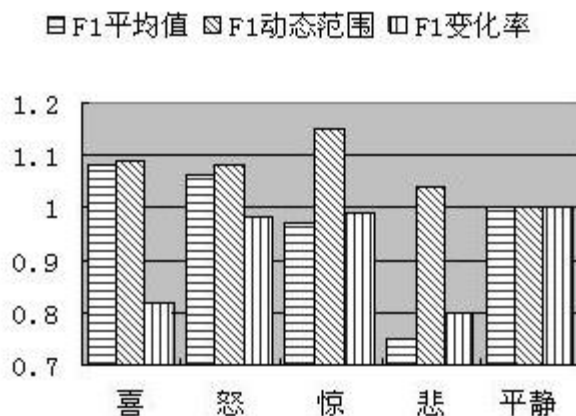
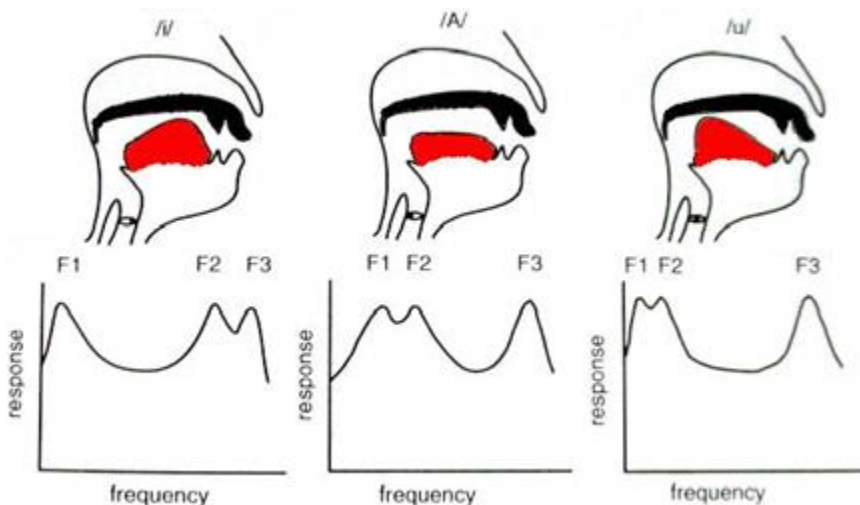




# 语音情感特征

## ■ 共振峰

- 共振峰是反映声道特性的一个重要参数。不同情感发音的共振峰的位置不同。分析时首先用LPC法求出声道的功率谱包络，在用峰值检出法算出个共振峰的频率。



# 语音情感特征

## ■ 分析

- 可以上面的分析对含有四种情感的语音信号进行分析比较，归纳如下表所示的结论：

	音色 T	基频 F <sub>0</sub>	F <sub>0</sub> range	F <sub>0</sub> rate	音强 A	A range	共振峰 F <sub>1</sub>	F <sub>1</sub> range	F <sub>1</sub> rate
喜	+	+	+	+	+	++	+	+	-
怒	-	+	+	++	+	++	+	+	-
惊	-	++	++	++	++	++	-	+	-
悲	++	-	-	--	-	+	--	+	--

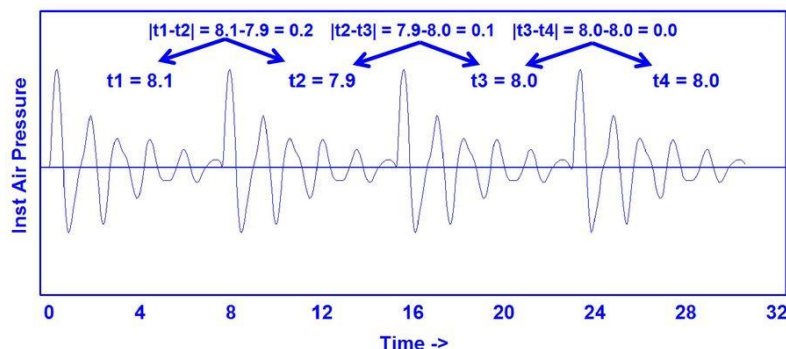
(上表中符号意义 +: 增加 ++: 较大增加 -: 减小 --: 较大减小 -: 无明显变化)

# 语音情感特征

## ■ 基频抖动（Jitter）

- 焦虑语音会出现“F0抖动”现象。Jitter是基频值的变化程度。
- F0 Jitter是由生理器官的作用才产生，比如情感的变化会导致声带肌肉紧张度，气流的体积速度，声道表面的坚硬或柔软等发生变化，从而产生基频抖动现象。

### How Jitter is Measured: Mean Jitter



$$\begin{aligned}\text{Mean Jitter} &= \text{sum of (abs) period diffs} / \text{number of diffs} \\ &= 0.2 + 0.1 + 0.0 / 3 = 0.3 / 3 = 0.1 \text{ ms}\end{aligned}$$

$$\text{In English: MeanJ} = \text{SumOfAbsDiffs} / \text{ndiffs}$$

# 语音情感特征

---

## ■ 线性预测倒谱系数（LPCC）

- LPCC是基于语音信号为自回归信号的假设，利用线性预测分析获得倒谱系数。
- 不同情感的发音会使声道有不同的变化，进而引起声道传输函数倒谱的变化。

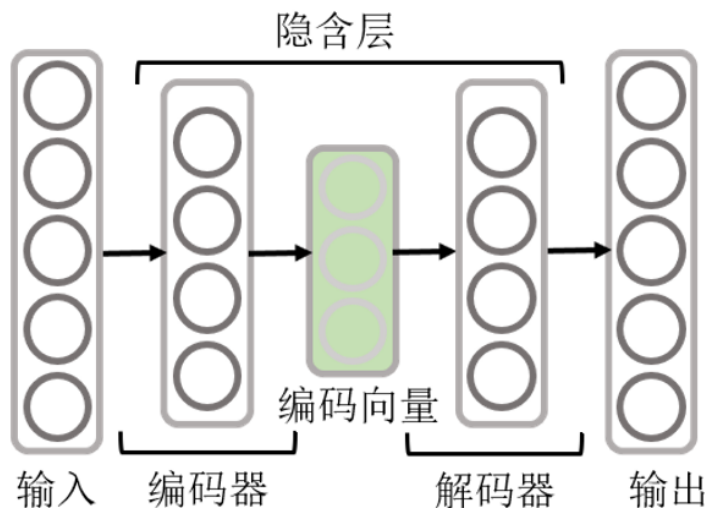
## ■ Mel频域倒谱系数（MFCC）

- MFCC考虑了人耳对不同频带的分辨率不同，充分融合了人耳的听觉特性。

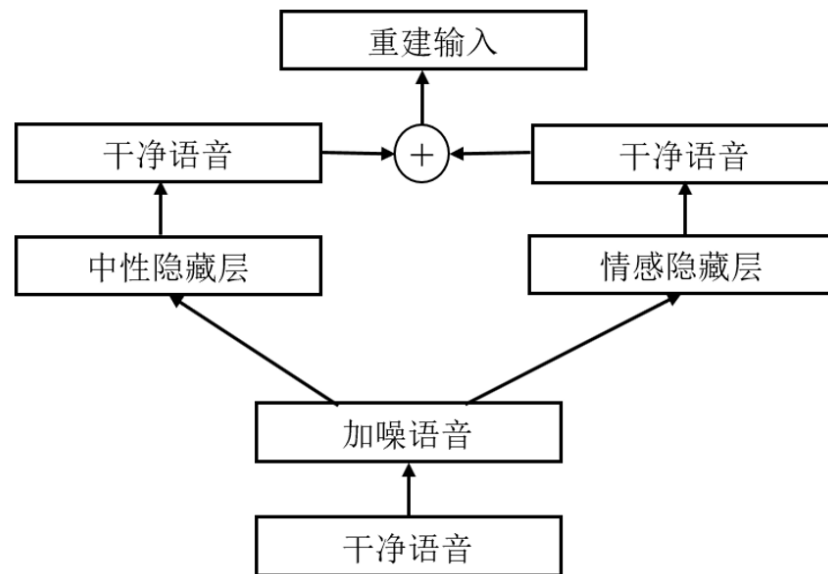
# 语音情感特征

## ■ 其他语音特征

- 通过无监督学习模型发现语音中的层次结构和内在分布，从而更好地编码原始语音，以期挖掘更好的情感特征。



自编码器

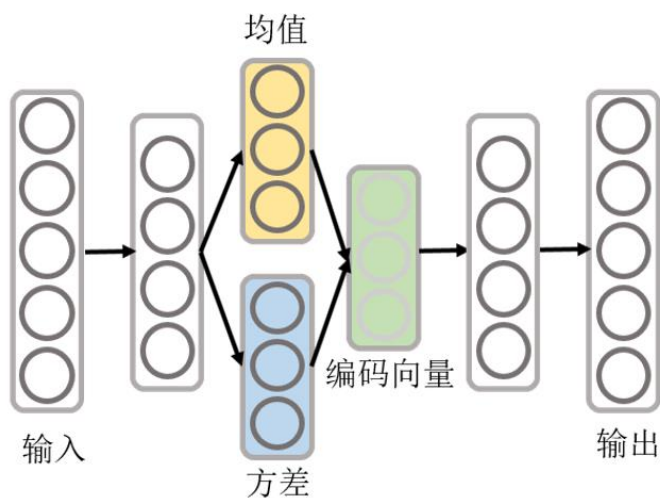


降噪自编码器

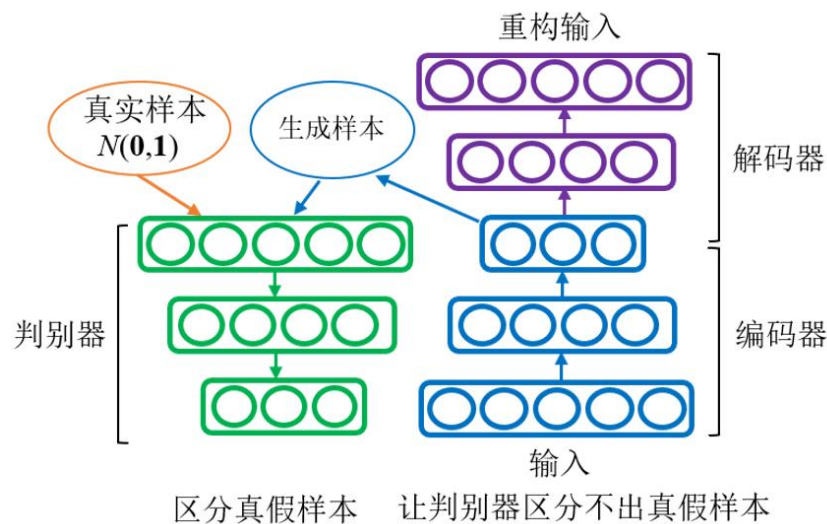
# 语音情感特征

## ■ 其他语音特征

- 通过无监督学习模型发现语音中的层次结构和内在分布，从而更好地编码原始语音，以期挖掘更好的情感特征。



变分自编码器



对抗自编码器

# 语音情感特征

---

## ■ 语音情感特征总结

- 韵律特征、音质特征和频谱特征相结合
- 分析情感语音和平静语音相对关系，找出这种相关特征的构造、特点和分布规律，以消除语音影响
- 局部特征 & 全局特征
- 多类特征组合

# 语音情感特征

## ■ 功能性副语言中携带了大量情感信息

功能性副语言	高兴	伤心	惊讶	生气	害怕	厌恶
笑声	Y	N	N	N	N	N
伤心的哭声	N	Y	N	N	N	N
质疑声	N	N	Y	N	N	N
叫喊声	N	N	N	Y	N	N
害怕的哭声	N	N	N	N	Y	N
叹息声	N	N	N	N	N	Y



# 语音情感特征

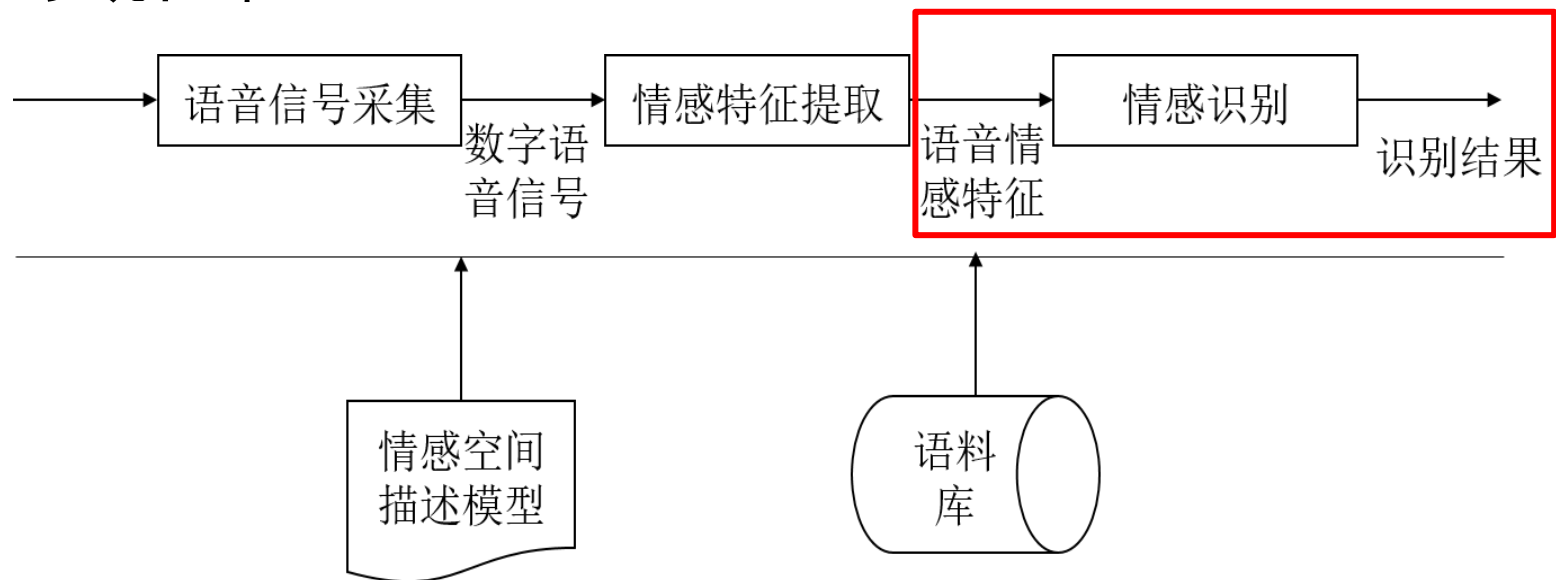
---

## ■ 特征集

- 任意类型特征都有各自的侧重点和使用范围，不同的特征之间具有互补性，因此在相当多的文献采用了混合参数构成特征向量
- the INTERSPEECH 2009 Emotion Challenge: 384
- the INTERSPEECH 2010 Paralinguistic Challenge: 1582
- the INTERSPEECH 2011 Speaker State Challenge: 4368
- the SPEECH 2012 Speaker Trait Challenge: 6125
- the INTERSPEECH 2013: 6373

# 语音情感识别

## ■ 系统框架

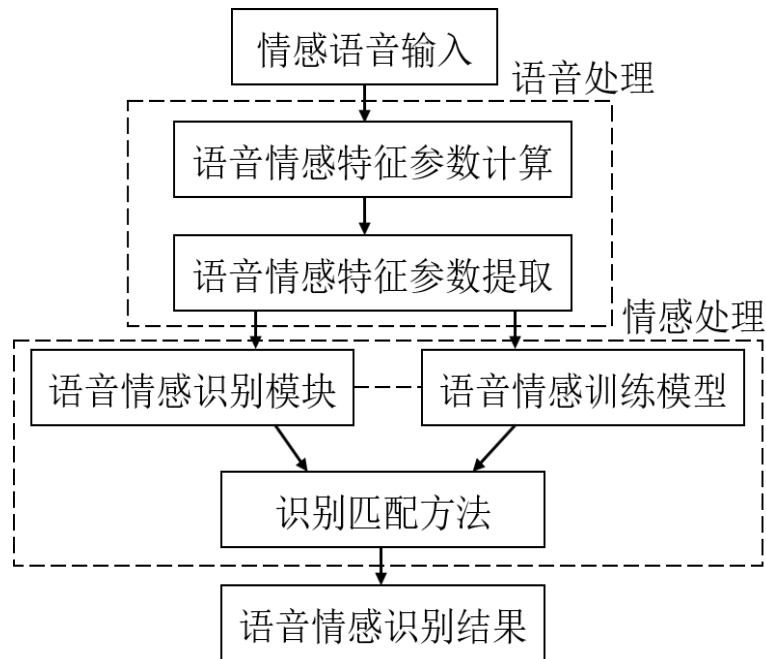


- 情感描述模型
- 语料库
- 语音信号采集
- 情感特征提取
- 情感识别模型

# 语音情感识别方法

## ■ 语音情感识别本质上属于模式识别

- 神经网络
- 高斯混合模型
- 隐马尔科夫模型
- 支持向量机
- 集成学习算法
- 深度神经网络



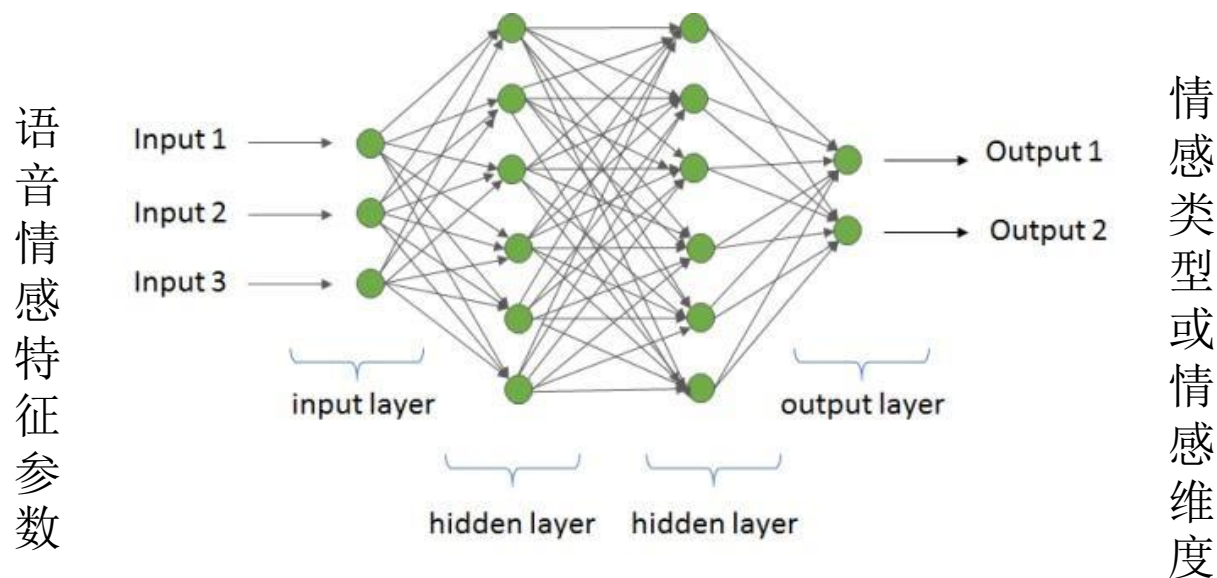
# 语音情感识别的分类

---

- 离散情感识别
- 连续情感识别

# 语音情感识别方法

## ■ 神经网络 (ANN)



运用神经网络可以达到70%的识别率

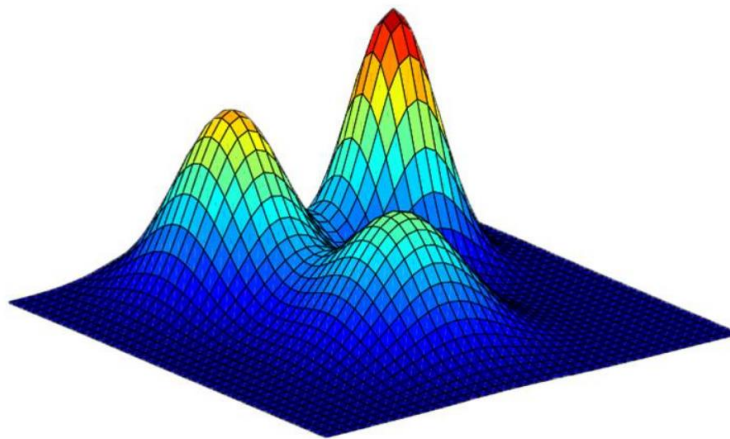
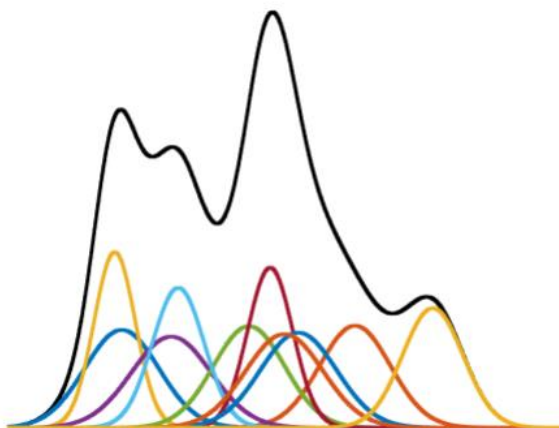
# 语音情感识别方法

## ■ 混合高斯模型法(GMM)

- 混合高斯模型是只有一个状态的模型，在这个状态里具有多个高斯分布函数。

$$P_k = \sum_{i=1}^N w_i f_i(\vec{Y})$$

- 其中  $f_i$  是一个高斯分布函数，不同高斯分布之间的加权系数  $w_i$  满足  $\sum_{i=1}^N w_i = 1$



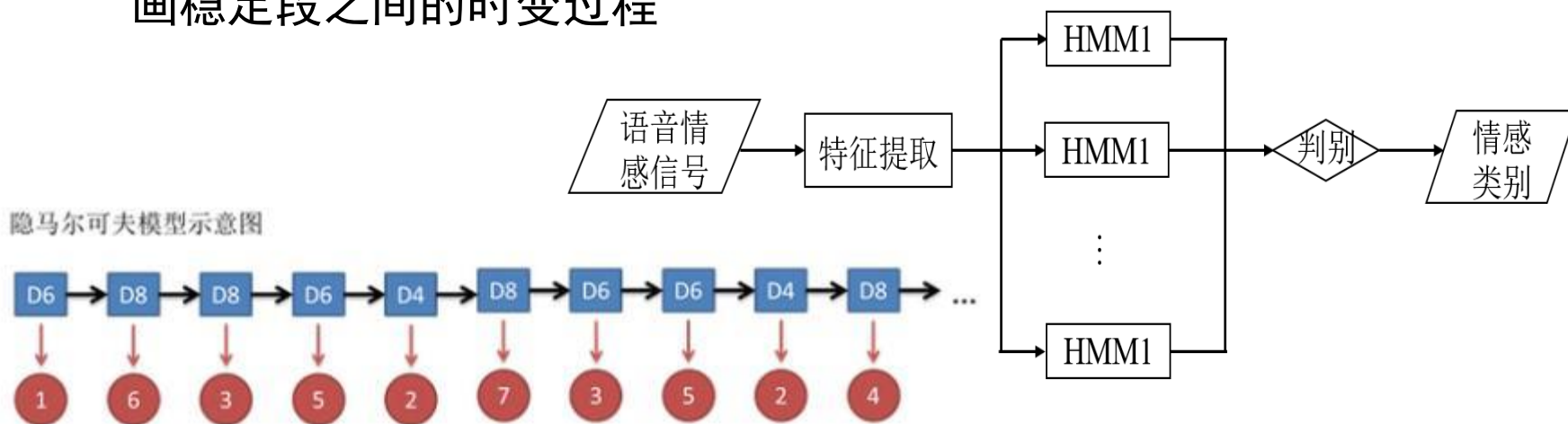
每一个情感类型训练一个高斯混合函数，高斯函数的输入参数为语音的情感特征参数。

2009年，基于混合高斯模型的识别系统在语音领域的著名的国际会议Interspeech上举行语音情感识别的评比中，在总体性能上获得了该次比赛的第一。

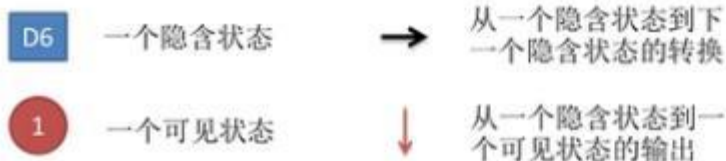
# 语音情感识别方法

## ■ 隐马尔科夫模型（HMM）

- HMM是一种基于转移概率和观测概率的随机模型，它既能用短时模型（状态）解决声学特征相对稳定段的描述，又能用状态转移规律刻画稳定段之间的时变过程

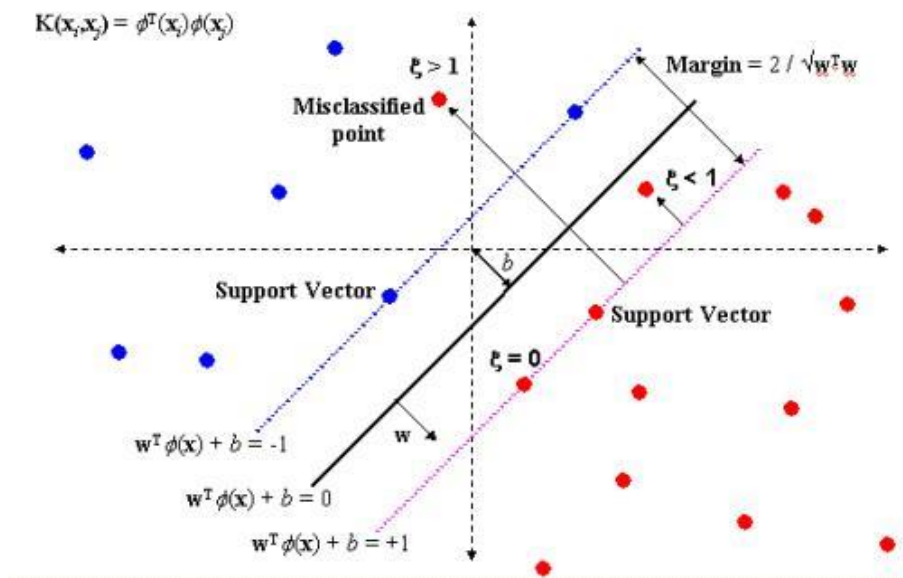


图例说明：



## ■ 支持向量机

- 基于结构风险最小化和统计学习理论提出了一种名为支持向量机（SVM）的机器学习方法，该方法在诸如函数拟合、非线性模式识别。小样本等领域都极具优势。



$$\min_{w, b, x} \left\{ -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j a_i a_j (x_i \cdot x_j) + \sum_{i=1}^n a_i \right\}$$
$$s.t. \quad \sum_{i=1}^n a_i y_i = 0, a_i \geq 0, i = 1, 2, \dots, n$$



# 语音情感识别方法

## ■ 常用统计方法比较

识别算法	对语音情感数据的拟合性能	识别率	优点	缺点
GMM	高	在 AIBO 数据库、本文数据库上表现较高	对数据的拟合能力较高	对训练数据依赖性强
SVM	较高	在柏林库上表现较高	适合于小样本训练集	多类分类问题中存在不足
KNN	较高	在柏林库上表现一般	易于实现,较符合语音情感数据的分布特性	计算量较大
HMM	一般	在柏林库上表现较高	适合于时序序列的识别	受到音位信息的影响较大
决策树	一般	在 AIBO 数据库上表现一般	易于实现,适合于离散情感类别的识别	识别率有待提高
ANN	较高	在日语情感语音上表现一般	逼近复杂的非线性关系	容易陷入局部极小特性和算法收敛速度较低的
混合蛙跳算法	较高	在汉语音情感数据上表现较高	优化能力强,有利于发现情感数据中潜在的模式	在迭代后期容易陷入局部最优,收敛速度较慢

# 语音情感识别方法

---

## ■ 混合模型

- 把上述若干模型融合起来，各自取长补短，形成混合模型
  - 并联融合
  - 串联融合

# 语音情感识别方法

---

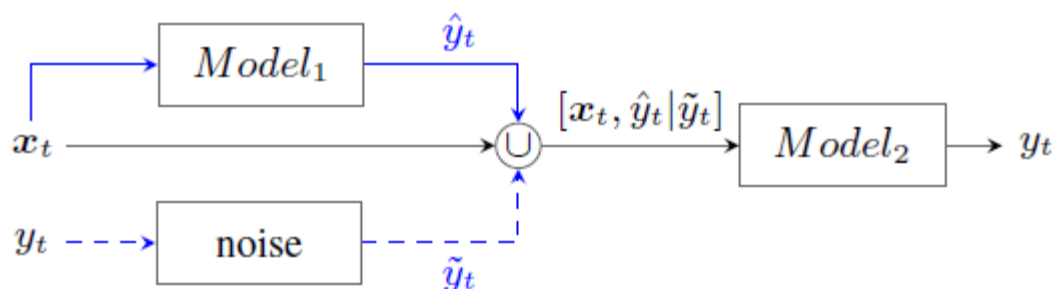
## ■ 并联融合

- 并联融合是将单项分别进行独立的匹配处理，得到各个匹配分数，通过融合算法将各匹配分数进行综合得到最终决策结果
- GMM/K最近邻的方法
- SVM/ANN的方法
- HMM/ANN的方法

# 语音情感识别方法

## ■ 串联融合

- 串联融合是将前面分类器的输出作为后面分类器的输入，最终决策结果由后面分类器决定
- GMM/SVM方法
- SVM/LSTM方法
- LSTM/SVM 方法



# 语音情感识别方法

---

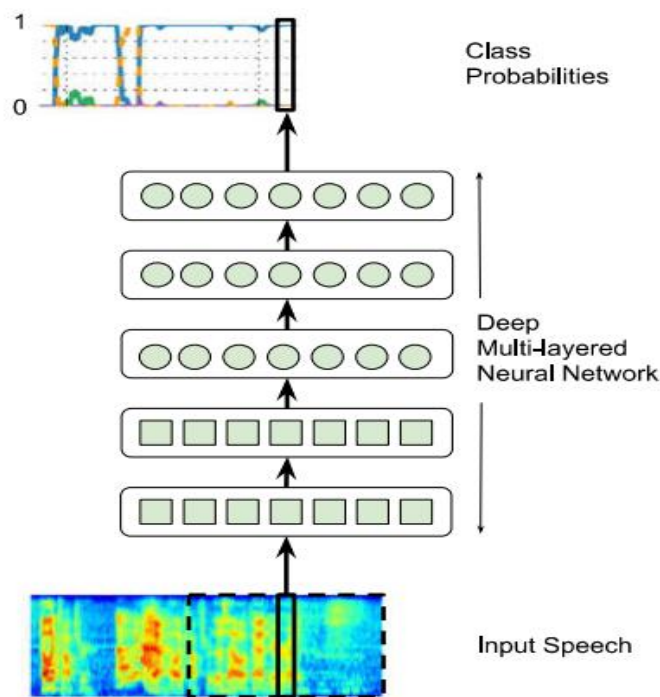
## ■ 混合模型

- **混合高斯模型-支持向量机**：该方法不仅拥有混合高斯模型统计能力强的优点，而且同时有支持向量机分类能力强的优点。
- **隐马尔科夫模型-人工神经网络模型**：首先用隐马尔科夫模型对情感特征向量进行整合，再用人工神经网络进行最终分类识别。
- **采用投票机制**将支持向量机、K最近邻算法、人工神经网络种分类器进行融合。

# 语音情感识别方法

## ■ 基于深度学习的语音情感识别模型

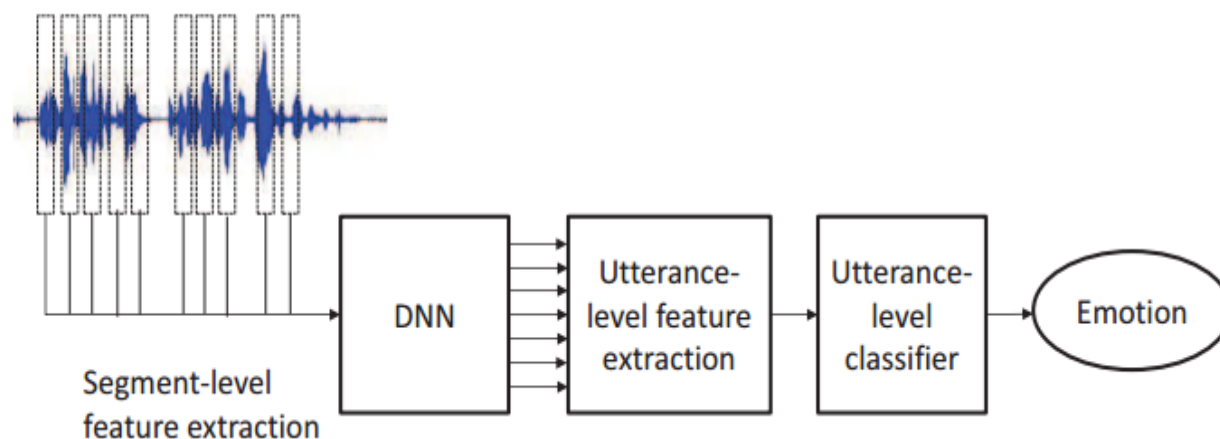
- 基于深度学习的语音情感识别模型输入是语音信号，经过多层网络结构，输出各个情感类别概率，得到预测结果。中间网络层可以是不同的网络结构，如**深度神经网络、卷积神经网络或循环神经网络**等。



# 语音情感识别方法

## ■ DNN

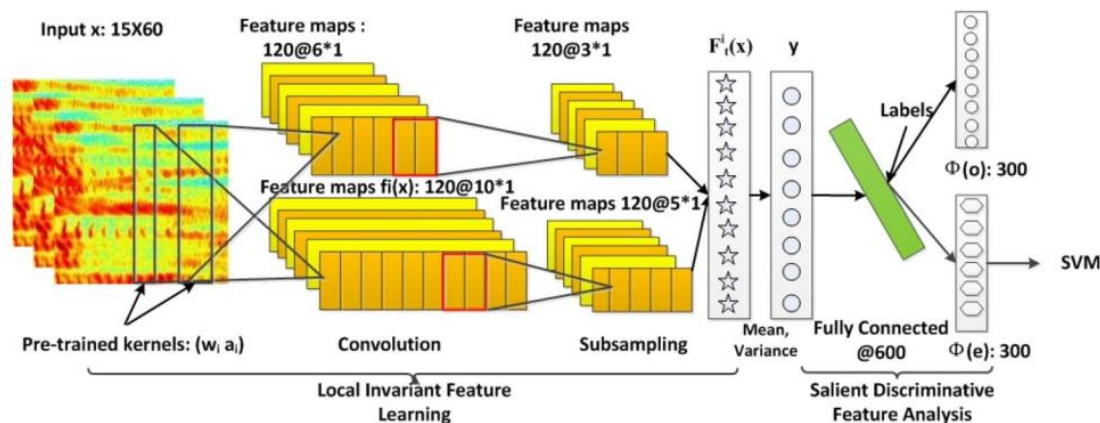
- 原始语音信号分段输入到网络中，提取局部的情感信息，然后经过处理得到全局情感特征，送到分类器中，得到预测的概率类别。输入一段语音情感信号，可以得到每一段对每个情感类别的预测概率值。



# 语音情感识别方法

## ■ CNN

- 卷积神经网络广泛应用到语音情感识别中，包括局部不变特征学习模块、情感区分特征分析模块和支持向量机模块三个部分。这种模型结构能够抽取出具有区分性的情感特征。



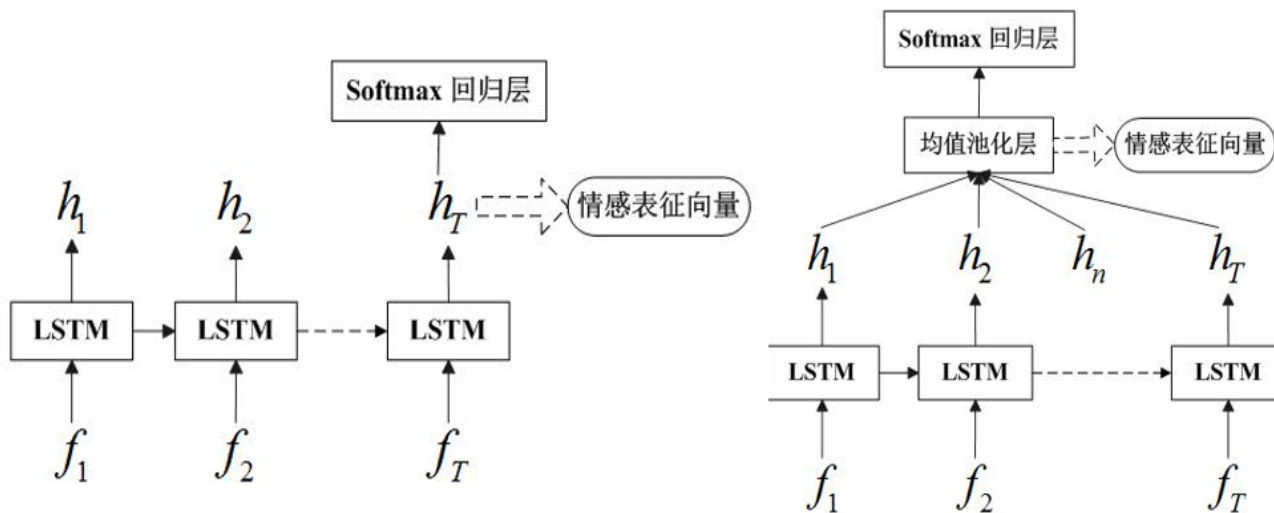


# 语音情感识别方法

## ■ RNN

RNN擅长对序列数据进行建模

■ 能够有效融合上下文信息进行音频情感建模

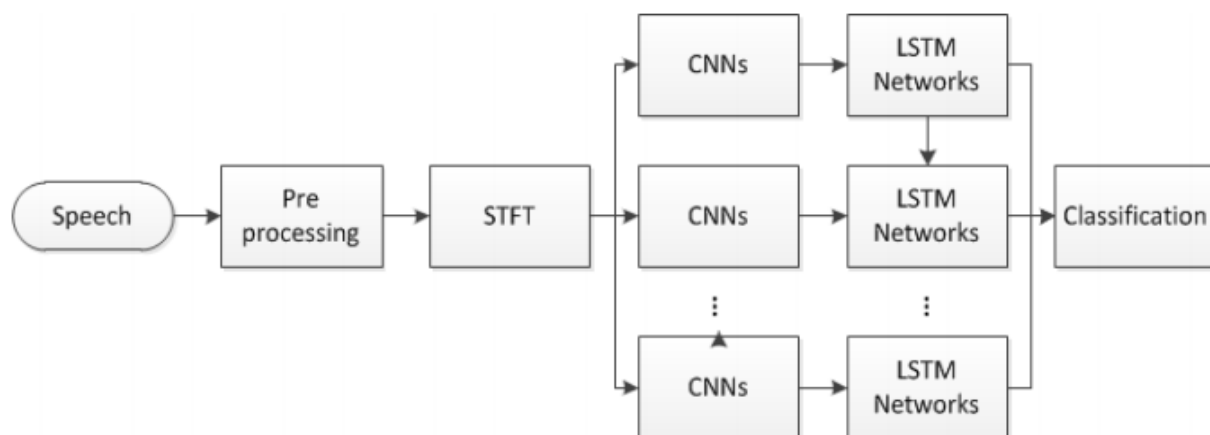


Encoding method	Network topology	Accuracy
LSTM-mean	2048/64/64/7	0.4420
LSTM-last	2048/64/64/7	0.3909
temporal mean pooling	2048/64/7	0.4394
temporal max-pooling	2048/64/7	0.4474
temporal max-pooling	2048/256/7	0.4528
max-pooling	2048/64/7	0.4339
mean pooling	2048/64/7	0.4367

# 语音情感识别方法

## ■ CNN-RNN

- 联合了CNN模型的音频信息表征能力和RNN模型的情感时序建模能力

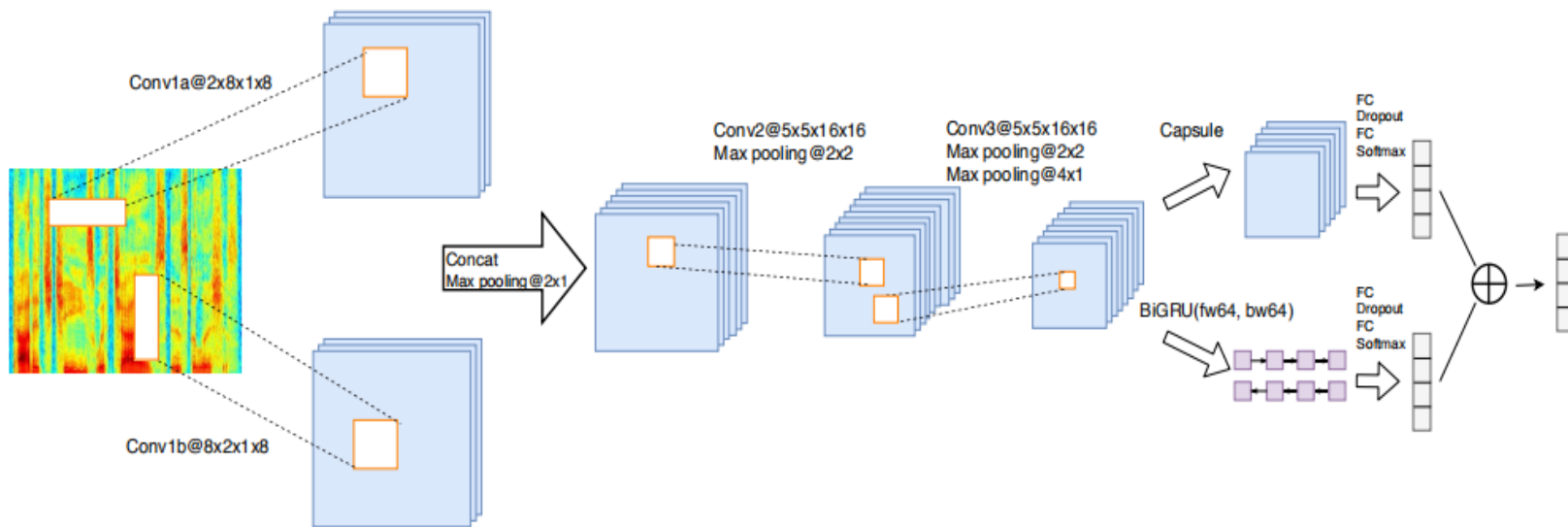


网络模型	精确率	召回率	F1
卷积神经网络	87.74%	86.32%	86.06%
循环神经网络	79.87%	78.83%	78.31%
卷积神经网络 和循环神经网络	88.01%	86.86%	86.65%

# 语音情感识别方法

## ■ 胶囊网络

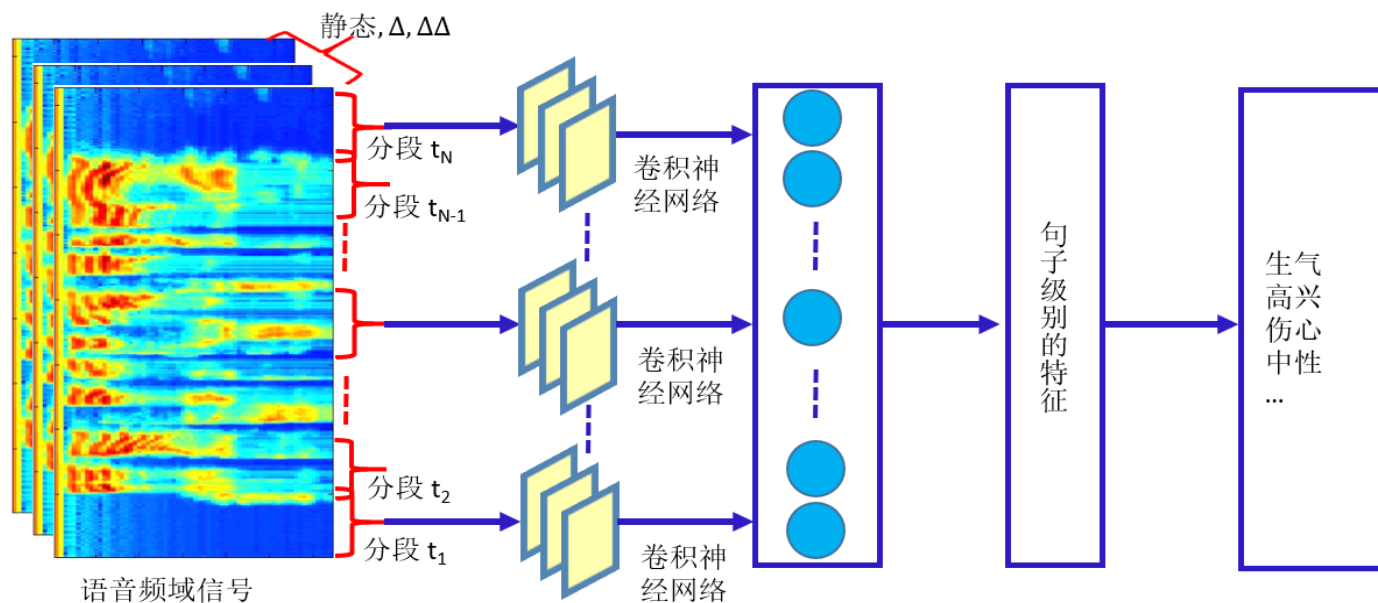
- 利用胶囊网络，**考虑音频特征在声谱图中的空间关系**，为获取语音全局特征提供了一种有效的汇聚方法



# 语音情感识别方法

## ■ 基于端到端音频情感识别

- 直接将语谱图或时域波形点作为输入，具有特征自学习能力的优势
- 其性能依赖于数据规模

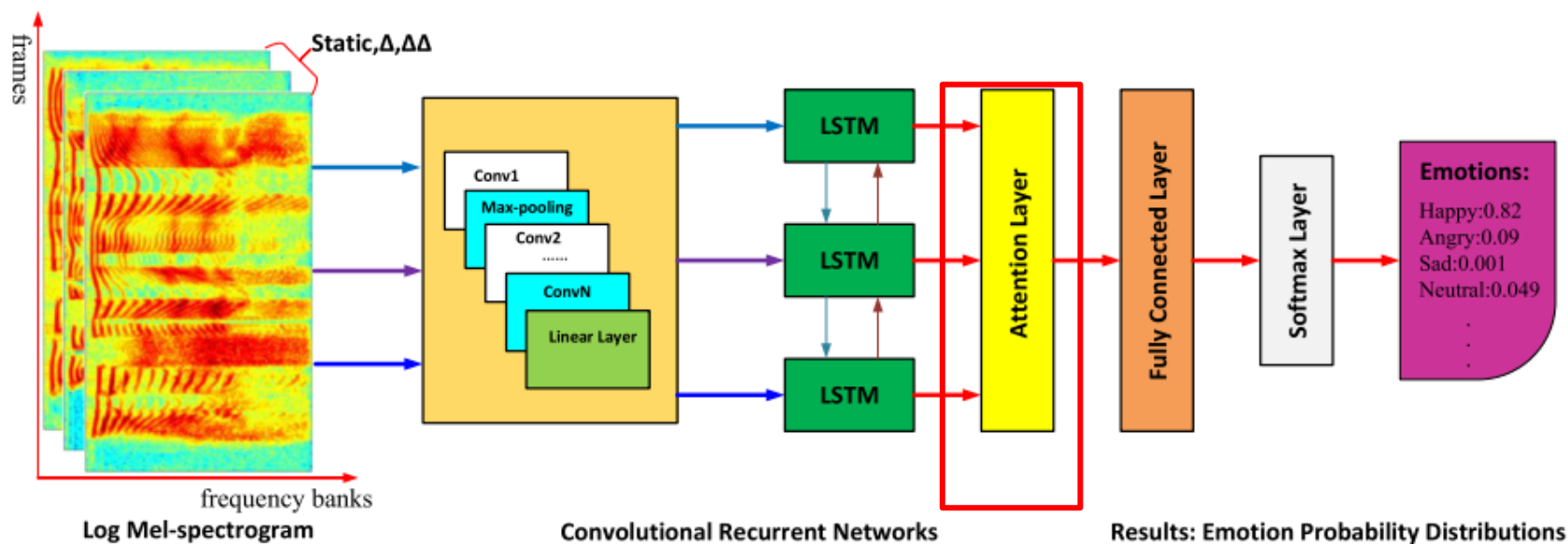


基于端到端卷积神经网络的语音情感识别系统

# 语音情感识别方法

## ■ 注意力机制

### ■ 有效挖掘不同音频片段对当前情感状态的贡献度

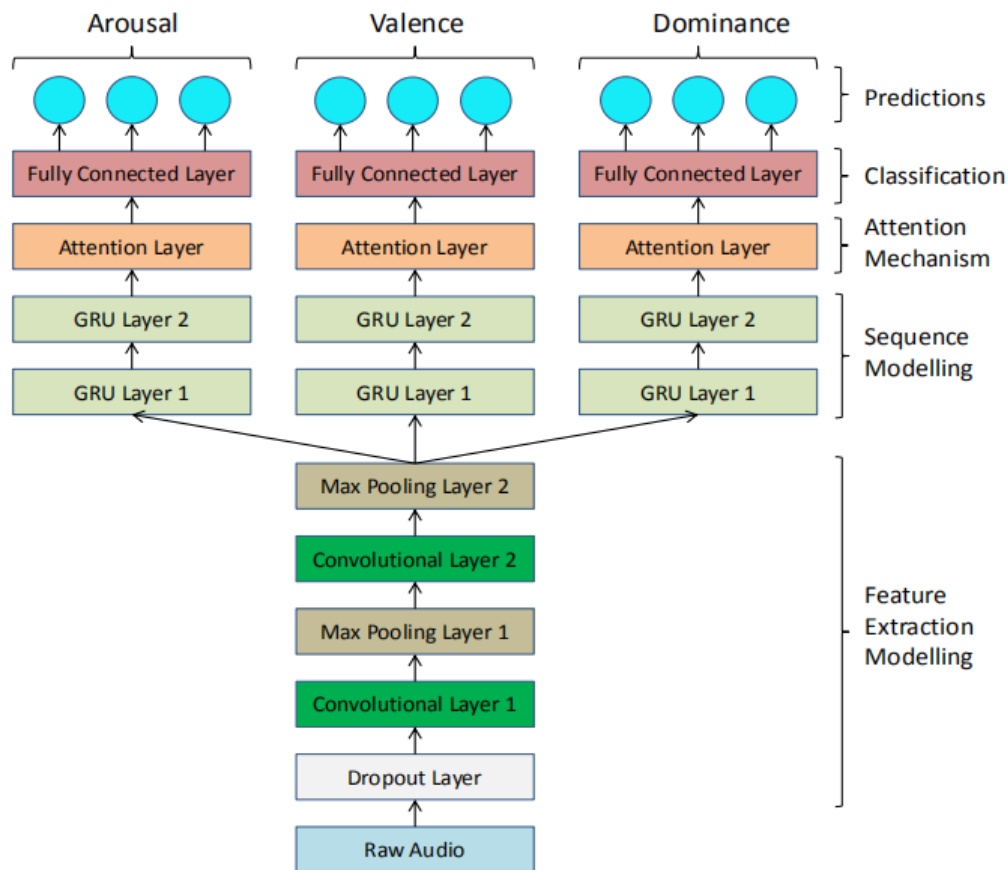


基于卷积神经网络和循环神经网络的注意力模型

# 语音情感识别方法

## ■ 多任务学习

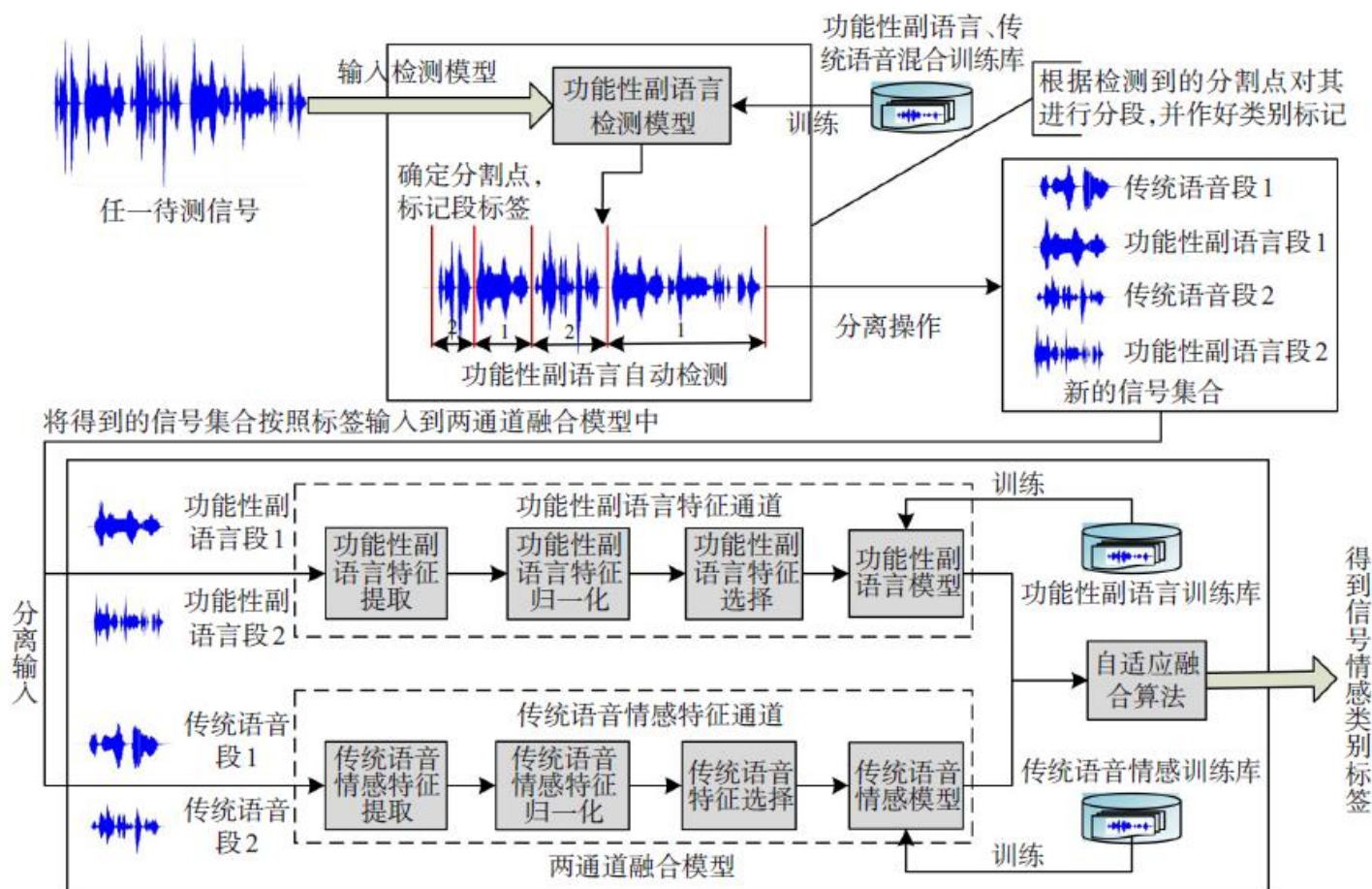
- 不同维度情感状态存在依存关系，利用多任务学习机制实现维度情感模型的协同优化



# 语音情感识别方法

## ■ 融合功能性副语言信息检测的识别模型

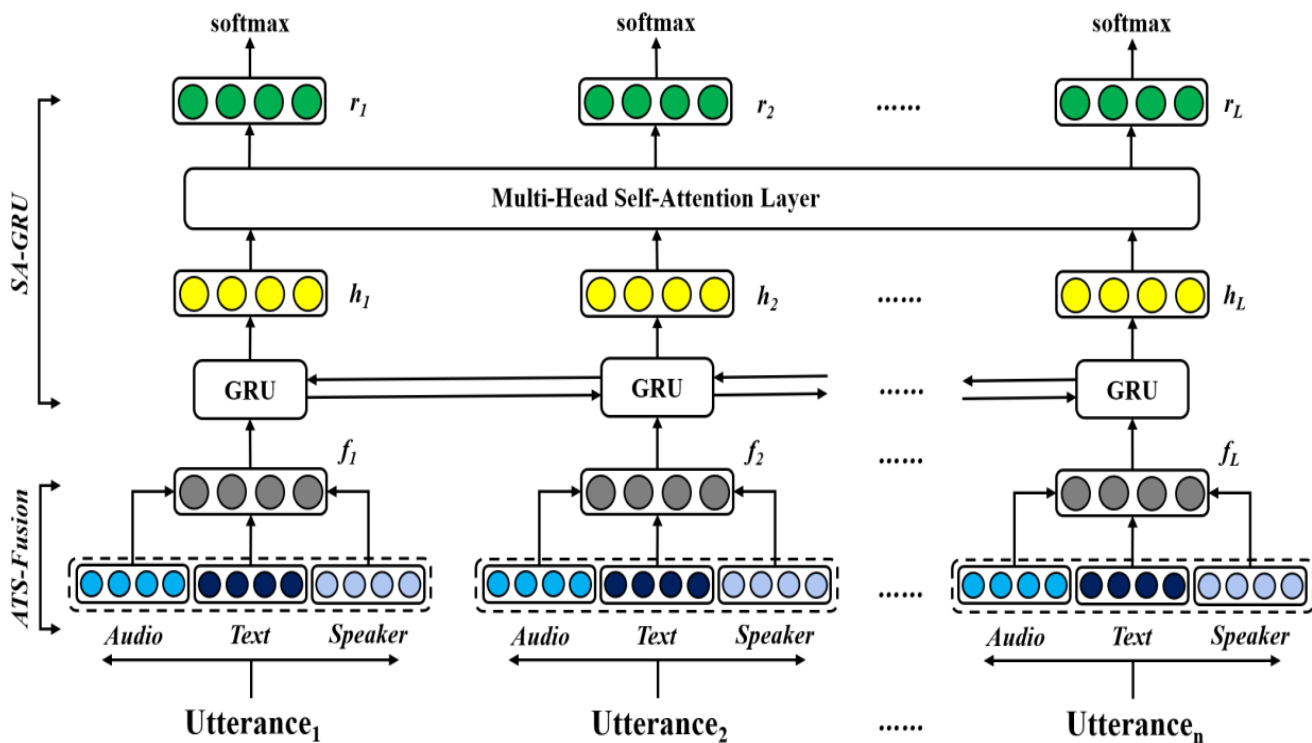
- 对副语言段和语音段进行区分性建模，有效利用副语言先验信息



# 语音情感识别方法

## ■ 融合说话人信息和文本信息的识别模型

■ 对音频信息、识别的文本信息、说话人信息进行时空融合

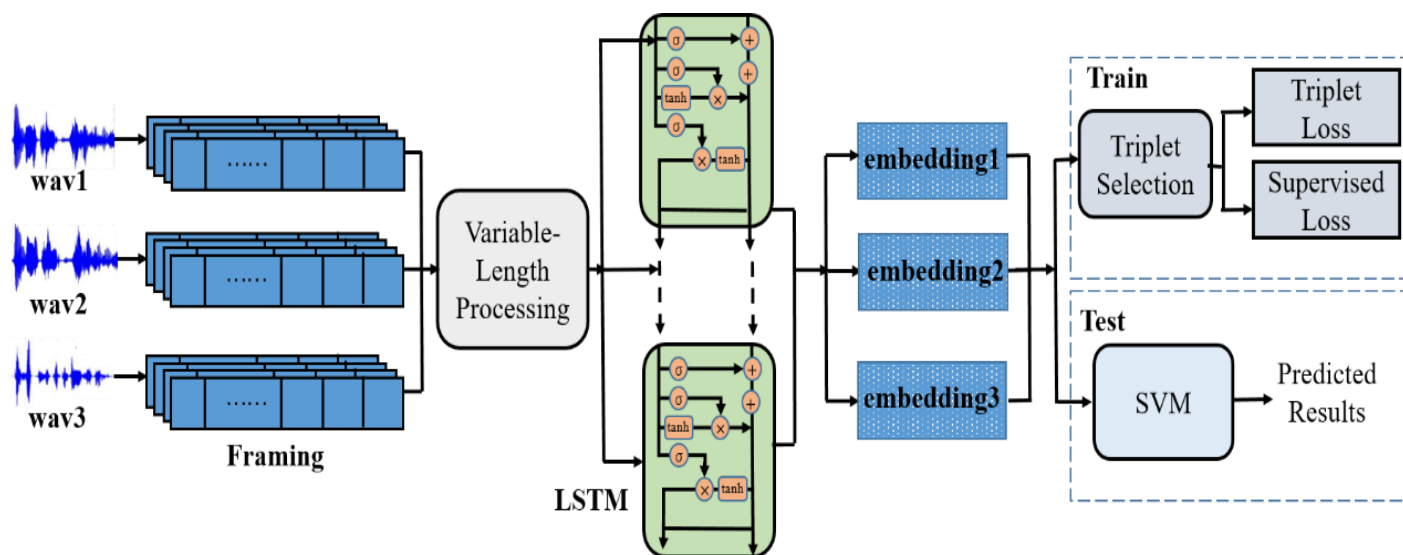




# 语音情感识别方法

## ■ 融合区分性训练准则的识别模型

- 利用三元损失函数解决不同情感状态边界模糊的问题



# 目录

---

- 背景及意义
- 研究现状与进展
- 音频情感数据库
- 语音情感识别
- 音乐情感识别
- 展望

# 音乐情感识别

---

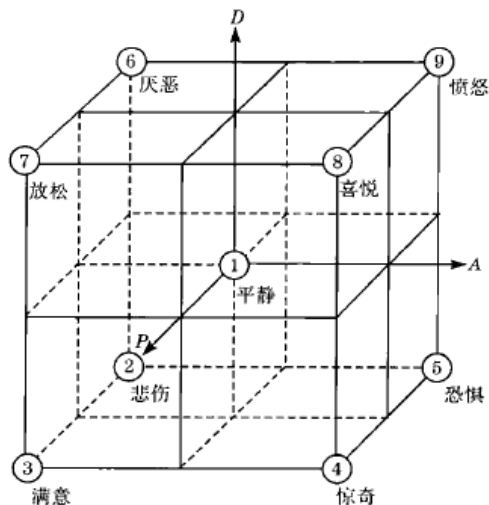
- 鉴于音乐表达和唤起情感的普遍共识，基于音乐情感属性来组织和检索音乐的需求是客观存在的
- 音乐情感自动识别是指，根据音乐的音频数据和其他相关信息构建计算模型，实现音乐情感自动判别的过程
- 已有十几年的历史

# 音乐情感模型

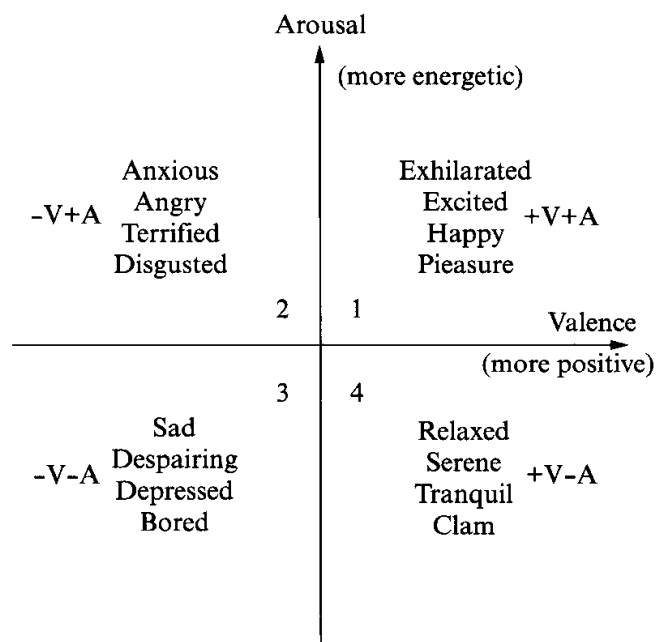
---

- 情感的三维空间模型在音乐情感识别中，使用比较多的是
- 通用连续维度情感模型
- 音乐表达情感离散类别模型
- 音乐唤起情感离散类别模型

# 音乐连续情感模型



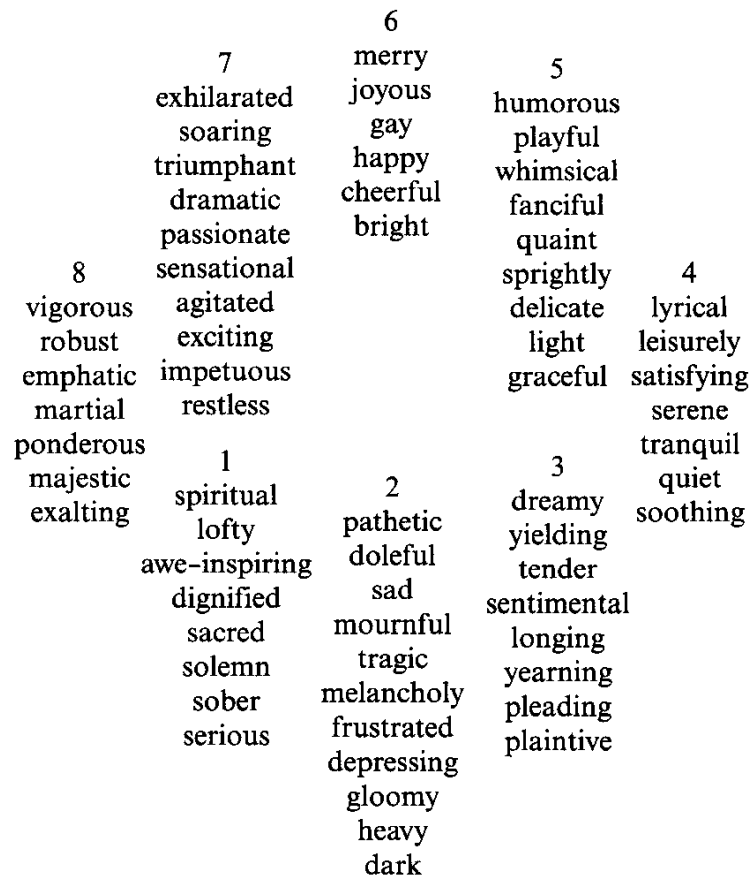
■ **PAD模型**：愉悦度Pleasure、激活度Arousal、优势度Dominance。



■ **VA模型**：Russell提出的Valence-Arousal情感模型作为音乐情感模型，情感状态是分布在一个包含 *Valence*（效价度）和 *Arousal*（激活度）。

# 音乐表达离散情感模型

- 1936年Hevner在“音乐中表达元素的实验研究”中提出的音乐情感离散类别模型。Hevner用67个情感形容词来描述音乐的情感空间，并且将这67个情感形容词分成8个类别：庄严的(dignified)、悲伤的(sad)、如梦的(dreamy)、宁静的(serene)、优雅的(graceful)、快乐的(happy)、激动的(exciting)、有力的(vigorous)



# 音乐情感模型

---

- **音乐唤起情感离散类别模型：**基于“唤起说”的音乐情感离散类别模型研究的主要问题是，选择哪些词来准确描述和区分音乐唤起的情感。EI内瓦情感音乐量表(the Geneva Emotional Music Scales, GEMS)被认为是第一个专门为度量音乐唤起的情感而设计的工具，是这方面研究的一个典型成果。
- GEMS-45包含45个情感标签，这45个情感状态又被分为9大类，即：wonder、transcendence、tenderness、nostalgia、peacefulness、power、joyful activation、tension、sadness(奇妙的、超越的、温柔的、怀旧的、歌舞升平的、强大的、快活的、紧张的、悲伤的)。相关实验表明，这些情感标签在描述音乐唤起的情感状态时，听众的选择具有一致性。

# 典型数据集

---

## ■ CAL500 (Computer Audio Lab) 数据集

- 情感离散类别模型：CAL500是一个包含500首西方流行音乐的公开数据集。采用了135个音乐相关的概念，涉及情感、曲风、乐器、场合和演唱特性等方面的174个语义关键词，对每首音乐进行标注，其中，情感相关的概念(关键词)有18个。

## ■ MIREX 2007 AMC (Audio Mood Classification) 数据集

- MIREX 2007 AMC数据集是2007年开始组织音乐情感分类算法评测活动的数据集，由600首音乐组成(均为30 S的音乐片段)，采用一种5类的类别模型来表示音乐情感，且情感类之间是互斥的。这个类别模型是通过对互联网音乐情感相关的社会标签做聚类分析而得来

## ■ MIREX 2013 K-POP Mood Classification数据集

- 从2013年开始，MIREX为音乐情感分类算法评测引入一个新的数据集。该数据集有1 437首韩国流行歌曲。采用的情感类别模型与MIREX 2007 AMC数据集所采用的模型相同，歌曲也是被分成互不重叠的5类。



# 典型数据集

---

## ■ MediaEval Emotion in Music任务数据集

- MediaEval Emotion in Music是一个动态(连续时间)音乐情感识别算法评测。这个评测所使用的数据集来源于Mohammad Soleymani等的研发成果。评测用数据集包含约1 744首音乐，均为45 S的片段。每段都标有一个段级的静态VA值和一组间隔为0. 5 S的动态VA值。

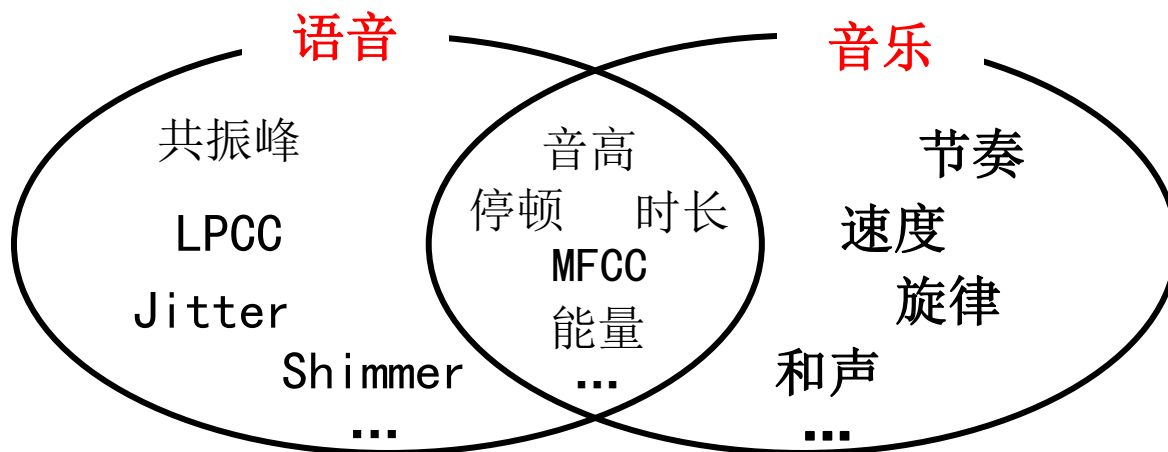
## ■ AMG1608数据集

- AMG1608数据集。包含1 608首当代西方音乐(均为30 S片段)。有665个标注者参与了标注。其中，46个标注者每人至少标注了150个片段。音乐情感模型采用VA维度模型，每个音乐片段只标注一个VA值。

# 音乐情感特征

## ■ 音乐情感特征

- 节奏 (Rhythm)
- 速度 (Tempo)
- 旋律 (Melody)
- 和声 (harmony)



# 音乐情感特征

---

## ■ 速度

- 指音乐速度的快慢
- 速度快的反映出高兴、生气、害怕
- 速度慢的反映出伤心、柔情、平静

## ■ 节奏

- 指音乐的抑扬顿挫、轻重缓急
- 规律性强的节奏反映出高兴、平静
- 规律性弱的节奏反映出生气

# 音乐情感特征

---

## ■ 旋律

- 乐音按一定的调式和节奏组织的序列
- 上升的旋律一般是高兴、惊讶
- 下降的旋律一般是伤心、厌恶

## ■ 和声

- 超过一个频率所组成的声音
- 悦耳的和声反映出高兴

# 音乐情感特征参数

---

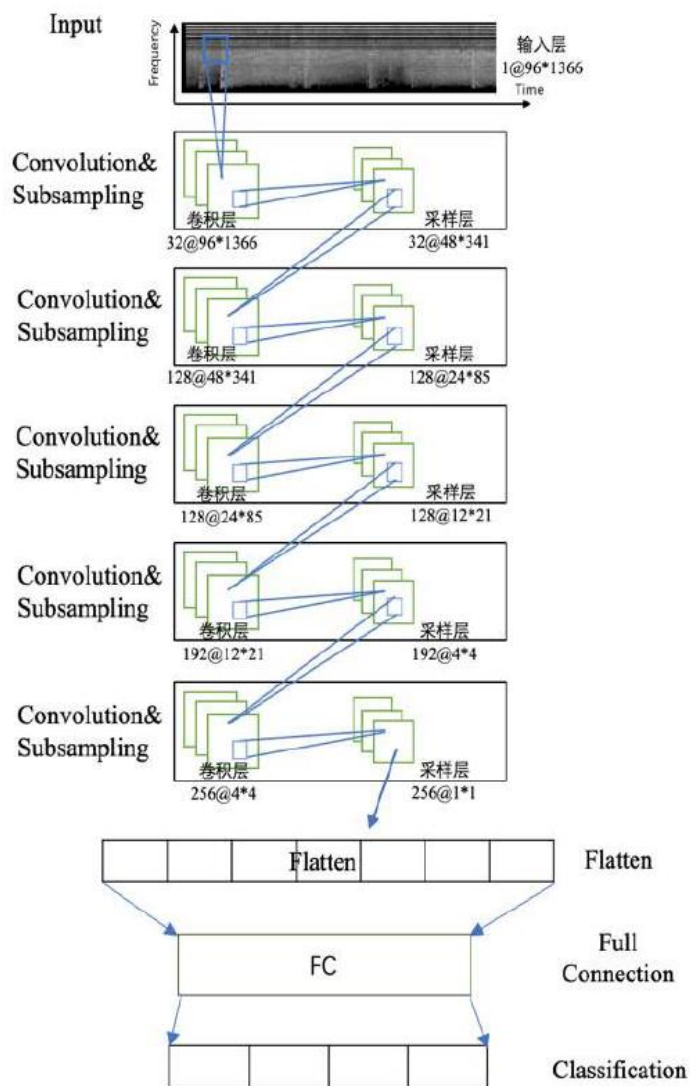
## ■ 从MIDI中获取

- 1) 节拍：是音乐中有规律地强拍和弱拍的反复；
- 2) 变化音的个数：所谓的变化音就是把固定的音升高或者降低，在MIDI文件中有相应的变音记号来表示，变化音会对乐曲造成冲突不和谐的感觉；
- 3) 最大音程：表示在整首乐曲中，音高最高的音符与音高最低的音符之间的音程差；
- 4) 音符密度：平均每小节包含的音符数；
- 5) 速度：每秒钟的音符数 (Note / s)；
- 6) 大和弦小节的比例：一般都认为大三和弦色彩明亮，而小三和弦情感色彩相对暗淡。

## ■ 从音频中获取

- 与语音情感特征类似

# 基于CNN的方法



训练次数	<i>accuracy</i>	<i>recall</i>	<i>precision</i>	<i>F1-score</i>
1	0.808	0.749	0.768	0.735
2	0.816	0.745	0.765	0.737
3	0.806	0.759	0.764	0.738
4	0.785	0.718	0.711	0.693
5	0.822	0.793	0.772	0.761
6	0.812	0.759	0.774	0.741
7	0.805	0.747	0.743	0.721
8	0.806	0.765	0.776	0.753
9	0.817	0.776	0.781	0.762
10	0.793	0.734	0.726	0.711
平均值	0.807	0.754	0.758	0.735

# 音乐情感识别的问题

---

- 与语音情感识别任务相比，音乐情感识别还处于初级阶段；
- 音乐本身是表达情感的，但这种情感是非常主观且难以量化的。音乐情感识别是个很困难的问题，主要是因为人的情感固有的模糊性；
- 音乐情感识别都依赖一个情感模型，但情感模型仍然是心理学研究的一个活跃课题；
- 音乐情感并不是完全包含在音频中。单靠音频数据本身，不能完全识别音乐情感；
- 基于音频的音乐情感识别是音乐信息检索研究中的一项长期目标。

# 目录

---

- 背景及意义
- 研究现状与进展
- 音频情感数据库
- 语音情感识别
- 音乐情感识别
- 展望



## ■ 情感语音语料

- 一个丰富、优质的情感语音数据库是开展语音情感计算研究的必要基础, 可以为研究工作提供可靠的训练。
- 情感语音数据的采集和整理工作非常困难, 进而导致了高质量的情感语料难以获取, 尤其是如何同时满足语料的自然度和情感的纯净度是其面临的最大挑战。
- 应该关注跨数据库的扩展性能的研究, 对不同民族之间和不同语种之间的情感表达的差异应该受到研究者的重视。

## ■ 语音情感特征

- 对于**情感语音的构成**进行进一步深入的分析，找出对于情感的表达有贡献的新的特征参数，并将其加入到识别参数中，以获得更高的识别率；
- 由于**语音情感变化引起语音的诸多特征发生变化**，将多种特征混合起来可以更全面地表示情感，多类特征组合是特征获取的一个研究方向。
- 在**高维情况下分类器的泛化性能反而会更弱**，需要进行针对性的情感声学特征降维和选择等方法的研究。

## ■ 语义的理解

- **利用语义的语音情感识别**：由于说话人表达情感有其特定的环境，这样在语音情感识别的研究中要考虑语义所具有的情感倾向性。
- 目前在汉语词汇中词义和情感之间的联系还没有得到研究，如何利用词义，将语音语义识别同情感识别相结合，在更高层次上把握说话人的情感是一个重要研究课题。

---

# 讨论&问答