如何从语音中判断人的情感

人类的情感是很玄妙的东西。早在两千年前，亚里士多德就意识到情感在社会交流中的作用。比如，他认为一个在合适场合能有点儿脾气的人值得敬重，没脾气的反而是傻瓜。换句话说，情感本身就是一种智能。Rosalind Picard 的《Affective Computing》一书是情感计算的起点。Picard 认为，如果我们希望计算机拥有通用智能并实现和人的自然交互，那么必须让它具有“识别”、“理解”，甚至“拥有”和“表达”情感的能力。到目前为止，让机器拥有情感这件事还很困难，但在交互中识别出人的情感还是有可能的。

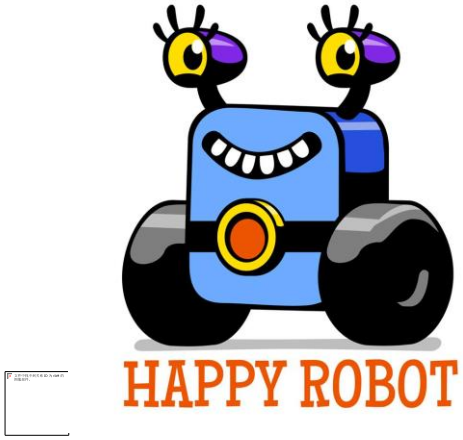

图 1：带有喜感的机器人[1]

目前对情感类别的定义并不统一。比较有代表性的是 Ekman 的离散情感理论，这一理论认为人类有六种基础情绪 [220, 221]：愤怒(Anger)、恐惧(Fear)、 厌恶(Disgust)、惊讶(Surprise)、快乐(Happiness)、悲伤(Sadness)。这些基础情绪是与生俱 来且与人种、文化无关的。基于这些基础情感，通过一定的比例互相混合可以派生出其它各种情感[3]。

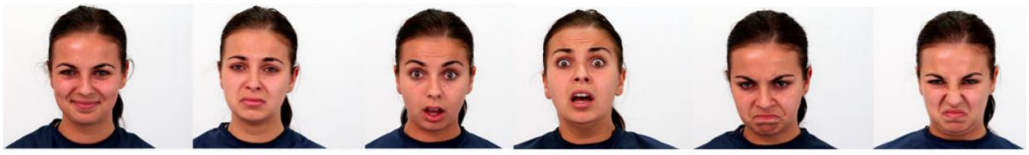

图 2：Ekman 定义的六种情感[3]

人的情感有多种表现方式，包括语言，声音，脸部表情，肢体动作等，都是情感的表达渠道。在这些表达渠道中，声音表达情感灵活自然，因此受到广泛关注。那么，如何让机器通过声音识别情感呢？

传统识别方法基于特征提取+统计建模的基础框架。首先从声音中提取和情感有关的特征，如发音能量、基频、共振峰位置、语速和停顿等。一些语音质量的变化，如压力感、沙哑、喘息、基频上的抖动等，也是出现极端情绪的表征。有了这些基础特征，通常还需要对这些特征进行统计，以确定在一句话中这些特征的分布情况，包括均值、方差等。有了这些统计量，就可以建立分类模型对情感进行预测。常用的统计模型包括高斯混合模型（GMM）、隐马尔可夫模型（HMM），支持向量机（SVM），神经网络（NN）等[4]。

深度学习兴起后，基于深度神经网络的情感识别方法得到广泛应用。图 3 是一个基于卷积神经网络（CNN）的情感识别系统，输入为一段语音信号，经过若干卷积和池化操作后输出情感类别。实验结果表明，在一个包括 10 位发音人 5 种情感的数据库上，这一方法可得到 40%的正确率。

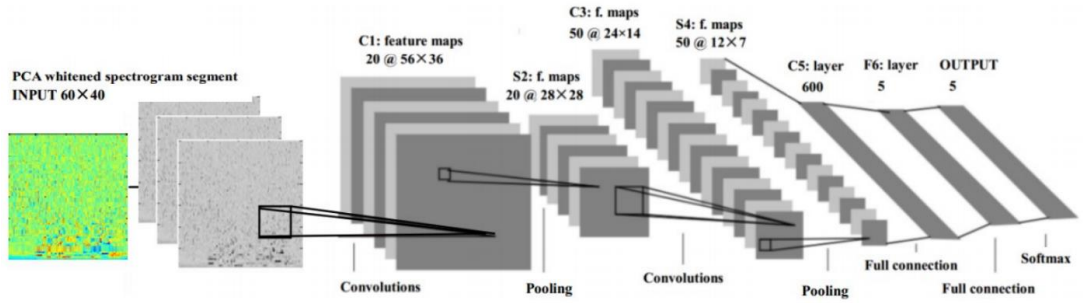

图 3：基于深度卷积网络的 5 种情感识别系统[2]

尽管取得了一定进步，情感识别依然是很困难的任务。一方面情感的定义本身就比较模糊，对一个人而言愤怒或悲伤的声音对另一个人也许感觉不到什么；同时，获取情感数据目前还比较困难，大部分由专业演员模拟生成，真实性有待提高。现在人们倾向认为，如果综合利用视频、音频和发音内容信息，有望显著提高情感识别的性能。

1. http://www.happyrobotgames.com/
2. WQ Zheng, JS Yu, and YX Zou. “An experimental study of speech emotion recognition based on deep convolutional neural networks”， 2015.
3. Edward R Morrison, Paul H Morris, and Kim A Bard. “The stability of facial attractiveness: is it what you’ve got or what you do with it?” In: Journal of Nonverbal Behavior 37.2 (2013), pages 59－67
4. 汤志远等，《语音识别基本法》，2021.2.