语音情绪识别

2019/1/13

如今随着人工智能技术的发展，计算机变得越来越“健谈”：Alexa，Cortana，Siri 以及更多的对话系统走进人们的生活，但是他们中几乎没有一个能注意到我们的情绪，并且像我们人类自己在聊天时对伙伴作出反应。而如果我们有一个足够出色的语音情绪识别系统（speech emotion recognition，SER）的话，那么情况就会好很多。

1. 什么是语音情绪识别?

人类的语音中包含了许多信息，其中就包括前文所述的一个人想要通过语音传递的语义信息、语音所属的说话人的身份信息、说话人所使用的语言信息，以及说话人的情绪信息。语音情绪识别是指通过计算机自动的识别说话人所说的语言中含有怎样的情绪。

语音中的情绪信息是反应人类情绪的一个十分重要的行为信号，同时识别语音中所包含的情绪信息是实现自然人机交互的重要一环。同样的一条语音内容，以不同的情绪说出来，其所携带的语义可能完全不同，只有计算机同时识别出语音的内容以及语音所携带的情绪，我们才能准确的理解语音的语义，因此理解语音的情绪才能让人机交互更为自然和流利。

人类之所以能够通过聆听语音捕捉对方情绪状态的变化，是因为人脑具备了感知和理解语音信号中的能够反映说话人情绪状态的信息(如特殊的语气词、语调的变化等)的能力。自动语音情绪识别则是计算机对人类上述情绪感知和理解过程的模拟，它的任务就是从采集到的语音信号中提取表达情绪的声学特征，并找出这些声学特征与人类情绪的映射关系。

1.1. 语音情绪的分类表示

基于语音信号的情绪识别在近几年得到了广泛的关注和研究。但对于情绪的分类，研究者们没有统一的标准，现阶段基于语音信号的情绪识别主要分为两大类：离散情绪和维度情绪，分类的依据是对情绪的不同表示方式。第一种表示方式是情绪的种类，大多数研究者认为人类具有六种离散的基本情绪，包括开心(happiness)，难过(sadness)，生气(anger)，厌恶(disgust)，害怕(fear)，惊讶(surprise)，如 Figure1 所示，但语音的情绪识别研究大多采用的是快乐、悲伤、愤怒和中性这四种区分度大的情绪；

Figure 1：六种基本情绪，图片源自[1]

不同于将情绪标识为离散的情绪类别，另有一些学者尝试用连续的维度来表示情绪，其中最著名的、也最为广大学者所接受的是唤醒度—愉悦度一控制度(Valence-Arousal-Power)三维情绪模型[2]。这种用维度描述情绪的思路与将情绪分类的思路并不矛盾。事实上，任何一种情绪类别都可以用这些维度来表示，从而在情绪空间中为其定位。例如图 2 所示。唤醒度（Arousal）代表情绪唤起程度的高低，愉悦度（valence）代表积极情绪的高低，这两个维度都可以通过数值来代表它的高低程度。比如 figure 2 中的数值区间[1,9]，1 代表非常低迷/消极，9 代表非常激动/积极。 这样，开心(happy)就可以用高唤醒度和高愉悦度来表示，而难过（sad）则可以用低唤醒度和低愉悦度来表示。几乎人类所有的情绪都可以用这两个维度所构成的空间来表示。

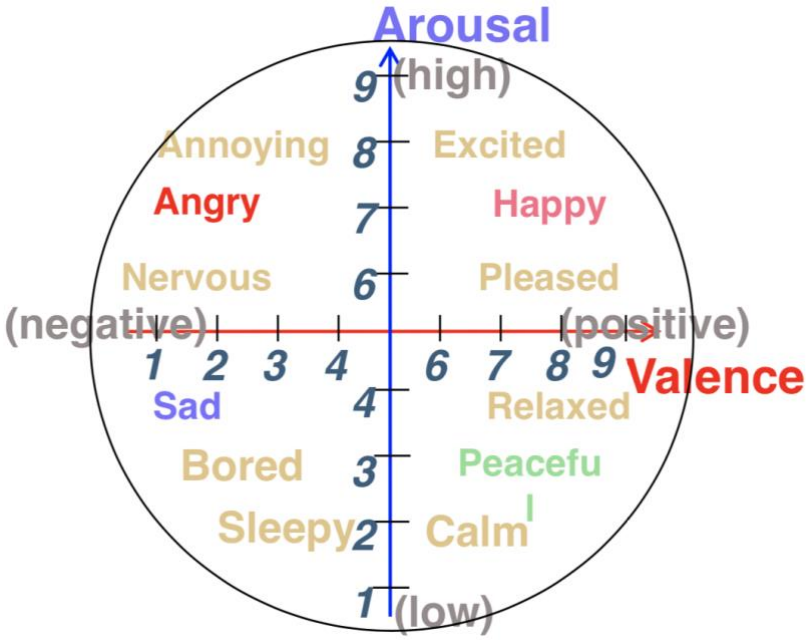

Figure 2：维度情绪模型，图片源自[3]

表 1 描述了上述两种表示方法区别，有研究比较发现[5]，不管采用哪种分类方法,最终的情绪识别率相近。其中对愤怒和悲伤这两种情绪的识别率最高。

Table 1：两种情绪分类方式

考察点	离散情绪描述模型	维度情绪描述模型
情绪描述方式	形容词标签	笛卡尔空间中的坐标点
情绪描述能力	有限的几个情绪类别	任意情绪类别
被应用到语音情绪识别领域的时期	1980s	2000s
优点	简洁、易懂、容易着手	无限的情绪描述能力
缺点	单一、有限的情绪描述能力无法满足对自发情绪的描述	将主观情绪量化为客观实数值的过程，是一个繁重且难以保证质量的过程

2. 语音情绪的语料设计

语音情绪识别研究的进展与一个优秀的情绪语音数据库有很大关系。情绪语音库的质量高低，直接影响了由其训练的语音情绪识别系统的性能，如果使用低质量的数据库，可能会得到不正确的结论。所以，设计一个情绪语音数据库要考虑以下两个因素。

2.1.情绪的自然度

所谓情绪的自然度就是表现出来的情绪和日常交流中自然产生的情绪之间的相似度,它对于情绪语音的基础研究和应用都具有重要影响。按照自然度的不同，可以把情绪语料大体分为三类，下面一一介绍这三类语料的获取方法及其优缺点。

2.1.1.自然情绪语料

自然情绪语料是指从自然生活中采集的、经过人工筛选的语料。这类语料是最直接、真实、可信的语料，具有最高的自然度。语料的获取方法，一般是在被录音者毫不知情的情况下进行录音，因此所录得的语音自然度最高。但这种录音方式的缺点是操作性差且涉及到隐私相关的法律问题。另一个存在的问题是，采集到的自然情绪语料是需要人为标注的，这样才能给识别系统使用，但由前面的介绍我们可以得知，情绪的分类本身存在争议，而自然状态下采集到的语料，标注人对语料的标注具有较强的主观性，不同的标注者可能对同一段语音有不同的看法，所以对自然情绪语料进行分类标注是件不太容易的事。[4][5][6]

2.1.2.模拟情绪语料

尽管自然情绪语料是最理想的，但考虑到获取方面的困难，在情绪识别领域，许多研究者选择邀请一些专业的播音员或善于表达情绪的人，进行情绪模仿录制语料，也就是让录音者模仿不同的情绪朗读指定的内容。这种方法是最为常见的语料获取方法，大多数情绪语音识别研究中所使用的语料都是采用这种方法录制的，因为这种方法有两个突出的优点：一是可操作性强，只需要一些简单的录音设备，再找一个安静的录音环境即可，可以在短时间内获取；其二，这样录制的语料符合情绪要求，可区分性强。

通过这种方法获取的语料的就是模拟情绪语料，是录音者伪装出来的，情绪的表达不受心理活动刺激，情绪的自然度完全取决于录音者的模仿能力。因此与自然情绪语料相比，模拟情绪语料中的情绪成分通常被夸大，并不能体现真实的情绪。一些实验表明，使用模拟情绪语料的语音情绪识别实验通常能获得较高的识别率，但是模拟情绪语料并不能完全代表自然情绪语音。情绪语音的自然程度和它的获取目前来讲是一对不容易调和的矛盾。[4][5][6]

2.1.3.诱导情绪语料

获取这类语料时，研究者通常会让录音者置于恰当的环境之中，诱导录音者产生某种情绪，完成录音，这种方法的可操作性也比较强，且能获得比模拟情绪语料更多的自然度，但这种方法无法确保录音者在设置的环境里出现诱导的情绪。

从上面的分析中我们不难看出语料的情绪自然度受较多因素影响，很难选出最完美的预料，语料的选择因研究方法和目标不同而变化。[4][5][6]

2.2.上下文相关

人类的情绪表达不仅仅体现在声学信号里，同时也体现在上下文的内容上，这里所指的上下文相关是说情绪还可以通过面部表情，甚至是一些身体语言的表现出来，所以如果能够建立一个包含多个维度的情绪上下文信息的语料库对情绪识别这一任务而言是更有帮助的，但由于客观条件的限制，做到这一点十分困难。[6]

3. 语音情绪的特征提取

在语音情绪识别系统里提取出合适的且有效的语音情绪特征是十分重要的事。一般来讲，语音情绪特征的提取有以下两个主要的问题：

1) 特征提取的作用域。一些研究人员认为应该先将语音分帧再进行特征提取，即提取局部特征；而另外一些研究人员更倾向于将整句语音的全部特征直接抽取出来。

2) 提取什么特征作为语音情绪识别任务的主要特征？例如韵律特征、基于频谱的特征、声音质量特征等。

3.1. 局部特征与全局特征

从每一帧中提取诸如音调和能量之类的韵律特征称为局部特征，而全局特征是指从一句话中提取的所有语音特征的统计结果。现阶段的相关研究表明全局特征在分类的准确度上往往比局部特征表现的要好，同时耗时也更少（特征量较少），然而全局特征也有许多缺点：

1) 全局特征只在激活度（arousal）相差较多的情绪中比较有效，而当激活度相差较小时，比如在分类 anger 和 joy 时，二者的激活度都较高，相差不多，这时全局特征就会失效；

2) 全局特征会丢失语音的短时信息。由于提取的特征不是在帧层级特征，所以短时的情绪信息容易丢失。

3) 当使用较为复杂的分类器（HMM，SVM 等）时，全局变量会因为特征较少而无法进行有效的训练。这时如果在复杂的模型中使用局部特征，模型的效果会更好。

还有一种做法是对语音信号根据音素进行分段而不是分帧。研究显示了把分段的特征和全局特征相结合可以一定程度提高是别的准确率。[5]

3.2. 提取什么特征?

1) 韵律学特征：韵律是指语音中凌驾于语义符号之上的音高、音长、快慢和轻重等方面的变化，是对语音流表达方式的一种结构性安排。它的存在与否并不影响我们对字、词、句的听辨，却决定着一句话是否听起来自然顺耳、抑扬顿挫。韵律学特征又被称为“超音段特征”或“超语言学特征”，它的情绪区分能力已得到语音情绪识别领域研究者们的广泛认可，使用非常普遍[7]，其中最为常用的韵律特征有时长(duration)、基频(pitch)、能量(energy)等。

2) 基于频谱的相关特征：被认为是声道(vocal tract)形状变化和发声运动(articulator movement)之间相关性的体现，已在包括语音识别、话者识别等在内的语音信号处理领域有着成功的运用。Nwe 等人[8]通过对情绪语音的相关谱特征进行研究发现，语音中的情绪内容对频谱能量在各个频谱区间的分布有着明显的影响。例如,表达高兴情绪的语音在高频段表现出高能量，而表达悲伤的语音在同样的频段却表现出差别明显的低能量。近年来，有越来越多的研究者们将谱相关特征运用到语音情绪的识别中来[8, 9, 10]，并起到了改善系统识别性能的作用，相关谱特征的情绪区分能力是不可忽视的。在语音情绪识别任务中使用的线性谱特征(linear-based spectral feature)一般有：LPC(linear predictor coefficient)[11]，LFPC(log-frequency power coefficient)[8]等；倒谱特征(cepstral-based spectral feature)一般有：LPCC(linear predictor cepstral coefficient)，OSALPCC(cepstral-based OSALPC)[9]，MFCC(mel-frequency cepstral coefficient)等。

3) 声音质量是人们赋予语音的一种主观评价指标，用于衡量语音是否纯净、清晰、容易辨识等[12]。对声音质量产生影响的声学表现有喘息、颤音、哽咽等，并且常常出现在说话者情绪激动、难以抑制的情形之下。语音情绪的听辨实验中，声音质量的变化被听辨者们一致认定为与语音情绪的表达有着密切的关系[12]。在语音情绪识别研究中,用于衡量声音质量的声学特征一般有：共振峰频率及其带宽(format frequency and

bandwidth)、频率微扰和振幅微扰(jitter and shimmer)[50]、声门参数(glottal parameter)等。

4) I-vetor 特征[13]。i-vector 在近些年来的说话人识别领域有着广泛的应用，是一项将高维高斯混合模型(Gaussian mixture models，GMM)超向量空间映射到低维总变异空间的技术，然而在语音情绪识别领域的应用还较为新颖。文献[14]提出使用串联结构的情绪 i-vector 特征用于语音情绪的识别，他们首先使用 openSMILE 提取出 1584 维的声学特征,并使用这些特征为自然情绪状态的语音训练得到一个通用背景模型（universal background model），然后在该通用模型的基础上为每类情绪状态生成各自的 GMM,继而得到每类情绪状态的 GMM 超向量用于提取 i-vector。最后，对应于各个情绪状态的 i-vector 被串连在一起作为支持向量机的输入，用于 angry, happy, neutral, sad 这 4 类语音情绪的识别，取得了优于原始 1584 维声学特征的识别性能。

4. 语音情绪的特征分类模型

在构建语音情绪识别系统时，第 2 节介绍过的两种不同的情绪表示方式造就了两种不同的分类模型。第一种识别情绪种类的系统是基于离散的情绪表示的模型，第二种识别情绪的系统是基于维度的一个回归系统，因为系统的输出是一个连续性的数字。这两种系统都是从声音信号里面提取出与情绪相关联的一系列特征向量。然后这些与情绪相关的特征向量会被用来训练分类器或者回归系统。

4.1. 离散情绪模型

目前大多数的研究是在离散语音情绪模型上展开的，研究者们提出了许多分类模型，常用的语音情绪识别领域的分类器，包括线性的有：Naive Bayes Classifier，Linear ANN(artificial neural network)，Linear SVM(support vector machine)等；非线性的有：Decision Trees，k-NN(k-nearest neighbor algorithm)，Non-linear ANN，Non-linear SVM，GMM (Gaussian mixture model)，HMM (hidden Markov model)以及 DNN（deep neural network）等，其中应用最为广泛有 HMM, GMM, SVM 以及 DNN。

Nwe 等人[8]使用基于 HMM 的识别器用于 6 类情绪的识别。其中 LFPC，MFCC 和 LPCC 被用作情绪特征，为每个话者的每类情绪构建一个四状态、全连接的 HMM，一个缅甸语语料库和一个汉语普通话语料库被分别用于 HMM 的训练和测试，系统识别的准确率分别可达到 78.5%和 75.5%。Lee 等人[15]分别以情绪类别和音素类别为单位建立 HMM 模型,并在说话人不相关的情形下对模型性能进行测试。实验结果表明，基于音素类别的 HMM 模型具有更优的表现。

GMM 是一种用于密度估计的概率模型，可以被看作是只包含一个状态的连续 HMM 模型。文献[16]中，GMM 分类器被用于对面向婴儿的(infant-directed)KISMET 数据库进行情绪分类，并使用一种基于峰态模型（kurtosis-based）的选择策略[17]对 Gaussian 成分的数量进行优化，由基频和能量的相关特征训练得到 GMM 模型识别准确率可达到 78.77%。Tang 等人[18]针对语音情绪识别构造了一种使用 Boosting 算法进行类条件分布估计的 GMM 模型，并称其为 Boosted-GMM，与传统的使用 EM(expectation maximization)方法进行分布估计的 EM-GMM 相比, Boosted-GMM 表现出更优的性能。

SVM 分类器的关键在于核函数的运用，它负责将原始特征以非线性的方式映射到高维空间中，从而提高数据的可分性。SVM 在语音情绪识别领域有着广泛的应用，这里以文献[19]为例进行说明.文中共有 3 种策略被用来构建基于二分类 SVM 的多分类模型：前两种策略中都首先为每类情绪构建一个二分类的 SVM,不同的是,第 1 种策略将待识别语句分配给距离其余情绪距离最远的情绪类型，而第 2 种策略则将各个二分类 SVM 的输出作为一个 3 层 MLP 网络的输入,通过进一步的计算做出最终的分配决定；第 3 种策略被称为多层次的分类模型(hierarchical classification model),各个 SVM 子分类器按照树形结构进行排列，从根节点开始由粗略到细致地实现情绪的逐步划分，在叶节点处给出最终识别结果。实验结果表明：在 FERMUS III 数据库[19]之上，3 种策略的识别率分别为 76.12%，75.45%和 81.29%，第 3 种策略表现最优。

语音情绪识别问题的一个难点就在于不清楚哪些特征对于情绪的识别更有效，而深度神经网络的出现恰好能对这一问题提供更多帮助，因为 DNN 模型具有极强特征选择能力，Kun Han 等人[20]首次提出用 DNN 模型来模拟每一句语音的情绪状态概率分布，并将通过 DNN 模型提取出来的句子层级的情绪语音特征送入一个极端学习机器（extreme learning machine，ELM）中来对语音进行分类，DNN 方法的流程如下图所示，实验结果表明 DNN 的方法可以取得比上面提到的传统方法更好的性能。

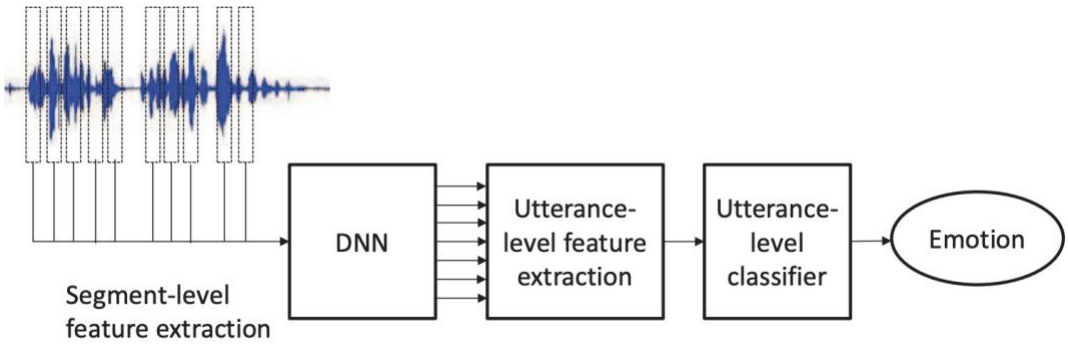

Figure 3 DNN 语音情绪识别流程，图片来自[20]

4.2.维度情绪模型

相比于离散语音情绪识别，基于维度情绪的模型较为新兴，但也已得到领域内研究者们越来越多的关注[21,22]。该方法一般被建模为标准的回归预测问题，即使用回归预测算法对情绪属性值进行估计，在当前的维度语音情绪识别领域使用较多的预测算法有:Linear Regression, k-NN, ANN, SVR(support vector regression)等。其中,SVR 因为性能稳定、训练时间短等优点应用得最为广泛。例如,Grimm 等人[21]在 VAM 数据库上对 k-NN 和 SVR 等方法，基于一个三维情绪属性模型的预测能力进行比较，结果表明，SVR 的预测能力更胜一筹.我们可以看出：相比离散情绪分类器的繁荣发展，维度情绪预测算法的研究较为薄弱，更多针对情绪识别任务的有效算法仍有待探索。

5. 小结

本节介绍了语音情绪识别的概念；语音情绪的分类表示，语音的情绪通常有两种表示方法，一种是离散的语音情绪表示，这种方法通常将语音分成基本的几个种类，另一种是基于维度的语音情绪表示方法，这种方法通常将语音的情绪划分为三个维度，这对于语音情绪的描述更加细致且连续；紧接着介绍了语音情绪特征数据库的设计通常需要考虑的一些问题，包括情绪的自然度以及上下文的内容；随后介绍了语音特征的抽取，具体说明了特征抽取的作用域和抽取什么特征；最后本文着重介绍了传统的基于离散语音情绪表示的分类模型有哪些，以及识别系统的性能。

References

[1] http://news.softpedia.com/news/Is-Emotion-Tracking-the-Future-in-Tech-or-Just-Down-Right-Creepy-434806.shtml

[2] Cowie, Roddy, et al. "Emotion recognition in human-computer interaction." *IEEE Signal processing magazine* 18.1 (2001): 32-80.

[3] Jiang, Bihan, et al. "A dynamic appearance descriptor approach to facial actions temporal modeling." *IEEE Trans. Cybernetics* 44.2 (2014): 161-174.

[4] Ververidis, Dimitrios, and Constantine Kotropoulos. "Emotional speech recognition: Resources, features, and methods." *Speech communication* 48.9 (2006): 1162-1181.

[5] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." *Pattern Recognition*44.3 (2011): 572-587.

[6] Ververidis, Dimitrios, and Constantine Kotropoulos. "A state of the art review on emotional speech databases." *Proceedings of 1st Richmedia Conference*. 2003.

[7] Lee, Chul Min, and Shrikanth S. Narayanan. "Toward detecting emotions in spoken dialogs." *IEEE transactions on speech and audio processing* 13.2 (2005): 293-303.

[8] Nwe, Tin Lay, Say Wei Foo, and Liyanage C. De Silva. "Speech emotion recognition using hidden Markov models." *Speech communication* 41.4 (2003): 603-623.

[9] Bou-Ghazale, Sahar E., and John HL Hansen. "A comparative study of traditional and newly proposed features for recognition of speech under stress." *IEEE Transactions on speech and audio processing* 8.4 (2000): 429-442.

[10] Wu, Siqing, Tiago H. Falk, and Wai-Yip Chan. "Automatic speech emotion recognition using modulation spectral features." *Speech communication* 53.5 (2011): 768-785.

[11] Rabiner, Lawrence R., and Ronald W. Schafer. *Digital processing of speech signals*. Vol. 100. Englewood Cliffs, NJ: Prentice-hall, 1978.

[12] Gobl, Christer, and A. N. Chasaide. "The role of voice quality in communicating emotion, mood and attitude. " *Speech Communication*40.1-2(2003):189-212.

[13] Dehak, Najim, et al. "Front-End Factor Analysis for Speaker Verification." *IEEE Transactions on Audio Speech & Language Processing*19.4(2011):788-798.

[14] Xia, Rui, and Yang Liu. "Using i-vector space model for emotion recognition." *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[15] Lee, Chul Min, et al. "Emotion recognition based on phoneme classes." *Eighth International Conference on Spoken Language Processing*. 2004.

[16] Breazeal, Cynthia, and Lijin Aryananda. "Recognition of affective communicative intent in

robot-directed speech." *Autonomous robots* 12.1 (2002): 83-104.

[17] Vlassis, Nikos, and Aristidis Likas. "A kurtosis-based dynamic approach to Gaussian mixture modeling." *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 29.4 (1999): 393-399.

[18] Tang, Hao, et al. "Emotion recognition from speech via boosted gaussian mixture models." *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 2009.

[19] Schuller, Björn. "Towards intuitive speech interaction by the integration of emotional aspects." *Systems, Man and Cybernetics, 2002 IEEE International Conference on*. Vol. 6. IEEE, 2002.

[20] Han, Kun, Dong Yu, and Ivan Tashev. "Speech emotion recognition using deep neural network and extreme learning machine." *Fifteenth annual conference of the international speech communication association*. 2014.

[21] Grimm, Michael, Kristian Kroschel, and Shrikanth Narayanan. "Support vector regression for automatic recognition of spontaneous emotions in speech." *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. Vol. 4. IEEE, 2007.

[22] Karadoğan, Seliz Gülsen, and Jan Larsen. "Combining semantic and acoustic features for valence and arousal recognition in speech." *Cognitive Information Processing (CIP)*, 2012 3rd International Workshop on. IEEE, 2012.