计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(3):146−151 [doi: 10.15888/j.cnki.csa.006794]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于轨迹的时空光谱特征语音情感识别算法①

朱艺伟, 宋泊东, 张立臣

(广东工业大学 计算机学院, 广州 510006)
通讯作者: 朱艺伟, E-mail: 757873403@qq.com

摘　要: 语音识别领域的发展日新月异. 同时,现有的研究结果表明声学特性集中存在较多的互补信息. 本文提出了一种基于轨迹的空间-时间谱特语音情感识别方法. 其核心思想是从语音频谱图中获得空间和时间上的描述符, 进行分类和维度情感识别. 本方法采用了穷举特征提取的实验表明: 与 MFCCs 和基频等特征提取方法相比, 提出的方法在噪声条件下, 更具鲁棒性. 通过在 4 类情感识别实验中获得了可比较的非加权平均回馈, 得到了较为准确的结果, 语音激活检测方面也具有显著的改进.

关键词: 情感识别; 语音处理; 时空描述符; 融合; 特征提取

引用格式: 朱艺伟,宋泊东,张立臣.基于轨迹的时空光谱特征语音情感识别算法.计算机系统应用,2019,28(3):146–151. http://www.c-s-a.org.cn/1003-3254/6794.html

Speech Emotion Recognition Based on Space Time Spectrum Trajectory Feature

ZHU Yi-Wei, SONG Bo-Dong, ZHANG Li-Chen

(Faculty of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: The development of speech recognition is changing with each passing day. At the same time, the existing research results show that there is more complementary information in acoustic characteristics. In this study, a trajectory based spatio temporal spectral speech emotion recognition method is proposed. Its core idea is to get spatial and temporal descriptors from the speech spectrum, classify and identify dimensional emotion. The experiment using the exhaustive feature extraction shows that the proposed method is more robust in the noise condition than the MFCCs and the fundamental frequency extraction methods. In the 4 classes of emotion recognition experiments, the comparison of non weighted average feedback is obtained, and more accurate results are obtained. And, the voice activation detection is also improved significantly.

Key words: emotion recognition; speech processing; spatial-temporal descriptors; mel-filter bank energy; feature extraction

引言

在过去的十年中, 情感计算的研究蓬勃发展, 已经开始使机器能够感知和具有情感表达行为[1]. 其技术广泛应用于人机界面[2]和交互式机器人设计[3]领域, 甚至是新兴的交叉研究领域, 如社会信号处理[4]和行为信号处理[5]等. 作为人类交流的自然编码信息, 语音可以反映人类信息[6], 例如: 情感、性别、年龄及人格等等. 因此, 开发语音情感识别算法, 仍然是一个流行的话题.

目前, 国内外在情感识别建模语音声学方面进行了大量的研究, 比如: 底层特征工程、机器学习算法、甚至是联合特征标签表示[7]. 这些研究大多数都依赖于提取一组常用的短时间特征 (声学低层描述符——LLDs),

① 基金项目: 国家自然科学基金 (61572142); 广东省自然科学基金 (2015A030313490)
Foundation item: National Natural Science Foundation of China (61572142); Natural Science Foundation of Guangdong Province (2015A030313490)
收稿时间: 2018-08-23; 修改时间: 2018-09-20, 2018-09-30; 采用时间: 2018-10-08; csa 在线出版时间: 2019-02-22

例如: 这些特征可以是相关的光谱特征 (如 MFCCs)、韵律特征 (如音高语调)、语音质量 (如抖动)、低能量算子等[8], 然后选择情感识别框架. 例如: 支持向量机[9]或深层神经网络[10]. 或者利用时间序列模型将短语音低级描述符特征的时间性特征纳入到表达水平的情感识别中. 如隐马尔科夫模型[11]. 有一些研究利用听觉感知激发的调制光谱轨迹的时间特征[12], 用于情感识别. 基于上述研究成果, 本文提出了一种基于轨迹的视频描述符提取方法. 该方法将音频文件本质上视为一组光谱图 (通常是 0.5–1), 通过跟踪重要节点提取一组轨迹. 然后通过对轨迹的时间过程和随时间的空间变化进行建模、计算, 获取这些描述符在事件[13]和运动识别特征[14]. 基于上述研究成果, 本文提出了一种基于轨迹的时空谱特征语音情感识别方法. 该方法的核心思想是从语音频谱图,获得空间和时间上的描述符, 进行分类和维度情感识别. 与 MFCCs 和基频等特征提取方法相比, 本文提出的方法在噪声条件下, 调制光谱特更具鲁棒性. 在 4 类情绪识别实验中获得了可比较的非加权平均值回馈, 在激活识别任务中显著优于 Conv-PS 和 Opem-Utt.

1 研究方法

1.1 情感语音数据库的选择

语音信号的特征是指它的声学特征、语音信号的时域波形、频谱特征以及语音信号的统计特性. 语音信号首先是一个时间序列, 进行语音分析时, 最直观的就是它的时域波形. 通过分析语音信号的时域波形, 提取情感特征, 就可以判断说话者的喜怒哀乐.

从语音信号中提取反映情感的参数较为困难, 因为语音信号中包含了多种特征信息, 不仅包括了说话者自身的特征信息、说话者的情感状态信息, 也包括了说话内容、词汇和语法信息等. 目前很多文献对如何提取语音中的情感特征参数做了大量的研究. 其中, 基频作为描述情感的最主要特征, 很多文献都采用基于基频的统计特征, 如峰值、均值、方差等. 虽然这些特征描述了语音信号在不同情感状态下的变化, 但是没有进一步详细描述基频曲线的变化趋势. 针对这种现状, 提出了一种基于轨迹的空间-时间谱特语音情感识别方法. 其核心思想是从语音频谱图, 获得空间和时间上的描述符, 进行分类和维度情感识别, 来提高情感的判断力.

本研究采用著名情感数据库: USC IEMOCAP 数据库[15]用于算法实验. 这个数据库由 10 个参与者组成, 他们两人一组, 进行面对面的互动. 二元互动的设计是为了从演员中引出自然的多模态情感表现. 话语都有明确的情感标签 (如: 愤怒、快乐、悲伤、神经等) 和维度表征 (如: 价感、激活和支配). 每句话的特征标签至少由 3 个评分者标注, 维度属性至少由 2 个评分者标注. 考虑到这个数据库的自发性和评估者之间的协议约为 0.4, 这个数据库对于算法的发展仍然是一个具有挑战性的情绪数据库. 在这项工作中, 我们在这个数据库上进行了两项不同的情绪识别任务:1) 四类情绪识别; 2) 三层的情感效价维度和激活维度识别. 对于分类情绪识别, 分别是快乐的、悲伤的、中性的和愤怒的, 可以认为样本与“兴奋”的标签是相同的“快乐”. 评价和激活的三个层次被定义为: 低 (0-1:67)、中 (1:67-3:33) 和高 (3:33-5), 其中每个样本的值是基于评分者的平均值计算的. 表 1 列出了每种类型标签的样本数量.

表 1 情感分类标签的样本数量

情感分类标签	531(快乐)	576(悲伤)	411(中性)
唤醒维度标签	331(低)	1228(中)	337(高)
情感价维标签	653(低)	820(中)	423(高)

1.2 基于轨迹的时空光谱特征提取

语音信号的振幅特征和各种情感信息也具有较强的相关性. 当说话者处于生气或者高兴时, 出现较大的幅值, 而悲伤情感的幅度值较低, 而且这些幅度差异越大, 体现出情感的变化也越大. 此外, 语音的共振峰频率也是表达情感的特征参数之一. 当同一人发出的带有不同情感而内容相同的语句时, 其声道会有不同的变化, 而语音的共振峰频率与声道的形状和大小有关, 每种形状都有一套共振峰频率作为其特征.

因此, 本研究试图从语速、基频 (范围、平均值、包络等)、谱信息 (共振峰位置, 带宽等)、 语音能量信息特征方面具体分析语音中的情感特征.

图 1 描述了基于轨迹的时空光谱特性的音频文件分析流程. 以下是特征提取的步骤: 空间时间谱特征提取: 话语框架, 代表了信号实现框架使用一个情感序列, 形成每个 MFB-系数轨迹, 计算基于网格的时空特征和获得额外导出轨迹. 如假设$p=(p_1,p_2,\cdots,p_k)$是一个语音信号的基础频率, 其中 k 为这个语音信号的基础频率帧数, 那么, 这个语音信号基础频率的最大值为: $p_{\max}=\max(p_1,p_2,\cdots,p_k)$; 最小值: $p_{\min}=\min(p_1,p_2,\cdots,p_k)$;

均值: $p_{均值}=\frac{1}{k}\sum_{i=j}^{k}p_i$; 动态范围为: $p_{range}=p_{max}-p_{min}$; 方差为: $p=\sqrt{\sum_{i=j}^{k}(p_i-p_{均值})^2}$.

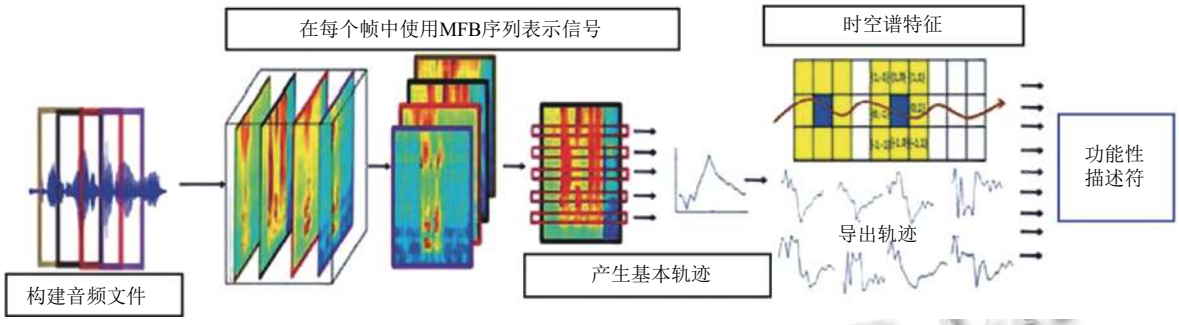

图 1 基于轨迹的时空光谱特征分析流程

语音信号的时空谱计算则可以表示为: $\Delta p_{前端}$, $\Delta p_{后端}$, $\Delta p_{争端语音信号帧}$. 通过计算统计函数轨迹, 就可获得框架水平特性.

(1) 框架的信号

将整个话语分割成帧的区域, 每个帧的长度为 L(L=250 ms, 150 ms). 帧之间有 50% 的重叠.

(2) 代表段

使用 26 个 Mel 滤波器能量组 (MFB) 输出的序列表示每一帧中的信号——也可以被成像为光谱图. MFB 的窗口大小设置为 25 ms, 重叠度为 50%. MFB 计算的频率上限为 3000 Hz.

(3) 形成基本轨迹

26 个滤波器输出的每个能量轮廓在每个帧的持续时间内形成一个基本轨迹.

(4) 计算时空特征

对于每个基本轨迹, 在 t= 1 时, 我们计算其相邻网格的一阶差分 (8total: 在图 1 中标记为黄色); 然后我们沿着时间轴移动, 计算这些网格差, 直到帧结束. 因此, 我们得到 8 个额外的轨迹 (所谓的派生轨迹), 为每帧 26 个滤镜输出 (一个轨迹的真实例子见图 1), 组成总共 9 个轨迹 (1 个基本轨迹+8 个派生轨迹).

(5) 框架水平时空描述符

我们通过应用 4 个统计功能, 即基于帧级轨迹的时空描述符, 得到最终的帧级轨迹. 即: 最大、最小、平均、标准偏差. 26×9 轨迹——每帧形成一组特性.

我们新提出的特性的基本思想本质上是跟踪光谱能量的变化在一个长期的框架内, 在频率轴 (空间) 和时间轴的方向上. 由于框架灵感来自于视频描述符的提取方法, 与语音生成/感知相关的物理意义虽然很难建立. 但是, 这个框架提供了一种简单的方法来量化语音信号的频谱-时间特性之间的各种相互关系, 直接从时间-频率表示, 而不需要进行更高级别的处理.

2 实验与结果分析

在本研究中, 我们对前文所述的情感识别任务进行了如下两个实验:

(1) 实验 I: 三种情绪识别实验中我们提出的带有 Conv-PS 和 OpEmo-Uttfeatures 的 Traj-ST 的比较和分析.

(2) 实验 II: 在三个情感识别实验中, Traj-ST 与 Conv-PS 和/ oropem-utt 特征融合后的识别精度分析.

其中, Conv-PS 特征提取方法与 Traj-ST 相似, 但不是计算 Mel-filter 输出轨迹的时空特征, 而是每 10 ms 计算基本频率 (f0)、强度 (INT)、MFCCs、它们的 delta 和 delta-delta -delta -delta -delta -delta 45 个低级描述符. 然后我们将 7 个统计函数 (max, min, mean, standard deviation, kurtosis, skewness, inter-quantile range) 应用到这些 LLD 特征上, 从而得到每一帧 Conv-PS 总共有 315 个特性. OpEmo-Utt 是一个详尽的语音级特性集. 在许多辅助语言识别任务中都有使用. 每句话包含 6668 个特征. 所有的特征都是针对单个说话者的. 所有的评价都是通过一对一的交叉验证进行的, 精度是用非加权平均的方法来衡量的. 基于 ANOVA 测试的单变量特征选择是针对 Traj-ST 和 Conv-PS 特性集进行的.

2.1 识别框架

在实验 I 中, 对于 Traj-ST 和 Conv-PS 特征集, 我们使用高斯混合模型 (M=32) 生成帧级每个类标签的概率分数 $p_{i,t}$, 然后使用以下简单规则进行帧级识别:

$$\arg\max_{i\in classes}\sum_{t=1}^{N}p_{i,t}$$

在提到的类标签中, t 指的是框架指数, 而 N 则指的是一个话语中的总帧数. 对于 OpEmo-Utt, 由于它是一个大维度的话语级特征向量, 我们在进行主成分分析 (90% 的方差) 和线性核支持向量机多类分类器后, 使用了基于 GMM 的方法.

在实验 II 中, Traj-ST 与 Conv-PS 和 OpEmo-Utt 的融合方法如图 2 所示. 融合框架基于逻辑回归. 对于 Traj-ST 和 Conv-PS, 融合是在统计功能上进行的, 即均值, 标准差, 最大值和最小值, 应用于 $p_{i,t}$; 对于 OpEmo-Utt, 融合是基于从一个 Vs-all 多类支持向量机输出的决策分数进行的.

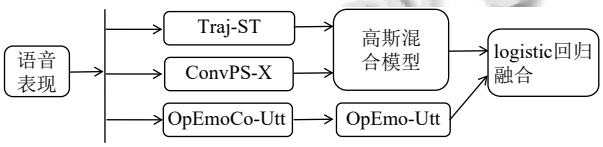

图 2 三个特征集融合方法

图 2 描述了三种特征集的融合方法. 基于框架的特征用 GMM 模型概率评分输出的统计功能进行融合, 使用 SVM 分类器的决策分数直接融合话语层次特征. 最后采用的融合模型是 logistic 回归.

2.2 实验 I: 结果和讨论

表 2 总结了 Exp i 的详细结果. 对于 Traj-ST 和 Conv-PS, 我们报告了使用不同帧长进行特征提取的 GMM 模型的 UARsof, 即 125 ms, 250 ms, 375 ms, 完整发音长度. 对于 OpEmo-Utt, 我们报告了使用 GMM 和 svm 模型的 UARs.

结果中有几点需要注意. 在四类情绪识别任务中, Traj-ST 与 OpEmo-Utt (47.5% vs. 47.7%) 进行了比较, 而最佳准确率为 Conv-PS(48.6%). 在三层价识别任务中, 使用 OpEmo-Utt(47.4%) 是最准确的, 在这一任务中, Traj-ST 和 Conv-PS 表现不佳. 最后, 我们建议的 Traj-ST 特性集在三层激活识别任务上的性能明显优于 Conv-PS 和 OpEmo-Utt. 它的识别率达到了 61.5%, 比 Conv-PS 提高了 1.7%, 比 OpEmo-Utt 提高了 2.9%. 通过三种类型的情绪识别任务的运行, 似乎可以明显地看出, 每一组这些特征确实具有不同数量和不同质量的情绪内容. Opem-Uttem 似乎对价性表现得最好, 这可能是由于对价度的感知的复杂性. 例如, 需要在话语层面提取语气特征. 虽然过去已经证明, 与声音有关的特征在激活维度中往往包含更多的信息, 但是我们仍然可以很肯定地看到我们提出的特征, Traj-ST, 在预测激活的整体感知方面比这两个其他特征集更有效.

识别任务: 4 级情绪识别, 3 级激活/情感效价识别. 对于 Traj-ST 和 Conv-PS, 采用具有不同框架长度的 GMM 模型的 UARs, 用于特征提取. 对于 OpEmo-Utt, 使用 GMM 和 SVM 模型的 UARs. 帧的持续时间也对获得最佳的精度 forTraj-ST(也适用于 Conv-PS) 起着重要的作用. 由此可见, 大约 250 ms 的持续时间是最理想的帧-持续时间.

这一结果证实了已有研究在情感识别中使用长期光谱特征的发现. 此外, Traj-ST 的特征选择输出结果表明, 时空特征的前三个方向分别为{0, 0}-基轨迹, {1, 0}-高时空等效方向轨迹, 以及{1, –1}-高时空-早时空方向轨迹. 这三种特征占选择产生的特征的 50%. 这些轨迹量化了光谱能量向高频段方向的变化, 具有较高的情感识别精度, 在 3 级激活识别中也表现显著.

2.3 实验 II: 结果和讨论

假设在实验 I 中, 每一组特征似乎都能识别不同的情绪表现. 为了进一步验证算法的可靠性, 本文融合这三种不同的特征. 表 2 列出了各种融合结果. OpEmo-Utt 是指融合 SVM 模型输出的决策分数. 表 3 总结了三个不同特征集的融合结果.

需要注意 Traj-ST, Conv-PS, OpEmo-Utt 为使用 UAR 计算所呈现的数目.

由表 3 可见. 首先, 不同特征集的融合都提高了最佳单特征集的结果数据. 具体表现在, 4 类情感识别的最佳融合精度是通过融合所有三组特征获得的 53.5% (相对于绝对单个特征集的 4.8% 的绝对改进); 3 级情感效价的最佳融合结果是 47.8% (1% 绝对改进优于最佳单特征集, OpE). 最后, 三级激活的最佳融合结果是 61.2% (相对于最佳单特征集 0.9% 的绝对改进, Traj-ST). 由此可见, 本文新提出的特征 Traj-ST 确实能够在该融合框架下进一步提高分类情感识别和激活水平检测的识别率, 这意味着我们的特征的互补信息在情感方面具有较高的一致性. 总之, 实验证明, 以轨迹为基础的空间时间谱特征可以结合利用两个不同的声学特征集, 提高情感识别率.

表 2 实验 I 输出了三种不同情绪的结果

	四级情感识别									
	Traj-ST: 提出功能				Conv-PS: MFCC + 智力+ F0				OpEmo-Utt6668features	
	125 ms	250 ms	375 ms	全部	125 ms	250 ms	375 ms	全部	GMM	支持向量机
快乐	35.5	34.2	41.2	34.4	40.6	44.2	40.1	42.9	45.9	44.6
悲伤的	65.4	65.6	64.6	43.1	73.1	73.2	71.8	55.7	54.3	59.6
中性	29.4	39.1	34.5	31.4	27.7	24.1	23.2	32.6	22.1	35.2
愤怒的	44.6	49.2	49.4	48.3	47.2	52.8	48.6	47.7	60.2	51.5
UAR	43.6	47.1	47.5	39.2	47.2	48.6	45.8	44.7	45.6	47.6
	维度属性分类激活三级									
低	76.1	74.2	67.6	44.5	72.5	66.2	56.7	29.3	22.3	61.2
中期	59.3	60.2	62.5	62.4	51.3	52.2	57.1	74.6	78.2	55.2
高	48.3	49.5	53.4	49.8	55.4	51.3	56.3	36.2	39.7	59.0
UAR	61.2	61.5	61.2	52.2	59.2	56.5	57.2	46.4	46.9	58.6
	空间属性分级 3 级化合价									
低	33.3	32.9	33.2	34.2	34.6	26.8	32.8	46.8	55.5	50.3
中期	61.5	61.8	60.5	59.2	57.4	58.5	54.8	47.6	50.2	50.2
高	28.8	29.7	30.2	30.1	52.4	46.8	42.0	31.6	26.9	46.5
UAR	41.2	41.5	41.2	40.5	45.2	46.0	43.2	42.1	44.2	47.4

表 3 实验 II 输出了三个不同特征集融合的分析结果

融合	情感	激活	情感效价
Traj-ST+Conv-PS	51.5	60.1	47.0
Traj-ST+OpEmo-Utt	52.0	61.2	47.0
Conv-PS +OpEmo-Utt	51.6	51.6	47.8
Traj-ST+ Conv-PS +OpEmo-Utt	53.5	61.5	47.8

3 结语

本文提出了一种低水平声学特征的语音情感识别方法, 以表征语音信号的长期时空信息. 我们利用所提出的特征对分类情感归因和维度表征进行情感识别实验. 实验表明, 所提出的特征集与已建立的低级声学描述符和最先进的穷举特征提取方法相比, 在分类情感识别方面具有更优秀的性能, 在激活水平识别的任务上优于现有的特征提取方法. 通过融合基于轨迹的时空特征, 提高了情感识别的整体精度.

参考文献

1 Bach-y-Rita P, Kercel SW. Sensory substitution and the human-machine interface. Trends in Cognitive Sciences, 2003, 7(12): 541–546. [doi: 10.1016/j.tics.2003.10.013]

2 Busso C, Bulut M, Lee CC, *et al*. IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2008, 42(4): 335–359. [doi: 10.1007/s10579-008-9076-6]

3 Calvo RA, D'Mello S, Gratch J, *et al*. The Oxford handbook of affective computing. Oxford, England: Oxford University Press, 2014.

4 Campbell WM, Sturim DE, Reynolds DA. Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters, 2006, 13(5): 308–311. [doi: 10.1109/LSP.2006.870086]

5 Chaspari T, Dimitriadis D, Maragos P. Emotion classification of speech using modulation features. Proceedings of the 22nd European Signal Processing Conference. Lisbon, Portugal. 2014. 1552–1556.

6 Chi TS, Yeh LY, Hsu CC. Robust emotion recognition by spectro-temporal modulation statistic features. Journal of Ambient Intelligence and Humanized Computing, 2012, 3(1): 47–60. [doi: 10.1007/s12652-011-0088-5]

7 Childers DG, Wu K. Gender recognition from speech. Part II: Fine analysis. The Journal of the Acoustical Society of America, 1991, 90(4): 1841–1856. [doi: 10.1121/1.401664]

8 Dobry G, Hecht RM, Avigal M, *et al*. Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 1975–1985. [doi: 10.1109/TASL.2011.2104955]

9 Eyben F, Wöllmer M, Schuller B. Opensmile: The Munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM International Conference on Multimedia. Firenze, Italy. 2010. 1459–1462. [doi: 10.1145/1873951.1874246]

10 Neville H, Krebs HI, Sharon A, *et al*. Interactive robotic therapist. USA: United States Patent, 1995.

11 Hollinger GA, Georgiev Y, Manfredi A, *et al*. Design of a

social mobile robot using emotion-based decision mechanisms. Proceedings of 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. Beijing, China. 2006. 3093–3098. [doi: 10.1109/IROS.2006.282327]

12 Kim Y, Lee H, Provost EM. Deep learning for robust feature generation in audiovisual emotion recognition. Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada. 2013. 3687–3691. [doi: 10.1109/icassp.2013.6638346]

13 Lee CM, Narayanan SS. Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing, 2005, 13(2): 293–303. [doi: 10.1109/TSA.2004.838534]

14 Lee CC, Mower E, Busso C, *et al*. Emotion recognition using a hierarchical binary decision tree approach. Speech Communication, 2011, 53(9–10): 1162–1171. [doi: 10.1016/j.specom.2011.06.004]

15 Li LF, Zhao Y, Jiang DM, *et al*. Hybrid deep neural network-hidden markov model (DNN-HMM) based speech emotion recognition. Proceedings of 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. Geneva, Switzerland. 2013. 312–317. [doi: 10.1109/ACII.2013.58]