

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 4 月 16 日 (16.04.2020)



(10) 国际公布号
WO 2020/073665 A1

(51) 国际专利分类号:
G10L 25/63 (2013.01) *G10L 21/0208* (2013.01)
G10L 25/30 (2013.01) *G10L 15/06* (2013.01)
G10L 25/45 (2013.01) *G10L 15/20* (2006.01)
G10L 25/18 (2013.01)

(21) 国际申请号: PCT/CN2019/089148

(22) 国际申请日: 2019 年 5 月 30 日 (30.05.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811183893.2 2018 年 10 月 11 日 (11.10.2018) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福

安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。

(72) 发明人: 刘博卿(LIU, Boqing); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。 贾雪丽(JIA, Xueli); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。 王健宗(WANG, Jianzong); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。 肖京(XIAO, Jing); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。

(74) 代理人: 北京鸿元知识产权代理有限公司 (GRANDER IP LAW FIRM); 中国北京市朝

(54) Title: METHOD AND SYSTEM FOR PERFORMING SPEECH EMOTION RECOGNITION USING SPECTRUM, AND STORAGE MEDIUM

(54) 发明名称: 在频谱上对语音进行情绪识别的方法、系统及存储介质

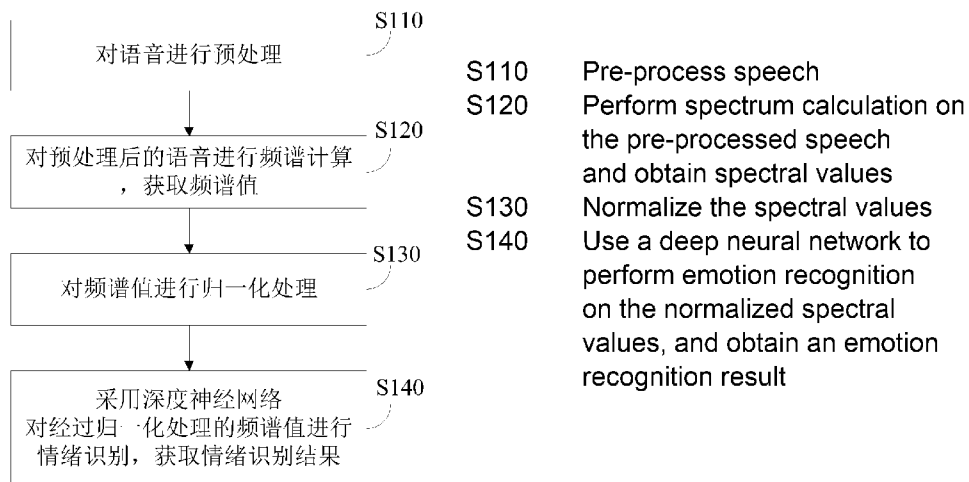


图 1

(57) Abstract: The present application relates to the technical field of artificial intelligence, and provides a method and system for performing speech emotion recognition using a spectrum, and a storage medium. The method comprises: pre-processing speech; performing spectrum calculation on the pre-processed speech and obtaining spectral values; normalizing the spectral values; and using a deep neural network to perform emotion recognition on the normalized spectral values, and obtaining an emotion recognition result. The present application can be used to resolve prior-art issues such as a low speech emotion recognition rate, time delay, and poor

阳区光华路7号汉威大厦东区18A6室,
Beijing 100004 (CN)。

- (81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。
- (84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

- 包括国际检索报告(条约第21条(3))。

robustness to unidentified environmental noise.

(57) 摘要: 本申请涉及人工智能技术领域, 提供一种在频谱上对语音进行情绪识别的方法、系统及存储介质, 其中方法包括: 对语音进行预处理; 对预处理后的语音进行频谱计算, 获取频谱值; 对频谱值进行归一化处理; 采用深度神经网络对经过归一化处理的频谱值进行情绪识别, 获取情绪识别结果。利用本申请, 能够解决目前对语音进行的情绪识别率低、延时以及对未知环境噪音的鲁棒性差等问题。

在频谱上对语音进行情绪识别的方法、系统及存储介质

5 本申请要求申请号为 201811183893.2，申请日为 2018 年 10 月 11 日，发明创造名称为“在频谱上对语音进行情绪识别的方法、系统及存储介质”的专利申请的优先权。

技术领域

10 本申请涉及人工智能技术领域，更为具体地，涉及一种在频谱上利用深度学习对语音进行情绪识别的方法、装置、系统及存储介质。

背景技术

15 情感计算的情绪识别是情感计算的一个重要组成部分，情绪识别研究的内容包括面部表情、语音、心率、行为、文本和生理信号识别等方面，通过以上内容来判断用户的情绪状态。其中，对语音进行的情绪识别可以被应用到很多方面，比如：人机交互，可以通过电子老师让电子学习变得更高效。

20 目前的一些困难与挑战有：提高识别率，降低延迟，提高对未知环境噪音的鲁棒性；其中，传统语音情绪识别的方法包括：首先对短时间的语音进行特征提取，然后在语句层面对其进行信息提取，最后再进行分类或者回归。近几年的方法是采用深度学习来进行上面所述步骤。虽然基于梅尔的频谱在语音识别上取得了很好的结果，但他去掉了一些音高的信息，但音高会反映情绪的信息，故采用基于梅尔的频谱(Mel-Frequency Cepstral Coefficients，简称 MFCC)在语音识别会丢失一些音高所反映的情绪信息。

25 因此，为了解决目前对语音进行的情绪识别率低、延时以及对未知环境噪音的鲁棒性差等问题，本申请提供了一种在频谱上对语音进行情绪识别的方法、装置、系统及存储介质。

发明内容

30 鉴于上述问题，本申请的目的是提供一种在频谱上对语音进行情绪识别

的方法、装置、系统及存储介质，采用代表语音谐波结构的线性频谱，实现高精度的情绪识别系统，同时降低延时性，增强对环境噪音的鲁棒性。

第一方面，本申请提供一种在频谱上对语音进行情绪识别的方法，应用于电子装置，包括：

- 5 对语音进行预处理；
- 对预处理后的语音进行频谱计算，获取频谱值，其中，采用重叠的海宁窗口对每一个子句进行操作，其中，窗移 10 毫秒，窗口大小 20 毫秒，
- 对于每一音频帧，计算一个维度为 800、网格分辨率为 20Hz 的离散傅里叶变换值，其中，所述离散傅里叶变换值即为频谱值，
- 10 将通过计算获得每一个子句的频谱值聚合在一起，形成一个 $N \times M$ 的矩阵，其中，根据语音句子的长度， $N \leq 300$ ，根据选定的频率网格分辨率， $M=200$ ；
- 对所述频谱值进行归一化处理；
- 采用深度神经网络对经过归一化处理的频谱值进行情绪识别，获取情绪识别结果。

- 15 第二方面，本申请还提供一种电子装置，该电子装置包括：存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如上述在频谱上对语音进行情绪识别的方法的步骤。

- 第三方面，本申请还提供一种在频谱上对语音进行情绪识别的系统，包括：
- 20

- 语音预处理单元，用于对语音进行预处理；
- 频谱值获取单元，用于对预处理后的语音进行频谱计算，获取频谱值；其中，所述频谱值获取单元包括：
- 海宁窗口处理模块，用于采用重叠的海宁窗口对每一个子句进行操作，
- 25 其中，窗移 10 毫秒，窗口大小 20 毫秒；
- 子句频谱值获取模块，用于对于每一帧，计算一个维度为 800、网格分辨率为 20Hz 的离散傅里叶变换值，其中，所述离散傅里叶变换值即为频谱值；
- 整句频谱值获取单元，用于将通过计算获得每一个子句的频谱值聚合在一起，形成一个 $N \times M$ 的矩阵，其中，
- 30 根据语音句子的长度， $N \leq 300$ ，根据选定的频率网格分辨率， $M=200$ ；

数据归一化处理单元，用于对所述频谱值进行归一化处理；

情绪识别获取单元，用于采用深度神经网络对经过归一化处理的频谱值进行情绪识别，获取情绪识别结果。

第四方面，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质中包括计算机程序，所述计算机程序被处理器执行时，实现如上述所述的在频谱上对语音进行情绪识别的方法的步骤。

从上面的技术方案可知，本申请提供的在频谱上利用深度学习对语音进行情绪识别的方法、装置、系统及存储介质，通过对语音进行频谱计算，并利用深度神经网络对经过频谱计算的语音进行情绪识别，从而解决传统的方法对语音情绪识别存在识别率低、延时，以及对未知环境噪音的鲁棒性差等问题。

为了实现上述以及相关目的，本申请的一个或多个方面包括后面将详细说明的特征。下面的说明以及附图详细说明了本申请的某些示例性方面。然而，这些方面指示的仅仅是可使用本申请的原理的各种方式中的一些方式。

此外，本申请旨在包括所有这些方面以及它们的等同物。

附图说明

通过参考以下结合附图的说明，并且随着对本申请的更全面理解，本申请的其它目的及结果将更加明白及易于理解。在附图中：

图 1 为根据本申请实施例的在频谱上对语音进行情绪识别的方法流程图；
图 2-1 为根据本申请实施例的对预处理后的语音进行频谱计算流程示意图；

图 2-2 为根据本申请实施例的对频谱值进行归一化处理流程示意图；

图 3 为根据本申请实施例的在频谱上对语音进行情绪识别的系统逻辑结构框图；

图 4 为根据本申请实施例的电子装置逻辑结构示意图。

在所有附图中相同的标号指示相似或相应的特征或功能。

具体实施方式

在下面的描述中，出于说明的目的，为了提供对一个或多个实施例的全

面理解，阐述了许多具体细节。然而，很明显，也可以在没有这些具体细节的情况下实现这些实施例。

以下将结合附图对本申请的具体实施例进行详细描述。

实施例 1

5 为了说明本申请提供的在频谱上对语音进行情绪识别的方法，图 1 示出了根据本申请实施例的在频谱上对语音进行情绪识别的方法流程。

如图 1 所示，本申请提供的在频谱上对语音进行情绪识别的方法包括：

S110：对语音进行预处理；

10 在步骤 S110 中，语音是指 IEMOCAP 数据库的语音，里面的语音采样频率是 16kHz，每一个句子时长从 1 秒到 20 秒不等，并且每一个句子都有一个情绪的标签。

对语音进行预处理就是将每一个长于 3 秒的句子分成基本上等长小于 3 秒的子句，每一个子句的标签和整个句子的标签一样的；每个句子可以分成若干个子句，并且子句情绪的标签于整句的情绪标签一致，当对整句的预测
15 进行评估时，同时对子句计算后验概率的平均值，从而在一定程度上解决延时的问题。

S120：对预处理后的语音进行频谱计算，获取频谱值。

可以理解地，频谱是指一个时域的信号在频域下为方式，可以针对语音信号进行傅立叶变换而得，所得的结果会是以分别以振幅及相位为纵轴，频率为横轴的两张图，不过有时也会省略相位的资讯，只有不同频率下对应振幅的资料。有时也以“振幅频谱”为振幅随频率变化的情形，“相位频谱”为相位随频率变化的情形，简单来说，频谱可以为一个讯号是由哪些频率的弦波所组成。在本实施例的采用对语音信号进行频谱计算，从而得到语音的傅立叶变换值，进一步获取语音的情绪特征，情绪特征是用于体现语音数据中
25 用词的词性特征，包括但不限于本实施例中的高兴词和悲伤词。词性特征与其他语音特征结合，有利于识别语音数据对应的说话人的情绪。

具体地，预处理后的语音是指将整句分成小于等于 3 秒的若干子句；计算若干子句的频谱，将所有的子句的频谱合在一起，形成一个频谱矩阵（N×M）。如图 2-1 所示，对预处理后的语音进行频谱计算流程如下：

30 S121：采用重叠的海宁窗口对预处理后的语音信号（小于等于 3 秒的子

句) 进行操作, 其中, 窗移 10 毫秒, 窗口大小 20 毫秒。

S122: 对于每一音频帧, 计算一个维度是 800、网格分辨率为 20Hz 的离散傅里叶变换 (DFT, Discrete Fourier Transformation) 值, DFT 值就是频谱值; 其中, 只考虑 0-4kHz 的语音, 忽略别的频率的语音。

5 S123: 将每个整句的所有子句的频谱都合到一起, 会得到一个 $N \times M$ 的矩阵, 其中, 根据整句的语音句子的长度, 得到 $N \leq 300$; 根据选定的频率网格分辨率得到 $M=200$ 。

信号 $x[n]$ 的 N -点 DFT 的公式如下:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j \frac{2\pi}{N} kn}, k = 0, \dots, N-1$$

10 其中, $x[n]$ 为预处理后的语音信号, j 代表的是虚数, $X[k]$ 为傅立叶变换后的语音信号。

S130: 对频谱值进行归一化处理。

具体地, 在本实施例中, 如图 2-2 所示, 对频谱值进行归一化处理的过程如下:

S131: 将频谱值转为能量频谱对数值;

15 S132: 采用经验值 E_{noise} 对所述能量频谱对数值进行限制, 即: 能量频谱对数值减去经验值 E_{noise} ;

S133: 采用归一化使得能量频谱对数值减去经验值 E_{noise} 的方差为 1。

20 其中, 在采用经验值 E_{noise} 对所述能量频谱对数值进行限制中, E_{noise} 是对语音进行静音检测, 提取出 200 小时的静音部分 (包括噪音部分), 对这些静音/噪音同样进行第一步(S110 步骤)和第二步 S120 (步骤) 的操作, 之后把得到的 DFT 数据同样转为能量频谱的对数形式 dB, 对这些数值进行平均得到 E_{noise} 。

归一化公式为:

$$X' = \frac{X}{X \text{ 的标准差}}$$

25 式中, X 是音频信号, x' 为归一化后音频信号。

其中, 在本申请的一个具体的实施例中, 在将频谱值转为能量频谱对数值的过程中, 首先, 采用开源的音高检测器, 获取语音每帧的基音频率; 然

后，对于每一个非静音的帧，经过噪音滤波器，获取改进版的能量频谱对数值。

改进版的能量频谱对数值公式为：

$$S(f) = E(f) - 0.5 * (1 - \cos(2\pi f / F_0)) * D(f)$$

5 其中， $S(f)$ 为改进过的短时能量频谱对数值； $E(f)$ 是短时频谱包络； F_0 为基音频率； $D(f)$ 从 20dB @0Hz 到 12dB @4KHz 是线性的； f 为频率， $0 \leq f \leq 4\text{KHz}$ 。

也就是说，对预处理后的语音进行频谱计算的过程具体包括对预处理后的语音数据进行采用重叠的海宁窗口处理，对处理后的语音数据进行分帧、
10 加窗运算，然后通过傅里叶变换以获取上述的语音特征；再将上述语音特征进行归一化处理作为输入深度神经网络模型进行情绪识别。

其中，神经网络的输入是定长的，因为在第二步（步骤 S120）中语音句子的长度 $N \leq 300$ ，因此将每个语音句子的长度都定为 300，所以对不足 300 个时间点的语音用填 0 的方法来满足需求。

15 S140：采用深度神经网络对经过归一化处理的频谱值进行情绪识别，获取情绪识别结果。

具体地，将经过归一化处理的频谱值输入卷积神经网络（Convolutional Neural Network, CNN），并将卷积神经网络的输出作为长短期记忆网络（LSTM, Long Short-Term Memory）的输入，从而对语音进行情绪识别，得到情绪识别
20 的结果。

其中，卷积神经网络是一种前馈神经网络，它的人工神经元可以响应一部分覆盖范围内的周围单元，对于大型图像处理有出色表现，它包括卷积层（convolutional layer）和池化层（pooling layer）。

CNN 的基本结构包括两层，其一为特征提取层，每个神经元的输入与前一层的局部接受域相连，并提取该局部的特征。一旦该局部特征被提取后，
25 它与其它特征间的位置关系也随之确定下来；其二是特征映射层，网络的每个计算层由多个特征映射组成，每个特征映射是一个平面，平面上所有神经元的权值相等。特征映射结构采用影响函数核小的 sigmoid 函数作为卷积网络的激活函数，使得特征映射具有位移不变性。

30 LSTM(Long Short-Term Memory)是长短期记忆网络，是一种时间递归神

神经网络，适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。基于 LSTM 的系统可以学习翻译语言、控制机器人、图像分析、文档摘要、语音识别图像识别、手写识别、控制聊天机器人、预测疾病、点击率和股票、合成音乐等任务。在本申请的实施例中，采用长短期记忆网络对语音进行识别，从而得出情绪识别的结果。

其中，表 1 示出了在本申请的实施例中卷积神经网络和长短期记忆网络训练学习的过程。

表 1

输入：频谱 200*300 (4000Hz*3sec)
卷积层 1:16 个滤波器，大小为 12*16 (240Hz*160sec)
最大化 2:1 100*150
卷积层 2:24 个滤波器，大小为 8*12 (320Hz*240sec)
最大化 2:1 50*75
卷积层 3:32 个滤波器，大小为 5*7 (400Hz*280sec)
最大化 2:1 25*37
LSTM：双向，128*2
稠密层：长度 64
Dropout:长度 64
SOFT-MAX:长度
输出：4 个后验概率

如表 1 所示，将 4000Hz*3sec 的频谱输入到卷积神经网络，分别在卷积神经网络的第一卷积层（16 个滤波器，大小为 12*16）、第一池化层（2:1 100*150）、第二卷积层（24 个滤波器，大小为 8*12）、第二池化层（2:1 50*75）、第三卷积层（32 个滤波器，大小为 5*7）、第三池化层（2:1 25*37）进行训练学习，并提取语音特征，将提取的语音特征在长短期记忆网络进行继续学习，最后输出四个后验概率。

在一个可选实施例中，所述深度神经网络通过下述步骤构建：
通过下式构建第一卷积层

$$a_t^{1,ch} = \text{ReLu} \left(b^{1,ch} + \sum_{j=-d_1}^{d_1} W_j^{1,ch} x_{t+j} \right)$$

其中, ch 为第一卷积层的第 ch 通道, d_1 为第一卷积层的卷积核的半径, $a_t^{1,ch}$ 为第一卷积层的第 ch 通道在第 t 个位置的神经元的激活值, $b^{1,ch}$ 为第一卷积层的第 ch 通道的卷积核的偏置值, $W_j^{1,ch}$ 为第一卷积层的卷积核权重矩阵 $W^{1,ch}$ 的第 j 个元素, $W^{1,ch} \in R^{2*d_1+1}$, x_{t+j} 为输入向量 X 的第 $t+j$ 个元素,

5 ReLu为线性整流函数;

通过下式构建第一池化层

$$z_t^{1,ch} = \max_{t \in \{2t-1, 2t\}} (a_{t-}^{1,ch})$$

其中, $z_t^{1,ch}$ 为第一池化层在第 ch 个通道上的第 t 个位置上的输出值;
通过下式构建第二卷积层

$$a_t^{2,ch} = \text{ReLu} \left(b^{2,ch} + \sum_{i=1}^{dim_2} \sum_{j=-d_2}^{d_2} W_{i,j+d_2+1}^{2,ch} z_{i,t+j}^1 \right)$$

10 其中, $a_t^{2,ch}$ 为第二卷积层的第 ch 通道上第 t 个位置上的神经元的输出值, dim_2 为第二卷积层的第 ch 通道的激活矩阵 $a^{2,ch}$ 的维数, d_2 为第二卷积层的卷积核的半径, $b^{2,ch}$ 为第二卷积层的第 ch 通道的卷积核的偏置值, $W_{i,j}^{2,ch}$ 为第二卷积层的卷积核权重矩阵 $W^{2,ch}$ 的在 (i, j) 位置上的元素, $z_{i,j}^1$ 为第一池化层输出矩阵 z^1 的 (i, j) 位置上的元素;

通过下式构建第二池化层

$$z_t^{2,ch} = \max_{t \in \{2t-1, 2t\}} (a_{t-}^{2,ch})$$

15 其中, $z_t^{2,ch}$ 为第二池化层在第 ch 个通道上的第 t 个位置上的输出值;
通过下式构建第三卷积层

$$a^3 = \text{ReLu} (b^3 + W^3 z^2)$$

其中, a^3 为第三卷积层的输出, b^3 为第三卷积层的卷积核的偏置, W^3 为第三卷积层的卷积核的权重参数, z^2 为第二池化层输出矩阵;

通过下式构建第三池化层

$$z_t^{3,ch} = \max_{t \in \{2t-1, 2t\}} (a_{t-}^{3,ch})$$

20 其中, $z_t^{3,ch}$ 为第三池化层在第 ch 个通道上的第 t 个位置上的输出值。

所述长短期记忆网络包括输入层、隐含层和输出层, 输入层设置有输入门、输出层设置有输出门和遗忘门, 所述长短期记忆网络的构建步骤包括:

通过下式构建输入门模型

$$S_i(t) = f(a^{(S)} x_i(t) + b^{(S)} o(t-1) + c^{(S)})$$

其中， $x_i(t)$ 为输入层第 i 个节点在时刻 t 的输入， $x_i(t) = \tanh(w_1 z^{3, \text{ch}}(t) + w_2 o(t-1) + w_3)$ ， $z^{3, \text{ch}}(t)$ 为时刻 t 的深度神经网络的输出， $S_i(t)$ 为 $x_i(t)$ 在输入门的输出、 w_1 、 w_2 、 w_3 、 $a^{(S)}$ 、 $b^{(S)}$ 和 $c^{(S)}$ 为输入层的权值参数， $o(t-1)$ 为时刻 $t-1$ 的输出门与第 i 个节点对应的输出， f 为激活函数；

通过下式构建隐含层输出模型

$$H_j(t) = f(a^{(H)} S_i(t) + c^{(H)})$$

其中， $H_j(t)$ 为 $x_i(t)$ 在隐含层第 j 个节点的输出， $a^{(H)}$ 和 $c^{(H)}$ 为隐含层的权值参数；

通过下式构建输出门模型

$$o(t) = f(a^{(O)} H_j(t) + c^{(O)})$$

其中， $o(t)$ 为 $x_i(t)$ 在输出门的输出， $a^{(O)}$ 和 $c^{(O)}$ 为输出门的权值参数；通过下式构建遗忘门模型

$$Y(t) = f(a^{(Y)} H_j(t) + c^{(Y)})$$

其中， $Y(t)$ 为 $x_i(t)$ 在遗忘门的输出， $a^{(Y)}$ 和 $c^{(Y)}$ 为遗忘门的权值参数。

在本申请的实施例中，通过卷积神经网络对输入的语音频谱进行训练，并对训练语音数据进行情绪标注，以使训练语音数据携带情绪标签；然后对携带情绪标签的训练语音数据进行特征提取，获取携带情绪标签的训练语音特征；然后在采用长短期记忆网络对携带情绪标签的训练语音特征进行学习训练，将高兴、愤怒、悲伤、烦躁和平静等情绪对应的特征从训练语音特征中提取出来，从而获取对应的情绪识别结果，并将情绪识别结果从输出层输出。

优选地，所述长短期记忆网络的训练步骤包括：

将已确定情感标签的多个句子的子句经过深度神经网络后的输出作为输入门训练集，将情感标签作为输出门测试集，将其他标签作为遗忘门测试集；

通过输入门训练集、输出门测试集和遗忘门测试集训练输入门模型、隐含层输出模型、输出门模型和遗忘门模型。在上述方法对在语音进行情绪识别，并获得情绪识别的结果，下面将介绍对情绪识别的结果进行评估。一般

来说,对情绪识别的结果有两个评估标准,这两个标准作为模型的评估标准,分别为整体正确率和类正确率。其中,整体正确率是指每个句子都有相同的权重,评估整体正确率的公式为:

$$\text{整体正确率} = \frac{S_n^c \text{ 的个数}}{N}$$

5 假设有 N 个句子,分别为: $s_1^x, s_2^x, s_3^x, s_4^x, s_5^x, \dots, s_N^x, x \in (c, w)$, 其中 s_n^c 代表句子 s_n 的情绪被识别正确, s_n^w 代表句子 s_n 的情绪没有被识别正确。

类正确率是指首先对每个情绪计算正确率,然后再做平均。

假设有 K 个情绪 $e_1, e_2, \dots, e_K$, 在对句子进行情绪识别后每个句子都会有一个情绪的标签 e_k 。比如: 句子 s_1, s_3, s_5, s_6 被识别后的情绪的标签为 e_2 , 其中 s_1, s_3, s_6 10 的真实情绪是 e_2 , 但 s_5 的真实的情绪应该为 e_1 , 所以有三个情绪识别的结果是正确的, 一个情绪识别的结果是错误的, 所以 e_2 这个类的准确率就是 3/4。对每一个类都算出这样一个准确率, 然后再作平均就能得到总体的类正确率。

另外,在本申请的一个具体的实施例中,在噪音环境下进行情绪识别时,可以直接从 log 频谱中移除噪音,这样做的优点是不需要提前知道噪音的来源, 15 除此之外如何处理噪音非常大的情况,比如信噪比为 0dB 的情况。在本申请的实施例中,有两种方式,具体为:

第一方式: 带噪信号 $\rightarrow$ 带噪频谱(不加噪音滤波器) $\rightarrow$ 正常的预测器

20 第二方式: 带噪信号 $\rightarrow$ 改进过的频谱(加噪音滤波器) $\rightarrow$ 改进过的预测器

其中,第二种方式进行语音识别的过程如下: 首先用一个开源的音高检测器,获取语音每帧的基音频率。然后,对于每一个非静音的帧,产生一个改进版的能量频谱对数值(加入噪音滤波器):

$$S(f) = E(f) - 0.5 * (1 - \cos(2\pi f / F0)) * D(f)$$

25 其中, $S(f)$ 是改进过的短时能量频谱对数值; $E(f)$ 是短时频谱包络; $F0$ 是基音频率; $D(f)$ 从 20dB @0Hz 到 12dB @4KHz 是线性的; f 是频率, $0 \leq f \leq 4\text{KHz}$ 。

用 7 种不同的噪音信号：三个音乐信号和四个人群噪音；用 IEMOCAP 的数据训练两个预测器：正常的预测器（深度学习神经网络，3 层卷积层以及 LSTM 10 Hz 网格分辨率，输入的频谱值不是改进过的频谱值）和改进过的预测器（和正常深度学习神经网络相似，但是输入的频谱值为用改进过的频谱值来进行训练的）；然后将把噪音加到语音上，信噪比设为 0dB，然后用这两个预测器来处理有噪音的信号，处理过程如下所示：

带噪信号 → 带噪频谱 → 正常的预测器

带噪信号 → 改进过的频谱 → 改进过的预测器

通过上述两种方式获取的语音情绪识别结果，通过对情绪识别结果进行评估后，可以得知，采用第二种方式对语音进行识别，其真确率会更准确，因此，当噪音非常大的情况下，可以采用第二种方式对语音进行识别。

本申请实施例提供的在频谱上对语音进行情绪识别的方法，对语音进行预处理；对预处理后的语音进行频谱计算，获取频谱值；对频谱值进行归一化处理；采用深度神经网络对经过归一化处理的频谱值进行情绪识别，获取情绪识别结果。采用代表语音谐波结构的线性频谱，实现高精度的情绪识别，同时提高识别率，降低延时性，增强对环境噪音的鲁棒性。

应理解，上述实施例中各步骤的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不对本申请实施例的实施过程构成任何限定。

实施例 2

与上述方法相对应，本申请还提供一种在频谱上对语音进行情绪识别的系统，图 3 示出了根据本申请实施例的在频谱上对语音进行情绪识别的系统逻辑结构。

如图 3 所示，本申请提供一种在频谱上对语音进行情绪识别的系统 300，包括：语音预处理单元 310、频谱值获取单元 320、数据归一化处理单元 330 和情绪识别获取单元 340。其中，语音预处理单元 310、频谱值获取单元 320、数据归一化处理单元 330 和情绪识别获取单元 340 的实现功能与实施例中频谱上对语音进行情绪识别的对应的步骤一一对应，为避免赘述，本实施例不一一详述。

语音预处理单元 310，用于对语音进行预处理；

频谱值获取单元 320, 用于对预处理后的语音进行频谱计算, 获取频谱值;
数据归一化处理单元 330, 用于对频谱值进行归一化处理;

情绪识别获取单元 340, 用于采用深度神经网络对经过归一化处理的频谱值进行情绪识别, 获取情绪识别结果。

5 优选地, 语音预处理单元 310 在对语音进行预处理的过程中, 在语音数据库中, 每一个语音句子时长从 1 秒到 20 秒不等, 每一个句子均有一个情绪的标签; 将每一个长于 3 秒的语音句子分成等于小于 3 秒的子句, 并且每一个子句的标签于整句的标签相同。

10 优选地, 频谱值获取单元 320 包括: 海宁窗口处理模块 321、子句频谱值获取模块 322 和整句频谱值获取单元 323。

其中, 海宁窗口处理模块 321, 用于采用重叠的海宁窗口对每一个子句进行操作, 其中, 窗移 10 毫秒, 窗口大小 20 毫秒;

15 子句频谱值获取模块 322, 用于对于每一帧, 计算一个维度为 800、网格分辨率为 20Hz 的离散傅里叶变换值, 其中, 所述离散傅里叶变换值即为频谱值;

整句频谱值获取单元 323, 用于将通过计算获得每一个子句的频谱值聚合在一起, 形成一个 $N \times M$ 的矩阵, 其中,

根据语音句子的长度, $N \leq 300$, 根据选定的频率网格分辨率, $M = 200$ 。

20 优选地, 数据归一化处理单元 330 包括: 能量频谱对数值转化模块 331、能量频谱对数值限制模块 332 和方差获取模块 333。

其中, 能量频谱对数值转化模块 331, 用于将频谱值转为能量频谱对数值;
能量频谱对数值限制模块 332, 用于采用经验值 E_{noise} 对能量频谱对数值进行限制, 即: 能量频谱对数值减去经验值 E_{noise} ;

25 方差获取模块 333, 用于采用归一化使得能量频谱对数值减去经验值 E_{noise} 的方差为 1。

优选地, 能量频谱对数值转化模块 331 包括基音频率获取单元 3311 和改进版的能量频谱对数值单元 3312。

其中, 基音频率获取单元 3311, 用于采用开源的音高检测器, 获取语音每帧的基音频率;

30 改进版的能量频谱对数值单元 3312, 用于对于每一个非静音的语音帧,

经过噪音滤波器，获取改进版的能量频谱对数值。

优选地，情绪识别获取单元 340 包括卷积神经网络训练模块 341 和卷积神经网络训练模块 341。

其中，卷积神经网络训练模块 341，用于将经过归一化处理的频谱值输入卷积神经网络，并进行训练；

长短期记忆网络训练模块 342，用于将卷积神经网络的输出作为长短期记忆网络的输入，从而对语音进行情绪识别，得到情绪识别的结果。

本申请实施例提供的在频谱上对语音进行情绪识别的系统，语音预处理单元 310，用于对语音进行预处理；频谱值获取单元 320，用于对预处理后的语音进行频谱计算，获取频谱值；数据归一化处理单元 330，用于对频谱值进行归一化处理；情绪识别获取单元 340，用于采用深度神经网络对经过归一化处理的频谱值进行情绪识别，获取情绪识别结果。采用代表语音谐波结构的线性频谱，实现高精度的情绪识别，同时提高识别率，降低延时性，增强对环境噪音的鲁棒性。

实施例 3

图 4 是本申请一实施例提供的电子装置逻辑结构的示意图。如图 4 所示，该实施例的电子装置 40 包括处理器 41、存储器 42 以及存储在存储器 42 中并可在处理器 41 上运行的计算机程序 43。处理器 41 执行计算机程序 43 时实现实施例 1 中在频谱上对语音进行情绪识别的方法的各个步骤，例如图 1 所示的步骤 S110 至 S140。或者，处理器 41 执行在频谱上对语音进行情绪识别的方法时实现上述各装置实施例中各模块/单元的功能，例如图 3 所示的语音预处理单元 310、频谱值获取单元 320、数据归一化处理单元 330 和情绪识别获取单元 340。

示例性的，计算机程序 43 可以被分割成一个或多个模块/单元，一个或者多个模块/单元被存储在存储器 42 中，并由处理器 41 执行，以完成本申请。一个或多个模块/单元可以是能够完成特定功能的一系列计算机程序指令段，该指令段用于描述计算机程序 43 在电子装置 40 中的执行过程。例如，情绪识别程序 23 可以被分割成实施例 2 中的语音预处理单元 310、频谱值获取单元 320、数据归一化处理单元 330 和情绪识别获取单元 340，其功能作用在实施例 2 中有详细描述，在此不一一赘述。

电子装置 40 可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。电子装置 40 可包括，但不仅限于，处理器 41、存储器 42。本领域技术人员可以理解，图 2 仅仅是电子装置 40 的示例，并不构成对电子装置 40 的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件，例如电子装置还可以包括输入输出设备、网络接入设备、总线等。

所称处理器 41 可以是中央处理单元(Central Processing Unit, CPU)，还可以是其他通用处理器、数字信号处理器(Digital Signal Processor, DSP)、专用集成电路(Application Specific Integrated Circuit, ASIC)、现成可编程门阵列(Field-Programmable Gate Array, FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

存储器 42 可以是电子装置 40 的内部存储单元，例如电子装置 40 的硬盘或内存。存储器 42 也可以是电子装置 40 的外部存储设备，例如电子装置 40 上配备的插接式硬盘，智能存储卡(Smart Media Card ,SMC)，安全数字(Secure Digital ,SD)卡，闪存卡(Flash Card)等。进一步地，存储器 42 还可以既包括电子装置 40 的内部存储单元也包括外部存储设备。存储器 42 用于存储计算机程序以及电子设备所需的其他程序和数据。存储器 42 还可以用于暂时地存储已经输出或者将要输出的数据。

实施例 4

本实施例提供一计算机可读存储介质，该计算机可读存储介质上存储有计算机程序，该计算机程序被处理器执行时实现实施例 1 中在频谱上对语音进行情绪识别的方法，为避免重复，这里不再赘述。或者，该计算机程序被处理器执行时实现实施例 2 中在频谱上对语音进行情绪识别的系统中各模块/单元的功能，为避免重复，这里不再赘述。

所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，仅以上述各功能单元、模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能单元、模块完成，即将装置的内部结构划分成不同的功能单元或模块，以完成以上描述的全部或者部分功能。实施例中的各功能单元、模块可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中，上述集成的单元既可

以采用硬件的形式实现，也可以采用软件功能单元的形式实现。另外，各功能单元、模块的具体名称也只是为了便于相互区分，并不用于限制本申请的保护范围。上述系统中单元、模块的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

5 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中沒有详述或记载的部分，可以参见其它实施例的相关描述。

 本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

 在本申请所提供的实施例中，应该理解到，所揭露的装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述模块或单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通讯连接可以是通过一些接口，装置或单元的间接耦合或通讯连接，可以是电性，机械或其它的形式。

 作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

 另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

 所述集成的模块/单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请实现上述实施例方法中的全部或部分流程，也可以通过计算机程序来指令相关的硬件来完成，所述的计算机程序可存储于一计算机可读存储介质

中，该计算机程序在被处理器执行时，可实现上述各个方法实施例的步骤。其中，所述计算机程序包括计算机程序代码，所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、
5 U 盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM, Read-Only Memory)、随机存取存储器(RAM, Random Access Memory)、电载波信号、电信信号以及软件分发介质等。需要说明的是，所述计算机可读介质包含的内容可以根据司法管辖区内立法和专利实践的要求进行适当的增减，例如在某些司法管辖区，根据立法和专利实践，计算机可读介质不包括是电载波信号
10 和电信信号。

以上所述实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱
15 离本申请各实施例技术方案的精神和范围，均应包含在本申请的保护范围之内。

权 利 要 求 书

1.一种在频谱上对语音进行情绪识别的方法，应用于电子装置，其特征在于，包括：

5 对语音进行预处理；

对预处理后的语音进行频谱计算，获取频谱值，其中，采用重叠的海宁窗口对每一个子句进行操作，其中，窗移 10 毫秒，窗口大小 20 毫秒，

对于每一音频帧，计算一个维度为 800、网格分辨率为 20Hz 的离散傅里叶变换值，其中，所述离散傅里叶变换值即为频谱值，

10 将通过计算获得每一个子句的频谱值聚合在一起，形成一个 $N \times M$ 的矩阵，其中，根据语音句子的长度， $N \leq 300$ ，根据选定的频率网格分辨率， $M=200$ ；

对所述频谱值进行归一化处理；

采用深度神经网络对经过归一化处理的频谱值进行情绪识别，获取情绪识别结果。

15 2.根据权利要求 1 所述的在频谱上对语音进行情绪识别的方法，其特征在于，在对语音进行预处理的过程中，

在语音数据库中，每一个语音句子时长从 1 秒到 20 秒不等，每一个句子均有一个情绪的标签；

20 将每一个长于 3 秒的语音句子分成等于小于 3 秒的子句，并且每一个子句的标签于整句的标签相同。

3.根据权利要求 1 所述的在频谱上对语音进行情绪识别的方法，其特征在于，在对所述频谱值进行归一化处理的过程中，

首先，将所述频谱值转为能量频谱对数值；

25 然后，采用经验值 E_{noise} 对所述能量频谱对数值进行限制，即：所述能量频谱对数值减去经验值 E_{noise} ；

最后，采用归一化使得所述能量频谱对数值减去经验值 E_{noise} 的方差为 1。

4. 根据权利要求 3 所述的在频谱上对语音进行情绪识别的方法，其特征在于，所述经验值 E_{noise} 的获取方法包括：

对语音进行静音检测，提取设定时间段的噪音部分和静音部分；

30 对静音部分和噪音部分进行预处理；

对预处理后的静音部分和噪音部分进行频谱计算，获取静音部分和噪音部分的 DFT 数据；

把得到的 DFT 数据转为能量频谱的对数形式，进行平均得到 E_{noise} 。

5 5. 根据权利要求 3 所述的在频谱上对语音进行情绪识别的方法，其特征在于，在将所述频谱值转为能量频谱对数值的过程中，

首先，采用开源的音高检测器，获取语音每帧的基音频率；

然后，对于每一个非静音的帧，经过噪音滤波器，获取改进版的能量频谱对数值。

10 6. 根据权利要求 5 所述的在频谱上对语音进行情绪识别的方法，其特征在于，通过下式获取改进版的能量频谱对数值

$$S(f) = E(f) - 0.5 * (1 - \cos(2\pi f / F_0)) * D(f)$$

其中， $S(f)$ 为改进过的短时能量频谱对数值； $E(f)$ 是短时频谱包络；

F_0 为基音频率； $D(f)$ 从 20dB @0Hz 到 12dB @4KHz 是线性的；

F 为频率， $0 \leq f \leq 4\text{KHz}$ 。

15 7. 根据权利要求 1 所述的在频谱上对语音进行情绪识别的方法，其特征在于，在采用深度神经网络对经过归一化处理的频谱值进行情绪识别，获取情绪识别结果过程中，

20 将经过归一化处理的频谱值输入卷积神经网络，并将所述卷积神经网络的输出作为长短期记忆网络的输入，从而对语音进行情绪识别，得到情绪识别的结果。

8. 根据权利要求 1 所述的在频谱上对语音进行情绪识别的方法，其特征在于，所述深度神经网络进行参数训练时，采用整体正确率和类正确率作为评估标准，所述类正确率是将一个情绪标签作为一类，所述情绪标签对应的句子的整体正确率作为所述情绪标签的类正确率

$$25 \quad \text{整体正确率} = \frac{S_n^c \text{ 的个数}}{N}$$

其中， N 为句子总数， s_n^c 代表句子 s_n 的情绪被识别正确。

9. 根据权利要求 7 所述的在频谱上对语音进行情绪识别的方法，其特征在于，所述深度神经网络为卷积神经网络，包括第一卷积层、第一池化层、第二卷积层、第二池化层、第三卷积层和第三池化层。

10.根据权利要求 9 所述的在频谱上对语音进行情绪识别的方法，其特征在于，所述深度神经网络通过下述步骤构建：

通过下式构建第一卷积层

$$a_t^{1,ch} = \text{ReLu} \left(b^{1,ch} + \sum_{j=-d_1}^{d_1} W_j^{1,ch} x_{t+j} \right)$$

其中，ch 为第一卷积层的第 ch 通道， d_1 为第一卷积层的卷积核的半径， $a_t^{1,ch}$ 为第一卷积层的第 ch 通道在第 t 个位置的神经元的激活值， $b^{1,ch}$ 为第一卷积层的第 ch 通道的卷积核的偏置值， $W_j^{1,ch}$ 为第一卷积层的卷积核权重矩阵 $W^{1,ch}$ 的第 j 个元素， $W^{1,ch} \in R^{2*d_1+1}$ ， x_{t+j} 为输入向量 X 的第 t+j 个元素，ReLu 为线性整流函数；

通过下式构建第一池化层

$$z_t^{1,ch} = \max_{t \in \{2t-1, 2t\}} (a_{t-}^{1,ch})$$

其中， $z_t^{1,ch}$ 为第一池化层在第 ch 个通道上的第 t 个位置上的输出值；通过下式构建第二卷积层

$$a_t^{2,ch} = \text{ReLu} \left(b^{2,ch} + \sum_{i=1}^{dim_2} \sum_{j=-d_2}^{d_2} W_{i,j+d_2+1}^{2,ch} z_{i,t+j}^1 \right)$$

其中， $a_t^{2,ch}$ 为第二卷积层的第 ch 通道上第 t 个位置上的神经元的输出值， dim_2 为第二卷积层的第 ch 通道的激活矩阵 $a^{2,ch}$ 的维数， d_2 为第二卷积层的卷积核的半径， $b^{2,ch}$ 为第二卷积层的第 ch 通道的卷积核的偏置值， $W_{i,j}^{2,ch}$ 为第二卷积层的卷积核权重矩阵 $W^{2,ch}$ 的在 (i, j) 位置上的元素， $z_{i,j}^1$ 为第一池化层输出矩阵 z^1 的 (i, j) 位置上的元素；

通过下式构建第二池化层

$$z_t^{2,ch} = \max_{t \in \{2t-1, 2t\}} (a_{t-}^{2,ch})$$

其中， $z_t^{2,ch}$ 为第二池化层在第 ch 个通道上的第 t 个位置上的输出值；通过下式构建第三卷积层

$$a^3 = \text{ReLu} (b^3 + W^3 z^2)$$

其中， a^3 为第三卷积层的输出， b^3 为第三卷积层的卷积核的偏置， W^3 为第三卷积层的卷积核的权重参数， z^2 为第二池化层输出矩阵；

通过下式构建第三池化层

$$z_t^{3,ch} = \max_{t \in \{2t-1, 2t\}} (a_{t-}^{3,ch})$$

其中， $z_t^{3,ch}$ 为第三池化层在第 ch 个通道上的第 t 个位置上的输出值。

11. 根据权利要求 10 所述的在频谱上对语音进行情绪识别的方法，其特征在于，所述长短期记忆网络包括输入层、隐含层和输出层，输入层设置有

输入门、输出层设置有输出门和遗忘门，所述长短期记忆网络的构建步骤包括：

通过下式构建输入门模型 $S_i(t) = f(a^{(S)}x_i(t) + b^{(S)}o(t-1) + c^{(S)})$

其中， $x_i(t)$ 为输入层第 i 个节点在时刻 t 的输入，
 $x_i(t) = \tanh(w_1 z^{3, \text{ch}}(t) + w_2 o(t-1) + w_3)$ ， $z^{3, \text{ch}}(t)$ 为时刻 t 的深度神经网络的输出， $S_i(t)$ 为 $x_i(t)$ 在输入门的输出、 w_1 、 w_2 、 w_3 、 $a^{(S)}$ 、 $b^{(S)}$ 和 $c^{(S)}$ 为输入层的权值参数， $o(t-1)$ 为时刻 $t-1$ 的输出门与第 i 个节点对应的输出， f 为激活函数；

通过下式构建隐含层输出模型

$$H_j(t) = f(a^{(H)}S_i(t) + c^{(H)})$$

其中， $H_j(t)$ 为 $x_i(t)$ 在隐含层第 j 个节点的输出， $a^{(H)}$ 和 $c^{(H)}$ 为隐含层的权值参数；

通过下式构建输出门模型

$$o(t) = f(a^{(O)}H_j(t) + c^{(O)})$$

其中， $o(t)$ 为 $x_i(t)$ 在输出门的输出， $a^{(O)}$ 和 $c^{(O)}$ 为输出门的权值参数；
 通过下式构建遗忘门模型

$$Y(t) = f(a^{(Y)}H_j(t) + c^{(Y)})$$

其中， $Y(t)$ 为 $x_i(t)$ 在遗忘门的输出， $a^{(Y)}$ 和 $c^{(Y)}$ 为遗忘门的权值参数。

12. 根据权利要求 11 所述的在频谱上对语音进行情绪识别的方法，其特征在于，所述长短期记忆网络的训练步骤包括：

将已确定情感标签的多个句子的子句经过深度神经网络后的输出作为输入门训练集，将情感标签作为输出门测试集，将其他标签作为遗忘门测试集；

通过输入门训练集、输出门测试集和遗忘门测试集训练输入门模型、隐含层输出模型、输出门模型和遗忘门模型。

13. 一种电子装置，该电子装置包括：存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如权利要求 1 至 12 任一项所述在频谱上对语音进行情绪识别的方法的步骤。

14. 一种在频谱上对语音进行情绪识别的系统，其特征在于，包括：

语音预处理单元，用于对语音进行预处理；

频谱值获取单元，用于对预处理后的语音进行频谱计算，获取频谱值；
其中，所述频谱值获取单元包括：

海宁窗口处理模块，用于采用重叠的海宁窗口对每一个子句进行操作，
5 其中，窗移 10 毫秒，窗口大小 20 毫秒；

子句频谱值获取模块，用于对于每一帧，计算一个维度为 800、网格分辨率为 20Hz 的离散傅里叶变换值，其中，所述离散傅里叶变换值即为频谱值；

整句频谱值获取单元，用于将通过计算获得每一个子句的频谱值聚合在一起，形成一个 $N \times M$ 的矩阵，其中，

10 根据语音句子的长度， $N \leq 300$ ，根据选定的频率网格分辨率， $M=200$ ；

数据归一化处理单元，用于对所述频谱值进行归一化处理；

情绪识别获取单元，用于采用深度神经网络对经过归一化处理的频谱值进行情绪识别，获取情绪识别结果。

15 15.根据权利要求 14 所述的在频谱上对语音进行情绪识别的系统，其特征在于，所述数据归一化处理单元包括：

能量频谱对数值转化模块，用于将频谱值转为能量频谱对数值；

能量频谱对数值限制模块，用于采用经验值 E_{noise} 对能量频谱对数值进行限制，即：能量频谱对数值减去经验值 E_{noise} ；

20 方差获取模块，用于采用归一化使得能量频谱对数值减去经验值 E_{noise} 的方差为 1。

16.根据权利要求 15 所述的在频谱上对语音进行情绪识别的系统，其特征在于，所述能量频谱对数值转化模块包括：

基音频率获取单元，用于采用开源的音高检测器，获取语音每帧的基音频率；

25 改进版的能量频谱对数值单元，用于对于每一个非静音的语音帧，经过噪音滤波器，获取改进版的能量频谱对数值。

17. 根据权利要求 14 所述的在频谱上对语音进行情绪识别的系统，其特征在于，所述情绪识别获取单元的深度神经网络为卷积神经网络，包括第一卷积层、第一池化层、第二卷积层、第二池化层、第三卷积层和第三池化层。

30 18. 根据权利要求 17 所述的在频谱上对语音进行情绪识别的系统，其特

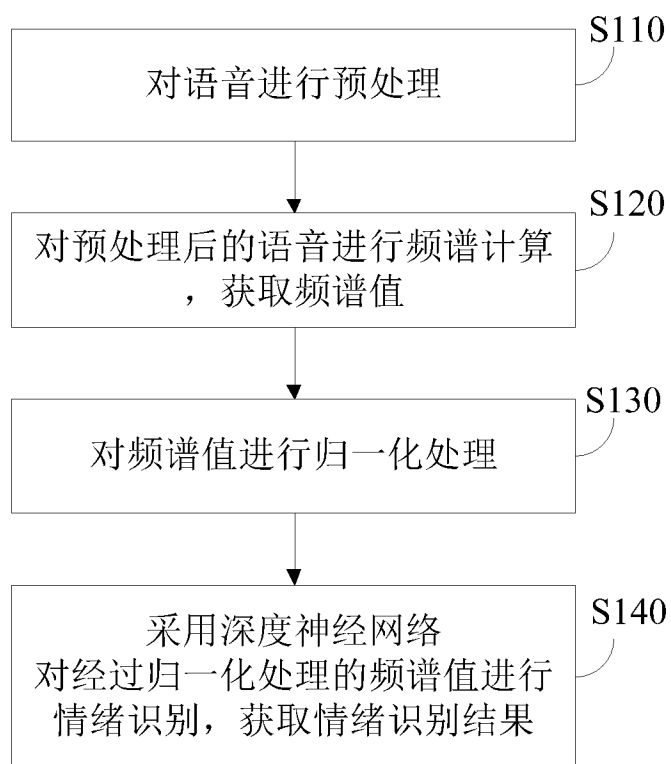
征在于，所述情绪识别获取单元的深度神经网络与述长短期记忆网络相结合，将所述卷积神经网络的输出作为长短期记忆网络的输入，从而对语音进行情绪识别，所述长短期记忆网络包括输入层、隐含层和输出层，输入层设置有输入门、输出层设置有输出门和遗忘门。

- 5 19. 根据权利要求 18 所述的在频谱上对语音进行情绪识别的系统，其特征在于，所述长短期记忆网络的训练步骤包括：

将已确定情感标签的多个句子的子句经过深度神经网络后的输出作为输入门训练集，将情感标签作为输出门测试集，将其他标签作为遗忘门测试集；

- 10 通过输入门训练集、输出门测试集和遗忘门测试集训练输入门模型、隐含层输出模型、输出门模型和遗忘门模型。

20. 一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中包括计算机程序，所述计算机程序被处理器执行时，实现如权利要求 1 至 12 中任一项所述的在频谱上对语音进行情绪识别的方法的步骤。

**图 1**

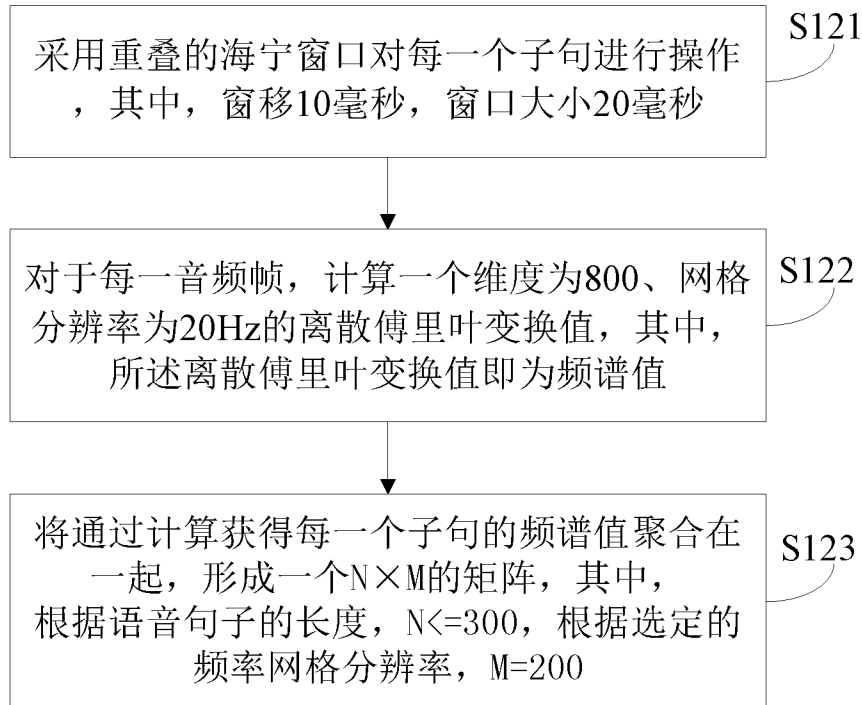


图 2-1

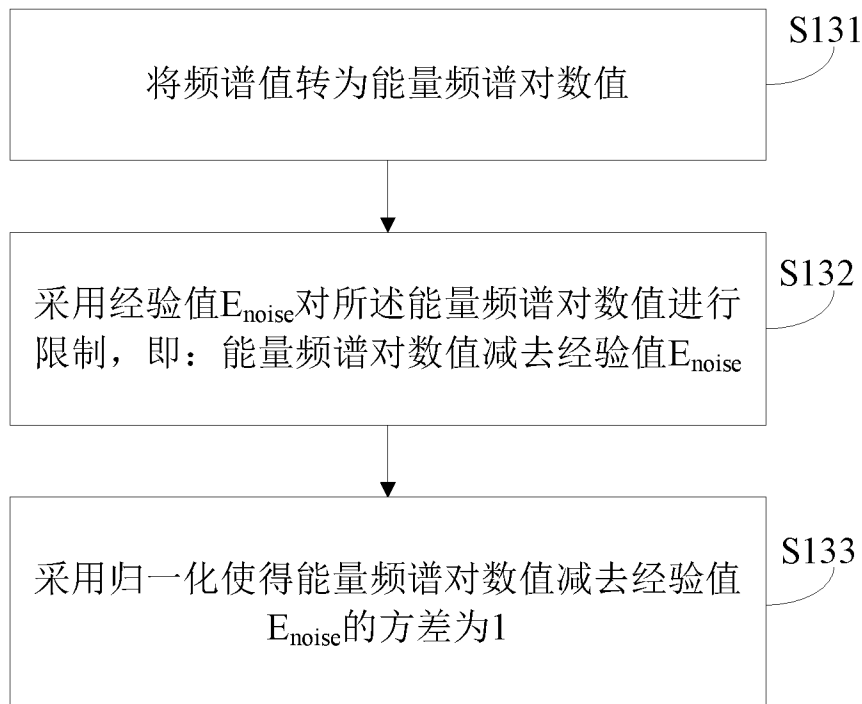


图 2-2

在频谱上对语音进行情绪识别的系统300

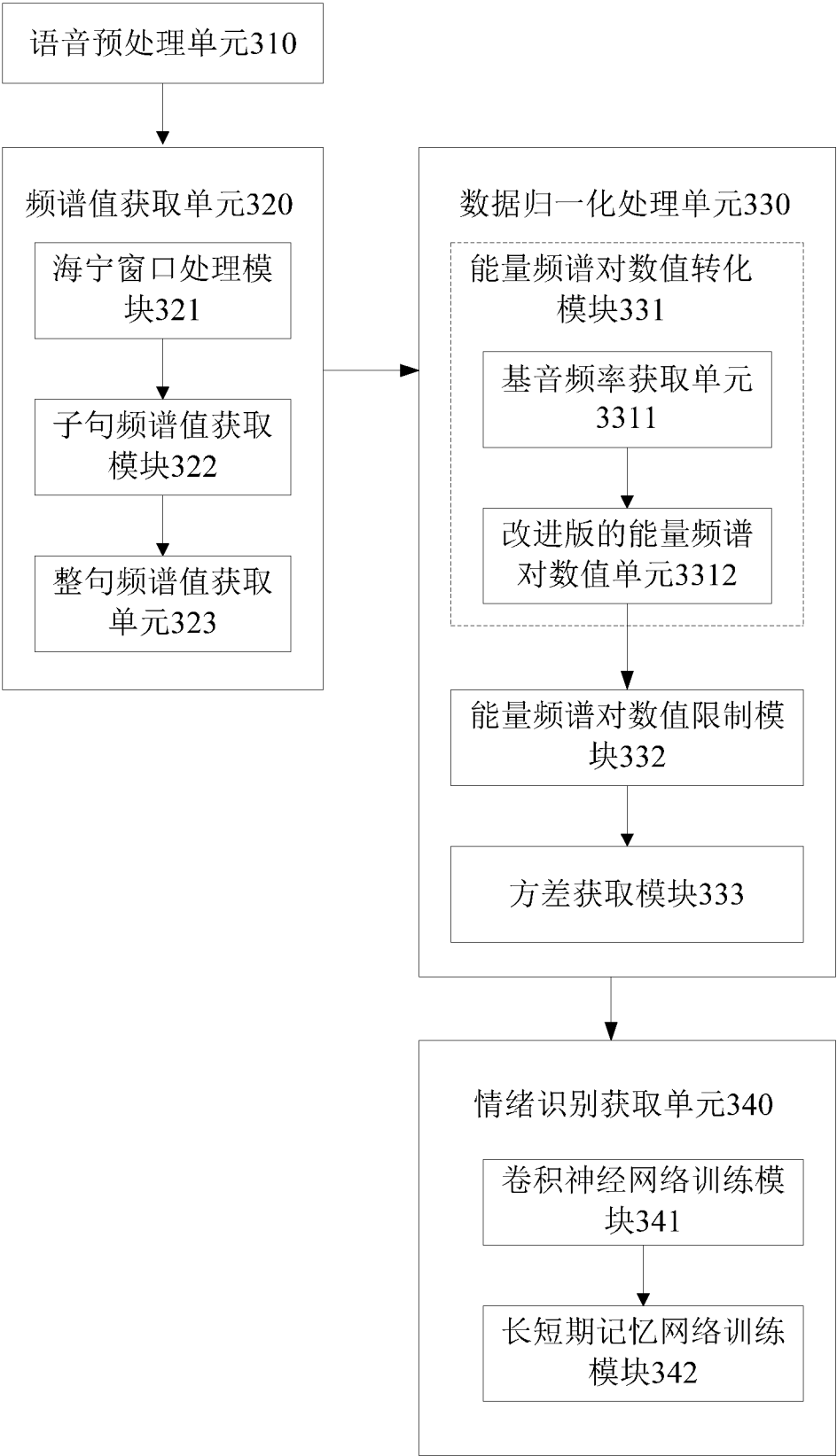


图 3

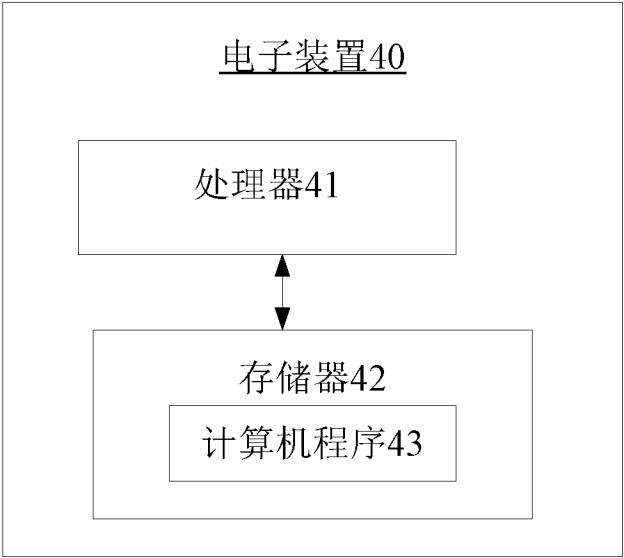


图 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/089148

A. CLASSIFICATION OF SUBJECT MATTER

G10L 25/63(2013.01)i; G10L 25/30(2013.01)i; G10L 25/45(2013.01)i; G10L 25/18(2013.01)i; G10L 21/0208(2013.01)i; G10L 15/06(2013.01)i; G10L 15/20(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; VEN; CNTXT; WOTXT; EPTXT; USTXT; CNKI; IEEE: 情绪识别, 语音, 加窗, 深度神经网络, 卷积神经网络, 长短期记忆网络, 归一化, voice, emotion recogni+, window, nature network, CNN, LSTM, normalizat+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109243491 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 18 January 2019 (2019-01-18) description, paragraphs [0036]-[0134], and figures 1-4	1-9, 13-20
A	CN 108597539 A (GUILIN UNIVERSITY OF ELECTRONIC TECHNOLOGY) 28 September 2018 (2018-09-28) description, paragraphs [0038]-[0086], and figures 1-3	1-20
A	CN 108172213 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 15 June 2018 (2018-06-15) entire document	1-20
A	CN 104835507 A (BOHAI UNIVERSITY) 12 August 2015 (2015-08-12) entire document	1-20
A	US 2017206915 A1 (XEROX CORP) 20 July 2017 (2017-07-20) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

19 July 2019

Date of mailing of the international search report

22 August 2019

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/089148

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	109243491	A	18 January 2019	None			
CN	108597539	A	28 September 2018	None			
CN	108172213	A	15 June 2018	None			
CN	104835507	A	12 August 2015	CN	104835507	B	16 January 2018
US	2017206915	A1	20 July 2017	US	9812154	B2	07 November 2017

国际检索报告

国际申请号

PCT/CN2019/089148

A. 主题的分类

G10L 25/63(2013.01)i; G10L 25/30(2013.01)i; G10L 25/45(2013.01)i; G10L 25/18(2013.01)i; G10L 21/0208(2013.01)i; G10L 15/06(2013.01)i; G10L 15/20(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G10L

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称,和使用的检索词(如使用))

CNABS;VEN;CNTXT;WOTXT;EPTXT;USTXT;CNKI;IEEE:情绪识别,语音,加窗,深度神经网络,卷积神经网络,长短期记忆网络,归一化,voice,emotion recogni+,window,nature network,CNN,LSTM,normalizat+

C. 相关文件

类 型*	引用文件,必要时,指明相关段落	相关的权利要求
PX	CN 109243491 A (平安科技深圳有限公司) 2019年 1月 18日 (2019 - 01 - 18) 说明书第[0036]-[0134]段,附图1-4	1-9、13-20
A	CN 108597539 A (桂林电子科技大学) 2018年 9月 28日 (2018 - 09 - 28) 说明书第[0038]-[0086]段,附图1-3	1-20
A	CN 108172213 A (北京百度网讯科技有限公司) 2018年 6月 15日 (2018 - 06 - 15) 全文	1-20
A	CN 104835507 A (渤海大学) 2015年 8月 12日 (2015 - 08 - 12) 全文	1-20
A	US 2017206915 A1 (XEROX CORP) 2017年 7月 20日 (2017 - 07 - 20) 全文	1-20

☐ 其余文件在C栏的续页中列出。

☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件,或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布,与申请不相抵触,但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件,单独考虑该文件,认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件,当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时,要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2019年 7月 19日

国际检索报告邮寄日期

2019年 8月 22日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)
中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

授权官员

颜彦

电话号码 86-(0512)-88997267

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/089148

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	109243491	A	2019年 1月 18日	无	
CN	108597539	A	2018年 9月 28日	无	
CN	108172213	A	2018年 6月 15日	无	
CN	104835507	A	2015年 8月 12日	CN 104835507	B 2018年 1月 16日
US	2017206915	A1	2017年 7月 20日	US 9812154	B2 2017年 11月 7日