(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 12 月 30 日 (30.12.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/258503 A1

(51) 国际专利分类号:
G10L 25/30 (2013.01) *G10H 1/00* (2006.01)
G10L 21/007 (2013.01) *G06K 9/46* (2006.01)

(21) 国际申请号: PCT/CN2019/103573

(22) 国际申请日: 2019 年 8 月 30 日 (30.08.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910566906.2 2019年6月27日 (27.06.2019) 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)。

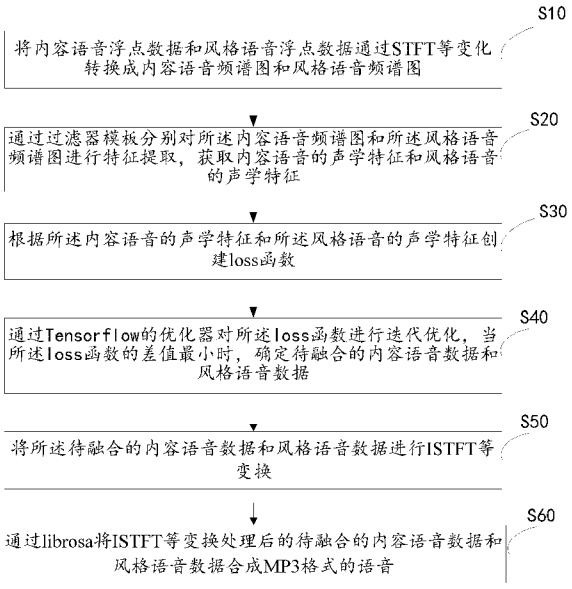
(72) 发明人: 张鹏 (ZHANG, Peng); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518033 (CN)。

(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,

(54) **Title:** TENSORFLOW-BASED VOICE FUSION METHOD, ELECTRONIC DEVICE, AND STORAGE MEDIUM

(54) **发明名称:** 基于Tensorflow的语音融合方法、电子装置及存储介质



S10 CONVERT, BY MEANS OF CHANGES SUCH AS STFT, CONTENT VOICE FLOAT DATA AND STYLE VOICE FLOAT DATA INTO A CONTENT VOICE SPECTROGRAM AND A STYLE VOICE SPECTROGRAM
S20 PERFORM, BY MEANS OF A FILTER TEMPLATE, FEATURE EXTRACTION ON THE CONTENT VOICE SPECTROGRAM AND THE STYLE VOICE SPECTROGRAM, RESPECTIVELY, TO ACQUIRE ACOUSTIC FEATURES OF CONTENT VOICE AND ACOUSTIC FEATURES OF STYLE VOICE
S30 CREATE A LOSS FUNCTION ACCORDING TO THE ACOUSTIC FEATURES OF THE CONTENT VOICE AND THE ACOUSTIC FEATURES OF THE STYLE VOICE
S40 ITERATIVELY OPTIMIZE THE LOSS FUNCTION BY MEANS OF AN OPTIMIZER OF THE TENSORFLOW, AND WHEN THE DIFFERENCE OF THE LOSS FUNCTION IS MINIMAL, DETERMINE CONTENT VOICE DATA AND STYLE VOICE DATA TO BE FUSED
S50 PERFORM CONVERSION SUCH AS ISTFT ON THE CONTENT VOICE DATA AND STYLE VOICE DATA TO BE FUSED
S60 SYNTHESIZE, BY MEANS OF LIBROSA, THE CONTENT VOICE DATA AND STYLE VOICE DATA TO BE FUSED AFTER BEING SUBJECTED TO CONVERSION PROCESSING SUCH AS ISTFT, INTO VOICE IN AN MP3 FORMAT

图 3

(57) **Abstract:** A Tensorflow-based voice fusion method, an electronic device and a computer readable storage medium. Said method comprises: converting, by means of changes such as STFT, content voice float data and style voice float data into a content voice spectrogram and a style voice spectrogram (S10); performing, by means of a filter template, feature extraction on the two voice spectrograms, respectively, to acquire acoustic features of the two voices (S20), and creating a loss function (S30); iteratively optimizing the loss function by means of an optimizer of the Tensorflow, and when the difference of the loss function is minimal, determining content voice data and style voice data to be fused (S40); performing conversion such as ISTFT on the voice data to be fused (S50); and synthesizing, by means of librosa, the content voice data and style voice data to be fused after being subjected to conversion processing

WO 2020/258503 A1

[见续页]

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))。

such as ISTFT, into voice in an MP3 format (S60). By fusing voices of different styles together on the basis of Tensorflow, background music is automatically added for speaking voice and new music is created intelligently, thereby providing users with more experience.

(57) 摘要：一种基于Tensorflow的语音融合方法、电子装置及计算机可读存储介质，其中的方法包括：将内容语音浮点数据和风格语音浮点数据通过STFT等变化转换成内容语音频谱图和风格语音频谱图（S10）；通过过滤器模板分别对两种语音频谱图进行特征提取，获取两种语音的声学特征（S20），并创建loss函数（S30）；通过Tensorflow的优化器对loss函数进行迭代优化，当loss函数的差值最小时，确定待融合的内容语音数据和风格语音数据（S40）；将待融合的语音数据进行ISTFT等变换（S50）；通过librosa将ISTFT等变换处理后的待融合的内容语音数据和风格语音数据合成MP3格式的语音（S60）。通过基于Tensorflow将不同风格的语音融合在一起，为说话声自动增加背景音乐、智能创造新音乐，从而为用户提供更多地体验。

基于 Tensorflow 的语音融合方法、电子装置及存储介质

本申请要求于2019年6月27日提交中国专利局，申请号为201910566906.2、发明名称为“基于Tensorflow的语音融合方法、电子装置及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及人工智能技术领域，尤其涉及一种基于Tensorflow的语音融合方法、电子装置及计算机可读存储介质。

背景技术

目前的语音融合为用户提供了很多需求和方便，但是在语音系统中还不能实现不同风格的语音进行融合。其中，随着深度学习技术的发展和应用，为语音融合提供可能。也就是说，如何分别捕获内容语音、风格语音文件的特征，通过深度学习，将两者的特征进行融合，再将其合成一条新语音成为一种可能。即：如何将一个语音风格迁移到另一条不同的语音。

目前为了满足用户将不同风格的语音进行融合的需求，亟需一种不同风格的语音融合的方法。

发明内容

本申请提供一种基于Tensorflow的语音融合方法、电子装置及计算机可读存储介质，其主要目的在于通过基于Tensorflow将不同风格的语音融合在一起，为说话声自动增加背景音乐、智能创造新音乐，从而为用户提供更多方便和更好地体验。

此外，为实现上述目的，本申请提供一种基于 Tensorflow 的语音融合方法，该方法包括：

将内容语音浮点数据和风格语音浮点数据通过 STFT 等变化转换成内容语音频谱图和风格语音频谱图；

通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行

特征提取，获取内容语音的声学特征和风格语音的声学特征；

根据所述内容语音的声学特征和所述风格语音的声学特征创建loss函数；

通过 Tensorflow 的优化器对所述 loss 函数进行迭代优化，当所述 loss 函数的差值最小时，确定待融合的内容语音数据和风格语音数据；

将所述待融合的内容语音数据和风格语音数据进行 ISTFT 等变换；

通过 librosa 将 ISTFT 等变换处理后的待融合的内容语音数据和风格语音数据合成 MP3 格式的语音。

为实现上述目的，本申请还提供一种电子装置，该装置包括：存储器、处理器，所述存储器中包括基于 Tensorflow 的语音融合程序，所述基于 Tensorflow 的语音融合程序被所述处理器执行时实现如下步骤：

将内容语音浮点数据和风格语音浮点数据通过 STFT 等变化转换成内容语音频谱图和风格语音频谱图；

通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，获取内容语音的声学特征和风格语音的声学特征；

根据所述内容语音的声学特征和所述风格语音的声学特征创建loss函数；

通过 Tensorflow 的优化器对所述 loss 函数进行迭代优化，当所述 loss 函数的差值最小时，确定待融合的内容语音数据和风格语音数据；

将所述待融合的内容语音数据和风格语音数据进行 ISTFT 等变换；

通过 librosa 将 ISTFT 等变换处理后的待融合的内容语音数据和风格语音数据合成 MP3 格式的语音。

此外，为实现上述目的，本申请还提供一种计算机非易失性可读存储介质，所述计算机非易失性可读存储介质中包括基于 Tensorflow 的语音融合程序，所述基于 Tensorflow 的语音融合程序被处理器执行时，实现如上所述的基于 Tensorflow 的语音融合方法中的任意步骤。

本申请提出的基于 Tensorflow 的语音融合方法、电子装置及计算机非易失性可读存储介质，通过 Tensorflow 的过滤器模板提取内容语音和风格语音的声学特征，并创建 loss 函数，通过 Tensorflow 的优化器对所述 loss 函数进行迭代优化，确定待融合的内容语音数据和风格语音数据，通过 librosa 将待融合的内容语音数据和风格语音数据进行融合，从而实现将不同风格的语音融合在一起，为说话声自动增加背景音乐、智能创造新音乐，从而为用户提

供更多方便和更好地体验。

附图说明

图1为本申请基于Tensorflow的语音融合方法较佳实施例的应用环境示意图；

图2为图1中基于Tensorflow的语音融合程序较佳实施例的模块示意图；

图3为本申请基于Tensorflow的语音融合方法较佳实施例的流程图。

本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

应当理解，此处所描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

本申请提供一种基于 Tensorflow 的语音融合方法，应用于一种电子装置 1。参照图 1 所示，为本申请基于 Tensorflow 的语音融合方法较佳实施例的应用环境示意图。

在本实施例中，电子装置 1 可以是服务器、智能手机、平板电脑、便携计算机、桌上型计算机等具有运算功能的终端设备。

该电子装置 1 包括：处理器 12、存储器 11、网络接口 14 及通信总线 15。

存储器 11 包括至少一种类型的可读存储介质。所述至少一种类型的可读存储介质可为如闪存、硬盘、多媒体卡、卡型存储器 11 等的非易失性存储介质。在一些实施例中，所述可读存储介质可以是所述电子装置 1 的内部存储单元，例如该电子装置 1 的硬盘。在另一些实施例中，所述可读存储介质也可以是所述电子装置 1 的外部存储器 11，例如所述电子装置 1 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。

在本实施例中，所述存储器 11 的可读存储介质通常用于存储安装于所述电子装置 1 的基于 Tensorflow 的语音融合程序 10、过滤器模板及 TensorFlow 的优化器等。所述存储器 11 还可以用于暂时地存储已经输出或者将要输出的数据。

处理器 12 在一些实施例中可以是一中央处理器（Central Processing Unit, CPU），微处理器或其他数据处理芯片，用于运行存储器 11 中存储的程序代码或处理数据，例如执行人物情绪分析程序 10 等。

网络接口 14 可选地可以包括标准的有线接口、无线接口（如 WI-FI 接口），通常用于在该电子装置 1 与其他电子设备之间建立通信连接。

通信总线 15 用于实现这些组件之间的连接通信。

图 1 仅示出了具有组件 11-15 的电子装置 1，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。

可选地，该电子装置 1 还可以包括用户接口，用户接口可以包括输入单元比如键盘（Keyboard）、语音输入装置比如麦克风（microphone）等具有语音识别功能的设备、语音输出装置比如音响、耳机等，可选地用户接口还可以包括标准的有线接口、无线接口。

可选地，该电子装置 1 还可以包括显示器，显示器也可以称为显示屏或显示单元。在一些实施例中可以是 LED 显示器、液晶显示器、触控式液晶显示器以及有机发光二极管（Organic Light-Emitting Diode，OLED）触摸器等。显示器用于显示在电子装置 1 中处理的信息以及用于显示可视化的用户界面。

可选地，该电子装置 1 还包括触摸传感器。所述触摸传感器所提供的供用户进行触摸操作的区域称为触控区域。此外，这里所述的触摸传感器可以为电阻式触摸传感器、电容式触摸传感器等。而且，所述触摸传感器不仅包括接触式的触摸传感器，也可包括接近式的触摸传感器等。此外，所述触摸传感器可以为单个传感器，也可以为例如阵列布置的多个传感器。

此外，该电子装置 1 的显示器的面积可以与所述触摸传感器的面积相同，也可以不同。可选地，将显示器与所述触摸传感器层叠设置，以形成触摸显示屏。该装置基于触摸显示屏侦测用户触发的触控操作。

可选地，该电子装置 1 还可以包括射频（Radio Frequency，RF）电路，传感器、音频电路等等，在此不再赘述。

在图 1 所示的装置实施例中，作为一种计算机存储介质的存储器 11 中可以包括操作系统、以及基于 Tensorflow 的语音融合程序 10；处理器 12 执行存储器 11 中存储的基于 Tensorflow 的语音融合程序 10 时实现如下步骤：

将内容语音浮点数据和风格语音浮点数据通过 STFT 等变化转换成内容

语音频谱图和风格语音频谱图；

通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，获取内容语音的声学特征和风格语音的声学特征；

根据所述内容语音的声学特征和所述风格语音的声学特征创建loss函数；

通过Tensorfow的优化器对所述loss函数进行迭代优化，当所述loss函数的差值最小时，确定待融合的内容语音数据和风格语音数据；

将所述待融合的内容语音数据和风格语音数据进行ISTFT等变换；

通过librosa将ISTFT等变换处理后的待融合的内容语音数据和风格语音数据合成MP3格式的语音。

在本申请中，在将内容语音数据和风格语音数据通过STFT等变化转换成内容语音频谱图和风格语音频谱图的步骤之前，还包括如下步骤：

通过librosa对内容语音和风格语音进行加载处理；

将所述内容语音转换为所述内容语音浮点数据，将所述风格语音转换为所述风格语音浮点数据。

其中，librosa是一个用于音频、音乐分析、处理的python工具包，一些常见的时频处理、特征提取、绘制声音图形等功能应有尽有，功能十分强大。

加载两条语音的目的是想明确所操作的语音是哪个，即想要融合的内容语音和风格语音各是什么。

语音本身是wav格式的文件，加载是需要把其中wav格式转化为的数据（float浮点数据）放到计算机内存中，以便后续的步骤。

接下来，将获取的内容语音浮点数据和风格语音浮点数据通过STFT等变化转换成内容语音频谱图和风格语音频谱图。其中，STFT（短时叶变换）是和傅里叶变换相关的一种数学变换，用以确定时变信号其局部区域正弦波的频率与相位。声音频谱图就是使用STFT绘制而成的，其实质就是STFT的能量，也就是模值的平方。

在本申请中，调用librosa的STFT函数实现浮点数据到频谱图的转换。即：通过调用librosa的STFT函数将内容语音和风格语音由浮点数据转化为内容语音频谱图和风格语音频谱图。

其中，通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，获取内容语音的声学特征和风格语音的声学特征。即：新

建一个过滤器模板，分别捕获出内容语音和风格语音的声学特征。

过滤器模板其实是对上述得到的频谱做特征提取，采用的方式是用卷积神经网络的 tf.nn.conv2d 方法，其中卷积核是自定义（如：随机生成一定维度的数组乘以一个系数），就像在更高纬度抽取一些数值来代表原来更大、更多信息的语音。内容语音和风格语音特征当前抽取还不是非常准确，但可以近似的生成融合后的语音，即融合后的新语音有分别来自原来内容语音、风格语音的内容。

具体地，tf.nn.conv2d 是 Tensorflow 里面实现卷积的函数，这是搭建卷积神经网络比较核心的一个方法。

tf.nn.conv2d(input，filter，strides，padding，use_cudnn_on_gpu=None，name=None)

除去 name 参数用以指定该操作的 name，一共五个参数：

第一个参数 input：指需要做卷积的输入图像，它要求是一个 Tensor，具有[batch， in_height， in_width，in_channels]这样的 shape，具体含义是[训练时一个 batch 的图片数量，图片高度，图片宽度，图像通道数]，注意这是一个 4 维的 Tensor，要求类型为 float32 和 float64 其中之一；

第二个参数 filter：相当于 CNN 中的卷积核，它要求是一个 Tensor，具有[filter_height, filter_width, in_channels, out_channels]这样的 shape，具体含义是[卷积核的高度，卷积核的宽度，图像通道数，卷积核个数]，要求类型与参数 input 相同，有一个地方需要注意，第三维 in_channels，就是参数 input 的第四维；

第三个参数 strides：卷积时在图像每一维的步长，这是一个一维的向量，长度 4；

第四个参数 padding：string 类型的量，只能是"SAME","VALID"其中之一，这个值决定了不同的卷积方式；

第五个参数：use_cudnn_on_gpu:bool 类型，是否使用 cudnn 加速，默认为 true。

结果返回一个 Tensor，这个输出。

接下来，建一个变量 X 列表，并根据声学特征建立一个 loss 函数。新建 X 列表就是初始一个列表变量 （也可认为是建立一个 Tensorflow 的张量），

它的数据代表着新生成融合语音的数据。只不过初始化时，值为空，要通过机器学习来修改里面的数据，最终逼近真正合成语音的数值。

其中，所述 loss 函数的公式如下：

$$l_{total}(a,x,p)=\alpha*l_{style}(a,x)+\beta*l_{content}(p,x)$$

其中，l_{total} 表示全部差值；l_{style} 表示风格语音差值；$l_{content}$ 表示内容语音差值；α 表示风格语音参数；β 表示内容语音参数；a 表示风格语音特征值；p 表示内容语音特征值；x 表示变量特征值。

当 x 的值达到某一个值时，此时，获取的 l_{total} 为最小（loss 函数 L，不一定要等于 0，也可以为 0，只要其在变小并且有收敛于某一值即可），x 的值约等于风格语音特征值，以及约等于内容语音特征值，其中，a 和 p 可以看做为提取的特征值。

根据获取的 l_{total} 为最小值时，确定 X 的值，从而确定融合语音的数据。

其中，所述通过 Tensorflow 的优化器对所述 loss 函数进行迭代优化的步骤包括：

通过改变 loss 函数的变量 x ，获取 l_{total} 值；

根据获取的 l_{total} 的最小值，确定 loss 函数的变量 x 值；

根据确定的 loss 函数的变量 x 值，确定内容语音特征值和风格语音特征值，进而确定待融合的内容语音数据和风格语音数据。

通过 Tensorflow 中的优化器对 loss 函数的行数进行迭代优化，使用 tf.contrib.optScipyOptimizerInterface 方法，具体的优化器是：L-BFGS-B。

简单说，就是改变变量 X 列表里面数值，同时使上面的损失函数 loss 的差值达到最小。

其中，tf.contrib.optScipyOptimizerInterface 这个接口提供了实现，只要在入参中指定 method= ‘L-BFGS-B’，Tensorflow 就会直接调用，具体的算法内

容是现有技术，在此不再做详述。

其中，将所述待融合的内容语音数据和风格语音数据进行 ISTFT 等变换。此过程是“将内容语音浮点数据和风格语音浮点数据通过 STFT 等变化转换成内容语音频谱图和风格语音频谱图”的过程的逆操作，将获取的待融合的内容语音数据和风格语音数据，将获取的语音的频域转换成时域，最终由时域来重新生成语音。

需要说明的是，对信号进行时域分析时，有时一些信号的时域参数相同，但并不能说明信号就完全相同。比如具有相同函数结构的两个信号可能并不相同，因为信号不仅随时间变化，还与频率、相位等信息有关，这就需要进一步分析信号的频率结构，并在频率域中对信号进行描述。

动态信号从时间域变换到频率域主要通过傅立叶级数和傅立叶变换实现。周期信号靠傅立叶级数，非周期信号靠傅立叶变换。很简单时域分析的函数是参数是 t，也就是 y=f(t)，频域分析时，参数是 w，也就是 y=F(w)两者之间可以互相转化。时域函数通过傅立叶或者拉普拉斯变换就变成了频域函数。

傅里叶变换可以将一个时域信号转换成在不同频率下对应的振幅及相位，其频谱就是时域信号在频域下的表现，而反傅里叶变换可以将频谱再转换回时域的信号。

上述实施例提出的电子装置 1，通过 Tensorflow 的过滤器模板提取内容语音和风格语音的声学特征，并创建 loss 函数，通过 Tensorflow 的优化器对所述 loss 函数进行迭代优化，确定待融合的内容语音数据和风格语音数据，通过 librosa 将待融合的内容语音数据和风格语音数据进行融合，从而实现将不同风格的语音融合在一起，为说话声自动增加背景音乐、智能创造新音乐，从而为用户提供更多方便和更好地体验。

在其他实施例中，基于 Tensorflow 的语音融合程序 10 还可以被分割为一个或者多个模块，一个或者多个模块被存储于存储器 11 中，并由处理器 12 执行，以完成本申请。本申请所称的模块是指能够完成特定功能的一系列计算机程序指令段。参照图 2 所示，为图 1 中基于 Tensorflow 的语音融合程序 10 较佳实施例的程序模块图。所述基于 Tensorflow 的语音融合程序 10 可以被分割为：频谱图获取模块 110、声学特征获取模块 120、loss 函数创建模块 130、待融合语音数据获取模块 140、ISTFT 等变换模块 150 和语音融合模块 160。

所述模块 110-160 所实现的功能或操作步骤均与上文类似，此处不再详述，示例性地，例如其中：

频谱图获取模块 110，用于将内容语音浮点数据和风格语音浮点数据通过 STFT 等变化转换成内容语音频谱图和风格语音频谱图；

声学特征获取模块 120，用于通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，获取内容语音的声学特征和风格语音的声学特征；

loss 函数创建模块 130，用于根据所述内容语音的声学特征和所述风格语音的声学特征创建 loss 函数；

待融合语音数据获取模块 140，用于通过 Tensorflow 的优化器对所述 loss 函数进行迭代优化，当所述 loss 函数的差值最小时，确定待融合的内容语音数据和风格语音数据；

ISTFT 等变换模块 150，用于将所述待融合的内容语音数据和风格语音数据进行 ISTFT 等变换；

语音融合模块 160，用于通过 librosa 将 ISTFT 等变换处理后的待融合的内容语音数据和风格语音数据合成 MP3 格式的语音。

优选地，基于 Tensorflow 的语音融合程序 10 还包括语音加载模块。其中，所述语音加载模块，用于通过 librosa 对内容语音和风格语音进行加载处理；将所述内容语音转换为所述内容语音浮点数据，将所述风格语音转换为所述风格语音浮点数据。

此外，本申请还提供一种基于 Tensorflow 的语音融合方法。参照图 3 所示，为本申请基于 Tensorflow 的语音融合方法较佳实施例的流程图。该方法可以由一个装置执行，该装置可以由软件和/或硬件实现。

在本实施例中，基于 Tensorflow 的语音融合方法，包括：步骤 S10-步骤 S60。

步骤 S10：将内容语音浮点数据和风格语音浮点数据通过 STFT 等变化转换成内容语音频谱图和风格语音频谱图；

步骤 S20：通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，获取内容语音的声学特征和风格语音的声学特征；

步骤 S30：根据所述内容语音的声学特征和所述风格语音的声学特征创建

loss 函数；

步骤 S40：通过 Tensorflow 的优化器对所述 loss 函数进行迭代优化，当所述 loss 函数的差值最小时，确定待融合的内容语音数据和风格语音数据；

步骤 S50：将所述待融合的内容语音数据和风格语音数据进行 ISTFT 等变换；

步骤 S60：通过 librosa 将 ISTFT 等变换处理后的待融合的内容语音数据和风格语音数据合成 MP3 格式的语音。

在本申请中在步骤 S10 之前还包括以下步骤：

通过 librosa 对内容语音和风格语音进行加载处理；

将所述内容语音转换为所述内容语音浮点数据，将所述风格语音转换为所述风格语音浮点数据。

其中，librosa 是一个用于音频、音乐分析、处理的 python 工具包，一些常见的时频处理、特征提取、绘制声音图形等功能应有尽有，功能十分强大。

加载两条语音的目的是想明确所操作的语音是哪个，即想要融合的内容语音和风格语音各是什么。

语音本身是 wav 格式的文件，加载是需要把其中 wav 格式转化为的数据（float 浮点数据）放到计算机内存中，以便后续的步骤。

在步骤 S10 中，将获取的内容语音浮点数据和风格语音浮点数据通过 STFT 等变化转换成内容语音频谱图和风格语音频谱图。其中，STFT（短时叶变换）是和傅里叶变换相关的一种数学变换，用以确定时变信号其局部区域正弦波的频率与相位。声音频谱图就是使用 stft 绘制而成的，其实质就是 stft 的能量，也就是模值的平方。

在本申请的实施例中，调用 librosa 的 stft 函数实现浮点数据到频谱图的转换。即：通过调用 librosa 的 stft 函数将内容语音和风格语音由浮点数据转化为内容语音频谱图和风格语音频谱图。

在步骤 S20 中，通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，获取内容语音的声学特征和风格语音的声学特征。即：新建一个过滤器模板，分别捕获出内容语音和风格语音的声学特征。

过滤器模板其实是对步骤 S10 得到的频谱做特征提取，采用的方式是用卷积神经网络的 tf.nn.conv2d 方法，其中卷积核是自定义（如：随机生成一定

维度的数组乘以一个系数），就像在更高纬度抽取一些数值来代表原来更大、更多信息的语音。内容语音和风格语音特征当前抽取还不是非常准确，但可以近似的生成融合后的语音，即融合后的新语音有分别来自原来内容语音、风格语音的内容。

具体地，tf.nn.conv2d 是 Tensorflow 里面实现卷积的函数，这是搭建卷积神经网络比较核心的一个方法。

tf.nn.conv2d(input，filter，strides，padding，use_cudnn_on_gpu=None，name=None)

除去 name 参数用以指定该操作的 name，一共五个参数：

第一个参数 input：指需要做卷积的输入图像，它要求是一个 Tensor，具有[batch， in_height， in_width，in_channels]这样的 shape，具体含义是[训练时一个 batch 的图片数量，图片高度，图片宽度，图像通道数]，注意这是一个 4 维的 Tensor，要求类型为 float32 和 float64 其中之一；

第二个参数 filter：相当于 CNN 中的卷积核，它要求是一个 Tensor，具有[filter_height, filter_width, in_channels, out_channels]这样的 shape，具体含义是[卷积核的高度，卷积核的宽度，图像通道数，卷积核个数]，要求类型与参数 input 相同，有一个地方需要注意，第三维 in_channels，就是参数 input 的第四维；

第三个参数 strides：卷积时在图像每一维的步长，这是一个一维的向量，长度 4；

第四个参数 padding：string 类型的量，只能是"SAME","VALID"其中之一，这个值决定了不同的卷积方式；

第五个参数：use_cudnn_on_gpu:bool 类型，是否使用 cudnn 加速，默认为 true。

结果返回一个 Tensor，这个输出。

在步骤 S30 中，根据所述内容语音的声学特征和所述风格语音的声学特征创建 loss 函数接下来。

在本申请的实施例中，建一个变量 X 列表，并根据声学特征建立一个 loss 函数。新建 X 列表就是初始一个列表变量 （也可认为是建立一个 tensorflow 的张量），它的数据代表着新生成融合语音的数据。只不过初始化时，值为空，

要通过机器学习来修改里面的数据，最终逼近真正合成语音的数值。

其中，所述 loss 函数的公式如下：

$$l_{total}(a,x,p)=\alpha*l_{style}(a,x)+\beta*l_{content}(p,x)$$

其中，l_{total} 表示全部差值；l_{style} 表示风格语音差值；$l_{content}$ 表示内容语音差值；α 表示风格语音参数；β 表示内容语音参数；a 表示风格语音特征值；p 表示内容语音特征值；x 表示变量特征值。

当 x 的值达到某一个值时，此时，获取的 l_{total} 为最小（loss 函数 L，不一定要等于 0，也可以为 0，只要其在变小并且有收敛于某一值即可），x 的值约等于风格语音特征值，以及约等于内容语音特征值，其中，a 和 p 可以看做为提取的特征值。

根据获取的 l_{total} 为最小值时，确定 X 的值，从而确定融合语音的数据。

步骤 S40：通过 Tensorflow 的优化器对所述 loss 函数进行迭代优化，当所述 loss 函数的差值最小时，确定待融合的内容语音数据和风格语音数据。

在本申请的实施例中，所述通过 Tensorfow 的优化器对所述 loss 函数进行迭代优化的步骤包括：

通过改变 loss 函数的变量 x ，获取 l_{total} 值；

根据获取的 l_{total} 的最小值，确定 loss 函数的变量 x 值；

根据确定的 loss 函数的变量 x 值，确定内容语音特征值和风格语音特征值，进而确定待融合的内容语音数据和风格语音数据。

通过 Tensorflow 中的优化器对 loss 函数行数进行迭代优化，使用 tf.contrib.optScipyOptimizerInterface 方法，具体的优化器是：L-BFGS-B。

简单说，就是改变变量 X 列表里面数值，同时使上面的损失函数 loss 达到最小。

其中，tf.contrib.optScipyOptimizerInterface 这个接口提供了实现，只要在

入参中指定 method= ‘L-BFGS-B’，tensorflow 就会直接调用，具体的算法内容是现有技术，在此不再做详述。

步骤 S50：将所述待融合的内容语音数据和风格语音数据进行 ISTFT 等变换。

其中，在本申请的实施例中将所述待融合的内容语音数据和风格语音数据进行 ISTFT 等变换。此过程是步骤 S10 的过程的逆操作，将获取的待融合的内容语音数据和风格语音数据，将获取的语音的频域转换成时域，最终由时域来重新生成语音。

需要说明的是，对信号进行时域分析时，有时一些信号的时域参数相同，但并不能说明信号就完全相同。比如具有相同函数结构的两个信号可能并不相同，因为信号不仅随时间变化，还与频率、相位等信息有关，这就需要进一步分析信号的频率结构，并在频率域中对信号进行描述。

动态信号从时间域变换到频率域主要通过傅立叶级数和傅立叶变换实现。周期信号靠傅立叶级数，非周期信号靠傅立叶变换。很简单时域分析的函数是参数是 t，也就是 y=f(t)，频域分析时，参数是 w，也就是 y=F(w)两者之间可以互相转化。时域函数通过傅立叶或者拉普拉斯变换就变成了频域函数。

傅里叶变换可以将一个时域信号转换成在不同频率下对应的振幅及相位，其频谱就是时域信号在频域下的表现，而反傅里叶变换可以将频谱再转换回时域的信号。

上述实施例提出的基于 Tensorflow 的语音融合方法，通过 Tensorflow 的过滤器模板提取内容语音和风格语音的声学特征，并创建 loss 函数，通过 Tensorflow 的优化器对所述 loss 函数的进行迭代优化，确定待融合的内容语音数据和风格语音数据，通过 librosa 将待融合的内容语音数据和风格语音数据进行融合，从而实现将不同风格的语音融合在一起，为说话声自动增加背景音乐、智能创造新音乐，从而为用户提供更多方便和更好地体验。

此外，本申请实施例还提出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基于 Tensorflow 的语音融合程序，所述基于 Tensorflow 的语音融合程序被处理器执行时实现如下操作：

将内容语音浮点数据和风格语音浮点数据通过 STFT 等变化转换成内容语音频谱图和风格语音频谱图；

通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，获取内容语音的声学特征和风格语音的声学特征；

根据所述内容语音的声学特征和所述风格语音的声学特征创建loss函数；

通过Tensorflow的优化器对所述loss函数进行迭代优化，当所述loss函数的差值最小时，确定待融合的内容语音数据和风格语音数据；

将所述待融合的内容语音数据和风格语音数据进行ISTFT等变换；

通过librosa将ISTFT等变换处理后的待融合的内容语音数据和风格语音数据合成MP3格式的语音。

优选地，在将内容语音数据和风格语音数据通过STFT等变化转换成内容语音频谱图和风格语音频谱图的步骤之前，还包括如下步骤：

通过librosa对内容语音和风格语音进行加载处理；

将所述内容语音转换为所述内容语音浮点数据，将所述风格语音转换为所述风格语音浮点数据。

优选地，所述通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取的步骤包括：

所述过滤器模板通过卷积神经网络的tf.nn.conv2d分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，其中，

所述卷积神经网络的tf.nn.conv2d为基于Tensorflow实现的卷积神经网络的函数。

优选地，所述loss函数的公式如下：

$$l_{total}(a,x,p)=\alpha * l_{style}(a,x)+\beta * l_{content}(p,x)$$

其中，l_{total} 表示全部差值；l_{style} 表示风格语音差值；

$l_{content}$ 表示内容语音差值；α 表示风格语音参数；

β 表示内容语音参数；a 表示风格语音特征值；

p 表示内容语音特征值；x 表示变量特征值。

优选地，所述通过 TensorFlow 的优化器对所述 loss 函数进行迭代优化的步骤包括：

通过改变 loss 函数的变量 x ，获取 l_{total} 值；

根据获取的 l_{total} 的最小值，确定 loss 函数的变量 x 值；

根据确定的 loss 函数的变量 x 值，确定内容语音特征值和风格语音特征值，进而确定待融合的内容语音数据和风格语音数据。

优选地，所述 Tensorflow 的优化器为 L-BFGS-B 算法。

本申请之计算机可读存储介质的具体实施方式与上述基于 Tensorflow 的语音融合方法、电子装置的具体实施方式大致相同，在此不再赘述。

需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、装置、物品或者方法不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、装置、物品或者方法所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、装置、物品或者方法中还存在另外的相同要素。

上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在如上所述的一个存储介质(如 ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，或者网络设备等)执行本申请各个实施例所述的方法。

以上仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内。

权 利 要 求 书

1、一种基于 Tensorflow 的语音融合方法，应用于电子装置，其特征在于，所述方法包括：

将内容语音浮点数据和风格语音浮点数据通过 STFT 等变化转换成内容语音频谱图和风格语音频谱图；

通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，获取内容语音的声学特征和风格语音的声学特征；

根据所述内容语音的声学特征和所述风格语音的声学特征创建 loss 函数；

通过 Tensorflow 的优化器对所述 loss 函数进行迭代优化，当所述 loss 函数的差值最小时，确定待融合的内容语音数据和风格语音数据；

将所述待融合的内容语音数据和风格语音数据进行 ISTFT 等变换；

通过 librosa 将 ISTFT 等变换处理后的待融合的内容语音数据和风格语音数据合成 MP3 格式的语音。

2、根据权利要求 1 所述的基于 Tensorflow 的语音融合方法，其特征在于，在将内容语音数据和风格语音数据通过 STFT 等变化转换成内容语音频谱图和风格语音频谱图的步骤之前，还包括如下步骤：

通过 librosa 对内容语音和风格语音进行加载处理；

将所述内容语音转换为所述内容语音浮点数据，将所述风格语音转换为所述风格语音浮点数据。

3、根据权利要求 1 所述的基于 Tensorflow 的语音融合方法，其特征在于，所述通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取的步骤包括：

所述过滤器模板通过卷积神经网络的 tf.nn.conv2d 分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，其中，

所述卷积神经网络的 tf.nn.conv2d 为基于 Tensorflow 实现的卷积神经网络的函数。

4、根据权利要求 2 所述的基于 Tensorflow 的语音融合方法，其特征在于，所述通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取的步骤包括：

所述过滤器模板通过卷积神经网络的 tf.nn.conv2d 分别对所述内容语音

频谱图和所述风格语音频谱图进行特征提取，其中，

所述卷积神经网络的 tf.nn.conv2d 为基于 Tensorflow 实现的卷积神经网络的函数。

5、根据权利要求 1 所述的基于 Tensorflow 的语音融合方法，其特征在于，所述 loss 函数的公式如下：

$$l_{total}(a,x,p)=\alpha*l_{style}(a,x)+\beta*l_{content}(p,x)$$

其中，l_{total} 表示全部差值；l_{style} 表示风格语音差值；

$l_{content}$ 表示内容语音差值；α 表示风格语音参数；

β 表示内容语音参数；a 表示风格语音特征值；

p 表示内容语音特征值；x 表示变量特征值。

6、根据权利要求 5 所述的基于 Tensorflow 的语音融合方法，其特征在于，所述通过 Tensorflow 的优化器对所述 loss 函数进行迭代优化的步骤包括：

通过改变 loss 函数的变量 x ，获取 l_{total} 值；

根据获取的 l_{total} 的最小值，确定 loss 函数的变量 x 值；

根据确定的 loss 函数的变量 x 值，确定内容语音特征值和风格语音特征值，进而确定待融合的内容语音数据和风格语音数据。

7、根据权利要求 1 所述的基于 Tensorflow 的语音融合方法，其特征在于，所述 Tensorflow 的优化器为 L-BFGS-B 算法。

8、一种电子装置，其特征在于，该电子装置包括：存储器、处理器，所述存储器中包括基于 Tensorflow 的语音融合程序，所述基于 Tensorflow 的语音融合程序被所述处理器执行时实现如下步骤：

将内容语音浮点数据和风格语音浮点数据通过 STFT 等变化转换成内容语音频谱图和风格语音频谱图；

通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，获取内容语音的声学特征和风格语音的声学特征；

根据所述内容语音的声学特征和所述风格语音的声学特征创建 loss 函数；

通过 Tensorflow 的优化器对所述 loss 函数进行迭代优化，当所述 loss 函数的差值最小时，确定待融合的内容语音数据和风格语音数据；

将所述待融合的内容语音数据和风格语音数据进行 ISTFT 等变换；

通过 librosa 将 ISTFT 等变换处理后的待融合的内容语音数据和风格语音数据合成 MP3 格式的语音。

9、根据权利要求 8 所述的电子装置，其特征在于，在将内容语音数据和风格语音数据通过 STFT 等变化转换成内容语音频谱图和风格语音频谱图的步骤之前，还包括如下步骤：

通过 librosa 对内容语音和风格语音进行加载处理；

将所述内容语音转换为所述内容语音浮点数据，将所述风格语音转换为所述风格语音浮点数据。

10、根据权利要求 8 所述的电子装置，其特征在于，所述通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取的步骤包括：

所述过滤器模板通过卷积神经网络的 tf.nn.conv2d 分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，其中，

所述卷积神经网络的 tf.nn.conv2d 为基于 Tensorflow 实现的卷积神经网络的函数。

11、根据权利要求 9 所述的电子装置，所述通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取的步骤包括：

所述过滤器模板通过卷积神经网络的 tf.nn.conv2d 分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，其中，

所述卷积神经网络的 tf.nn.conv2d 为基于 Tensorflow 实现的卷积神经网络的函数。

12、根据权利要求 8 所述的电子装置，其特征在于，所述 loss 函数的公

式如下：

$$l_{total}(a,x,p)=\alpha*l_{style}(a,x)+\beta*l_{content}(p,x)$$

其中，l_{total} 表示全部差值；l_{style} 表示风格语音差值；

$l_{content}$ 表示内容语音差值；α 表示风格语音参数；

β 表示内容语音参数；a 表示风格语音特征值；

p 表示内容语音特征值；x 表示变量特征值。

13、根据权利要求 12 所述的电子装置，其特征在于，所述通过 TensorFlow 的优化器对所述 loss 函数进行迭代优化的步骤包括：

通过改变 loss 函数的变量 x ，获取 l_{total} 值；

根据获取的 l_{total} 的最小值，确定 loss 函数的变量 x 值；

根据确定的 loss 函数的变量 x 值，确定内容语音特征值和风格语音特征值，进而确定待融合的内容语音数据和风格语音数据。

14、一种计算机非易失性可读存储介质，其特征在于，所述计算机非易失性可读存储介质中包括基于 Tensorflow 的语音融合程序，所述基于 Tensorflow 的语音融合程序被处理器执行时，实现如下步骤：

将内容语音浮点数据和风格语音浮点数据通过 STFT 等变化转换成内容语音频谱图和风格语音频谱图；

通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，获取内容语音的声学特征和风格语音的声学特征；

根据所述内容语音的声学特征和所述风格语音的声学特征创建 loss 函数；

通过 Tensorflow 的优化器对所述 loss 函数进行迭代优化，当所述 loss 函

数的差值最小时，确定待融合的内容语音数据和风格语音数据；

将所述待融合的内容语音数据和风格语音数据进行 ISTFT 等变换；

通过librosa将ISTFT等变换处理后的待融合的内容语音数据和风格语音数据合成 MP3 格式的语音。

15、根据权利要求 14 所述的计算机非易失性可读存储介质，其特征在于，在将内容语音数据和风格语音数据通过 STFT 等变化转换成内容语音频谱图和风格语音频谱图的步骤之前，还包括如下步骤：

通过 librosa 对内容语音和风格语音进行加载处理；

将所述内容语音转换为所述内容语音浮点数据，将所述风格语音转换为所述风格语音浮点数据。

16、根据权利要求 14 所述的计算机非易失性可读存储介质，其特征在于，所述通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取的步骤包括：

所述过滤器模板通过卷积神经网络的 tf.nn.conv2d 分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，其中，

所述卷积神经网络的 tf.nn.conv2d 为基于 Tensorflow 实现的卷积神经网络的函数。

17、根据权利要求 15 所述的计算机非易失性可读存储介质，其特征在于，所述通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取的步骤包括：

所述过滤器模板通过卷积神经网络的 tf.nn.conv2d 分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，其中，

所述卷积神经网络的 tf.nn.conv2d 为基于 Tensorflow 实现的卷积神经网络的函数。

18、根据权利要求 14 所述的计算机非易失性可读存储介质，其特征在于，所述 loss 函数的公式如下：

$$l_{total}(a,x,p)=\alpha * l_{style}(a,x)+\beta * l_{content}(p,x)$$

其中，l_{total} 表示全部差值；l_{style} 表示风格语音差值；

$l_{content}$ 表示内容语音差值；α 表示风格语音参数；

β 表示内容语音参数；a 表示风格语音特征值；

p 表示内容语音特征值；x 表示变量特征值。

19、根据权利要求 18 所述的计算机非易失性可读存储介质，其特征在于，所述通过 Tensorflow 的优化器对所述 loss 函数进行迭代优化的步骤包括：

通过改变 loss 函数的变量 x ，获取 l_{total} 值；

根据获取的 l_{total} 的最小值，确定 loss 函数的变量 x 值；

根据确定的 loss 函数的变量 x 值，确定内容语音特征值和风格语音特征值，进而确定待融合的内容语音数据和风格语音数据。

20、根据权利要求 14 所述的计算机非易失性可读存储介质，其特征在于，所述 Tensorflow 的优化器为 L-BFGS-B 算法。

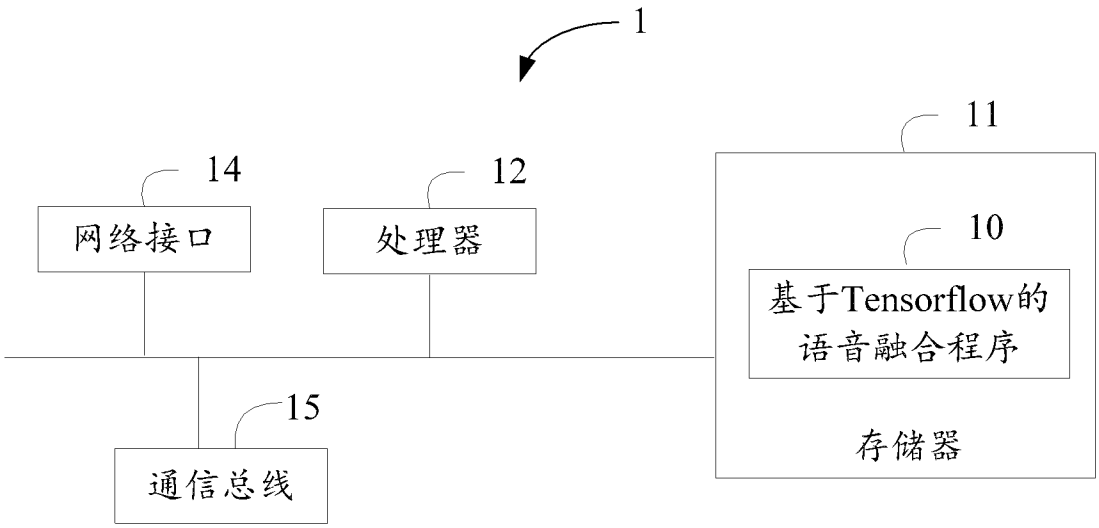

1
11
14
12
网络接口
处理器
10
基于Tensorflow的
语音融合程序
15
通信总线
存储器

图 1


基于Tensorflow的语音融合程序10
频谱图获取模块 110
声学特征获取模块 120
loss函数创建模块 130
待融合语音数据获取模块 140
ISTFT等变换模块 150
语音融合模块 160

图 2

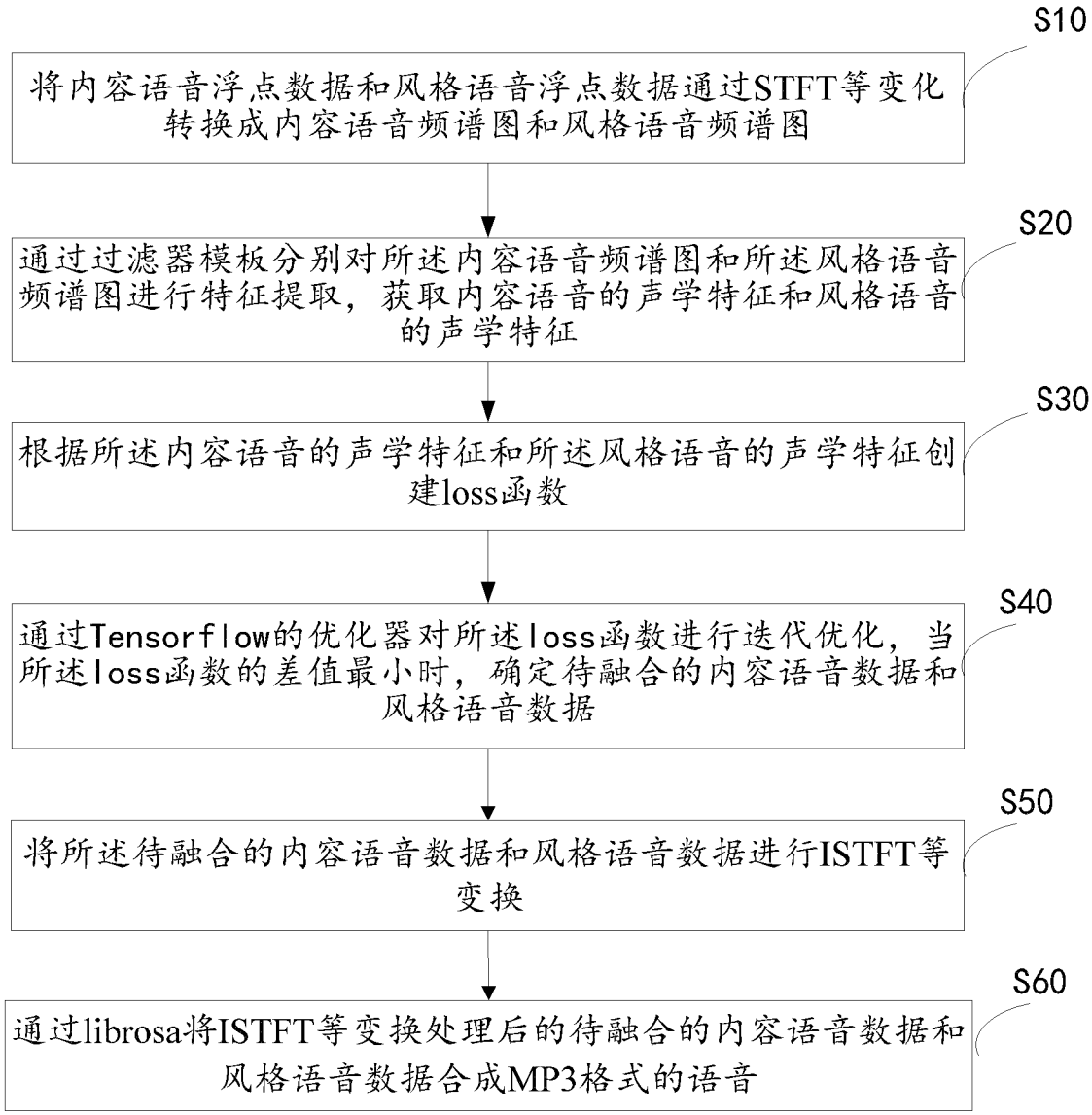

S10
将内容语音浮点数据和风格语音浮点数据通过STFT等变化转换成内容语音频谱图和风格语音频谱图
S20
通过过滤器模板分别对所述内容语音频谱图和所述风格语音频谱图进行特征提取，获取内容语音的声学特征和风格语音的声学特征
S30
根据所述内容语音的声学特征和所述风格语音的声学特征创建loss函数
S40
通过Tensorflow的优化器对所述loss函数进行迭代优化，当所述loss函数的差值最小时，确定待融合的内容语音数据和风格语音数据
S50
将所述待融合的内容语音数据和风格语音数据进行ISTFT等变换
S60
通过librosa将ISTFT等变换处理后的待融合的内容语音数据和风格语音数据合成MP3格式的语音

图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/103573

A. CLASSIFICATION OF SUBJECT MATTER

G10L 25/30(2013.01)i; G10L 21/007(2013.01)i; G10H 1/00(2006.01)i; G06K 9/46(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; G10H; G06N; G06K

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, CNTXT, CNKI, 万方, WANFANG, VEN, WOTXT, EPTXT, USTXT, IEEE, GOOGLE: 语音, 音频, 音乐, 风格迁移, 语音克隆, 语音转换, 语谱图, 内容, 风格, 损失, 代价, voice, music, tensorflow, style transfer, voice clone, voice conversion, STFT, spectrogram, content, style, loss

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Baris Demirdelen. "Neural Style Transfer on Audio" *MSc Artificial Intelligence Master Thesis of UNIVERSITY OF AMSTERDAM,* 31 August 2017 (2017-08-31), sections 2.1, 2.3.2, 3.3, 4.3, 5.2, and 5.3	1-20
X	CN 106847294 A (BAIDU ONLINE NETWORK TECHNOLOGY BEIJING CO., LTD.) 13 June 2017 (2017-06-13) claims 1-20, description, paragraphs 0045-0099, and figures 4 and 5	1-20
X	TOMCZAK, M. et al. "Audio Style Transfer With Rhythmic Constraints" *Proceedings of the 21st International Conference on Digital Audio Effects (DAFx-18),* 08 September 2018 (2018-09-08), sections 1.1 and 2	1-20
X	Anonymous authors Paper under double-blind review. ""STYLE" TRANSFER FOR MUSICAL AUDIO USING MULTIPLE TIME-FREQUENCY REPRESENTATIONS" *Under review as a conference paper at ICLR 2018,* 31 December 2018 (2018-12-31), sections 2-4	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	Date of mailing of the international search report
19 March 2020	**27 March 2020**
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

International application No.
PCT/CN2019/103573

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	GRINSTEIN, E. et al. "Audio Style Transfer" *2018 IEEE International Conference on Acoustics , Speech and Signal Processing (ICASSP)*, 20 April 2018 (2018-04-20), ISSN: 2379-190X, pp. 586-590	1-20

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/103573

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 106847294 A	13 June 2017	CN 106847294 B	30 November 2018
		US 10192163 B2	29 January 2019
		US 2018204121 A1	19 July 2018

国际检索报告

国际申请号
PCT/CN2019/103573

A. 主题的分类

G10L 25/30(2013.01)i; G10L 21/007(2013.01)i; G10H 1/00(2006.01)i; G06K 9/46(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; G10H; G06N; G06K

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, CNTXT, CNKI,万方,VEN, WOTXT, EPTXT, USTXT, IEEE, GOOGLE: 语音，音频，音乐，风格迁移，语音克隆，语音转换，语谱图，内容，风格，损失，代价，voice, music, tensorflow, style transfer, voice clone, voice conversion, STFT, spectrogram, content, style, loss

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
X	Baris Demirdelen. "Neural Style Transfer on Audio" MSc Artificial Intelligence Master Thesis of UNIVERSITY OF AMSTERDAM, 2017年 8月 31日 (2017-08-31), 第2.1、2.3.2、3.3、4.3、5.2、5.3节	1-20
X	CN 106847294 A (百度在线网络技术北京有限公司) 2017年 6月 13日 (2017-06-13) 权利要求1-20，说明书第0045-0099段及图4-5	1-20
X	Maciek Tomczak 等. "Audio Style Transfer With Rhythmic Constraints" Proceedings of the 21st International Conference on Digital Audio Effects (DAFx-18), 2018年 9月 8日 (2018-09-08), 第1.1、2节	1-20
X	Anonymous authors Paper under double-blind review. "“STYLE” TRANSFER FOR MUSICAL AUDIO USING MULTIPLE TIME-FREQUENCY REPRESENTATIONS" Under review as a conference paper at ICLR 2018, 2018年 12月 31日 (2018-12-31), 第2-4节	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	国际检索报告邮寄日期
2020年 3月 19日	2020年 3月 27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	受权官员 喻新 电话号码 86-010-62089308

国际检索报告

国际申请号
PCT/CN2019/103573

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
A	Eric Grinstein 等. "Audio Style Transfer" 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018年 4月 20日 (2018-04-20), ISSN：2379-190X, 第586-590页	1-20

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/103573

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 106847294 A	2017年 6月 13日	CN 106847294 B	2018年 11月 30日
		US 10192163 B2	2019年 1月 29日
		US 2018204121 A1	2018年 7月 19日