(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 17 日 (17.12.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/248485 A1

(51) 国际专利分类号:
G10L 21/0272 (2013.01)

(21) 国际申请号: PCT/CN2019/117096

(22) 国际申请日: 2019 年 11 月 11 日 (11.11.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910511337.1 2019年6月13日 (13.06.2019) 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 518000 (CN)。

(72) 发明人: 吴冀平 (WU, Jiping); 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 518000 (CN)。 亢祖衡 (KANG, Zuheng); 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 518000 (CN)。 彭俊清 (PENG, Junqing); 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 518000 (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 518000 (CN)。

(74) 代理人: 深圳众鼎专利商标代理事务所（普通合伙） (SHENZHEN ZHONGDING INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市罗湖区笋岗东路3012号中民时代广场B座7楼701室, Guangdong 518000 (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:** MONOPHONIC SEPARATION METHOD AND APPARATUS OF AUDIO, COMPUTER DEVICE AND STORAGE MEDIUM

(54) **发明名称**: 音频单音色分离方法、装置、计算机设备及存储介质

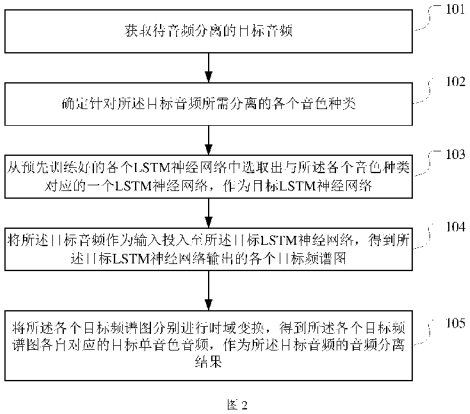


图 2

101 Acquire a target audio to be subjected to audio separation
102 Determine, for the target audio, each tone type needing to be separated
103 Select, from respective pretrained LSTM neural networks, one LSTM neural network corresponding to each tone type as a target LSTM neural network
104 Put the target audio serving as an input into the target LSTM neural network to obtain each target spectrogram output by the target LSTM neural network
105 Perform time domain transformation on each target spectrogram to obtain a target monophonic audio corresponding to each target spectrogram and use as an audio separation result of the target audio

(57) **Abstract:** A monophonic separation method and apparatus of an audio, a computer device and a storage medium, which are applied to the technical field of audio processing and are used for solving the problems, in the prior art, that monophonic separation cannot be realized. The method comprises: acquiring a target audio to be subjected to audio separation (101); determining, for the target audio, each tone type needing to be separated (102); selecting, from respective pretrained LSTM neural networks, one LSTM neural network corresponding to each tone type as a target LSTM neural network (103), wherein each LSTM neural network is obtained by pretraining audio samples corresponding to different tone type combinations, and each tone type combination consists of two or more tone types; putting the target audio serving as an input into the target LSTM neural network to obtain each output target spectrogram (104); and performing time domain transformation on each target spectrogram to obtain a target monophonic audio corresponding to each target spectrogram and use as an audio separation result of the target audio (105).

WO 2020/248485 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))。

(57) 摘要：一种音频单音色分离方法、装置、计算机设备及存储介质，应用于音频处理技术领域，用于解决现有技术无法实现单音色分离的问题。方法包括：获取待音频分离的目标音频(101)；确定针对目标音频所需分离的各个音色种类(102)；从预先训练好的各个LSTM神经网络中选取出与各个音色种类对应的一个LSTM神经网络，作为目标LSTM神经网络(103)，各个LSTM神经网络分别采用不同的音色种类组合所对应的音频样本预先训练得到，每个音色种类组合由两个以上音色种类组成；将目标音频作为输入投入至目标LSTM神经网络，得到输出的各个目标频谱图(104)；将各个目标频谱图分别进行时域变换，得到各个目标频谱图各自对应的目标单音色音频，作为目标音频的音频分离结果(105)。

音频单音色分离方法、装置、计算机设备及存储介质

本申请以 2019 年 06 月 13 日提交的申请号为 201910511337.1，名称为“音频单音色分离方法、装置、计算机设备及存储介质”的中国发明专利申请为基础，并要求其优先权。

技术领域

本申请涉及音频处理技术领域，尤其涉及音频单音色分离方法、装置、计算机设备及存储介质。

背景技术

在音乐库的开发中，音乐的内容分析是尤为重要的。一般自然采集得到的音频中常常是混合多种乐器和人声，若能实现对音频中各种乐器、人声的单音色分离，则可以对单一乐器、人声的内容分析提供强力的素材支持，并且促进音乐的音高识别技术和自动转录技术快速发展。可见，对音频实现单音色分离具有巨大的意义和价值。

因此，寻找一种能够实现音频单音色分离的方法一直是本领域技术人员亟需解决的问题。

发明内容

本申请实施例提供一种音频单音色分离方法、装置、计算机设备及存储介质，以解决现有技术无法实现单音色分离的问题。

一种音频单音色分离方法，包括：

获取待音频分离的目标音频；

确定针对所述目标音频所需分离的各个音色种类；

从预先训练好的各个 LSTM 神经网络中选取出与所述各个音色种类对应的一个 LSTM 神经网络，作为目标 LSTM 神经网络，所述各个 LSTM 神经网络分别采用不同的音色种类组合所对应的音频样本预先训练得到，每个音色种类组合由两个以上音色种类组成；

将所述目标音频作为输入投入至所述目标 LSTM 神经网络，得到所述目标 LSTM 神经网络输出的各个目标频谱图；

将所述各个目标频谱图分别进行时域变换，得到所述各个目标频谱图各自对应的目标单音色音频，作为所述目标音频的音频分离结果。

一种音频单音色分离装置，包括：

音频获取模块，用于获取待音频分离的目标音频；

音色种类确定模块，用于确定针对所述目标音频所需分离的各个音色种类；

神经网络选取模块，用于从预先训练好的各个 LSTM 神经网络中选取出与所述各个音色种类对应的一个 LSTM 神经网络，作为目标 LSTM 神经网络，所述各个 LSTM 神经网络分别采用不同的音色种类组合所对应的音频样本预先训练得到，每个音色种类组合由两个以上音色种类组成；

目标音频输入模块，用于将所述目标音频作为输入投入至所述目标 LSTM 神经网络，得到所述目标 LSTM 神经网络输出的各个目标频谱图；

时域变换模块，用于将所述各个目标频谱图分别进行时域变换，得到所述各个目标频谱图各自对应的目标单音色音频，作为所述目标音频的音频分离结果。

一种计算机设备，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可读指令，所述处理器执行所述计算机可读指令时实现上述音频单音色

分离方法的步骤。

一个或多个存储有计算机可读指令的可读存储介质，所述计算机可读存储介质存储有计算机可读指令，使得所述一个或多个处理器执行上述音频单音色分离方法的步骤。

本申请的一个或多个实施例的细节在下面的附图和描述中提出，本申请的其他特征和优点将从说明书、附图以及权利要求变得明显。

附图说明

为了更清楚地说明本申请实施例的技术方案，下面将对本申请实施例的描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

图 1 是本申请一实施例中音频单音色分离方法的一应用环境示意图；

图 2 是本申请一实施例中音频单音色分离方法的一流程图；

图 3 是本申请中 LSTM 神经网络对一个混合音频进行音色分离得到各个分离音频的原理示意图；

图 4 是本申请一实施例中音频单音色分离方法步骤 102 在一个应用场景下的流程示意图；

图 5 是本申请一实施例中音频单音色分离方法在一个应用场景下预先训练 LSTM 神经网络的流程示意图；

图 6 是本申请一实施例中音频单音色分离方法在一个应用场景下合成混合音频样本的流程示意图；

图 7 是本申请一实施例中音频单音色分离方法步骤 304 在一个应用场景下的流程示意图；

图 8 是本申请一实施例中音频单音色分离装置在一个应用场景下的结构示意图；

图 9 是本申请一实施例中音色种类确定模块的结构示意图；

图 10 是本申请一实施例中音频单音色分离装置在另一个应用场景下的结构示意图；

图 11 是本申请一实施例中计算机设备的一示意图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

本申请提供的音频单音色分离方法，可应用在如图 1 的应用环境中，其中，客户端通过网络与服务器进行通信。其中，该客户端可以但不限于各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备。服务器可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

在一实施例中，如图 2 所示，提供一种音频单音色分离方法，以该方法应用在图 1 中的服务器为例进行说明，包括如下步骤：

101、获取待音频分离的目标音频；

本实施例中，服务器首先需要获取到待音频分离的目标音频。可以理解的是，服务器可以通过多种方式获取到目标音频，比如，可以由负责音频分离的工作人员上传至服务器；也可以由服务器根据定时任务从指定数据库中提取得到音频文件，并将这些提取到的音频文件确定为待音频分离的目标音频；等等。

102、确定针对所述目标音频所需分离的各个音色种类；

需要说明的是，本实施例中使用的 LSTM 神经网络需要预先训练得到，一个 LSTM 神经网络在训练时以哪几种音色种类为目标进行音频分离的，在该LSTM神经网络训练完成后，该LSTM神经网络一般仅能用于分离得到这几种音色种类的单音色音频。例如，假设在训练某 LSTM 神经网络时，针对钢琴、小提琴、鼓三种音色种类进行学习训练，该LSTM神经网络训练好之后，使用该LSTM神经网络对某目标音频进行音色分离时，会得到钢琴、小提琴、鼓三种音色种类下的单音色音频。因此，在本实施例中针对不同的音色种类组合分别训练了多个LSTM神经网络，以便这些预先训练好的LSTM神经网络能够覆盖尽可能多的应用场景，例如针对钢琴、小提琴和鼓三种音色种类的组合训练一个LSTM 神经网络，记为 1 号网络；针对钢琴、小提琴和口琴三种音色种类的组合训练一个LSTM 神经网络，记为 2 号网络；针对二胡、风琴和吉他三种音色种类的组合训练一个LSTM 神经网络，记为 3 号网络；等等。

在上述内容基础上，为了准确分离该目标音频中的单音色音频数据，服务器还需要确定针对所述目标音频所需分离的各个音色种类。具体地，一种方式是，负责音色分离的工作人员可以手动输入指令，告知服务器对该目标音频所需分类的音色种类是哪几种，这样服务器即可确定所述各个音色种类；另一种方式是，服务器也可以根据该目标音频的来源场合自动判断该目标音频中大概包含的音色种类组合，从而确定该目标音频所需分离的各个音色种类。

可以理解的是，对目标音频的采集一般根据来源场合的不同会具有某些共同的音色种类，部分来源场合甚至仅有一种音色种类组合。比如，歌舞厅场合中采集到的音频往往包含人声、鼓、电子琴等几种音色种类，学校会议室场合中采集到的音频则往往仅包含人声和噪声。为便于理解，如图 4 所示，进一步地，步骤 102 可以包括：

201、获取所述目标音频的来源场合；

202、根据预设的场合音色对应关系确定与所述目标音频的来源场合对应的音色种类组合，所述场合音色对应关系记录了场合与音色种类组合之间的对应关系；

203、将确定出的所述音色种类组合中的各个音色种类确定为针对所述目标音频所需分离的各个音色种类。

对于上述步骤 201，可以理解的是，目标音频在被采集时可以一并记录该目标音频的来源场合，从而服务器获取到该目标音频时，可以同时获取到该目标音频的来源场合。例如，可以在目标音频上打上数字标签“001”，该数字标签标示该目标音频的来源场合为学校会议室，服务器从目标音频上读取到该数字标签即可获取到该来源场合。

对于步骤 202，服务器上可以预先设置场合音色对应关系，所述场合音色对应关系记录了场合与音色种类组合之间的对应关系，服务器在获取到目标音频的来源场合之后，可以根据预设的场合音色对应关系确定与所述目标音频的来源场合对应的音色种类组合。例如，服务器获取到该目标音频的来源场合为“001”，即学校会议室，在该场合音色对应关系中记录了“001”与“人声、噪声”的音色种类组合对应，则服务器可以确定出该“人声、噪声”的音色种类组合。

对于步骤 203，容易理解的是，音色种类组合由一个以上的音色种类组成，且每个音色种类组合由哪些音色种类组成均预先设置好。因此，服务器可以直接获取到确定出的所述音色种类组合中的各个音色种类，并将这些音色种类确定为针对所述目标音频所需分离的各个音色种类。

103、从预先训练好的各个LSTM神经网络中选取出与所述各个音色种类对应的一个LSTM神经网络，作为目标LSTM神经网络，所述各个LSTM神经网络分别采用不同的音色种类组合所对应的音频样本预先训练得到，每个音色种类组合由两个以上音色种类组成；

由上述内容可知，本实施例中预先训练好各个 LSTM 神经网络，这些 LSTM 神经网

络分别采用不同的音色种类组合所对应的音频样本预先训练得到，也即与不同音色种类组合相对应，其中每个音色种类组合由两个以上音色种类组成。

服务器在确定出目标音频所需分离的各个音色种类之后，可以从预先训练好的各个LSTM神经网络中选取出与所述各个音色种类对应的一个LSTM神经网络，作为用于为该目标音频进行音色分离的目标LSTM神经网络。

为便于理解，下面将对每个音色种类组合对应的 LSTM 神经网络的训练过程进行详细描述。如图5所示，进一步地，每个音色种类组合对应的LSTM神经网络通过以下步骤预先训练得到：

301、获取待训练的LSTM神经网络对应的音色种类组合包含的各个音色种类，作为各个样本音色种类；

302、分别采集所述各个样本音色种类各自对应的单音色音频样本；

303、根据所述各个样本音色种类各自对应的单音色音频样本合成得到各个混合音频样本，每个混合音频样本由所述各个样本音色种类各自对应的一个单音色音频样本合成得到；

304、针对每个混合音频样本，将所述每个混合音频样本作为输入投入至所述 LSTM 神经网络，得到所述 LSTM 神经网络输出的各个样本频谱图；

305、使用预设的代价函数计算所述各个样本频谱图与各个单音色频谱图之间的误差，所述各个单音色频谱图是指所述每个混合音频样本对应的各个单音色音频样本进过频域变换得到的频谱图；

306、以所述代价函数的计算结果为目标，调整所述LSTM神经网络的网络参数，直到所述代价函数的计算结果收敛，然后确定所述 LSTM 神经网络已训练完成。

对于步骤301，可以理解的是，当准备训练某个音色种类组合对应的LSTM神经网络时，首先要获取这个音色种类组合包含的音色种类，记为各个样本音色种类。例如，准备训练二胡、吉他和口琴这一组合对应的LSTM神经网络，服务器首先获取该LSTM神经网络的各个样本音色种类为“二胡、吉他和口琴”。

特别地，在某些应用场景下，有时候仅需从音频中分离出特定的几种音色种类的音频即可，为了满足这种情况的需求，也可以将所有与特定音色种类不同的音频数据划分为“其它”音色种类，在训练LSTM神经网络时考虑上“其它”音色种类。例如，若某个LSTM神经网络需要被训练成用于分离二胡、吉他和口琴三种音色种类的音频数据，可以将二胡、吉他、口琴和“其它”四种音色种类设定为一个音色种类组合，作为该LSTM神经网络的各个样本音色种类。

对于步骤302，训练时，为了保证样本的纯净度，使用于LSTM神经网络训练的数据均为准确的音频数据，因此，需要分别针对各个样本音色种类采集各自对应的单音色音频样本。这里所说的单音色音频样本是指该音频样本中仅包含有一种音色种类的音频数据，比如“二胡”这一音色种类对应的单音色音频样本可以在无噪音环境下采集二胡乐器奏响的音频，这样采集得到的音频数据可以认为是“二胡”的单音色音频数据，因此可以用作单音色音频样本。

对于步骤 303，在采集得到单音色音频样本之后，服务器可以根据所述各个样本音色种类各自对应的单音色音频样本合成得到各个混合音频样本，其中，每个混合音频样本由所述各个样本音色种类各自对应的一个单音色音频样本合成得到。

为便于理解，如图6所示，更进一步地，步骤303中每个混合音频样本通过以下步骤合成得到：

401、针对每个所述样本音色种类对应的单音色音频样本，从所述单音色音频样本中选取出一个单音色音频样本，作为待混音样本；

402、将所述各个样本音色种类各自对应的待混音样本进行混音处理，得到一个混合

音频样本。

对于步骤401和步骤402，服务器可以从每个样本音色种类的单音色音频样本中取一个样本，即所述待混音样本，再将取出的各个待混音样本进行混音处理，得到一个混合音频样本。例如，步骤301获取到3个样本音色种类，分别记为种类1、种类2和种类3，步骤302为每个样本音色种类分别采集了10个单音色音频样本，则服务器可以从种类1的10个单音色音频样本中取出1个单音色音频样本作为待混音样本，从种类2的10个单音色音频样本中取出1个单音色音频样本作为待混音样本，从种类3的10个单音色音频样本中取出1个单音色音频样本作为待混音样本，然后把取出的3个待混音样本进行混音处理，得到1个混音音频样本。这就是一个混音音频样本的合成过程，重复上述步骤401和步骤402，可以得到多个混音音频样本。

需要说明的是，为了提高样本训练的有效性，在步骤303中应当注意任意两个混音音频样本所用的合成的单音色音频样本的组合是不同的，应当知道，若两个单音色音频样本的组合相同，则这两个组合各自合成得到的混音音频样本也相同，两个相同的样本投入到后续的训练，一般仅增加了训练所需的运算负担，对LSTM神经网络的训练完成度没有帮助。为此，在重复执行上述步骤401和步骤402时，可以采用组合的方式选取待混音样本的组合，这样，由各个组合混音后得到的各个混合音频样本各不相同。

对于步骤304，服务器在合成得到各个混合音频样本之后，可以将这些混音音频样本投入到LSTM神经网络中，对该LSTM神经网络进行训练。针对每个混音音频样本，服务器将其作为输入投入至所述LSTM神经网络，得到所述LSTM神经网络输出的各个样本频谱图。

为便于理解，下面将对每个混音音频样本输入LSTM神经网络后，在LSTM神经网络中的处理过程进行详细描述。更进一步地，如图7所示，步骤304可以包括：

501、将所述每个混合音频样本进行频域变换，得到所述每个混合音频样本的混合频谱图；

502、对所述混合频谱图加窗，且对加窗得到的每帧数据进行短时傅里叶变换，得到各个频谱特征向量；

503、对所述各个频谱特征向量进行重叠分组，得到各组频谱特征片段，任意相邻两组频谱特征片段之间存在重叠的频谱特征向量；

504、将每组频谱特征片段分别与预设卷积核进行卷积计算，得到降维后的各个片段向量，每个片段向量对应一组频谱特征片段；

505、将所述各个片段向量输入至LSTM，得到所述LSTM输出的各个音频信息向量；

506、针对预设数量个多层感知器中的每个多层感知器，将所述各个音频信息向量输入至每个多层感知器，得到所述每个多层感知器输出的分离特征向量，所述预设数量等于所述各个样本音色种类的数量；

507、针对每个多层感知器输出的分离特征向量，将所述分离特征向量输入至所述LSTM进行特征还原，得到每个多层感知器对应的音频特征向量；

508、针对每个多层感知器对应的音频特征向量，将所述音频特征向量分别与所述预设卷积核进行转置卷积计算，得到升维后的、每个多层感知器对应的各个音频特征片段；

509、分别拼合各个多层感知器各自对应的各个音频特征片段，得到各个多层感知器各自对应的单音色频谱图，作为所述LSTM神经网络输出的各个样本频谱图。

对于上述步骤501-509，可以结合图3进行理解，图3示出了LSTM神经网络对一个混合音频进行音色分离得到各个分离音频的原理示意图。

对于步骤501，首先，服务器可以将所述每个混合音频样本进行频域变换，得到所

述每个混合音频样本的混合频谱图。

对于步骤 502，在得到混合频谱图之后，服务器可以对该混合频谱图加窗。具体地，可以对该混合频谱图加汉明窗，通过汉明窗可以取得该混合频谱图上的一帧帧的数据，服务器再对每帧数据进行短时傅里叶变换，得到各个频谱特征向量。需要说明的是，通过汉明窗从该混合频谱图上读取数据时可以设定一定的重叠率，比如可以设定50%-80%区间中的一个重叠率，每帧数据的时间长度可以设定为 20 毫秒左右。

对于步骤 503，在得到各个频谱特征向量之后，可以对这些频谱特征向量进行重叠分组，也即将多个频谱特征向量捆绑成一组频谱特征片段。分组时，每组频谱特征片段具体可以划分有预设数量个频谱特征向量，该预设数量可以根据实际使用情况设定，本实施例对此不作限定。

需要说明的是，分组与分组之间存在一定的数据重叠，也即任意相邻两组频谱特征片段之间存在重叠的频谱特征向量。这样分组的意义在于抽取瞬态音频的变化，分组之间重叠的时间长度要求概括较为简单的音频变化。由于还需要把瞬时变化的部分考虑进去，在分组的过程中，可以采用 50%重叠率的重叠分组方式。为了便于理解，重叠分组时，首先把各个频谱特征向量进行编号，从 0 到 n。然后可以定义从 0 到 a 号的频谱特征向量划分为第一分组，即第一组频谱特征片段；从 a/2 到 3*a/2 号的频谱特征向量划分为第二分组，即第二组频谱特征片段；从 a 到 2*a 号的频谱特征向量划分为第三分组，即第三组频谱特征片段；以此类推，直至对所有频谱特征向量完成分组位置（其中 a 可以为每组的帧数，且保证 a 为一个偶数）。

对于步骤 504，在得到各组频谱特征片段之后，为了将这些频谱特征片段降维，便于后续进行序列学习，服务器可以将每组频谱特征片段分别与预设卷积核进行卷积计算，得到降维后的各个片段向量，其中，每个片段向量对应一组频谱特征片段。具体地，该预设卷积核可以设定为在时间维度上较小，在频率维度上较大，从而对频谱特征片段进行卷积之后，得到的片段向量将在频域方向上展平，一般来说，可以使得降维后的片段向量为一个一维向量。

需要说明的是，为了提高 LSTM 神经网络对模型的表达能力，可以在每层卷积层后面使用激活函数进行激活，从而在该 LSTM 神经网络中加入非线性因素，解决线性模型在表达能力上不足的问题。

对于步骤 505，参阅图 3，在进行步骤 504 处理后，服务器得到的各个片段向量可以为被拉直的一个个一维向量，然后，服务器将这些片段向量输入一个长短期记忆网络（LSTM），进行一次从序列到序列（seq2seq）的学习，并且 LSTM 中每个网络单元输出的向量作为某时间节点被抽取的音频信息，即得到 LSTM 输出的各个音频信息向量。

对于步骤 506，本实施例中的 LSTM 神经网络上还设置有预设数量个多层感知器（MLP，Multi-Layer Perceptron），多层感知器的数量等于所述各个样本音色种类的数量，比如，本次训练的 LSTM 神经网络对应的音色种类组合存在 3 个样本音色种类，则该 LSTM 神经网络中预设有 3 个多层感知器。本实施例中多层感知器的功能类似于信息过滤器，用来过滤出某一音色的特征信息，实现了音频信息的分离。

服务器针对每个多层感知器，将步骤 505 得到的各个音频信息向量输入至该多层感知器，得到该多层感知器输出的分离特征向量。这相当于，服务器将各个音频信息向量分别输入每个多层感知器一次，比如，假设有 3 个多层感知器，分别为感知器 a、感知器 b 和感知器 c，服务器执行步骤 505 得到 N 个音频信息向量，则服务器可以将 N 个音频信息向量输入至感知器 a 中，得到感知器 a 输出的分离特征向量；并且，服务器将 N 个音频信息向量输入至感知器 b 中，得到感知器 b 输出的分离特征向量；另外，服务器还将 N 个音频信息向量输入至感知器 c 中，得到感知器 c 输出的分离特征向量。这样，服务器可以得到 3 个多层感知器分别输出的结果。

对于步骤 507，可以理解的是，服务器经过上述步骤 501-506 的处理，基本已经实现了对混合音频样本中单音色音频信息的分离，但这些分离得到的音频信息以一种仅被 LSTM 神经网络识别的数据形式存在，即上述的分离特征向量。为了使得这些分离后的音频信息能够被识别和使用，还需要将这些分离特征向量经过对偶的逆向过程处理，实现音频信息的还原。

参阅图 3，并结合上述步骤 502-504 可知，步骤 502-504 是对音频在数据形式上的处理过程，使得混合音频数据转换为更容易被神经网络理解和识别的数据形式，因此，步骤 507-509 为与 502-504 对偶的逆向处理过程，能够将分离特征向量在数据形式上还原为与混合频谱图相同。

具体地，服务器针对每个多层感知器输出的分离特征向量，将所述分离特征向量输入同一个 LSTM 进行 seq2seq 学习，每个网络单元输出的向量作为某时间节点被还原的分离的音频特征信息，从而得到每个多层感知器对应的音频特征向量。

对于步骤 508，服务器在得到各个音频特征向量之后，可以针对每个多层感知器对应的音频特征向量，将所述音频特征向量分别与所述预设卷积核进行转置卷积计算，得到升维后的、每个多层感知器对应的各个音频特征片段。

对于步骤 509，服务器在得到升维后的、每个多层感知器对应的各个音频特征片段，可以分别拼合各个多层感知器各自对应的各个音频特征片段，得到各个多层感知器各自对应的单音色频谱图，作为所述 LSTM 神经网络输出的各个样本频谱图。

对于步骤 305 和步骤 306，本实施例中，为了评估该 LSTM 神经网络训练的完成度，可以预先设置代价函数来计算所述各个样本频谱图与各个单音色频谱图之间的误差，然后在调整网络参数时，以所述代价函数的计算得到的误差为目标，不断调整 LSTM 神经网络中的各个网络参数，直到所述代价函数的计算结果收敛，然后确定所述 LSTM 神经网络已训练完成。具体地，在训练时，可以使用随机梯度下降法（SGD）促使该 LSTM 神经网络快速收敛。

特别地，上述的代价函数可以为均方误差（MSE，Mean Squared Error）。

优选地，服务器可以预先将采集得到的训练数据样本划分为训练数据集和测试数据集，其中训练数据集占样本数量的 80%，测试数据集占样本数量的 20%。在该 LSTM 神经网络训练完成之后，服务器可以使用测试数据集中的各个样本对该 LSTM 神经网络进行测试和评估，评估时，可以由负责测试的工作人员对该 LSTM 神经网络输出的单音色音频进行试听，并以试听的效果作为对该 LSTM 神经网络的评判，评判通过，再确定该 LSTM 神经网络确已完成训练，可以投入使用；反之，若评判不通过，可以考虑对该 LSTM 神经网络重新训练。

优选地，本实施例中，还可以将步骤 304 输出的各个样本频谱图与混合音频样本对应的各个单音色音频样本的频谱图进行对比校验，若均校验一致，则可以确定神经网络训练完成。

104、将所述目标音频作为输入投入至所述目标 LSTM 神经网络，得到所述目标 LSTM 神经网络输出的各个目标频谱图；

本实施例中，服务器在获取到目标音频，确定出目标 LSTM 神经网络之后，可以将所述目标音频作为输入投入至所述目标 LSTM 神经网络，得到所述目标 LSTM 神经网络输出的各个目标频谱图。

105、将所述各个目标频谱图分别进行时域变换，得到所述各个目标频谱图各自对应的目标单音色音频，作为所述目标音频的音频分离结果。

服务器在得到所述目标 LSTM 神经网络输出的各个目标频谱图之后，为了便于音频数据的管理和存储，也为了方便后续对单音色音频的使用，可以将所述各个目标频谱图分别进行时域变换，得到所述各个目标频谱图各自对应的目标单音色音频，作为所述目标音

频的音频分离结果。可以认为，最后得到的各个目标单音色音频，就是该目标音频中所包含的各个单音色音频数据各自从目标音频中分离出来的音频，同时也是该目标音频在所需分离的各个音色种类的划分下的音频分离结果。

本实施例中，首先，获取待音频分离的目标音频；然后，确定针对所述目标音频所需分离的各个音色种类；接着，从预先训练好的各个 LSTM 神经网络中选取出与所述各个音色种类对应的一个 LSTM 神经网络，作为目标 LSTM 神经网络，所述各个 LSTM 神经网络分别采用不同的音色种类组合所对应的音频样本预先训练得到，每个音色种类组合由两个以上音色种类组成；再之，将所述目标音频作为输入投入至所述目标 LSTM 神经网络，得到所述目标 LSTM 神经网络输出的各个目标频谱图；最后，将所述各个目标频谱图分别进行时域变换，得到所述各个目标频谱图各自对应的目标单音色音频，作为所述目标音频的音频分离结果。可见，本申请通过预先训练好的 LSTM 神经网络能够将目标音频分离成各个目标单音色音频，且可以根据所需分离得到的音色种类选取出对应的 LSTM 神经网络来决定最终分离得到的目标单音色音频的音色种类，不仅实现音频的单音色分离，还使得单音色分离的结果在一定程度上可控，在某些应用场景下为音频的内容分析提供更多的支持和帮助。

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，而不应对本申请实施例的实施过程构成任何限定。

在一实施例中，提供一种音频单音色分离装置，该音频单音色分离装置与上述实施例中音频单音色分离方法一一对应。如图 8 所示，该音频单音色分离装置包括音频获取模块 601、音色种类确定模块 602、神经网络选取模块 603、目标音频输入模块 604 和时域变换模块 605。各功能模块详细说明如下：

音频获取模块 601，用于获取待音频分离的目标音频；

音色种类确定模块 602，用于确定针对所述目标音频所需分离的各个音色种类；

神经网络选取模块 603，用于从预先训练好的各个 LSTM 神经网络中选取出与所述各个音色种类对应的一个 LSTM 神经网络，作为目标 LSTM 神经网络，所述各个 LSTM 神经网络分别采用不同的音色种类组合所对应的音频样本预先训练得到，每个音色种类组合由两个以上音色种类组成；

目标音频输入模块 604，用于将所述目标音频作为输入投入至所述目标 LSTM 神经网络，得到所述目标 LSTM 神经网络输出的各个目标频谱图；

时域变换模块 605，用于将所述各个目标频谱图分别进行时域变换，得到所述各个目标频谱图各自对应的目标单音色音频，作为所述目标音频的音频分离结果。

如图 9 所示，优选地，所述音色种类确定模块 602 可以包括：

来源场合获取单元 6021，用于获取所述目标音频的来源场合；

种类组合确定单元 6022，用于根据预设的场合音色对应关系确定与所述目标音频的来源场合对应的音色种类组合，所述场合音色对应关系记录了场合与音色种类组合之间的对应关系；

音色种类确定单元 6023，用于将确定出的所述音色种类组合中的各个音色种类确定为针对所述目标音频所需分离的各个音色种类。

如图 10 所示，优选地，每个音色种类组合对应的 LSTM 神经网络可以通过以下模块预先训练得到：

样本种类获取模块 606，用于获取待训练的 LSTM 神经网络对应的音色种类组合包含的各个音色种类，作为各个样本音色种类；

音频样本采集模块 607，用于分别采集所述各个样本音色种类各自对应的单音色音频样本；

混合音频样本合成模块608，用于根据所述各个样本音色种类各自对应的单音色音频样本合成得到各个混合音频样本，每个混合音频样本由所述各个样本音色种类各自对应的一个单音色音频样本合成得到；

样本输入模块609，用于针对每个混合音频样本，将所述每个混合音频样本作为输入投入至所述LSTM神经网络，得到所述LSTM神经网络输出的各个样本频谱图；

误差计算模块610，用于使用预设的代价函数计算所述各个样本频谱图与各个单音色频谱图之间的误差，所述各个单音色频谱图是指所述每个混合音频样本对应的各个单音色音频样本进过频域变换得到的频谱图；

网络参数调整模块611，用于以所述代价函数的计算结果为目标，调整所述LSTM神经网络的网络参数，直到所述代价函数的计算结果收敛，然后确定所述LSTM神经网络已训练完成。

优选地，所述样本输入模块可以包括：

频域变换单元，用于将所述每个混合音频样本进行频域变换，得到所述每个混合音频样本的混合频谱图；

加窗单元，用于对所述混合频谱图加窗，且对加窗得到的每帧数据进行短时傅里叶变换，得到各个频谱特征向量；

重叠分组单元，用于对所述各个频谱特征向量进行重叠分组，得到各组频谱特征片段，任意相邻两组频谱特征片段之间存在重叠的频谱特征向量；

卷积计算单元，用于将每组频谱特征片段分别与预设卷积核进行卷积计算，得到降维后的各个片段向量，每个片段向量对应一组频谱特征片段；

片段向量输入单元，用于将所述各个片段向量输入至LSTM，得到所述LSTM输出的各个音频信息向量；

多层感知器处理单元，用于针对预设数量个多层感知器中的每个多层感知器，将所述各个音频信息向量输入至每个多层感知器，得到所述每个多层感知器输出的分离特征向量，所述预设数量等于所述各个样本音色种类的数量；

特征还原单元，用于针对每个多层感知器输出的分离特征向量，将所述分离特征向量输入至所述LSTM进行特征还原，得到每个多层感知器对应的音频特征向量；

转置卷积计算单元，用于针对每个多层感知器对应的音频特征向量，将所述音频特征向量分别与所述预设卷积核进行转置卷积计算，得到升维后的、每个多层感知器对应的各个音频特征片段；

特征片段拼合单元，用于分别拼合各个多层感知器各自对应的各个音频特征片段，得到各个多层感知器各自对应的单音色频谱图，作为所述LSTM神经网络输出的各个样本频谱图。

优选地，每个混合音频样本通过以下模块合成得到：

待混音样本选取模块，用于针对每个所述样本音色种类对应的单音色音频样本，从所述单音色音频样本中选取出一个单音色音频样本，作为待混音样本；

混音处理模块，用于将所述各个样本音色种类各自对应的待混音样本进行混音处理，得到一个混合音频样本。

关于音频单音色分离装置的具体限定可以参见上文中对于音频单音色分离方法的限定，在此不再赘述。上述音频单音色分离装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处理器调用执行以上各个模块对应的操作。

在一个实施例中，提供了一种计算机设备，该计算机设备可以是服务器，其内部结构

图可以如图 11 所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中，该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括可读存储介质、内存储器。该可读存储介质存储有操作系统、计算机可读指令和数据库。该内存储器为可读存储介质中的操作系统和计算机可读指令的运行提供环境。该计算机设备的数据库用于存储音频单音色分离方法中涉及到的数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机可读指令被处理器执行时以实现一种音频单音色分离方法。本实施例所提供的可读存储介质包括非易失性可读存储介质和易失性可读存储介质。

在一个实施例中，提供了一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机可读指令，处理器执行计算机可读指令时实现上述实施例中音频单音色分离方法的步骤，例如图 2 所示的步骤 101 至步骤 105。或者，处理器执行计算机可读指令时实现上述实施例中音频单音色分离装置的各模块/单元的功能，例如图 8 所示模块 601 至模块 605 的功能。为避免重复，这里不再赘述。

在一个实施例中，提供了一种计算机可读存储介质，该一个或多个存储有计算机可读指令的可读存储介质，计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行计算机可读指令时实现上述方法实施例中音频单音色分离方法的步骤，或者，该一个或多个存储有计算机可读指令的可读存储介质，计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行计算机可读指令时实现上述装置实施例中音频单音色分离装置中各模块/单元的功能。为避免重复，这里不再赘述。本实施例所提供的可读存储介质包括非易失性可读存储介质和易失性可读存储介质。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机可读指令来指令相关的硬件来完成，所述的计算机可读指令可存储于一计算机可读取存储介质中，该计算机可读指令在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。存储器可包括只读存储器（ROM）、可编程 ROM（PROM）、电可编程 ROM（EPROM）、电可擦除可编程 ROM（EEPROM）或闪存。易失性存储器可包括随机存取存储器（RAM）或者外部高速缓冲存储器。作为说明而非局限，RAM 以多种形式可得，诸如静态 RAM（SRAM）、动态 RAM（DRAM）、同步 DRAM（SDRAM）、双数据率 SDRAM（DDRSDRAM）、增强型 SDRAM（ESDRAM）、同步链路（Synchlink） DRAM（SLDRAM）、存储器总线（Rambus）直接 RAM（RDRAM）、直接存储器总线动态 RAM（DRDRAM）、以及存储器总线动态 RAM（RDRAM）等。

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，即将所述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。

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. 一种音频单音色分离方法，其特征在于，包括：

获取待音频分离的目标音频；

确定针对所述目标音频所需分离的各个音色种类；

从预先训练好的各个 LSTM 神经网络中选取出与所述各个音色种类对应的一个 LSTM 神经网络，作为目标 LSTM 神经网络，所述各个 LSTM 神经网络分别采用不同的音色种类组合所对应的音频样本预先训练得到，每个音色种类组合由两个以上音色种类组成；

将所述目标音频作为输入投入至所述目标 LSTM 神经网络，得到所述目标 LSTM 神经网络输出的各个目标频谱图；

将所述各个目标频谱图分别进行时域变换，得到所述各个目标频谱图各自对应的目标单音色音频，作为所述目标音频的音频分离结果。

2. 根据权利要求 1 所述的音频单音色分离方法，其特征在于，所述确定针对所述目标音频所需分离的各个音色种类包括：

获取所述目标音频的来源场合；

根据预设的场合音色对应关系确定与所述目标音频的来源场合对应的音色种类组合，所述场合音色对应关系记录了场合与音色种类组合之间的对应关系；

将确定出的所述音色种类组合中的各个音色种类确定为针对所述目标音频所需分离的各个音色种类。

3. 根据权利要求 1 或 2 所述的音频单音色分离方法，其特征在于，每个音色种类组合对应的 LSTM 神经网络通过以下步骤预先训练得到：

获取待训练的 LSTM 神经网络对应的音色种类组合包含的各个音色种类，作为各个样本音色种类；

分别采集所述各个样本音色种类各自对应的单音色音频样本；

根据所述各个样本音色种类各自对应的单音色音频样本合成得到各个混合音频样本，每个混合音频样本由所述各个样本音色种类各自对应的一个单音色音频样本合成得到；

针对每个混合音频样本，将所述每个混合音频样本作为输入投入至所述 LSTM 神经网络，得到所述 LSTM 神经网络输出的各个样本频谱图；

使用预设的代价函数计算所述各个样本频谱图与各个单音色频谱图之间的误差，所述各个单音色频谱图是指所述每个混合音频样本对应的各个单音色音频样本进过频域变换得到的频谱图；

以所述代价函数的计算结果为目标，调整所述 LSTM 神经网络的网络参数，直到所述代价函数的计算结果收敛，然后确定所述 LSTM 神经网络已训练完成。

4. 根据权利要求 3 所述的音频单音色分离方法，其特征在于，所述针对每个混合音频样本，将所述每个混合音频样本作为输入投入至所述 LSTM 神经网络，得到所述 LSTM 神经网络输出的各个样本频谱图包括：

将所述每个混合音频样本进行频域变换，得到所述每个混合音频样本的混合频谱图；

对所述混合频谱图加窗，且对加窗得到的每帧数据进行短时傅里叶变换，得到各个频谱特征向量；

对所述各个频谱特征向量进行重叠分组，得到各组频谱特征片段，任意相邻两组频谱特征片段之间存在重叠的频谱特征向量；

将每组频谱特征片段分别与预设卷积核进行卷积计算，得到降维后的各个片段向量，每个片段向量对应一组频谱特征片段；

将所述各个片段向量输入至 LSTM，得到所述 LSTM 输出的各个音频信息向量；

针对预设数量个多层感知器中的每个多层感知器，将所述各个音频信息向量输入至每个多层感知器，得到所述每个多层感知器输出的分离特征向量，所述预设数量等于所述各个样本音色种类的数量；

针对每个多层感知器输出的分离特征向量，将所述分离特征向量输入至所述 LSTM 进行特征还原，得到每个多层感知器对应的音频特征向量；

针对每个多层感知器对应的音频特征向量，将所述音频特征向量分别与所述预设卷积核进行转置卷积计算，得到升维后的、每个多层感知器对应的各个音频特征片段；

分别拼合各个多层感知器各自对应的各个音频特征片段，得到各个多层感知器各自对应的单音色频谱图，作为所述 LSTM 神经网络输出的各个样本频谱图。

5. 根据权利要求 3 所述的音频单音色分离方法，其特征在于，每个混合音频样本通过以下步骤合成得到：

针对每个所述样本音色种类对应的单音色音频样本，从所述单音色音频样本中选取出一个单音色音频样本，作为待混音样本；

将所述各个样本音色种类各自对应的待混音样本进行混音处理，得到一个混合音频样本。

6. 一种音频单音色分离装置，其特征在于，包括：

音频获取模块，用于获取待音频分离的目标音频；

音色种类确定模块，用于确定针对所述目标音频所需分离的各个音色种类；

神经网络选取模块，用于从预先训练好的各个 LSTM 神经网络中选取出与所述各个音色种类对应的一个 LSTM 神经网络，作为目标 LSTM 神经网络，所述各个 LSTM 神经网络分别采用不同的音色种类组合所对应的音频样本预先训练得到，每个音色种类组合由两个以上音色种类组成；

目标音频输入模块，用于将所述目标音频作为输入投入至所述目标 LSTM 神经网络，得到所述目标 LSTM 神经网络输出的各个目标频谱图；

时域变换模块，用于将所述各个目标频谱图分别进行时域变换，得到所述各个目标频谱图各自对应的目标单音色音频，作为所述目标音频的音频分离结果。

7. 根据权利要求 6 所述的音频单音色分离装置，其特征在于，所述音色种类确定模块包括：

来源场合获取单元，用于获取所述目标音频的来源场合；

种类组合确定单元，用于根据预设的场合音色对应关系确定与所述目标音频的来源场合对应的音色种类组合，所述场合音色对应关系记录了场合与音色种类组合之间的对应关系；

音色种类确定单元，用于将确定出的所述音色种类组合中的各个音色种类确定为针对所述目标音频所需分离的各个音色种类。

8. 根据权利要求 6 或 7 所述的音频单音色分离装置，其特征在于，每个音色种类组合对应的 LSTM 神经网络通过以下模块预先训练得到：

样本种类获取模块，用于获取待训练的 LSTM 神经网络对应的音色种类组合包含的各个音色种类，作为各个样本音色种类；

音频样本采集模块，用于分别采集所述各个样本音色种类各自对应的单音色音频样本；

混合音频样本合成模块，用于根据所述各个样本音色种类各自对应的单音色音频样本合成得到各个混合音频样本，每个混合音频样本由所述各个样本音色种类各自对应的一个单音色音频样本合成得到；

样本输入模块，用于针对每个混合音频样本，将所述每个混合音频样本作为输入投入至所述 LSTM 神经网络，得到所述 LSTM 神经网络输出的各个样本频谱图；

误差计算模块，用于使用预设的代价函数计算所述各个样本频谱图与各个单音色频谱图之间的误差，所述各个单音色频谱图是指所述每个混合音频样本对应的各个单音色音频样本进过频域变换得到的频谱图；

网络参数调整模块，用于以所述代价函数的计算结果为目标，调整所述 LSTM 神经网络的网络参数，直到所述代价函数的计算结果收敛，然后确定所述 LSTM 神经网络已训练完成。

9. 根据权利要求 8 所述的音频单音色分离装置，其特征在于，所述样本输入模块包括：

频域变换单元，用于将所述每个混合音频样本进行频域变换，得到所述每个混合音频样本的混合频谱图；

加窗单元，用于对所述混合频谱图加窗，且对加窗得到的每帧数据进行短时傅里叶变换，得到各个频谱特征向量；

重叠分组单元，用于对所述各个频谱特征向量进行重叠分组，得到各组频谱特征片段，任意相邻两组频谱特征片段之间存在重叠的频谱特征向量；

卷积计算单元，用于将每组频谱特征片段分别与预设卷积核进行卷积计算，得到降维后的各个片段向量，每个片段向量对应一组频谱特征片段；

片段向量输入单元，用于将所述各个片段向量输入至 LSTM，得到所述 LSTM 输出的各个音频信息向量；

多层感知器处理单元，用于针对预设数量个多层感知器中的每个多层感知器，将所述各个音频信息向量输入至每个多层感知器，得到所述每个多层感知器输出的分离特征向量，所述预设数量等于所述各个样本音色种类的数量；

特征还原单元，用于针对每个多层感知器输出的分离特征向量，将所述分离特征向量输入至所述 LSTM 进行特征还原，得到每个多层感知器对应的音频特征向量；

转置卷积计算单元，用于针对每个多层感知器对应的音频特征向量，将所述音频特征向量分别与所述预设卷积核进行转置卷积计算，得到升维后的、每个多层感知器对应的各个音频特征片段；

特征片段拼合单元，用于分别拼合各个多层感知器各自对应的各个音频特征片段，得到各个多层感知器各自对应的单音色频谱图，作为所述 LSTM 神经网络输出的各个样本频谱图。

10. 根据权利要求 8 所述的音频单音色分离装置，其特征在于，每个混合音频样本通过以下模块合成得到：

待混音样本选取模块，用于针对每个所述样本音色种类对应的单音色音频样本，从所述单音色音频样本中选取出一个单音色音频样本，作为待混音样本；

混音处理模块，用于将所述各个样本音色种类各自对应的待混音样本进行混音处理，得到一个混合音频样本。

11. 一种计算机设备，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可读指令，其特征在于，所述处理器执行所述计算机可读指令时实现如下步骤：

获取待音频分离的目标音频；

确定针对所述目标音频所需分离的各个音色种类；

从预先训练好的各个 LSTM 神经网络中选取出与所述各个音色种类对应的一个 LSTM 神经网络，作为目标 LSTM 神经网络，所述各个 LSTM 神经网络分别采用不同的音色种类组合所对应的音频样本预先训练得到，每个音色种类组合由两个以上音色种类组成；

将所述目标音频作为输入投入至所述目标 LSTM 神经网络，得到所述目标 LSTM 神

经网络输出的各个目标频谱图；

将所述各个目标频谱图分别进行时域变换，得到所述各个目标频谱图各自对应的目标单音色音频，作为所述目标音频的音频分离结果。

12. 根据权利要求 11 所述的计算机设备，其特征在于，所述确定针对所述目标音频所需分离的各个音色种类包括：

获取所述目标音频的来源场合；

根据预设的场合音色对应关系确定与所述目标音频的来源场合对应的音色种类组合，所述场合音色对应关系记录了场合与音色种类组合之间的对应关系；

将确定出的所述音色种类组合中的各个音色种类确定为针对所述目标音频所需分离的各个音色种类。

13. 根据权利要求 11 或 12 所述的计算机设备，其特征在于，每个音色种类组合对应的 LSTM 神经网络通过以下步骤预先训练得到：

获取待训练的 LSTM 神经网络对应的音色种类组合包含的各个音色种类，作为各个样本音色种类；

分别采集所述各个样本音色种类各自对应的单音色音频样本；

根据所述各个样本音色种类各自对应的单音色音频样本合成得到各个混合音频样本，每个混合音频样本由所述各个样本音色种类各自对应的一个单音色音频样本合成得到；

针对每个混合音频样本，将所述每个混合音频样本作为输入投入至所述 LSTM 神经网络，得到所述 LSTM 神经网络输出的各个样本频谱图；

使用预设的代价函数计算所述各个样本频谱图与各个单音色频谱图之间的误差，所述各个单音色频谱图是指所述每个混合音频样本对应的各个单音色音频样本进过频域变换得到的频谱图；

以所述代价函数的计算结果为目标，调整所述 LSTM 神经网络的网络参数，直到所述代价函数的计算结果收敛，然后确定所述 LSTM 神经网络已训练完成。

14. 根据权利要求 13 所述的计算机设备，其特征在于，所述针对每个混合音频样本，将所述每个混合音频样本作为输入投入至所述 LSTM 神经网络，得到所述 LSTM 神经网络输出的各个样本频谱图包括：

将所述每个混合音频样本进行频域变换，得到所述每个混合音频样本的混合频谱图；

对所述混合频谱图加窗，且对加窗得到的每帧数据进行短时傅里叶变换，得到各个频谱特征向量；

对所述各个频谱特征向量进行重叠分组，得到各组频谱特征片段，任意相邻两组频谱特征片段之间存在重叠的频谱特征向量；

将每组频谱特征片段分别与预设卷积核进行卷积计算，得到降维后的各个片段向量，每个片段向量对应一组频谱特征片段；

将所述各个片段向量输入至 LSTM，得到所述 LSTM 输出的各个音频信息向量；

针对预设数量个多层感知器中的每个多层感知器，将所述各个音频信息向量输入至每个多层感知器，得到所述每个多层感知器输出的分离特征向量，所述预设数量等于所述各个样本音色种类的数量；

针对每个多层感知器输出的分离特征向量，将所述分离特征向量输入至所述 LSTM 进行特征还原，得到每个多层感知器对应的音频特征向量；

针对每个多层感知器对应的音频特征向量，将所述音频特征向量分别与所述预设卷积核进行转置卷积计算，得到升维后的、每个多层感知器对应的各个音频特征片段；

分别拼合各个多层感知器各自对应的各个音频特征片段，得到各个多层感知器各自对应的单音色频谱图，作为所述 LSTM 神经网络输出的各个样本频谱图。

15. 根据权利要求 13 所述的计算机设备，其特征在于，每个混合音频样本通过以下

步骤合成得到：

针对每个所述样本音色种类对应的单音色音频样本，从所述单音色音频样本中选取出一个单音色音频样本，作为待混音样本；

将所述各个样本音色种类各自对应的待混音样本进行混音处理，得到一个混合音频样本。

16. 一个或多个存储有计算机可读指令的可读存储介质，其特征在于，所述计算机可读指令被一个或多个处理器执行时，使得所述一个或多个处理器执行如下步骤：

获取待音频分离的目标音频；

确定针对所述目标音频所需分离的各个音色种类；

从预先训练好的各个 LSTM 神经网络中选取出与所述各个音色种类对应的一个 LSTM 神经网络，作为目标 LSTM 神经网络，所述各个 LSTM 神经网络分别采用不同的音色种类组合所对应的音频样本预先训练得到，每个音色种类组合由两个以上音色种类组成；

将所述目标音频作为输入投入至所述目标 LSTM 神经网络，得到所述目标 LSTM 神经网络输出的各个目标频谱图；

将所述各个目标频谱图分别进行时域变换，得到所述各个目标频谱图各自对应的目标单音色音频，作为所述目标音频的音频分离结果。

17. 根据权利要求 16 所述的可读存储介质，其特征在于，所述确定针对所述目标音频所需分离的各个音色种类包括：

获取所述目标音频的来源场合；

根据预设的场合音色对应关系确定与所述目标音频的来源场合对应的音色种类组合，所述场合音色对应关系记录了场合与音色种类组合之间的对应关系；

将确定出的所述音色种类组合中的各个音色种类确定为针对所述目标音频所需分离的各个音色种类。

18. 根据权利要求 16 或 17 所述的可读存储介质，其特征在于，每个音色种类组合对应的 LSTM 神经网络通过以下步骤预先训练得到：

获取待训练的 LSTM 神经网络对应的音色种类组合包含的各个音色种类，作为各个样本音色种类；

分别采集所述各个样本音色种类各自对应的单音色音频样本；

根据所述各个样本音色种类各自对应的单音色音频样本合成得到各个混合音频样本，每个混合音频样本由所述各个样本音色种类各自对应的一个单音色音频样本合成得到；

针对每个混合音频样本，将所述每个混合音频样本作为输入投入至所述 LSTM 神经网络，得到所述 LSTM 神经网络输出的各个样本频谱图；

使用预设的代价函数计算所述各个样本频谱图与各个单音色频谱图之间的误差，所述各个单音色频谱图是指所述每个混合音频样本对应的各个单音色音频样本进过频域变换得到的频谱图；

以所述代价函数的计算结果为目标，调整所述 LSTM 神经网络的网络参数，直到所述代价函数的计算结果收敛，然后确定所述 LSTM 神经网络已训练完成。

19. 根据权利要求 18 所述的可读存储介质，其特征在于，所述针对每个混合音频样本，将所述每个混合音频样本作为输入投入至所述 LSTM 神经网络，得到所述 LSTM 神经网络输出的各个样本频谱图包括：

将所述每个混合音频样本进行频域变换，得到所述每个混合音频样本的混合频谱图；

对所述混合频谱图加窗，且对加窗得到的每帧数据进行短时傅里叶变换，得到各个频谱特征向量；

对所述各个频谱特征向量进行重叠分组，得到各组频谱特征片段，任意相邻两组频谱特征片段之间存在重叠的频谱特征向量；

将每组频谱特征片段分别与预设卷积核进行卷积计算，得到降维后的各个片段向量，每个片段向量对应一组频谱特征片段；

将所述各个片段向量输入至 LSTM，得到所述 LSTM 输出的各个音频信息向量；

针对预设数量个多层感知器中的每个多层感知器，将所述各个音频信息向量输入至每个多层感知器，得到所述每个多层感知器输出的分离特征向量，所述预设数量等于所述各个样本音色种类的数量；

针对每个多层感知器输出的分离特征向量，将所述分离特征向量输入至所述 LSTM 进行特征还原，得到每个多层感知器对应的音频特征向量；

针对每个多层感知器对应的音频特征向量，将所述音频特征向量分别与所述预设卷积核进行转置卷积计算，得到升维后的、每个多层感知器对应的各个音频特征片段；

分别拼合各个多层感知器各自对应的各个音频特征片段，得到各个多层感知器各自对应的单音色频谱图，作为所述 LSTM 神经网络输出的各个样本频谱图。

20. 根据权利要求 18 所述的可读存储介质，其特征在于，每个混合音频样本通过以下步骤合成得到：

针对每个所述样本音色种类对应的单音色音频样本，从所述单音色音频样本中选取出一个单音色音频样本，作为待混音样本；

将所述各个样本音色种类各自对应的待混音样本进行混音处理，得到一个混合音频样本。

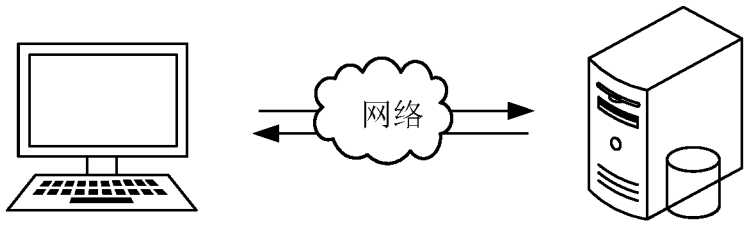

网络
客户端
服务器

图 1

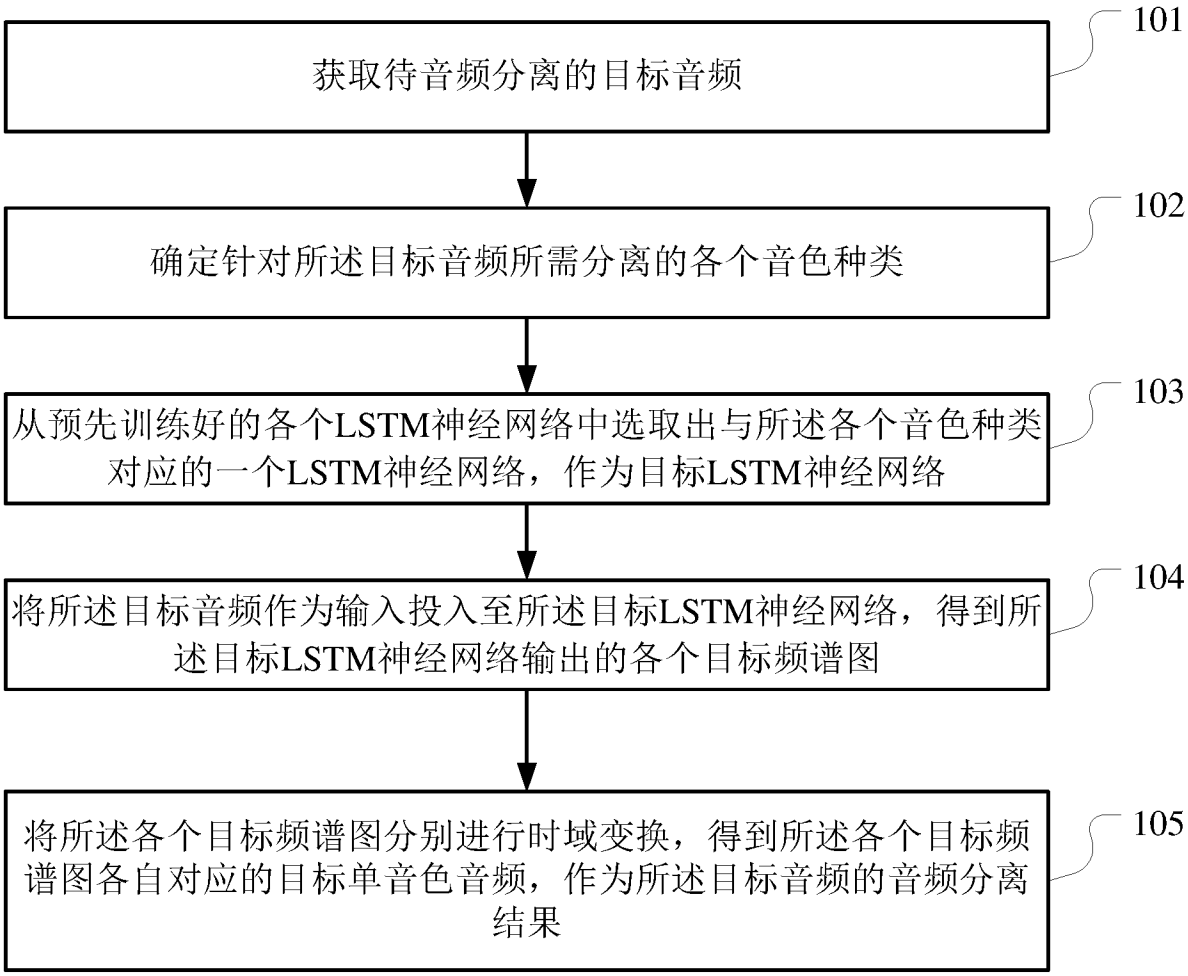

101
获取待音频分离的目标音频
102
确定针对所述目标音频所需分离的各个音色种类
103
从预先训练好的各个LSTM神经网络中选取出与所述各个音色种类对应的一个LSTM神经网络，作为目标LSTM神经网络
104
将所述目标音频作为输入投入至所述目标LSTM神经网络，得到所述目标LSTM神经网络输出的各个目标频谱图
105
将所述各个目标频谱图分别进行时域变换，得到所述各个目标频谱图各自对应的目标单音色音频，作为所述目标音频的音频分离结果

图 2

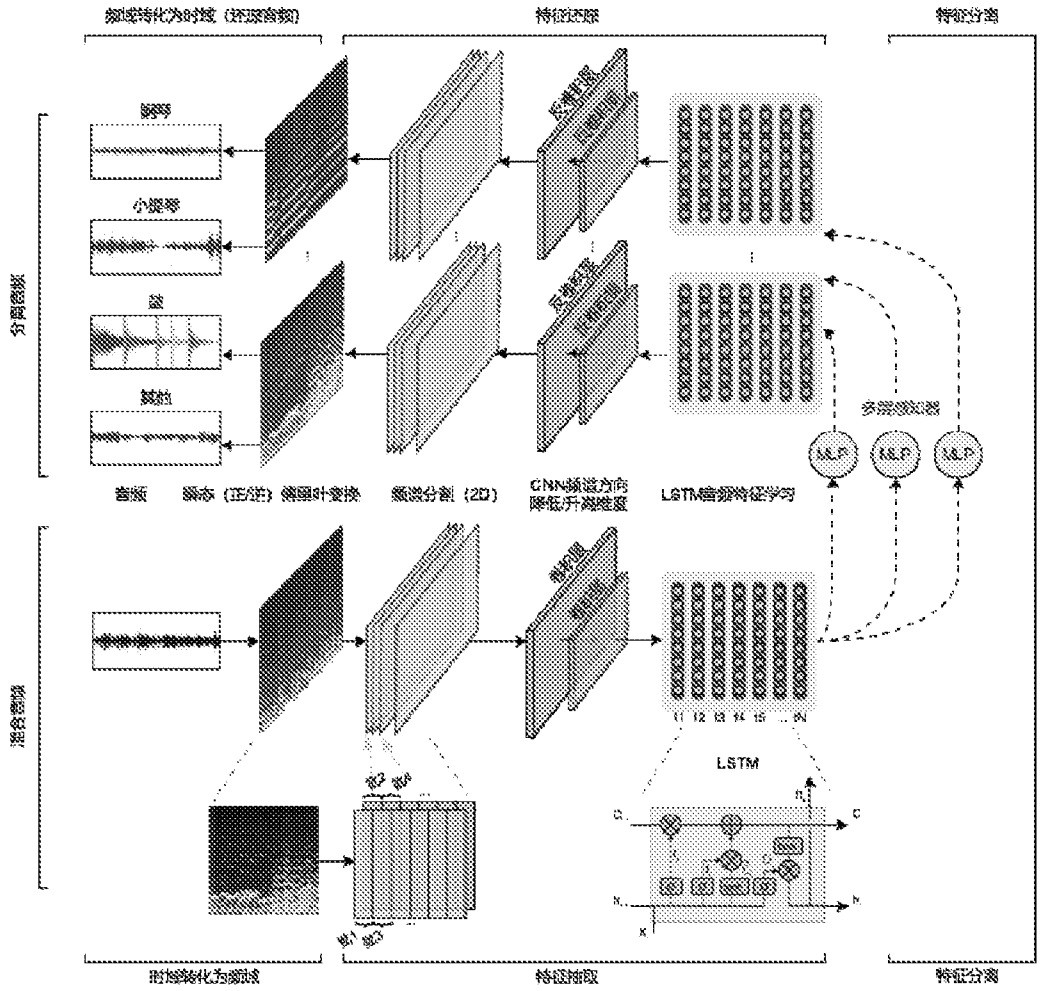

特征还原
特征分离
LSTM
MLP
MLP
MLP

图 3

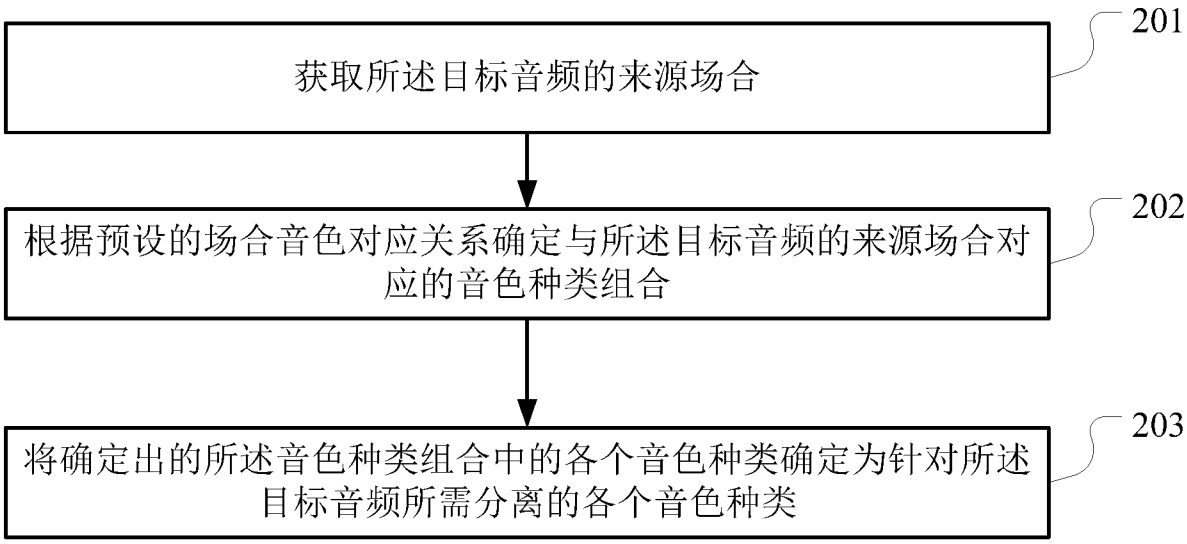

201
获取所述目标音频的来源场合
202
根据预设的场合音色对应关系确定与所述目标音频的来源场合对应的音色种类组合
203
将确定出的所述音色种类组合中的各个音色种类确定为针对所述目标音频所需分离的各个音色种类

图 4

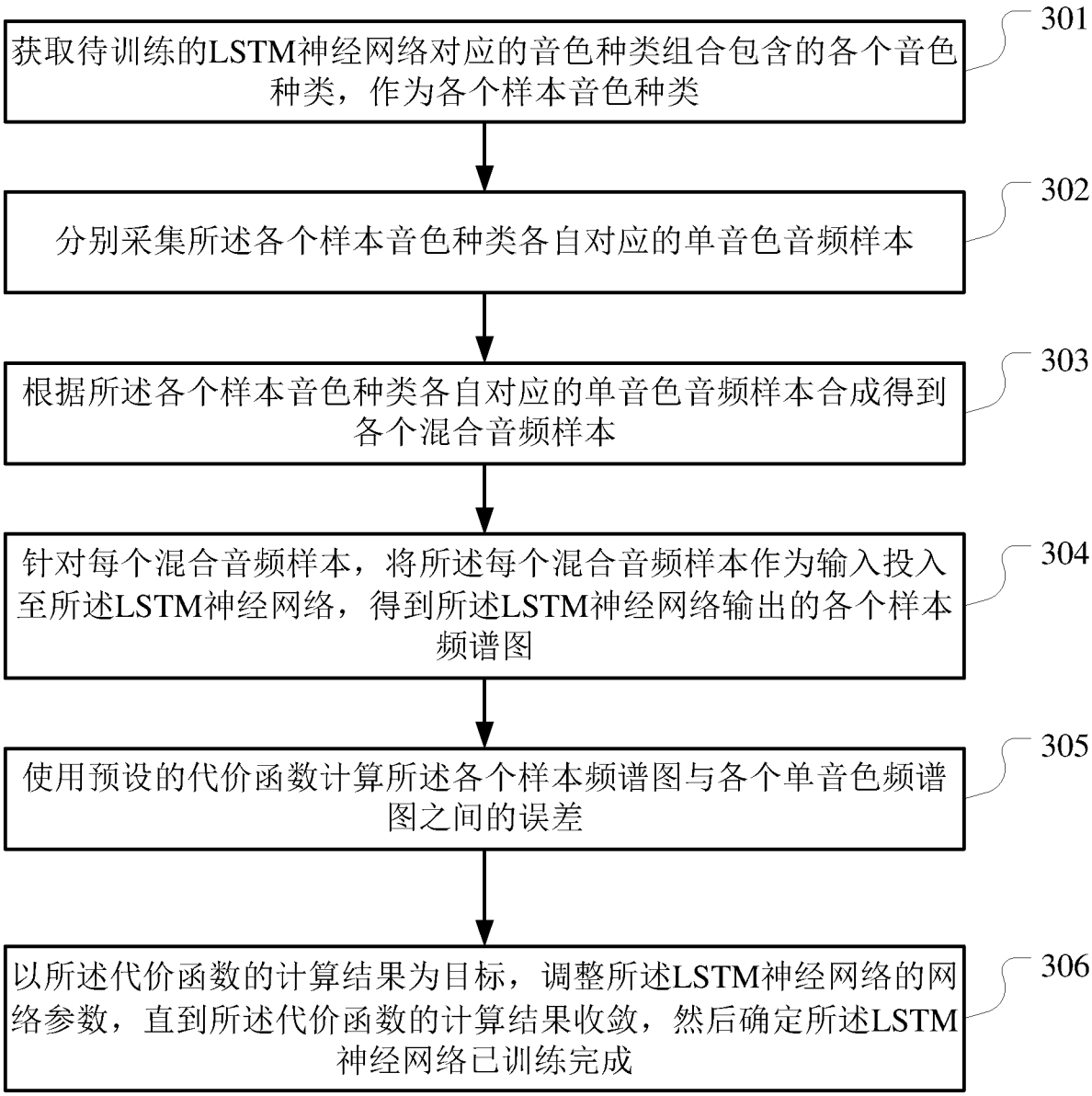

获取待训练的LSTM神经网络对应的音色种类组合包含的各个音色种类，作为各个样本音色种类
301
分别采集所述各个样本音色种类各自对应的单音色音频样本
302
根据所述各个样本音色种类各自对应的单音色音频样本合成得到各个混合音频样本
303
针对每个混合音频样本，将所述每个混合音频样本作为输入投入至所述LSTM神经网络，得到所述LSTM神经网络输出的各个样本频谱图
304
使用预设的代价函数计算所述各个样本频谱图与各个单音色频谱图之间的误差
305
以所述代价函数的计算结果为目标，调整所述LSTM神经网络的网络参数，直到所述代价函数的计算结果收敛，然后确定所述LSTM神经网络已训练完成
306

图 5

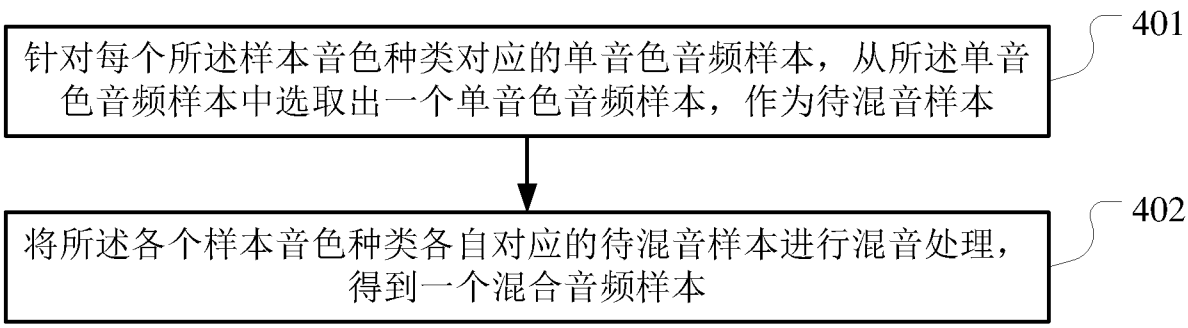

针对每个所述样本音色种类对应的单音色音频样本，从所述单音色音频样本中选取出一个单音色音频样本，作为待混音样本
401
将所述各个样本音色种类各自对应的待混音样本进行混音处理，得到一个混合音频样本
402

图 6

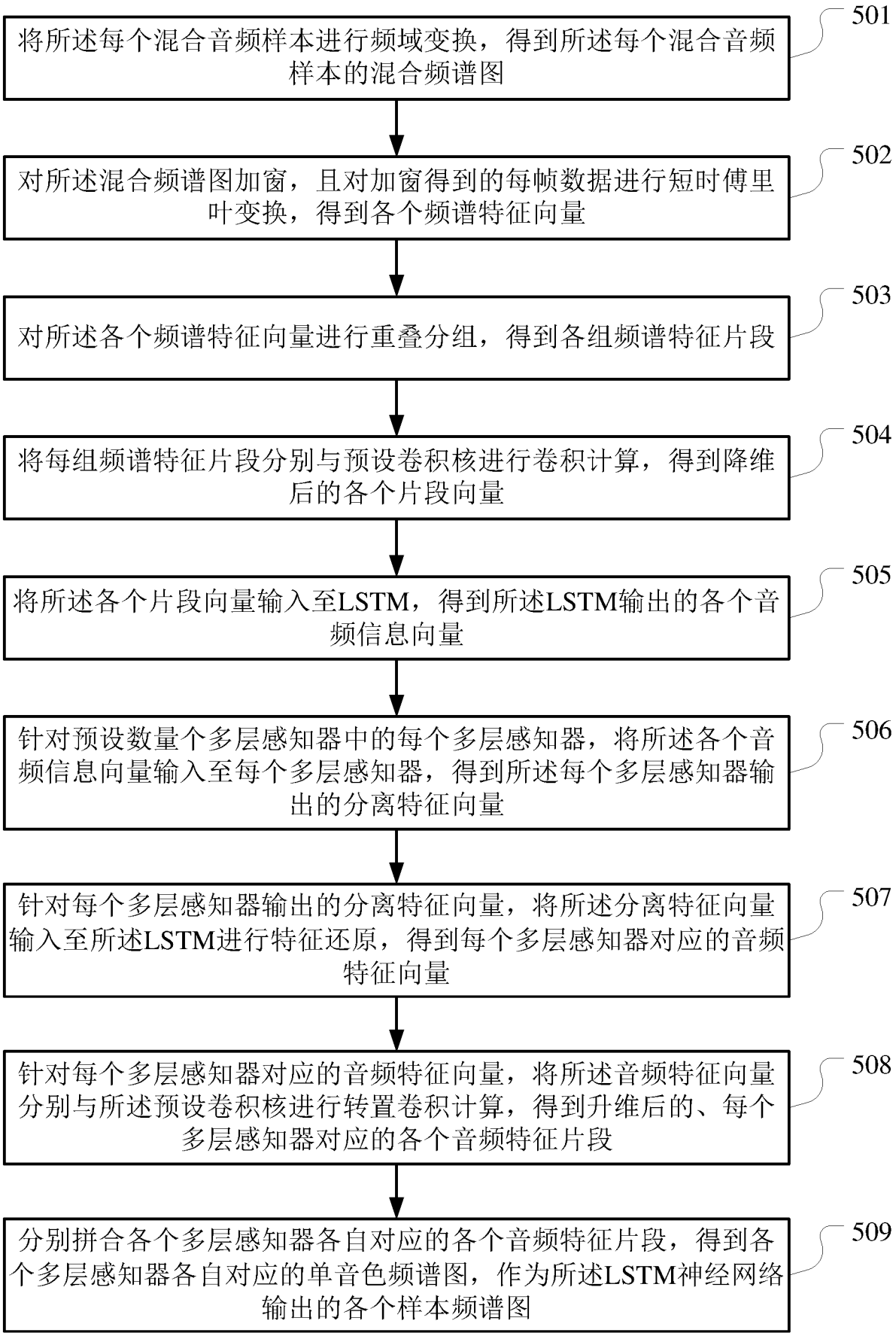

501
将所述每个混合音频样本进行频域变换，得到所述每个混合音频样本的混合频谱图
502
对所述混合频谱图加窗，且对加窗得到的每帧数据进行短时傅里叶变换，得到各个频谱特征向量
503
对所述各个频谱特征向量进行重叠分组，得到各组频谱特征片段
504
将每组频谱特征片段分别与预设卷积核进行卷积计算，得到降维后的各个片段向量
505
将所述各个片段向量输入至LSTM，得到所述LSTM输出的各个音频信息向量
506
针对预设数量个多层感知器中的每个多层感知器，将所述各个音频信息向量输入至每个多层感知器，得到所述每个多层感知器输出的分离特征向量
507
针对每个多层感知器输出的分离特征向量，将所述分离特征向量输入至所述LSTM进行特征还原，得到每个多层感知器对应的音频特征向量
508
针对每个多层感知器对应的音频特征向量，将所述音频特征向量分别与所述预设卷积核进行转置卷积计算，得到升维后的、每个多层感知器对应的各个音频特征片段
509
分别拼合各个多层感知器各自对应的各个音频特征片段，得到各个多层感知器各自对应的单音色频谱图，作为所述LSTM神经网络输出的各个样本频谱图

图 7

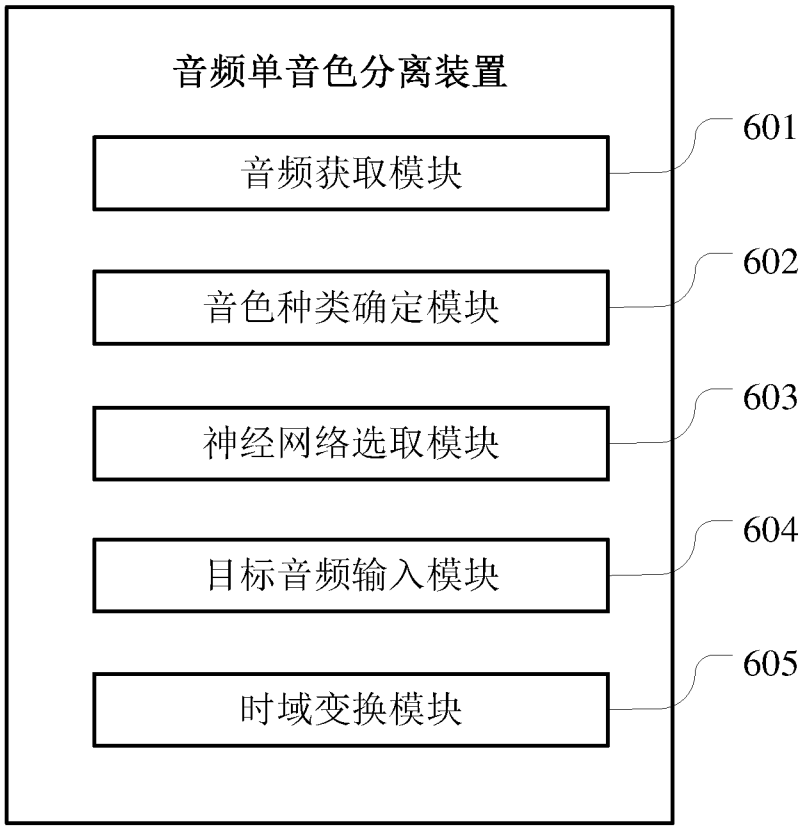

音频单音色分离装置
音频获取模块
601
音色种类确定模块
602
神经网络选取模块
603
目标音频输入模块
604
时域变换模块
605

图 8

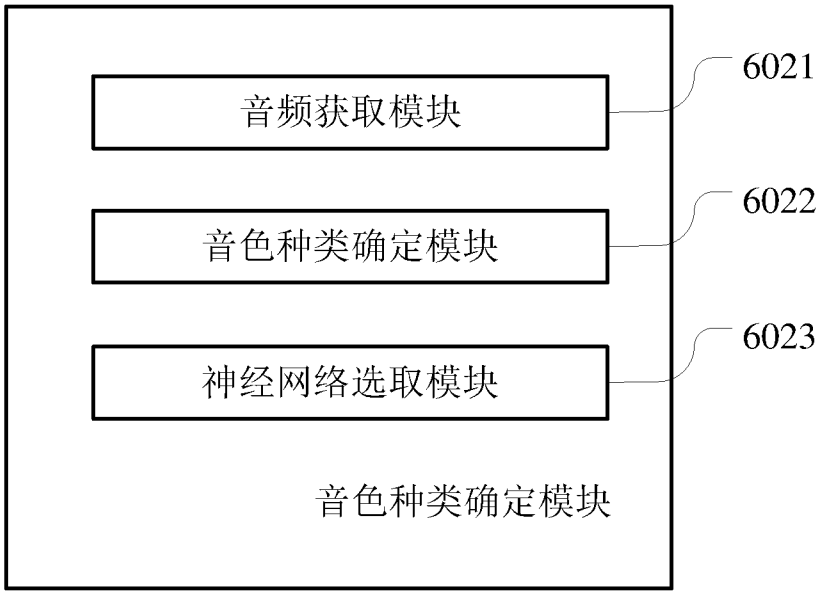

音频获取模块
6021
音色种类确定模块
6022
神经网络选取模块
6023
音色种类确定模块

图 9

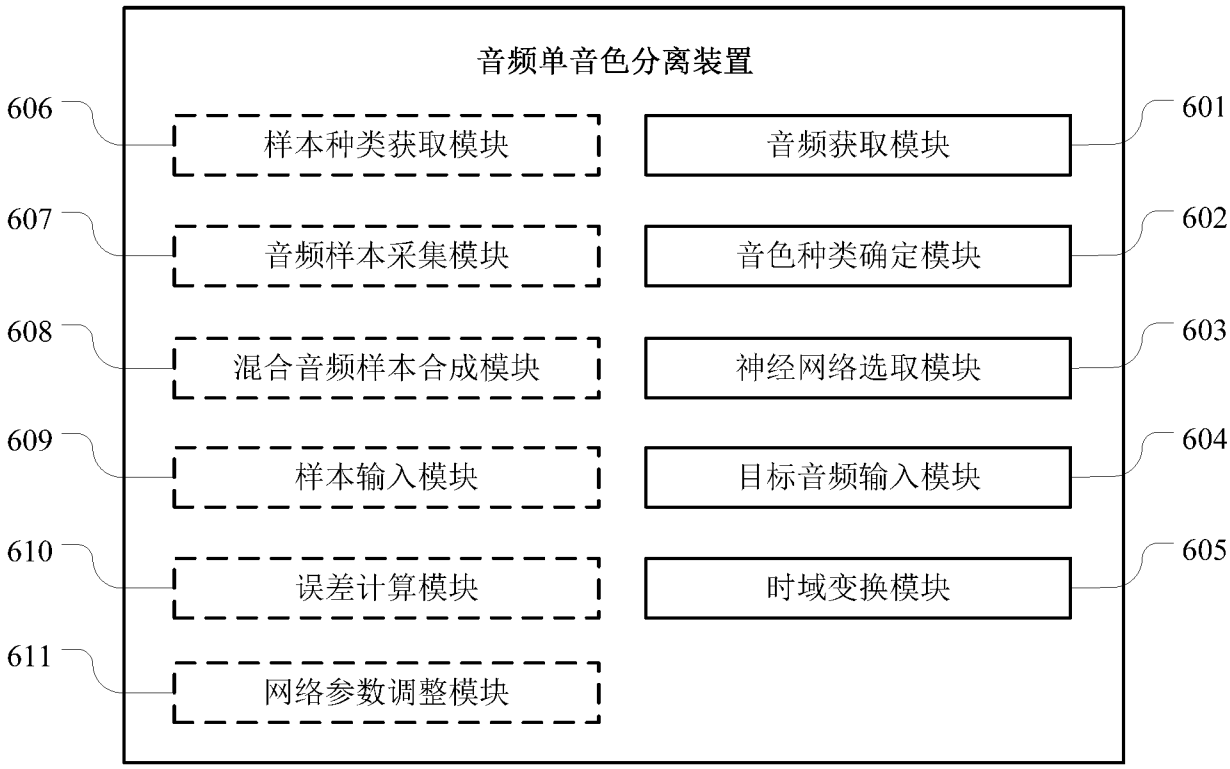

音频单音色分离装置
606
样本种类获取模块
601
音频获取模块
607
音频样本采集模块
602
音色种类确定模块
608
混合音频样本合成模块
603
神经网络选取模块
609
样本输入模块
604
目标音频输入模块
610
误差计算模块
605
时域变换模块
611
网络参数调整模块

图 10

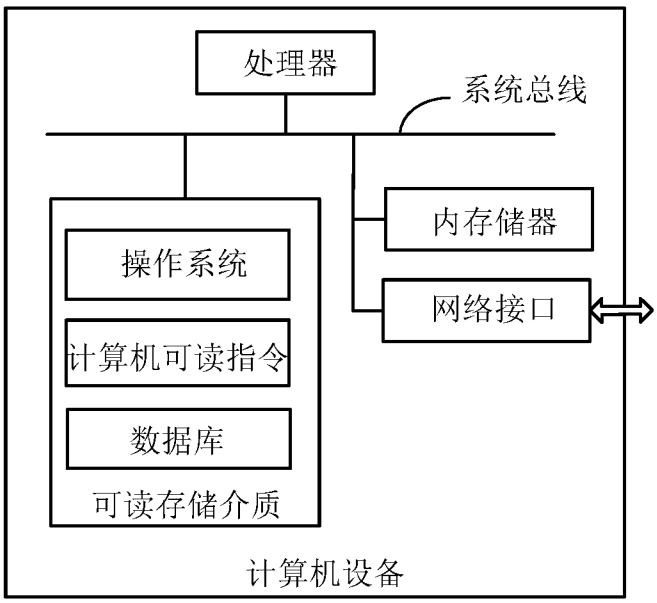

处理器
系统总线
内存储器
操作系统
网络接口
计算机可读指令
数据库
可读存储介质
计算机设备

图 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/117096

A. CLASSIFICATION OF SUBJECT MATTER

G10L 21/0272(2013.01)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)

CNPAT, WPI, EPODOC, CNKI: 音频, 语音, 声音, 分离, 音色, 场景, 场合, 频率, 频谱, 神经网络, 模型, sound, audio, speech, separate, scene, tone, frequency, spectrum, band, neutral network, model, LSTM

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110335622 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 15 October 2019 (2019-10-15) claims 1-10, description, paragraphs [0033]-[0137], and figures 1-11	1-20
X	CN 109378010 A (GREE ELECTRIC APPLIANCES, INC. OF ZHUHAI) 22 February 2019 (2019-02-22) description, paragraphs [0058]-[0144], and figures 1-4	1-3, 5-8, 10-13, 15-18, 20
A	CN 109801644 A (BEIJING DAJIA INTERCONNECTION INFORMATION TECHNOLOGY CO., LTD.) 24 May 2019 (2019-05-24) entire document	1-20
A	CN 109584903 A (INSTITUTE OF ACOUSTICS, CHINESE ACADEMY OF SCIENCES) 05 April 2019 (2019-04-05) entire document	1-20
A	US 2018158456 A1 (POSTECH ACADEMY-INDUSTRY FOUNDATION) 07 June 2018 (2018-06-07) 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	Date of mailing of the international search report
27 February 2020	**12 March 2020**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/117096

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 110335622 A	15 October 2019	None	
CN 109378010 A	22 February 2019	None	
CN 109801644 A	24 May 2019	None	
CN 109584903 A	05 April 2019	None	
US 2018158456 A1	07 June 2018	KR 20180062859 A	11 June 2018

国际检索报告

国际申请号
PCT/CN2019/117096

A. 主题的分类

G10L 21/0272(2013.01)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））

CNPAT, WPI, EPODOC, CNKI：音频，语音，声音，分离，音色，场景，场合，频率，频谱，神经网络，模型，sound, audio, speech, separate, scene, tone, frequency, spectrum, band, neutral network, model, LSTM

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 110335622 A (平安科技深圳有限公司) 2019年 10月 15日 (2019-10-15) 权利要求1-10，说明书第[0033]-[0137]段，附图1-11	1-20
X	CN 109378010 A (珠海格力电器股份有限公司) 2019年 2月 22日 (2019-02-22) 说明书第[0058]-[0144]段，附图1-4	1-3，5-8，10-13，15-18，20
A	CN 109801644 A (北京达佳互联信息技术有限公司) 2019年 5月 24日 (2019-05-24) 全文	1-20
A	CN 109584903 A (中国科学院声学研究所) 2019年 4月 5日 (2019-04-05) 全文	1-20
A	US 2018158456 A1 (POSTECH ACADEMY-INDUSTRY FOUNDATION) 2018年 6月 7日 (2018-06-07) 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	国际检索报告邮寄日期
2020年 2月 27日	2020年 3月 12日

ISA/CN的名称和邮寄地址	受权官员
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	陈成
传真号 (86-10)62019451	电话号码 86-(10)-53962548

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/117096

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 110335622 A	2019年 10月 15日	无	
CN 109378010 A	2019年 2月 22日	无	
CN 109801644 A	2019年 5月 24日	无	
CN 109584903 A	2019年 4月 5日	无	
US 2018158456 A1	2018年 6月 7日	KR 20180062859 A	2018年 6月 11日