

人声音符识别模型的训练方法、人声音符识别方法及设备

技术领域

本申请实施例涉及人工智能技术领域，特别涉及一种人声音符识别模型的训练方法、人声音符识别方法及设备。

背景技术

歌曲的人声音符识别是指根据带伴奏的歌曲，得到该歌曲的人声音符序列。

歌曲里除了包含人声之外，通常还包含各种乐器演奏组成的伴奏，有些现场歌曲里还包含各种背景噪声或混响，这给歌曲人声音符识别带来了较大的挑战。相关技术中，通过人声伴奏分离算法将歌曲中的人声音频分离出来，再通过人声音符识别模型对人声音频进行处理，得到歌曲的人声音符序列。

然而，上述方法需要在人声伴奏分离算法的基础上进行人声音符识别，计算复杂度较高。

发明内容

本申请实施例提供了一种人声音符识别模型的训练方法、人声音符识别方法及设备。所述技术方案如下：

根据本申请实施例的一个方面，提供了一种人声音符识别模型的训练方法，所述方法包括：

获取至少一个标注人声音频、各个所述标注人声音频分别对应的人声音符标注结果、至少一个纯人声音频以及至少一个伴奏音频；

基于所述标注人声音频、所述伴奏音频和所述标注人声音频对应的人声音符标注结果，对第一网络进行训练，得到训练后的第一网络；所述第一网络用于根据所述标注人声音频和所述伴奏音频的合成音频，输出所述标注人声音频对应的人声音符识别结果；

基于所述训练后的第一网络、所述纯人声音频和所述伴奏音频，对第二网络进行训练，得到人声音符识别模型；所述第二网络用于根据所述纯人声音频和所述伴奏音频的合成音频，输出所述纯人声音频对应的人声音符识别结果。

根据本申请实施例的一个方面，提供了一种人声音符识别方法，所述方法包括：

获取带伴奏的目标音频，所述目标音频中包含人声和伴奏；

获取所述目标音频的音频特征，所述音频特征包括所述目标音频在时频域上相关的特征；

通过人声音符识别模型对所述音频特征进行处理，得到所述目标音频的音符特征，所述音符特征包括与所述目标音频的人声音符相关的特征；

通过所述人声音符识别模型对所述音符特征进行处理，得到所述目标音频的人声音符序列；

其中，所述人声音符识别模型是基于训练后的第一网络、纯人声音频和伴奏音频，对第二网络进行训练得到的；所述第一网络用于根据标注人声音频和所述伴奏音频的合成音频，输出所述标注人声音频对应的人声音符识别结果；所述第二网络用于根据所述纯人声音频和所述伴奏音频的合成音频，输出所述纯人声音频对应的人声音符识别结果。

根据本申请实施例的一个方面，提供了一种人声音符识别模型的训练装置，所述装置包括：

样本获取模块，用于获取第一训练样本集、第二训练样本集和第三训练样本集，所述第一训练样本集中包括至少一个标注人声音频以及所述标注人声音频对应的人声音符标注结果，所述第二训练样本集中包括至少一个纯人声音频，所述第三训练样本集中包括至少一个伴奏

音频；

第一网络训练模块，用于基于所述标注人声音频、所述伴奏音频和所述标注人声音频对应的人声音符标注结果，对第一网络进行训练，得到训练后的第一网络；所述第一网络用于根据所述标注人声音频和所述伴奏音频的合成音频，输出所述标注人声音频对应的人声音符识别结果；

第二网络训练模块，用于基于所述训练后的第一网络、所述纯人声音频和所述伴奏音频，对第二网络进行训练，得到人声音符识别模型；所述第二网络用于根据所述纯人声音频和所述伴奏音频的合成音频，输出所述纯人声音频对应的人声音符识别结果。

根据本申请实施例的一个方面，提供了一种人声音符识别装置，所述装置包括：

音频获取模块，用于获取带伴奏的目标音频，所述目标音频中包含人声和伴奏；

特征获取模块，用于获取所述目标音频的音频特征，所述音频特征包括所述目标音频在时频域上相关的特征；

特征提取模块，用于通过人声音符识别模型对所述音频特征进行处理，得到所述目标音频的音符特征，所述音符特征包括与所述目标音频的人声音符相关的特征；

结果得到模块，用于通过所述人声音符识别模型对所述音符特征进行处理，得到所述目标音频的人声音符序列；

其中，所述人声音符识别模型是基于训练后的第一网络、纯人声音频和伴奏音频，对第二网络进行训练得到的；所述第一网络用于根据标注人声音频和所述伴奏音频的合成音频，输出所述标注人声音频对应的人声音符识别结果；所述第二网络用于根据所述纯人声音频和所述伴奏音频的合成音频，输出所述纯人声音频对应的人声音符识别结果。

根据本申请实施例的一个方面，提供了一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器中存储有计算机程序，所述处理器执行所述计算机程序以实现上述人声音符识别模型的训练方法，或者以实现上述人声音符识别方法。

根据本申请实施例的一个方面，提供了一种计算机可读存储介质，所述存储介质中存储有计算机程序，所述计算机程序用于被处理器执行，以实现上述人声音符识别模型的训练方法，或者以实现上述人声音符识别方法。

根据本申请实施例的一个方面，提供了一种计算机程序产品，所述计算机程序产品包括计算机指令，所述计算机指令存储在计算机可读存储介质中，处理器从所述计算机可读存储介质读取并执行所述计算机指令，以实现上述人声音符识别模型的训练方法，或者以实现上述人声音符识别方法。

本申请实施例提供的技术方案可以包括如下有益效果：

通过上述训练方法得到的人声音符识别模型，能够直接从带伴奏的目标音频中识别出对应的人声音符序列，因而在模型使用阶段，无需调用人声伴奏分离算法从目标音频中提取出人声音频，降低了人声音符识别的计算复杂度。另外，本申请采用了半监督训练的方法，通过少量标注样本对第一网络进行训练，然后通过第一网络和大量未标注样本对第二网络进行训练，这样仅需要少量标注样本，即可训练出泛化性能强的模型，降低了训练样本的获取成本。

附图说明

图1是本申请一个实施例提供的方案实施环境的示意图；

图2是本申请一个实施例提供的人声音符识别模型的训练方法的流程图；

图3是本申请另一个实施例提供的人声音符识别模型的训练方法的流程图；

图4是本申请另一个实施例提供的人声音符识别模型的训练方法的流程图；

图5是本申请一个实施例提供的人声音符识别模型的训练方法的示意图；

图6是本申请一个实施例提供的人声音符识别方法的流程图；

图 7 是本申请一个实施例提供的人声音符识别模型的示意图；
图 8 是本申请一个实施例提供的人声音符识别模型的训练装置的框图；
图 9 是本申请另一个实施例提供的人声音符识别模型的训练装置的框图；
图 10 是本申请一个实施例提供的人声音符识别装置的框图；
图 11 是本申请一个实施例提供的计算机设备的结构示意图。

具体实施方式

为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

请参考图 1，其示出了本申请一个实施例提供的方案实施环境的示意图。该方案实施环境可以包括：模型使用设备 10 和模型训练设备 20。

模型使用设备 10 用于执行本申请实施例中的声音符识别方法。模型使用设备 10 可以是终端设备 11，也可以是服务器 12。终端设备 11 可以是诸如手机、平板电脑、游戏主机、电子书阅读器、多媒体播放设备、可穿戴设备、PC (Personal Computer, 个人计算机)、车载终端等电子设备。终端设备 11 中可以运行有目标应用程序，或者目标应用程序的客户端。在本申请实施例中，上述目标应用程序是指提供声音符识别功能的应用程序。可选地，目标应用程序可以是系统级应用程序，如操作系统或者操作系统提供的原生应用程序；也可以是第三方应用程序，如用户自行下载安装的第三方应用程序，本申请实施例对此不作限定。

服务器 12 可以是上述目标应用程序的后台服务器，用于为终端设备 11 中的目标应用程序提供后台服务。服务器 12 可以是一台服务器，也可以是由多台服务器组成的服务器集群，或者是一个云计算服务中心。可选地，服务器 12 同时为多个终端设备 11 中的目标应用程序提供后台服务。

终端设备 11 和服务器 12 之间可通过网络 13 进行互相通信。该网络 13 可以有有线网络，也可以是无网络。

本申请实施例提供的声音符识别方法，各步骤的执行主体可以是计算机设备，所述计算机设备是指具备数据计算、处理和存储能力的电子设备。例如，可以由终端设备 11 执行声音符识别方法（如终端设备 11 中安装运行的目标应用程序的客户端执行该声音符识别方法），也可以由服务器 12 执行该声音符识别方法，或者由终端设备 11 和服务器 12 交互配合执行，本申请对此不作限定。例如，由终端设备 11 获取目标音频，并将目标音频发送给服务器 12，由服务器 12 执行该声音符识别方法，得到声音符序列。

模型训练设备 20 用于执行本申请实施例中的声音符识别模型的训练方法。模型训练设备 20 可以是服务器，也可以是计算机设备，所述计算机设备是指具备数据计算、处理和存储能力的电子设备。由模型训练设备 20 对声音符识别模型进行训练，将训练好的声音符识别模型部署在模型使用设备 10 中。

请参考图 2，其示出了本申请一个实施例提供的声音符识别模型的训练方法的流程图。该方法可以包括如下步骤 210~230 中的至少一个步骤。

步骤 210，获取至少一个标注人声音频、各个标注人声音频分别对应的声音符标注结果、至少一个纯人声音频以及至少一个伴奏音频。

在一些实施例中，可以获取第一训练样本集、第二训练样本集和第三训练样本集，第一训练样本集中包括至少一个标注人声音频以及标注人声音频对应的声音符标注结果，第二训练样本集中包括至少一个纯人声音频，第三训练样本集中包括至少一个伴奏音频。

人声是指歌曲中歌词、和声等由人声演唱的部分。非人声是指歌曲中除人声部分之外的部分，如伴奏、混响、噪音等。

标注人声音频是指无伴奏的清唱音频，且标注了音频包含的各个音频帧对应的声音符。标注人声音频对应的声音符标注结果是指标注人声音频包含的各个音频帧对应的声音符

构成的人声音符序列。

纯人声音频是指在带伴奏的歌曲音频中分离得到的仅包含人声的音频。

伴奏音频是指在带伴奏的歌曲音频中分离得到的仅包含伴奏的音频。

在一些实施例中，可以采用人声伴奏分离算法，从带伴奏的歌曲中分离得到纯人声音频和伴奏音频。通过对多首歌曲执行上述分离操作，可以得到多个纯人声音频用来构建第二训练样本集，且可以得到多个伴奏音频用来构建第三训练样本集。

在一些实施例中，第一训练样本集中包含的标注人声音频的数量，远少于第二训练样本集中包含的纯人声音频的数量。示例性地，第一训练样本集中包含 100 首标注人声音频，第二训练样本集中包含 10000 首纯人声音频。

对于第三训练样本集中的伴奏音频的数量，本申请不作限定。例如，第三训练样本集中的伴奏音频的数量可以与第二训练样本集中的纯人声音频的数量相同，也可以不同。

步骤 220，基于标注人声音频、伴奏音频和标注人声音频对应的人声音符标注结果，对第一网络进行训练，得到训练后的第一网络；第一网络用于根据标注人声音频和伴奏音频的合成音频，输出标注人声音频对应的人声音符识别结果。

第一网络是指初始化的人声音符识别模型。在一些实施例中，第一网络也可以称为教师网络，第二网络也可以称为学生网络。

在一些实施例中，采用伴奏音频与标注人声音频进行合成，得到标注人声音频对应的合成音频；基于标注人声音频对应的合成音频以及标注人声音频对应的人声音符标注结果，对第一网络进行训练，得到训练后的第一网络。

在一些实施例中，标注人声音频对应的合成音频包含伴奏音频和标注人声音频。

在一些实施例中，通过第一网络，对标注人声音频对应的合成音频进行处理，得到标注人声音频对应的人声音符识别结果，作为人声音符第一识别结果；根据人声音符第一识别结果以及人声音符标注结果，对第一网络进行训练，得到训练后的第一网络。

人声音符第一识别结果是指通过第一网络得到的纯人声音频的人声音符序列。通过将标注人声音频对应的合成音频输入至第一网络，由第一网络对标注人声音频对应的合成音频进行处理，输出得到标注人声音频对应的人声音符第一识别结果。在一些实施例中，根据损失函数，对第一网络进行训练，得到训练后的第一网络。对于具体的损失函数，本申请不作限定。示例性地，可以采用交叉熵损失函数、指数损失函数、log 对数损失函数、绝对值损失函数、Focal-Loss 损失函数等。

在一些实施例中，通过计算人声音符第一识别结果，与人声音符标注结果之间的损失函数值，对第一网络的参数进行调整，得到训练后的第一网络。

在一些实施例中，通过计算人声音符第一识别结果，与人声音符标注结果之间的损失函数值，调整第一网络的参数，对第一网络进行训练。

在一些实施例中，第一网络包含输入层、中间层和输出层。输入层用于输入标注人声音频对应的合成音频的音频特征；中间层用于根据音频特征，提取标注人声音频对应的合成音频的音符特征；输出层用于根据音符特征，得到标注人声音频对应的合成音频的人声音符序列。

在一些实施例中，输入层根据标注人声音频对应的合成音频，获取标注人声音频对应的合成音频的音频特征，并传输给中间层。

在一些实施例中，输入层直接获取标注人声音频对应的合成音频的音频特征，并传输给中间层。

在一些实施例中，输出层还用于识别音符特征的人声部分和非人声部分。

在一些实施例中，根据音符特征的人声部分、人声音符第一识别结果以及人声音符标注结果，对第一网络进行训练，得到训练后的第一网络。

在一些实施例中，第一网络为神经网络，对于具体的网络结构，本申请不作限定。

步骤 230, 基于训练后的第一网络、纯人声音频和伴奏音频, 对第二网络进行训练, 得到人声音符识别模型; 第二网络用于根据纯人声音频和伴奏音频的合成音频, 输出纯人声音频对应的人声音符识别结果。

在一些实施例中, 基于训练后的第一网络、纯人声音频和伴奏音频, 对第二网络进行训练。

第二网络是指初始化的人声音符识别模型。在一些实施例中, 第二网络为神经网络, 对于具体的网络结构, 本申请不作限定。

在一些实施例中, 第二网络与第一网络为结构相同, 初始化参数相同的两个网络。

在一些实施例中, 通过训练后的第一网络对纯人声音频进行处理, 得到纯人声音频对应的人声音符识别结果, 作为人声音符第二识别结果; 将人声音符第二识别结果, 确定为纯人声音频对应的伪标签信息; 根据纯人声音频、伴奏音频和纯人声音频对应的伪标签信息, 对第二网络进行训练。

在一些实施例中, 可以将人声音符第二识别结果直接确定为伪标签信息。方案简单易行, 计算成本低。

在一些实施例中, 对人声音符第二识别结果进行修正, 将修正后得到的人声音符序列确定为伪标签信息。对人声音符第二识别结果进行修正, 提高了伪标签信息的准确性, 进一步提高了训练后得到的人声音符识别模型的准确性。

在一些实施例中, 采用伴奏音频与纯人声音频进行合成, 得到纯人声音频对应的合成音频; 根据纯人声音频对应的合成音频和伪标签信息, 对第二网络进行训练。

在一些实施例中, 纯人声音频对应的合成音频包含伴奏音频与纯人声音频。

在一些实施例中, 通过第二网络对纯人声音频对应的合成音频进行处理, 得到纯人声音频对应的人声音符识别结果, 作为人声音符第三识别结果; 根据人声音符第三识别结果以及伪标签信息, 对第二网络进行训练。人声音符第三识别结果是指通过第二网络得到的纯人声音频的人声音符序列。通过纯人声音频对应的合成音频输入至第二网络, 由第二网络对纯人声音频对应的合成音频进行处理, 输出得到人声音符第三识别结果。

在一些实施例中, 根据损失函数, 对第二网络进行训练。对于具体地损失函数, 本申请不作限定。示例性地, 可以采用交叉熵损失函数、指数损失函数、log 对数损失函数、绝对值损失函数、Focal-Loss 损失函数等。

在一些实施例中, 通过计算人声音符第三识别结果, 与伪标签信息之间的损失函数值, 对第二网络的参数进行调整, 得到人声音符识别模型。

在一些实施例中, 通过计算人声音符第三识别结果, 与伪标签信息之间的损失函数值, 调整第二网络的参数, 对第二网络进行训练。

在一些实施例中, 第二网络包含输入层、中间层和输出层。输入层用于输入纯人声音频对应的合成音频的音频特征; 中间层用于根据音频特征, 提取纯人声音频对应的合成音频的音符特征; 输出层用于根据音符特征, 得到纯人声音频对应的合成音频的人声音符序列。

在一些实施例中, 输出层还用于识别音符特征的人声部分和非人声部分。

在一些实施例中, 输入层用于根据纯人声音频对应的合成音频, 获取纯人声音频对应的合成音频的音频特征, 并传输给中间层。

在一些实施例中, 输入层用于直接获取纯人声音频对应的合成音频的音频特征, 并传输给中间层。

在一些实施例中, 根据音符特征的人声部分、人声音符第二识别结果以及伪标签信息, 对第二网络进行训练。

在一些实施例中, 对第一网络进行训练的损失函数, 和对第二网络进行训练的损失函数可以相同, 也可以不同, 本申请对此不作限定。示例性地, 对第一网络进行训练的损失函数, 和对第二网络进行训练的损失函数均为交叉熵损失函数。示例性地, 对第一网络进行训练的

损失函数为交叉熵损失函数，对第二网络进行训练的损失函数为绝对值损失函数。

人声音符序列是指表征人声的音高区间的音符序列，其中包含不同音高区间的起点、偏移点和音高值。偏移点是指该音高区间的终点，可以采用其相对于起点的偏移量来表示，因此称为偏移点。音高是指各种音调高低不同的声音，即音的高度，是音的基本特征之一。音高区间是指具有相同音高的一段音频区间。

在一些实施例中，人声音符序列为 MIDI（Musical Instrument Digital Interface，乐器数字接口）序列。

在一些实施例中，停止训练条件为第二网络收敛，即通过第二网络得到的纯人声音频对应的人声音符第二识别结果，无限接近纯人声音频对应的伪标签信息。

在一些实施例中，根据损失函数判断第二网络是否满足停止训练条件。例如，第二网络的停止训练条件为损失函数值取得最小值。

在一些实施例中，停止训练条件可以设置为迭代次数，达到设定迭代次数即为满足停止训练条件。迭代次数可以根据步骤 230 的执行次数进行计算。

在一些实施例中，如图 3 所示，该方法还包括步骤 232，判断第二网络是否满足停止训练条件；若是，则将训练后的第二网络确定为人声音符识别模型，若否，则将训练后的第二网络确定为训练后的第一网络，并再次执行上述步骤 230。即，在第二网络未满足停止训练条件的情况下，将训练后的第二网络确定为训练后的第一网络，并再次从基于训练后的第一网络、纯人声音频和伴奏音频，对第二网络进行训练的步骤（步骤 230）开始执行。

示例性地，第二网络在第 n 次训练后满足停止训练条件，对于 n 次训练中的第 i 次训练，将第 $i-1$ 次训练后的第二网络确定为第 i 次训练的第一网络，并再次从基于训练后的第一网络、纯人声音频和伴奏音频，对第二网络进行训练的步骤（步骤 230）开始执行，其中， n 为大于 2 的整数， i 为大于 1 的整数。

本申请实施例提供的技术方案，通过上述训练方法得到的人声音符识别模型，能够直接从带伴奏的目标音频中识别出对应的人声音符序列，因而在模型使用阶段，无需调用人声伴奏分离算法从目标音频中提取出人声音频，降低了人声音符识别的计算复杂度。另外，本申请采用了半监督训练的方法，通过少量标注样本对第一网络进行训练，然后通过第一网络和大量未标注样本对第二网络进行训练，这样仅需要少量标注样本，即可训练出泛化性能强的模型，降低了训练样本的获取成本。

请参考图 4，其示出了本申请另一个实施例提供的人声音符识别模型的训练方法的流程图。该方法可以包括如下步骤 410~440 中的至少一个步骤。

步骤 410，获取至少一个标注人声音频、各个所述标注人声音频分别对应的人声音符标注结果、至少一个纯人声音频以及至少一个伴奏音频。

在一些实施例中，获取清唱数据集和歌曲数据集，清唱数据集中包括至少一个无伴奏的清唱音频以及清唱音频对应的人声音符标注结果，歌曲数据集中包括至少一个带伴奏的歌曲音频。

清唱音频是指在无伴奏环境中演唱的人声音频。清唱音频对应的人声音符标注结果是指清唱音频包含的各个音频帧对应的人声音符构成的人声音符序列。

歌曲音频是指由歌词和伴奏相结合的音频，其中包含伴奏和人声。在一些实施例中，歌曲音频还包含噪音和混响。

在一些实施例中，根据清唱音频以及清唱音频对应的人声音符标注结果，生成标注人声音频以及标注人声音频对应的人声音符标注结果，构建得到第一训练样本集。

在一些实施例中，对清唱音频进行检测，得到清唱音频中的静音部分和清音部分；将清唱音频确定为标注人声音频；从清唱音频对应的人声音符标注结果中，删除静音部分对应的人声音符标注结果和清音部分对应的人声音符标注结果，生成标注人声音频对应的人声音符

标注结果, 构建得到第一训练样本集。

在一些实施例中, 通过人声检测算法, 对清唱音频进行检测, 得到清唱音频中的静音部分和清音部分。

采用上述方式, 确保清唱音频对应的人声音符标注结果只在人声部分有音高, 静音部分和清音部分无音高, 保证清唱音频对应的人声音符标注结果的准确性。

在一些实施例中, 对歌曲音频进行人声分离操作, 得到人声音频和伴奏音频; 根据人声音频, 生成纯人声音频, 构建得到第二训练样本集; 根据伴奏音频, 构建得到第三训练样本集。

对于对歌曲音频进行人声分离操作的具体方式, 本申请不作限定。例如, 通过人声伴奏分离算法, 对歌曲音频进行人声分离操作, 得到人声音频和伴奏音频。

在一些实施例中, 对人声音频进行检测, 得到人声音频中的非人声部分; 删除人声音频中的非人声部分, 生成纯人声音频; 根据纯人声音频, 构建得到第二训练样本集。

在一些实施例中, 通过人声检测算法对人声音频进行检测, 得到人声音频中的非人声部分, 删除人声音频中的非人声部分, 生成纯人声音频。示例性地, 通过人声检测算法对人声音频进行检测, 得到人声音频中的非人声部分, 删除人声音频中超过 3 秒的非人声部分, 生成纯人声音频。一般歌曲中人声只占据其中的一部分, 而训练所需要的第二训练样本集中的训练样本的数量大, 删除人声音频中的非人声部分, 可以提升训练效率, 节省第二训练样本集所需要的存储空间。

在一些实施例中, 将得到的所有纯人声音频, 构建得到第二训练样本集。

由于人声伴奏分离算法不能保证完美地将每一首歌曲的人声和伴奏分离开, 因此需要对纯人声音频进行清洗, 将残留有伴奏的纯人声音频剔除掉。

在一些实施例中, 对纯人声音频中的每一个音频帧, 检测音频帧是否为人声音频帧, 并计算音频帧的能量; 若音频帧不是人声音频帧, 且音频帧的能量小于第二阈值, 则将音频帧确定为无效帧; 若纯人声音频中的无效帧数量在纯人声音频包含的音频帧总数中的占比大于第三阈值, 则将该纯人声音频确定为无效纯人声音频; 根据除无效纯人声音频之外的纯人声音频, 生成纯人声音频。

在一些实施例中, 第二阈值与第三阈值的具体取值可以根据实际需要进行设定, 本申请不作限定。示例性地, 对于不同风格的歌曲, 第二阈值的取值可以不同, 例如摇滚歌曲的第二阈值高于古风歌曲的第二阈值。

示例性地, 第三阈值的取值设为 30%, 若纯人声音频中的无效帧数量在纯人声音频包含的音频帧总数中的占比大于 30%, 则将该纯人声音频确定为无效纯人声音频。

在一些实施例中, 将得到的除无效纯人声音频之外的所有纯人声音频, 生成纯人声音频。

步骤 420, 采用伴奏音频与标注人声音频进行合成, 得到标注人声音频对应的合成音频。

在一些实施例中, 从至少一个伴奏音频中随机选择伴奏音频作为目标伴奏音频; 对标注人声音频进行数据增强处理, 得到处理后的标注人声音频; 其中, 数据增强处理包括以下至少之一: 添加混响、改变基频; 将目标伴奏音频与处理后的标注人声音频进行合成, 得到标注人声音频对应的合成音频。

在一些实施例中, 从第三训练样本集中随机选择伴奏音频作为目标伴奏音频。

声波在传播中遇到障碍物时, 会被障碍物反射, 每反射一次都要被障碍物吸收一些。这样, 当声源停止发声后, 声波还会经过多次反射和吸收, 最后才消失, 我们就感觉到声源停止发声后还有若干个声波混合持续一段时间, 这种现象叫做混响。对标注人声音频添加混响, 能够改变标注人声音频的音质。

改变基频是指在一定范围内改变标注人声音频的基频, 以及该标注人声音频对应的人声音符标注结果。对于改变基频的范围, 本申请不作限定。示例性地, 在 -200~+300 音分的范围内改变标注人声音频的基频, 并将该标注人声音频对应的人声音符标注结果调整到对应的

音高。例如，将标注人声音频的基频调高 200 音分，并将该标注人声音频对应的人声音符标注结果的音高也调高 200 音分。

在一些实施例中，可以改变标注人声音频中包含的各个音频帧的任意一个或多个音频帧的基频，以及该一个或多个音频帧对应的人声音符标注结果的音高。

步骤 430，基于标注人声音频对应的合成音频以及标注人声音频对应的人声音符标注结果，对第一网络进行训练，得到训练后的第一网络。

在一些实施例中，通过第一网络对标注人声音频对应的合成音频进行处理，得到标注人声音频对应的人声音符识别结果，作为人声音符第一识别结果；根据人声音符第一识别结果和人声音符标注结果，确定第一网络的损失函数值；根据第一网络的损失函数值，对第一网络的参数进行调整，得到训练后的第一网络。

在一些实施例中，采用交叉熵损失函数对第一网络进行训练。

在一些实施例中，基于标注人声音频对应的合成音频以及人声音符标注结果，对第一网络进行训练，直至收敛，得到训练后的第一网络。

步骤 440，基于训练后的第一网络、纯人声音频和伴奏音频，对第二网络进行训练，得到人声音符识别模型。

在一些实施例中，通过训练后的第一网络对纯人声音频进行处理，得到纯人声音频对应的人声音符识别结果，作为人声音符第二识别结果；将人声音符第二识别结果确定为纯人声音频对应的伪标签信息；根据纯人声音频、伴奏音频和伪标签信息，对第二网络进行训练。

在一些实施例中，提取纯人声音频的基频；根据纯人声音频的基频，对人声音符第二识别结果进行修正，得到纯人声音频对应的伪标签信息。

在一些实施例中，通过基频提取算法，提取纯人声音频的基频。

在一些实施例中，对于人声音符第二识别结果中包含的每一个音符，计算音符与音符对应的发音位置的基频之间的音高差；若音高差大于第一阈值，则将该音符的音高修正为音符对应的发音位置的基频的音高；若音高差小于或等于第一阈值，则保持音符的音高不变。

在一些实施例中，对于第一阈值的取值，本申请不作限定。

示例性地，第一阈值的取值为 3 个 MIDI 值，则若音符与音符对应的发音位置的基频之间的音高差大于 3 个 MIDI 值，则将该音符的音高修正为音符对应的发音位置的基频的音高；若音高差小于或等于 3 个 MIDI 值，则保持该音符的音高不变。

例如，该音符对应的发音位置的基频为 5 个 MIDI 值，若该音符的音高小于 2 个 MIDI 值，或者该音符的音高大于 8 个 MIDI 值，则将该音符的音高修正为 5 个 MIDI 值；若该音符的音高位于 2 个 MIDI 值至 8 个 MIDI 值之间，则保持该音符的音高不变。

通过上述方式对人声音符第二识别结果进行修正，保证了纯人声音频对应的伪标签信息的准确性，使得半监督训练的方法更加高效、稳定。

在一些实施例中，采用伴奏音频与纯人声音频进行合成，得到纯人声音频对应的合成音频；通过第二网络对纯人声音频对应的合成音频进行处理，得到纯人声音频对应的人声音符识别结果，作为人声音符第三识别结果；根据人声音符第三识别结果和伪标签信息，对第二网络进行训练。

在一些实施例中，根据人声音符第三识别结果和伪标签信息，确定第二网络的损失函数值；根据第二网络的损失函数值，对第二网络的参数进行调整，得到人声音符识别模型。

在一些实施例中，采用交叉熵损失函数对第二网络进行训练。

在一些实施例中，第二网络还可以对纯人声音频对应的合成音频进行人声识别，得到纯人声音频对应的合成音频的人声部分和纯人声音频对应的合成音频的非人声部分，进而根据纯人声音频对应的合成音频的人声部分、纯人声音频对应的合成音频的非人声部分和纯人声音频对第二网络进行训练。

在一些实施例中，可以通过全连接层，对纯人声音频对应的合成音频进行人声识别，得

到纯人声音频对应的合成音频的人声部分和纯人声音频对应的合成音频的非人声部分。示例性地,可以采用 Softmax 作为分类器,对纯人声音频对应的合成音频的人声部分和纯人声音频对应的合成音频的非人声部分进行分类。

在一些实施例中,该方法还包括步骤 442,判断第二网络是否满足停止训练条件;若是,则将训练后的第二网络确定为人声音符识别模型;若否,则将训练后的第二网络确定为训练后的第一网络,并再次执行上述步骤 440。

示例性地,请参考图 5,其示出了本申请一个实施例提供的人声音符识别模型的训练方法的示意图。

步骤一:从第三训练样本集(也可以称为数据集 3) 511 中随机选择伴奏音频,作为目标伴奏音频;对第一训练样本集(也可以称为数据集 1) 512 中的标注人声音频进行数据增强处理,得到处理后的标注人声音频;将目标伴奏音频与处理后的标注人声音频进行合成,得到标注人声音频对应的合成音频。

通过教师网络 513 对标注人声音频对应的合成音频进行处理,得到标注人声音频对应的人声音符识别结果,作为人声音符第一识别结果;根据人声音符第一识别结果和标注人声音频对应的人声音符标注结果,确定教师网络的损失函数值 514(交叉熵损失函数);根据教师网络的损失函数值 514(交叉熵损失函数),对教师网络 513 进行训练,得到训练后的教师网络 521。

步骤二:通过训练后的教师网络 521 对第二训练样本集(也可以称为数据集 2) 522 中的纯人声音频进行处理,得到纯人声音频对应的人声音符识别结果,作为人声音符第二识别结果(也可以称为纯人声音频对应的伪标签) 523;基于人声音符第二识别结果 523,确定纯人声音频对应的伪标签信息(也可以称为纯人声音频对应的伪标签纠正) 524。

步骤三:从第三训练样本集 511 中随机选择伴奏音频,作为目标伴奏音频;对至少一个纯人声音频 522 中的纯人声音频进行数据增强处理,得到处理后的纯人声音频;将目标伴奏音频与处理后的纯人声音频进行合成,得到纯人声音频对应的合成音频。

通过学生网络 525 对纯人声音频对应的合成音频进行处理,得到纯人声音频对应的人声音符学生识别结果,作为人声音符第三识别结果(也可以称为纯人声音频对应的预测) 526。

步骤四:根据纯人声音频对应的人声音符学生识别结果 526 和纯人声音频对应的伪标签信息 524,确定学生网络的损失函数值 527(交叉熵损失函数);根据学生网络的损失函数值 527(交叉熵损失函数),对学生网络 525 进行训练,得到训练后的学生网络 531。

推理:在训练后的学生网络 531 未满足停止训练条件的情况下,将训练后的学生网络 531 确定为训练后的教师网络,并再次从步骤 2 开始执行。即将步骤 2 中的训练后的教师网络 521 替换为训练后的学生网络 531,再次从步骤 2 开始执行。

在训练后的学生网络 531 满足停止训练条件的情况下,将训练后的学生网络 531 确定为人声音符识别模型。输入带伴奏的歌曲,人声音符识别模型对带伴奏的歌曲进行处理,可以得到带伴奏的歌曲对应的人声音符序列 533。

本申请实施例提供的技术方案,通过随机数据扩增的策略,在已有的训练样本的基础上,进一步扩大训练样本的数量来对人声音符识别模型进行训练,进一步提升了人声音符识别模型的鲁棒性。

请参考图 6,其示出了本申请一个实施例提供的人声音符识别方法的流程图。该方法可以包括如下步骤 610~640 中的至少一个步骤。

步骤 610,获取带伴奏的目标音频,目标音频中包含人声和伴奏。

在一些实施例中,目标音频中还包括噪音和混响。

在一些实施例中,对于带伴奏的目标音频的种类本申请不作限定。示例性地,目标音频可以是带伴奏的歌曲,也可以是现场歌曲录音。

步骤 620, 获取目标音频的音频特征, 音频特征包括目标音频在时频域上相关的特征。

在一些实施例中, 对目标音频进行时频变换, 得到目标音频的频域特征; 对频域特征进行滤波处理, 得到目标音频的音频特征。

对于对目标音频进行时频变换的具体方法, 本申请不作限定。示例性地, 可以采用 CWT-ESS (Continuous Wavelet Transform, 连续小波变换) 算法、STFT-ESS (Short-Time Fourier Transform, 短时傅里叶变换) 算法、OpenGAN 算法等。

对于对频域特征进行滤波处理的方法, 本申请不作限定。示例性地, 可以采用低通滤波、高通滤波、带通滤波、带阻滤波等。

步骤 630, 通过人声音符识别模型对音频特征进行处理, 得到目标音频的音符特征, 音符特征包括与目标音频的人声音符相关的特征。

人声音符识别模型是基于训练后的第一网络、纯人声音频和伴奏音频, 对第二网络进行训练得到的; 第一网络用于根据标注人声音频和伴奏音频的合成音频, 输出标注人声音频对应的人声音符识别结果; 第二网络用于根据纯人声音频和所述伴奏音频的合成音频, 输出纯人声音频对应的人声音符识别结果。

在一些实施例中, 对于目标音频包含的每个音频帧, 通过人声音符识别模型对音频帧的音频特征, 和音频帧的音频特征的上下文信息进行处理, 得到音频帧对应的第一中间特征; 根据音频帧对应的第一中间特征, 提取音频帧对应的第二中间特征; 根据音频帧对应的第二中间特征, 和音频帧对应的第二中间特征的上下文信息, 得到音频帧对应的音符特征; 其中, 目标音频的音符特征包括目标音频包含的各个音频帧分别对应的音符特征。

音频帧对应的第一中间特征包含音频帧对应的音频特征以及音频帧对应的音频特征的上下文信息。

音频帧对应的第二中间特征用于表征音频帧的音高特征。

音频帧对应的音符特征包含音频帧对应的第二中间特征以及音频帧对应的第二中间特征的上下文信息。

上下文信息是指目标音频帧与邻近音频帧之间的关联信息。邻近音频帧是指目标音频帧的相邻音频帧和/或相近音频帧。相邻音频帧是指与目标音频帧之间不包含其他音频帧的音频帧。相近音频帧是指在目标音频帧一定范围内的音频帧。例如目标音频帧的前后五帧音频帧可以称邻近音频帧。对于确定相近音频帧的范围, 本申请不作限定。

对于根据音频帧的音频特征, 和音频帧的音频特征的上下文信息, 得到音频帧对应的第一中间特征的方法, 本申请不作限定。示例性地, 可以采用递归神经网络实现。例如, 可以通过 LSTM(Long Short Term Memory Network, 长短时记忆网络)模型实现, 也可以通过 GRU (Gate Recurrent Unit, 门控循环单元) 模型实现。

对于根据音频帧对应的第一中间特征, 提取音频帧对应的第二中间特征的方法, 本申请不作限定。示例性地, 可以通过卷积神经网络实现。例如, 可以通过 CNN(Convolutional Neural Network, 卷积神经网络) 实现, 也可以通过残差卷积神经网络(ResNet)实现。

对于根据音频帧对应的第二中间特征, 和音频帧对应的第二中间特征的上下文信息, 得到音频帧对应的音符特征的方法, 本申请不作限定。示例性地, 可以采用递归神经网络实现。例如, 可以通过 LSTM(Long Short Term Memory Network, 长短时记忆网络)模型实现, 也可以通过 GRU (Gate Recurrent Unit, 门控循环单元) 模型实现。

步骤 640, 通过人声音符识别模型对音符特征进行处理, 得到目标音频的人声音符序列。

在一些实施例中, 通过人声音符识别模型对目标音频的音符特征进行分类处理, 得到目标音频的人声音符序列。

在一些实施例中, 根据目标音频的音符特征的音高, 对目标音频的音符特征进行分类处理, 得到目标音频的人声音符序列。

示例性地, 目标音频的人声音符序列为 MIDI 序列, 根据目标音频的音符特征的音高,

将目标音频的音符特征分类为不同的 MIDI 值，得到目标音频的 MIDI 序列。

在一些实施例中，人声音符识别模型包括：输入层、中间层和输出层。

输入层用于输入目标音频的音频特征。

中间层用于根据音频特征，提取目标音频的音符特征。

中间层包括第一中间特征提取层、第二中间特征提取层和音符特征提取层。

对于目标音频包含的每个音频帧，第一中间特征提取层用于根据音频帧的音频特征，和音频帧的音频特征的上下文信息，得到音频帧对应的第一中间特征。第二中间特征提取层用于根据音频帧对应的第一中间特征，提取音频帧对应的第二中间特征。音符特征提取层用于根据音频帧对应的第二中间特征，和音频帧对应的第二中间特征的上下文信息，得到音频帧对应的音符特征。

在一些实施例中，第一特征提取层为双向的 LSTM 模型，第二特征提取层为 CNN 模型，音符特征提取层为双向的 LSTM 模型。在一些实施例中，第二特征提取层可以根据实际需要设置一个或多个 CNN 网络构成 CNN 模型，本申请对此不作限定。例如，由 5 层 CNN 网络构成 CNN 模型。

输出层用于根据音符特征，得到目标音频的人声音符序列。

在一些实施例中，输出层为全连接层。在一些实施例中，输出层采用 Softmax 作为分类器。

示例性地，如图 7 所示，人声音符识别模型 700 包括输入层 710、中间层 720 和输出层 730。中间层 720 包含第一中间特征提取层 721、第二中间特征提取层 722 和音符特征提取层 730。

需要说明的是，上述人声音符识别方法实施例与上述人声音符识别模型的训练方法实施例属于相同构思，请参考上述人声音符识别模型的训练方法实施例，此处不再一一赘述。

本申请实施例提供的技术方案，通过人声音符识别模型，可以将带伴奏的目标音符的人声音符序列识别出来，无需调用人声伴奏分离算法，降低计算的复杂度，进而降低生产成本，同时准确率也不受人声伴奏分离算法的影响，保证了人声音符序列的准确性。

下述为本申请装置实施例，可以用于执行本申请方法实施例。对于本申请装置实施例中未披露的细节，请参照本申请方法实施例。

请参考图 8，其示出了本申请一个实施例提供的人声音符识别模型的训练装置的框图。该装置具有实现上述方法示例的功能，所述功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。该装置可以是上文介绍的终端设备，也可以设置在终端设备中。如图 8 所示，所述装置 800 可以包括：样本获取模块 810、第一网络训练模块 820、第二网络训练模块 830。

样本获取模块 810，用于获取至少一个标注人声音频、各个所述标注人声音频分别对应的人声音符标注结果、至少一个纯人声音频以及至少一个伴奏音频。

第一网络训练模块 820，用于基于所述标注人声音频、所述伴奏音频和所述标注人声音频对应的人声音符标注结果，对第一网络进行训练，得到训练后的第一网络；所述第一网络用于根据所述标注人声音频和所述伴奏音频的合成音频，输出所述标注人声音频对应的人声音符识别结果。

第二网络训练模块 830，用于基于所述训练后的第一网络、所述纯人声音频和所述伴奏音频，对第二网络进行训练，得到人声音符识别模型；所述第二网络用于根据所述纯人声音频和所述伴奏音频的合成音频，输出所述纯人声音频对应的人声音符识别结果。

在一些实施例中，如图 9 所示，所述第一网络训练模块 820，包括第一合成单元 821 和第一训练单元 822。

第一合成单元 821，用于采用所述伴奏音频与所述标注人声音频进行合成，得到所述标注人声音频对应的合成音频；

第一训练单元 822, 用于基于所述标注人声音频对应的合成音频以及所述标注人声音频对应的人声音符标注结果, 对所述第一网络进行训练, 得到所述训练后的第一网络。

在一些实施例中, 所述第一合成单元 821, 用于从所述至少一个伴奏音频中随机选择伴奏音频作为目标伴奏音频; 对所述标注人声音频进行数据增强处理, 得到处理后的标注人声音频; 其中, 所述数据增强处理包括以下至少之一: 添加混响、改变基频; 将所述目标伴奏音频与所述处理后的标注人声音频进行合成, 得到所述标注人声音频对应的合成音频。

在一些实施例中, 所述第一训练单元 822, 用于通过所述第一网络对所述标注人声音频对应的合成音频进行处理, 得到所述标注人声音频对应的人声音符识别结果, 作为人声音符第一识别结果; 根据所述人声音符第一识别结果和所述人声音符标注结果, 确定所述第一网络的损失函数值; 根据所述第一网络的损失函数值, 对所述第一网络的参数进行调整, 得到所述训练后的第一网络。

在一些实施例中, 如图 9 所示, 所述第二网络训练模块 830, 包括第一处理单元 831、确定单元 832、第二合成单元 833、第二处理单元 834 和第二训练单元 835。

第一处理单元 831, 用于通过所述训练后的第一网络对所述纯人声音频进行处理, 得到所述纯人声音频对应的人声音符识别结果, 作为人声音符第二识别结果。

确定单元 832, 用于将所述人声音符第二识别结果确定为所述纯人声音频对应的伪标签信息。

第二合成单元 833, 用于采用所述伴奏音频与所述纯人声音频进行合成, 得到所述纯人声音频对应的合成音频。

第二处理单元 834, 用于通过所述第二网络对所述纯人声音频对应的合成音频进行处理, 得到所述纯人声音频对应的人声音符识别结果, 作为人声音符第三识别结果。

第二训练单元 835, 用于根据所述人声音符第三识别结果和所述纯人声音频对应的伪标签信息, 对所述第二网络进行训练, 得到人声音符识别模型。

在一些实施例中, 所述确定单元 832, 用于提取所述纯人声音频的基频; 根据所述纯人声音频的基频, 对所述人声音符第二识别结果进行修正, 得到所述纯人声音频对应的伪标签信息。

在一些实施例中, 确定单元 832, 用于对于所述人声音符第二识别结果中包含的每一个音符, 计算所述音符与所述音符对应的发音位置的基频之间的音高差; 若所述音高差大于第一阈值, 则将所述音符的音高修正为所述音符对应的发音位置的基频的音高; 若所述音高差小于或等于所述第一阈值, 则保持所述音符的音高不变; 将音高调整后的所述人声音符第二识别结果, 确定为所述纯人声音频对应的伪标签信息。

在一些实施例中, 所述第二训练单元 835, 用于根据所述人声音符第三识别结果和所述伪标签信息, 确定所述第二网络的损失函数值; 根据所述第二网络的损失函数值, 对所述第二网络的参数进行调整, 得到所述人声音符识别模型。

在一些实施例中, 所述第二网络训练模块 830, 还用于在所述第二网络未满足停止训练条件的情况下, 将训练后的第二网络确定为所述训练后的第一网络, 并再次从所述基于所述训练后的第一网络、所述纯人声音频和所述伴奏音频, 对第二网络进行训练的步骤开始执行。

在一些实施例中, 所述样本获取模块 810, 用于获取至少一个无伴奏的清唱音频、各个所述清唱音频分别对应的人声音符标注结果, 以及至少一个带伴奏的歌曲音频; 根据所述清唱音频以及所述清唱音频对应的人声音符标注结果, 生成所述标注人声音频以及所述标注人声音频对应的人声音符标注结果; 对所述歌曲音频进行人声分离操作, 得到人声音频和伴奏音频; 根据所述人声音频, 生成所述纯人声音频。

在一些实施例中, 所述样本获取模块 810, 用于对所述清唱音频进行检测, 得到所述清唱音频中的静音部分和清音部分; 将所述清唱音频确定为所述标注人声音频; 从所述清唱音频对应的人声音符标注结果中, 删除所述静音部分对应的人声音符标注结果和所述清音部分

对应的人声音符标注结果,生成所述标注人声音频对应的人声音符标注结果。

在一些实施例中,所述样本获取模块 810,用于对所述人声音频进行检测,得到所述人声音频中的非人声部分;删除所述人声音频中的所述非人声部分,生成纯人声音频;对所述纯人声音频中的每一个音频帧,检测所述音频帧是否为人声音频帧,并计算所述音频帧的能量;若所述音频帧不是所述人声音频帧,且所述音频帧的能量小于第二阈值,则将所述音频帧确定为无效帧;若所述纯人声音频中的无效帧数量在所述纯人声音频包含的音频帧总数中的占比大于第三阈值,则将所述纯人声音频确定为无效纯人声音频;根据除所述无效纯人声音频之外的纯人声音频,生成所述纯人声音频。

本申请实施例提供的技术方案,通过上述训练方法得到的人声音符识别模型,能够直接从带伴奏的目标音频中识别出对应的人声音符序列,因而在模型使用阶段,无需调用人声伴奏分离算法从目标音频中提取出人声音频,降低了人声音符识别的计算复杂度。另外,本申请采用了半监督训练的方法,通过少量标注样本对第一网络进行训练,然后通过第一网络和大量未标注样本对第二网络进行训练,这样仅需要少量标注样本,即可训练出泛化性能强的模型,降低了训练样本的获取成本。

请参考图 10,其示出了本申请一个实施例提供的人声音符识别装置的框图。该装置具有实现上述方法示例的功能,所述功能可以通过硬件实现,也可以通过硬件执行相应的软件实现。该装置可以是上文介绍的终端设备,也可以设置在终端设备中。如图 10 所示,所述装置 1000 可以包括:音频获取模块 1010、特征获取模块 1020、特征提取模块 1030 和结果得到模块 1040。

音频获取模块 1010,用于获取带伴奏的目标音频,所述目标音频中包含人声和伴奏。

特征获取模块 1020,用于获取所述目标音频的音频特征,所述音频特征包括所述目标音频在时频域上相关的特征。

特征提取模块 1030,用于通过人声音符识别模型对所述音频特征进行处理,得到所述目标音频的音符特征,所述音符特征包括与所述目标音频的人声音符相关的特征。

结果得到模块 1040,用于通过所述人声音符识别模型对所述音符特征进行处理,得到所述目标音频的人声音符序列;其中,所述人声音符识别模型是基于训练后的第一网络、纯人声音频和伴奏音频,对第二网络进行训练得到的;所述第一网络用于根据标注人声音频和所述伴奏音频的合成音频,输出所述标注人声音频对应的人声音符识别结果;所述第二网络用于根据所述纯人声音频和所述伴奏音频的合成音频,输出所述纯人声音频对应的人声音符识别结果。

在一些实施例中,所述特征提取模块 1030,用于对于所述目标音频包含的每个音频帧,通过所述人声音符识别模型根据所述音频帧的音频特征,和所述音频帧的音频特征的上下文信息,得到所述音频帧对应的第一中间特征;根据所述音频帧对应的第一中间特征,提取所述音频帧对应的第二中间特征;根据所述音频帧对应的第二中间特征,和所述音频帧对应的第二中间特征的上下文信息,得到所述音频帧对应的音符特征;其中,所述目标音频的音符特征包括所述目标音频包含的各个音频帧分别对应的音符特征。

在一些实施例中,所述特征获取模块 1020,用于对所述目标音频进行时频变换,得到所述目标音频的频域特征;对所述频域特征进行滤波处理,得到所述目标音频的音频特征。

在一些实施例中,所述结果得到模块 1040,用于通过所述人声音符识别模型对所述目标音频的音符特征进行分类处理,得到所述目标音频的人声音符序列。

在一些实施例中,所述人声音符序列由人声音符识别模型得到,所述人声音符识别模型包括:输入层、中间层和输出层;所述输入层用于输入所述目标音频的音频特征;所述中间层用于根据所述音频特征,提取所述目标音频的音符特征;所述输出层用于根据所述音符特征,得到所述目标音频的人声音符序列。

本申请实施例提供的技术方案，通过人声音符识别模型，可以将带伴奏的目标音符的人声音符序列识别出来，无需调用人声伴奏分离算法，降低计算的复杂度，同时准确率也不受人声伴奏分离算法的影响，保证了人声音符序列的准确性。

需要说明的是，上述实施例提供的装置在实现其功能时，仅以上述各个功能模块的划分进行举例说明，实际应用中，可以根据实际需要而将上述功能分配由不同的功能模块完成，即将设备的内容结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。

关于上述实施例中的装置，其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述，此处将不做详细阐述说明。

请参考图 11，其示出了本申请一个实施例提供的计算机设备的结构示意图。该计算机设备可以是任何具备数据计算、处理和存储功能的电子设备。该计算机设备可用于实施上述实施例中提供的人声音符识别模型的训练方法，或者用于实施上述实施例中提供的人声音符识别方法。具体来讲：

该计算机设备 1100 包括中央处理单元（如 CPU（Central Processing Unit，中央处理器）、GPU（Graphics Processing Unit，图形处理器）和 FPGA（Field Programmable Gate Array，现场可编程逻辑门阵列）等）1101、包括 RAM（Random-Access Memory，随机存储器）1102 和 ROM（Read-Only Memory，只读存储器）1103 的系统存储器 1104，以及连接系统存储器 1104 和中央处理单元 1101 的系统总线 1105。该计算机设备 1100 还包括帮助服务器内的各个器件之间传输信息的基本输入/输出系统（Input Output System，I/O 系统）1106，和用于存储操作系统 1113、应用程序 1114 和其他程序模块 1111 的大容量存储设备 1107。

在一些实施例中，该基本输入/输出系统 1106 包括有用于显示信息的显示器 1108 和用于用户输入信息的诸如鼠标、键盘之类的输入设备 1109。其中，该显示器 1108 和输入设备 1109 都通过连接到系统总线 1105 的输入输出控制器 1110 连接到中央处理单元 1101。该基本输入/输出系统 1106 还可以包括输入输出控制器 1110 以用于接收和处理来自键盘、鼠标、或电子触控笔等多个其他设备的输入。类似地，输入输出控制器 1110 还提供输出到显示屏、打印机或其他类型的输出设备。

该大容量存储设备 1107 通过连接到系统总线 1105 的大容量存储控制器（未示出）连接到中央处理单元 1101。该大容量存储设备 1107 及其相关联的计算机可读介质为计算机设备 1100 提供非易失性存储。也就是说，该大容量存储设备 1107 可以包括诸如硬盘或者 CD-ROM（Compact Disc Read-Only Memory，只读光盘）驱动器之类的计算机可读介质（未示出）。

不失一般性，该计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括以用于存储诸如计算机可读指令、数据结构、程序模块或其他数据等信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括 RAM、ROM、EPROM（Erasable Programmable Read-Only Memory，可擦写可编程只读存储器）、EEPROM（Electrically Erasable Programmable Read-Only Memory，电可擦写可编程只读存储器）、闪存或其他固态存储技术，CD-ROM、DVD（Digital Video Disc，高密度数字视频光盘）或其他光学存储、磁带盒、磁带、磁盘存储或其他磁性存储设备。当然，本领域技术人员可知该计算机存储介质不局限于上述几种。上述的系统存储器 1104 和大容量存储设备 1107 可以统称为存储器。

根据本申请实施例，该计算机设备 1100 还可以通过诸如因特网等网络连接到网络上的远程计算机运行。也即计算机设备 1100 可以通过连接在该系统总线 1105 上的网络接口单元 1111 连接到网络 1112，或者说，也可以使用网络接口单元 1111 来连接到其他类型的网络或远程计算机系统（未示出）。

所述存储器中存储有计算机程序，所述计算机程序由所述处理器加载并执行以实现上述人声音符识别模型的训练方法，或者以实现上述人声音符识别方法。

在示例性实施例中，还提供了一种计算机可读存储介质，所述计算机可读存储介质中存储有计算机程序，所述计算机程序由处理器加载并执行以实现上述人声音符识别模型的训练方法，或者以实现上述人声音符识别方法。

可选地，该计算机可读存储介质可以包括：ROM（Read-Only Memory，只读存储器）、RAM（Random-Access Memory，随机存储器）、SSD（Solid State Drives，固态硬盘）或光盘等。其中，随机存取记忆体可以包括 ReRAM（Resistance Random Access Memory，电阻式随机存取记忆体）和 DRAM（Dynamic Random Access Memory，动态随机存取存储器）。

在示例性实施例中，还提供了一种计算机程序产品，所述计算机程序产品包括计算机程序，所述计算机程序存储在计算机可读存储介质中，处理器从所述计算机可读存储介质读取并执行所述计算机程序，以实现上述人声音符识别模型的训练方法，或者以实现上述人声音符识别方法。

在本申请实施例的描述中，术语“对应”可表示两者之间具有直接对应或间接对应的关系，也可以表示两者之间具有关联关系，也可以是指示与被指示、配置与被配置等关系。

在本文中提及的“多个”是指两个或两个以上。“和/或”，描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。字符“/”一般表示前后关联对象是一种“或”的关系。

另外，本文中描述的步骤编号，仅示例性示出了步骤间的一种可能的执行先后顺序，在一些其它实施例中，上述步骤也可以不按照编号顺序来执行，如两个不同编号的步骤同时执行，或者两个不同编号的步骤按照与图示相反的顺序执行，本申请实施例对此不作限定。

另外，本文中提供的实施例可以任意组合，以形成新的实施例，这都在本申请的保护范围之内。

本领域技术人员应该可以意识到，在上述一个或多个示例中，本申请实施例所描述的功能可以用硬件、软件、固件或它们的任意组合来实现。当使用软件实现时，可以将这些功能存储在计算机可读介质中或者作为计算机可读介质上的一个或多个指令或代码进行传输。计算机可读介质包括计算机存储介质和通信介质，其中通信介质包括便于从一个地方向另一个地方传送计算机程序的任何介质。存储介质可以是通用或专用计算机能够存取的任何可用介质。

以上所述仅为本申请的示例性实施例，并不用以限制本申请，凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

权利要求书

1.一种人声音符识别模型的训练方法，其特征在于，所述方法包括：

获取至少一个标注人声音频、各个所述标注人声音频分别对应的人声音符标注结果、至少一个纯人声音频以及至少一个伴奏音频；

基于所述标注人声音频、所述伴奏音频和所述标注人声音频对应的人声音符标注结果，对第一网络进行训练，得到训练后的第一网络；所述第一网络用于根据所述标注人声音频和所述伴奏音频的合成音频，输出所述标注人声音频对应的人声音符识别结果；

基于所述训练后的第一网络、所述纯人声音频和所述伴奏音频，对第二网络进行训练，得到人声音符识别模型；所述第二网络用于根据所述纯人声音频和所述伴奏音频的合成音频，输出所述纯人声音频对应的人声音符识别结果。

2.根据权利要求1所述的方法，其特征在于，所述基于所述标注人声音频、所述伴奏音频和所述标注人声音频对应的人声音符标注结果，对第一网络进行训练，得到训练后的第一网络，包括：

采用所述伴奏音频与所述标注人声音频进行合成，得到所述标注人声音频对应的合成音频；

基于所述标注人声音频对应的合成音频以及所述标注人声音频对应的人声音符标注结果，对所述第一网络进行训练，得到所述训练后的第一网络。

3.根据权利要求2所述的方法，其特征在于，所述采用所述伴奏音频与所述标注人声音频进行合成，得到所述标注人声音频对应的合成音频，包括：

从所述至少一个伴奏音频中随机选择伴奏音频作为目标伴奏音频；

对所述标注人声音频进行数据增强处理，得到处理后的标注人声音频；其中，所述数据增强处理包括以下至少之一：添加混响、改变基频；

将所述目标伴奏音频与所述处理后的标注人声音频进行合成，得到所述标注人声音频对应的合成音频。

4.根据权利要求2所述的方法，其特征在于，所述基于所述标注人声音频对应的合成音频以及所述标注人声音频对应的人声音符标注结果，对所述第一网络进行训练，得到所述训练后的第一网络，包括：

通过所述第一网络对所述标注人声音频对应的合成音频进行处理，得到所述标注人声音频对应的人声音符识别结果，作为人声音符第一识别结果；

根据所述人声音符第一识别结果和所述人声音符标注结果，确定所述第一网络的损失函数值；

根据所述第一网络的损失函数值，对所述第一网络的参数进行调整，得到所述训练后的第一网络。

5.根据权利要求1所述的方法，其特征在于，所述基于所述训练后的第一网络、所述纯人声音频和所述伴奏音频，对第二网络进行训练，得到人声音符识别模型，包括：

通过所述训练后的第一网络对所述纯人声音频进行处理，得到所述纯人声音频对应的人声音符识别结果，作为人声音符第二识别结果；

将所述人声音符第二识别结果确定为所述纯人声音频对应的伪标签信息；

采用所述伴奏音频与所述纯人声音频进行合成，得到所述纯人声音频对应的合成音频；

通过所述第二网络对所述纯人声音频对应的合成音频进行处理，得到所述纯人声音频对

应的人声音符识别结果，作为人声音符第三识别结果；

根据所述人声音符第三识别结果和所述伪标签信息，对所述第二网络进行训练，得到人声音符识别模型。

6.根据权利要求 5 所述的方法，其特征在于，所述将所述人声音符第二识别结果确定为所述纯人声音频对应的伪标签信息，包括：

提取所述纯人声音频的基频；

根据所述纯人声音频的基频，对所述人声音符第二识别结果进行修正，得到所述纯人声音频对应的伪标签信息。

7.根据权利要求 6 所述的方法，其特征在于，所述根据所述纯人声音频的基频，对所述人声音符第二识别结果进行修正，得到所述纯人声音频对应的伪标签信息，包括：

对于所述人声音符第二识别结果中包含的每一个音符，计算所述音符与所述音符对应的发音位置的基频之间的音高差；

若所述音高差大于第一阈值，则将所述音符的音高修正为所述音符对应的发音位置的基频的音高；

若所述音高差小于或等于所述第一阈值，则保持所述音符的音高不变；

将音高调整后的所述人声音符第二识别结果，确定为所述纯人声音频对应的伪标签信息。

8.根据权利要求 5 所述的方法，其特征在于，所述根据所述人声音符第三识别结果和所述伪标签信息，对所述第二网络进行训练，得到人声音符识别模型，包括：

根据所述人声音符第三识别结果和所述伪标签信息，确定所述第二网络的损失函数值；

根据所述第二网络的损失函数值，对所述第二网络的参数进行调整，得到所述人声音符识别模型。

9.根据权利要求 1 所述的方法，其特征在于，所述方法还包括：

在所述第二网络未满足停止训练条件的情况下，将训练后的第二网络确定为所述训练后的第一网络，并再次从所述基于所述训练后的第一网络、所述纯人声音频和所述伴奏音频，对第二网络进行训练的步骤开始执行。

10.根据权利要求 1 所述的方法，其特征在于，所述获取至少一个标注人声音频、各个所述标注人声音频分别对应的人声音符标注结果、至少一个纯人声音频以及至少一个伴奏音频，包括：

获取至少一个无伴奏的清唱音频、各个所述清唱音频分别对应的人声音符标注结果，以及至少一个带伴奏的歌曲音频；

根据所述清唱音频以及所述清唱音频对应的人声音符标注结果，生成所述标注人声音频以及所述标注人声音频对应的人声音符标注结果；

对所述歌曲音频进行人声分离操作，得到人声音频和所述伴奏音频；

根据所述人声音频，生成所述纯人声音频。

11.根据权利要求 10 所述的方法，其特征在于，所述根据所述清唱音频以及所述清唱音频对应的人声音符标注结果，生成所述标注人声音频以及所述标注人声音频对应的人声音符标注结果，包括：

对所述清唱音频进行检测，得到所述清唱音频中的静音部分和清音部分；

将所述清唱音频确定为所述标注人声音频；

从所述清唱音频对应的人声音符标注结果中，删除所述静音部分对应的人声音符标注结

果和所述清音部分对应的人声音符标注结果，生成所述标注人声音频对应的人声音符标注结果。

12.根据权利要求 10 所述的方法，其特征在于，所述根据所述人声音频，生成所述纯人声音频，包括：

对所述人声音频进行检测，得到所述人声音频中的非人声部分；

删除所述人声音频中的所述非人声部分，生成纯人声音频；

对所述纯人声音频中的每一个音频帧，检测所述音频帧是否为人声音频帧，并计算所述音频帧的能量；

若所述音频帧不是所述人声音频帧，且所述音频帧的能量小于第二阈值，则将所述音频帧确定为无效帧；

若所述纯人声音频中的无效帧数量在所述纯人声音频包含的音频帧总数中的占比大于第三阈值，则将所述纯人声音频确定为无效纯人声音频；

根据除所述无效纯人声音频之外的纯人声音频，生成所述纯人声音频。

13.一种人声音符识别方法，其特征在于，所述方法包括：

获取带伴奏的目标音频，所述目标音频中包含人声和伴奏；

获取所述目标音频的音频特征，所述音频特征包括所述目标音频在时频域上相关的特征；

通过人声音符识别模型对所述音频特征进行处理，得到所述目标音频的音符特征，所述音符特征包括与所述目标音频的人声音符相关的特征；

通过所述人声音符识别模型对所述音符特征进行处理，得到所述目标音频的人声音符序列；

其中，所述人声音符识别模型是基于训练后的第一网络、纯人声音频和伴奏音频，对第二网络进行训练得到的；所述第一网络用于根据标注人声音频和所述伴奏音频的合成音频，输出所述标注人声音频对应的人声音符识别结果；所述第二网络用于根据所述纯人声音频和所述伴奏音频的合成音频，输出所述纯人声音频对应的人声音符识别结果。

14.根据权利要求 13 所述的方法，其特征在于，所述通过所述人声音符识别模型根据所述音频特征，提取所述目标音频的音符特征，包括：

对于所述目标音频包含的每个音频帧，通过所述人声音符识别模型对所述音频帧的音频特征，和所述音频帧的音频特征的上下文信息进行处理，得到所述音频帧对应的第一中间特征；

根据所述音频帧对应的第一中间特征，提取所述音频帧对应的第二中间特征；

根据所述音频帧对应的第二中间特征，和所述音频帧对应的第二中间特征的上下文信息，得到所述音频帧对应的音符特征；

其中，所述目标音频的音符特征包括所述目标音频包含的各个音频帧分别对应的音符特征。

15.根据权利要求 13 所述的方法，其特征在于，所述获取所述目标音频的音频特征，包括：

对所述目标音频进行时频变换，得到所述目标音频的频域特征；

对所述频域特征进行滤波处理，得到所述目标音频的音频特征。

16.根据权利要求 13 所述的方法，其特征在于，所述通过所述人声音符识别模型根据所述音符特征，得到所述目标音频的人声音符序列，包括：

通过所述人声音符识别模型对所述目标音频的音符特征进行分类处理，得到所述目标音

频的人声音符序列。

17.根据权利要求 13 所述的方法，其特征在于，所述人声音符识别模型包括：输入层、中间层和输出层；

所述输入层用于输入所述目标音频的音频特征；

所述中间层用于根据所述音频特征，提取所述目标音频的音符特征；

所述输出层用于根据所述音符特征，得到所述目标音频的人声音符序列。

18.一种人声音符识别模型的训练装置，其特征在于，所述装置包括：

样本获取模块，用于获取至少一个标注人声音频、各个所述标注人声音频分别对应的人声音符标注结果、至少一个纯人声音频以及至少一个伴奏音频；

第一网络训练模块，用于基于所述标注人声音频、所述伴奏音频和所述标注人声音频对应的人声音符标注结果，对第一网络进行训练，得到训练后的第一网络；所述第一网络用于根据所述标注人声音频和所述伴奏音频的合成音频，输出所述标注人声音频对应的人声音符识别结果；

第二网络训练模块，用于基于所述训练后的第一网络、所述纯人声音频和所述伴奏音频，对第二网络进行训练，得到人声音符识别模型；所述第二网络用于根据所述纯人声音频和所述伴奏音频的合成音频，输出所述纯人声音频对应的人声音符识别结果。

19.一种人声音符识别装置，其特征在于，所述装置包括：

音频获取模块，用于获取带伴奏的目标音频，所述目标音频中包含人声和伴奏；

特征获取模块，用于获取所述目标音频的音频特征，所述音频特征包括所述目标音频在时频域上相关的特征；

特征提取模块，用于通过人声音符识别模型对所述音频特征进行处理，得到所述目标音频的音符特征，所述音符特征包括与所述目标音频的人声音符相关的特征；

结果得到模块，用于通过所述人声音符识别模型对所述音符特征进行处理，得到所述目标音频的人声音符序列；

其中，所述人声音符识别模型是基于训练后的第一网络、纯人声音频和伴奏音频，对第二网络进行训练得到的；所述第一网络用于根据标注人声音频和所述伴奏音频的合成音频，输出所述标注人声音频对应的人声音符识别结果；所述第二网络用于根据所述纯人声音频和所述伴奏音频的合成音频，输出所述纯人声音频对应的人声音符识别结果。

20.一种计算机设备，其特征在于，所述计算机设备包括处理器和存储器，所述存储器中存储有计算机程序，所述处理器执行所述计算机程序以实现如权利要求 1 至 12 任一项所述的方法，或者实现如权利要求 13 至 17 任一项所述的方法。

21.一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存储有计算机程序，所述计算机程序用于被处理器执行，以实现如权利要求 1 至 12 任一项所述的方法，或者实现如权利要求 13 至 17 任一项所述的方法。

22.一种计算机程序产品，其特征在于，所述计算机程序产品包括计算机程序，所述计算机程序存储在计算机可读存储介质中，处理器从所述计算机可读存储介质读取并执行所述计算机程序，以实现如权利要求 1 至 12 任一项所述的方法，或者实现如权利要求 13 至 17 任一项所述的方法。