

## 语音包推荐方法、装置、设备及存储介质

本申请要求在 2020 年 5 月 27 日提交中国专利局、申请号为 202010463433.6 的中国专利申请的优先权，该申请的全部内容通过引用结合在本申请中。

### 技术领域

本申请涉及数据处理技术领域，例如涉及智能搜索技术。

### 背景技术

在包含语音播报功能的应用程序中，通常设置有多种语音包供用户选择使用。为了提高语音包选取效率以及用户操作便捷度，通常会采用个性化推荐的方式，向用户推荐感兴趣的语音包。

然而，个性化推荐严重依赖于用户的历史行为数据。当应用程序开发前期没有注册用户或注册用户行为较少时，根本无法提供足够多的行为数据来针对用户进行个性化推荐。

### 发明内容

以下是对本文详细描述的主题的概述。本概述并非是为了限制权利要求的保护范围。

本申请实施例提供了一种冷启动情况下，实现语音包自动推荐的一种语音包推荐方法、装置、设备及存储介质。

根据本申请的一方面，提供了一种语音包推荐方法，包括：

根据视频推荐场景下的第一样本用户的第一用户行为数据和所述第一用户行为数据关联的第一视频数据，构建第一视频训练样本；

根据所述第一样本用户的样本搜索数据和对第一样本语音包的历史交互数据，构建用户训练样本；

根据所述第一视频训练样本和所述用户训练样本，对神经网络模型进行预训练；

采用第二样本语音包关联的样本视频和样本标注数据，对预训练好的神经网络模型进行再训练，得到语音包推荐模型。

根据本申请的另一方面，提供了另一种语音包推荐装置，包括：

第一视频训练样本构建模块，用于根据视频推荐场景下的第一样本用户的第一用户行为数据和所述第一用户行为数据关联的第一视频数据，构建第一视频训练样本；

用户训练样本构建模块，用于根据所述第一样本用户的样本搜索数据和对第一样本语音包的历史交互数据，构建用户训练样本；

预训练模块，用于根据所述第一视频训练样本和所述用户训练样本，对神经网络模型进行预训练；

再训练模块，用于采用第二样本语音包关联的样本视频和样本标注数据，对预训练好的神经网络模型进行再训练，得到语音包推荐模型。

根据本申请的又一方面，提供了又一种电子设备，包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行本申请任一实施例提供的一种语音包推荐方法。

根据本申请的在一方面，提供了再一种存储有计算机指令的非瞬态计算机可读存储介质，所述计算机指令用于使所述计算机执行本申请任一实施例提供的一种语音包推荐方法。

本申请实施例根据视频推荐场景下的第一样本用户的第一用户行为数据和第一用户行为数据关联的第一视频数据，构建第一视频训练样本；根据第一样本用户的样本搜索数据和对第一样本语音包的历史交互数据，构建用户训练样本；根据第一视频训练样本和用户训练样本，对神经网络模型进行预训练；采用第二样本语音包关联的样本视频和样本标注数据，对预训练好的神经网络模型进行再训练，得到语音包推荐模型。本申请实施例能够实现在冷启动情况下，进行神经网络模型的训练，从而使得神经网络模型具备了冷启动情况下的语音包自动推荐能力。

应当理解，本部分所描述的内容并非旨在标识本公开的实施例的关键

或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

在阅读并理解了附图和详细描述后，可以明白其他方面。

### 附图说明

附图设置为更好地理解本方案，不构成对本申请的限定。其中：

图 1 是本申请实施例提供的一种语音包推荐方法的流程图；

图 2 是本申请实施例提供的另一种语音包推荐方法的流程图；

图 3 是本申请实施例提供的另一种语音包推荐方法的流程图；

图 4A 是本申请实施例提供的另一种语音包推荐方法的流程图；

图 4B 是本申请实施例提供的一种视频向量表示网络的结构图；

图 4C 是本申请实施例提供的一种用户偏好向量表示网络的结构图；

图 4D 是本申请实施例提供的一种匹配模型的结构图；

图 4E 是本申请实施例提供的一种预训练后的匹配网络的结构图；

图 5 是本申请实施例提供的一种语音包推荐装置的结构图；

图 6 是用来实现本申请实施例的语音包推荐方法的电子设备的框图。

### 具体实施方式

以下结合附图对本申请的示范性实施例做出说明，其中包括本申请实施例的每一种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出每一种改变和修改，而不会背离本申请的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

本申请实施例提供的每一语音包推荐方法和语音包推荐装置，适用于对包含语音播报功能的应用程序，在冷启动情况下，进行语音包自动推荐能力训练的情况。本申请实施例所提供的每一语音包推荐方法，由语音包推荐装置执行，该装置采用软件、或硬件、或软件和硬件实现，并具体配置于电子设备中。

其中，冷启动可以理解为在应用程序中添加新项目或增加新用户，由于数据量不足导致无法向用户进行项目推荐的情况。例如，在本申请实施

例中，可以是在具备语音播报功能的应用程序中，添加语音包推荐功能，但是无大量历史数据可供使用的情况，致使无法使用历史数据对语音包推荐模型进行直接训练。

图1是本申请实施例提供的一种语音包推荐方法的流程图，该方法包括：

S101、根据视频推荐场景下的第一样本用户的第一用户行为数据和所述第一用户行为数据关联的第一视频数据，构建第一视频训练样本。

其中，视频推荐场景可以是具备视频推荐需求的任意场景，例如在短视频社交应用、视频播放应用和购物推广应用等至少一种应用中，向用户进行视频推荐的场景。

为了保证后续预训练模型与最终训练好的语音包推荐模型的匹配度，也即为了缩短再训练的训练周期，同时保证语音包推荐模型的模型精度，本申请实施例中的视频推荐场景，与本申请中语音包推荐场景的推荐机制相同或相似。

其中，第一用户行为数据，可以是第一样本用户对第一视频产生交互行为对应的数据。其中，交互行为可以是浏览、点赞、收藏、下载、评论和分享等行为中的至少一种。为了提高后续对神经网络模型进行预训练的训练结果的准确度，进而提高语音包推荐模型的模型精度，同时缩短语音包推荐模型再训练的训练周期，在本申请实施例的一种可选实施方式中，第一用户行为数据，可以包括用户浏览完成且点赞收藏行为的行为数据。示例性地，可以将第一样本用户浏览完成且点赞收藏行为的第一视频数据作为正样本；第一样本用户的其他浏览行为的第一视频数据作为负样本；基于每一正样本和负样本，构建第一视频训练样本。为了避免后续所训练模型出现过拟合现象，同时保证所训练模型的灵敏度和特异性，通常会将正负样本数量设置合理比例。其中，比例值可以由技术人员根据需要或经验值设定，或通过大量试验反复确定。

其中，第一视频数据可以是与第一视频所关联的数据，用于表征视频内容、视频特点和视频属性等中的至少一种。例如，第一视频数据可以包括第一视频的视频内容和描述文本等数据中的至少一种。

其中，第一视频可以是视频推荐场景中的任意视频，还可以是视频推

荐场景中，与样本语音包所关联的视频。

S102、根据所述第一样本用户的样本搜索数据和对第一样本语音包的历史交互数据，构建用户训练样本。

其中，样本搜索数据和历史交互数据，可以是在语音包搜索场景下第一样本用户的搜索行为所产生的数据。其中，语音包搜索场景可以是用户在具备语音播报功能的应用中进行语音包搜索，以及对所搜索的语音包产生交互行为的场景。其中，交互行为可以是浏览、点赞、收藏、下载、评论、分享和使用等行为中的至少一种。为了提高后续对神经网络模型进行预训练的训练结果的准确度，进而提高语音包推荐模型的模型精度，同时缩短语音包推荐模型再训练的训练周期，在本申请实施例的一种可选实施方式中，历史交互数据可以是语音包使用数据。其中，样本搜索数据，可以是第一样本用户进行语音包搜索时所使用的搜索词。

S103、根据所述第一视频训练样本和所述用户训练样本，对神经网络模型进行预训练。

示例性地，在对神经网络模型进行预训练时，可以将第一样本用户的第一视频训练样本中的第一视频数据、用户训练样本中的样本搜索数据和历史交互数据，输入至神经网络模型中，通过神经网络模型的输出结果拟合用户交互行为，并根据拟合的用户交互行为和第一用户行为数据之间的差异，优化调整神经网络模型的网络参数，从而实现对神经网络模型的预训练。

S104、采用第二样本语音包关联的样本视频和样本标注数据，对预训练好的神经网络模型进行再训练，得到语音包推荐模型。

其中，第二样本语音包关联的样本视频可以包括语音提供者的形象、声音和字幕等中的至少一种，用于表征语音包中语音提供者的形象特征以及语音特征。其中，形象特征包括萝莉、御姐、大叔、IP（Intellectual Property，知识产权）形象等中的至少一种。其中，语音特征包括音质特点、或语音风格、或音质特点和语音风格等。其中，音质特点包括男生、女生、甜美和沙哑等中的至少一种；语音风格包括播音腔和幽默等中的至少一种。

其中，样本标注数据可以是基于对第二样本语音包的感兴趣情况，对第二样本语音包人工标注的标签数据。

需要说明的是，第二样本语音包可以与前述第一样本语音包相同或不同。

示例性地，根据第二样本语音包关联的样本视频和样本标注数据，构建语音包训练样本；根据语音包训练样本，对预训练好的神经网络模型进行再训练，以对预训练好的神经网络模型的至少部分网络参数进行优化调整，得到语音包推荐模型。

例如，将第二样本语音包关联的样本视频的关联数据和第二样本语音包的语音包关联数据，输入至预训练好的神经网络模型，通过神经网络模型拟合用户对第二样本语音包的用户交互行为；根据所拟合的用户交互行为与样本标注数据之间的差异，优化调整神经网络模型中的网络参数，从而实现对神经网络模型的再训练。

其中，样本视频的关联数据可以是样本视频的视频数据，如视频内容和描述文本等数据中的至少一种；语音包关联数据可以是与第二样本语音包对应的搜索数据和第二样本语音包的语音包标识等数据中的至少一种。

可以理解的是，由于冷启动情况下样本语音包关联的样本视频和样本标注数据的数据量较少，而仅采用较少的训练样本对神经网络模型进行训练，将会使所训练模型出现欠拟合的情况，严重影响模型的泛化能力。为了避免上述情况的发生，本申请实施例通过将相似场景（如视频推荐场景，又如前述语音包搜索场景）下所构建的第一视频训练样本、或用户训练样本、或第一视频训练样本和用户训练样本，迁移至语音包推荐场景中，从而基于相似场景下的训练样本首先对神经网络模型进行预训练，避免出现模型欠拟合的情况。

然而，由于相似场景与本申请涉及的语音包推荐场景存在一定的差异，因而仅通过相似场景下所构建的训练样本进行神经网络模型的预训练，无法保证预训练好的神经网络模型，能够在语音包推荐场景下，同样具备较高的泛化能力，因此，需要通过语音包推荐场景下的标注样本对预训练好的神经网络模型进行再训练，从而实现对预训练好的神经网络模型中至少部分网络参数的调整。

可以理解的是，通常在神经网络中会设置全连接层，用于将模型学习到的分布式特征表示映射到样本标记空间，一定程度上表征网络的分类识

别能力。为了保证神经网络模型的模型精度，同时缩短模型的训练周期，在本申请实施例的一种可选实施方式中，在对预训练好的神经网络模型进行再训练时，可以将样本视频和样本标注数据输入至预训练好的神经网络模型，以对神经网络模型中的全连接层的网络参数进行调整，得到最终的语音包推荐模型。

本申请实施例根据视频推荐场景下的第一样本用户的第一用户行为数据和第一用户行为数据关联的第一视频数据，构建第一视频训练样本；根据第一样本用户的样本搜索数据和对第一样本语音包的历史交互数据，构建用户训练样本；根据第一视频训练样本和用户训练样本，对神经网络模型进行预训练；采用第二样本语音包关联的样本视频和样本标注数据，对预训练好的神经网络模型进行再训练，得到语音包推荐模型。本申请实施例通过将视频推荐场景下构建的第一视频训练样本，迁移至语音包推荐场景中，从而丰富了语音包推荐场景下进行模型训练时的训练样本，进而能够在训练样本样本量不足，也即冷启动情况下，能够实现对语音包推荐模型的训练，使得所得到的语音包推荐模型具备了冷启动情况下的语音包自动推荐能力，同时保证了语音包推荐模型在冷启动情况下的推荐结果的准确度。

在上述每一实施例的技术方案的基础上，在得到语音包推荐模型之后，还可以采用得到的语音包推荐模型，向用户进行语音包自动推荐。示例性地，将待推荐用户的每一候选展现视频、每一所述候选展现视频的描述文本、历史检索词和所用的历史语音包，输入至所述语音包推荐模型；根据所述语音包推荐模型的模型输出结果，向所述待推荐用户推荐包括目标语音包下载信息的目标展现视频。

其中，下载信息可以通过网址或者携带网址信息的二维码加以实现。其中，目标展现视频的数量为至少一个。

当目标展现视频存在至少两个时，可以通过滑动切换的方式，实现视频的顺序播放，从而使用户操作更加便捷。

可以理解的是，通过语音包推荐模型实现了在冷启动情况下，对待推荐用户进行语音包的自动推荐，实现了从用户找语音包到语音包主动找用户的转变。另外，借助与语音包关联的视频进行语音包推荐，无需用户频

繁进行语音包的试听，提高了用户获取语音包的便捷度以及语音包的获取效率。同时，通过视频展现的方式，能够让用户更加直观全面的获取到语音包的特点，且强化用户对语音包的印象，让用户更加轻松地获取到语音包特点信息，增强了用户的浏览体验和使用体验。并且，通过视频携带语音包下载信息，缩短了语音包下载步长。

在一实施例中，为了进一步增强视频交互性，还可以在目标展现视频中添加分享、点赞和评论外漏功能，从而缩短用户的视频交互或用户间交互的步长，提升用户参与度，同时提高了视频在用户间的传播效率。

需要说明的是，对神经网络模型进行预训练、再训练以及进行语音包推荐模型使用的电子设备可以相同或不同。

图2是本申请实施例提供的另一种语音包推荐方法的流程图，该方法在上述每一技术方案的基础上，进行了优化改进。

在一实施例中，在执行“根据所述第一视频训练样本和所述用户训练样本，对神经网络模型进行预训练”之前，追加“对预先构建的视频特征向量表示网络进行训练；根据训练好的视频特征向量表示网络，构建所述神经网络模型”，以完善神经网络模型的构建机制。

如图2所示的一种语音包推荐方法，包括：

S201、根据视频推荐场景下的第一样本用户的第一用户行为数据和所述第一用户行为数据关联的第一视频数据，构建第一视频训练样本。

S202、根据所述第一样本用户的样本搜索数据和对第一样本语音包的历史交互数据，构建用户训练样本。

S203、对预先构建的视频特征向量表示网络进行训练。

其中，视频特征向量表示网络，能够将所输入的视频数据转化为结构化表示的特征向量。

在本申请实施例的一种可选实施方式中，对预先构建的视频特征向量表示网络进行训练，可以是根据视频推荐场景下的第二样本用户的第二用户行为数据和第二用户行为数据关联的第二视频数据，构建第二视频训练样本；根据第二视频训练样本，对预先构建的视频特征向量表示网络进行训练。

其中，视频推荐场景与前述视频推荐场景相同。其中，第二样本用户

可以与第一样本用户相同或不同。

其中，第二用户行为数据，可以是第二样本用户对第二视频产生交互行为对应的数据。其中，交互行为可以是浏览、点赞、收藏、下载、评论和分享等行为中的至少一种。为了保证所训练的视频特征向量表示网络的网络性能，进而提高对基于视频特征向量表示网络所构建的神经网络模型，进行预训练时的训练效率和准确度，在本申请实施例的一种可选实施方式中，第二用户行为数据可以包括用户浏览完成或点赞收藏行为的行为数据。示例性地，可以将第二样本用户浏览完成且点赞收藏行为的第二视频数据作为正样本；第二样本用户的其他浏览行为的第二视频数据作为负样本。基于每一正样本和负样本，构建第二视频训练样本。为了避免所训练的视频特征向量表示网络出现过拟合现象，同时保证所训练模型的灵敏度和特异性，通常会将正负样本数量设置合理比例。其中，比例值可以由技术人员根据需要或经验值设定，或通过大量试验反复确定。

其中，第二视频数据可以是与第二视频所关联的数据，用于表征视频内容、视频特征和视频属性等数据中的至少一种。例如，第二视频数据可以包括第二视频的视频内容和描述文本等数据中的至少一种。

例如，将第二视频训练样本中的第二用户行为数据关联的第二视频数据输入至视频特征向量表示网络，得到视频特征向量；根据视频特征向量表示网络所关联的分类器，对视频特征向量进行处理，拟合第二样本用户对应的用户行为数据；根据拟合得到的用户行为数据与第二视频训练样本中的第二用户行为数据之间的差异，优化调整视频特征向量表示网络和所关联分类器中的网络参数，从而实现对视频特征向量表示网络的训练。

需要说明的是，通过第二用户行为数据和第二用户行为数据关联的第二视频数据构建第二视频训练样本，进行视频特征向量表示网络的训练，能够借助用户来刻画视频的特征向量，从而使得所训练的视频特征向量表示网络，能够提取出能够反映用户行为的特征数据，为后续语音包推荐模型的推荐结果的准确度，奠定了基础。

S204、根据训练好的视频特征向量表示网络，构建所述神经网络模型。

可以理解的是，根据训练好的视频特征向量表示模型，构建神经网络模型，从而使得所构建的神经网络模型中存在部分非随机初始化的网络参

数，减少了对神经网络模型的网络参数随机初始化的数量。

需要说明的是，本申请实施例对 S203 ~ S204 与 S201 ~ S202 的先后顺序不做任何限定，也即 S203 ~ S204 可以在 S201 ~ S202 之前执行，也可以在 S201 ~ S202 之后执行，还可以与 S201 ~ S202 同时执行或交叉执行。

S205、根据所述第一视频训练样本和所述用户训练样本，对神经网络模型进行预训练。

S206、采用第二样本语音包关联的样本视频和样本标注数据，对预训练好的神经网络模型进行再训练，得到语音包推荐模型。

需要说明的是，进行视频特征向量表示网络训练、神经网络模型预训练和神经网络模型再训练，所采用的电子设备可以相同或不同。

本申请实施例通过追加对预先构建的视频特征向量表示网络进行训练；根据训练好的视频特征向量表示网络，构建神经网络模型，从而通过前置训练的方式，对神经网络模型中的部分网络参数进行初始化，避免了随机初始化神经网络模型中的全部网络参数，造成对神经网络模型训练过程需要投入大量的时间成本，同时避免了所训练的神经网络模型出现过拟合现象，提高了模型的鲁棒性。

图 3 是本申请实施例提供的另一种语音包推荐方法的流程图，该方法在前述每一技术方案的基础上，进行了优化改进。

在一实施例中，在执行语音包推荐方法的过程中，追加“确定所述第二样本语音包的候选样本视频；根据每一所述候选样本视频的视频来源优先级，或者，根据每一所述候选样本视频与所述第二样本语音包的相似度，或者，根据每一所述候选样本视频的视频来源优先级和每一所述候选样本视频与所述第二样本语音包的相似度，确定与所述第二样本语音包关联的所述样本视频”，以完善语音包与样本视频关联关系的构建机制。

如图 3 所示的一种语音包推荐方法，包括：

S301、根据视频推荐场景下的第一样本用户的第一用户行为数据和所述第一用户行为数据关联的第一视频数据，构建第一视频训练样本。

S302、根据所述第一样本用户的样本搜索数据和对第一样本语音包的历史交互数据，构建用户训练样本。

S303、根据所述第一视频训练样本和所述用户训练样本，对神经网络

模型进行预训练。

S304、确定所述第二样本语音包的候选样本视频。

在本申请实施例的一种可选实施方式中，可以通过第二样本语音包的语音包提供方，直接进行视频录制的方式，生成第二样本语音包的候选样本视频。可以理解的是，由于语音包提供方更清楚自己的语音包的风格特点，因此会录制出更能凸显语音包特点的视频，从而使得候选样本视频与第二样本语音包更加贴合。

为了提高候选样本视频的生成效率，减少候选样本视频生成所投入的人力物力成本，在本申请实施例的另一可选实施方式中，还可以根据第二样本语音包的语音包提供者的推广图片，确定第二样本语音包的推广文本；基于所述语音包提供者的声学合成模型，根据所述推广文本生成推广音频和推广字幕；根据所述推广图片、所述推广音频和所述推广字幕，生成所述候选样本视频。

示例性地，可以根据推广图片中包含的语音包提供者，确定第二样本语音包的推广文本。例如，将语音包提供者的介绍信息作为推广文本。基于语音包提供者的声学合成模型，根据推广文本生成推广音频，以及生成与推广音频相对应的推广字幕。为了使推广音频和推广字幕更加具备语音包宣传功能，还可以在根据推广文本生成推广音频和推广字幕时，根据预先构建的宣传话术模板，进行推广字幕的生成，并基于语音包提供者的声学合成模型，合成与推广字幕相对应的推广音频，从而达到模拟语音包提供者的声音的目的，得到语音包提供者的声音播放的推广字幕。

其中，宣传话术模板可以由技术人员根据需要或宣传经验进行构建，例如在电子地图对应的语音包中，可以采用以下宣传话术模板“【人物简介】，欢迎大家使用我的语音包，【人物名称】伴您安全出行”。

可以理解的是，通过上述模板化制作的方式进行视频生成，无需进行视频的录制，提高了视频生成效率，同时减少了视频生成的人力物力成本。

为了提高候选样本视频的生成效率，减少候选样本视频生成所投入的人力物力成本，在本申请实施例的又一可选实施方式中，还可以根据第二样本语音包的语音包提供者信息构建视频搜索词；根据所述视频搜索词，搜索与所述语音包提供者的视频作为所述候选样本视频。

其中，语音包提供者信息包括语音包提供者的特点描述信息，例如声音特点，包括甜美、沙哑、贴心等；还可以包括播音风格，包括幽默、搞笑等。

可以理解的是，通过全网挖掘的方式，搜索与语音包提供者信息相关联的视频，无需进行视频的录制，提高了视频生成效率，同时减少了视频生成的人力物力成本。

S305、根据每一所述候选样本视频的视频来源优先级，或者，根据每一所述候选样本视频与所述第二样本语音包的相似度，或者，根据每一所述候选样本视频的视频来源优先级和每一所述候选样本视频与所述第二样本语音包的相似度，确定与所述第二样本语音包关联的所述样本视频。

在一实施例中，预先设置有不同视频来源对应的视频来源优先级，从而针对不同来源的候选样本视频，能够根据每一候选样本视频的视频来源优先级，确定与第二样本语音包关联的样本视频。其中，视频来源优先级能够表征样本语音包与样本视频的关联性，优先级越高，则关联性越大。可以理解的是，通过引入视频来源优先级，保证了样本语音包与样本视频之间的关联程度，为后续进行样本语音包关联的样本视频进行神经网络模型的再训练，奠定了基础。

示例性地，视频来源可以包括语音包提供者录制、模板化制作以及全网挖掘等中的至少一种。其中，视频来源优先级可以由技术人员根据需要或经验进行设定。当候选样本视频中存在视频来源变动时，相应的，还可以由技术人员根据需要或经验，在视频来源优先级中进行视频来源的编辑，并对每一视频来源的优先级顺序进行调整。其中，视频来源变动可以包括视频来源的添加或删除；相应的，对视频来源的编辑可以是添加视频来源或删除视频来源。

举例说明，当视频来源包括语音包提供者录制、模板化制作以及全网挖掘时，所设置的视频来源优先级由高到低，依次可以是“语音包提供者录制、模板化制作、以及全网挖掘”。

在一实施例中，针对相同或不同来源的候选样本视频，还可以确定每一候选样本视频与第二样本语音包的相似度，并根据相似度确定与第二样本语音包关联的样本视频。可以理解的是，通过引入相似度，辅助构建样

本语音包和样本视频的关联关系，从而保证了样本语音包与样本视频之间的关联程度，为后续进行样本语音包关联的样本视频进行神经网络模型的再训练，奠定了基础。

示例性地，可以通过神经网络的方式计算出第二样本语音包的声音和候选样本视频中的余弦相似度；对每一候选样本视频的余弦相似度进行排序；选择设定数量阈值、或满足设定数值条件、或选择设定数量阈值以及满足设定数值条件的候选样本视频作为第二样本语音包关联的样本视频。其中，设定数量阈值或设定数值条件可以由技术人员根据需要或经验值进行设定。

其中，神经网络的训练阶段，可有通过人工标注的方式构建训练语料，得到样本语音包和与样本语音包对应的正负样本视频；相应的，通过训练语料对神经网络进行训练，从而对神经网络中的网络参数进行优化调整。

在本申请实施例的一种可选实施方式中，可以将第二样本语音包和与第二样本语音包关联的样本视频，关联存储在电子设备本地、或与电子设备关联的其他存储设备中。为了提高存储效率可以通过键值对 (key-value) 的方式进行第二样本语音包和样本视频关联关系的存储。在一实施例中，可以采用正排拉链，以语音包标识为键域 (key)，以样本视频的视频数据为值域 (value) 进行存储；或者，在一实施例中，还可以采用倒排拉链，以样本视频的视频标识为 key，以语音包标识为 value 进行存储。

为了进一步保证第二样本语音包与样本视频之间的关联性，同时减少构建第二样本语音包与样本视频关联关系时的数据运算量，在一实施例中，还可以首先根据每一候选样本视频的视频来源优先级，对候选样本视频进行初步筛选；根据每一筛选后的候选样本视频与第二样本语音包的相似度，对筛选后的候选样本视频进行再次筛选，得到与第二样本语音包关联的样本视频。

需要说明的是，本申请实施例对 S304 ~ S305 与 S301 ~ S303 的先后顺序不做任何限定，也即 S304 ~ S305 可以在 S301 ~ S303 之前执行，也可以在 S301 ~ S303 之后执行，还可以与 S301 ~ S303 同时执行或交叉执行。

S306、采用第二样本语音包关联的样本视频和样本标注数据，对预训练好的神经网络模型进行再训练，得到语音包推荐模型。

本申请实施例通过将语音包推荐过程中，追加确定第二样本语音包的候选样本视频；并根据每一候选样本视频的视频来源优先级，或者，根据每一候选样本视频与第二样本语音包的相似度，或者，根据每一候选样本视频的视频来源优先级和每一候选样本视频与第二样本语音包的相似度，确定与第二样本语音包关联的样本视频。采用上述技术方案，完善了第二样本语音包与样本视频的关联关系的构建机制，为后续进行神经网络模型的再训练，提供了数据支撑。同时，通过视频来源优先级、或样本视频与第二样本语音包的相似度、或视频来源优先级和样本视频与第二样本语音包的相似度，对候选样本视频进行筛选，得到与第二样本语音包关联的样本视频，保证了第二样本语音包与样本视频之间的关联度，从而为所训练的语音包推荐模型的语音包推荐结果的准确度提供了保障。

图 4A 是本申请实施例提供的另一种语音包推荐方法的流程图，该方法在前述每一技术方案的基础上，提供了一种优选实施方式。

如图 4A 所示的一种语音包推荐方法，包括：语音包视频特征抽取和表示、用户偏好特征抽取和表示、语音包推荐模型训练及语音包推荐四个阶段。

## 1、语音包视频特征抽取和表示

### 1) 语音包视频生成

#### a、候选样本视频生成

语音包视频的来源主要有三类：专家生产、全网挖掘、模板化制作。具体如下：

专家生产：主要通过语音包提供方录制视频的方式生产候选样本视频。语音包提供方更清楚自己的语音包的特点（音色、风格等等），会录制出凸显出语音包特点的视频。以小 A 的语音包视频制作为例，这是一款年轻漂亮的女生、声音甜美并且很贴心的语音包时，通过在视频中甜美的装扮、外加一些很贴心的台词，将该语音包的特点表现的淋漓尽致。

全网挖掘：主要通过构造关键词的方式挖掘出视频。同样以小 A 的语音包视频制作为例，根据模板自动构造出“小 A 贴心视频”、“小 A 甜美视频”等搜索词，并通过搜索词在搜索引擎中进行搜索，获取大量候选样本视频。

模板化制作：主要通过将相关图片和台词（通过该语音包声音进行播报）融合的方式制作出视频。仍以小 A 的语音包视频制作为例，将小 A 的个人简介通过宣传话术模板生成推广字幕，譬如，“【简介】，欢迎大家使用我的语音包，【人称】伴你平安出行”等等。根据小 A 的声学合成模型，合成推广字幕对应的推广音频；根据推广字幕、推广音频和小 A 的个人照片，制作候选样本视频。

#### b、语音包与视频关联

通过以上方式会构造出大量的候选样本视频，需要根据候选样本视频与样本语音包的相关性进行排序，并根据排序结果，选取至少一个候选样本视频作为样本视频。具体方式如下：

##### a) 针对不同视频来源的视频选取

可以预先定义优先级规则，以定义不同来源视频的优先级。例如，优先级由高到低依次可以是：专家生产、模板化制作、以及全网挖掘。从而根据视频来源优先级，选取至少一个候选样本视频作为样本视频。

##### b) 针对相同来源的视频选取

主要通过第一神经网络的方式计算出语音包声音和视频的余弦相似度；对余弦相似度进行排序，并根据排序结果，选取至少一个候选样本视频作为样本视频。

#### 2) 视频向量表示网络训练

通过迁移学习的手段，首先基于相似场景数据（如视频推荐场景中的用户行为数据）训练产出模型，然后通过该模型对每个语音包视频的特征向量表示进行编码。之所以这样主要由于在冷启动时，语音包的视频无大量的用户行为数据，我们无法对其进行直接建模，基于此，我们选取了与该场景接近的视频推荐场景的用户行为数据进行视频向量表示网络的预训练。

参见图 4B 所示的一种视频向量表示网络的结构示意图。首先，利用视频推荐场景下的用户行为数据，对视频向量表示网络进行预训练，得到网络参数。在对视频向量表示网络进行训练的过程中，假设某种类型的视频的受众群体相对固定，那么通过用浏览完视频或点赞评论过的用户，来刻画这个视频的特征向量。

例如，在模型训练过程中，采用样本视频的视频内容和描述文本作为模型的训练输入；将浏览完视频或点赞评论过的视频的用户作为模型的输出标签。

其中，视频向量表示网络包括第一特征提取网络、第二特征提取网络、特征融合层、全连接层和分类器。其中，第一特征提取网络，设置为提取视频内容中的第一特征数据；第二特征提取网络，设置为提取描述文本中的第二特征数据；特征融合层，设置为对第二特征数据和第二特征数据进行融合处理，得到融合特征数据；全连接层，设置为将融合特征数据映射到样本标记空间；分类器，设置为根据融合特征数据的映射结果，进行样本分类，拟合用户行为。

在对视频向量表示网络训练完成后，去掉视频向量表示网络中的分类器，设置为通过处理后的视频向量表示网络，具备对视频的视频内容和描述文本得到视频特征向量表示的能力。

## 2、用户偏好特征抽取和表示

用户偏好的特征向量表示主要是基于用户的检索行为日志和用户历史选择过的语音包数据进行编码。比如某个用户搜索过“甜美的语音包”、“漂亮小姐姐”等，这个用户会更倾向于人美声甜的语音包；再比如某个用户使用过“小 A 的语音包”，这个用户就是喜欢“小 A”这种甜美风格的语音包。

参见图 4C 所示的用户偏好向量表示网络的结构示意图，该用户偏好向量表示网络包括语义表示层、嵌入层、特征融合层和全连接层。其中，语义表示层，设置为对检索行为日志中的检索词进行编码处理，得到第一编码向量；嵌入层，设置为对用户历史选择过的语音包的语音包标识，例如语音包表示进行编码处理，得到第二编码向量；特征融合层，设置为对第一编码向量和第二编码向量进行融合处理，得到融合编码向量；全连接层，设置为将融合编码向量进行映射，得到用户偏好特征向量。其中，语义表示层可以采用训练好的语音表示模型加以实现。

## 3、语音包推荐模型训练

### 1) 匹配模型构建

参见图 4D 所示的匹配模型的结构示意图可知，根据前述视频向量表

示网络和用户偏好向量表示网络，添加激活层，进行匹配模型构建。

其中，激活层，设置为对视频向量表示网络输出的视频特征向量和用户偏好向量表示网络输出的用户偏好特征向量进行激活处理，并根据激活处理结果拟合用户浏览行为。其中，激活层可以通过点乘计算的方式实现。

然而，由于视频向量表示网络和用户偏好向量表示网络各自训练得到，使得每一网络输出的视频特征向量和用户偏好特征向量之间不具备相关性，导致最终构建的匹配模型的匹配效果较差，也即基于当前的匹配模型进行语音包推荐时，所推荐语音包与用户的匹配程度较低。因此，需要对所构建的匹配网络进一步进行训练。

## 2) 匹配模型预训练

对图 4D 所示的匹配网络进行预训练。其中，训练数据通过视频推荐场景下产生的用户浏览的很多视频，将视频的视频内容、描述文本，以及这个用户的历史检索词和所使用语音包，作为训练样本，输入至匹配网络中，拟合用户的浏览行为，并根据所拟合行为与实际浏览行为标签值之间的差异，调整匹配模型参数，实现对匹配模型的预训练。其中，用户浏览完成且点赞收藏行为对应正样本标签值；用户的其他浏览行为对应负样本标签值。

## 3) 匹配模型再训练

通过迁移学习的手段进行匹配模型的预训练，使得匹配模型可能无法适配语音包推荐场景。因此，还需要在语音包推荐场景下进行少量样本标注，利用所标注样本对预训练后的匹配模型进行再训练，得到具备语音包推荐能力的语音包推荐模型。

参加图 4E 所示的预训练后的匹配网络的结构图，对匹配模型中的最上层，也即全连接层的网络参数进行微调，底层网络参数固定不变，得到能够应用于语音包视频个性化推荐场景的语音包推荐模型。

其中，将标注语音包关联的样本视频的视频内容和描述文本，以及标注语音包对应的检索词和语音包标识作为训练样本，输入至预训练好的匹配模型，拟合用户标注数据；根据用户标注数据好所标注的标签值之间的差异，对全连接层的网络参数进行微调。其中，用户标注数据可以是浏览或不浏览所对应的标签值。

#### 4、语音包推荐

将待推荐用户的每一候选展现视频、每一候选展现视频的描述文本、历史检索词和所用的历史语音包，输入至语音包推荐模型，得到每一候选展现视频对应的概率值；根据概率值对每一候选展现视频进行排序；选取设定数量、或满足设定概率阈值、或设定数量和满足设定概率阈值的至少一个候选展现视频作为目标展现视频；向待推荐用户推荐目标展现视频；其中，目标展现视频中添加有与目标展现视频对应的目标语音包的下载信息。

通过目标语音包关联的目标展现视频，向用户推荐目标语音包，让用户更加直观全面的获取到语音包的特点且印象深刻，大幅度提升用户的选择效率，并且视频形式的浏览体验更好，让用户能够更轻松地获取信息。

通过视频交互形式展现目标语音包，具体包含了三大方面：首先分享、点赞、评论功能外漏，使得交互方式更加便捷；动态生成该语音包下载的二维码图片，放到目标视频的右上角展现，缩短了用户分享下载步长，大幅度提升用户传播效率；支持滑动切换等便捷交互操作。

图5是本申请实施例提供的一种语音包推荐装置的结构图，该语音包推荐装置500，包括：第一视频训练样本构建模块501、用户训练样本构建模块502、预训练模块503和再训练模块504。其中，

第一视频训练样本构建模块501，设置为根据视频推荐场景下的第一样本用户的第一用户行为数据和所述第一用户行为数据关联的第一视频数据，构建第一视频训练样本；

用户训练样本构建模块502，设置为根据所述第一样本用户的样本搜索数据和对第一样本语音包的历史交互数据，构建用户训练样本；

预训练模块503，设置为根据所述第一视频训练样本和所述用户训练样本，对神经网络模型进行预训练；

再训练模块504，设置为采用第二样本语音包关联的样本视频和样本标注数据，对预训练好的神经网络模型进行再训练，得到语音包推荐模型。

本申请实施例通过第一视频训练样本构建模块根据视频推荐场景下的第一样本用户的第一用户行为数据和第一用户行为数据关联的第一视频数据，构建第一视频训练样本；通过用户训练样本构建模块根据第一样

本用户的样本搜索数据和对第一样本语音包的历史交互数据，构建用户训练样本；通过预训练模块根据第一视频训练样本和用户训练样本，对神经网络模型进行预训练；通过再训练模块采用第二样本语音包关联的样本视频和样本标注数据，对预训练好的神经网络模型进行再训练，得到语音包推荐模型。本申请实施例通过将视频推荐场景下构建的第一视频训练样本，迁移至语音包推荐场景中，从而丰富了语音包推荐场景下进行模型训练时的训练样本，进而能够在训练样本样本量不足，也即冷启动情况下，能够实现对语音包推荐模型的训练，使得所得到的语音包推荐模型具备了冷启动情况下的语音包自动推荐能力，同时保证了语音包推荐模型在冷启动情况下的推荐结果的准确度。

在一实施例中，该装置，还包括：

视频特征向量表示网络训练模块，设置为对预先构建的视频特征向量表示网络进行训练；

神经网络模型构建模块，设置为根据训练好的视频特征向量表示网络，构建所述神经网络模型。

在一实施例中，所述视频特征向量表示网络训练模块，包括：

第二视频训练样本构建单元，设置为根据所述视频推荐场景下的第二样本用户的第二用户行为数据和所述第二用户行为数据关联的第二视频数据，构建第二视频训练样本；

视频特征向量表示网络训练单元，设置为根据所述第二视频训练样本，对所述预先构建的所述视频特征向量表示网络进行训练。

在一实施例中，所述再训练模块 504，包括：

网络参数调整单元，设置为将所述样本视频和所述样本标注数据输入至预训练好的神经网络模型，以对所述神经网络模型中的全连接层的网络参数进行调整。

在一实施例中，该装置，还包括：

候选样本视频确定模块，设置为确定所述第二样本语音包的候选样本视频；

样本视频确定模块，设置为根据每一所述候选样本视频的视频来源优先级，确定与所述第二样本语音包关联的所述样本视频。

在一实施例中，该装置，还包括：

候选样本视频确定模块，设置为确定所述第二样本语音包的候选样本视频；

样本视频确定模块，设置为根据每一所述候选样本视频与所述第二样本语音包的相似度，确定与所述第二样本语音包关联的所述样本视频。

在一实施例中，所述候选样本视频确定模块，包括：

推广文本确定单元，设置为根据所述第二样本语音包的语音包提供者的推广图片，确定所述第二样本语音包的推广文本；

音频字幕确定单元，设置为基于所述语音包提供者的声学合成模型，根据所述推广文本生成推广音频和推广字幕；

候选样本视频生成单元，设置为根据所述推广图片、所述推广音频和所述推广字幕，生成所述候选样本视频。

在一实施例中，所述候选样本视频确定模块，包括：

视频搜索词构建单元，设置为根据所述第二样本语音包的语音包提供者信息，构建视频搜索词；

候选样本视频确定单元，设置为根据所述视频搜索词，搜索所述语音包提供者的视频作为所述候选样本视频。

在一实施例中，该装置，还包括：

数据输入模块，设置为将待推荐用户的每一候选展现视频、每一所述候选展现视频的描述文本、历史检索词和所用的历史语音包，输入至所述语音包推荐模型；

目标语音包推荐模块，设置为根据所述语音包推荐模型的模型输出结果，向所述待推荐用户推荐包括目标语音包下载信息的目标展现视频。

在一实施例中，所述第一用户行为数据包括用户浏览完成且点赞收藏行为的行为数据；所述第一视频数据包括所述第一用户行为数据关联的第一视频的视频内容和描述文本；所述历史交互数据为语音包使用数据。

上述语音包推荐装置可执行本申请任意实施例提供的语音包推荐方法，具备执行语音包推荐方法相应的功能模块和有益效果。

根据本申请的实施例，本申请还提供了一种电子设备和一种可读存储介质。

如图 6 所示,是实现本申请实施例的语音包推荐方法的电子设备的框图。电子设备旨在表示每一种形式的数字计算机,诸如,膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示每一种形式的移动装置,诸如,个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例,并且不意在限制本文中描述的或者要求的本申请的实现。

如图 6 所示,该电子设备包括:一个或多个处理器 601、存储器 602,以及设置为连接每一部件的接口,包括高速接口和低速接口。每一个部件利用不同的总线互相连接,并且可以被安装在公共主板上或者根据需要以其它方式安装。处理器可以对在电子设备内执行的指令进行处理,包括存储在存储器中或者存储器上以在外部输入/输出装置(诸如,耦合至接口的显示设备)上显示 GUI 的图形信息的指令。在其它实施方式中,若需要,可以将多个处理器、或多条总线、或多个服务器和多条总线,与多个存储器和多个存储器一起使用。同样,可以连接多个电子设备,每一个设备提供部分必要的操作(例如,作为服务器阵列、一组刀片式服务器、或者多处理器系统)。图 6 中以一个处理器 601 为例。

存储器 602 即为本申请所提供的非瞬时计算机可读存储介质。其中,所述存储器存储有可由至少一个处理器执行的指令,以使所述至少一个处理器执行本申请所提供的语音包推荐方法。本申请的非瞬时计算机可读存储介质存储计算机指令,该计算机指令设置为使计算机执行本申请所提供的语音包推荐方法。

存储器 602 作为一种非瞬时计算机可读存储介质,可设置为存储非瞬时软件程序、非瞬时计算机可执行程序以及模块,如本申请实施例中的语音包推荐方法对应的程序指令/模块(例如,附图 5 所示的第一视频训练样本构建模块 501、用户训练样本构建模块 502、预训练模块 503 和再训练模块 504)。处理器 601 通过运行存储在存储器 602 中的非瞬时软件程序、指令以及模块,从而执行服务器的每一种功能应用以及数据处理,即实现上述方法实施例中的语音包推荐方法。

存储器 602 可以包括存储程序区和存储数据区,其中,存储程序区可

存储操作系统、至少一个功能所需要的应用程序；存储数据区可存储实现语音包推荐方法的电子设备的使用所创建的数据等。此外，存储器 602 可以包括高速随机存取存储器，还可以包括非瞬时存储器，例如至少一个磁盘存储器件、闪存器件、或其他非瞬时固态存储器件。在一些实施例中，存储器 602 可选包括相对于处理器 601 远程设置的存储器，这些远程存储器可以通过网络连接至实现语音包推荐方法的电子设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

实现语音包推荐方法的电子设备还可以包括：输入装置 603 和输出装置 604。处理器 601、存储器 602、输入装置 603 和输出装置 604 可以通过总线或者其他方式连接，图 6 中以通过总线连接为例。

输入装置 603 可接收输入的数字或字符信息，以及产生与所述语音包推荐方法电子设备的用户设置以及功能控制有关的键信号输入，例如触摸屏、小键盘、鼠标、轨迹板、触摸板、指示杆、一个或者多个鼠标按钮、轨迹球、操纵杆等输入装置。输出装置 604 可以包括显示设备、辅助照明装置（例如，LED）和触觉反馈装置（例如，振动电机）等。该显示设备可以包括但不限于，液晶显示器（LCD）、发光二极管（LED）显示器和等离子体显示器。在一些实施方式中，显示设备可以是触摸屏。

此处描述的系统和技术中的每一种实施方式可以在数字电子电路系统、集成电路系统、专用 ASIC（专用集成电路）、计算机硬件、固件、软件、或它们的组合中实现。这些每一种实施方式可以包括：实施在一个或者多个计算机程序中，该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行、或解释、或执行和解释，该可编程处理器可以是专用或者通用可编程处理器，可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令，并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

这些计算程序（也称作程序、软件、软件应用、或者代码）包括可编程处理器的机器指令，并且可以利用高级过程或面向对象的编程语言、或汇编/机器语言来实施这些计算程序。如本文使用的，术语“机器可读介质”和“计算机可读介质”指的是设置为将机器指令或数据提供给可编程处理器的任何计算机程序产品、设备、或装置（例如，磁盘、光盘、存储器、

可编程逻辑装置（PLD）），包括，接收作为机器可读信号的机器指令的机器可读介质。术语“机器可读信号”指的是设置为将机器指令或数据提供给可编程处理器的任何信号。

为了提供与用户的交互，可以在计算机上实施此处描述的系统和技术，该计算机具有：设置为向用户显示信息的显示装置（例如，CRT（阴极射线管）或者 LCD（液晶显示器）监视器）；以及键盘和指向装置（例如，鼠标或者轨迹球），用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以设置为提供与用户的交互；例如，提供给用户的反馈可以是任何形式的传感反馈（例如，视觉反馈、听觉反馈、或者触觉反馈）；并且可以用任何形式（包括声输入、语音输入或者、触觉输入）来接收来自用户的输入。

可以将此处描述的系统和技术实施在包括后台部件的计算系统（例如，作为数据服务器）、或者包括中间件部件的计算系统（例如，应用服务器）、或者包括前端部件的计算系统（例如，具有图形用户界面或者网络浏览器的用户计算机，用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互）、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信（例如，通信网络）来将系统的部件相互连接。通信网络的示例包括：局域网（LAN）、广域网（WAN）和互联网。

计算机系统可以包括客户端和服务端。客户端和服务端一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务端关系的计算机程序来产生客户端和服务端的关系。

根据本申请实施例的技术方案，通过根据视频推荐场景下的第一样本用户的第一用户行为数据和第一用户行为数据关联的第一视频数据，构建第一视频训练样本；根据第一样本用户的样本搜索数据和对第一样本语音包的历史交互数据，构建用户训练样本；根据第一视频训练样本和用户训练样本，对神经网络模型进行预训练；采用第二样本语音包关联的样本视频和样本标注数据，对预训练好的神经网络模型进行再训练，得到语音包推荐模型。本申请实施例通过将视频推荐场景下构建的第一视频训练样本，迁移至语音包推荐场景中，从而丰富了语音包推荐场景下进行模型训练时

的训练样本，进而能够在训练样本样本量不足，也即冷启动情况下，能够实现语音包推荐模型的训练，使得所得到的语音包推荐模型具备了冷启动情况下的语音包自动推荐能力，同时保证了语音包推荐模型在冷启动情况下的推荐结果的准确度。

应该理解，可以使用上面所示的每一种形式的流程，重新排序、增加或删除步骤。例如，本申请中记载的每一步骤可以并行地执行也可以顺序地执行也可以不同的次序执行，只要能够实现本申请公开的技术方案所期望的结果，本文在此不进行限制。

上述具体实施方式，并不构成对本申请保护范围的限制。本领域技术人员应该明白的是，根据设计要求和因素，可以进行每一种修改、组合、子组合和替代。任何在本申请的精神和原则之内所作的修改、等同替换和改进等，均应包含在本申请保护范围之内。

## 权 利 要 求 书

### 1. 一种语音包推荐方法，包括：

根据视频推荐场景下的第一样本用户的第一用户行为数据和所述第一用户行为数据关联的第一视频数据，构建第一视频训练样本；

根据所述第一样本用户的样本搜索数据和对第一样本语音包的历史交互数据，构建用户训练样本；

根据所述第一视频训练样本和所述用户训练样本，对神经网络模型进行预训练；

采用第二样本语音包关联的样本视频和样本标注数据，对预训练好的神经网络模型进行再训练，得到语音包推荐模型。

### 2. 根据权利要求1所述的方法，还包括：

对预先构建的视频特征向量表示网络进行训练；

根据训练好的视频特征向量表示网络，构建所述神经网络模型。

### 3. 根据权利要求2所述的方法，其中，所述“对预先构建的视频特征向量表示网络进行训练”，包括：

根据所述视频推荐场景下的第二样本用户的第二用户行为数据和所述第二用户行为数据关联的第二视频数据，构建第二视频训练样本；

根据所述第二视频训练样本，对预先构建的视频特征向量表示网络进行训练。

### 4. 根据权利要求1所述的方法，其中，所述“采用第二样本语音包关联的样本视频和样本标注数据，对预训练好的神经网络模型进行再训练”，包括：

将所述样本视频和所述样本标注数据输入至预训练好的神经网络模型，以对所述神经网络模型中的全连接层的网络参数进行调整。

### 5. 根据权利要求1所述的方法，还包括：

确定所述第二样本语音包的候选样本视频；

根据每一所述候选样本视频的视频来源优先级，确定与所述第二样本语音包关联的所述样本视频。

### 6. 根据权利要求1所述的方法，还包括：

确定所述第二样本语音包的候选样本视频；

根据每一所述候选样本视频与所述第二样本语音包的相似度，确定与所述第二样本语音包关联的所述样本视频。

7. 根据权利要求5或6所述的方法，其中，所述“确定所述第二样本语音包的候选样本视频”，包括：

根据所述第二样本语音包的语音包提供者的推广图片，确定所述第二样本语音包的推广文本；

基于所述语音包提供者的声学合成模型，根据所述推广文本生成推广音频和推广字幕；

根据所述推广图片、所述推广音频和所述推广字幕，生成所述候选样本视频。

8. 根据权利要求5或6所述的方法，其中，所述“确定所述第二样本语音包的候选样本视频”，包括：

根据所述第二样本语音包的语音包提供者信息，构建视频搜索词；

根据所述视频搜索词，搜索所述语音包提供者的视频作为所述候选样本视频。

9. 根据权利要求1-6任一项所述的方法，还包括：

将待推荐用户的每一候选展现视频、每一所述候选展现视频的描述文本、历史检索词和所用的历史语音包，输入至所述语音包推荐模型；

根据所述语音包推荐模型的模型输出结果，向所述待推荐用户推荐包括目标语音包下载信息的目标展现视频。

10. 根据权利要求1-6任一项所述的方法，其中，所述第一用户行为数据包括用户浏览完成且点赞收藏行为的行为数据；所述第一视频数据包括所述第一用户行为数据关联的第一视频的视频内容和描述文本；所述历史交互数据为语音包使用数据。

11. 一种语音包推荐装置，包括：

第一视频训练样本构建模块，设置为根据视频推荐场景下的第一样本用户的第一用户行为数据和所述第一用户行为数据关联的第

一视频数据，构建第一视频训练样本；

用户训练样本构建模块，设置为根据所述第一样本用户的样本搜索数据和对第一样本语音包的历史交互数据，构建用户训练样本；

预训练模块，设置为根据所述第一视频训练样本和所述用户训练样本，对神经网络模型进行预训练；

再训练模块，设置为采用第二样本语音包关联的样本视频和样本标注数据，对预训练好的神经网络模型进行再训练，得到语音包推荐模型。

12. 根据权利要求 11 所述的装置，还包括：

视频特征向量表示网络训练模块，设置为对预先构建的视频特征向量表示网络进行训练；

神经网络模型构建模块，设置为根据训练好的视频特征向量表示网络，构建所述神经网络模型。

13. 根据权利要求 12 所述的装置，其中，所述视频特征向量表示网络训练模块，包括：

第二视频训练样本构建单元，设置为根据所述视频推荐场景下的第二样本用户的第二用户行为数据和所述第二用户行为数据关联的第二视频数据，构建第二视频训练样本；

视频特征向量表示网络训练单元，设置为根据所述第二视频训练样本，对预先构建的视频特征向量表示网络进行训练。

14. 根据权利要求 11 所述的装置，其中，所述再训练模块，包括：

网络参数调整单元，设置为将所述样本视频和所述样本标注数据输入至预训练好的神经网络模型，以对所述神经网络模型中的全连接层的网络参数进行调整。

15. 根据权利要求 11 所述的装置，还包括：

候选样本视频确定模块，设置为确定所述第二样本语音包的候选样本视频；

样本视频确定模块，设置为根据每一所述候选样本视频的视频来源优先级，确定与所述第二样本语音包关联的所述样本视频。

16. 根据权利要求 11 所述的装置，还包括：

候选样本视频确定模块，设置为确定所述第二样本语音包的候选样本视频；

样本视频确定模块，设置为根据每一所述候选样本视频与所述第二样本语音包的相似度，确定与所述第二样本语音包关联的所述样本视频。

17. 根据权利要求 15 或 16 所述的装置，其中，所述候选样本视频确定模块，包括：

推广文本确定单元，设置为根据所述第二样本语音包的语音包提供者的推广图片，确定所述样本语音包的推广文本；

音频字幕确定单元，设置为基于所述语音包提供者的声学合成模型，根据所述推广文本生成推广音频和推广字幕；

候选样本视频生成单元，设置为根据所述推广图片、所述推广音频和所述推广字幕，生成所述候选样本视频。

18. 根据权利要求 15 或 16 所述的装置，其中，所述候选样本视频确定模块，包括：

视频搜索词构建单元，设置为根据所述第二样本语音包的语音包提供者信息，构建视频搜索词；

候选样本视频确定单元，设置为根据所述视频搜索词，搜索所述语音包提供者的视频作为所述候选样本视频。

19. 根据权利要求 11-16 任一项所述的装置，还包括：

数据输入模块，设置为将待推荐用户的每一候选展现视频、每一所述候选展现视频的描述文本、历史检索词和所用的历史语音包，输入至所述语音包推荐模型；

目标语音包推荐模块，设置为根据所述语音包推荐模型的模型输出结果，向所述待推荐用户推荐包括目标语音包下载信息的目标展现视频。

20. 根据权利要求 11-16 任一项所述的装置，其中，所述第一用户行为数据包括用户浏览完成且点赞收藏行为的行为数据；所述第一视频数据包括所述第一用户行为数据关联的第一视频的视频内容

和描述文本；所述历史交互数据为语音包使用数据。

21. 一种电子设备，包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求 1-10 中任一项所述的一种语音包推荐方法。

22. 一种存储有计算机指令的非瞬时计算机可读存储介质，所述计算机指令设置为使所述计算机执行权利要求 1-10 中任一项所述的一种语音包推荐方法。