

多媒体处理方法、装置、电子设备及存储介质

本申请的交叉引用

5 本公开要求于 2021 年 7 月 15 日提交的、申请号为 202110802038.0、名称为“多媒体处理方法、装置、电子设备及存储介质”的中国专利申请的优先权，其全部内容通过引用并入本文。

技术领域

10 本公开实施例涉及计算机技术领域，尤其涉及一种多媒体处理方法、装置、电子设备及存储介质。

背景技术

15 目前，随着多媒体制作软件、多媒体制作平台的快速发展，大众进行多媒体资源制作的技术难度和门槛大大降低，使基于多媒体的内容创造和分享进入全民化阶段。在多媒体资源制作过程中，为了使多媒体资源具有更好的表现力，往往需要作者在录制多媒体资源后，对多媒体资源进行后期的裁剪，去掉例如口误、停顿等语音表达不流畅、不正确的音频、视频片段，从而提高多媒体资源的整体播放流畅性。

20 然而，现有技术中，对多媒体资源进行裁剪，以提高多媒体资源的整体播放流畅性的过程，只能通过专业的编辑工具进行人工处理，专业要求高、操作难度大，造成了多媒体资源的剪裁过程效率低、效果差，以及剪裁后的多媒体资源播放流畅性低等问题。

发明内容

本公开实施例提供一种多媒体处理方法、装置、电子设备及存储介质，以克服多媒体资源的剪裁过程效率低、效果差，以及剪裁后的多媒体资源播放流畅性低等问题。

25 第一方面，本公开实施例提供一种多媒体处理方法，包括：

30 获取第一多媒体资源；对所述第一多媒体资源的音频数据进行语音识别，确定所述第一多媒体资源对应的初始文本内容，其中，所述第一多媒体资源的音频数据包含所述初始文本内容的语音数据；确定所述初始文本内容中的无效文本内容，其中，所述无效文本内容为语义上无信息表达作用的文本内容；确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置；基于所述第一播放位置，对所述第一多媒体资源进行裁剪，得到第二多媒体资源，其中，所述第二多媒体资源的音频数据中包含目标文本内容的语音数据且不包含所述无效文本内容的语音数据，所述目标文本内容为所述初始文本内容中除所述无效文本内容外的其他文本内容。

35 第二方面，本公开实施例提供一种多媒体处理装置，包括：

语音识别模块，用于获取第一多媒体资源，并对所述第一多媒体资源的音频数据进行语音识别，确定所述第一多媒体资源对应的初始文本内容，其中，所述第一多媒体资源的音频数据包含所述初始文本内容的语音数据；

第一确认模块，用于确定所述初始文本内容中的无效文本内容，其中，所述无效文本内容为语义上无信息表达作用的文本内容；

第二确认模块，确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置；

- 5 生成模块，用于基于所述第一播放位置，对所述第一多媒体资源进行裁剪，得到第二多媒体资源，其中，所述第二多媒体资源的音频数据中包含目标文本内容的语音数据且不包含所述无效文本内容的语音数据，所述目标文本内容为所述初始文本内容中除所述无效文本内容外的其他文本内容。

- 10 第三方面，本公开实施例提供一种电子设备，包括：至少一个处理器和存储器；
所述存储器存储计算机执行指令；

所述至少一个处理器执行所述存储器存储的所述计算机执行指令，使得所述至少一个处理器执行如上第一方面以及第一方面各种可能的设计所述的多媒体处理方法。

- 15 第四方面，本公开实施例提供一种计算机可读存储介质，所述计算机可读存储介质中存储有计算机执行指令，当处理器执行所述计算机执行指令时，实现如上第一方面以及第一方面各种可能的设计所述的多媒体处理方法。

第五方面，本公开实施例提供一种计算机程序产品，包括计算机程序，该计算机程序被处理器执行时实现如上第一方面以及第一方面各种可能的设计所述的多媒体处理方法。

- 20 第六方面，本公开实施例提供一种计算机程序，所述计算机程序被处理器执行时实现如上第一方面以及第一方面各种可能的设计所述的多媒体处理方法。

- 25 本实施例提供的多媒体处理方法、装置、电子设备及存储介质，通过获取第一多媒体资源；对所述第一多媒体资源的音频数据进行语音识别，确定所述第一多媒体资源对应的初始文本内容，其中，所述第一多媒体资源的音频数据包含所述初始文本内容的语音数据；确定所述初始文本内容中的无效文本内容，其中，所述无效文本内容为语义上无信息表达作用的文本内容；确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置；基于所述第一播放位置，对所述第一多媒体资源进行裁剪，得到第二多媒体资源，其中，所述第二多媒体资源的音频数据中包含目标文本内容的语音数据且不包含所述无效文本内容的语音数据，所述目标文本内容为所述初始文本内容中除所述无效文本内容外的其他文本内容。基于第一多媒体资源的语音识别和文字处理，实现了对第一多媒体资源中的无效文本内容的自动定位和裁剪，提高了对多媒体资源中的无效内容的剪裁效率和剪裁效果，提高剪裁后的多媒体资源的播放流畅度。
- 30

附图说明

- 35 为了更清楚地说明本公开实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本公开的一些实施例，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为本公开实施例提供的多媒体处理方法的一种应用场景图；

- 图 2 为本公开实施例提供的多媒体处理方法的另一种应用场景图；
 图 3 为本公开实施例提供的多媒体处理方法的流程示意图一；
 图 4 为本公开实施例提供的一种通过第一多媒体资源得到第二多媒体资源的过程示意图；
 图 5 为本公开实施例提供的多媒体处理方法流程示意图二；
 5 图 6 为图 5 所示实施例中步骤 S203 的实现方式流程图；
 图 7 为图 5 所示实施例中步骤 S204 的实现方式流程图；
 图 8 为本公开实施例提供的一种交互界面示意图；
 图 9 为本公开实施例提供的另一种通过第一多媒体资源得到第二多媒体资源的过程示意图；
 10 图 10 为本公开实施例提供的多媒体处理装置的结构框图；
 图 11 为本公开实施例提供的一种电子设备的结构示意图；
 图 12 为本公开实施例提供的电子设备的硬件结构示意图。

具体实施方式

- 15 为使本公开实施例的目的、技术方案和优点更加清楚，下面将结合本公开实施例中的附图，对本公开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公开一部分实施例，而不是全部的实施例。基于本公开中的实施例，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，都属于本公开保护的范围。

- 20 下面对本公开实施例的应用场景进行解释：

- 图 1 为本公开实施例提供的多媒体处理方法的一种应用场景图，本实施例提供的多媒体处理方法可以应用于多媒体视频资源录制后的后期剪裁处理的应用场景下，其中，多媒体视频资源例如为包含音轨的视频，具体地，本实施例提供的方法可以应用于终端设备，如图 1 所示，用户通过终端设备 11 录制一段多媒体初始视频后，该终端设备 11
 25 通过执行本实施例提供的多媒体处理方法，对多媒体初始视频进行处理，自动去除该多媒体初始视频中出现的口误、停顿、口头语等无意义的内容对应的视频片段和音频片段，生成多媒体剪裁视频，由于多媒体剪裁视频中的无意义内容已被去除，因此相对于多媒体初始视频，多媒体剪裁视频在内容上更加连贯和顺畅，多媒体视频资源的内容品质更高。之后，终端设备 11 将该多媒体剪裁视频发送至视频平台服务器 12，视频平台服务器
 30 12 将该多媒体剪裁视频作为多媒体视频资源发布在视频平台，其他终端设备 13 可以通过与视频平台服务器 12 通信实现对该多媒体视频资源的观看。

- 图 2 为本公开实施例提供的多媒体处理方法的另一种应用场景图，在一种可能的实现方式中，本实施例提供的方法也可以应用于视频平台服务器，即用户通过终端设备录制一段多媒体初始视频后，将该多媒体初始视频发送至视频平台服务器，视频平台服务器
 35 通过执行本实施例提供的多媒体处理方法，对多媒体初始视频进行处理，生成如图 1 中所示的多媒体剪裁视频，并将该多媒体剪裁视频作为多媒体视频资源发布在视频平台，其他终端设备可以通过与视频平台服务器通信实现对该多媒体视频资源的观看。

现有技术中，对多媒体资源进行裁剪，以提高多媒体资源的整体播放流畅性的过程，只能通过专业的编辑工具进行人工处理，专业要求高、操作难度大，造成了多媒体资源的剪裁过程效率低、效果差，以及剪裁后的多媒体资源播放流畅性低等问题。这是由于，由用户自己录制的多媒体资源，由于不同用户的发音特点、语言表述风格之间的差异，以及多媒体资源的内容差异，难以实现多媒体资源中无意义内容的准确识别和定位，因此，通常只能依靠人工结合多媒体资源的画面和声音，综合判断后，进行人工裁剪，不仅效率低下，而且影响剪裁后的多媒体资源的播放流畅性。

本公开实施例提供一种多媒体处理方法以解决上述问题。

图 3 为本公开实施例提供的多媒体处理方法的流程示意图一。本实施例的方法可以应用在服务器或终端设备中，本实施例中，示例性地以终端设备作为本实施例的方法的执行主体进行说明，该多媒体处理方法包括：

S101：获取第一多媒体资源。

示例性地，多媒体通常指是多种媒体的综合，一般包括文本，声音和图像等多种媒体形式。在计算机系统中，多媒体指组合两种或两种以上媒体的一种人机交互式信息交流和传播媒体。本实施例中，第一多媒体资源可以是带有音轨的有声视频，更具体地，第一多媒体资源可以是一种保护视频数据和音频数据的多媒体资源或文件。

进一步地，第一多媒体资源可以是用户通过终端设备，例如智能手机的录像功能录制的有声视频，在本实施例中，该有声视频是包含有人声的视频，例如解说视频、教学视频、产品介绍视频等。当然，可以理解的是，该第一多媒体资源还可以是终端设备通过接收其他电子设备传输的数据而获得的，此处不再一一举例。

S102：对第一多媒体资源的音频数据进行语音识别，确定第一多媒体资源对应的初始文本内容，其中，第一多媒体资源的音频数据包含初始文本内容的语音数据。

示例性地，第一多媒体资源中至少包括音频数据，在获得第一多媒体资源后，根据预设的算法，对第一多媒体资源的音频数据进行语音识别，确定第一多媒体资源中音频数据所表达的内容对应的文本，即初始文本内容。其中，示例性地，音频数据中可以仅包括语音数据，或者，同时包括语音数据和非语音数据，语音数据为视频中录制的人的语音对应的音频数据；对应的，非语音数据为视频中录制的非人的语音对应的音频数据。更具体地，例如第一多媒体资源为一段产品介绍视频，其中，视频中介绍产品的人的语音所对应的音频数据，为语音数据。通过对第一多媒体资源的音频数据进行语音识别，所得到的文本即为初始文本内容，也即为视频中对产品进行介绍的人声语音转化的文字。

S103：确定初始文本内容中的无效文本内容，其中，无效文本内容为语义上无信息表达作用的文本内容。

进一步地，示例性地，获得初始文本内容后，基于初始文本内容中的字、词、句、段落，对初始文本内容进行识别，可以确定出其中的无效文本内容和目标文本内容。其中，无效文本内容为语义上无信息表达作用的文本内容，更具体地，例如说话过程中的非正常的停顿、重复、多余的语气词等无信息表达作用，但会影响语言表达的流畅性字、词、短句等。

示例性地，确定初始文本内容中的无效文本内容的实现方式有多种，例如，基于预设的无效文本内容库，确定初始文本内容中的无效文本内容，具体地，无效文本内容库

中包括无效文本内容对应的字、词、短句等预设要素，基于无效文本内容库中的字、词、短句，检测初始文本内容中是否含有上述无效文本内容库中的字、词、短句，进而确定初始文本内容中的无效文本内容。

5 在在另一种可能的实现方式中，可以通过对初始文本内容进行语义分析，获得初始文本内容对应的语义信息；再根据语义信息，确定初始文本内容中的无效文本内容。其中，具体地，通过对初始文本内容进行语义分析，可以确定初始文本内容中每一字、词元素的语义含义，其中，无效文本内容，也是语义含义的一个分类。因此通过对初始文本内容进行语义分析，可以确定初始文本内容中的无效文本内容。

10 其中，对初始文本内容进行语义分析，可以通过预训练的语言处理模型实现，用于语义分析的语言处理模型的使用和训练为本领域技术人员知晓的现有技术，此处不再进行追溯。

S104：确定无效文本内容的语音数据在第一多媒体资源中的第一播放位置。

15 示例性地，初始文本内容中的每一个字或词，对应有一段音频片段，以及一段音频数据，音频数据中包含有播放时间戳，在确定初始文本内容中的无效文本内容后，根据无效文本内容中所包含的字、词，确定各字、词对应的音频数据，并获取音频数据的播放时间戳，进而，根据音频数据的播放时间戳，可以确定无效文本内容的语音数据在第一多媒体资源中的第一播放位置。示例性地，在一种可能的实现方式中，无效文本内容在第一多媒体资源中对应是一段连续的音频数据，此种情况下，第一播放位置，可以仅包括一组包含起始点和终止点的播放时间戳。在另一种可能的实现方式中，无效文本内容20 在第一多媒体资源中对应是多段不连续的音频数据，此种情况下，第一播放位置，可以包括多组包含起始点和终止点的播放时间戳。

S105：基于第一播放位置，对第一多媒体资源进行裁剪，得到第二多媒体资源，其中，第二多媒体资源的音频数据中包含目标文本内容的语音数据且不包含无效文本内容的语音数据，目标文本内容为初始文本内容中除无效文本内容外的其他文本内容。

25 示例性地，在确定第一播放位置后，相当于已经对无效文本内容对应的语音数据进行了识别和定位，为了提高多媒体视频播放的整体流畅度，基于第一播放位置，将无效文本内容对应的语音数据从第一多媒体资源的全部音频数据中删除，保留目标文本内容对应的语音数据，从而减少由于停顿、重复、多余的语气词等无信息表达作用的字、词对多媒体视频内容表达的影响。具体地，基于第一播放位置对无效文本内容对应的语音数据30 在第一多媒体资源中的播放位置的描述，确定无效文本内容对应的语音数据的起始点和终止点，进而，将起始点和终止点中间的数据删除，将起始点之前和终止点之后的目标文本内容的语音数据进行拼接，生成剪裁后的音频数据。

进一步地，在本实施例中，第一多媒体资源还包括视频数据，因此，类似地，基于第一播放位置，对第一多媒体资源中的视频数据进行对应处理，剪裁无效文本内容对应的35 的视频数据，生成与剪裁后的音频数据在播放时间轴上对应的剪裁后的视频数据，进而，得到第二多媒体资源。

图 4 为本公开实施例提供的一种通过第一多媒体资源得到第二多媒体资源的过程示意图，如图 4 所示，第一多媒体资源中包括音频数据和视频数据，音频数据中包括语音数据，语音数据中包括第一语音数据和第二语音数据，其中，第一语音数据为无效文本

内容对应的语音数据；第二语音数据为目标文本内容对应的语音数据。在根据第一播放位置，确定语音数据中的第一语音数据和第二语音数据的位置后，对第一语音数据进行裁剪，将第二语音数据保留，并重新进行拼接，生成目标音频数据；相应的，根据第一播放位置，将对应的视频数据进行裁剪和拼接，生成目标视频数据。进而，根据目标音频数据和目标视频数据，生成第二多媒体资源。

- 5 在本实施例中，通过获取第一多媒体资源；对第一多媒体资源的音频数据进行语音识别，确定第一多媒体资源对应的初始文本内容，其中，第一多媒体资源的音频数据包含初始文本内容的语音数据；确定初始文本内容中的无效文本内容，其中，无效文本内容为语义上无信息表达作用的文本内容；确定无效文本内容的语音数据在第一多媒体资源中的第一播放位置；基于第一播放位置，对第一多媒体资源进行裁剪，得到第二多媒体资源，其中，第二多媒体资源的音频数据中包含目标文本内容的语音数据且不包含无效文本内容的语音数据，目标文本内容为初始文本内容中除无效文本内容外的其他文本内容。基于第一多媒体资源的语音识别和文字处理，实现了对第一多媒体资源中的无效文本内容的自动定位和裁剪，提高了对多媒体资源中的无效内容的剪裁效率和剪裁效果，
- 10 提高剪裁后的多媒体资源的播放流畅度。

图 5 为本公开实施例提供的多媒体处理方法流程示意图二。本实施例在图 3 所示实施例的基础上，对步骤 S102-S105 进一步细化，该多媒体处理方法包括：

S201：获取第一多媒体资源。

- 20 S202：通过语音端点检测算法，识别第一多媒体资源的音频数据中的语音数据和非语音数据。

- 示例性地，语音端点检测（Voice Activity Detection, VAD）算法，又称为语音活动检测，目的是从声音信号流里识别静音期（即非人声信号），本实施例中，通过语音端点检测算法，对第一多媒体资源的音频数据进行处理，可以识别出其中的人的语音所对应的语音数据和非人的语音对应的非语音数据，从而实现后续的基于语音数据的处理过程。
- 25 其中，语音端点检测算法的具体实现方法，为本领域技术人员知晓的现有技术，此处不再赘述。

S203：对第一多媒体资源的音频数据中的语音数据进行语音识别，确定第一多媒体资源对应的初始文本内容，其中，初始文本内容包括多个片段内容。

示例性地，如图 6 所示，S203 包括 S2031、S2032 两个具体的实现步骤，

- 30 S2031：根据自动语音识别技术，对第一多媒体资源的音频数据进行语音识别，获得多个语音字，以及各语音字对应的时间戳，时间戳表征语音字对应的音频数据在第一多媒体资源中的播放位置。

S2032：根据多个语音字，生成初始文本内容。

- 示例性地，自动语音识别技术（ASR, Automatic Speech Recognition），是一种将人的语音转换为文本的技术，是基于声学、语音学、语言学及计算机科学等多技术学科实现的技术，通过对音频信号进行预处理、特征提取、后处理、特征识别等多个环节，实现由语音向文字的转换，其中，各处理所涉及的方法也有多种，为本领域技术人员知晓的现有技术，可根据具体的需要对各处理环节中涉及的具体算法进行配置，此处不再一一举例说明。
- 35

进一步地，通过自动语音识别技术对第一多媒体资源的音频数据进行语音识别后，输出字级别的识别结果，即语音字，其中，每一个语音字对应一个段音频数据，即一个音频片段。根据第一多媒体资源的音频数据中各音频片段对应的播放时序，依次对各音频片段对应的语音字进行排列，所生成的包含多个语音字的文本，即为初始文本内容。

- 5 此外，示例性地，在通过自动语音识别技术识别语音字的过程中，还会确定每一语音字对应的时间戳，用以表征语音字对应的音频数据（即上述语音片段）在第一多媒体资源中的播放位置，该过程是通过预设的自动语音识别算法实现的，此处不再赘述。

其中，示例性地，每一个或多个语音字，可以构成一个片段内容，用于表征特定的语义。例如由两个语音字构成的词，或者由四个语音字构成的成语等。而对于初始文本内容中的无效文本内容，可以是以一个语音字的形式出现的，例如“嗯”，也可能是有两个语音字的形式出现的，例如“那个”，上述示例仅是针对中文的语音中所常出现的语义上无信息表达作用的内容，对于其他语言是类似的，此处不再一一举例。

S204：从初始文本内容的多个片段内容中，确定至少一个无效片段内容。

- 15 示例性地，如图7所示，S204包括S2041、S2042、S2043、S2044、S2045五个具体的实现步骤，

S2041：基于预设的无效文本内容库，确定初始文本内容中的无效片段内容。

S2042：若无效文本内容库中不存在无效片段内容，则对初始文本内容进行语义分析，获得初始文本内容的各片段内容对应的语义信息。

- 20 示例性地，无效文本内容库中预设有多个语义上无信息表达作用的语音字和或语音字的组合，根据无效文本内容库中预设的语义上无信息表达作用的语音字和或语音字的组合，对初始文本内容中进行一致性的检测，即可确定初始文本内容中的相同的语义上无信息表达作用的语音字和或语音字的组合，即无效片段内容。此时，即可实现步骤S204。其中，通过无效文本内容库确定无效片段内容，由于省略了语义分析的步骤，因此效率更高，计算资源占用更少，可以提高定位和剪裁无效文本内容的效率。

- 25 而若无效文本内容库中不存在无效片段内容，即根据无效文本内容库中预设的多个语义上无信息表达作用的语音字和或语音字的组合，对初始文本内容中进行一致性的检测后，在初始文本内容中未检测到无效文本内容库中预设的多个语义上无信息表达作用的语音字和或语音字的组合，或者，在初始文本内容中检测到的无效文本内容库中预设的多个语义上无信息表达作用的语音字和或语音字的组合的数量小于预设值，则进行对
30 初始文本内容进行语义分析，通过语义分析，确定初始文本内容的各片段内容对应的语义信息，进而在后续步骤中，通过语义信息确定无效片段内容。

S2043：根据初始文本内容的各片段内容对应的语义信息，确定初始文本内容中至少一个片段内容的可信度系数，可信度系数用于表征片段内容为无效文本内容的可信度。

- 35 S2044：根据片段内容的可信度系数以及预设的可信度阈值，从至少一个片段内容中确定至少一个无效片段内容。

示例性地，在基于预设的语义分析算法或模型对初始文本内容的各片段内容进行语义分析后，输出的语音信息中，包括该片段内容对应的语义类型的置信度，该置信度表征语义分析模型对片段内容进行语义分类结果的可信度评估，即置信度即为可信度系数。可信度系数越高，说明该片段内容对应的语义类型越可信。而“无效内容”也是对应一

个语义分类，在确定“无效内容”这一语义分类的过程中，通过片段内容的可信度系数以及预设的可信度阈值，将大于可信度阈值的可信度系数对应的片段内容，确定为该“无效内容”的语义分类，即无效片段内容。

5 本实施例步骤中，通过各片段内容的可信度系数确定无效片段内容，可以提高无效片段内容的识别准确性，减少误判。

S2045：将基于语义信息确定的无效片段内容，添加至无效文本内容库。

10 示例性地，在通过语音信息确定无效片段内容后，由于无效文本库中不包括上述无效片段内容对应的字或字的组合，因此将该无效片段内容添加至无效文本内容库，扩充无效文本内容库的内容，提高后续使用无效文本内容库进行无效文本内容判断的精度有效性。进而提高定位和剪裁无效文本内容的效率。

S205：根据至少一个无效片段内容，确定初始文本内容中的无效文本内容。

示例性地，无效文本内容中可以包括一个或多个无效片段内容，在确定初始文本内容中的各无效片段内容后，即可确定对应的无效文本内容。

15 在一种可能的实现方式中，无效片段内容包括至少一个语音字，确定初始文本内容中的无效文本内容，包括：根据语音字对应的时间戳，获取各语音字的播放时长；根据预设的标准时长，以及语音字的播放时长，将播放时长大于标准时长第一时长阈值的语音字，或者，播放时长小于标准时长第二时长阈值的语音字，确定为初始文本内容中的无效文本内容。

20 示例性地，无效片段内容中的语音字，是通过对语音数据转换而生成的，而语音数据对应人的语音，但在实际的应用过程中，对于同一个语音字，人的语音的发音时长可能存在差异性，即人的语音在发音时长维度上存在差异性。而不同的发音时长，可以表达不同的语义，进而导致了语音字是否能够表达语义信息。例如，用户发出“嗯”的语音两秒，此语境下该语音字“嗯”为一个语气停顿词，并没有具体含义，应属于无效片段内容；而在用户对话的语境下：“你吃早饭了吗”，“嗯，吃了”，此时时长 0.2 秒的语音“嗯”表征了具体的语义，因此不属于无效片段内容。

25 进而，基于以上举例说明，各语音字具有预设的标准时长，例如均为 0.2 秒。当根据语音字对应的时间戳，确定的语音字的播放时远远大于该标准时长或远小于该标准时长时，说明该语音字很可能为语气停顿词，不表达具体语义，因此可以将其确定为无效文本内容。

30 本实施例中，通过语音字的标准时长，和语音字的播放时长进行对比，检测无效片段内容中的语音字是否为由于语气停顿等原因生成的无效字，进而减少对此类在不同语境下具有不同含义的语音词的误判，提高无效文本内容的识别精确度。

S206：确定无效文本内容中的各无效片段内容的语音数据在第一多媒体资源的音频数据中的起始点和终止点。

35 S207：根据无效文本内容中的各无效片段内容对应的起始点和终止点，确定无效文本内容的语音数据在第一多媒体资源中的第一播放位置。

示例性地，无效片段内容包括至少一个语音字，根据无效文本内容中各语音字对应的时间戳，确定无效文本内容的语音数据在第一多媒体资源中的第一播放位置。

S208：显示初始文本内容中的无效文本内容。

S209: 响应于针对无效文本内容的操作指令, 播放无效文本内容对应的音频片段。

示例性地, 本实施例提供的方法所应用的终端设备, 具有可触摸显示屏, 终端设备内运行有用于对第一多媒体资源进行编辑的应用 (Application, APP), 该终端设备通过可触摸显示屏显示该 APP 的交互界面。图 8 为本公开实施例提供的一种交互界面示意图, 参考图 8, 在终端设备的交互界面上, 显示有初始文本内容 (图 8 中示例性的显示了其中一部分 “下面为您介绍的呢是一款本月 16 日那个最新发布的机械键盘”), 并以高亮的方式 (图 8 中以字体加黑表示), 在初始文本内容中显示无效文本内容 (即 “嗯嗯” 和 “那个”)。用户通过终端设备的可触摸显示屏, 触摸无效文本内容对应的显示屏区域时, 终端设备播放无效文本内容对应的音频和或视频片段, 以使用户确定该无效文本内容对应的音频片段是否确实为无信息表达作用的停顿、重复等无意义内容。

S210: 根据非语音数据的起始点和终止点, 确定非语音数据在第一多媒体资源中的第二播放位置。

S211: 基于第一播放位置和第二播放位置, 获取初始文本内容中除无效片段内容外的其他文本内容, 其中, 其他文本内容包括至少一个目标片段内容。

非语音数据是第一多媒体资源中, 非语音部分所对应的音频数据, 例如产品介绍的视频中, 未开始介绍前的空白部分, 以及介绍完毕后的空白部分。非语音数据在步骤 S202 中通过语音端点检测算法得到, 此处不再赘述。类似地, 在确定非语音数据后, 对应的可获得非语音数据在第一多媒体资源中的播放位置, 即第二播放位置。

进一步地, 根据第二播放位置, 可以实现对非语音数据的定位, 基于第一播放位置和第二播放位置, 对第一多媒体资源进行剪裁, 将第一多媒体资源中的无效片段内容对应的语音数据, 以及非语音数据去除, 留下其他文本内容对应的音频数据, 即目标片段内容对应的音频数据。

S212: 在至少一个目标片段内容对应的语音数据的起始点添加淡入效果, 和或, 在至少一个目标片段内容对应的语音数据的结束点添加淡出效果, 生成与目标片段内容对应的过渡语音数据。

S213: 按照第一播放位置和第二播放位置拼接过渡语音数据, 生成第二多媒体资源。

进一步地, 留下其他文本内容中包括至少一个目标片段内容, 为了提高剪裁后的音频播放流畅度, 对目标片段内容对应的语音数据添加淡入淡出效果。具体地, 例如, 在至少一个目标片段内容对应的语音数据的起始点添加淡入效果, 在至少一个目标片段内容对应的语音数据的添加淡出效果。其中, 淡入效果和淡出效果, 是指在语音数据的起始点和结束点, 进行时域加窗, 使一个目标片段内容对应的语音数据开始播放是, 音量由小逐渐变大 (淡入) 或者由大逐渐变小 (淡出), 减少音频数据剪裁后的突兀感。具体对一个音频片段添加淡入淡出的方法, 为本领域现有技术, 此处不再赘述。

进一步地, 对目标片段内容对应的语音数据进行淡入淡出处理后, 生成对应的过渡语音数据, 之后, 根据第一播放位置和第二播放位置, 对各过渡语音数据进行拼接, 生成目标音频数据, 类似的, 根据第一播放位置和第二播放位置, 获取目标音频数据对应的目标视频数据, 进而生成第二多媒体资源。该第二多媒体资源, 仅包括由目标片段内容构成的目标文本内容对应的多媒体视频, 而不包含无效文本内容对应的多媒体视频,

以及，非语音数据对应的多媒体视频。由于第二多媒体资源去除了口误、停顿、口头语等无信息表达作用的内容，因此能够提高第二多媒体资源的内容播放流畅度。

图 9 为本公开实施例提供的另一种通过第一多媒体资源得到第二多媒体资源的过程示意图，如图 9 所示，通过对第一多媒体资源的音频数据进行语音端点检测后，确定其中的语音数据和非语音数据，之后，对语音数据进行语音识别、语义分析等处理步骤后，确定多个无效片段内容，其中，无效片段内容对应第一语音数据，除无效片段内容之外的目标片段内容对应第二语音数据，根据无效片段内容对应的第一语音数据对应的第一播放位置和非语音数据对应的第二播放位置，进行音频数据剪裁，去除第一语音数据以及非语音数据，生成目标文本内容对应的目标音频数据。类似地，根据第一播放位置和第二播放位置对视频数据进行剪裁处理，生成目标视频数据。再将目标音频数据和目标视频数据进行合并，生成第二多媒体资源。

本实施例中，步骤 S201 的实现方式与本公开图 3 所示实施例中的步骤 S101 的实现方式相同，在此不再一一赘述。

对应于上文实施例的多媒体处理方法，图 10 为本公开实施例提供的多媒体处理装置的结构框图。为了便于说明，仅示出了与本公开实施例相关的部分。参照图 10，多媒体处理装置 3 包括：

语音识别模块 31，用于获取第一多媒体资源，并对第一多媒体资源的音频数据进行语音识别，确定第一多媒体资源对应的初始文本内容，其中，第一多媒体资源的音频数据包含初始文本内容的语音数据；

第一确认模块 32，用于确定初始文本内容中的无效文本内容，其中，无效文本内容为语义上无信息表达作用的文本内容；

第二确认模块 33，确定无效文本内容的语音数据在第一多媒体资源中的第一播放位置；

生成模块 34，用于基于第一播放位置，对第一多媒体资源进行裁剪，得到第二多媒体资源，其中，第二多媒体资源的音频数据中包含目标文本内容的语音数据且不包含无效文本内容的语音数据，目标文本内容为初始文本内容中除无效文本内容外的其他文本内容。

在本公开的一个实施例中，第一确认模块 32，具体用于：对初始文本内容进行语义分析，获得初始文本内容对应的语义信息；根据语义信息，确定初始文本内容中的无效文本内容。

在本公开的一个实施例中，初始文本内容包括多个片段内容，第一确认模块 32 在根据语义信息，确定初始文本内容中的无效文本内容时，具体用于：根据初始文本内容对应的语义信息，确定初始文本内容中至少一个片段内容的可信度系数，可信度系数用于表征片段内容为无效文本内容的可信度；根据片段内容的可信度系数以及预设的可信度阈值，从至少一个片段内容中确定至少一个无效片段内容；根据至少一个无效片段内容，确定初始文本内容中的无效文本内容。

在本公开的一个实施例中，第二确认模块 33，具体用于：确定各无效片段内容的语音数据在第一多媒体资源的音频数据中的起始点和终止点；根据各无效片段内容对应的起始点和终止点，确定无效文本内容的语音数据在第一多媒体资源中的第一播放位置。

在本公开的一个实施例中，生成模块 34，具体用于：基于第一播放位置，获取初始文本内容中除无效片段内容外的其他文本内容，其中，其他文本内容包括至少一个目标片段内容；在至少一个目标片段内容对应的语音数据的起始点添加淡入效果，和/或，在至少一个目标片段内容对应的语音数据的结束点添加淡出效果，生成与目标片段内容对应的过渡语音数据；按照第一播放位置拼接过渡语音数据，生成第二多媒体资源。

在本公开的一个实施例中，在对初始文本内容进行语义分析，获得初始文本内容对应的语义信息之前，第一确认模块 32 在，还用于：基于预设的无效文本内容库，确定初始文本内容中的无效文本内容；第一确认模块 32 在对初始文本内容进行语义分析，获得初始文本内容对应的语义信息时，具体用于：若无效文本内容库中不存在无效文本内容，则对初始文本内容进行语义分析，获得初始文本内容对应的语义信息；第一确认模块 32 在根据语义信息，确定初始文本内容中的无效文本内容之后，还用于：将基于语义信息确定的无效文本内容，添加至无效文本内容库。

在本公开的一个实施例中，在基于第一播放位置，对第一多媒体资源进行裁剪，得到第二多媒体资源之前，生成模块 34，还用于：显示初始文本内容中的无效文本内容；响应于针对无效文本内容的操作指令，播放无效文本内容对应的音频片段。

在本公开的一个实施例中，语音识别模块 31 对第一多媒体资源的音频数据进行语音识别，确定第一多媒体资源对应的初始文本内容时，具体用于：通过语音端点检测 VAD 算法，识别第一多媒体资源的音频数据中的语音数据和非语音数据；对第一多媒体资源的音频数据中的语音数据进行语音识别，确定第一多媒体资源对应的初始文本内容。

在本公开的一个实施例中，第二确认模块 33，还用于：根据非语音数据的起始点和终止点，确定非语音数据在第一多媒体资源中的第二播放位置；生成模块 34，具体用于：基于第一播放位置和第二播放位置，对第一多媒体资源进行裁剪，得到第二多媒体资源，其中，第二多媒体资源不包含非语音数据。

在本公开的一个实施例中，语音识别模块 31 在对第一多媒体资源的音频数据进行语音识别，确定第一多媒体资源对应的初始文本内容时，具体用于：根据自动语音识别 ASR 技术，对第一多媒体资源的音频数据进行语音识别，获得多个语音字，以及各语音字对应的时间戳，时间戳表征语音字对应的音频数据在第一多媒体资源中的播放位置；根据多个语音字，生成初始文本内容；第二确认模块 33，具体用于：根据无效文本内容中各语音字对应的时间戳，确定无效文本内容的语音数据在第一多媒体资源中的第一播放位置。

在本公开的一个实施例中，第一确认模块 32，具体用于：根据语音字对应的时间戳，获取各语音字的播放时长；根据预设的标准时长，以及语音字的播放时长，将播放时长大于标准时长第一时长阈值的语音字，或者，播放时长小于标准时长第二时长阈值的语音字，确定为初始文本内容中的无效文本内容。

在本公开的一个实施例中，第一多媒体资源还包括视频数据，生成模块 34，具体用于：基于第一播放位置，对第一多媒体资源的音频数据和视频数据进行裁剪，得到第二多媒体资源。

本实施例提供的多媒体处理装置 3，可用于执行上述方法实施例的技术方案，其实现原理和技术效果类似，本实施例此处不再赘述。

图 11 为本公开实施例提供的一种电子设备的结构示意图，如图 11 所示，该电子设备 4 包括至少一个处理器 41 和存储器 42；

存储器 42 存储计算机执行指令；

至少一个处理器 41 执行存储器 42 存储的计算机执行指令，使得至少一个处理器 41 5 执行如图 2-图 7 所示实施例中的多媒体处理方法。

其中，处理器 41 和存储器 42 通过总线 43 连接。

相关说明可以对应参见图 2-图 7 所对应的实施例中的步骤所对应的相关描述和效果进行理解，此处不做过多赘述。

参考图 12，其示出了适于用来实现本公开实施例的电子设备 900 的结构示意图，该 10 电子设备 900 可以为终端设备或服务器。其中，终端设备可以包括但不限于诸如移动电话、笔记本电脑、数字广播接收器、个人数字助理（Personal Digital Assistant，简称 PDA）、平板电脑（Portable Android Device，简称 PAD）、便携式多媒体播放器（Portable Media Player，简称 PMP）、车载终端（例如车载导航终端）等等的移动终端以及诸如数字电视（Television，TV）、台式计算机等等的固定终端。图 12 示出的电子设备仅仅是一个示例， 15 不对本公开实施例的功能和使用范围带来任何限制。

如图 12 所示，电子设备 900 可以包括处理装置（例如中央处理器、图形处理器等） 901，其可以根据存储在只读存储器（Read Only Memory，简称 ROM）902 中的程序或者从存储装置 908 加载到随机访问存储器（Random Access Memory，简称 RAM）903 中的程序而执行各种适当的动作和处理。在 RAM 903 中，还存储有电子设备 900 操作所需的 20 各种程序和数据。处理装置 901、ROM 902 以及 RAM 903 通过总线 904 彼此相连。输入/输出（Input/Output，简称 I/O）接口 905 也连接至总线 904。

通常，以下装置可以连接至 I/O 接口 905：包括例如触摸屏、触摸板、键盘、鼠标、摄像头、麦克风、加速度计、陀螺仪等的输入装置 906；包括例如液晶显示器（Liquid Crystal Display，简称 LCD）、扬声器、振动器等输出装置 907；包括例如磁带、硬盘等的存储 25 装置 908；以及通信装置 909。通信装置 909 可以允许电子设备 900 与其他设备进行无线或有线通信以交换数据。虽然图 12 示出了具有各种装置的电子设备 900，但是应理解的是，并不要求实施或具备所有示出的装置。可以替代地实施或具备更多或更少的装置。

特别地，根据本公开的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本公开的实施例包括一种计算机程序产品，其包括承载在计算机可读介 30 质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信装置 909 从网络上被下载和安装，或者从存储装置 908 被安装，或者从 ROM 902 被安装。在该计算机程序被处理装置 901 执行时，执行本公开实施例的方法中限定的上述功能。

需要说明的是，本公开上述的计算机可读介质可以是计算机可读信号介质或者计算机 35 可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是一——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（Erasable Programmable ROM，简称 EPROM 或闪存）、光纤、便

携式紧凑磁盘只读存储器 (Compact Disc ROM, 简称 CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开中, 计算机可读存储介质可以是任何包含或存储程序的有形介质, 该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开中, 计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号, 其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式, 包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读信号介质还可以是计算机可读存储介质以外的任何计算机可读介质, 该计算机可读信号介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输, 包括但不限于: 电线、光缆、射频 (Radio Frequency, 简称 RF) 等等, 或者上述的任意合适的组合。

上述计算机可读介质可以是上述电子设备中所包含的; 也可以是单独存在, 而未装配入该电子设备中。

上述计算机可读介质承载有一个或者多个程序, 当上述一个或者多个程序被该电子设备执行时, 使得该电子设备执行上述实施例所示的方法。

可以以一种或多种程序设计语言或其组合来编写用于执行本公开的操作的计算机程序代码, 上述程序设计语言包括面向对象的程序设计语言—诸如 Java、Smalltalk、C++, 还包括常规的过程式程序设计语言—诸如 “C” 语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中, 远程计算机可以通过任意种类的网络——包括局域网 (Local Area Network, 简称 LAN) 或广域网 (Wide Area Network, 简称 WAN) 一连接到用户计算机, 或者, 可以连接到外部计算机 (例如利用因特网服务提供商来通过因特网连接)。

附图中的流程图和框图, 图示了按照本公开各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上, 流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分, 该模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意, 在有些作为替换的实现中, 方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如, 两个接连地表示的方框实际上可以基本并行地执行, 它们有时也可以按相反的顺序执行, 这依所涉及的功能而定。也要注意, 框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合, 可以用执行规定的功能或操作的专用的基于硬件的系统来实现, 或者可以用专用硬件与计算机指令的组合来实现。

描述于本公开实施例中所涉及到的单元可以通过软件的方式实现, 也可以通过硬件的方式来实现。其中, 单元的名称在某种情况下并不构成对该单元本身的限定, 例如, 第一获取单元还可以被描述为 “获取至少两个网际协议地址的单元”。

本文中以上描述的功能可以至少部分地由一个或多个硬件逻辑部件来执行。例如, 非限制性地, 可以使用的示范类型的硬件逻辑部件包括: 现场可编程门阵列 (Field Programmable Gate Array, 简称 FPGA)、专用集成电路 (Application Specific Integrated Circuit, 简称 ASIC)、专用标准产品 (Application Specific Standard Parts, 简称 ASSP)、

片上系统（System on a Chip，简称 SOC）、复杂可编程逻辑设备（Complex Programmable Logic Device，简称 CPLD）等等。

在本公开的上下文中，机器可读介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦除可编程只读存储器（EPROM 或快闪存储器）、光纤、便捷式紧凑盘只读存储器（CD-ROM）、光学储存设备、磁储存设备、或上述内容的任何合适组合。

第一方面，根据本公开的一个或多个实施例，提供了一种多媒体处理方法，包括：获取第一多媒体资源；对所述第一多媒体资源的音频数据进行语音识别，确定所述第一多媒体资源对应的初始文本内容，其中，所述第一多媒体资源的音频数据包含所述初始文本内容的语音数据；确定所述初始文本内容中的无效文本内容，其中，所述无效文本内容为语义上无信息表达作用的文本内容；确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置；基于所述第一播放位置，对所述第一多媒体资源进行裁剪，得到第二多媒体资源，其中，所述第二多媒体资源的音频数据中包含目标文本内容的语音数据且不包含所述无效文本内容的语音数据，所述目标文本内容为所述初始文本内容中除所述无效文本内容外的其他文本内容。

根据本公开的一个或多个实施例，确定所述初始文本内容中的无效文本内容，包括：对所述初始文本内容进行语义分析，获得所述初始文本内容对应的语义信息；根据所述语义信息，确定所述初始文本内容中的无效文本内容。

根据本公开的一个或多个实施例，所述初始文本内容包括多个片段内容，根据所述语义信息，确定所述初始文本内容中的无效文本内容，包括：根据所述初始文本内容对应的语义信息，确定所述初始文本内容中至少一个片段内容的可信度系数，所述可信度系数用于表征所述片段内容为所述无效文本内容的可信度；根据所述片段内容的可信度系数以及预设的可信度阈值，从所述至少一个片段内容中确定至少一个无效片段内容；根据所述至少一个无效片段内容，确定所述初始文本内容中的无效文本内容。

根据本公开的一个或多个实施例，确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置，包括：确定各所述无效片段内容的语音数据在所述第一多媒体资源的音频数据中的起始点和终止点；根据各所述无效片段内容对应的所述起始点和所述终止点，确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置。

根据本公开的一个或多个实施例，基于所述第一播放位置，对所述第一多媒体资源进行裁剪，得到第二多媒体资源，包括：基于所述第一播放位置，获取所述初始文本内容中除所述无效片段内容外的其他文本内容，其中，所述其他文本内容包括至少一个目标片段内容；在至少一个所述目标片段内容对应的语音数据的起始点添加淡入效果，和/或，在至少一个所述目标片段内容对应的语音数据的结束点添加淡出效果，生成与所述

目标片段内容对应的过渡语音数据；按照所述第一播放位置拼接所述过渡语音数据，生成所述第二多媒体资源。

根据本公开的一个或多个实施例，在对所述初始文本内容进行语义分析，获得所述初始文本内容对应的语义信息之前，所述方法还包括：基于预设的无效文本内容库，确定所述初始文本内容中的无效文本内容；对所述初始文本内容进行语义分析，获得所述初始文本内容对应的语义信息，包括：若所述无效文本内容库中不存在所述无效文本内容，则对所述初始文本内容进行语义分析，获得所述初始文本内容对应的语义信息；在根据所述语义信息，确定所述初始文本内容中的无效文本内容之后，所述方法还包括：将基于所述语义信息确定的所述无效文本内容，添加至所述无效文本内容库。

- 5 根据本公开的一个或多个实施例，在基于所述第一播放位置，对所述第一多媒体资源进行裁剪，得到第二多媒体资源之前，还包括：显示所述初始文本内容中的无效文本内容；响应于针对所述无效文本内容的操作指令，播放所述无效文本内容对应的音频片段。

- 10 根据本公开的一个或多个实施例，对所述第一多媒体资源的音频数据进行语音识别，确定所述第一多媒体资源对应的初始文本内容，包括：通过语音端点检测算法，识别所述第一多媒体资源的音频数据中的语音数据和非语音数据；对所述第一多媒体资源的音频数据中的语音数据进行语音识别，确定所述第一多媒体资源对应的初始文本内容。

- 15 根据本公开的一个或多个实施例，所述方法还包括：根据所述非语音数据的起始点和终止点，确定所述非语音数据在所述第一多媒体资源中的第二播放位置；基于所述第一播放位置，对所述第一多媒体资源进行裁剪，得到第二多媒体资源，包括：基于所述第一播放位置和所述第二播放位置，对所述第一多媒体资源进行裁剪，得到所述第二多媒体资源，其中，所述第二多媒体资源不包含所述非语音数据。

- 20 根据本公开的一个或多个实施例，对所述第一多媒体资源的音频数据进行语音识别，确定所述第一多媒体资源对应的初始文本内容，包括：根据自动语音识别技术，对所述第一多媒体资源的音频数据进行语音识别，获得多个语音字，以及各所述语音字对应的时间戳，所述时间戳表征所述语音字对应的音频数据在所述第一多媒体资源中的播放位置；根据所述多个语音字，生成所述初始文本内容；确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置，包括：根据所述无效文本内容中各语音字对应的时间戳，确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置。

- 25 根据本公开的一个或多个实施例，确定所述初始文本内容中的无效文本内容，包括：根据所述语音字对应的时间戳，获取各所述语音字的播放时长；根据预设的标准时长，以及所述语音字的播放时长，将所述播放时长大于所述标准时长第一时长阈值的语音字，或者，所述播放时长小于所述标准时长第二时长阈值的语音字，确定为所述初始文本内容中的无效文本内容。

- 30 根据本公开的一个或多个实施例，所述第一多媒体资源还包括视频数据，基于所述第一播放位置，对所述第一多媒体资源进行裁剪，得到第二多媒体资源，包括：基于所述第一播放位置，对所述第一多媒体资源的音频数据和视频数据进行裁剪，得到所述第二多媒体资源。

第二方面，根据本公开的一个或多个实施例，提供了一种多媒体处理装置，包括：
语音识别模块，用于获取第一多媒体资源，并对所述第一多媒体资源的音频数据进行语音识别，确定所述第一多媒体资源对应的初始文本内容，其中，所述第一多媒体资源的音频数据包含所述初始文本内容的语音数据；

5 第一确认模块，用于确定所述初始文本内容中的无效文本内容，其中，所述无效文本内容为语义上无信息表达作用的文本内容；

第二确认模块，确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置；

10 生成模块，用于基于所述第一播放位置，对所述第一多媒体资源进行裁剪，得到第二多媒体资源，其中，所述第二多媒体资源的音频数据中包含目标文本内容的语音数据且不包含所述无效文本内容的语音数据，所述目标文本内容为所述初始文本内容中除所述无效文本内容外的其他文本内容。

根据本公开的一个或多个实施例，第一确认模块，具体用于：对所述初始文本内容进行语义分析，获得所述初始文本内容对应的语义信息；根据所述语义信息，确定所述
15 初始文本内容中的无效文本内容。

根据本公开的一个或多个实施例，所述初始文本内容包括多个片段内容，第一确认模块在根据所述语义信息，确定所述初始文本内容中的无效文本内容时，具体用于：根据所述初始文本内容对应的语义信息，确定所述初始文本内容中至少一个片段内容的可信度系数，所述可信度系数用于表征所述片段内容为所述无效文本内容的可信度；根据
20 所述片段内容的可信度系数以及预设的可信度阈值，从所述至少一个片段内容中确定至少一个无效片段内容；根据所述至少一个无效片段内容，确定所述初始文本内容中的无效文本内容。

根据本公开的一个或多个实施例，所述第二确认模块，具体用于：确定各所述无效片段内容的语音数据在所述第一多媒体资源的音频数据中的起始点和终止点；根据各所述
25 无效片段内容对应的所述起始点和所述终止点，确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置。

根据本公开的一个或多个实施例，所述生成模块，具体用于：基于所述第一播放位置，获取所述初始文本内容中除所述无效片段内容外的其他文本内容，其中，所述其他文本内容包括至少一个目标片段内容；在至少一个所述目标片段内容对应的语音数据的
30 起始点添加淡入效果，和/或，在至少一个所述目标片段内容对应的语音数据的结束点添加淡出效果，生成与所述目标片段内容对应的过渡语音数据；按照所述第一播放位置拼接所述过渡语音数据，生成所述第二多媒体资源。

根据本公开的一个或多个实施例，在对所述初始文本内容进行语义分析，获得所述初始文本内容对应的语义信息之前，第一确认模块还用于：基于预设的无效文本内容库，
35 确定所述初始文本内容中的无效文本内容；第一确认模块在对所述初始文本内容进行语义分析，获得所述初始文本内容对应的语义信息时，具体用于：若所述无效文本内容库中不存在所述无效文本内容，则对所述初始文本内容进行语义分析，获得所述初始文本内容对应的语义信息；第一确认模块在根据所述语义信息，确定所述初始文本内容中的

无效文本内容之后，还用于：将基于所述语义信息确定的所述无效文本内容，添加至所述无效文本内容库。

根据本公开的一个或多个实施例，在基于所述第一播放位置，对所述第一多媒体资源进行裁剪，得到第二多媒体资源之前，所述生成模块，还用于：显示所述初始文本内容中的无效文本内容；响应于针对所述无效文本内容的操作指令，播放所述无效文本内容对应的音频片段。

根据本公开的一个或多个实施例，所述语音识别模块对所述第一多媒体资源的音频数据进行语音识别，确定所述第一多媒体资源对应的初始文本内容时，具体用于：通过语音端点检测算法，识别所述第一多媒体资源的音频数据中的语音数据和非语音数据；对所述第一多媒体资源的音频数据中的语音数据进行语音识别，确定所述第一多媒体资源对应的初始文本内容。

根据本公开的一个或多个实施例，所述第二确认模块，还用于：根据所述非语音数据的起始点和终止点，确定所述非语音数据在所述第一多媒体资源中的第二播放位置；所述生成模块，具体用于：基于所述第一播放位置和所述第二播放位置，对所述第一多媒体资源进行裁剪，得到所述第二多媒体资源，其中，所述第二多媒体资源不包含所述非语音数据。

根据本公开的一个或多个实施例，语音识别模块在对所述第一多媒体资源的音频数据进行语音识别，确定所述第一多媒体资源对应的初始文本内容时，具体用于：根据自动语音识别技术，对所述第一多媒体资源的音频数据进行语音识别，获得多个语音字，以及各所述语音字对应的时间戳，所述时间戳表征所述语音字对应的音频数据在所述第一多媒体资源中的播放位置；根据所述多个语音字，生成所述初始文本内容；所述第二确认模块，具体用于：根据所述无效文本内容中各语音字对应的时间戳，确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置。

根据本公开的一个或多个实施例，所述第一确认模块，具体用于：根据所述语音字对应的时间戳，获取各所述语音字的播放时长；根据预设的标准时长，以及所述语音字的播放时长，将所述播放时长大于所述标准时长第一时长阈值的语音字，或者，所述播放时长小于所述标准时长第二时长阈值的语音字，确定为所述初始文本内容中的无效文本内容。

根据本公开的一个或多个实施例，所述第一多媒体资源还包括视频数据，所述生成模块，具体用于：基于所述第一播放位置，对所述第一多媒体资源的音频数据和视频数据进行裁剪，得到所述第二多媒体资源。

第三方面，根据本公开的一个或多个实施例，提供了一种电子设备，包括：至少一个处理器和存储器；所述存储器存储计算机执行指令；所述至少一个处理器执行所述存储器存储的所述计算机执行指令，使得所述至少一个处理器执行如上第一方面以及第一方面各种可能的设计所述的多媒体处理方法。

第四方面，根据本公开的一个或多个实施例，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执行指令，当处理器执行所述计算机执行指令时，实现如上第一方面以及第一方面各种可能的设计所述的多媒体处理方法。

第五方面，本公开实施例提供一种计算机程序产品，包括计算机程序，该计算机程序被处理器执行时实现如上第一方面以及第一方面各种可能的设计所述的多媒体处理方法。

5 第六方面，本公开实施例提供一种计算机程序，该计算机程序被处理器执行时实现如上第一方面以及第一方面各种可能的设计所述的多媒体处理方法。

10 以上描述仅为本公开的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本公开中所涉及的公开范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离上述公开构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本公开中公开的（但不限于）具有类似功能的技术特征进行互相替换而形成的技术方案。

15 此外，虽然采用特定次序描绘了各操作，但是这不应理解为要求这些操作以所示出的特定次序或以顺序次序执行来执行。在一定环境下，多任务和并行处理可能是有利的。同样地，虽然在上面论述中包含了若干具体实现细节，但是这些不应被解释为对本公开的范围的限制。在单独的实施例的上下文中描述的某些特征还可以组合地实现在单个实施例中。相反地，在单个实施例的上下文中描述的各种特征也可以单独地或以任何合适的子组合的方式实现在多个实施例中。

20 尽管已经采用特定于结构特征和/或方法逻辑动作的语言描述了本主题，但是应当理解所附权利要求书中所限定的主题未必局限于上面描述的特定特征或动作。相反，上面所描述的特定特征和动作仅仅是实现权利要求书的示例形式。

权 利 要 求 书

- 1、一种多媒体处理方法，其特征在于，包括：
获取第一多媒体资源；
5 对所述第一多媒体资源的音频数据进行语音识别，确定所述第一多媒体资源对应的初始文本内容，其中，所述第一多媒体资源的音频数据包含所述初始文本内容的语音数据；
确定所述初始文本内容中的无效文本内容，其中，所述无效文本内容为语义上无信息表达作用的文本内容；
确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置；
10 基于所述第一播放位置，对所述第一多媒体资源进行裁剪，得到第二多媒体资源，其中，所述第二多媒体资源的音频数据中包含目标文本内容的语音数据且不包含所述无效文本内容的语音数据，所述目标文本内容为所述初始文本内容中除所述无效文本内容外的其他文本内容。
- 2、根据权利要求1所述的方法，其特征在于，确定所述初始文本内容中的无效文本内容，包括：
15 对所述初始文本内容进行语义分析，获得所述初始文本内容对应的语义信息；
根据所述语义信息，确定所述初始文本内容中的无效文本内容。
- 3、根据权利要求2所述的方法，其特征在于，所述初始文本内容包括多个片段内容，根据所述语义信息，确定所述初始文本内容中的无效文本内容，包括：
20 根据所述初始文本内容对应的语义信息，确定所述初始文本内容中至少一个片段内容的可信度系数，所述可信度系数用于表征所述片段内容为所述无效文本内容的可信度；
根据所述片段内容的可信度系数以及预设的可信度阈值，从所述至少一个片段内容中确定至少一个无效片段内容；
根据所述至少一个无效片段内容，确定所述初始文本内容中的无效文本内容。
- 25 4、根据权利要求3所述的方法，其特征在于，确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置，包括：
确定各所述无效片段内容的语音数据在所述第一多媒体资源的音频数据中的起始点和终止点；
根据各所述无效片段内容对应的所述起始点和所述终止点，确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置。
- 30 5、根据权利要求3或4所述的方法，其特征在于，基于所述第一播放位置，对所述第一多媒体资源进行裁剪，得到第二多媒体资源，包括：
基于所述第一播放位置，获取所述初始文本内容中除所述无效片段内容外的其他文本内容，其中，所述其他文本内容包括至少一个目标片段内容；
35 在至少一个所述目标片段内容对应的语音数据的起始点添加淡入效果，和/或，在至少一个所述目标片段内容对应的语音数据的结束点添加淡出效果，生成与所述目标片段内容对应的过渡语音数据；
按照所述第一播放位置拼接所述过渡语音数据，生成所述第二多媒体资源。
- 6、根据权利要求2-5任一项所述的方法，其特征在于，在对所述初始文本内容进行

语义分析, 获得所述初始文本内容对应的语义信息之前, 所述方法还包括:

基于预设的无效文本内容库, 确定所述初始文本内容中的无效文本内容;

对所述初始文本内容进行语义分析, 获得所述初始文本内容对应的语义信息, 包括:

若所述无效文本内容库中不存在所述无效文本内容, 则对所述初始文本内容进行语

5 义分析, 获得所述初始文本内容对应的语义信息;

在根据所述语义信息, 确定所述初始文本内容中的无效文本内容之后, 所述方法还包括:

将基于所述语义信息确定的所述无效文本内容, 添加至所述无效文本内容库。

7、根据权利要求 1-6 任一项所述的方法, 其特征在于, 在基于所述第一播放位置,
10 对所述第一多媒体资源进行裁剪, 得到第二多媒体资源之前, 还包括:

显示所述初始文本内容中的无效文本内容;

响应于针对所述无效文本内容的操作指令, 播放所述无效文本内容对应的音频片段。

8、根据权利要求 1-7 任一项所述的方法, 其特征在于, 对所述第一多媒体资源的音频数据进行语音识别, 确定所述第一多媒体资源对应的初始文本内容, 包括:

15 通过语音端点检测 VAD 算法, 识别所述第一多媒体资源的音频数据中的语音数据和非语音数据;

对所述第一多媒体资源的音频数据中的语音数据进行语音识别, 确定所述第一多媒体资源对应的初始文本内容。

9、根据权利要求 8 所述的方法, 其特征在于, 所述方法还包括:

20 根据所述非语音数据的起始点和终止点, 确定所述非语音数据在所述第一多媒体资源中的第二播放位置;

基于所述第一播放位置, 对所述第一多媒体资源进行裁剪, 得到第二多媒体资源, 包括:

25 基于所述第一播放位置和所述第二播放位置, 对所述第一多媒体资源进行裁剪, 得到所述第二多媒体资源, 其中, 所述第二多媒体资源不包含所述非语音数据。

10、根据权利要求 1-7 任一项所述的方法, 其特征在于, 对所述第一多媒体资源的音频数据进行语音识别, 确定所述第一多媒体资源对应的初始文本内容, 包括:

根据自动语音识别 ASR 技术, 对所述第一多媒体资源的音频数据进行语音识别, 获得多个语音字, 以及各所述语音字对应的时间戳, 所述时间戳表征所述语音字对应的音频数据在所述第一多媒体资源中的播放位置;

30 根据所述多个语音字, 生成所述初始文本内容;

确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置, 包括:

根据所述无效文本内容中各语音字对应的时间戳, 确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置。

35 11、根据权利要求 10 所述的方法, 其特征在于, 确定所述初始文本内容中的无效文本内容, 包括:

根据所述语音字对应的时间戳, 获取各所述语音字的播放时长;

根据预设的标准时长, 以及所述语音字的播放时长, 将所述播放时长大于所述标准时长第一时长阈值的语音字, 或者, 所述播放时长小于所述标准时长第二时长阈值的语

音字，确定为所述初始文本内容中的无效文本内容。

12、根据权利要求 1-11 任一项所述的方法，其特征在于，所述第一多媒体资源还包括视频数据，基于所述第一播放位置，对所述第一多媒体资源进行裁剪，得到第二多媒体资源，包括：

- 5 基于所述第一播放位置，对所述第一多媒体资源的音频数据和视频数据进行裁剪，得到所述第二多媒体资源。

13、一种多媒体处理装置，其特征在于，包括：

- 语音识别模块，用于获取第一多媒体资源，并对所述第一多媒体资源的音频数据进行语音识别，确定所述第一多媒体资源对应的初始文本内容，其中，所述第一多媒体资源
10 的音频数据包含所述初始文本内容的语音数据；

 第一确认模块，用于确定所述初始文本内容中的无效文本内容，其中，所述无效文本内容为语义上无信息表达作用的文本内容；

 第二确认模块，确定所述无效文本内容的语音数据在所述第一多媒体资源中的第一播放位置；

- 15 生成模块，用于基于所述第一播放位置，对所述第一多媒体资源进行裁剪，得到第二多媒体资源，其中，所述第二多媒体资源的音频数据中包含目标文本内容的语音数据且不包含所述无效文本内容的语音数据，所述目标文本内容为所述初始文本内容中除所述无效文本内容外的其他文本内容。

14、一种电子设备，其特征在于，包括：至少一个处理器和存储器；

- 20 所述存储器存储计算机执行指令；

 所述至少一个处理器执行所述存储器存储的所述计算机执行指令，使得所述至少一个处理器执行如权利要求 1-12 任一项所述的多媒体处理方法。

- 15、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执行指令，当处理器执行所述计算机执行指令时，实现如权利要求 1-12 任一项所述的多媒体处理方法。
25

16、一种计算机程序产品，包括计算机程序，该计算机程序被处理器执行时实现权利要求 1-12 中任一项所述的方法。

17、一种计算机程序，所述计算机程序被处理器执行时实现权利要求 1-12 任一项所述的方法。
30