(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
2024 年 12 月 19 日 (19.12.2024)

WIPO | PCT

(10) 国际公布号
WO 2024/255461 A1

(51) 国际专利分类号:
G10L 21/028 (2013.01)

(21) 国际申请号: PCT/CN2024/089862

(22) 国际申请日: 2024 年 4 月 25 日 (25.04.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202310699993.5 2023年6月13日 (13.06.2023) CN

(71) 申请人: 腾 讯 科 技 (深 圳) 有 限 公 司 (**TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED**) [CN/CN]; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(72) 发明人: 冯鑫 (**FENG, Xin**); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(74) 代理人: 广州三环专利商标代理有限公司 (**SCIHEAD IP LAW FIRM**); 中国广东省广州市越秀区先烈中路80号汇华商贸大厦1508室, Guangdong 510070 (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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ,

(54) **Title:** SPEECH PROCESSING METHOD AND APPARATUS, DEVICE, MEDIUM, AND PROGRAM PRODUCT

(54) 发明名称: 一种语音处理方法、装置、设备、介质及程序产品

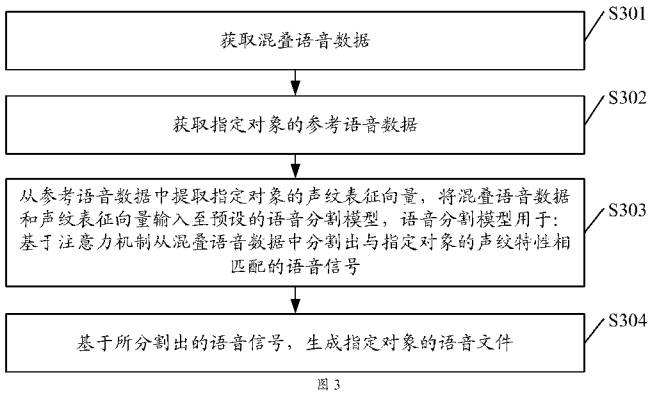

图 3

(57) **Abstract:** A speech processing method and apparatus, a computer device, a storage medium, and a program product. The method comprises: acquiring overlapping speech data (S301); acquiring reference speech data of a specified object (S302); extracting a voiceprint representation vector of the specified object from the reference speech data, wherein the voiceprint representation vector is used for representing a voiceprint characteristic of the specified object, and inputting the overlapping speech data and the voiceprint representation vector into a preset speech segmentation model, wherein the speech segmentation model is used for segmenting the overlapping speech data on the basis of an attention mechanism to obtain a target speech signal matched with the voiceprint characteristic (S303); and generating a speech file of the specified object on the basis of the speech signal obtained by segmentation (S304). According to the method, the overlapping speech data can be segmented to obtain a clearer speech signal of any specified object.

(57) 摘要: 一种语音处理方法、装置、计算机设备、存储介质及程序产品，该方法包括：获取混叠语音数据（S301）；获取指定对象的参考语音数据（S302）；从参考语音数据中提取指定对象的声纹表征向量，声纹表征向量用于表征指定对象的声纹特性，将混叠语音数据和声纹表征向量输入预设的语音分割模型，语音分割模型用于：基于注意力机制从混叠语音数据中分割出与声纹特性相匹配的目标语音信号（S303）；基于所分割出的语音信号，生成指定对象的语音文件（S304）。该方法能够从混叠语音数据中分割出任意的指定对象的纯净的语音信号。

WO 2024/255461 A1

[见续页]

IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, 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))。

一种语音处理方法、装置、设备、介质及程序产品

本申请要求于2023年06月13日提交中国专利局、申请号为：202310699993.5、申请名称为“一种语音处理方法、装置、设备、介质及程序产品”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及计算机技术领域，尤其涉及人工智能领域，具体涉及一种语音处理方法、一种语音处理装置、一种计算机设备、一种计算机可读存储介质及一种计算机程序产品。

背景技术

混叠语音数据（或者称为混叠语音）是混叠着多种声音源（即产生声音的对象）产生的语音信号的语音数据。例如，在会议场景中录音设备从物理环境中录取的混叠语音数据中可以包括多个参会者产生的语音信号，还可以包括该物理环境中某些设备（如播放会议视频的设备）产生的语音信号。

目前，针对混叠语音数据提供的信源分离的方法包括：1、人耳分离混叠语音数据，这种人工听取方式导致分割过程较长，效率较低。2、依赖音色频率分离混叠语音数据，当存在多个音色频率较为相似的对象时，不能实现精准分割。3、根据声音源的距离分离混叠语音数据，会使语音分割受限于各声音源的距离不同。4、采用指定对象的专属语音分割模型分离混叠语音数据，该方法不可移植，无法做到通用性。

发明内容

本申请实施例提供一种语音处理方法、装置、设备、介质及程序产品，能够从混叠语音数据中分割出任意的指定对象的纯净语音信号，具有通用性。

一方面，本申请实施例提供了一种语音处理方法，该方法由计算机设备执行，该方法包括：

获取混叠语音数据，混叠语音数据中包含至少两个对象中每个对象产生的语音信号；

获取指定对象的参考语音数据；指定对象是指至少两个对象中的任一个；参考语音数据中包含指定对象的参考语音信号；

从参考语音数据中提取指定对象的声纹表征向量，所述声纹表征向量用于表征指定对象的声纹特性；

将混叠语音数据和声纹表征向量输入预设的语音分割模型，所述语音分割模型用于：基于注意力机制从混叠语音数据中分割出与声纹特性相匹配的目标语音信号；

基于所分割出的目标语音信号，生成指定对象的语音文件。

另一方面，本申请实施例提供了一种语音处理装置，该装置包括：

获取单元，用于获取混叠语音数据，混叠语音数据中包含至少两个对象中每个对象产生的语音信号；

获取单元，还用于获取指定对象的参考语音数据；指定对象是指至少两个对象中的任一个；参考语音数据中包含指定对象的参考语音信号；

处理单元，用于从参考语音数据中提取指定对象的声纹表征向量，所述声纹表征向量用于表征指定对象的声纹特性；

处理单元，还用于将混叠语音数据和声纹表征向量输入预设的语音分割模型，所述语音分割模型用于：基于注意力机制从混叠语音数据中分割出与声纹特性相匹配的目标语音信号；

处理单元，还用于基于所分割出的目标语音信号，生成指定对象的语音文件。

另一方面，本申请实施例提供了一种计算机设备，该计算机设备包括：

处理器，用于加载并执行计算机程序；

计算机可读存储介质，该计算机可读存储介质中存储有计算机程序，该计算机程序被处理器执行时，实现上述语音处理方法。

另一方面，本申请实施例提供了一种计算机可读存储介质，计算机可读存储介质存储有计算机程序，该计算机程序适于由处理器加载并执行上述语音处理方法。

另一方面，本申请实施例提供了一种计算机程序产品，该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，实现上述语音处理方法。

本申请实施例中，获取待分割的混叠语音数据，混叠语音数据中包含至少两个对象中每个对象产生的语音信号；如果具有对该至少两个对象中的指定对象所产生的语音信号的分割需求，可以获取该指定对象的一段参考语音数据（如该指定对象产生的几秒语音）；该指定对象可以是至少两个对象中的任一个。从该参考语音数据中提取该指定对象的声纹表征向量，该声纹表征向量能够表征指定对象的声纹特性，该声纹特性具有唯一性，能够表征该指定对象的身份。这样，可以将能够唯一表征指定对象身份的声纹表征向量和待分割的混叠语音数据输入至预设的语音分割模型，使得语音分割模型能够基于注意力机制从混叠语音数据中分割出与指定对象的声纹特性相匹配的目标语音信号，从而基于所分割出的目标语音信号，生成该指定对象的单独的语音文件。由此可见，一方面，本申请实施例支持从指定对象的纯净的参考语音数据中提取出表征指定对象的声纹特性的声纹表征向量，并将该声纹表征向量作为参考，利用语音分割模型提供的注意力机制，从混叠语音数据中清晰准确的计算和提取出指定对象的目标语音信号，提升目标语音信号的提取纯净性，达到更准确的语音分离效果。另一方面，本申请实施例只需要获取指定对象的参考语音数据，就能够从混叠语音数据中分割出该指定对象的目标语音信号；如果想要获取其他对象的语音数据，只需要更换待分割的对象的声纹表征向量，不需要针对每个对象训练一个专属网络，大大提高便捷和可迁移，提升本方案的通用性。

附图说明

图 1 是本申请一个示例性实施例提供的一种语音处理系统的架构示意图；

图 2 是本申请一个示例性实施例提供的一种语音处理场景的架构示意图；

图 3 是本申请一个示例性实施例提供的一种语音处理方法的流程示意图；

图 4 是本申请一个示例性实施例提供的一种由用户输入指定对象的参考语音数据的界面示意图；

图 5 是一种现有的 Unet 网络的结构示意图；

图 6 是本申请一个示例性实施例提供的一种在 Unet 网络中的每个网络层中均加入 Attention 机制，所构建的语音分割模型的结构示意图；

图 7a 是本申请一个示例性实施例提供的一种目标网络层为卷积层或卷积连接层时语音分割的示意图；

图 7b 是本申请一个示例性实施例提供的一种目标网络层为上采样层时语音分割的示意图；

图 8 是本申请一个示例性实施例提供的另一种语音处理方法的流程示意图；

图 9 是本申请一个示例性实施例提供的一种声纹向量提取的流程示意图；

图 10 是本申请一个示例性实施例提供的一种改进型 PANNS 的结构示意图；

图 11 是本申请一个示例性实施例提供的一种 transformer 网络的结构示意图；

图 12 是本申请一个示例性实施例提供的一种语音处理装置的结构示意图；

图 13 是本申请一个示例性实施例提供的一种计算机设备的结构示意图。

具体实施方式

在本申请实施例中，提供了一种语音处理方案，具体是提供了一种针对混叠语音数据进行信源分离的语音分离方案。其中，混叠语音数据可以简称为混叠语音或者混合音频信号，是一条掺杂着多种语音信号（或者称为音频信号）的音频；即所谓“混叠”可以理解为多种语音信号混合/糅杂在一起。在实际应用场景中，该混叠语音数据可以理解为：通过收音设备（如麦克风）直接从环境中采集到的，包含多种声音源产生的语音信号的语音数据。其中，多种语音信号可以是由不同的对象（或者称为声音源）所产生的，此处的对象可以包括但是不限于：人类、动物或者实体设备（如汽车）等；本申请实施例对混叠语音数据所包含的多种语音信号的来源不作限定。举例来说，在多人参与讨论的会议场景中，采集的语音数据中通常包含不同参与者产生的语音信号；当然，如果在会议场景中还包含播放音视频的设备，那么采集的语音数据中还包含该设备发出的语音信号；如此，可以将在会议场景中采集到的语音数据称为混叠语音数据，该混叠语音数据中包含会话场景中的多个对象所产生的语音信号。

进一步的，信源分离是指：从混叠语音数据中分离出某个指定对象的语音信号的过程。换句话说，信源分离可以简单理解为：通过信号处理或者其他算法将混叠语音数据进行分离，以实现从混叠语音数据中分割出指定对象的目标语音信号，最终生成该指定对象的单独的音频文件（或语音文件）的技术。举例来说，在户外嘈杂场景中采集了一段混叠语音数据后，可以通过信源分离的技术从该混叠语音数据中提取出某个指定对象所产生的目标语音信号，以生成该指定对象的语音文件；这样，播放该语音文件时只存在该指定对象所产生的语音，以此达到识别某个指定对象所产生的语音的目的。

基于对混叠语音数据和信源分离的概念的简单介绍，本申请实施例提供一种新的语音处理方案，该方案主要包括：获取待分割的混叠语音数据，该混叠语音数据中包含至少两个对象中每个对象产生的语音信号，如混叠语音数据中包括的语音信号包括：对象 1 产生的语音信号和对象 2 产生的语音信号；如果用户想要从混叠语音数据中提取出指定对象（如至少两个对象中的任一对象）所产生的目标语音信号，则可以获取一段包含该指定对象的参考语音信号的参考语音数据。这样，可以基于参考语音数据提取到该指定对象的声纹表征向量，该声纹表征向量能够表征指定对象的声纹特性，声纹特性可以理解为指定对象的声音特色，如指定对象独特的音高或音色等。如此，将指定对象的声纹表征向量和混叠语音数据输入至语音分割模型，就可以利用语音分割模型中的注意力机制，从混叠语音数据中分割提取出与指定对象的声纹特性相匹配的目标语音信号，从而基于该目标语音信号为该指定对象生成单独的语音文件。

由此可见，一方面，本申请实施例依赖于每个用户的声纹特性的唯一性，只需提供任一指定对象的一段参考语音数据来提取表征该指定对象的声纹特性的声纹特征向量，就能够基于该声纹表征向量从混叠语音数据中分离提取出该任一指定对象的目标语音信号；不仅达到从混叠语音数据中精准分离指定对象的目标语音信号的目的，而且对于混叠语音数据所包含的任意语音信号所属的对象均可以实现信源分离，做到了高度的可复用和可移植，降低了用户输入操作复杂度，让整个系统更通用化。另一方面，本申请实施例基于注意力机制来进行对指定对象的声纹特性和混叠语音数据进行计算，大大提高从混叠语音数据中提取指定对象的目标语音信号的清晰性和准确性，避免提取出的目标语音信号中包含有太多杂音，实现更准确纯净的语音分离效果。

本申请实施例主要通过基于声纹向量嵌入的可复用型指定说话人语音分割系统实现语音处理方案，即该系统部署了本申请实施例提供的语音处理方案；这样，任意用户具有对混叠语音数据进行语音信号的分离需求时，可以调用该系统自动从混叠语音数据中分离提取出指定对象对应的语音文件。其中，系统的示例性架构示意图可以参见图 1；如图 1 所示，该系统主要包括两个模块，分别为：声纹向量提取模型和语音分割模型；下面对这两个模块进行简单介绍，其中：

（1）声纹向量提取模型，可以称为声纹向量提取器或者声纹识别网络等。声纹向量提取模型主要用于：对待分割的指定对象的身份进行识别，并提取该指定对象的身份语义向量；此处的身份语义向量在本

申请实施例中称为声纹表征向量（或简称为声纹向量），用于表征该指定对象的声纹特性。

参见图 1 可见，声纹向量提取模型是基于改进的音频神经网络（Pretrained Audio Neural Networks，PANNS）网络和转换器（Transformer）网络构建的。声纹向量提取模型是使用开源的大规模说话人数据集（包含丰富的语音数据的数据集）进行充分训练得到的，训练好的声纹向量提取模型具有充分表达对象的声纹特性的能力；这样，该声纹向量提取模型可以作为整个系统的声纹向量提取器；在推理的阶段，加载提前用大规模数据说话人数据集训练好的模型参数后，可以采用训练好的声纹向量提取模型对指定对象的参考语音数据（如一小段（如几秒或十几秒）语音）计算出该指定对象的声纹表征向量，该声纹表征向量用于表征该指定对象的声纹特性。由此可见，通过大规模说话人数据集对声纹向量提取模型进行训练的方式，无需针对混叠语音数据中每个对象特意收集相关训练数据，摆脱针对混叠语音数据中对象进行数据提取来构建某个对象专属的模型的依赖。

其中：①声纹向量提取模型中改进型 PANNS 是对传统 PANNS 改进得到的；该改进主要体现于：设计了时域链路和频域链路之间的信息交流的链路，使得在声纹表征向量提取的过程中存在多次时域和频域的信息交流，从而实现时域和频域保持信息上的互补，能够让高层网络充分感知底层网络信息，提高声纹向量提取的准确性。其中，PANNS 是一种基于大型音频数据集（包含大规模说话人的语音数据）训练得到的音频神经网络；通常用于音频模式识别或者音频帧级别的向量化（embedding），作为模型前端的编码网络。②转换器网络是一个依赖于注意力机制（Attention）来计算输入和输出的转换模型；Transformer 网络抛弃了卷积模型结构，仅仅通过注意力机制和前向神经网络（Feed Forward Neural Network），不需要使用序列对齐的循环架构就实现了较好的表现。

（2）语音分割模型，可以称为语义分割网络或分割网络等。该语音分割网络主要用于：接收声纹向量提取模型输入的关于指定对象的声纹特性（具体接收的是声纹表征向量），并基于该声纹表征向量利用注意力机制从混叠语音数据中提取出与指定对象的声纹特性相匹配的语音信号。

参见图 1 可见，语音分割模型是融合了注意力机制的分割模型。通过在分割网络中引入注意力机制进行改造，能够在分割网络针对混叠语音数据进行特征处理的过程中，结合注意力机制计算出与指定对象的声纹特征相关的目标语音信号，能够从混叠语音数据中分割出指定对象的目标语音信号；这样，能够更清晰准确的计算提取出混叠语音数据中指定对象的目标语音信号，并将指定对象纯净的目标语音信号分离出来，达到更准确纯净的语音分离效果。其中，注意力机制（Attention 机制）是模仿人类注意力而提出的一种解决问题的办法；简单地说，就是模仿人类注意力从大量信息中快速筛选出想要关注的信息。主要用于解决时序模型输入序列较长时很难获得难以获取合理的向量表示问题，做法是保留时序模型的中间结果，用新的模型对其进行学习并将其与输出进行关联，从而达到信息筛选的目的。

综上所述，本申请实施例提供的系统中包含两个模块，声纹向量提取模型在对指定对象的参考语音数据进行提取出能够表征该指定对象的声纹特性的声纹表征向量后，可以将声纹表征向量嵌入至语音分割模型；这样，语音分割模型可以基于注意力机制从混叠语音数据中提取并分离出与该声纹特性相匹配的语音信号，实现较好地信号分离效果。一方面，参见图 1 可知本申请实施例提供的系统是基于多个深度学习神经网络（如改进型 PANNS 网络、转换网络和融合有注意力机制的分割网络等）进行构建的全自动分割系统；对于该全自动分割系统而言，只需要用户往该全自动分割系统中输入指定对象的参考语音数据和待分割的混叠语音数据，全自动分割系统就能够自动快速地从混叠语音数据中提取出指定对象的语音信号，极大的提升语音分割的效率，彻底摆脱人工的参与，形成快速的标准化。另一方面，通过创新性地将声纹向量提取模型提取的声纹特性嵌入至语音分割模型的模型架构，能够让系统中的语音分离模型可复用；其中，可复用是指每次系统进行信源分离时，只需要更换提取的对象的声纹特性，不需要针对于每个对象均训练一个单独的分割网络，能够做到网络的便捷可迁移，使得整个系统具有极高的通用性。

图 1 所示的系统可以部署于计算机设备中，具体可以是部署于计算机设备中运行的应用程序（如以插件形式部署于应用程序）中；也就是说，由计算机设备中运行的应用程序来提供本方案。其中：①应用程序可以是指为完成某项或多项特定工作的计算机程序。按照不同维度（如应用程序的运行方式、功能等）对应用程序进行归类，可得到同一应用程序在不同维度下的类型。例如：按照应用程序的运行方式分类，应用程序可包括但不限于：安装在终端中的客户端、无需下载安装即可使用的小程序（作为客户端的子程序）、通过浏览器打开的 web（World WideWeb，全球广域网）应用程序等等。再如：按照应用程序的功能类型分类，应用程序可包括但不限于：IM（Instant Messaging，即时通信）应用程序、内容交互应用程序、音频应用程序或者视频应用程序等等。其中，即时通信应用程序是指基于互联网的即时交流消息和社交交互的应用程序，即时通信应用程序可以包括但不限于：包含通信功能的社交应用程序、包含社交交互功能的地图应用程序、游戏应用程序等等。内容交互应用程序是指能够实现内容交互的应用程序，例如可以是网银、分享平台、个人空间、新闻等应用程序。音频应用程序是指基于互联网实现音频功能的应用程序，音频应用程序可以包括但是不限于：具备音乐播放和编辑能力的音乐类应用程序，具备电台播放能力的电台类应用程序或者具备直播能力的直播类应用程序等等。视频应用程序是指能够播放画面的应用程序，视频应用程序可以包括但是不限于：具备短视频（视频长度往往较短，如几秒或几分钟等）的应用程序，具备长视频（如类似电影或电视剧这种播放时常较长的视频）的应用程序等等。

②计算机设备可以包括终端和/或服务器。其中，终端可以包括但是不限于：智能手机（如部署安卓（Android）系统的智能手机，或部署互联网操作系统（Internetworking Operating System，IOS）的智能手机）、平板电脑、便携式个人计算机、移动互联网设备（Mobile Internet Devices，MID）、车载设备、头戴设备、智能电视或智能家居等设备，本申请实施例并不对终端的类型进行限定，在此说明。该终端中部署有图 1 所示的系统或提供该系统的应用程序（或插件）等。服务器可以是独立的物理服务器，也可以是多个物理服务器构成的服务器集群或者分布式系统，还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络（Content Delivery Network，CDN）、以及大数据和人工智能平台等基础云计算服务的云服务器。

由此可见，本申请实施例可以由终端或者服务器执行，还可以由终端和服务器共同执行。一种示例性的由终端和服务器共同执行语音处理方案的系统架构示意图可以参见图 2；如图 2 所示，终端 201 为具有语音分离需求的用户所持有的设备。在用户具有从混叠语音数据中分离出指定对象的语音文件的需求时，用户可以通过终端 201 将待分割的混叠语音数据和指定对象的参考语音数据发送至服务器 202。这样，服务器 202 在接收到指定对象的参考语音数据和待分割的混叠语音数据后，可以先通过系统中的声纹向量提取模型对参考语音数据进行身份识别，得到用于表征指定对象的身份的声纹表征向量，然后将该声纹表征向量嵌入至系统中的语音分割模型；语音分割模型接收到指定对象的声纹表征向量和待分割的混叠语音数据后，能够基于注意力机制从混叠语音数据中提取到与声纹表征向量所表征的声纹特性相匹配的纯净地目标语音信号，从而基于该目标语音信号生成指定对象的语音文件。如此，服务器 202 将指定对象的语音文件返回至终端 201，使得用户可以通过该终端 201 播放只包含该指定对象的语音数据的语音文件。

应当理解的是，上述以计算机设备为终端和服务器为例，对语音处理方案的流程进行了简单介绍；但计算机设备为终端或服务器时，计算机设备执行语音处理方案的思路与上述描述流程是类似的，只是执行主体有所不同，在此不作赘述。此外，图 2 所示的终端 201 和服务器 202 之间可以通过有线或无线通信方式进行直接或间接地连接，本申请在此不做限制。

进一步的，本申请实施例提供的语音处理方案可以应用于任意具有语音分离需求的应用场景；根据所应用的场景不同，提供本方案的计算机设备也有由不同，对此不作限定。其中，应用场景可以包括但是不限于以下至少一种：影视剧场景、音视频创作场景和会话场景等。

可选的，应用场景为影视剧场景。示例性地，影视剧场景为影视剧中针对角色的配音场景。具体地：

在影视剧制作阶段，往往需要配音演员针对影视剧中的某个角色进行配音（如收音录制的混叠语音数据送审后，存在部分的台词不符合规定，需要进行重新录音）。然而，在影视剧拍摄过程或者后期制作过程中进行的收音录制所得到的语音数据通常是包含多个语音信号的混叠语音数据。因此，在配音前，需要将混叠语音数据中除了待重新配音的指定演员的语音信号之外的其他演员的语音信号进行纯净提取，以便于将配音后的该指定演员的语音信号和提取到的其他演员的纯净语音信号进行混合，生成新的混叠语音数据，加入到影视剧中。由此可见，在影视剧场景中，通过本申请实施例能够提供精确的语音分割，而且无需使用所有演员的大量数据进行专属分割网络的训练，可以快速高效的提取分割出纯净的演员语音。

可选的，应用场景为音视频创作场景。示例性地，音视频创作场景为针对音视频的二创场景（即针对已存在的音视频再次进行创作）。具体地：在二创场景中，用户喜欢提取指定演员在多个音视频中部分台词进行台词对话剪辑，即将指定演员在不同音视频中的语音数据剪辑到同一音视频中。这就会涉及提取该指定演员在多个音视频中的纯净语音信号；考虑到各音视频中的台词带有背景音乐或其他对象的语音信号，因此需要从音视频中剔除掉背景音乐得到该指定演员的纯净语音信号，以便于将提取到的多个纯净的语音信号进行融合，生成该指定演员对应的剪辑语音文件。

可选的，应用场景为会话场景。示例性地，会话场景为在线会议场景。具体地：在线会议场景中往往具有语音转录文本的需求，即将录音到的语音数据转换为文本形式；但是在多人参与的在线会议场景中，包含多人的语音信号的混叠语音数据的转录一直是一个难题，转录是指将多人中某个指定人的语音信号转换为文本的过程。采用本申请实施例可以先按照参与在线会议的每个对象的声纹特性，从混叠语音数据中分割提取出每个对象的语音信号，然后再分别将每个对象的语音信号输入至语音识别系统中实现文本转录，能够极高的提升会话语音混叠转录的准确性。

值得说明的是，本申请实施例提供的语音处理方案所适用的应用场景并不仅限于上述几种；并且随着应用场景的不同，承载语音处理方案的应用或平台有所不同。

还需说明的是，本申请实施例中相关数据收集处理应该严格根据相关法律法规的要求，获取个人信息需得到个人主体的知情或同意（或具备信息获取的合法性基础），并在法律法规及个人信息主体的授权范围内，开展后续数据使用及处理行为。例如，本申请实施例运用到具体产品或技术中时，如获取指定对象的参考语音数据时，需要获得该指定对象的许可或者同意，且相关数据的收集、使用和处理（如对象发布的弹幕的收集和发布等）需要遵守相关地区的相关法律法规和标准。

基于上述描述的语音处理方案，本申请实施例提出更为详细的语音处理方法，下面将结合附图对本申请实施例提出的语音处理方法进行详细介绍。

图 3 示出了本申请一个示例性实施例提供的一种语音处理方法的流程示意图；该语音处理方法可以由前述提及的系统中的计算机设备来执行，如计算机设备为终端和/或服务器；该语音处理方法可包括但不限于步骤 S301-S304：

S301：获取待分割的混叠语音数据。

S302：获取至少两个对象中的指定对象的参考语音数据。

步骤 S301-S302 中，待分割的混叠语音数据中包含至少两个对象中每个对象所产生的语音信号。例如，一首重金属音乐中包括由“歌手”产生的“歌词”语音信号，由“吉他”产生的“旋律”语音信号以及由“架子鼓”产生的“旋律”语音信号等；因此，确定该重金属音乐是混叠语音数据，该混叠语音数据中包含的对象分别为：歌手、吉他和架子鼓，该混叠语音数据中包含的语音信号分别为“歌手”产生的语音信号、“吉他”产生的语音信号和“架子鼓”产生的语音信号。

进一步的，如果用户具有从混叠语音数据中分离提取出某个对象的语音信号的需求，可以获取该某个对象的参考语音数据。此时，将该某个对象称为指定对象，该指定对象的参考语音数据和混叠语音数据不

同，但该参考语音数据中包括指定对象的纯净的参考语音信号；这样，指定对象的参考语音数据中包含的参考语音信号可以作为参考的信号，用于从混叠语音数据中包含的至少两个对象对应的至少两个语音信号中分离出指定对象的目标语音信号。下面对指定对象和参考语音数据进行简单介绍，其中：①指定对象可以是混叠语音数据中包含的至少两个对象中，用户想要提取语音信号的任一对象；由前述描述可知，对象可以是指人类、动物或实体设备；为便于阐述，以指定对象的对象类型为人类为例进行介绍，特在此说明。例如，上述重金属音乐的例子中，如果用户想要从嘈杂重金属音乐中提取出“歌手”产生的“歌词”，那么确定该“歌手”为指定对象，此时需要将重金属音乐中各乐器产生的语音信号和该“歌手”产生的语音信号进行分离，并提取出“歌手”的纯净的语音信号。

②参考语音数据是指包含该指定对象的参考语音信号的一段语音数据。为确保能够从参考语音数据中提取到指定对象的较为纯净地声纹特性，参考语音数据应当是包含指定对象的一段较为纯净地语音数据。例如，参考语音数据中只包含指定对象的参考语音信号；再如，参考语音数据中同时包含指定对象的参考语音信号和其他语音信号，但要确保容易从掺杂有其他语音信号的参考语音数据中提取指定对象的参考语音信号（如其他语音信号的信号频率较低，而指定对象的参考语音信号的信号频率相对较高等），这样有利于对纯净的参考语音数据进行分析，以提取到指定对象较为准确的声纹特性。本申请实施例对该参考语音数据的类型、时长和来源不作限定。示例性地：参考语音数据的类型可以包括但是不限于：指定对象朗读文章所产生的一段音频，指定对象说话所产生的一段音频或者指定对象清唱产生的一段音频等。参考语音数据的时长可以为几秒或十几秒等。参考语音数据的来源可以包括但是不限于：在指定对象与具有语音分离需求的用户为不同用户时，该参考语音数据可以是由指定对象发送给该用户的，或者，该用户通过某些途径（如历史的语音信息）下载或录音得到的；在指定对象与具有语音分离需求的用户为相同用户时，该参考语音数据可以是由指定对象实时录入的，即通过用户持有的终端中部署的麦克风实时采集的。

一种示例性地由用户输入指定对象的参考语音数据的界面示意图可以参见图 4；如图 4 所示，在用户持有的终端的终端屏幕中显示有语音获取界面 401，该语音获取界面 401 中包含关于参考语音数据的获取区域 402。详细地，在该获取区域 402 中可以显示至少两种语音获取入口，如采集入口 4021 和上传入口 4022。当采集入口 4021 被触发时，表示用户想要通过实时采集的方式输入指定对象（指定对象为该用户，或者指定对象和该用户处于同一物理环境）的参考语音数据，那么终端的麦克风被打开，以便于能够实时采集用户所处物理环境中的参考语音信号生成参考语音数据。当上传入口 4022 被触发时，表示用户想要通过上传文件的方式输入指定对象的参考语音数据，那么用户可以从存储空间（如终端的本地存储空间，云存储空间或者服务器存储空间等）中将关于指定对象的参考语音数据进行上传。

应当理解的是，语音获取界面的界面元素（如界面所包含的界面内容）和界面样式并不仅限于图 4 所示。例如，在语音获取界面中还可以显示混叠语音数据的上传入口，通过该上传入口用户可以实现更换待分割的混叠语音数据。再如，在语音获取界面中还可以添加文本转换控件（或称为组件、按键、选项等），这样用户可以在语音分离前或语音分离后通过触发该文本转换控件，实现将分离出的语音信号一键转换为文本形式，在一定程度上缩短文本转换路径，从而提升文本转换效率。

S303：从参考语音数据中提取指定对象的声纹表征向量，将混叠语音数据和声纹表征向量输入预设的语音分割模型，语音分割模型用于：基于注意力机制从混叠语音数据中分割出与指定对象的声纹特性相匹配的语音信号。

S304：基于所分割出的语音信号，生成指定对象的语音文件。

步骤 S303-S304 中，声纹(Voiceprint)是携带语音信息的声波频谱，是由波长、频率以及强度等多种特征维度组成的生物特征；声纹具有稳定性、可测量性和唯一性等特点，可以用来唯一标识对象的声音特点，即声纹可以用于表征对象的身份。因此，本申请实施例在获取到指定对象的较为纯净的参考语音数据后，支持从该参考语音数据中提取指定对象的声纹特性，以便于后续基于该唯一的声纹特性进行语音信号的分

离提取。

由前述图 1 所示的系统可知，通过声纹向量提取模型对参考语音数据进行分析，可以实现从参考语音数据中提取出用于表征指定对象的身份的声纹特性。在实际应用中，该声纹向量提取模型输出的是该指定对象的声纹表征向量（或简称为声纹向量），即声纹向量提取模型对参考语音数据进行分析，得到的是能够用于表征指定对象的声纹特性的声纹表征向量。进一步的，声纹向量提取模型在提取到指定对象的声纹表征向量后，可以创新性的采用向量嵌入的方式进行声纹信息表征的传递，将声纹表征向量输入到语音分割模型中参与注意力机制的计算，以便于从混叠语音数据中分割出与指定对象的声纹特性相匹配的目标语音信号。这种创新性的向量嵌入机制能够让语音分割模型不依赖于任何对象的历史语音数据进行额外的训练，只需要对少量的参考语音数据进行声纹表征向量的提取，能够摆脱对大规模语音数据的依赖，从而做到系统的高度可复用可移植，让整个系统更高效，并降低用户输入操作复杂度，提高系统的通用化。

本申请实施例提供的语音分割模型是采用注意力机制，对传统的语音分割网络进行改进得到的；具体是将注意力机制融合至传统的语音分割网络所得到的。其中，本申请实施例涉及的传统的语音分割网络为 Unet（或表示为 U-net、U-Net 等）网络，Unet 是使用全卷积网络进行语义分割的算法之一；其主要使用包含压缩路径和扩展路径的对称 U 形结构。

一种示例性地 Unet 网络的网络结构的示意图可以参见图 5。如图 5 所示，Unet 网络是一个 U 型对称网络结构，该对称网络结构中包括左右对称的特征提取子网络和上采样子网络，且特征提取子网络和上采样子网络之间通过卷积连接层进行连接。其中：①特征提取子网络可以简单理解为下采样层或者编码网络，其包含层级分布（上一层级的卷积层输出的特征图（Feature map）作为相邻下一层级的卷积层的输入）的 m（图 5 中 m=4）个卷积层（Convolutional layer），m 为正整数；层级分布的 m 个卷积层是指，m 个卷积层依次连接，m 个卷积层中任意相邻的两个卷积层中上一个卷积层作为上一层级的卷积层，下一个卷积层作为下一层级的卷积层，且上一层级的卷积层输出的特征图作为相邻下一层级的卷积层的输入。如图 5 所示，在每个卷积层之后可以部署池化函数；通过这种先采用卷积层中的卷积网络针对混叠语音数据执行特征提取后，再采用池化函数（pool）进一步抽取更高阶的特征的方式，有效保留混叠语音数据中想要突出的特征；其中，本申请实施例对池化函数的类型不作限定，如池化函数为最大池化（max pool），其倾向于卷积层输出的特征图中池化窗口（如窗口大小为 2*2）内的最大特征。②相应地，特征提取子网络和上采样子网络具有对称性，上采样子网络可以简单理解为解码网络，其包含特征提取子网络中每个卷积层对应的上采样层（up sampling layer）。如图 5 所示，卷积连接层和每个上采样层之后还部署有卷积核为 2*2 的转置卷积（up-Conv），以通过转置卷积实现上采样功能。由此可见，Unet 网络这种对称网络结构既可以从头实现网络并进行权重的初始化，然后进行模型的训练；也可以借用现有一些网络的卷积层结构（如 resnet（残差神经网络）中的 vgg（一种卷积网络））和对应的已训练好的权重文件，再加上后面的上采样层进行训练计算等；这样，在深度学习的模型训练中使用已有的权重模型文件，可以大大加快模型训练的速度。

进一步的，每个卷积层、卷积连接层和上采样层中包括顺序连接的多个卷积网络；如图 5 所示，特征提取子网络、卷积连接层和上采样层均可以包含三个卷积核为 3*3 的卷积网络。其中，卷积网络或称为卷积神经网络（Convolutional Neural Network，CNN）；卷积神经网络是一种前馈神经网络，主要由一个或多个卷积层和顶端的全连通层组成，同时也包括关联权重和池化层（pooling layer）。如图 5 所示，每个卷积网络之后可以部署激活函数，以通过激活函数为模型加入非线性因素，以便于训练好的模型能够解决线性模型所不能解决的问题。本申请实施例对激活函数的类型不作限定，如激活函数可以为 ReLu 函数（ReLuSigmoidTanh，线性整流函数）等。

更进一步的，Unet 网络还可以通过跳跃连接（skip-connection，或称为 copy and crop）有效结合高级特征图和低级特征图来得到最终的特征图。其中，跳跃连接的具体过程可以包括：特征提取子网络中的每个卷积层得到的特征图都会拼接（concatenate）到，上采样子网络中对应的上采样层中；从而实现对每层

特征图都有效使用到后续计算中。这种跳跃连接不同维度的特征图的方式，相比于未实现跳跃连接的网络结构而言，可以有效避免直接在高级特征图中进行监督和损失计算，有效结合低级特征图中的特征，从而使得最终所得到的特征图既包含了高维度的特征，也包含很多低维度的特征，实现了不同规模下特征的融合，提高模型的结果精确度。

上述图 5 对传统的 Unet 网络的网络结构进行了详细介绍，本申请实施例提供的语音分割模型是对该 Unet 网络的网络结构进行改进得到的。本申请实施例针对 Unet 网络的网络结构的改进主要包括：在 Unet 网络的网络结构中的全部或部分网络层（如卷积层、卷积连接层和上采样层）中融合注意力机制。也就是说，基于注意力机制对 Unet 网络改进得到的语音分割模型中全部或部分网络中融合有注意力机制。为便于阐述，以在 Unet 网络中的每个网络层中均加入 Attention 机制来作为语音分割模型为例，通过给每个网络层均加入注意力机制，能够将指定对象的声纹表征向量嵌入到 Unet 网络中的每一网络层，可以让网络中每一网络层能够深度感受该声纹表征向量所表征的声纹信息或声纹特性，从而让最终输出的语音信号更贴近指定对象，确保提取的语音信号更纯净。

示例性地，在 Unet 网络中的每个网络层中均加入 Attention 机制时所构建的语音分割模型的结构示意图可以参见图 6。如图 6 所示，改进得到的语音分割模型相比于 Unet 网络而言，基本的网络架构与原始的 Unet 网络架构相同，但是在 Unet 网络架构中的每一个层级中均加入一个注意力机制，且该注意力机制的输入信息为：指定对象的声纹表征向量和该注意力机制的上一层级所输出的特征图。通过让指定对象的声纹表征向量嵌入到每一网络层中，主要是与每一网络层的特征图进行注意力计算，能够让整个模型深度的感受学习到提取的声纹表征向量，从而能够让每一个层级的计算都朝着声纹表征向量进行靠拢，确保最终提取的语音信号是与声纹表征向量所表征的声纹特性相匹配的。值得注意的是，注意力机制在网络层对应的多个卷积网络中的融合位置是不固定的，图 6 所示的融合位置是示例性地。

基于上述图 6 对语音分割模型的网络结构的相关介绍，下面以语音分割模型中每个网络层均融合有注意力机制为例，对语音分割模型基于注意力机制，从混叠语音数据中分割出与指定对象的声纹特性相匹配的语音信号，并基于语音信号生成指定对象的语音文件的具体实施过程进行介绍；该过程可以包括但不限于步骤（1）-（4），其中：

（1）将混叠语音数据从时域转换至频域，得到混叠语音数据对应的语音频谱特征，即该语音频谱特征是混叠语音数据在频域上的特征表现。

考虑到 Unet 网络的输入信息属于频域，因此需要将混叠语音数据进行转换，得到该混叠语音数据对应的语音频谱特征；这样该语音频谱特征可以作为图片输入至语音分割网络中。其中，时域和频域是音频应用中常用的两个概念，也是衡量音频特征的两个维度概念；时域是通过将语音信号的采样点在时间上进行展示处理，即与时间进行相关绑定；频域是通过将语音信号在各个频带上进行能量分布的一种特征表现；通过转换公式（如傅里叶变换(Fourier Transform)、拉普拉斯变换(Laplace Transform)或者索尔兹变换(Z Transform)等）可以实现语音信号从时域转换至频域，或者从频域转换为时域。

（2）基于注意力机制对声纹表征向量和语音频谱特征进行相关度计算，得到与声纹特性相匹配的语音频谱特征分段。

如图 6 所示，在获取到声纹向量提取模型提取的指定对象的声纹表征向量后，将该声纹表征向量输入至语音分割网络中的每个网络层，具体是每个网络层中融合的注意力机制。这样，按照每个网络层中融合的注意力机制，可以对声纹表征向量和相应网络层的第一特征图进行相关度计算，得到相应网络层输出的第二特征图；值得注意的是，根据网络层在语音分割网络中的位置不同，该网络层的第一特征图也有不同，在后续实施例对此进行介绍。然后，可以将语音分割模型中第 2m+1 个网络层（即上采样子网络中的最后一个上采样层）输出的第二特征图，作为与声纹表征向量相匹配的语音频谱特征分段；该语音频谱特征分段具体是语音频谱特征中，与声纹特性相匹配的分段，即语音频谱特性中属于指定对象的分段。由此可见，

通过让用于表征指定对象的声纹特性的声纹表征向量嵌入至语音分割模型中的每一网络层以参与特征提取处理，可以让语音分割网络中的每一网络层都能够深度感受到该指定对象的声纹信息，从而让最终分割输出的语音信号更靠近指定对象的声纹特性，保证提取的语音信号更纯净且更精确。

不难理解的是，根据注意力机制在语音分割模型中的融合位置不同，特征提取子网络和上采样子网络中的网络层按照融合的注意力机制，对声纹表征向量和相应网络层的上一层级所输出的第一特征图进行相关度计算的具体实施过程有所不同。下面以语音分割模型中融合有注意力机制的任一网络层表示为目标网络层为例，目标网络层按照注意力机制进行相关度计算进行示例性说明，其中：

在一种实现方式中，如图 7a 所示，假设目标网络层为语音分割模型中的卷积层或卷积连接层，且注意力机制在目标网络层对应的多个卷积网络中的融合位置为：目标网络层包括的顺序连接的多个卷积网络中，首个卷积网络和与首个卷积网络相邻的第二个卷积网络之间的位置；也就是说，注意力机制在目标网络层对应的多个卷积网络中的融合位置为：目标网络层包括的顺序连接的多个卷积网络中的首个卷积网络，和顺序连接的多个卷积网络中与首个卷积网络相邻且位于首个卷积网络之后的卷积网络之间的位置。此实现方式下，该目标网络层对声纹表征向量和该目标网络层的第一特征图进行相关度计算，以得到该目标网络层输出的第二特征图的具体实施过程可以包括：首先，采用目标网络层中的首个卷积网络，对目标网络层的第一特征图进行特征提取处理，得到该目标网络层的第三特征图；此处特征提取处理是指从第一特征图中提取出有用的信息（即特征），以供后续的分类、聚类和回归等任务使用的过程；特征提取处理的过程可以包括：对第一特征图进行预处理（如去噪、归一化或标准化等处理），并对预处理后的第一特征图进行特征提取，以提取出有用的特征，并从提取出的特征中筛选具有代表性或区分度的特征，筛选后的特征作为特征提取处理后的特征。其中，该目标网络层为语音分割模型（具体是特征提取子网络）中层级分布的首个卷积层 701 时，该目标网络层的第一特征图为对混叠语音数据进行频域转换所得到的语音频谱特征；目标网络层为语音分割模型中除首个卷积层 701 外的其他卷积层（如卷积层 702）时，该目标网络层的第一特征图是对该目标网络层相邻的上一层级网络层（如卷积层 701）输出的特征图进行池化处理得到的；池化处理由目标网络层中的池化层执行的，池化处理旨在通过并行处理或数据压缩等手段，减小上一层级网络层输出的特征图的尺寸和参数量，从而降低计算量。然后，按照目标网络层中融合的注意力机制，对声纹表征向量和目标网络层的第三特征图进行相关度计算，得到该目标网络层的第四特征图；该目标网络层的第三特征图的特征维度和第四特征图的特征维度相同；其中，特征图（如第三特征图和第四特征图）可以表现为向量形式，因此特征图的特征维度可以是向量的维度，该向量中的每一个维度对应一个特征，也就是说，注意力机制计算前后的特征维度是相同的。最后，采用目标网络层中除首个卷积网络外的其他卷积网络对第四特征图进行特征提取处理，得到目标网络层输出的第二特征图。

由此可见，在目标网络层为语音分割模型中的卷积层或卷积连接层的情况下，在卷积层或卷积连接层所包含的多个卷积网络中融合注意力机制，这样在采用卷积层或卷积连接层中包含的多个卷积网络对混叠语音数据进行特征提取的过程中，能够基于注意力机制在混叠语音数据中聚焦与指定对象的声纹表征向量相匹配的特征，从而通过特征提取过程中的注意力机制，分析得到与声纹表征向量相匹配的第二特征图，进而基于第二特征图能够从混叠语音数据中准确地分割出指定对象的目标语音信号。

其他实现方式中，如图 7b 所示，假设目标网络层为语音分割模型中的上采样层，且注意力机制在目标网络层对应的多个卷积网络中的融合位置为：顺序连接的多个卷积网络中的最后一个卷积网络之后的位置。此实现方式下，该目标网络层对声纹表征向量和该目标网络层的第一特征图进行相关度计算，以得到该目标网络层输出的第二特征图的具体实施过程可以包括：首先，采用目标网络层中顺序连接的多个卷积网络对目标特征图进行特征提取处理，得到目标网络层的第一特征图。此处的目标特征图是将目标网络层对应的卷积层输出的特征图，和目标网络层的上一层级网络层输出的特征图进行特征拼接得到的；如图 7b 所示，上采样子网络中的首个上采样层 703 的输入信息为目标特征图，该目标特征图是该首个上采样层 703

的上一层级卷积连接层 704 输出的特征图，和该首个上采样层 703 对应的卷积层 705 输出的特征图进行特征拼接得到的。然后，采用目标网络层中融合的注意力机制，对声纹表征向量和目标网络层的第一特征图进行相关度计算，得到目标网络层输出的第二特征图；该第二特征图的特征维度和第一特征图的特征维度相同。

由此可见，在目标网络层为语音分割模型中的上采样层的情况下，在上采样层所包含的多个卷积网络中融合注意力机制，这样在采用上采样层中包含的多个卷积网络对上一层级网络层输出的特征图和对应的卷积层输出的特征图进行特征提取后，能够基于注意力机制在特征提取后的第一特征图中聚焦与指定对象的声纹表征向量相匹配的特征，从而分析得到与指定对象的声纹表征向量相匹配的第二特征图，进而基于第二特征图能够从混叠语音数据中准确地分割出指定对象的目标语音信号。

应当理解的是，上述图 7a 和图 7b 只是注意力机制分别融合至特征提取模块、卷积连接层和上采样子网络层中的一种示例性融合位置时，目标网络层进行相关度计算的示例性过程；在注意力机制融合至目标网络层中的不同融合位置时，目标网络层执行相关度计算的具体实施过程有所不同。

（3）将与声纹特性相匹配的语音频谱特征分段从频域转换至时域，得到与声纹特性相匹配的目标语音信号。

在基于前述步骤，从混叠语音数据对应的语音频谱特征中，提取到与指定对象的声纹特性相匹配的语音频谱特征分段后，还需要将该语音频谱特征分段从频域转换至时域，以得到符合数据传输格式的目标语音信号。其中，将语音频谱特征分段从频域转换至时域的方式可以包含但是不限于前述提及的傅里叶变换(Fourier Transform)、拉普拉斯变换(Laplace Transform)或者索尔兹变换(Z Transform)等等，对此不作限定。

（4）基于与声纹特性相匹配的目标语音信号生成指定对象的语音文件。其中，该语音文件的文件格式可以按照用户的个性化需求进行设置，本申请实施例对语音文件的文件格式不作限定。例如，语音文件为文本文件时，支持使用语音识别算法或工具将与指定对象的声纹特性相匹配的语音信号转换为文本，并对转换后的文本进行文本处理（如纠正拼错、符合添加和文本清晰等处理），然后将文本处理后的文本保存为文本格式（如.doc 格式）的文本文件。再如，语音文件为音频文件时，可以直接将与指定对象的声纹特性相匹配的语音信号，保存为音频格式（如.WAV 格式）的音频文件。

基于上述步骤（1）-（4）所阐述的语音信号提取过程可知，本申请实施例支持将混叠语音数据从时域转换到频域，得到混叠语音数据在频域上的语音频谱特征，这样将混叠语音数据转换为能够和声纹表征向量进行相关度计算且属于频域的语音频谱特征后，就可以利用语音分割模型中的每一网络层的注意力机制，对同属于频域的声纹表征向量和语音频谱特征进行相关度计算，确保相关度计算的可行性；考虑到最终想要分割得到的是时域上的信号，因而需要将语音频谱特征分段从频域转换到时域，得到与声纹特性相匹配的目标语音信号，确保最终提取出的是能够被设备理解和读取的时域信号。考虑到语音分割网络中每一网络层的特征维度不同，具体是每个网络层中各卷积网络的特征维度不同。因此，在将指定对象的声纹表征向量输入至语音分割网络中的每一网络层之前，还需要先采用一层网络层对声纹表征向量进行维度变换，得到维度变换后的声纹表征向量。其中，维度变换是指对声纹表征向量的特征维度进行变化，使得维度变换后的声纹表征向量的特征维度，与待输入至相应网络层中融合的注意力机制的特征图的特征维度相同，从而将维度变换后的声纹表征向量输入至语音分割模型时，语音分割模型才能对该声纹表征向量进行有效处理，避免维度不同造成的声纹表征向量的不可用性。其中，待输入至相应网络层中融合的注意力机制的特征图，可以是指前述描述的第三特征图。此外，注意力机制可以插入至网络层中顺序连接的多个卷积网络中的任意两个卷积网络之间；网络层的第一特征图可以是指该网络层中与注意力机制相邻且位于该注意力机制之前的卷积网络所输出的特征图。

综上所述，本申请实施例创新性的构建了一种全自动的语音处理方案，该方案基于声纹向量嵌入的方式实现语音信号的分割；对于用户而言，只需要输入指定对象的一小段参考语音数据和待分割的混叠语音

数据，就能够实现自动快速的从混叠语音数据从分离出指定对象的语音信号，能够极大的提升语音分割的效率，彻底摆脱人工的参与，形成快速的标准化。在本方案中，一方面，采用了声纹向量嵌入的方式，将用于表征指定对象的声纹特性的声纹表征向量输入至语音分割模型中参与 Attention 的计算，能够让语音分割模型不依赖于任何对象的历史语音数据进行额外的训练，能够摆脱对对象的大规模语音数据的依赖，从而做到系统的高度可复用可移植，让整个系统更通用化。另一方面，通过基于注意力机制对 Unet 网络进行改进，使得输入的声纹表征向量能够和 Unet 网络中的每一网络层进行 Attention 机制计算，从而确保语音分割模型输出的特征图和指定对象的声纹表征向量所表征的声纹特性更加贴合，避免提取出的指定对象的语音信号中包含有太多杂音，提高语音信号的纯净性，提高语音分割模型的分割准确性。

请参见图 8，图 8 示出了本申请一个示例性实施例提供的另一种语音处理方法的流程示意图；该语音处理方法可以由前述提及的系统中的计算机设备来执行，如计算机设备为终端和/或服务器；该语音处理方法可包括但不限于步骤 S801-S806：

S801：获取待分割的混叠语音数据。

S802：获取至少两个对象中的指定对象的参考语音数据。

需要说明的是，步骤 S801-S802 所示的具体实施过程，可以参见前述图 3 所示实施例中步骤 S301-S302 所示的具体实施过程的相关描述，在此不作赘述。

S803：对指定对象的参考语音数据进行短时相关分析。

S804：对所述指定对象的参考语音数据进行长时相关分析，得到指定对象的声纹表征向量。

步骤 S803-S804 中，为了能够从指定对象的参考语音数据中学习到指定对象较为清晰的声纹特性，本申请实施例支持结合短时相关和长时相关对参考语音数据进行分析，以提取能够充分表达指定对象的声纹特性的声纹表征向量。其中，针对参考语音数据的短时相关分析可以简单理解为：对参考语音数据中一段较短（如 20 毫秒）语音信号进行特征分析的过程；考虑到在较短时间内，参考语音数据中的语音信号通常是不发生变化的，因此对参考语音数据进行离散化后，可以利用每段较短语音信号在时域和频域的信息分布，提取该语音信号在该较短时间内的特征，从而实现对参考语音数据中每段语音信号的特征分析。简而言之，短时相关分析关注于对参考语音数据中分段的语音信号进行特征分析。不同的是，针对参考语音数据的长时相关分析可以简单理解为：对整个参考语音数据进行特征分析的过程；也就是说，长时相关分析关注于对参考语音数据的整个信号序列进行语义表达。

上述提及的短时相关分析和长时相关分析，主要是依赖于图 1 所示系统中的声纹向量提取模型实现的。正如前述所描述的，声纹向量提取模型中包含改进型的音频神经网络（PANNS）网络和转换器（Transformer）网络。其中：改进型的音频神经网络（PANNS）网络可以简称为改进型 PANNS，主要用于对参考语音数据进行短时相关分析；转换器（Transformer）网络可以简称为 Transformer 网络，主要用于对参考语音数据进行长时相关分析。

一种示例性地采用改进型 PANNS 和 Transformer 网络对参考语音数据进行短时相关分析和长时相关分析的示意图可以参见图 9。如图 9 所示，首先，将参考语音数据从时域转换至频域，得到该参考语音数据对应的参考语音频谱特征；其中，时域转换为频域的转换公式可以参见前述相关描述，在此不作赘述。然后，为实现短时相关分析，还可以对参考语音频谱特征进行分段处理，得到每个语音数据分段对应的参考语音频谱特征分段；本申请实施例涉及的参考语音频谱特征分段可以为对数梅尔频谱（Log-mel 或者 Logmel）。其中，mel（梅尔）频谱是一种基于人耳对等距（即频带等距离的分布在梅尔尺度上）的音高(pitch)变化的感官判断而确定的非线性频率刻度，音高是指声音的高低；在进行信号处理时，更能够迎合人耳的听觉感受变化来人为设定。

然后，采用分段输入的方式，将每个语音数据分段对应的参考语音频谱特征分段输入至改进型 PANNS

网络中，以及，将完整的参考语音数据输入至改进型 PANNS 中；这样，改进型 PANNS 可以按照参考语音频谱特征分段时遵循的分段规则，对接收到的参考语音数据进行分段处理，得到参考语音数据对应的多个语音数据分段。并且，改进型 PANNS 会分别基于每个语音数据分段和相应的参考语音频谱特征分段，对每个语音数据分段进行短时相关分析，得到每个语音数据分段对应的声纹语义特征向量；该声纹语义特征向量用于表征相应语音数据分段的语义特性。

最后，将每个语音数据分段对应的声纹语义特征向量所组成的向量序列（或称为声纹语义特征向量序列）输入至 Transformer 网络，这样 Transformer 网络可以对声纹语义特征向量序列进行长时相关分析，得到指定对象的声纹表征向量，此处的声纹语义特征向量序列中包括每个语音数据分段对应的声纹语义特征向量。其中，Transformer 网络作为一个序列网络，其输入是一个整体的向量序列，其输出也是一个序列（即指定对象的声纹表征向量为一个序列）；在 Transformer 网络输出一个序列后，可以将该序列中最后一个向量称为 state，该 state 包含了整体序列的语义融合；这样，可以将 Transformer 网络输出的整体序列求取平均（mean），并将平均结果和 state 进行叠加，生成一个综合了整体序列所表达的语义特征的声纹表征向量。上述这种综合 Transformer 网络输出序列的平均结果所表征的语音特征和整体序列所表征的语音特征，来得到声纹表征向量的方式，可以有效确保声纹表征向量能够较为充分地体现指定对象的声纹特性，确保声纹特性提取的准确性。

在上述对包含改进型 PANNS 和 Transformer 网络的声纹向量提取模型进行整体介绍的基础上，下面分别对改进型 PANNS 和 Transformer 网络的结构和流程进行介绍，其中：

（1）改进型 PANNS。

改进型 PANNS 的示例性结构示意图可以参见图 10；如图 10 所示，该改进型 PANNS 的输入信息为参考语音数据，即该改进型 PANNS 的输入使用的是原语音采样点序列，即音频信号的原始序列。该改进型 PANNS 可以分为两个支路，分别为时域支路（或称为时域处理支路）和频域支路（或称为频域处理支路）。其中：时域支路的输入信息是参考语音数据，频域支路的输入信息是参考语音数据对应的参考语音频谱特征；该参考语音频谱特征是将时域信号“参考语音数据”从时域转换至频域所得到的。进一步的，由前述描述可知，为了充分提取和表征指定对象的声纹特性，该改进型 PANNS 着重于对参考语音数据进行短时相关分析，即支持通过分段输入的方式采用改进型 PANNS 进行处理；具体是改进型 PANNS 每次只对参考语音数据中的一段语音数据和该语音数据对应的参考语音频谱特征进行处理。

也就是说，每次输入时域支路的输入信息是参考语音数据中的一段语音数据分段；同理，每次输入频域支路的输入信息是参考语音数据中的一段语音数据对应的在频域中的参考语音频谱特征分段。其中，将参考语音数据输入至改进型 PANNS 后，可以将该参考语音数据从时域转换至频域，得到参考语音数据对应的参考语音频谱特征，并对参考语音频谱特征进行分段处理，得到每个语音数据分段对应的参考语音频谱特征分段；以及，对参考语音数据进行分段处理，得到参考语音数据对应的多个语音数据分段。值得注意的是，参考语音数据进行分段处理所遵循的分段规则，和参考语音频谱特征进行分段处理所遵循的分段规则是相同的；例如，分段规则可以包括：参考语音数据进行分段处理时，是周期性的采集 20 毫秒时长的语音数据为一个分段，则参考语音频谱特征进行分段处理时，每个参考语音频谱特征分段转换至时域后对应的语音数据分段时长为 20 毫秒。通过这种时域分段和频域分段相对应的方式，确保每次特性分析的准确性。

继续参见图 10，在分段输入的情况下，改进型 PANNS 可以分别基于每个语音数据分段和相应的参考语音频谱特征分段，对每个语音数据分段进行短时相关分析，得到每个语音数据分段对应的声纹语义特征向量；该声纹语义特征向量用于表征相应语音数据分段的语义特性。为便于阐述，以多个语音数据分段中的任一语音数据分段表示为目标语音数据分段为例，对上述描述的短时相关分析的具体过程进行介绍，其中：

①时域支路中包含多个一维卷积层（Conv 1D）和最大池化层（Max pooling）；其中，Conv 为卷积 Convolutional 的缩写，D 为维度 Dimension 的缩写，且最大池化层的维度为 1，步长 stride（简写为 s）是一个一维向量，其长度为 4。如图 10 所示，假设时域支路中依次包括：一维卷积层➡一维卷积块（Conv 1D block，由一个或多个一维卷积层组成）➡最大池化层（维度为 1，且步长 s 为 4）➡一维卷积块➡最大池化层➡一维卷积块➡最大池化层；那么通过卷积层对目标语音数据分段进行特征提取处理后，再由该卷积层后相邻的最大池化层对特征提取处理后的特征图进行特征挑选，有利于进一步提取出特征图中想要关注的特征。通过多个一维卷积层和最大池化层之间层层递进的方式，可以对参考语音数据（具体是目标语音数据分段）进行多次特征提取，得到一维序列（resize），并将该一维序列通过维度变换（Reshape）得到多个二维的时域特征图（或称为二维图谱 Wavegram），时域特征图是在时域中描述参考语音数据中的参考语音信号随时间变化的图形表示，可以直观地展示参考语音信号的基本特征（如参考语音信号的周期性、频率成分和相位关系等）。此处的维度转换的目的是为了使得转换后的时域特征图能够和频域支路输出的频域特征图进行融合。上述过程中，通过在时域支路使用大量一维卷积层能够使得通过时域支路对参考语音数据进行特征提取处理时，直接学习到参考语音数据中语音信号的时域特性（如音频响度和采样点幅度这类信息）。

②频域支路中包含多个二维卷积层（Conv 2D）和最大池化层（Max pooling）；其中，频域支路的最大池化层的维度为 2。如图 10 所示，假设频域支路中依次包括：二维卷积块（由一个或多个二维卷积层组成）➡最大池化层（维度为 2）➡二维卷积块➡最大池化层➡二维卷积块；通过该多个二维卷积层和最大池化层可以对目标语音数据分段对应的参考语音频谱特征分段（Logmel）进行特征提取处理，得到多个频域特征图（Feature maps），频域特征图是在频域中描述参考语音数据中参考语音信号的频率成分的图形表示，可以显示可以参考语音信号中包含的各种频率成分及其对应的幅度或强度。值得注意的是，此处的频域特征图的特征维度和时域支路输出的时域特征图的特征维度相同。上述过程中，通过在频域支路使用大量二维卷积层能够使得通过频域支路对参考语音频谱特征分段进行特征提取处理时，直接学习到参考语音数据中语音信号的频域特性。

③在基于前述步骤得到时域支路输出的时域特征图（Wavegram），和频域支路输出的频域特征图（Feature map）后，可以将时域特征图和频域特征图进行融合处理，生成目标语音数据分段对应的声纹语义特征向量。详细地，如图 10 所示的在时域支路和频域支路之间还存在多次时域和频域之间的信息交流，分别是将时域支路的信息特征进行维度变换（Reshape），然后与频域支路的特征进行融合（concat），并将融合结果经过二维卷积块（Conv 2D block）卷积后输入到更高层的融合模块进行融合。考虑到时域处理中能够获得参考语音数据在时序上的关联性，而频域处理中可以获得参考语音数据在不同频率上的关联性，两者属于不同领域；那么通过上述描述的两个域（时域和频域）之间的信息交互，能够实现时域和频域保持信息上的互补，使得高层网络能够感知到底层网络信息，从而实现针对参考语音数据的充分学习。

基于此，假设在时域支路和频域支路进行的特征提取处理的次数为 k，k 为大于 1 的整数，且任一次特征提取处理表示为第 i 次特征提取处理，那么上述提及的时域支路输出的时域特征图和频域支路输出的频域特征图之间的融合具体可以包括：首先，当 i=1 时（即首次特征提取处理时），将时域支路在第 1 次特征提取处理得到的中间时域特征图，和频域支路在第 1 次特征提取处理得到的中间频域特征图进行融合处理，生成第 1 次特征提取处理后的第一中间特征向量。然后，当 $1<i\leqslant k$ 时，将时域支路在第 i 次特征提取处理得到的中间时域特征图，频域支路在第 i 次特征提取处理得到的中间频域特征图，以及第 i-1 次特征提取处理得到的第 i-1 中间特征向量进行融合处理，生成第 i 次特征提取处理后的第 i 中间特征向量。最后，基于 i=k 时第 k 次特征提取处理后的第 k 中间特征向量，生成目标语义数据分段对应的声纹语义特征向量。其中，此处生成的具体过程可以包括：将第 k 中间特征向量输入到二维卷积神经网络（2D CNN layers）中进行特征提取，得到向量序列，并采用该向量序列进行平均运算求取平均值（mean）以及进行最大值运算

求取最大值（max），然后将求得的平均值与最大值进行相加（sum）后再经过一层激活函数（Rule）得到特征向量（vector），并采用归一化函数（softmax）对该特征向量进行归一化处理，以将特征向量转换为表示概率分布的声纹语义特征向量，即目标语音数据分段对应的声纹语义特征向量。

综上所述，整个参考语音数据进行分帧并分别输入到改进型 PANNS 中，该改进型 PANNS 选择使用整条序列的最后一个向量和整条序列的均值一起融合，从而生成最后的声纹表征向量，能够得到代表整个参考语音数据的多频带语义特征向量序列（包括每个语音数据分段对应的声纹语义特征向量）。通过该改进型 PANNS 能够实现针对参考语音数据的短时相关分析，充分学习参考语音数据中各分段语音信号的声纹特性，能够让提取的声纹表征向量充分表达指定对象的声纹信息。

（2）Transformer 网络。

考虑到改进型 PANNS 关注于短时相关性，也就是特征是按照分段进行计算的；因此为了能够充分从参考语音数据中充分学习指定对象的声纹特性，本申请实施例还引入 Transformer 网络来对参考语音数据进行长时关联信息的学习。Transformer 网络拥有的注意力机制能够更好地关注到参考语言数据关于指定对象的声纹特性，较好地实现针对参考语言数据的全局特征信息的提取。

示例性地，Transformer 网络的网络结构示意图可以参见图 11；如图 11 所示，Transformer 网络采用了 encoder（编码）-decoder（解码）架构。其中，encoder 侧和 decoder 侧均由 N 个 encoder 层（即图 11 中的“N×”）堆叠形成。encoder 层主要包含两个子层，其中：第一个子层中包含多头注意力机制（Multi-Head Attention）、残差和归一化（Add&Norm）；第二个子层中包含前馈神经网络（Feed Forward）、残差和归一化的层。其中，第一个子层所包含的多头注意力机制能够帮助获取到语音数据分段对应的声纹语义特征向量的上下文语义。decoder 层主要包含三个子层，其中：第一个子层中包含遮罩多头注意力机制（Masked Multi-Head Attention）、残差和归一化的层，第一个子层中包含多头注意力机制（Multi-Head Attention）、残差和归一化（Add&Norm）的层，第一个子层中包含前馈神经网络（Feed Forward）、残差和归一化的层；解码层通过这种三层结构能够帮助获取到需要关注的重点内容。

具体实现中，Transformer 网络中的 encoder 侧的输入信息是改进型 PANNS 输出的声纹语义特征向量序列（Input Embedding），该声纹语义特征向量序列中包含每个语义数据分段对应的声纹语义特征向量。然后，对输入的声纹语义特征向量序列进行位置编码（Positional Encoding），以实现对该声纹语义特征向量序列的预处理；其中，位置编码是一种用词的位置信息对向量序列中的每个词进行二次表示的方法，能够让输入至编码侧的数据携带词的位置信息。进一步的，将位置编码后的数据输入至编码侧，由编码侧中的各层（如前述对编码侧结构的相关描述）对输入的数据进行编码处理，得到编码结果。然后，解码侧在获取到编码侧输出的编码结果后，支持结合该编码结果和解码侧在之前时刻的输出特征（shifted right），作为解码侧此时的输入数据，并由解码侧对输入数据进行解码处理。最后，解码侧的输出特征经过线性变换（Linear）和分类（softmax）后，得到指定对象的声纹表征向量（Output Probabilities）。通过上述过程可以让声纹语义特征向量序列经过 Transformer 网络计算，能实现让整个声纹语义特征向量序列对整个参考语音数据的长时语义表达更为清晰，即最终输出的声纹表征向量能够充分且清晰地表达指定对象的声纹特性。

S805：将混叠语音数据和声纹表征向量输入至预设的语音分割模型，该语音分割模型用于：基于注意力机制从混叠语音数据中分割出与指定对象的声纹特性相匹配的目标语音信号。

需要说明的是，步骤 S805 所示的具体实施过程，可以参见前述图 3 所示实施例中，步骤 S302 中关于基于注意力机制从混叠语音数据中分割出与指定对象的声纹特性相匹配的语音信号这部分具体实施过程的相关描述，在此不作赘述。

由前述步骤 S302 所示的相关描述可知，本申请实施例主要是采用改进后的 Unet 网络（即语音分割网络）来实现基于注意力机制对混叠语音数据进行分割，以得到与指定对象的声纹特性相匹配的语音信号。

并且，该语音分割网络主要是将注意力机制融合至传统的Unet网络中，如在传统的Unet网络中的每一网络层中均加入注意力机制实现针对传统的 Unet 网络的改进。考虑到语音分割系统中可能会引入较多的Attention机制从而引起整个模型的参数量较大，因此本申请实施例还支持对融合了注意力机制的语音分割模型进行模型蒸馏，得到模型蒸馏后的语音分割模型；这样，前述提及的相关度计算可以由模型蒸馏后的语音分割模型实现，以此缩小整个系统的规模，降低整体参数量和耗时。声纹表征向量和语音频谱特征均表现为向量形式，那么模型蒸馏后的语音分割模型对声纹表征向量和语音频谱特征进行相关度计算的方式可以包括但是不限于点积方式；其中，通过点积方式计算声纹表征向量和语音频谱特征的相关度包括：声纹表征向量的模和语音频谱特征的模的乘积，与声纹表征向量和语音频谱特征夹角的余弦值的乘积；如声纹表征向量为$\vec{a}$，语音频谱特征为$\vec{b}$，声纹表征向量和语音频谱特征的夹角为θ，那么声纹表征向量和语音频谱特征的点积为$|\vec{a}||\vec{b}|\cos\theta$，该点积结果作为声纹表征向量和语音频谱特征的相似度。其中，模型蒸馏是一种对参数量较多的大模型(teacher model)进行学习以得到参数量较小的更为紧凑的小模型(student model)的方法。针对语音分割模型的模型蒸馏主要通过剪枝（pruning）和知识蒸馏等技术实现针对大模型的模型蒸馏；其中：剪枝称为模型剪枝（Model Pruning）是一种模型压缩技术，旨在通过删除语音分割模型中的一些不重要的参数或结构，从而减少语音分割模型的复杂度，提高语音分割模型的推理速度，并减少语音分割模型的存储需求。知识蒸馏可以理解为将相对复杂的语音分割模型作为一个教师模型，训练一个参数量较小且结构简单的学生模型，并训练过程学生模型学习和模仿教师模型的输出，使得训练好的学生模型不仅具有计算量小的优势，而且具有和教师模型相同的模型性能；在本申请实施例中，训练好的学习模型为知识增量后的语音分割模型。上述提及的模型剪枝和知识蒸馏是模型蒸馏的主要实现技术，但模型蒸馏还可以包括其他技术，本申请实施例对模型蒸馏的具体实施方式不作限定。

S806：基于所分割出的目标语音信号，生成指定对象的语音文件。

需要说明的是，步骤S806所示的具体实施过程，可以参见前述图3所示实施例中步骤S304所示的具体实施过程的相关描述，在此不作赘述。

综上所述，本申请实施例提供了一种声纹向量嵌入的可复用性指定对象的语音分割方法，该方法采用改进型PANNS和Transformer网络所组成的声纹向量提取模型，来对指定对象的一小段参考语音数据进行声纹提取，通过结合改进型PANNS的短时相关分析和transformer网络的长时相关分析，能够使得提取到的指定对象的声纹表征向量更为充分和清晰地表达出指定对象的声纹特性。此外，本申请实施例还通过将注意力机制融合至Unet网络中，使得融合了注意力机制的Unet网络（即语音分割模型）基于注意力机制来实现混叠语音数据的分割，能够更清晰准确的计算提取出混叠语音数据中指定对象的语音信号，确保提取的语音信号的纯净，实现更准确纯净的语音分离效果。此外，本申请实施例只需获取指定对象的参考语音数据，就能实现从混叠语音数据中分割出该指定对象的语音信号；如果想要获取其他对象的语音数据，则只需更换待分割的对象的声纹特性即可，不需要针对每个对象训练一个专属网络，能够做到方案的便捷可迁移，使得本方案具有极高的通用性。

上述详细阐述了本申请实施例的方法，为了便于更好地实施本申请实施例的上述方案，相应地，下面提供了本申请实施例的装置。

图12示出了本申请一个示例性实施例提供的一种语音处理装置的结构示意图；该语音处理装置可以用于执行图3或图8所示的方法实施例中的部分或全部步骤。请参见图12，该语音处理装置包括如下单元：

获取单元1201，用于获取混叠语音数据，混叠语音数据中包含至少两个对象中每个对象产生的语音信号；

获取单元1201，还用于获取指定对象的参考语音数据；指定对象是指至少两个对象中的任一个；参考语音数据中包含指定对象的参考语音信号；

处理单元 1202，用于从参考语音数据中提取指定对象的声纹表征向量，所述声纹表征向量用于表征指定对象的声纹特性；

处理器 1202，还用于将混叠语音数据和声纹表征向量输入预设的语音分割模型，所述语音分割模型用于：基于注意力机制从混叠语音数据中分割出与声纹特性相匹配的目标语音信号；

处理单元 1202，还用于基于所分割出的目标语音信号，生成指定对象的语音文件。

在一种实现方式中，基于注意力机制从混叠语音数据中分割出与声纹特性相匹配的目标语音信号的过程，包括：

将混叠语音数据从时域转换至频域，得到混叠语音数据对应的语音频谱特征；语音频谱特征是混叠语音数据在频域上的特征表现；

基于注意力机制对声纹表征向量和语音频谱特征进行相关度计算，得到与声纹特性相匹配的语音频谱特征分段；所述语音频谱特征分段是所述语音频谱特征中，与所述声纹特性相匹配的分段；

将语音频谱特征分段从频域转换至时域，得到与声纹特性相匹配的目标语音信号。

在一种实现方式中，相关度计算是通过语音分割模型实现的；语音分割模型中包括特征提取子网络和上采样子网络，特征提取子网络和上采样子网络之间通过卷积连接层进行连接；

特征提取子网络和上采样子网络具有对称性；特征提取子网络中包含层级分布的 m 个卷积层，上采样子网络中包含 m 个卷积层中每个卷积层对应的上采样层，m 为正整数；卷积层、卷积连接层和上采样层中均包括顺序连接的多个卷积网络；

其中，语音分割模型中的全部或部分网络层中融合有注意力机制，注意力机制在网络层包括的多个卷积网络中的融合位置不固定；网络层包括卷积层、上采样层和卷积连接层。

在一种实现方式中，语音分割模型中每个网络层均融合有注意力机制；处理单元 1202，用于基于注意力机制对声纹表征向量和语音频谱特征进行相关度计算，得到与声纹特性相匹配的语音频谱特征分段时，具体用于：

将声纹表征向量输入至语音分割模型中的每个网络层；

基于每个网络层中融合的注意力机制，对声纹表征向量和相应网络层的第一特征图进行相关度计算，以得到相应网络层输出的第二特征图；第二特征图所表征的声纹特性和声纹表征向量所表征的声纹特性相匹配；

将语音分割模型中第 2m+1 个网络层输出的第二特征图，作为与声纹表征向量所表征的声纹特性相匹配的语音频谱特征分段；语音分割模型中的第 2m+1 个网络层为上采样子网络中的最后一个上采样层。

在一种实现方式中，语音分割模型中融合有注意力机制的任一网络层表示为目标网络层；目标网络层为卷积层或者卷积连接层；注意力机制在目标网络层中的融合位置为：目标网络层包括的顺序连接的多个卷积网络中，首个卷积网络和与首个卷积网络相邻的第二个卷积网络之间的位置；

处理单元 1202，用于基于每个网络层中融合的注意力机制，对声纹表征向量和相应网络层的第一特征图进行相关度计算，以得到相应网络层输出的第二特征图时，具体用于：

采用目标网络层中的首个卷积网络，对目标网络层的第一特征图进行特征提取处理，得到目标网络层的第三特征图；其中，目标网络层为特征提取子网络中层级分布的首个卷积层时，目标网络层的第一特征图为语音频谱特征；目标网络层为语音分割模型中除首个卷积层外的其他卷积层时，目标网络层的第一特征图是对与目标网络层相邻的上一层级网络层输出的特征图进行池化处理得到的；

按照目标网络层中融合的注意力机制，对声纹表征向量和目标网络层的第三特征图进行相关度计算，得到目标网络层的第四特征图；第三特征图的特征维度和第四特征图的特征维度相同；

采用目标网络层中除首个卷积网络外的其他卷积网络对第四特征图进行特征提取处理，得到目标网络层输出的第二特征图。

在一种实现方式中，语音分割模型中融合有注意力机制的任一网络层表示为目标网络层；目标网络层为上采样层；注意力机制在目标网络层对应的多个卷积网络中的融合位置为：目标网络层中顺序连接的多个卷积网络中的最后一个卷积网络之后的位置；

处理单元1202，用于基于每个网络层中融合的注意力机制，对声纹表征向量和相应网络层的第一特征图进行相关度计算，以得到相应网络层输出的第二特征图时，具体用于：

采用目标网络层中顺序连接的多个卷积网络，对目标特征图进行特征提取处理，得到目标网络层的第一特征图；目标特征图是将目标网络层在特征提取子网络中对应的卷积层输出的特征图，和目标网络层的上一层级网络层输出的特征图进行特征拼接得到的；

采用目标网络层中融合的注意力机制，对声纹表征向量和目标网络层的第一特征图进行相关度计算，得到目标网络层输出的第二特征图；第二特征图的特征维度和第一特征图的特征维度相同。

在一种实现方式中，处理单元1202，还用于：

对声纹表征向量进行维度变换，得到维度变换后的声纹表征向量；

其中，维度变换后的声纹表征向量的特征维度，与待输入至相应网络层中融合的注意力机制的特征图的特征维度相同。

在一种实现方式中，若语音分割模型中融合有注意力机制的网络层的数量大于数量阈值，则处理单元1202，还用于：

对语音分割模型进行模型蒸馏，得到模型蒸馏后的语音分割模型；

其中，相关度计算由模型蒸馏后的语音分割模型实现。

在一种实现方式中，处理单元1202，用于从参考语音数据中提取指定对象的声纹表征向量时，具体用于：

对参考语音数据进行分段处理，得到参考语音数据对应的多个语音数据分段；

将参考语音数据从时域转换至频域，得到参考语音数据对应的参考语音频谱特征；

对参考语音频谱特征进行分段处理，得到每个语音数据分段对应的参考语音频谱特征分段；

分别基于每个语音数据分段和相应的参考语音频谱特征分段，对每个语音数据分段进行短时相关分析，得到每个语音数据分段对应的声纹语义特征向量；声纹语义特征向量用于表征语音数据分段的语义特性；

对声纹语义特征向量序列进行长时相关分析，得到指定对象的声纹表征向量；声纹语义特征向量序列中包括每个语音数据分段对应的声纹语义特征向量。

在一种实现方式中，多个语音数据分段中的任一语音数据分段表示为目标语音数据分段；处理单元1202，用于分别基于每个语音数据分段和相应的参考语音频谱特征分段，对每个语音数据分段进行短时相关分析，得到每个语音数据分段对应的声纹语义特征向量时，具体用于：

对目标语音数据分段进行特征提取处理，得到时域特征图；

对目标语音数据分段对应的参考语音频谱特征分段进行特征提取处理，得到频域特征图；

将时域特征图和频域特征图进行融合处理，生成目标语音数据分段对应的声纹语义特征向量。

在一种实现方式中，特征提取处理的次数为k次，k为大于1的整数；任一次特征提取处理表示为第i次特征提取处理；处理单元1202，用于将时域特征图和频域特征图进行融合处理，生成目标语音数据分段对应的声纹语义特征向量时，具体用于：

当i=1时，将第1次特征提取处理得到的中间时域特征图和中间频域特征图进行融合处理，生成第1次特征提取处理后的第一中间特征向量；

当$1<i\leqslant k$时，将第i次特征提取处理得到的中间时域特征图和中间频域特征图，以及第i-1次特征提取处理得到的第i-1中间特征向量进行融合处理，生成第i次特征提取处理后的第i中间特征向量；

基于i=k时第k次特征提取处理后的第k中间特征向量，生成目标语音数据分段对应的声纹语义特征

向量。

根据本申请的一个实施例，图 12 所示的语音处理装置中的各个单元可以分别或全部合并为一个或若干个另外的单元来构成，或者其中的某个（些）单元还可以再拆分为功能上更小的多个单元来构成，这可以实现同样的操作，而不影响本申请的实施例的技术效果的实现。上述单元是基于逻辑功能划分的，在实际应用中，一个单元的功能也可以由多个单元来实现，或者多个单元的功能由一个单元实现。在本申请的其它实施例中，该语音处理装置也可以包括其它单元，在实际应用中，这些功能也可以由其它单元协助实现，并且可以由多个单元协作实现。根据本申请的另一个实施例，可以通过在包括中央处理单元（CPU）、随机存取存储介质（RAM）、只读存储介质（ROM）等处理元件和存储元件的例如计算机的通用计算设备上运行能够执行如图 3 及图 8 所示的相应方法所涉及的各步骤的计算机程序（包括程序代码），来构造如图 12 中所示的语音处理装置，以及来实现本申请实施例的语音处理方法。计算机程序可以记载于例如计算机可读记录介质上，并通过计算机可读记录介质装载于上述计算设备中，并在其中运行。

基于同一发明构思，本申请实施例中提供的语音处理装置解决问题的原理与有益效果与本申请方法实施例中语音处理方法解决问题的原理和有益效果相似，可以参见方法的实施的原理和有益效果，为简洁描述，在这里不再赘述。

图 13 示出了本申请一个示例性实施例提供的一种计算机设备的结构示意图。请参见图 13，该计算机设备包括处理器 1301、通信接口 1302 以及计算机可读存储介质 1303。其中，处理器 1301、通信接口 1302 以及计算机可读存储介质 1303 可通过总线或者其它方式连接。其中，通信接口 1302 用于接收和发送数据。计算机可读存储介质 1303 可以存储在计算机设备的存储器中，计算机可读存储介质 1303 用于存储计算机程序，处理器 1301 用于执行计算机可读存储介质 1303 存储的计算机程序。处理器 1301（或称 CPU（Central Processing Unit，中央处理器））是计算机设备的计算核心以及控制核心，其适于实现一条或多条计算机程序，具体适于加载并执行一条或多条计算机程序从而实现相应方法流程或相应功能。

本申请实施例还提供了一种计算机可读存储介质（Memory），计算机可读存储介质是计算机设备中的记忆设备，用于存放程序和数据。可以理解的是，此处的计算机可读存储介质既可以包括计算机设备中的内置存储介质，当然也可以包括计算机设备所支持的扩展存储介质。计算机可读存储介质提供存储空间，该存储空间存储了计算机设备的处理系统。并且，在该存储空间中还存放了适于被处理器 1301 加载并执行的一条或多条计算机程序。需要说明的是，此处的计算机可读存储介质可以是高速 RAM 存储器，也可以是非不稳定的存储器（non-volatile memory），例如至少一个磁盘存储器；可选的，还可以是至少一个位于远离前述处理器的计算机可读存储介质。

在一个实施例中，该计算机设备可以是前述实施例提到的终端或服务器；该计算机可读存储介质中存储有一条或多条计算机程序；由处理器 1301 加载并执行计算机可读存储介质中存放的一条或多条计算机程序，以实现上述语音处理方法实施例中的相应步骤；具体实现中，计算机可读存储介质中的一条或多条计算机程序，由处理器 1301 加载并执行本申请各实施例的步骤；其中，本申请各实施例的步骤可以参见前述各实施例的相关描述，在此不作赘述。

基于同一发明构思，本申请实施例中提供的计算机设备解决问题的原理与有益效果与本申请方法实施例中语音处理方法解决问题的原理和有益效果相似，可以参见方法的实施的原理和有益效果，为简洁描述，在这里不再赘述。

本申请实施例还提供一种计算机程序产品，该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，实现上述语音处理方法。

本领域普通技术人员可以意识到，结合本申请中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，

取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用，使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时，可以全部或部分地以计算机程序产品的形式实现。计算机程序产品包括计算机程序（一个或多个）。在计算机设备上加载和执行计算机程序时，计算机程序执行本申请实施例上述的流程或功能。计算机设备可以是通用计算机、专用计算机、计算机网络、或者其他可编程设备。计算机程序可以存储在计算机可读存储介质中，或者通过计算机可读存储介质进行传输。计算机程序可以从一个网站站点、计算机设备、服务器或数据中心通过有线（例如，同轴电缆、光纤、数字用户线（DSL））或无线（例如，红外、无线、微波等）方式向另一个网站站点、计算机设备、服务器或数据中心进行传输。计算机可读存储介质可以是计算机设备能够存取的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。可用介质可以是磁性介质(例如，软盘、硬盘、磁带)、光介质(例如，DVD)、或者半导体介质（例如，固态硬盘(Solid State Disk，SSD)）等。

以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

权利要求书

1、一种语音处理方法，其特征在于，所述方法由计算机设备执行，所述方法包括：

获取混叠语音数据，所述混叠语音数据中包含至少两个对象中每个所述对象产生的语音信号；

获取指定对象的参考语音数据，所述指定对象是指所述至少两个对象中的任一个；所述参考语音数据中包含所述指定对象的参考语音信号；

从所述参考语音数据中提取所述指定对象的声纹表征向量，所述声纹表征向量用于表征所述指定对象的声纹特性；

将所述混叠语音数据和所述声纹表征向量输入预设的语音分割模型，所述语音分割模型用于：基于注意力机制从所述混叠语音数据中分割出与所述声纹特性相匹配的目标语音信号；

基于所分割出的所述目标语音信号，生成所述指定对象的语音文件。

2、如权利要求 1 所述的方法，其特征在于，基于注意力机制从所述混叠语音数据中分割出与所述声纹特性相匹配的目标语音信号的过程，包括：

将所述混叠语音数据从时域转换至频域，得到所述混叠语音数据对应的语音频谱特征；所述语音频谱特征是所述混叠语音数据在所述频域上的特征表现；

基于注意力机制对所述声纹表征向量和所述语音频谱特征进行相关度计算，得到与所述声纹特性相匹配的语音频谱特征分段；所述语音频谱特征分段是所述语音频谱特征中，与所述声纹特性相匹配的分段；

将所述语音频谱特征分段从所述频域转换至所述时域，得到与所述声纹特性相匹配的所述目标语音信号。

3、如权利要求 1 或 2 所述的方法，其特征在于，所述相关度计算是通过所述语音分割模型实现的；所述语音分割模型中包括特征提取子网络和上采样子网络，所述特征提取子网络和所述上采样子网络之间通过卷积连接层进行连接；

所述特征提取子网络和所述上采样子网络具有对称性；所述特征提取子网络中包含层级分布的 m 个卷积层，所述上采样子网络中包含所述 m 个卷积层中每个所述卷积层对应的上采样层，m 为正整数；所述卷积层、所述卷积连接层和所述上采样层中均包括顺序连接的多个卷积网络；

其中，所述语音分割模型中的全部或部分网络层中融合有所述注意力机制，所述注意力机制在所述网络层包括的多个卷积网络中的融合位置不固定；所述网络层包括所述卷积层、所述上采样层和所述卷积连接层。

4、如权利要求 1-3 任一项所述的方法，其特征在于，所述语音分割模型中每个所述网络层均融合有所述注意力机制；所述基于注意力机制对所述声纹表征向量和所述语音频谱特征进行相关度计算，得到与所述声纹特性相匹配的语音频谱特征分段，包括：

将所述声纹表征向量输入至所述语音分割模型中的每个所述网络层；

基于每个所述网络层中融合的所述注意力机制，对所述声纹表征向量和相应所述网络层的第一特征图进行相关度计算，以得到相应所述网络层输出的第二特征图；所述第二特征图所表征的声纹特性和所述声纹表征向量所表征的声纹特性相匹配；

将所述语音分割模型中第 2m+1 个网络层输出的第二特征图，作为与所述声纹表征向量所表征的声纹特性相匹配的语音频谱特征分段；所述第 2m+1 个网络层为所述上采样子网络中的最后一个上采样层。

5、如权利要 1-4 任一项所述的方法，其特征在于，所述语音分割模型中融合有所述注意力机制的任一网络层表示为目标网络层；所述目标网络层为所述卷积层或者所述卷积连接层；所述注意力机制在所述目标网络层中的融合位置为：所述目标网络层包括的顺序连接的多个所述卷积网络中，首个所述卷积网络和与首个所述卷积网络相邻的第二个所述卷积网络之间的位置；

所述基于每个所述网络层中融合的所述注意力机制，对所述声纹表征向量和相应所述网络层的第一特征图进行相关度计算，以得到相应所述网络层输出的第二特征图，包括：

采用所述目标网络层中的首个所述卷积网络，对所述目标网络层的第一特征图进行特征提取处理，得到所述目标网络层的第三特征图；其中，所述目标网络层为所述特征提取子网络中层级分布的首个卷积层时，所述目标网络层的第一特征图为所述语音频谱特征；所述目标网络层为所述语音分割模型中除首个所述卷积层外的其他所述卷积层时，所述目标网络层的第一特征图是对与所述目标网络层相邻的上一层级所述网络层输出的特征图进行池化处理得到的；

按照所述目标网络层中融合的所述注意力机制，对所述声纹表征向量和所述目标网络层的第三特征图进行相关度计算，得到所述目标网络层的第四特征图；所述第三特征图的特征维度和所述第四特征图的特征维度相同；

采用所述目标网络层中除首个所述卷积网络外的其他卷积网络对所述第四特征图进行特征提取处理，得到所述目标网络层输出的第二特征图。

6、如权利要求 1-5 任一项所述的方法，其特征在于，所述语音分割模型中融合有注意力机制的任一网络层表示为目标网络层；所述目标网络层为所述上采样层；所述注意力机制在所述目标网络层中的融合位置为：所述目标网络层中顺序连接的多个卷积网络中的最后一个卷积网络之后的位置；

所述基于每个所述网络层中融合的所述注意力机制，对所述声纹表征向量和相应所述网络层的第一特征图进行相关度计算，以得到相应所述网络层输出的第二特征图，包括：

采用所述目标网络层中顺序连接的多个卷积网络，对目标特征图进行特征提取处理，得到所述目标网络层的第一特征图；所述目标特征图是将所述目标网络层在所述特征提取子网络中对应的卷积层输出的特征图，和所述目标网络层的上一层级网络层输出的特征图进行特征拼接得到的；

采用所述目标网络层中融合的注意力机制，对所述声纹表征向量和所述目标网络层的第一特征图进行相关度计算，得到所述目标网络层输出的第二特征图；所述第二特征图的特征维度和所述第一特征图的特征维度相同。

7、如权利要求 1-6 任一项所述的方法，其特征在于，所述基于每个所述网络层中融合的所述注意力机制，对所述声纹表征向量和相应所述网络层的第一特征图进行相关度计算，以得到相应所述网络层输出的第二特征图之前，还包括：

对所述声纹表征向量进行维度变换，得到维度变换后的所述声纹表征向量；

其中，维度变换后的所述声纹表征向量的特征维度，与待输入至相应所述网络层中融合的注意力机制的特征图的特征维度相同。

8、如权利要求 1-7 任一项所述的方法，其特征在于，若所述语音分割模型中融合有所述注意力机制的网络层的数量大于数量阈值，则所述方法还包括：

对所述语音分割模型进行模型蒸馏，得到模型蒸馏后的语音分割模型；

其中，所述相关度计算由模型蒸馏后的所述语音分割模型实现。

9、如权利要求 1-8 任一项所述的方法，其特征在于，所述从所述参考语音数据中提取所述指定对象的声纹表征向量，包括：

对所述参考语音数据进行分段处理，得到所述参考语音数据对应的多个语音数据分段；

将所述参考语音数据从时域转换至频域，得到所述参考语音数据对应的参考语音频谱特征；

对所述参考语音频谱特征进行分段处理，得到每个所述语音数据分段对应的参考语音频谱特征分段；

分别基于每个所述语音数据分段和相应的所述参考语音频谱特征分段，对每个所述语音数据分段进行短时相关分析，得到每个所述语音数据分段对应的声纹语义特征向量；所述声纹语义特征向量用于表征所述语音数据分段的语义特性；

对声纹语义特征向量序列进行长时相关分析，得到所述指定对象的声纹表征向量；所述声纹语义特征向量序列中包括每个所述语音数据分段对应的声纹语义特征向量。

10、如权利要求 1-9 任一项所述的方法，其特征在于，多个所述语音数据分段中的任一语音数据分段表示为目标语音数据分段；所述分别基于每个所述语音数据分段和相应的所述参考语音频谱特征分段，对每个所述语音数据分段进行短时相关分析，得到每个所述语音数据分段对应的声纹语义特征向量，包括：

对所述目标语音数据分段进行特征提取处理，得到时域特征图；

对所述目标语音数据分段对应的参考语音频谱特征分段进行特征提取处理，得到频域特征图；

将所述时域特征图和所述频域特征图进行融合处理，生成所述目标语音数据分段对应的声纹语义特征向量。

11、如权利要求 1-10 任一项所述的方法，其特征在于，所述特征提取处理的次数为 k 次，k 为大于 1 的整数；任一次特征提取处理表示为第 i 次特征提取处理；所述将所述时域特征图和所述频域特征图进行融合处理，生成所述目标语音数据分段对应的声纹语义特征向量，包括：

当 i=1 时，将首次特征提取处理得到的中间时域特征图和中间频域特征图进行融合处理，生成所述首次特征提取处理后的第一中间特征向量；

当 1<i≤k 时，将所述第 i 次特征提取处理得到的中间时域特征图和中间频域特征图，以及第 i-1 次特征提取处理得到的第 i-1 中间特征向量进行融合处理，生成所述第 i 次特征提取处理后的第 i 中间特征向量；

基于 i=k 时第 k 次特征提取处理后的第 k 中间特征向量，生成所述目标语音数据分段对应的声纹语义特征向量。

12、一种语音处理装置，其特征在于，所述语音处理装置搭载于计算机设备，所述语音处理装置包括：

获取单元，用于获取混叠语音数据，所述混叠语音数据中包含至少两个对象中每个所述对象产生的语音信号；

所述获取单元，还用于获取指定对象的参考语音数据，所述指定对象是指所述至少两个对象中的任一个；所述参考语音数据中包含所述指定对象的参考语音信号；

处理单元，用于从所述参考语音数据中提取所述指定对象的声纹表征向量，所述声纹表征向量用于表征所述指定对象的声纹特性；

所述处理单元，还用于将所述混叠语音数据和所述声纹表征向量输入预设的语音分割模型，所述语音分割模型用于：基于注意力机制从所述混叠语音数据中分割出与所述声纹特性相匹配的目标语音信号；

所述处理单元，还用于基于所分割出的所述目标语音信号，生成所述指定对象的语音文件。

13、一种计算机设备，其特征在于，

处理器，适于执行计算机程序；

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，所述计算机程序被所述处理器执行时，实现如权利要求 1-11 任一项所述的语音处理方法。

14、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存储有计算机程序，所述计算机程序适于被处理器加载并执行如权利要求 1-11 任一项所述的语音处理方法。

15、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，所述计算机程序被处理器执行时，实现如权利要求 1-11 任一项所述的语音处理方法。

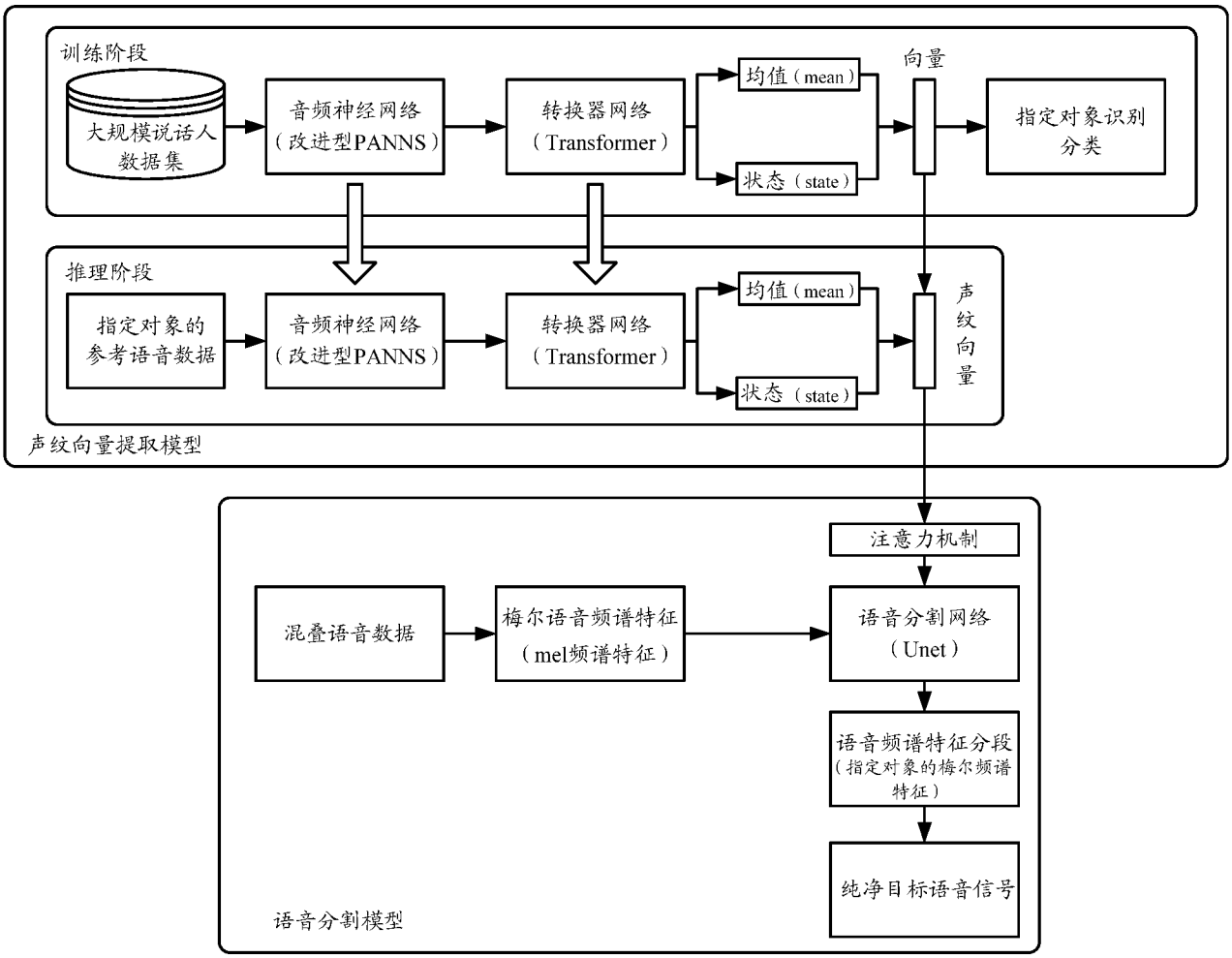
训练阶段
大规模说话人数据集
音频神经网络（改进型PANNS）
转换器网络（Transformer）
均值（mean）
状态（state）
向量
指定对象识别分类
推理阶段
指定对象的参考语音数据
音频神经网络（改进型PANNS）
转换器网络（Transformer）
均值（mean）
状态（state）
声纹向量
声纹向量提取模型
注意力机制
混叠语音数据
梅尔语音频谱特征（mel频谱特征）
语音分割网络（Unet）
语音频谱特征分段（指定对象的梅尔频谱特征）
纯净目标语音信号
语音分割模型

图 1

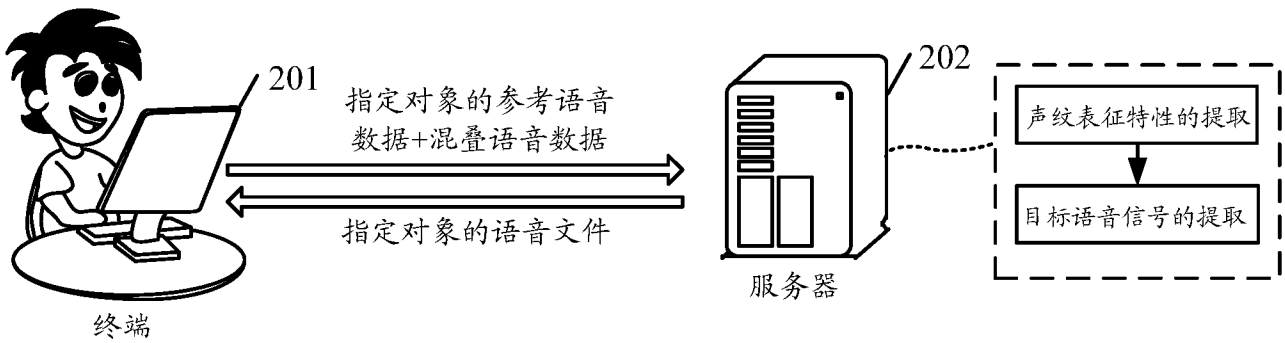
201
指定对象的参考语音数据+混叠语音数据
指定对象的语音文件
终端
202
服务器
声纹表征特性的提取
目标语音信号的提取

图 2

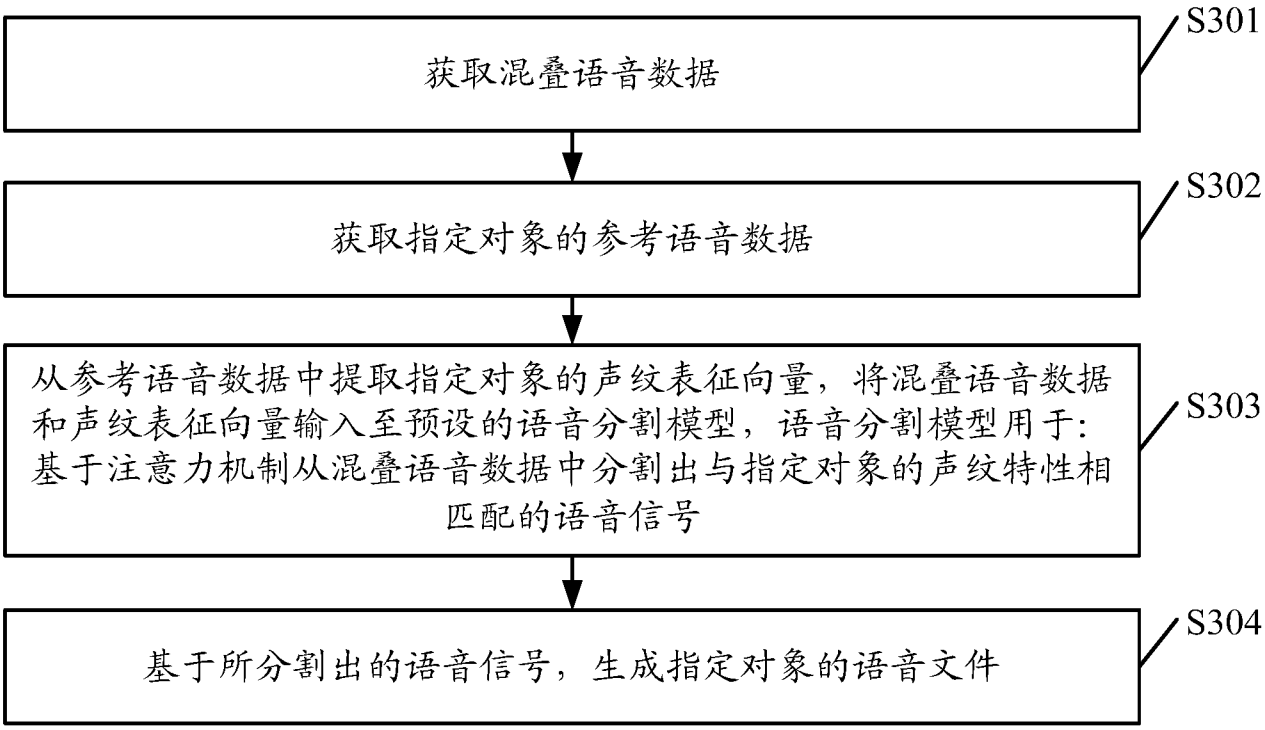
S301
获取混叠语音数据
S302
获取指定对象的参考语音数据
S303
从参考语音数据中提取指定对象的声纹表征向量，将混叠语音数据和声纹表征向量输入至预设的语音分割模型，语音分割模型用于：基于注意力机制从混叠语音数据中分割出与指定对象的声纹特性相匹配的语音信号
S304
基于所分割出的语音信号，生成指定对象的语音文件

图 3

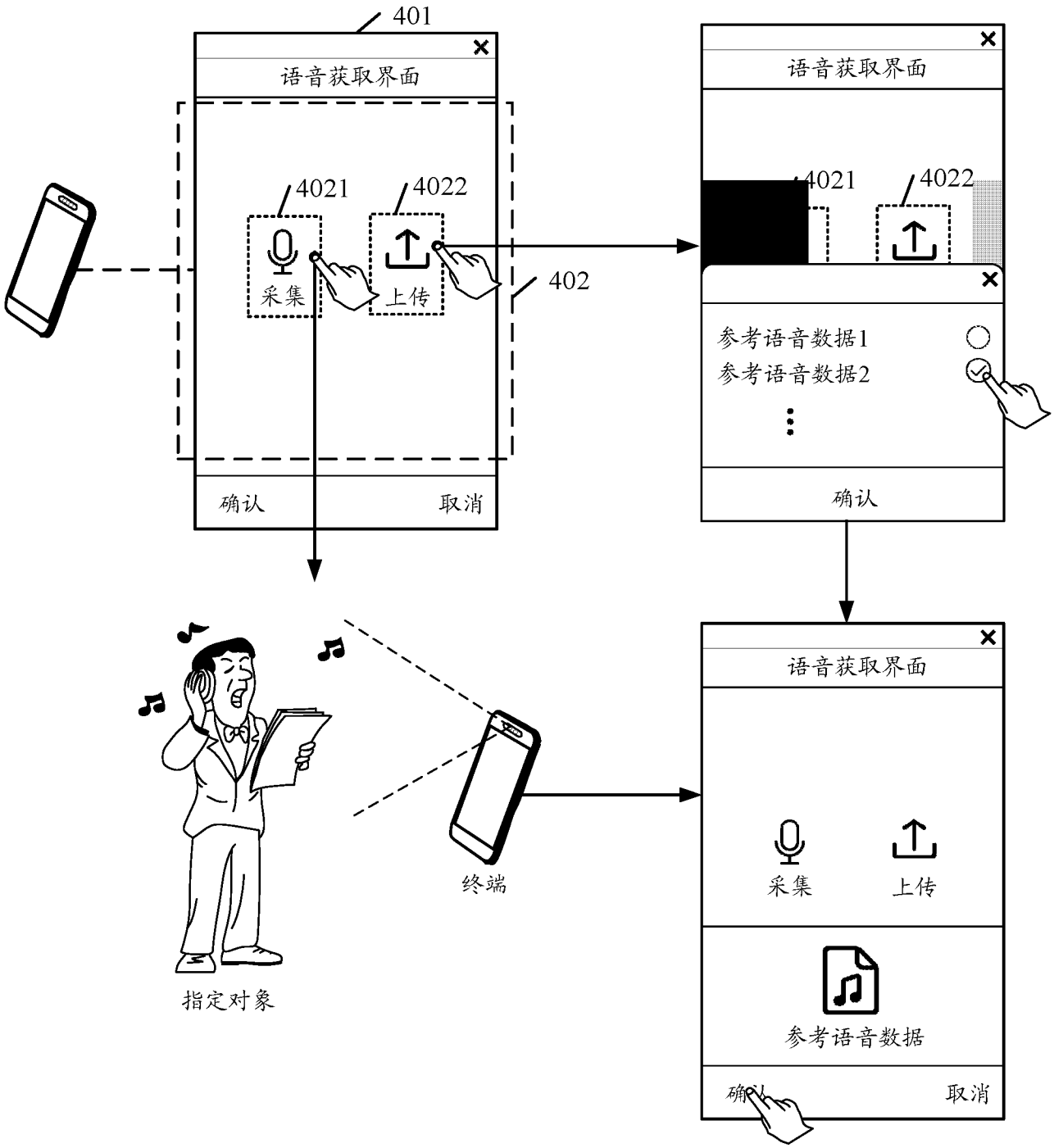
401
语音获取界面
4021
4022
采集
上传
402
确认
取消
语音获取界面
4021
4022
参考语音数据1
参考语音数据2
确认
指定对象
终端
语音获取界面
采集
上传
参考语音数据
确认
取消

图 4

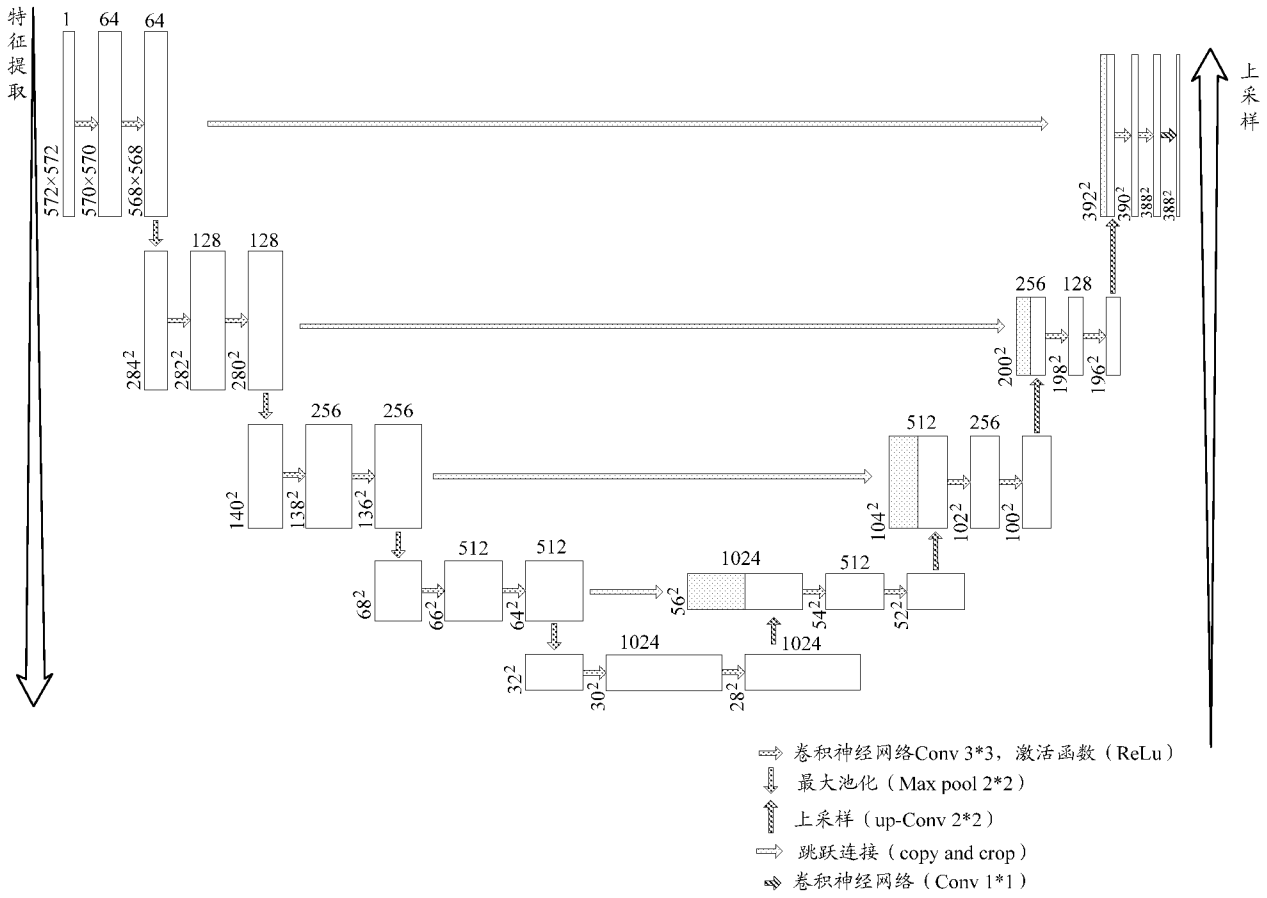
特征提取
上采样
卷积神经网络Conv 3*3，激活函数（ReLu）
最大池化（Max pool 2*2）
上采样（up-Conv 2*2）
跳跃连接（copy and crop）
卷积神经网络（Conv 1*1）

图 5

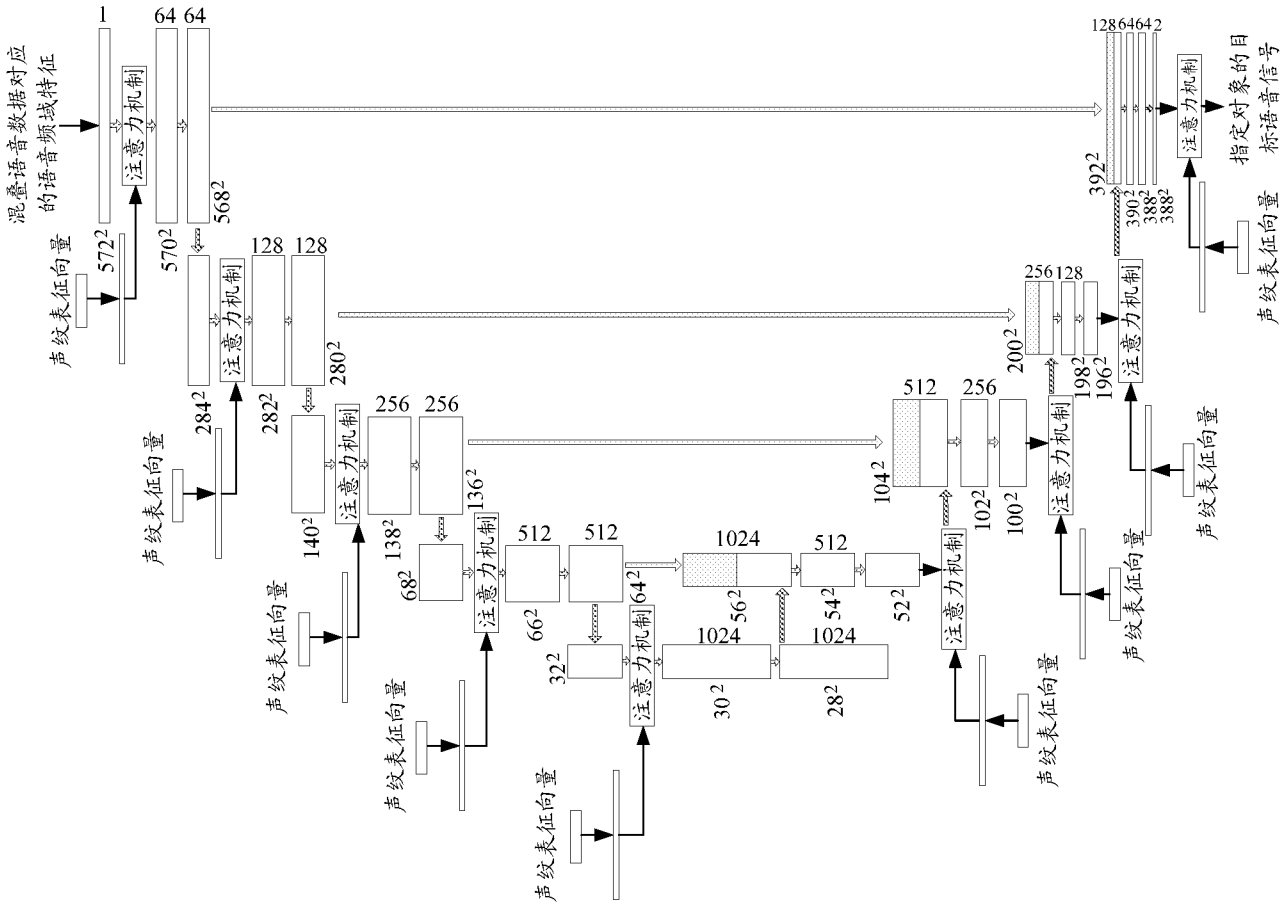
混叠语音数据对应的语音频域特征
注意力机制
声纹表征向量
指定对象的目标语音信号

图 6

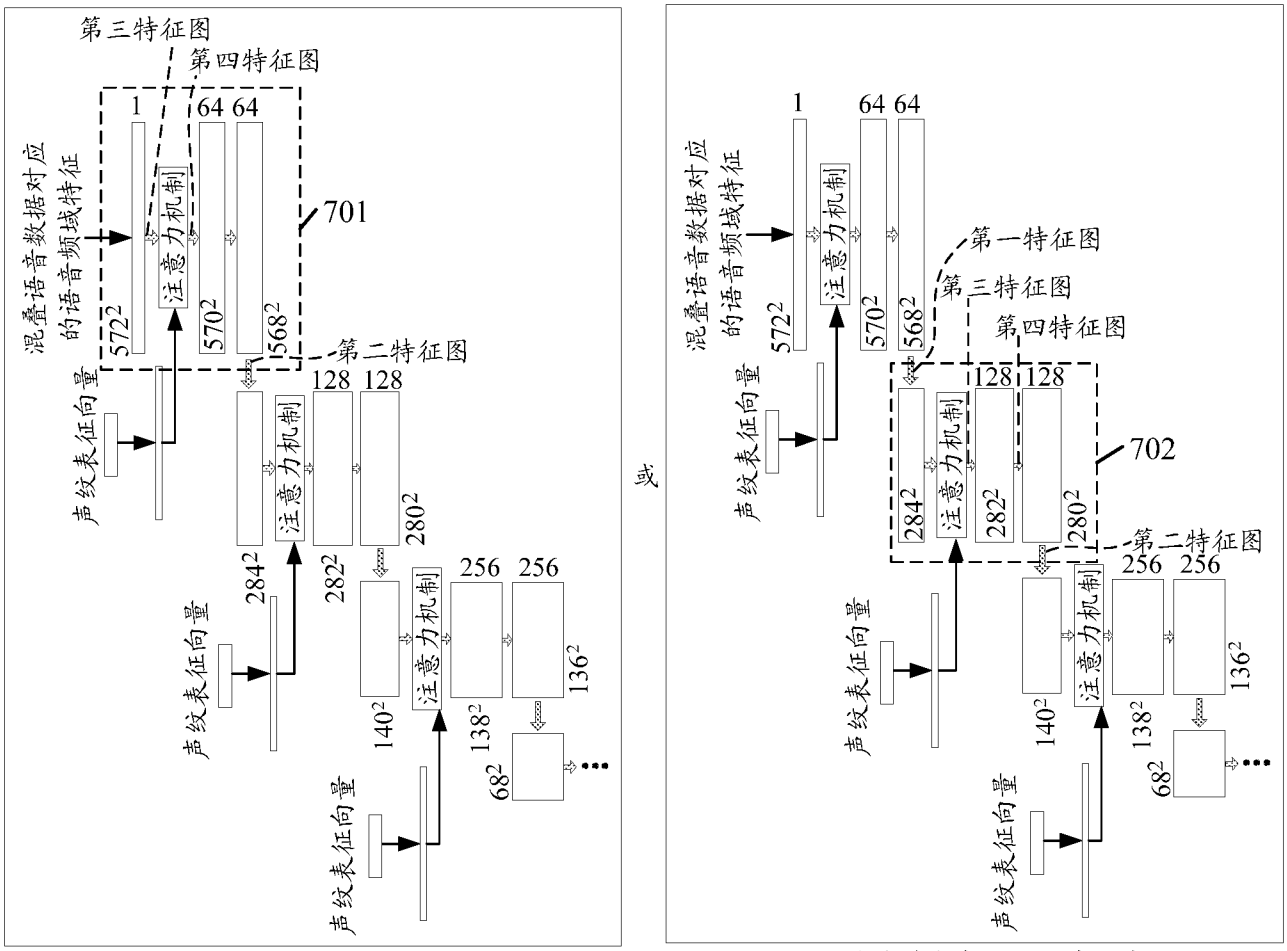
第三特征图
第四特征图
混叠语音数据对应的语音频域特征
声纹表征向量
注意力机制
701
第二特征图
第一特征图
702
或
目标网络层为首个卷积层
目标网络层为除首个卷积层之外的其他网络层

图 7a

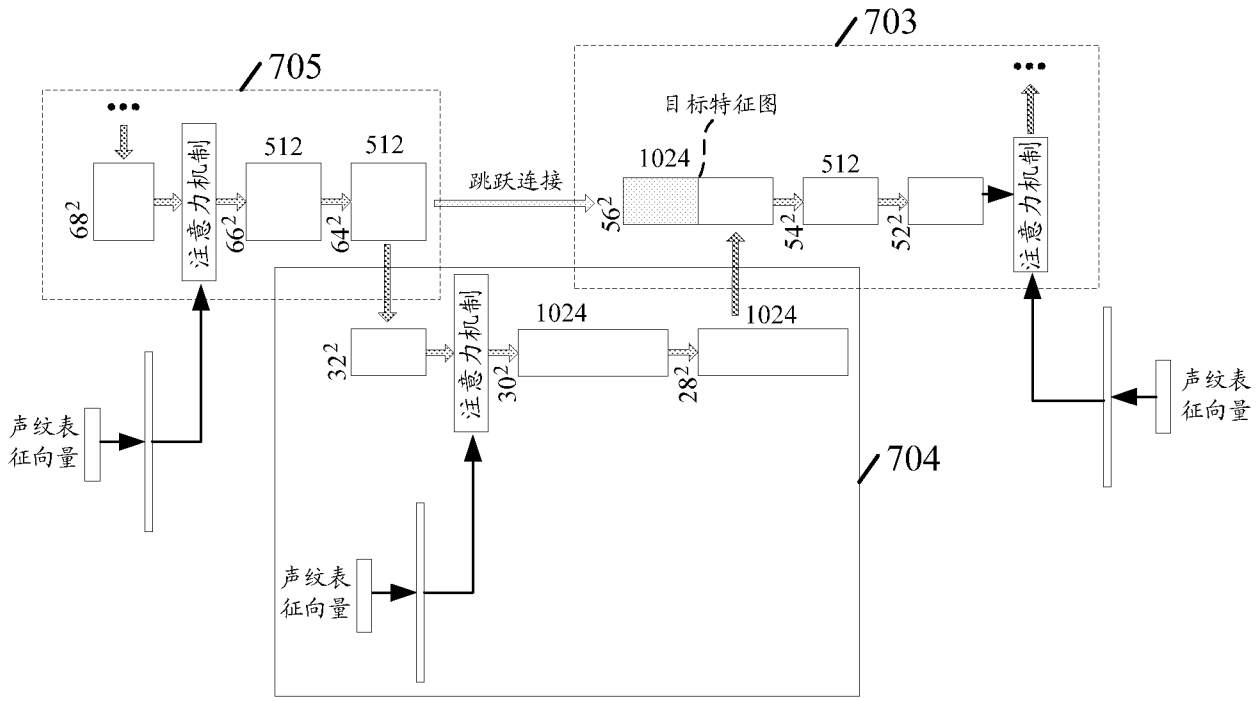
703
705
704
目标特征图
跳跃连接
注意力机制
声纹表征向量

图 7b

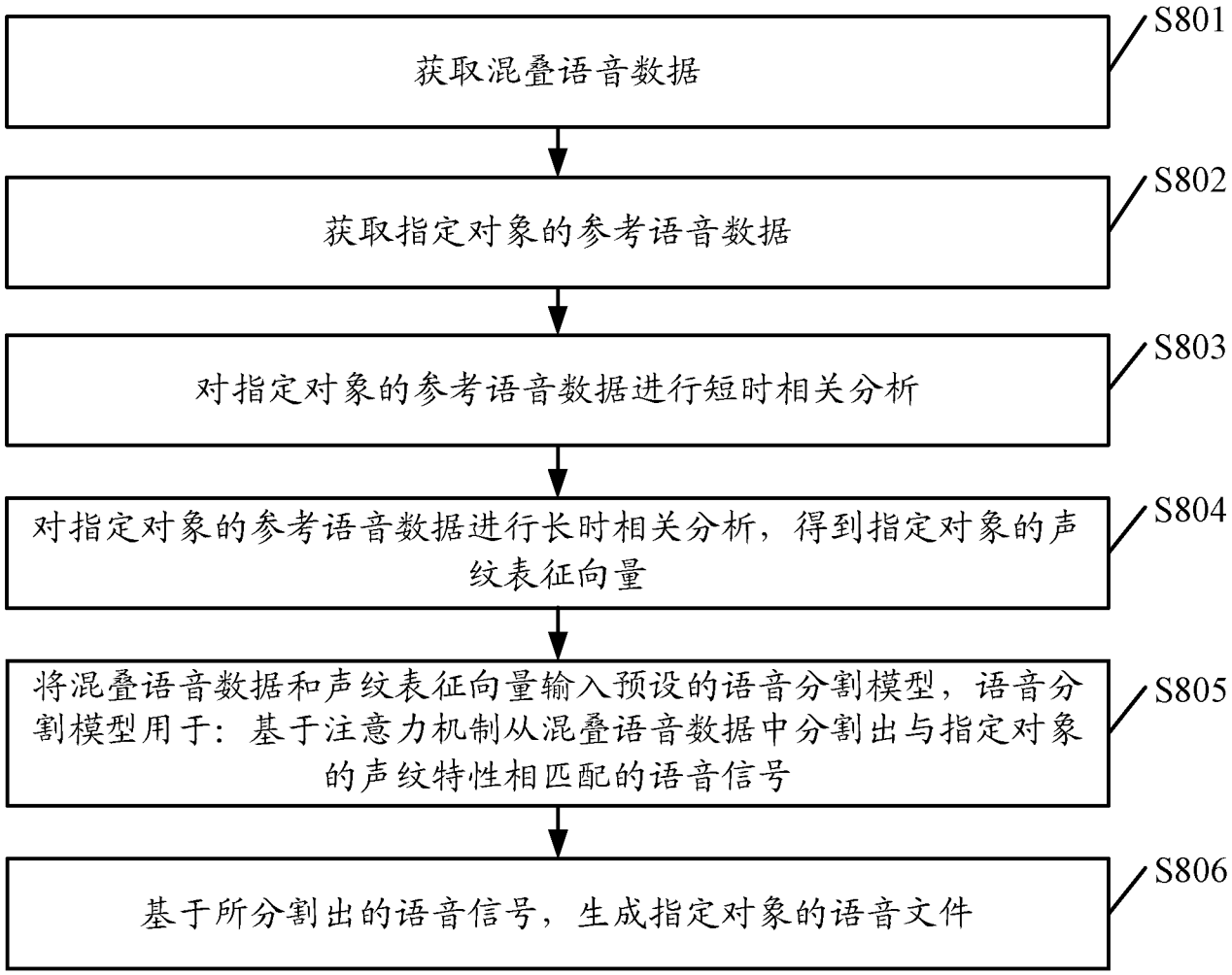
S801
获取混叠语音数据
S802
获取指定对象的参考语音数据
S803
对指定对象的参考语音数据进行短时相关分析
S804
对指定对象的参考语音数据进行长时相关分析，得到指定对象的声纹表征向量
S805
将混叠语音数据和声纹表征向量输入预设的语音分割模型，语音分割模型用于：基于注意力机制从混叠语音数据中分割出与指定对象的声纹特性相匹配的语音信号
S806
基于所分割出的语音信号，生成指定对象的语音文件

图 8

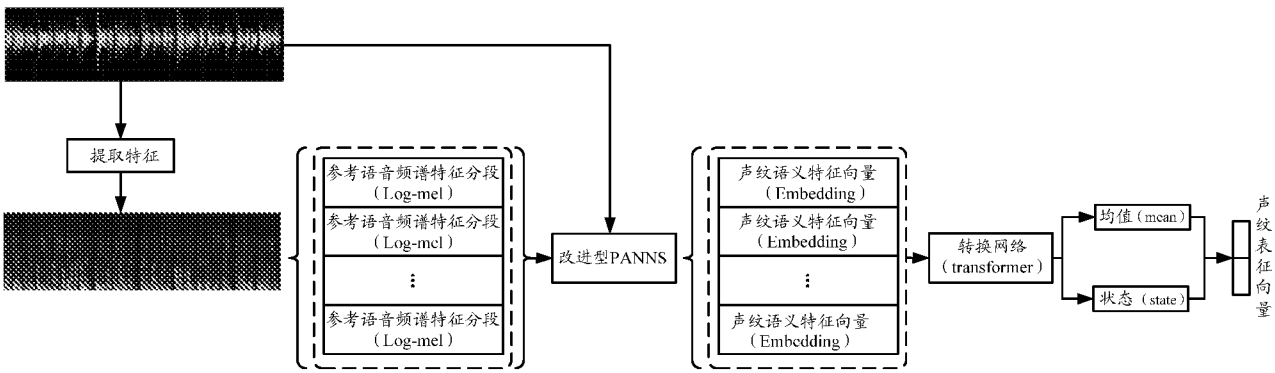
提取特征
参考语音频谱特征分段（Log-mel）
参考语音频谱特征分段（Log-mel）
参考语音频谱特征分段（Log-mel）
改进型PANNS
声纹语义特征向量（Embedding）
声纹语义特征向量（Embedding）
声纹语义特征向量（Embedding）
转换网络（transformer）
均值（mean）
状态（state）
声纹表征向量

图 9

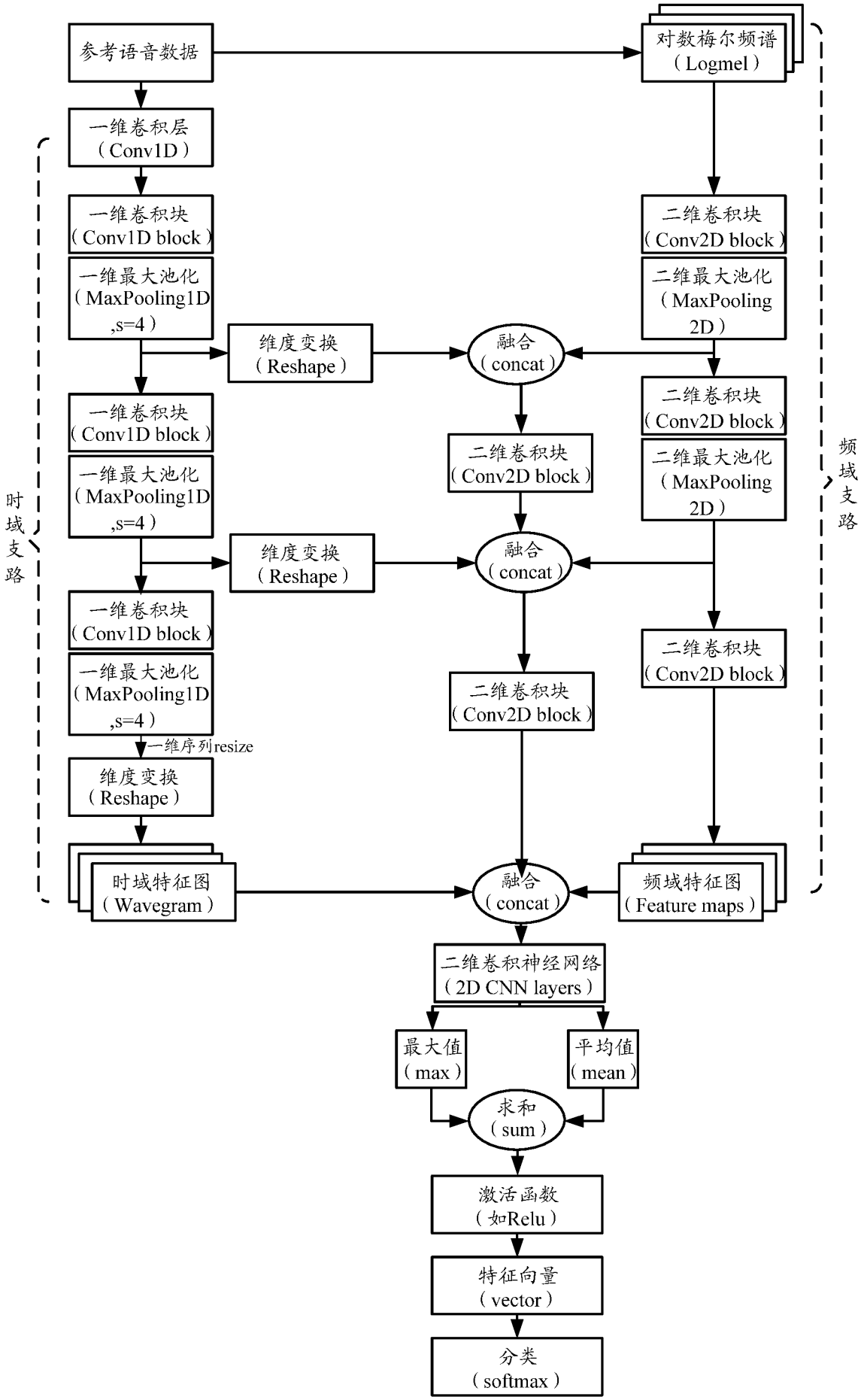
参考语音数据
对数梅尔频谱
（Logmel）
一维卷积层
（Conv1D）
一维卷积块
（Conv1D block）
二维卷积块
（Conv2D block）
一维最大池化
（MaxPooling1D
,s=4）
二维最大池化
（MaxPooling
2D）
维度变换
（Reshape）
融合
（concat）
二维卷积块
（Conv2D block）
一维卷积块
（Conv1D block）
一维最大池化
（MaxPooling1D
,s=4）
二维最大池化
（MaxPooling
2D）
时域支路
频域支路
维度变换
（Reshape）
融合
（concat）
一维卷积块
（Conv1D block）
二维卷积块
（Conv2D block）
一维最大池化
（MaxPooling1D
,s=4）
二维卷积块
（Conv2D block）
一维序列resize
维度变换
（Reshape）
时域特征图
（Wavegram）
融合
（concat）
频域特征图
（Feature maps）
二维卷积神经网络
（2D CNN layers）
最大值
（max）
平均值
（mean）
求和
（sum）
激活函数
（如Relu）
特征向量
（vector）
分类
（softmax）

图 10

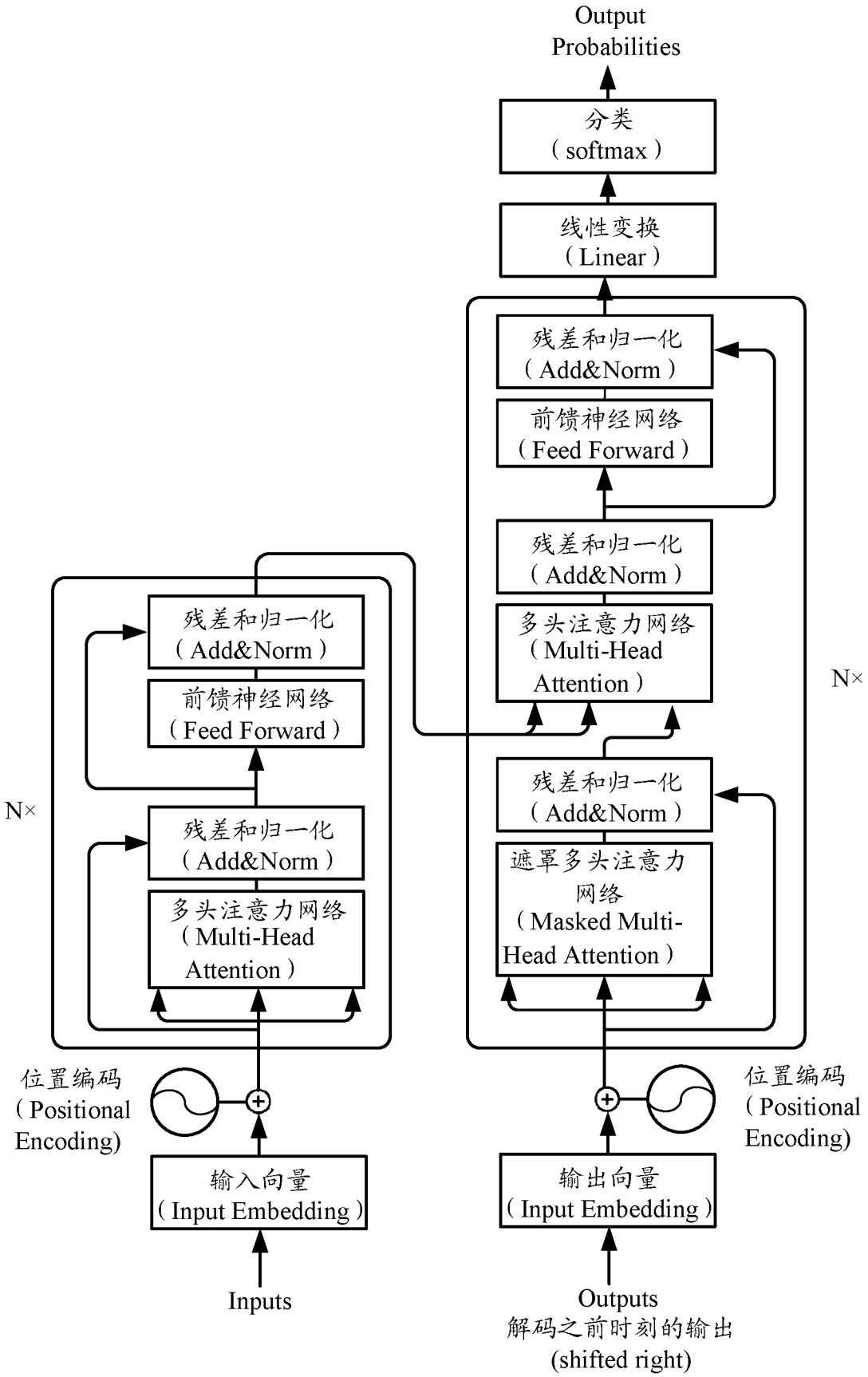
Output Probabilities
分类 (softmax)
线性变换 (Linear)
残差和归一化 (Add&Norm)
前馈神经网络 (Feed Forward)
残差和归一化 (Add&Norm)
多头注意力网络 (Multi-Head Attention)
残差和归一化 (Add&Norm)
遮罩多头注意力网络 (Masked Multi-Head Attention)
N×
残差和归一化 (Add&Norm)
前馈神经网络 (Feed Forward)
残差和归一化 (Add&Norm)
多头注意力网络 (Multi-Head Attention)
N×
位置编码 (Positional Encoding)
位置编码 (Positional Encoding)
输入向量 (Input Embedding)
输出向量 (Input Embedding)
Inputs
Outputs 解码之前时刻的输出 (shifted right)

图 11

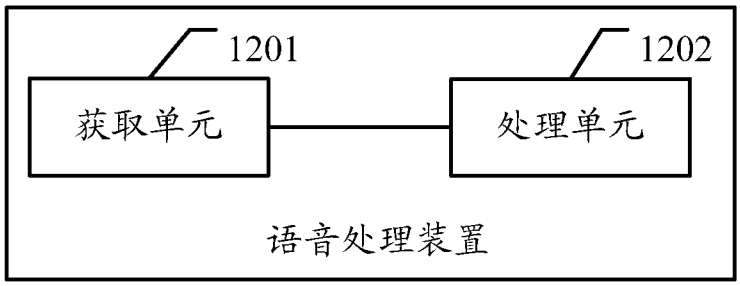
1201
1202
获取单元
处理单元
语音处理装置

图 12

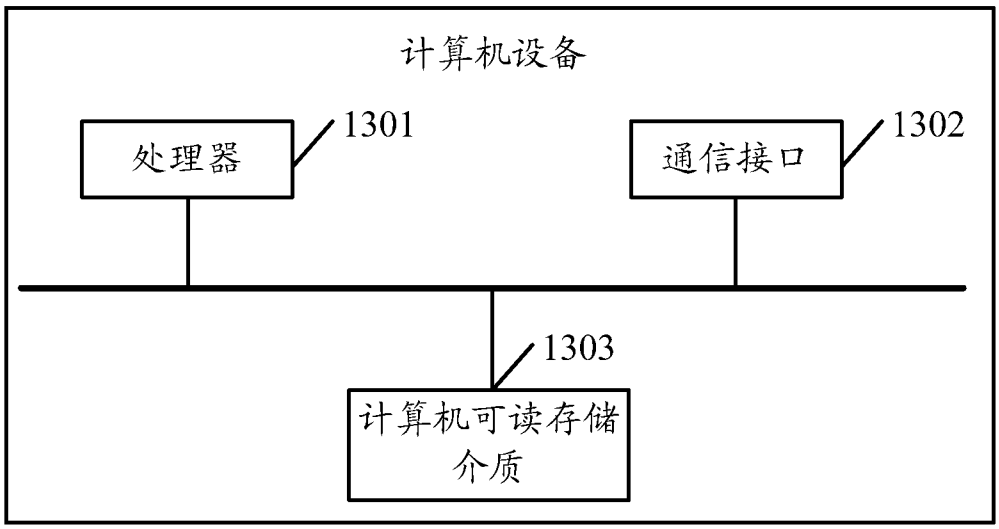
计算机设备
处理器
1301
通信接口
1302
1303
计算机可读存储介质

图 13

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2024/089862

A. CLASSIFICATION OF SUBJECT MATTER

G10L21/028(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)

IPC: 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)

CNABS, CNTXT, WPABSC, ENTXTC, WPABS: 语音, 音频, 混叠, 混合, 分割, 分离, 注意力, 声纹, 频谱, 相关度, 相关性, speech, voice, separat+, attention, voiceprint, spectrum, relevancy

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 108109619 A (INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES) 01 June 2018 (2018-06-01) description, paragraphs 42-43 and 55-107, and figures 1-7	1-2, 9-15
A	CN 111429937 A (BEIJING SOUNDAI TECHNOLOGY CO., LTD.) 17 July 2020 (2020-07-17) entire document	1-15
A	CN 115376541 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 22 November 2022 (2022-11-22) entire document	1-15
A	CN 115116448 A (SICHUAN QIRUIKE TECHNOLOGY CO., LTD.) 27 September 2022 (2022-09-27) entire document	1-15
A	WO 2022048239 A1 (HUAWEI TECHNOLOGIES CO., LTD. et al.) 10 March 2022 (2022-03-10) entire document	1-15

☐ 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
"D" document cited by the applicant in the international application
"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
07 August 2024	**08 August 2024**

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

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2024/089862

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 108109619 A	01 June 2018	WO 2019096149 A1	23 May 2019
		US 2020227064 A1	16 July 2020
		US 10818311 B2	27 October 2020
CN 111429937 A	17 July 2020	None	
CN 115376541 A	22 November 2022	None	
CN 115116448 A	27 September 2022	None	
WO 2022048239 A1	10 March 2022	US 2023206928 A1	29 June 2023
		EP 4198807 A1	21 June 2023
		EP 4198807 A4	20 March 2024

国际检索报告

国际申请号
PCT/CN2024/089862

A. 主题的分类

G10L21/028(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)

IPC：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））

CNABS，CNTXT，WPABSC，ENTXTC，WPABS：语音，音频，混叠，混合，分割，分离，注意力，声纹，频谱，相关度，相关性，speech，voice，separat+，attention，voiceprint，spectrum，relevancy

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
X	CN 108109619 A (中国科学院自动化研究所) 2018年6月1日 (2018 - 06 - 01) 说明书第42-43、55-107段，图1-7	1-2、9-15
A	CN 111429937 A (北京声智科技有限公司) 2020年7月17日 (2020 - 07 - 17) 全文	1-15
A	CN 115376541 A (平安科技（深圳）有限公司) 2022年11月22日 (2022 - 11 - 22) 全文	1-15
A	CN 115116448 A (四川启睿克科技有限公司) 2022年9月27日 (2022 - 09 - 27) 全文	1-15
A	WO 2022048239 A1 (HUAWEI TECH. CO., LTD. 等) 2022年3月10日 (2022 - 03 - 10) 全文	1-15

☐ 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

“D” 申请人在国际申请中引证的文件

“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
2024年8月7日	2024年8月8日
ISA/CN的名称和邮寄地址 中国国家知识产权局 中国北京市海淀区蓟门桥西土城路6号 100088	受权官员 房倩 电话号码 (+86) 010-62089621

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2024/089862

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	108109619	A	2018年6月1日	WO	2019096149	A1	2019年5月23日
				US	2020227064	A1	2020年7月16日
				US	10818311	B2	2020年10月27日
CN	111429937	A	2020年7月17日		无		
CN	115376541	A	2022年11月22日		无		
CN	115116448	A	2022年9月27日		无		
WO	2022048239	A1	2022年3月10日	US	2023206928	A1	2023年6月29日
				EP	4198807	A1	2023年6月21日
				EP	4198807	A4	2024年3月20日