(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
2023 年 11 月 2 日 (02.11.2023)

WIPO | PCT

(10) 国际公布号
WO 2023/207638 A1

(51) 国际专利分类号:
G10L 15/06 (2013.01) *G10L 13/02* (2013.01)

(21) 国际申请号: PCT/CN2023/088492

(22) 国际申请日: 2023 年 4 月 14 日 (14.04.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202210448585.8 2022年4月26日 (26.04.2022) CN

(71) 申请人:北京有竹居网络技术有限公司(**BEIJING YOUZHUJU NETWORK TECHNOLOGY CO., LTD.**) [CN/CN]; 中国北京市平谷区林荫北街13号信息大厦802室, Beijing 101299 (CN)。

(72) 发明人:董倩倩(**DONG, Qianqian**); 中国北京市朝阳区七圣中街12号院融中心B1小邮局, Beijing 100028 (CN)。岳凤鹏(**YUE, Fengpeng**); 中国北京市朝阳区七圣中街12号院融中心B1小邮局, Beijing 100028 (CN)。高汝霆(**KO, Yu Ting**); 中国北京市朝阳区七圣中街12号院融中心B1小邮局, Beijing 100028 (CN)。王明轩(**WANG, Mingxuan**); 中国北京市朝阳区七圣中街12号院融中心B1小邮局, Beijing 100028 (CN)。白奇丙(**BAI, Qibing**); 中国北京市朝阳区七圣中街12号院融中心B1小邮局, Beijing 100028 (CN)。

(74) 代理人:北京知帆远景知识产权代理有限公司(**ZHIFAN & PARTNERS**); 中国北京市海淀区阜成路73号裕惠大厦B座805, Beijing 100142 (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, 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,

(54) **Title:** MODEL TRAINING METHOD AND APPARATUS, SPEECH-TO-SPEECH TRANSLATION METHOD AND APPARATUS, AND MEDIUM

(54) 发明名称: 模型训练方法、语音到语音翻译方法、装置及介质

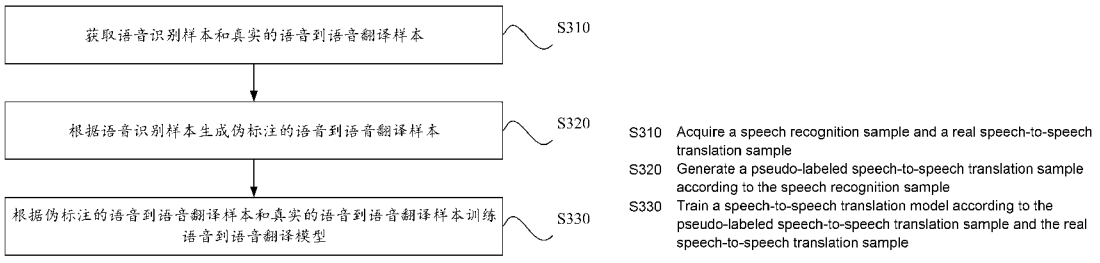


图 3

(57) **Abstract:** A model training method and apparatus, a speech-to-speech translation method and apparatus, and a medium. The model training method comprises: acquiring a speech recognition sample and a real speech-to-speech translation sample (S310); generating a pseudo-labeled speech-to-speech translation sample according to the speech recognition sample (S320); and training a speech-to-speech translation model according to the pseudo-labeled speech-to-speech translation sample and the real speech-to-speech translation sample (S330). The model training method can solve the problem of low model training precision caused by lack of translation sample data.

(57) 摘要: 一种模型训练方法、语音到语音翻译方法、装置及介质，该方法包括：获取语音识别样本和真实的语音到语音翻译样本（S310）；根据语音识别样本生成伪标注的语音到语音翻译样本（S320）；根据伪标注的语音到语音翻译样本和真实的语音到语音翻译样本训练语音到语音翻译模型（S330）。该方法能够解决翻译样本数据匮乏导致模型训练精度较低的问题。

WO 2023/207638 A1

[见续页]

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))。

模型训练方法、语音到语音翻译方法、装置及介质

优先权信息

本申请要求于 2022 年 04 月 26 日提交的，申请名称为“模型训练方法、语音到语音翻译方法、装置及介质”的、中国专利申请号“2022104485858”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

本申请实施例涉及机器学习技术领域，尤其涉及一种模型训练方法、语音到语音翻译方法、装置及介质。

背景技术

语音到语音翻译（Speech-to-Speech Translation，S2ST）模型旨在将源语言语音翻译为目标语言语音，其广泛应用于视频翻译、跨国会议演讲、翻译对讲机等各种场景。通常语音到语音翻译模型需要通过大量数据训练得到，而目前在现实场景中很难收集成对的语音到语音的翻译样本，这种数据匮乏的情况导致模型训练精度较低的问题。

发明内容

本申请提供一种模型训练方法、语音到语音翻译方法、装置及介质，从而可以提高模型训练精度。

第一方面，提供一种模型训练方法，包括：获取语音识别样本和真实的语音到语音翻译样本；根据语音识别样本生成伪标注的语音到语音翻译样本；根据伪标注的语音到语音翻译样本和真实的语音到语音翻译样本训练语音到语音翻译模型。

第二方面，提供一种语音到语音翻译方法，包括：获取源语言语音特征；将源语言语音特征输入至如第一方面或第一方面的可选方式训练得到的语音到语音翻译模型，得到源语言语音特征对应的目标语言语音特征。

第三方面，提供一种模型训练装置，包括：获取模块、生成模块和训练模块，获取模块用于获取语音识别样本和真实的语音到语音翻译样本；生成模块用于根据语音识别样本生成伪标注的语音到语音翻译样本；训练模块用于根据伪标注的语音到语音翻译样本和真实的语音到语音翻译样本训练语音到语音翻译模型。

第四方面，提供语音到语音翻译装置，包括：获取模块和处理模块，其中，获取模块用于获取源语言语音特征；处理模块用于将源语言语音特征输入至如第一方面或第一方面的可选方式训练得到的语音到语音翻译模型，得到源语言语音特征对应的目标语言语音特征。

第五方面，提供了一种电子设备，包括：处理器和存储器，该存储器用于存储计算机程序，该处理

器用于调用并运行该存储器中存储的计算机程序，以执行第一方面或第二方面所述的方法。

第六方面，提供了一种计算机可读存储介质，用于存储计算机程序，该计算机程序使得计算机执行第一方面或第二方面所述的方法。

综上，虽然在现实场景中很难收集成对的语音到语音的翻译样本，但是语音识别样本却比较好收集，基于该语音识别样本可以生成伪标注的语音到语音的翻译样本，从而扩充了语音到语音的翻译样本，进而可以提高模型训练精度。

附图说明

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

图 1 为 Transformer 的框架图；

图 2 为本申请实施例涉及的一种系统架构示意图；

图 3 为本申请实施例提供的一种模型训练方法的流程图；

图 4 为本申请实施例提供的另一种模型训练方法的流程图；

图 5 为本申请实施例提供的一种语音到语音翻译模型的示意图；

图 6 为本申请实施例提供的另一种语音到语音翻译模型的示意图；

图 7 为本申请实施例提供的一种语音到语音翻译方法的流程图；

图 8 为本申请实施例提供的一种模型训练装置 800 的示意图；

图 9 为本申请实施例提供的一种语音到语音翻译装置 900 的示意图；

图 10 是本申请实施例提供的电子设备 1000 的示意性框图。

具体实施方式

下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例，都属于本发明保护的范围。

需要说明的是，本发明的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本发明的实施例能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或服务器不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

在介绍本申请技术方案之前，下面首先对本申请技术方案的相关知识进行阐述：

一、编码器（Encoder）和解码器（Decoder）：

编码器（Encoder），用于处理源语言语音特征，并将源语言语音特征压缩成固定长度的隐藏表示，该隐藏表示也被称为上下文向量（context）、语义编码、语义向量等，期望该隐藏表示能够比较好的表示语言语音特征的信息。

解码器（Decoder），利用隐藏表示初始化解码器以得到目标语言语音特征。

二、转换器（Transformer）：

图 1 为 Transformer 的框架图，如图 1 所示，编码器可以由 N=6 个一模一样的单元组成。每个单元包含两个子单元。第一个是采用多头自注意力机制（multi-head self-attention mechanism）的自注意力网络，第二个是全连接的前馈网络，激活函数是 ReLU。这两个子单元都是用了残差连接（residual connection，ADD）和层归一化（layer normalization，Norm）。解码器与编码器几乎一样，只不过可以在中间多增加了一层多头注意力机制（encoder-decoder attention）来处理编码器的输出。同时，解码器的第一个单元即采用多头自注意力机制的第一个单元为了确保解码器不会读取当前位置之后的信息进行了遮挡（masking）操作。

下面对本申请所要解决的技术问题和发明构思进行说明：

如上所述，通常语音到语音翻译模型需要通过大量数据训练得到，而目前在现实场景中很难收集成对的语音到语音的翻译样本，这种数据匮乏的情况导致模型训练精度较低的问题。

为了解决上述技术问题，本申请提出了扩充训练数据，以提高模型训练精度。

在一些实施例中，本申请实施例的系统架构如图 2 所示。

图 2 为本申请实施例涉及的一种系统架构示意图，用户设备 201、数据采集设备 202、训练设备 203、执行设备 204、数据库 205 和内容库 206。

其中，数据采集设备 202 用于从内容库 206 中读取训练数据，并将读取的训练数据存储至数据库 205 中。本申请实施例涉及的训练数据包括伪标注的语音到语音翻译样本和真实的语音到语音翻译样本。

训练设备 203 基于数据库 205 中维护的训练数据，对语音到语音翻译模型进行训练，使得训练后的语音到语音翻译模型可以有效地实现源语言语音到目标语言语音的翻译。

另外，参考图 2，执行设备 204 配置有 I/O 接口 207，与外部设备进行数据交互。比如通过 I/O 接口接收用户设备 201 发送的源语言语音特征。执行设备 204 中的计算模块 208 使用训练后的语音到语音翻译模型对输入的源语言语音特征进行处理，输出目标语言语音特征，具体可以是目标语言语音特征，并通过 I/O 接口将相应的结果发送至用户设备 201。

其中，用户设备 201 可以包括手机、平板电脑、笔记本电脑、掌上电脑、移动互联网设备（mobile internet device，MID）、台式电脑、或其他具有安装浏览器功能的终端设备。

执行设备 204 可以为服务器。

示例性的，服务器可以是机架式服务器、刀片式服务器、塔式服务器或机柜式服务器等计算设备。该服务器可以是独立的测试服务器，也可以是多个测试服务器所组成的测试服务器集群。

本实施例中，执行设备 204 通过网络与用户设备 201 连接。所述网络可以是企业内部网(Intranet)、互联网(Internet)、全球移动通讯系统(Global System of Mobile communication，GSM)、宽带码分多址(Wideband Code Division Multiple Access，WCDMA)、第四代（the 4rd Generation，4G）网络、第五代（the 5rd Generation，5G）网络、蓝牙(Bluetooth)、无线保真（wireless fidelity，Wi-Fi）、通话网络等无线或有线网络。

需要说明的是，图 2 仅是本申请实施例提供的一种系统架构的示意图，图中所示设备、器件、模块等之间的位置关系不构成任何限制。在一些实施例中，上述数据采集设备 202 与用户设备 201、训练设备 203 和执行设备 204 可以为同一个设备。上述数据库 205 可以分布在一个服务器上也可以分布在多个服务器上，上述的内容库 206 可以分布在一个服务器上也可以分布在多个服务器上。

下面通过一些实施例对本申请实施例的技术方案进行详细说明。下面这几个实施例可以相互结合，对于相同或相似的概念或过程可能在某些实施例不再赘述。

图 3 为本申请实施例提供的一种模型训练方法的流程图，该方法可以由手机、平板电脑、笔记本电脑、掌上电脑、MID、台式电脑等任何电子设备执行，例如可以由图 2 中的训练设备执行，本申请对此不做限制，如图 3 所示，该方法包括：

S310：获取语音识别样本和真实的语音到语音翻译样本；

S320：根据语音识别样本生成伪标注的语音到语音翻译样本；

S330：根据伪标注的语音到语音翻译样本和真实的语音到语音翻译样本训练语音到语音翻译模型。

可选地，本申请中的语音到语音翻译模型可以是基于多任务学习（Multi-Task Learning，MTL）的语音到语音翻译模型，当然也可以是基于单任务学习的语音到语音翻译模型，本申请对此不做限制。其中，多任务学习是机器学习中一个很有前景的领域，其目标是利用多个学习任务中所包含的有用信息来帮助为每个任务学习得到更为准确的学习器。并且在多任务学习中，由于不同任务之间共享归纳偏好，所以任务之间一般是可以相互提升而避免单任务容易陷入局部最优。

为了方便起见，下面将真实的语音到语音翻译样本，又或者是伪标注的语音到语音翻译样本统称为语音到语音翻译样本。应理解的是，语音到语音翻译样本所包括的数据元素的个数与语音到语音翻译模型基于多任务学习或者单任务学习有关。例如：若语音到语音翻译模型是基于单任务学习的语音到语音翻译模型，则语音到语音翻译样本可以是一个二元组，包括：源语言语音特征和目标语言语音特征。若语音到语音翻译模型是基于多任务学习的语音到语音翻译模型，且多任务包括：一个主任务和两个辅助任务，主任务是语音到语音的翻译任务；两个辅助任务分别是语音识别任务和语音到文本的翻译任务，该语音识别任务用于将源语言语音特征转换为源语言文本，该语音到文本的翻译任务用于将源语言语音特征转换为源语言文本，并将该源语言文本转换为目标语言文本，这种情况下，语音到语音翻译样本可

以是一个四元组，包括：源语言语音特征、源语言文本、目标语言语音特征和目标语言文本。

可选地，真实的语音到语音翻译样本包括：第一源语言语音特征、第一源语言文本、第一目标语言语音特征和第一目标语言文本。

应理解的是，第一源语言语音特征是真实的源语言语音特征，第一源语言文本是真实的源语言文本，第一目标语言文本也是真实的目标语言文本，第一目标语言语音特征是电子设备对第一目标语言文本合成得到的目标语言语音特征，例如：电子设备可以将第一目标语言文本输入至语音合成模型中，得到第一目标语言语音特征。

可选地，语音识别样本包括：第二源语言语音特征和第二源语言文本。其中，该第二源语言语音特征是真实的源语言语音特征，该第二源语言文本也是真实的源语言文本。

应理解的是，所谓真实的源语言语音特征指的是在真实场景中可以得到的源语言语音特征，例如：电子设备可以通过麦克风采集某用户的语音，并提取该语音的特征。真实的源语言文本可以是通过人工方式得到的语言文本，例如：用户可以将一段语音记录下来，以形成该语音对应的语言文本。真实的目标语言文本也可以是通过人工方式得到的语言文本，例如：用户将源语言文本中的内容翻译为目标语言文本。

可选地，上述语音识别样本可以是一个或多个，上述真实的语音到语音翻译样本可以是一个或多个。

可选地，电子设备可以对第二源语言文本进行翻译，得到第二目标语言语音特征；对第二目标语言语音特征进行合成，得到第二目标语言语音特征；其中，伪标注的语音到语音翻译样本可以是一个四元组，包括：第二源语言语音特征、第二源语言文本、第二目标语言文本和第二目标语言语音特征。换句话讲，伪标注的语音到语音翻译样本中的前两项，即第二源语言语音特征、第二源语言文本均是真实的。

可选地，电子设备可以将第二源语言文本输入至机器翻译（Machine Translation，MT）模型中，得到第二目标语言文本。电子设备可以将第二目标语言文本输入至语音合成（Text-To-Speech，TTS）模型中，得到第二目标语言语音特征。

应理解的是，按照上述伪标注的语音到语音翻译样本的构造过程可知，真实的语音到语音翻译样本与伪标注的语音到语音翻译样本的区别主要在目标语言文本，例如：真实的语音到语音翻译样本是一个四元组，包括：$\{s_{src}, t_{src}, t_{tgt}, s_{tgt}\}$，其中，$s_{src}$ 表示真实的源语言语音特征、t_{src} 表示真实的源语言文本、t_{tgt} 表示真实的目标语言文本、s_{tgt} 表示对 t_{tgt} 合成之后得到的目标语言语音特征。伪标注的语音到语音翻译样本是一个四元组，包括：$\{s_{src}, t_{src}, t'_{tgt}, s_{tgt}\}$，其中，$s_{src}$ 表示真实的源语言语音特征、t_{src} 表示真实的源语言文本、t'_{tgt} 表示将真实的源语言文本输入至 MT 之后得到的目标语言文本、s_{tgt} 表示 t'_{tgt} 合成之后得到的目标语言语音特征。

本申请可以将基于这种伪标注的语音到语音翻译样本所进行的语音到语音的翻译称为伪翻译标签

适配（Pseudo Translation Label Adaptation，PTLA）。

应理解的是，上述伪标注的语音到语音翻译样本是在语音识别样本的基础上的得到的，实际上，也可以基于源语言语音特征构造伪标注的语音到语音翻译样本。例如：电子设备可以获取真实的源语言语音特征，将该源语言语音特征输入至自动语音识别（Automatic Speech Recognition，ASR）模型，得到该源语言语音特征对应的源语言文本，再将该源语言文本输入至 MT 模型中得到目标语言文本，最后可以将该目标语言文本输入至 TTS 模型中，得到目标语言语音特征，基于此，这些源语言语音特征、源语言文本、目标语言文本和目标语言语音特征构成伪标注的语音到语音翻译样本。

可选地，在本申请中，源语言语音特征可以是源语言语音的对数梅尔谱图（log-mel spectrogram），该对数梅尔谱图可以是 80 通道的对数梅尔谱图，但不限于此。

可选地，在本申请中，源语言语音特征可以是目标语言语音的线性频率频谱图（linear freq spectrogram），但不限于此。

应理解的是，电子设备对语音到语音翻译模型训练过程包括：预训练阶段（pre-training）和微调阶段（fine tuning）。

预训练指的是预先训练模型或者预先训练模型的过程。微调指的是将预训练过的模型作用于某任务的数据集，并使参数适应该任务的数据集的过程。

例如：当需要搭建一个网络模型来完成一个特定图像分类任务时，首先，电子设备需要随机初始化参数，然后开始训练网络模型，不断调整该模型的参数，使得网络模型的损失越来越小，直到满足训练停止条件为止，该过程就是预训练过程。当获取到一个与上述特定图像分类任务类似的图像分类任务时，电子设备可以直接使用之前训练的网络模型，将该网络模型的参数来作为这一任务的初始化参数，然后训练网络模型，不断调整该模型的参数，使得网络模型的损失越来越小，直到满足训练停止条件为止，该过程就是微调过程。

应理解的是，上述真实的语音到语音翻译样本也可以被称为原始的语音到语音翻译样本。伪标注的语音到语音翻译样本也可以被称为衍生的语音到语音翻译样本。其中，该真实的语音到语音翻译样本可以作用于语音到语音翻译模型的预训练阶段，也可以作用于该模型的微调阶段。该伪标注的语音到语音翻译样本可以作用于语音到语音翻译模型的预训练阶段，也可以作用于该模型的微调阶段，本申请对此不做限制。

综上，虽然在现实场景中很难收集成对的语音到语音的翻译样本，但是语音识别样本却比较好收集，基于该语音识别样本可以生成伪标注的语音到语音的翻译样本，从而扩充了语音到语音的翻译样本，进而可以提高模型训练精度。

下面将示例性地阐述若干模型训练方法：

如图 4 所示，上述 S330 包括：

S410：根据伪标注的语音到语音翻译样本预训练语音到语音翻译模型；

S420：根据真实的语音到语音翻译样本微调预训练后的语音到语音翻译模型。

应理解的是，由于本申请引入了伪标注的语音到语音翻译样本，为了提高模型训练精度，可以将伪标注的语音到语音翻译样本应用于模型的预训练阶段，而将真实的语音到语音翻译样本应用于模型的微调阶段。

可选地，电子设备可以直接通过真实的语音到语音翻译样本微调预训练后的语音到语音翻译模型，也就是说，电子设备只通过真实的语音到语音翻译样本微调预训练后的语音到语音翻译模型。

可选地，电子设备也可以根据真实的语音到语音翻译样本和伪标注的语音到语音翻译样本微调预训练后的语音到语音翻译模型。基于该方式所进行的模型训练方法可以称为混合训练方法。

应理解的是，这种混合训练算法可以最大限度地保留伪标注的语音到语音翻译样本收益。由于语音识别样本的规模相对于真实的语音到语音翻译样本的规模要大很多，基于此，伪标注的语音到语音翻译样本的规模相对于真实的语音到语音翻译样本的规模要大很多，为了防止伪标注的语音到语音翻译样本误导模型优化结果，在本申请中，可以对真实的语音到语音翻译样本进行上采样，以扩充真实的语音到语音翻译样本的规模，进而可以通过上采样后的语音到语音翻译样本和伪标注的语音到语音翻译样本微调预训练后的语音到语音翻译模型。

可选地，电子设备在根据真实的语音到语音翻译样本和伪标注的语音到语音翻译样本微调预训练后的语音到语音翻译模型之前，还可以对真实的语音到语音翻译样本标注第一标签，第一标签用于标识真实的语音到语音翻译样本为真实样本，可以用 real 表示；对伪标注的语音到语音翻译样本标注第二标签，第二标签用于标识伪标注的语音到语音翻译样本为伪标注样本，可以用 pseudo 表示。基于该方式所进行的模型训练方法可以称为提示（prompt）训练方法，基于这种提示训练方式，可以使得模型更好地区分真实的语音到语音翻译样本和伪标注的语音到语音翻译样本。

在本申请实施例中，电子设备可以根据伪标注的语音到语音翻译样本预训练语音到语音翻译模型；根据真实的语音到语音翻译样本微调预训练后的语音到语音翻译模型，即伪标注数据主要应用于预训练过程，通过这种方式可以防止伪标注的语音到语音翻译样本误导模型优化结果。

进一步地，由于真实的语音到语音翻译样本匮乏，所以伪标注数据也可以参与进微调过程，但是为了防止伪标注的语音到语音翻译样本误导模型优化结果，可以对真实的语音到语音翻译样本进行上采样，进而可以通过上采样后的语音到语音翻译样本和伪标注的语音到语音翻译样本微调预训练后的语音到语音翻译模型，通过该方法一方面可以解决真实的语音到语音翻译样本匮乏而导致的模型训练精度较低的问题，另一方面可以防止伪标注的语音到语音翻译样本误导模型优化结果。

更进一步地，电子设备在根据真实的语音到语音翻译样本和伪标注的语音到语音翻译样本微调预训练后的语音到语音翻译模型之前，还可以对真实的语音到语音翻译样本和伪标注的语音到语音翻译样本标注对应标签，从而可以使得模型更好地区分真实的语音到语音翻译样本和伪标注的语音到语音翻译样本。

可选地，在本申请中，语音到语音翻译模型可以是现有的翻译器（Translatotron）模型或者可以如图 5 所示的语音到语音翻译模型，本申请对此不做限制。

图 5 为本申请实施例提供的一种语音到语音翻译模型的示意图，如图 5 所示，该模型包括：编码器模块 510、第一注意力模块 520、第一解码器模块 530、N 个第二注意力模块 540 和 N 个第二解码器模块 550，N 为正整数，N 个第二注意力模块和 N 个第二解码器模块一一对应，图 5 以 N=2 为例，当然 N 也可以等于 1、3 等等。

可选地，该模型可以是基于多任务学习的语音到语音翻译模型，且多任务包括：一个主任务和 N 个辅助任务，主任务是语音到语音的翻译任务，其中，上述第二注意力模块、第二解码器模块的数量与辅助任务数量一致，例如：N=2，即存在两个辅助任务，两个辅助任务可以分别是语音识别任务和语音到文本的翻译任务，但不限于此。该语音识别任务用于将源语言语音特征转换为源语言文本，该语音到文本的翻译任务用于将源语言语音特征转换为源语言文本，并将该源语言文本转换为目标语言文本。再例如：N=1，即存在一个辅助任务，该辅助任务可以是语音识别任务或语音到文本的翻译任务，但不限于此。第一注意力模块 520 和第一解码器模块 530 对应于主任务，而下面的每组第二注意力模块 540 和第二解码器模块 550 对应于一个辅助任务。第一解码器模块 530 主要用于预测合成目标语言语音特征。训练时，两个辅助任务接受编码器模块 510 的输入，并将预测的损失值以加权和的形式加入到主任务，测试时，第二解码器模块 550 不被使用。

编码器模块 510 用于获取源语言语音特征，并对源语言语音特征进行处理，得到源语言语音特征对应的多组第一隐藏状态表示。

可选地，图 6 为本申请实施例提供的另一种语音到语音翻译模型的示意图，如图 6 所示，编码器模块 510 包括：卷积神经网络子模块 5101 和第一转换器模块 5102；卷积神经网络子模块 5101 用于获取源语言语音特征，并对源语言语音特征进行处理，得到源语言语音特征对应的第二隐藏状态表示；第一转换器模块 5102 用于获取第二隐藏状态表示，并对第二隐藏状态表示进行处理，得到多组第一隐藏状态表示。

可选地，卷积神经网络子模块 5101 可以包括两层卷积神经网络层，但不限于此。例如，80 通道的对数梅尔谱图输入卷积神经网络子模块 5101 之后，两层卷积神经网络层可以将 80 通道的对数梅尔谱图的长度映射为原来的四分之一，即假设之前 80 通道的对数梅尔谱图是通过 100 个向量表示，每个向量是 80 维的，那么经过两层卷积神经网络层处理之后，得到的是 25 个向量，为了与第一转换器模块 5102 中的隐藏单元数量保持一致，如隐藏单元数量是 512，那么经过两层卷积神经网络层处理之后的 25 个向量的维数也是 512，其中，这里可以将 25 个 512 维的向量理解为上述第二隐藏状态表示。

可选地，第一转换器模块 5102 可以类似于图 1 所示的编码器结构，即该第一转换器模块 5102 可以包括 6 个转换器层，也可以包括 12 个转换器层，本申请对此不做限制，每个转换器层可以带有 512 隐藏单元，也就是说，经过通过转换器层输出的隐藏表示可以是 512 维的。每个转换器层可以包括包含两

个子单元。第一个是采用多头自注意力机制的自注意力网络，这里的多头自注意力机制的自注意力网络可以是 8 头自注意力机制的自注意力网络，本申请对此不做限制。第二个是全连接的前馈网络，其中，前馈网络可以使用 2048 维的内部状态。这两个子单元都是用了残差连接和层归一化。经过第一转换器模块 5102 对第二隐藏状态表示的处理，可以得到 N 组第一隐藏状态表示，这里的 N 表示第一转换器模块 5102 所包括的转换器层的层数。

结合上面的示例，假设卷积神经网络子模块 5101 输出的是 25 个 512 维的向量，那么经过第一转换器模块 5102 处理后，可以得到 N 组第一隐藏状态表示，每组第一隐藏状态表示也是 25 个 512 维的向量，其中，第一转换器模块 5102 的最后一层所得到的 25 个 512 维的向量可以输出给第一注意力模块 520，中间层所得到的 25 个 512 维的向量可以输出给第二注意力模块 540。

第一注意力模块 520 用于获取多组第一隐藏状态表示中的一组第一隐藏状态表示以及第一解码器输出的各个时间步对应的第一向量，并对该组第一隐藏状态表示和各个时间步对应的第一向量进行处理，得到各个时间步对应的第一注意力表示。

第一解码器模块 530 用于获取各个时间步对应的第二向量，并对各个时间步对应的第二向量进行处理，得到各个时间步对应的第一向量，将各个时间步对应的第一向量输出给第一注意力模块 520，获取各个时间步对应的第一注意力表示，并对各个时间步对应的第一注意力表示进行处理，得到源语言语音特征对应的目标语言语音特征。

可选地，如图 6 所示，第一解码器模块 530 包括：前处理网络（prenet）5301、第二转换器模块 5302 和后处理网络（postnet）5303；前处理网络 5301 用于获取各个时间步对应的第二向量，并对各个时间步对应的第二向量进行处理，得到各个时间步对应的第一向量，将各个时间步对应的第一向量输出给第一注意力模块；第二转换器模块 5302 用于获取各个时间步对应的第一注意力表示，并对各个时间步对应的第一注意力表示进行处理，得到各个时间步上的目标语言语音特征；后处理网络 5303 用于对各个时间步上的目标语言语音特征进行处理，得到源语言语音特征对应的目标语言语音特征。

可选地，前处理网络 5301 的瓶颈（bottleneck）维数可以是 32。

结合上面的示例，在训练阶段，前处理网络 5301 可以获取 80 维的全 0 向量，该全 0 向量是第一个时间步对应的第二向量，前处理网络 5301 可以对全 0 向量进行处理，得到 512 维的全 0 向量，该全 0 向量是第一个时间步对应的第一向量，进一步地，前处理网络 5301 可以将该全 0 向量输入至第一注意力模块 520，第一注意力模块 520 可以对该全 0 向量以及从编码器模块 510 获取到的 25 个 512 维的向量进行处理，得到第一个时间步对应的第一注意力表示，更进一步地，第一注意力模块 520 将第一个时间步对应的第一注意力表示输入至第二转换器模块 5302，第二转换器模块 5302 可以对第一个时间步对应的第一注意力表示进行处理，得到第一个时间步上的目标语言语音特征，该第一个时间步上的目标语言语音特征是预测得到的第一个时间步上的目标语言语音特征。此外，前处理网络 5301 还可以获取第一个时间步上的实际地目标语言语音特征，该第一个时间步上的实际地目标语言语音特征可以被理解为第

二个时间步对应的第二向量，前处理网络 5301 可以对第二个时间步对应的第二向量进行处理，得到 512 维的向量，该 512 维的向量是第二个时间步对应的第一向量，进一步地，前处理网络 5301 可以将该第二个时间步对应的第一向量输入至第一注意力模块 520，第一注意力模块 520 可以对该第二个时间步对应的第一向量以及从编码器模块 510 获取到的 25 个 512 维的向量进行处理，得到第二个时间步对应的第一注意力表示，更进一步地，第一注意力模块 520 将第二个时间步对应的第一注意力表示输入至第二转换器模块 5302，第二转换器模块 5302 可以对第二个时间步对应的第一注意力表示进行处理，得到第二个时间步上的目标语言语音特征，该第二个时间步上的目标语言语音特征是预测得到的第二个时间步上的目标语言语音特征。

总之，前处理网络 5301 可以获取第 i 个时间步上的实际地目标语言语音特征，i=1,2……M，M 是总共的时间步数。该第 i 个时间步上的实际地目标语言语音特征可以被理解为第 i+1 个时间步对应的第二向量，前处理网络 5301 可以对第 i+1 个时间步对应的第二向量进行处理，得到 512 维的向量，该 512 维的向量是第 i+1 个时间步对应的第一向量，进一步地，前处理网络 5301 可以将该第 i+1 个时间步对应的第一向量输入至第一注意力模块 520，第一注意力模块 520 可以对该第 i+1 个时间步对应的第一向量以及从编码器模块 510 获取到的 25 个 512 维的向量进行处理，得到第 i+1 个时间步对应的第一注意力表示，更进一步地，第一注意力模块 520 将第 i+1 个时间步对应的第一注意力表示输入至第二转换器模块 5302，第二转换器模块 5302 可以对第 i+1 个时间步对应的第一注意力表示进行处理，得到第 i+1 个时间步上的目标语言语音特征。

在测试阶段，前处理网络 5301 可以获取 80 维的全 0 向量，该全 0 向量是第一个时间步对应的第二向量，前处理网络 5301 可以对全 0 向量进行处理，得到 512 维的全 0 向量，该全 0 向量是第一个时间步对应的第一向量，进一步地，前处理网络 5301 可以将该全 0 向量输入至第一注意力模块 520，第一注意力模块 520 可以对该全 0 向量以及从编码器模块 510 获取到的 25 个 512 维的向量进行处理，得到第一个时间步对应的第一注意力表示，更进一步地，第一注意力模块 520 将第一个时间步对应的第一注意力表示输入至第二转换器模块 5302，第二转换器模块 5302 可以对第一个时间步对应的第一注意力表示进行处理，得到第一个时间步上的目标语言语音特征，该第一个时间步上的目标语言语音特征是预测得到的第一个时间步上的目标语言语音特征。进一步地，前处理网络 5301 可以对第一个时间步上的预测地目标语言语音特征并对其进行处理，得到 512 维的向量，其中，第一个时间步上的预测地目标语言语音特征可以被理解为上述第二个时间步对应的第二向量，该 512 维的向量是第二个时间步对应的第一向量，进一步地，前处理网络 5301 可以将该第二个时间步对应的第一向量输入至第一注意力模块 520，第一注意力模块 520 可以对该第二个时间步对应的第一向量以及从编码器模块 510 获取到的 25 个 512 维的向量进行处理，得到第二个时间步对应的第一注意力表示，更进一步地，第一注意力模块 520 将第二个时间步对应的第一注意力表示输入至第二转换器模块 5302，第二转换器模块 5302 可以对第二个时间步对应的第一注意力表示进行处理，得到第二个时间步上的目标语言语音特征。

总之，前处理网络 5301 可以获取第 i 个时间步上的预测地目标语言语音特征，i=1,2……M，M 是总共的时间步数。该第 i 个时间步上的预测地目标语言语音特征可以被理解为第 i+1 个时间步对应的第二向量，前处理网络 5301 可以对第 i+1 个时间步对应的第二向量进行处理，得到 512 维的向量，该 512 维的向量是第 i+1 个时间步对应的第一向量，进一步地，前处理网络 5301 可以将该第 i+1 个时间步对应的第一向量输入至第一注意力模块 520，第一注意力模块 520 可以对该第 i+1 个时间步对应的第一向量以及从编码器模块 510 获取到的 25 个 512 维的向量进行处理，得到第 i+1 个时间步对应的第一注意力表示，更进一步地，第一注意力模块 520 将第 i+1 个时间步对应的第一注意力表示输入至第二转换器模块 5302，第二转换器模块 5302 可以对第 i+1 个时间步对应的第一注意力表示进行处理，得到第 i+1 个时间步上的目标语言语音特征。

可选地，第二转换器模块 5302 可以将各个时间步上的目标语言语音特征输入给后处理网络 5303，后处理网络 5303 可以对各个时间步上的目标语言语音特征进行加权求和，得到源语言语音特征对应的目标语言语音特征。

可选地，第二转换器模块 5302 可以类似于图 1 所示的解码器结构，即该第二转换器模块 5302 可以包括 6 个转换器层，本申请对此不做限制，第二转换器模块 5302 可以与第一转换器模块 5102 采用相同的超参数。

在对语音到语音翻译模型的训练阶段，第二注意力模块 540 用于获取多组第一隐藏状态表示中的一组第一隐藏状态表示以及第二注意力模块 540 对应的第二解码器输出的各个时间步对应的第三向量，并对该组第一隐藏状态表示和各个时间步对应的第三向量进行处理，得到各个时间步对应的第二注意力表示；

第二注意力模块 540 对应的第二解码器模块 550 用于获取各个时间步对应的第四向量，并对各个时间步对应的第四向量进行处理，得到各个时间步对应的第三向量，将各个时间步对应的第三向量输出给第二注意力模块 540，获取各个时间步对应的第二注意力表示，并对各个时间步对应的第二注意力表示进行处理，得到源语言语音特征对应的辅助表示。

可选地，第二解码器模块 550 可以包括：前处理网络、第三转换器模块和后处理网络；前处理网络用于获取各个时间步对应的第四向量，并对各个时间步对应的第四向量进行处理，得到各个时间步对应的第三向量，将各个时间步对应的第三向量输出给第二注意力模块 540；第三转换器模块用于获取各个时间步对应的第二注意力表示，并对各个时间步对应的第二注意力表示进行处理，得到各个时间步上的辅助表示；后处理网络用于对各个时间步上的辅助表示进行处理，得到源语言语音特征对应的辅助表示。

可选地，前处理网络的瓶颈（bottleneck）维数可以是 32。

示例性地，在训练阶段，前处理网络 5301 可以获取 80 维的嵌入向量，该嵌入向量是第一个时间步对应的第四向量，前处理网络可以对向量进行处理，得到 512 维的向量，该向量是第一个时间步对应的第三向量，进一步地，前处理网络可以将该向量输入至第二注意力模块 540，第二注意力模块 540 可以

对该向量以及从编码器模块 510 获取到的 25 个 512 维的向量进行处理，得到第一个时间步对应的第二注意力表示，更进一步地，第二注意力模块 540 将第一个时间步对应的第二注意力表示输入至第三转换器模块，第三转换器模块可以对第一个时间步对应的第二注意力表示进行处理，得到第一个时间步上的辅助表示，该第一个时间步上的辅助表示是预测得到的第一个时间步上的辅助表示。此外，前处理网络还可以获取第一个时间步上的实际地辅助表示，该第一个时间步上的实际地辅助表示可以被理解为第二个时间步对应的第四向量，前处理网络可以对第二个时间步对应的第四向量进行处理，得到 512 维的向量，该 512 维的向量是第二个时间步对应的第三向量，进一步地，前处理网络可以将该第二个时间步对应的第三向量输入至第二注意力模块 540，第二注意力模块 540 可以对该第二个时间步对应的第三向量以及从编码器模块 510 获取到的 25 个 512 维的向量进行处理，得到第二个时间步对应的第二注意力表示，更进一步地，第二注意力模块 540 将第二个时间步对应的第二注意力表示输入至第三转换器模块，第三转换器模块可以对第二个时间步对应的第二注意力表示进行处理，得到第二个时间步上的辅助表示，该第二个时间步上的辅助表示是预测得到的第二个时间步上的辅助表示。

总之，前处理网络可以获取第 i 个时间步上的实际地辅助表示，i=1,2……M，M 是总共的时间步数。该第 i 个时间步上的实际地辅助表示可以被理解为第 i+1 个时间步对应的第四向量，前处理网络可以对第 i+1 个时间步对应的第四向量进行处理，得到 512 维的向量，该 512 维的向量是第 i+1 个时间步对应的第三向量，进一步地，前处理网络可以将该第 i+1 个时间步对应的第三向量输入至第二注意力模块 540，第二注意力模块 540 可以对该第 i+1 个时间步对应的第三向量以及从编码器模块 510 获取到的 25 个 512 维的向量进行处理，得到第 i+1 个时间步对应的第二注意力表示，更进一步地，第二注意力模块 540 将第 i+1 个时间步对应的第二注意力表示输入至第三转换器模块，第三转换器模块可以对第 i+1 个时间步对应的第二注意力表示进行处理，得到第 i+1 个时间步上的辅助表示。

可选地，第三转换器模块可以将各个时间步上的辅助表示输入给后处理网络，后处理网络可以对各个时间步上的辅助表示进行加权求和，得到源语言语音特征对应的辅助表示。

可选地，当辅助任务是语音识别任务时，上述辅助表示可以是语音识别结果，如源语言语音对应的源语言文本。当辅助任务是语音到文本的翻译任务时，上述辅助表示可以是语音翻译结果，如目标语言文本。

应理解的是，本申请提供的语音到语音翻译模型是对现有的翻译器（Translatotron）模型进行了相应的改进，具体是将翻译器（Translatotron）模型中的长短期记忆网络（Long Short-Term Memory，LSTM）替换为转换器模块，在本申请中，可以将该语音到语音翻译模型称为基于转换器的翻译器模型（Transformer-based Translatotron）。在 LSTM 中每个时间步上的计算都是局部计算，而在转换器模块中，每个时间步上的计算都是全局计算，从而可以提高模型精度。

应理解的是，本申请所提供的 Transformer-based Translatotron 可以不基于伪标注的语音到语音翻译样本进行训练，将采用这种训练方式的语音到语音模型可以称为基线系统。当然本申请所提供的

Transformer-based Translatotron 也可以基于伪标注的语音到语音翻译样本进行训练，将采用这种训练方式的语音到语音模型可以称为基线系统+PTLA。

本申请提供了语音翻译中常用的 TEDEn2Zh 数据集（英到中）来测试基线系统和基线系统+PTLA 的性能，具体如表 1 所示：

表 1

方法	S-PER	Tp-BLEU	Dev-BLEU	test-BLEU
基线系统	13.06	44.57	7.0	11.2
基线系统+PTLA	11.93	48.64	10.4	18.0

其中，S-PER 表示语音识别任务在测试集上的音素识别错误率。Tp-BLEU 表示语音到文本翻译任务在测试集上的基于的音素计算的双语互译质量辅助工具（Bilingual Evaluation Understudy，BLEU）。Dev-BLER 表示主任务在开发集上的基于的音素计算的 BLEU。test-BLEU 表示主任务在测试集上的基于的音素计算的 BLEU。

从表 1 中可以看出，基线系统可以在复杂的语向翻译上取得良好性能，基线系统+PTLA 方案可以有效提升模型表现。

下面将提供一种语音到语音翻译方法：

图 7 为本申请实施例提供的一种语音到语音翻译方法的流程图，该方法可以由手机、平板电脑、笔记本电脑、掌上电脑、MID、台式电脑等任何电子设备执行，例如可以由图 2 中的执行设备执行，本申请对此不做限制，如图 7 所示，该方法包括：

S710：获取源语言语音特征；

S720：将源语言语音特征输入至语音到语音翻译模型，得到源语言语音特征对应的目标语言语音特征。

应理解的是，该语音到语音翻译模型可以由上述模型训练方法训练得到，由于通过上述训练方法所得到的语音到语音翻译模型精度更高，基于此，可以更好地实现语音到语音翻译。

图 8 为本申请实施例提供的一种模型训练装置 800 的示意图，如图 8 所示，该装置 800 包括：获取模块 810、生成模块 820 和训练模块 830，其中，获取模块 810 用于获取语音识别样本和真实的语音到语音翻译样本；生成模块 820 用于根据语音识别样本生成伪标注的语音到语音翻译样本；训练模块 830 用于根据伪标注的语音到语音翻译样本和真实的语音到语音翻译样本训练语音到语音翻译模型。

可选地，训练模块 830 具体用于：根据伪标注的语音到语音翻译样本预训练语音到语音翻译模型，并根据真实的语音到语音翻译样本微调预训练后的语音到语音翻译模型。

可选地，训练模块 830 具体用于：通过真实的语音到语音翻译样本微调预训练后的语音到语音翻译模型；或者，根据真实的语音到语音翻译样本和伪标注的语音到语音翻译样本微调预训练后的语音到语音翻译模型。

可选地，该装置 800 还包括：标注模块 840，用于在根据真实的语音到语音翻译样本和伪标注的语音到语音翻译样本微调预训练后的语音到语音翻译模型之前，对真实的语音到语音翻译样本标注第一标签，第一标签用于标识真实的语音到语音翻译样本为真实样本；对伪标注的语音到语音翻译样本标注第二标签，第二标签用于标识伪标注的语音到语音翻译样本为伪标注样本。

可选地，训练模块 830 具体用于：对真实的语音到语音翻译样本进行上采样，得到上采样后的语音到语音翻译样本；通过上采样后的语音到语音翻译样本和伪标注的语音到语音翻译样本微调预训练后的语音到语音翻译模型。

可选地，真实的语音到语音翻译样本包括：第一源语言语音特征、第一源语言文本、第一目标语言语音特征和第一目标语言文本；语音识别样本包括：第二源语言语音特征和第二源语言文本。

可选地，生成模块 820 具体用于：对第二源语言文本进行翻译，得到第二目标语言文本；对第二目标语言文本进行合成，得到第二目标语言语音特征；其中，伪标注的语音到语音翻译样本包括：第二源语言语音特征、第二源语言文本、第二目标语言文本和第二目标语言语音特征。

可选地，语音到语音翻译模型包括：编码器模块、第一注意力模块、第一解码器模块、N 个第二注意力模块和 N 个第二解码器模块，N 为正整数，N 个第二注意力模块和 N 个第二解码器模块一一对应；

编码器模块用于获取源语言语音特征，并对源语言语音特征进行处理，得到源语言语音特征对应的多组第一隐藏状态表示；

第一注意力模块用于获取多组第一隐藏状态表示中的一组第一隐藏状态表示以及第一解码器输出的各个时间步对应的第一向量，并对该组第一隐藏状态表示和各个时间步对应的第一向量进行处理，得到各个时间步对应的第一注意力表示；

第一解码器模块用于获取各个时间步对应的第二向量，并对各个时间步对应的第二向量进行处理，得到各个时间步对应的第一向量，将各个时间步对应的第一向量输出给第一注意力模块，获取各个时间步对应的第一注意力表示，并对各个时间步对应的第一注意力表示进行处理，得到源语言语音特征对应的目标语言语音特征；

在对语音到语音翻译模型的训练阶段，第二注意力模块用于获取多组第一隐藏状态表示中的一组第一隐藏状态表示以及第二注意力模块对应的第二解码器输出的各个时间步对应的第三向量，并对该组第一隐藏状态表示和各个时间步对应的第三向量进行处理，得到各个时间步对应的第二注意力表示；

第二注意力模块对应的第二解码器模块用于获取各个时间步对应的第四向量，并对各个时间步对应的第四向量进行处理，得到各个时间步对应的第三向量，将各个时间步对应的第三向量输出给第二注意力模块，获取各个时间步对应的第二注意力表示，并对各个时间步对应的第二注意力表示进行处理，得到源语言语音特征对应的辅助表示。

可选地，编码器模块包括：卷积神经网络子模块和第一转换器模块；卷积神经网络子模块用于获取源语言语音特征，并对源语言语音特征进行处理，得到源语言语音特征对应的第二隐藏状态表示；第一

转换器模块用于获取第二隐藏状态表示，并对第二隐藏状态表示进行处理，得到多组第一隐藏状态表示。

可选地，第一解码器模块包括：前处理网络、第二转换器模块和后处理网络；前处理网络用于获取各个时间步对应的第二向量，并对各个时间步对应的第二向量进行处理，得到各个时间步对应的第一向量，将各个时间步对应的第一向量输出给第一注意力模块；第二转换器模块用于获取各个时间步对应的第一注意力表示，并对各个时间步对应的第一注意力表示进行处理，得到各个时间步上的目标语言语音特征；后处理网络用于对各个时间步上的目标语言语音特征进行处理，得到源语言语音特征对应的目标语言语音特征。

应理解的是，装置实施例与方法实施例可以相互对应，类似的描述可以参照方法实施例。为避免重复，此处不再赘述。具体地，图 8 所示的装置 800 可以执行上述模型训练方法实施例，并且装置 800 中的各个模块的前述和其它操作和/或功能分别为了实现上述模型训练方法中的相应流程，为了简洁，在此不再赘述。

上文中结合附图从功能模块的角度描述了本申请实施例的装置 800。应理解，该功能模块可以通过硬件形式实现，也可以通过软件形式的指令实现，还可以通过硬件和软件模块组合实现。具体地，本申请实施例中的方法实施例的各步骤可以通过处理器中的硬件的集成逻辑电路和/或软件形式的指令完成，结合本申请实施例公开的方法的步骤可以直接体现为硬件译码处理器执行完成，或者用译码处理器中的硬件及软件模块组合执行完成。可选地，软件模块可以位于随机存储器，闪存、只读存储器、可编程只读存储器、电可擦写可编程存储器、寄存器等本领域的成熟的存储介质中。该存储介质位于存储器，处理器读取存储器中的信息，结合其硬件完成上述模型训练方法实施例中的步骤。

图 9 为本申请实施例提供的一种语音到语音翻译装置 900 的示意图，如图 9 所示，该装置 900 包括：获取模块 910 和处理模块 920，其中，获取模块 910 用于获取源语言语音特征；处理模块 920 用于将源语言语音特征输入至通过上述模型训练方法训练得到的语音到语音翻译模型，得到源语言语音特征对应的目标语言语音特征。

应理解的是，装置实施例与方法实施例可以相互对应，类似的描述可以参照方法实施例。为避免重复，此处不再赘述。具体地，图 9 所示的装置 900 可以执行上述语音到语音翻译方法实施例，并且装置 900 中的各个模块的前述和其它操作和/或功能分别为了实现上述语音到语音翻译方法中的相应流程，为了简洁，在此不再赘述。

上文中结合附图从功能模块的角度描述了本申请实施例的装置 900。应理解，该功能模块可以通过硬件形式实现，也可以通过软件形式的指令实现，还可以通过硬件和软件模块组合实现。具体地，本申请实施例中的方法实施例的各步骤可以通过处理器中的硬件的集成逻辑电路和/或软件形式的指令完成，结合本申请实施例公开的方法的步骤可以直接体现为硬件译码处理器执行完成，或者用译码处理器中的硬件及软件模块组合执行完成。可选地，软件模块可以位于随机存储器，闪存、只读存储器、可编程只读存储器、电可擦写可编程存储器、寄存器等本领域的成熟的存储介质中。该存储介质位于存储器，处

理器读取存储器中的信息，结合其硬件完成上述语音到语音翻译方法实施例中的步骤。

图 10 是本申请实施例提供的电子设备 1000 的示意性框图。

如图 10 所示，该电子设备 1000 可包括：

存储器 1010 和处理器 1020，该存储器 1010 用于存储计算机程序，并将该程序代码传输给该处理器 1020。换言之，该处理器 1020 可以从存储器 1010 中调用并运行计算机程序，以实现本申请实施例中的方法。

例如，该处理器 1020 可用于根据该计算机程序中的指令执行上述方法实施例。

在本申请的一些实施例中，该处理器 1020 可以包括但不限于：

通用处理器、数字信号处理器（Digital Signal Processor，DSP）、专用集成电路（Application Specific Integrated Circuit，ASIC）、现场可编程门阵列（Field Programmable Gate Array，FPGA）或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等等。

在本申请的一些实施例中，该存储器 1010 包括但不限于：

易失性存储器和/或非易失性存储器。其中，非易失性存储器可以是只读存储器（Read-Only Memory，ROM）、可编程只读存储器（Programmable ROM，PROM）、可擦除可编程只读存储器（Erasable PROM，EPROM）、电可擦除可编程只读存储器（Electrically EPROM，EEPROM）或闪存。易失性存储器可以是随机存取存储器（Random Access Memory，RAM），其用作外部高速缓存。通过示例性但不是限制性说明，许多形式的 RAM 可用，例如静态随机存取存储器（Static RAM，SRAM）、动态随机存取存储器（Dynamic RAM，DRAM）、同步动态随机存取存储器（Synchronous DRAM，SDRAM）、双倍数据速率同步动态随机存取存储器（Double Data Rate SDRAM，DDR SDRAM）、增强型同步动态随机存取存储器（Enhanced SDRAM，ESDRAM）、同步连接动态随机存取存储器（synch link DRAM，SLDRAM）和直接内存总线随机存取存储器（Direct Rambus RAM，DR RAM）。

在本申请的一些实施例中，该计算机程序可以被分割成一个或多个模块，该一个或者多个模块被存储在该存储器 1010 中，并由该处理器 1020 执行，以完成本申请提供的方法。该一个或多个模块可以是能够完成特定功能的一系列计算机程序指令段，该指令段用于描述该计算机程序在该电子设备中的执行过程。

如图 10 所示，该电子设备还可包括：

收发器 1030，该收发器 1030 可连接至该处理器 1020 或存储器 1010。

其中，处理器 1020 可以控制该收发器 1030 与其他设备进行通信，具体地，可以向其他设备发送信息或数据，或接收其他设备发送的信息或数据。收发器 1030 可以包括发射机和接收机。收发器 1030 还可以进一步包括天线，天线的数量可以为一个或多个。

应当理解，该电子设备中的各个组件通过总线系统相连，其中，总线系统除包括数据总线之外，还包括电源总线、控制总线和状态信号总线。

本申请还提供了一种计算机存储介质，其上存储有计算机程序，该计算机程序被计算机执行时使得该计算机能够执行上述方法实施例的方法。或者说，本申请实施例还提供一种包含指令的计算机程序产品，该指令被计算机执行时使得计算机执行上述方法实施例的方法。

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。该计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行该计算机程序指令时，全部或部分地产生按照本申请实施例该的流程或功能。该计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。该计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一个计算机可读存储介质传输，例如，该计算机指令可以从一个网站站点、计算机、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（digital subscriber line，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、服务器或数据中心进行传输。该计算机可读存储介质可以是计算机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。该可用介质可以是磁性介质（例如，软盘、硬盘、磁带）、光介质（例如数字视频光盘（digital video disc，DVD））、或者半导体介质（例如固态硬盘（solid state disk，SSD））等。

本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的模块及算法步骤，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来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。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统、装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，该模块的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个模块或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或模块的间接耦合或通信连接，可以是电性，机械或其它的形式。

作为分离部件说明的模块可以是或者也可以不是物理上分开的，作为模块显示的部件可以是或者也可以不是物理模块，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。例如，在本申请各个实施例中的各功能模块可以集成在一个处理模块中，也可以是各个模块单独物理存在，也可以两个或两个以上模块集成在一个模块中。

以上该，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该权利要求的保护范围为准。

权利要求书

1、一种模型训练方法，其特征在于，包括：

获取语音识别样本和真实的语音到语音翻译样本；

根据所述语音识别样本生成伪标注的语音到语音翻译样本；

根据所述伪标注的语音到语音翻译样本和所述真实的语音到语音翻译样本训练语音到语音翻译模型。

2、根据权利要求 1 所述的方法，其特征在于，所述根据所述伪标注的语音到语音翻译样本和所述真实的语音到语音翻译样本训练语音到语音翻译模型，包括：

根据所述伪标注的语音到语音翻译样本预训练语音到语音翻译模型，并根据所述真实的语音到语音翻译样本微调预训练后的语音到语音翻译模型。

3、根据权利要求 2 所述的方法，其特征在于，所述根据所述真实的语音到语音翻译样本微调预训练后的语音到语音翻译模型，包括：

通过所述真实的语音到语音翻译样本微调预训练后的语音到语音翻译模型；或者，

根据所述真实的语音到语音翻译样本和所述伪标注的语音到语音翻译样本微调预训练后的语音到语音翻译模型。

4、根据权利要求 3 所述的方法，其特征在于，所述根据所述真实的语音到语音翻译样本和所述伪标注的语音到语音翻译样本微调预训练后的语音到语音翻译模型之前，还包括：

对所述真实的语音到语音翻译样本标注第一标签，所述第一标签用于标识所述真实的语音到语音翻译样本为真实样本；

对所述伪标注的语音到语音翻译样本标注第二标签，所述第二标签用于标识所述伪标注的语音到语音翻译样本为伪标注样本。

5、根据权利要求 3 或 4 所述的方法，其特征在于，所述根据所述真实的语音到语音翻译样本和所述伪标注的语音到语音翻译样本微调预训练后的语音到语音翻译模型，包括：

对所述真实的语音到语音翻译样本进行上采样，得到上采样后的语音到语音翻译样本；

通过所述上采样后的语音到语音翻译样本和所述伪标注的语音到语音翻译样本微调预训练后的语音到语音翻译模型。

6、根据权利要求 1-4 任一项所述的方法，其特征在于，所述真实的语音到语音翻译样本包括：第一源语言语音特征、第一源语言文本、第一目标语言语音特征和第一目标语言文本；所述语音识别样本包括：第二源语言语音特征和第二源语言文本。

7、根据权利要求 6 所述的方法，其特征在于，所述根据所述语音识别样本生成伪标注的语音到语音翻译样本，包括：

对所述第二源语言文本进行翻译，得到第二目标语言文本；

对所述第二目标语言文本进行合成，得到第二目标语言语音特征；

其中，所述伪标注的语音到语音翻译样本包括：所述第二源语言语音特征、所述第二源语言文本、所述第二目标语言文本和所述第二目标语言语音特征。

8、根据权利要求 1-4 任一项所述的方法，其特征在于，所述语音到语音翻译模型包括：编码器模块、第一注意力模块、第一解码器模块、N 个第二注意力模块和 N 个第二解码器模块，N 为正整数，所述 N 个第二注意力模块和所述 N 个第二解码器模块一一对应；

所述编码器模块用于获取源语言语音特征，并对所述源语言语音特征进行处理，得到所述源语言语音特征对应的多组第一隐藏状态表示；

所述第一注意力模块用于获取所述多组第一隐藏状态表示中的一组第一隐藏状态表示以及所述第一解码器输出的各个时间步对应的第一向量，并对该组第一隐藏状态表示和各个时间步对应的第一向量进行处理，得到各个时间步对应的第一注意力表示；

所述第一解码器模块用于获取所述各个时间步对应的第二向量，并对所述各个时间步对应的第二向量进行处理，得到各个时间步对应的第一向量，将各个时间步对应的第一向量输出给所述第一注意力模块，获取各个时间步对应的第一注意力表示，并对各个时间步对应的第一注意力表示进行处理，得到所述源语言语音特征对应的目标语言语音特征；

在对所述语音到语音翻译模型的训练阶段，所述第二注意力模块用于获取所述多组第一隐藏状态表示中的一组第一隐藏状态表示以及所述第二注意力模块对应的第二解码器输出的各个时间步对应的第三向量，并对该组第一隐藏状态表示和各个时间步对应的第三向量进行处理，得到各个时间步对应的第二注意力表示；

所述第二注意力模块对应的第二解码器模块用于获取各个时间步对应的第四向量，并对各个时间步对应的第四向量进行处理，得到各个时间步对应的第三向量，将各个时间步对应的第三向量输出给所述第二注意力模块，获取各个时间步对应的第二注意力表示，并对各个时间步对应的第二注意力表示进行处理，得到所述源语言语音特征对应的辅助表示。

9、根据权利要求 8 所述的方法，其特征在于，所述编码器模块包括：卷积神经网络子模块和第一转换器模块；

所述卷积神经网络子模块用于获取所述源语言语音特征，并对所述源语言语音特征进行处理，得到所述源语言语音特征对应的第二隐藏状态表示；

所述第一转换器模块用于获取所述第二隐藏状态表示，并对所述第二隐藏状态表示进行处理，得到所述多组第一隐藏状态表示。

10、根据权利要求 9 所述的方法，其特征在于，所述第一解码器模块包括：前处理网络、第二转换器模块和后处理网络；

所述前处理网络用于获取各个时间步对应的第二向量，并对各个时间步对应的第二向量进行处理，得到各个时间步对应的第一向量，将各个时间步对应的第一向量输出给所述第一注意力模块；

所述第二转换器模块用于获取各个时间步对应的第一注意力表示，并对各个时间步对应的第一注意力表示进行处理，得到各个时间步上的目标语言语音特征；

所述后处理网络用于对各个时间步上的目标语言语音特征进行处理，得到所述源语言语音特征对应的目标语言语音特征。

11、一种语音到语音翻译方法，其特征在于，包括：

获取源语言语音特征；

将所述源语言语音特征输入至如权利要求 1 至 10 中任一项方法训练得到的语音到语音翻译模型，得到所述源语言语音特征对应的目标语言语音特征。

12、一种模型训练装置，其特征在于，包括：

获取模块，用于获取语音识别样本和真实的语音到语音翻译样本；

生成模块，用于根据所述语音识别样本生成伪标注的语音到语音翻译样本；

训练模块，用于根据所述伪标注的语音到语音翻译样本和所述真实的语音到语音翻译样本训练语音到语音翻译模型。

13、一种语音到语音翻译装置，其特征在于，包括：

获取模块，用于获取源语言语音特征；

处理模块，用于将所述源语言语音特征输入至如权利要求 1 至 10 中任一项方法训练得到的语音到语音翻译模型，得到所述源语言语音特征对应的目标语言语音特征。

14、一种电子设备，其特征在于，包括：

处理器和存储器，所述存储器用于存储计算机程序，所述处理器用于调用并运行所述存储器中存储的计算机程序，以执行权利要求 1 至 11 中任一项所述的方法。

15、一种计算机可读存储介质，其特征在于，用于存储计算机程序，所述计算机程序使得计算机执行如权利要求 1 至 11 中任一项所述的方法。

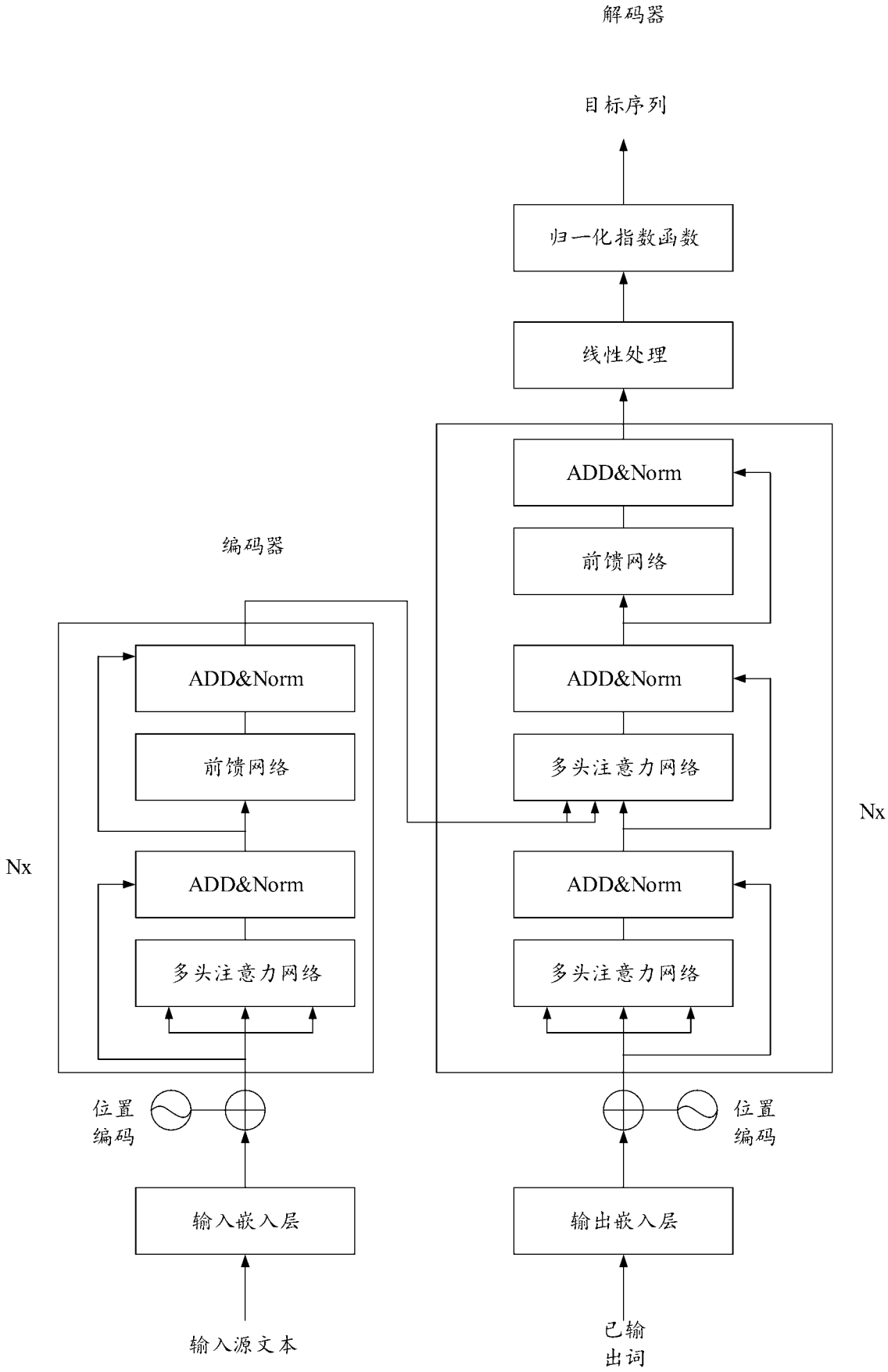

解码器
目标序列
归一化指数函数
线性处理
ADD&Norm
前馈网络
ADD&Norm
多头注意力网络
ADD&Norm
多头注意力网络
Nx
编码器
ADD&Norm
前馈网络
ADD&Norm
多头注意力网络
Nx
位置编码
位置编码
输入嵌入层
输出嵌入层
输入源文本
已输出词

图 1

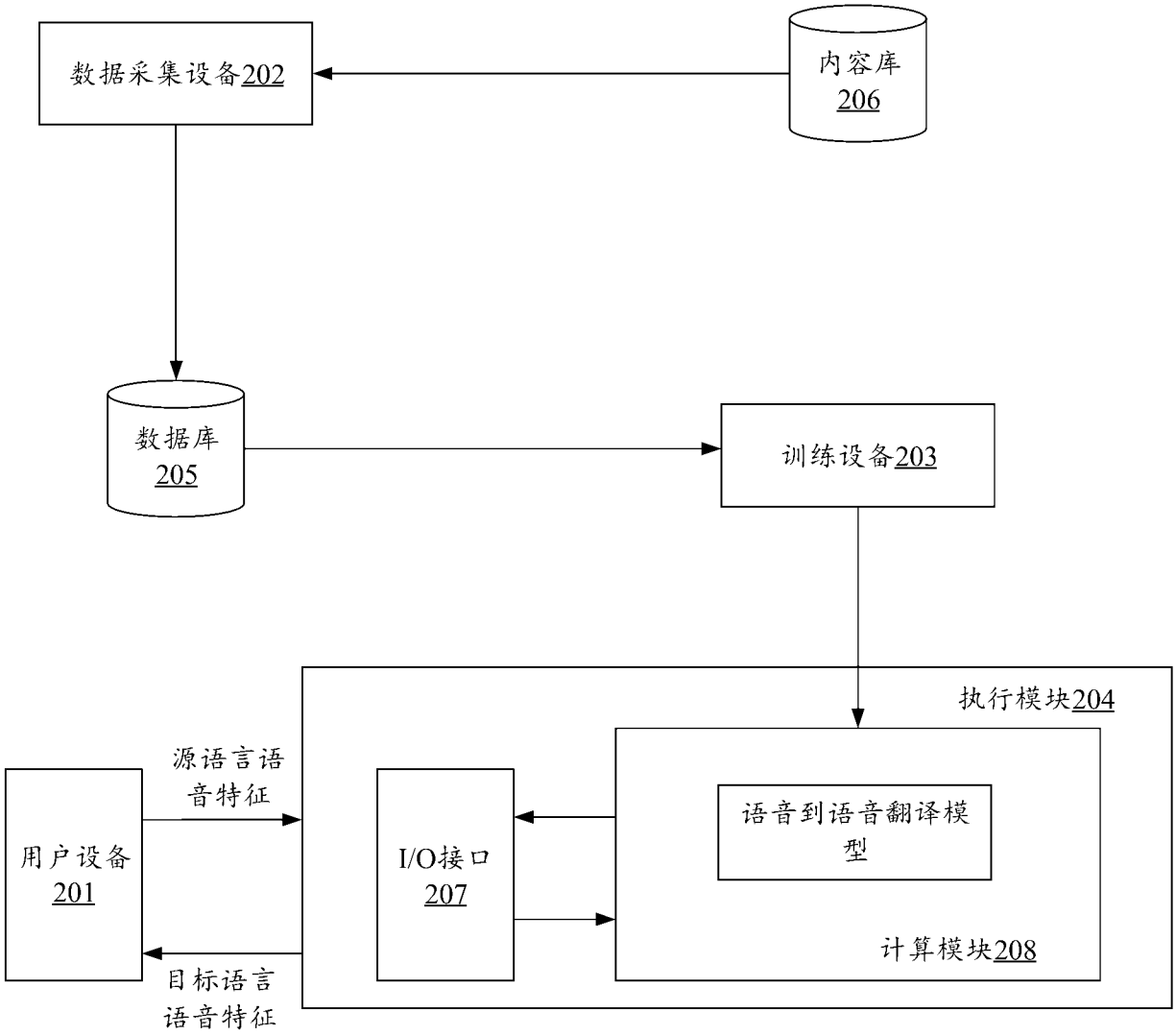

数据采集设备202
内容库
206
数据库
205
训练设备203
执行模块204
源语言语
音特征
用户设备
201
I/O接口
207
语音到语音翻译模
型
计算模块208
目标语言
语音特征

图 2

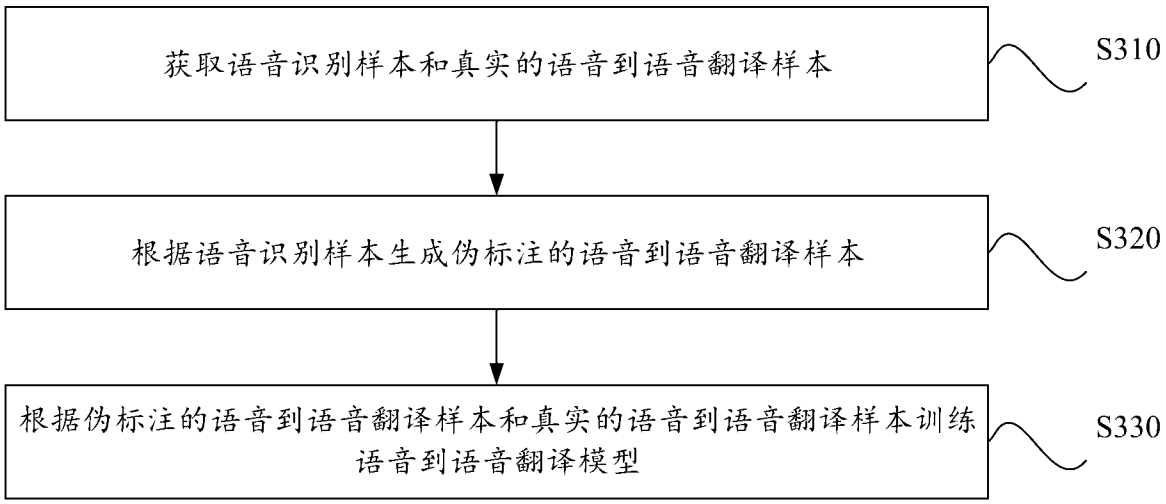

获取语音识别样本和真实的语音到语音翻译样本
S310
根据语音识别样本生成伪标注的语音到语音翻译样本
S320
根据伪标注的语音到语音翻译样本和真实的语音到语音翻译样本训练
语音到语音翻译模型
S330

图 3

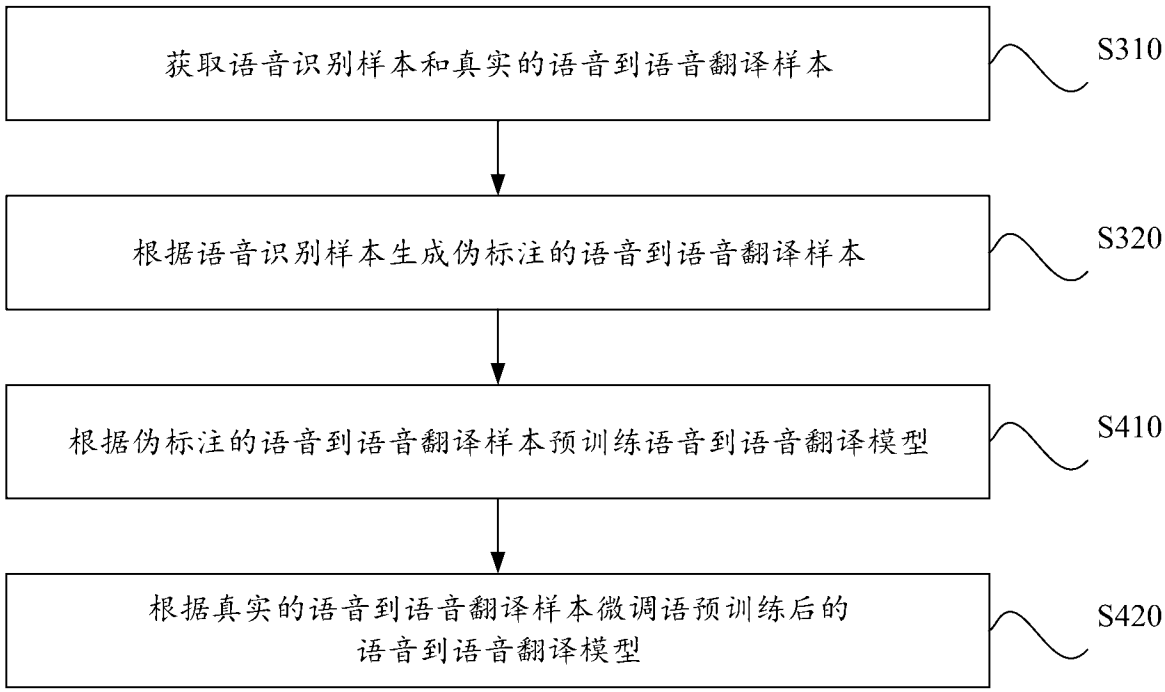

获取语音识别样本和真实的语音到语音翻译样本
S310
根据语音识别样本生成伪标注的语音到语音翻译样本
S320
根据伪标注的语音到语音翻译样本预训练语音到语音翻译模型
S410
根据真实的语音到语音翻译样本微调语预训练后的
语音到语音翻译模型
S420

图 4

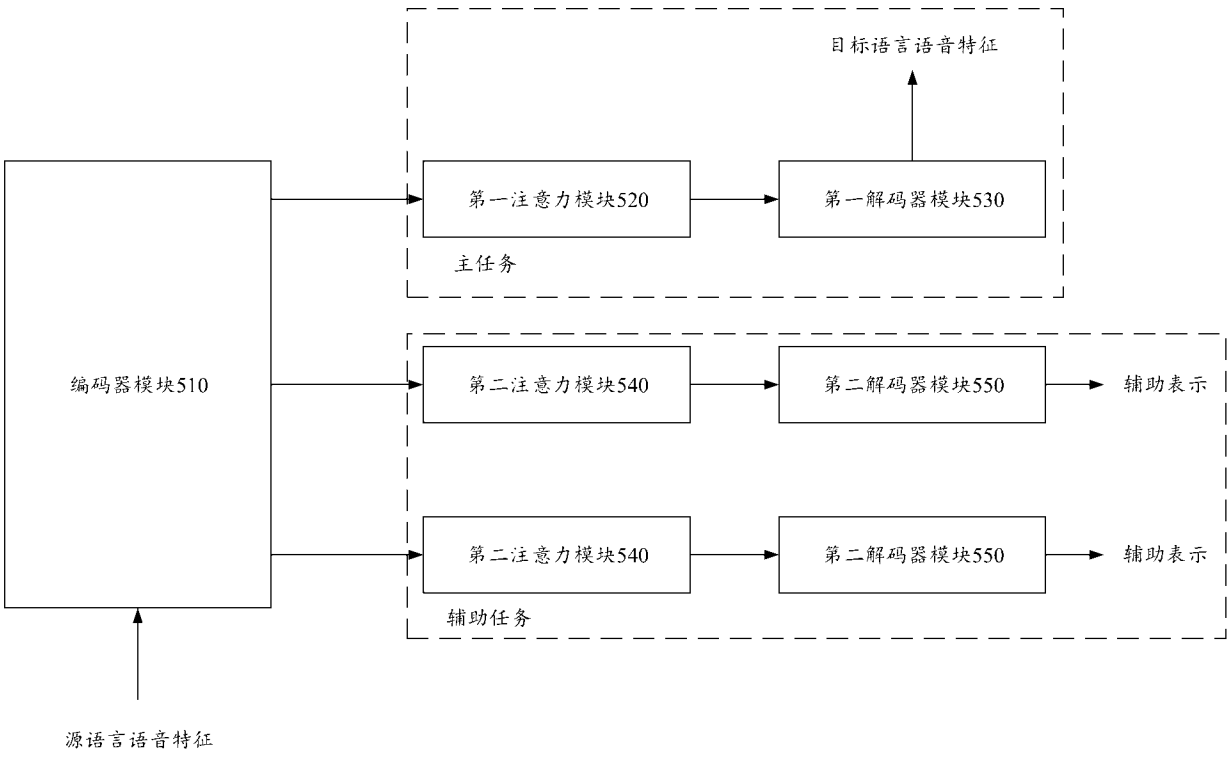

目标语言语音特征
第一注意力模块520
第一解码器模块530
主任务
编码器模块510
第二注意力模块540
第二解码器模块550
辅助表示
第二注意力模块540
第二解码器模块550
辅助表示
辅助任务
源语言语音特征

图 5

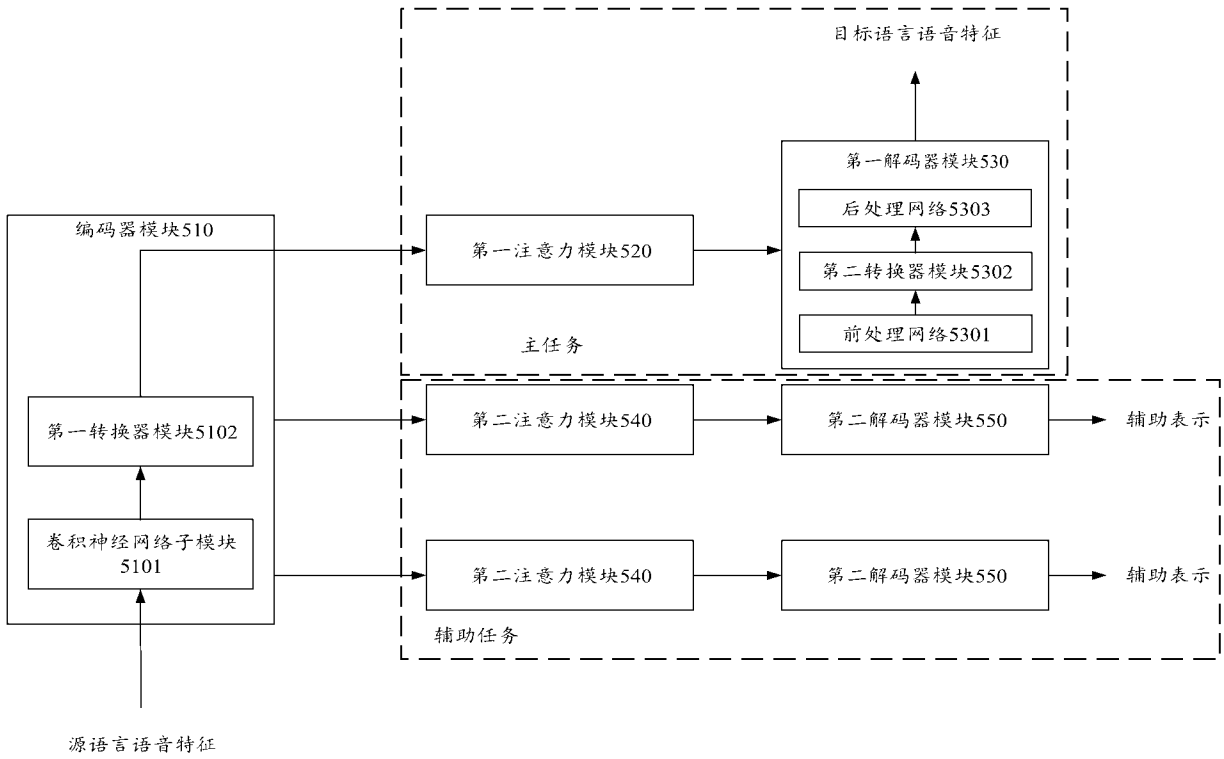

目标语言语音特征
第一解码器模块530
后处理网络5303
第二转换器模块5302
前处理网络5301
编码器模块510
第一注意力模块520
主任务
第一转换器模块5102
第二注意力模块540
第二解码器模块550
辅助表示
卷积神经网络子模块5101
第二注意力模块540
第二解码器模块550
辅助表示
辅助任务
源语言语音特征

图 6

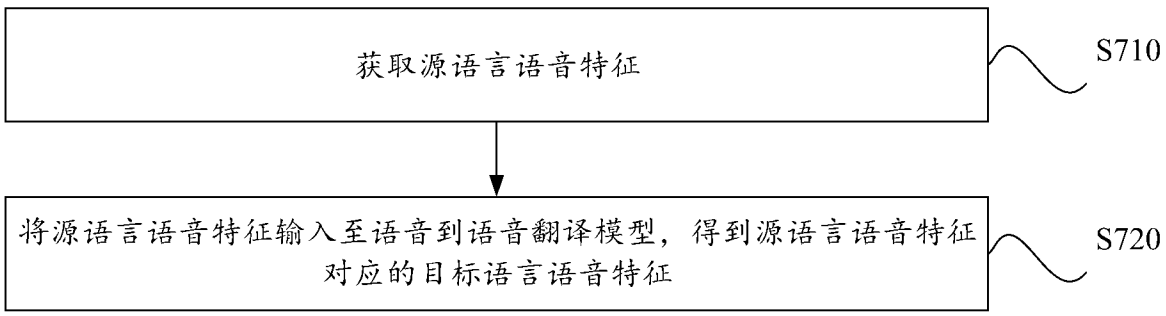

获取源语言语音特征
S710
将源语言语音特征输入至语音到语音翻译模型，得到源语言语音特征对应的目标语言语音特征
S720

图 7

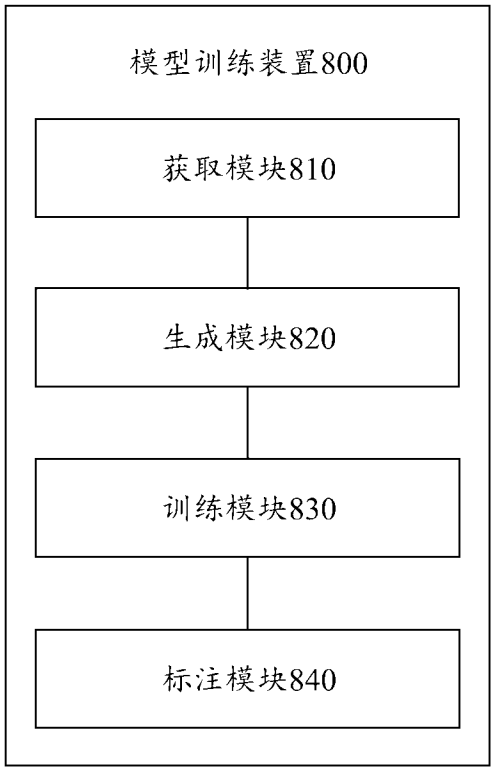

模型训练装置800
获取模块810
生成模块820
训练模块830
标注模块840

图 8

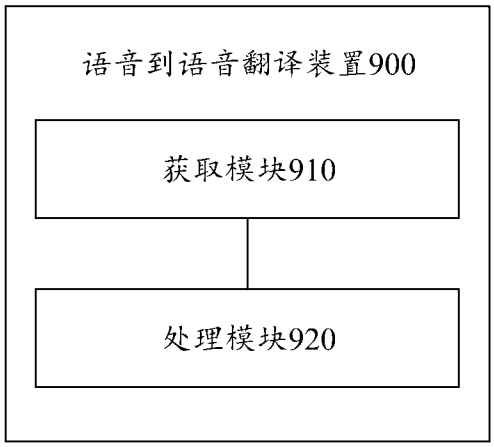

语音到语音翻译装置900
获取模块910
处理模块920

图 9

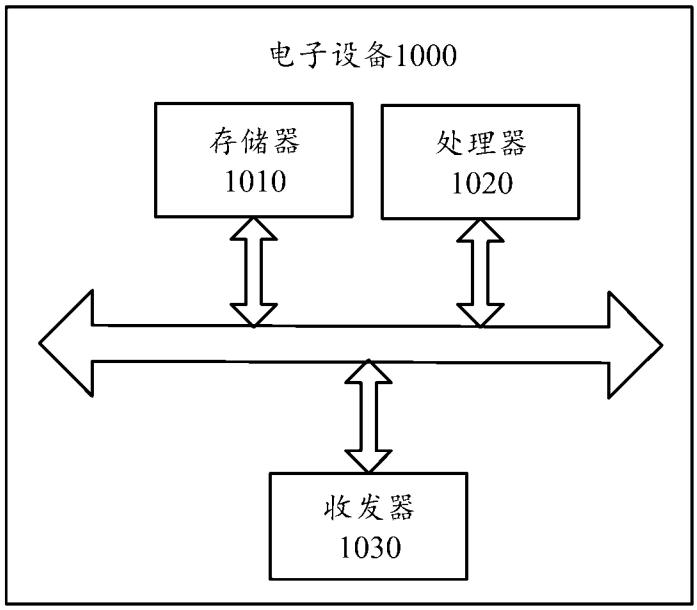


图 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/088492

A. CLASSIFICATION OF SUBJECT MATTER

G10L 15/06(2013.01)i; G10L13/02(2013.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, CNTXT, DWPI, CNKI, 百度学术, BAIDU SCHOLAR, IEEE: 翻译, 模型, 训练, 数据, 语料, 样本, 缺失, 匮乏, 稀少, speech to speech translation, S2ST, training data, paucity, data scarcity

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 114822499 A (BEIJING YOUZHUJU NETWORK TECHNOLOGY CO., LTD.) 29 July 2022 (2022-07-29) claims 1-15	1-15
A	CN 111738025 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.) 02 October 2020 (2020-10-02) entire document	1-15
A	CN 112966529 A (GLOBAL TONE COMMUNICATION TECHNOLOGY CO., LTD.) 15 June 2021 (2021-06-15) entire document	1-15
X	YE, Jia et al. "LEVERAGING WEAKLY SUPERVISED DATA TO IMPROVE END-TO-END SPEECH-TO-TEXT TRANSLATION" *ICASSP 2019*, 17 April 2019 (2019-04-17), sections 3-5	1-7, 12, 14-15

☑ 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
26 July 2023	**31 July 2023**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088**	Authorized officer Telephone No.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	YE, Jia et al. "LEVERAGING WEAKLY SUPERVISED DATA TO IMPROVE END-TO-END SPEECH-TO-TEXT TRANSLATION" *ICASSP 2019,* 17 April 2019 (2019-04-17), sections 3-5	11, 13
Y	YE, Jia et al. "TRANSLATOTRON 2: ROBUST DIRECT SPEECH-TO-SPEECH TRANSLATION" *arXiv: 2107.08661v4,* 03 December 2021 (2021-12-03), section 3	11, 13

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/088492

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 114822499 A	29 July 2022	None	
CN 111738025 A	02 October 2020	None	
CN 112966529 A	15 June 2021	None	

国际检索报告

国际申请号
PCT/CN2023/088492

A. 主题的分类

G10L 15/06(2013.01)i; G10L13/02(2013.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G10L

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNABS, CNTXT, DWPI, CNKI, 百度学术, IEEE: 翻译, 模型, 训练, 数据, 语料, 样本, 缺失, 匮乏, 稀少, speech to speech translation, S2ST, training data, paucity, data scarcity

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PX	CN 114822499 A (北京有竹居网络技术有限公司) 2022年7月29日 (2022-07-29) 权利要求1-15	1-15
A	CN 111738025 A (腾讯科技（深圳）有限公司) 2020年10月2日 (2020-10-02) 全文	1-15
A	CN 112966529 A (中译语通科技股份有限公司) 2021年6月15日 (2021-06-15) 全文	1-15
X	Ye Jia 等. "LEVERAGING WEAKLY SUPERVISED DATA TO IMPROVE END-TO-END SPEECH-TO-TEXT TRANSLATION" ICASSP 2019, 2019年4月17日 (2019-04-17), 第3-5部分	1-7,12,14-15
Y	Ye Jia 等. "LEVERAGING WEAKLY SUPERVISED DATA TO IMPROVE END-TO-END SPEECH-TO-TEXT TRANSLATION" ICASSP 2019, 2019年4月17日 (2019-04-17), 第3-5部分	11,13

☑ 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
2023年7月26日	2023年7月31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 游晓梅 电话号码 (+86) 010-62089539

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
Y	Ye Jia 等. "TRANSLATOTRON 2: ROBUST DIRECT SPEECH-TO-SPEECH TRANSLATION" arXiv:2107.08661v4, 2021年12月3日 (2021-12-03), 第3部分	11,13

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/088492

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	114822499	A	2022年7月29日	无	
CN	111738025	A	2020年10月2日	无	
CN	112966529	A	2021年6月15日	无	