

说明书

一种数字人唇形生成方法及装置

本申请要求于2023年07月12日提交中国国家知识产权局、申请号为2023108555002、
申请名称为“一种数字人唇形生成方法及装置”的中国专利申请的优先权，其全部内容通
5 过引用结合在本申请中。

技术领域

本说明书一个或多个实施例涉及虚拟数字人领域，尤其涉及一种数字人唇形生成方法
及装置。

10

背景技术

近年来，虚拟数字人在虚拟现实、数字营销、元宇宙、医疗、教育等领域得到了广泛
的应用。评价数字人效果的重要指标是各个维度上的体验是否自然，包含形象质感、语音、
动作和唇形等等。其中，唇形是指数字人在语音播报时，是否在静默的时候准确地执行“闭
15 口”动作，是否在读不同发音文字的时候执行正确的“嘴型”。人工手动为数字人播报的
每一段音频配置唇形的成本过高，且不具备可扩展性。因此，在对数字人进行唇形配置的
环节中需要实现自动编排。

发明内容

20 本说明书一个或多个实施例描述了一种数字人唇形生成方法及装置，旨在对数字人唇
形实现自动化配置。

第一方面，提供了一种数字人唇形生成方法，包括：

获取文本数据与对应的音频数据；

25 将所述文本数据与音频数据在时间维度上对齐，得到所述文本数据对应的文本单位集
合，以及所述文本单位集合中任一文本单位所对应的音频片段和时间段；

对于在目标时间段上的目标文本单位，根据第一字典，确定所述目标文本单位所对应
的目标口型信息，其中，所述第一字典中包含任一文本单位所对应的口型信息；

根据所述目标文本单位所对应的目标音频片段，确定所述目标口型信息对应的目标幅
度信息；

30 根据所述目标口型信息和目标幅度信息，确定在所述目标时间段上的目标唇形。

在一种可能的实施方式中，所述音频数据为预先录制音频，所述文本单位为字；将所
述文本数据与音频数据在时间维度上对齐，得到所述文本数据对应的文本单位集合，以及

所述文本单位集合中任一文本单位所对应的音频片段和时间段，包括：

根据所述文本数据与音频数据，使用强制对齐模型生成字集合，以及所述字集合中任一字所对应的音频片段和时间段。

- 5 在一种可能的实施方式中，所述音频数据为预先录制音频，所述文本单位为音素；将所述文本数据与音频数据在时间维度上对齐，得到所述文本数据对应的文本单位集合，以及所述文本单位集合中任一文本单位所对应的音频片段和时间段，包括：

根据所述文本数据中的任一字，确定所述字对应的音素，得到音素集合；

根据所述音素集合与音频数据，使用强制对齐模型生成所述音素集合中任一音素所对应的音频片段和时间段。

- 10 在一种可能的实施方式中，所述音频数据为使用文本转语音模型根据所述文本数据生成的，所述文本转语音模型在生成所述音频数据的同时，还生成了所述文本数据对应的文本单位集合以及所述文本单位集合中任一文本单位所对应的音频片段和时间段；

将所述文本数据与音频数据在时间维度上对齐，得到所述文本数据对应的文本单位集合，以及所述文本单位集合中任一文本单位所对应的音频片段和时间段，包括：

- 15 从所述文本转语音模型的生成结果中获取所述文本单位集合以及所述文本单位集合中任一文本单位所对应的音频片段和时间段。

在一种可能的实施方式中，根据所述目标文本单位所对应的目标音频片段，确定所述目标口型信息对应的目标幅度信息，包括：

- 20 根据所述目标音频片段的音频特征，确定所述目标幅度信息，所述音频特征包括音量特征、基频特征和共振峰特征中的至少一种。

在一种可能的实施方式中，所述音频特征为音量特征；根据所述目标音频片段的音频特征，确定所述目标幅度信息，包括：

根据所述目标文本单位所在的特定文本范围中的最大音量值和最小音量值，确定所述目标幅度信息。

- 25 在一种可能的实施方式中，所述目标文本单位所在的特定文本范围为所述目标文本单位所在句子或段落。

在一种可能的实施方式中，还包括：

根据多个所述目标唇形组成唇形集合，所述唇形集合包含在时间维度上相邻的第一唇形和第二唇形；

- 30 根据所述第一唇形和第二唇形，对所述第一唇形与所述第二唇形之间的空白时间段进

行插值操作，得到第一唇形序列。

在一种可能的实施方式中，所述插值操作为线性插值。

在一种可能的实施方式中，根据所述第一唇形和第二唇形，对所述第一唇形与所述第二唇形之间的空白时间段进行插值操作，得到第一唇形序列，包括：

- 5 根据所述第一唇形和第二唇形，确定所述空白时间段中任一最小时间单位所对应的唇形，得到第一唇形序列。

在一种可能的实施方式中，还包括：

对所述第一唇形、第二唇形和第一唇形序列所组成的目标唇形序列在时间维度上进行平滑操作，得到目标平滑唇形序列。

- 10 在一种可能的实施方式中，所述平滑操作为指数平滑或一维高斯滤波平滑。

第二方面，提供了一种数字人唇形生成装置，包括：

获取单元，配置为，获取文本数据与对应的音频数据；

对齐单元，配置为，将所述文本数据与音频数据在时间维度上对齐，得到所述文本数据对应的文本单位集合，以及所述文本单位集合中任一文本单位所对应的音频片段和时间
15 段；

口型确定单元，配置为，对于在目标时间段上的目标文本单位，根据第一字典，确定所述目标文本单位所对应的目标口型信息，其中，所述第一字典中包含任一文本单位所对应的口型信息；

- 20 幅度确定单元，配置为，根据所述目标文本单位所对应的目标音频片段，确定所述目标口型信息对应的目标幅度信息；

唇形确定单元，配置为，根据所述目标口型信息和目标幅度信息，确定在所述目标时间段上的目标唇形。

在一种可能的实施方式中，还包括：

- 25 集合确定单元，配置为，根据多个所述目标唇形组成唇形集合，所述唇形集合包含在时间维度上相邻的第一唇形和第二唇形；

插值单元，配置为，根据所述第一唇形和第二唇形，对所述第一唇形与所述第二唇形之间的空白时间段进行插值操作，得到第一唇形序列。

在一种可能的实施方式中，还包括：

- 30 平滑单元，配置为，对所述第一唇形、第二唇形和第一唇形序列所组成的目标唇形序列在时间维度上进行平滑操作，得到目标平滑唇形序列。

第三方面，提供了一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行第一方面的方法。

第四方面，提供了一种计算设备，包括存储器和处理器，其中，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现第一方面的方法。

5 本说明书实施例提出的一种数字人唇形生成方法及装置，针对数字人离线视频制作和实时播报两类不同的应用场景，在不使用标注的音频和文本对齐信息的情况下，实现对开闭口和各种唇型的精细控制。

附图说明

10 为了更清楚地说明本说明书披露的多个实施例的技术方案，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说明书披露的多个实施例，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它的附图。

- 图 1 示出根据一个实施例的数字人唇形生成方法的实施场景示意图；
- 15 图 2 示出根据一个实施例的数字人唇形生成方法的流程图；
- 图 3 示出根据一个实施例的唇形插值的示意图；
- 图 4 示出根据另一个实施例的唇形插值的示意图；
- 图 5 示出根据一个实施例的数字人唇形生成装置的示意性框图。

20 具体实施方式

下面结合附图，对本说明书提供的方案进行描述。

数字人又称虚拟形象，是运用数字技术创造出来的，与人类形象接近的数字化人物形象。唇形驱动是数字人的关键能力，指的是在数字人进行语音播报的时候，嘴部动画随音频进行合理的变化，达到音频与嘴部动画同步的效果。由于在客服播报、智能助手等数字人场景中，文本和音频都是不断变化、无法枚举的，因此无法用人工制作的方式，复刻全部的唇形情况。此时，就需要使用一定的方法自动化地根据音频和文本内容，实时生成唇形序列。

图 1 示出根据一个实施例的数字人唇形生成的实施场景示意图。在图 1 的示例中，输入数据包括文本数据和与其对应的音频数据，其中，音频数据可以是预先录制的，也可以是机器根据文本数据使用文本转语音（Text To Speech, TTS）模型生成的。将文本数据和音频数据输入到对齐模块之后，对齐模块首先判断音频数据是预先录制的，还是根据文本

使用 TTS 实时生成的。如果音频数据是预先录制的, 此时的应用场景为离线视频制作场景, 则对齐模块使用强制对齐 (Forced Alignment) 模型将文本和音频对齐, 将文本数据拆分为多个文本单位, 同时获得每个文本单位对应的音频片段和时间段; 如果音频数据是使用 TTS 模型生成的, 此时的应用场景为在线实时播报场景, TTS 模型在生成音频的同时就能输出任一文本单位对应的时间段, 对齐模块只需要从 TTS 模型的生成结果中获取任一文本单位对应的时间段即可, 时间段可以根据时间段的开始时间戳和结束时间戳表示。在获得文本单位对应的时间段后, 便可以根据时间段将音频拆分为对应的音频片段, 此时便能得到任一文本单位所对应的音频片段和时间段。

在得到任一文本单位所对应的音频片段和时间段之后, 使用预先配置好的文本-口型字典, 根据文本单位获得对应的口型信息; 根据音频片段的特定音频特征获取该口型对应的幅度信息, 综合口型信息和幅度信息, 便可以得到一个时间段上的对应的唇形。将不同时间段上的唇形按照时间顺序进行排列后, 便可以得到初步的唇形序列。

对这个唇形序列在相邻唇形之间的空白时间段进行插值和平滑操作, 实现不同唇形之间的过渡, 从而使全局的唇形动画更加流畅自然, 得到最终的平滑唇形序列。在得到最终的平滑唇形序列后, 将平滑唇形序列发送给数字人唇形动画生成模块 (图中未示出), 便可以根据唇形序列生成数字人在每个时间点对应的正确且流畅自然的唇形动画。

以下结合具体的实施例, 描述上述唇形生成方法的具体实施步骤。图 2 示出根据一个实施例的数字人唇形生成方法的流程图, 所述方法的执行主体可以为任何具有计算、处理能力的平台或服务器或设备集群等。

在步骤 202, 获取文本数据与对应的音频数据。

音频数据可以是预先录制的, 也可以是机器根据文本数据使用 TTS 模型实时生成的。

在使用 TTS 模型根据文本数据实时生成音频数据时, 首先将文本数据拆分成若干音素。音素是根据语音的自然属性划分出来的基本语音单位, 指的是一组具有区别性的语音单元, 是构成语言语音的最小可辨别的音段。在将文本数据拆分成若干音素之后, TTS 模型通过预测任一音素在整个音频中相应的时间段, 利用数字信号处理和语音合成相关技术, 生成相应的音频数据。音素的时间段可以由开始时间戳和结束时间戳表示。根据开始时间戳和结束时间戳, 便可以得到该音素在音频数据中对应的音频片段。在一个实施例中, 使用非回归 TTS 模型中的时长预测模块 (duration predictor) 预测任一音素的帧数以及开始帧和结束帧, 通过将帧数乘以每一帧的持续时长, 便能得到该音素的持续时长。

当音频数据为预先录制的时, 无法同时获取到音素对应的时间段。预先录制音频可以

是由人工朗读文本，然后经由音频采集设备将人声录制下来，得到相应的音频数据；也可以是使用有限制的 TTS 模型根据文本数据生成的。在使用有限制的 TTS 模型时，无法使用 TTS 模型中与时长相关的具体模块。例如，TTS 模型为上游服务商提供的接口，或者 TTS 模型已经被导出为可执行文件的形式，此时，无法根据 TTS 模型得到音素对应的时间段。

在步骤 204，将所述文本数据与音频数据在时间维度上对齐，得到所述文本数据对应的文本单位集合，以及所述文本单位集合中任一文本单位所对应的音频片段和时间段。

文本单位是在实施例中由人为设定的，不可再次拆分的最小的文本单元，例如，在中文文本中，文本单位可以是单个汉字，也可以是中文的音素，在英文文本中，文本单位可以是单词，也可以是英文的音素。

在一个实施例中，音频数据为预先录制的音频，文本单位为字。字指的是，在所述文本数据所包含的任意一种语言中，比该语言的音素大一级的语言单位，例如，在中文中指的是单个汉字，在英文中指的是单个英文单词。根据所述文本数据与音频数据，使用强制对齐模型生成字集合，以及所述字集合中任一字所对应的音频片段和时间段。

强制对齐模型可以是任意现有的模型，例如，可以是 MFA (Montreal Forced Aligner，蒙特利尔强制对齐器)，或者预训练声学模型 Wav2Vec2 中与强制对齐有关的模块，这里不做限定。强制对齐模型通过统计音频数据的特征与文本序列的共现和相关性，将文本中的词组、字或音素与音频中对应的语音片段进行对齐，从而使得文本和语音之间的对应关系更加准确和精确。

将文本数据和音频数据输入到强制对齐模型中，得到文本数据中任一字的开始时间戳和结束时间戳后，根据任一字的开始时间戳和结束时间戳，便可以从音频数据中分割出该字对应的音频片段。

在另一个实施例中，音频数据为预先录制的音频，文本单位为音素。

首先，根据所述文本数据中的任一字，确定所述字对应的音素，得到音素集合。

根据文本数据中不同语言的文本，使用相应的方法将字拆分为音素。

在一个具体的实施例中，当文本语言为中文时，由于中文中存在多音字，对于任一字，首先使用多音字集合判断该字为单音字还是多音字。由于汉字中的多音字数量是有限的，所以可以预先构造多音字集合。

当该字为单音字时，使用拼音库将单音字转换为对应的汉语拼音，然后再根据预设的拆分规则，将拼音拆分为对应的音素。例如，先使用拼音库将单音字“你”转换为“ni”，

再根据预设的拆分规则拆分为音素“n”和“i”。所使用的拼音库可以是现有的开源库，例如 pypinyin，也可以是根据具体需求专门定制的，这里不做限定。

需要说明的是，由于汉语拼音是有限的，而汉语中的音素的数量也是有限的，所以可以预先制定拆分规则，建立任一汉语拼音与其对应的音素之间的对应关系，形成拆分规则。

5 同时，由于声调对数字人唇形的影响十分有限，故在这个过程中不对不同的声调做区分处理。

当该字为多音字时，无法直接使用拼音库将字转换为对应的汉语拼音。此时，使用多音字预测模型，根据该字所处的上下文信息，预测该字的读音。例如，使用预训练的基于 BERT 的文本分类模型，将该字所在的句子或段落输入到文本分类模型中，以预测该字的
10 读音。在得到该多音字的正确读音后，使用与前述对单音字相似的处理方案将该多音字拆分为对应的音素。

在一个具体的实施例中，当文本语言为英文时，首先将英文文本拆分为英文单词的集合，对于任一英文单词，使用音素转写模型将其拆分为对应的音素。音素转写模型可以是现有的模型，例如 CMU Pronouncing Dictionary, Arpabet 等，也可以是根据具体需求专门
15 定制的，这里不做限定。

当文本数据中有其它语言时，也可以使用该语言相应的现有技术将字拆分为音素，在此不再赘述。

在得到音素集合后，根据所述音素集合与音频数据，使用强制对齐模型生成所述音素集合中任一音素所对应的音频片段和时间段。此处和上一个文本单位为字的实施例中的方案类似，参见上一个实施例的部分说明即可，在此不再赘述。
20

在又一个实施例中，音频数据为使用 TTS 模型根据文本数据实时生成的。由于 TTS 模型在步骤 202 中生成所述音频数据的同时，已经生成了音频数据中任一音素对应的时间段以及该时间段对应的音频片段。所以在此处只需从 TTS 模型的生成结果中获取相应数据即可。如果文本单位为字，则将任一字对应的若干音素的时间段拼接起来，即可得到该字
25 的时间段；将任一字对应的若干音频片段拼接起来，或者根据时间段对音频数据重新分割，即可得到该字对应的音频片段。

在步骤 206，对于在目标时间段上的目标文本单位，根据第一字典，确定所述目标文本单位所对应的目标口型信息，其中，所述第一字典中包含任一文本单位所对应的口型信息。

30 第一字典可以是现有的文本-口型字典，也可以是根据具体需求专门定制的，这里不做

限定。当定制时，让人穿上动作捕捉设备，通过动作捕捉设备捕捉和解析人在读出任一文本单位时所对应的口型信息，建立第一字典。第一字典可以基于 blend shape 建立，也可以基于其它任意动画技术建立，这里不做限定。

在步骤 208，根据所述目标文本单位所对应的目标音频片段，确定所述目标口型信息
5 对应的目标幅度信息。

具体地，根据所述目标音频片段的音频特征，确定所述目标幅度信息，所述音频特征包括音量特征、基频特征和共振峰特征中的至少一种。

在一个实施例中，单独使用文本单位对应的音量特征来确定目标幅度信息。对于任一文本单位，首先确定其所在一个特定的文本范围内的最大音量值 X_{max} 和最小音量值 X_{min} 。
10 特定的文本范围可以是该文本单位所在的句子或段落，或其它预设的范围。由于在一段较长的音频中，可能会存在某一个片段的音频特征与某一个片段的音频特征差距较大的情况，例如，对于音量特征而言，存在某一个片段整体音量较大，而另一个片段的整体音量较小的情况，如果只考虑整个音频的最大音量值 X_{max} 和最小音量值 X_{min} ，则在音量较小的部分会出现目标幅度过小，导致最终生成的动画口部运动不明显的情况。通过设定特定的文本
15 范围，分段确定音频特征的最大值和最小值，可以避免以上的问题。

在确定了特定的文本范围内的 X_{max} 和 X_{min} 之后，对于在该文本范围内的音量值为 X 的文本单位，其幅度信息可以用幅度值 P 表示，计算 P 的方法如公式（1）所示：

$$P = \alpha + \alpha \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

其中， α 为一超参数，用于控制幅度值 P 的最小值。在公式（1）中，P 的取值范围为
20 $[\alpha, 2\alpha]$ 。优选地， α 的取值范围为 0.5-0.8。

在其它实施例中，还可以参照上述实施例中的方法，使用基频特征、共振峰特征、或者它们与音量特征之间任意的组合特征，确定目标口型信息对应的目标幅度信息。

在步骤 210，根据所述目标口型信息和目标幅度信息，确定在所述目标时间段上的目标唇形。

25 对于目标文本单位，在确定其对应的目标口型信息和目标幅度信息后，便可以确定其对应的目标唇形。

在一个实施例中，使用类似 one-hot 独热编码的方式对口型信息进行编码，得到一个一维向量，向量在某个维度为 1，其余的维度为 0。向量的维度与口型总数相等。例如，口型 A 对应[1, 0, 0, ..., 0]，口型 B 对应[0, 1, 0, ..., 0]，口型 H 对应[0, 0, ..., 1, 0]。

30 然后，将口型信息对应的向量与目标幅度相乘，得到目标唇形对应的向量。例如，某

一文本单位的目标口型为口型 A，目标幅度为 0.8，则其目标唇形对应的向量为[0.8, 0, 0, ...,0]；另一文本单位的目标口型为口型 B，目标幅度为 0.5，则其目标唇形对应的向量为[0, 0.5, 0, ...,0]。

对于任一唇形对应的向量，向量在任一维度上的值代表该唇形在该维度对应的口型上的幅度。例如，唇形向量[0, 0.5, 0, ...,0]代表口型 B 的幅度为 0.5，而其余口型的幅度均为 0；唇形向量[0.2, 0.4, 0, ...,0]代表口型 A 的幅度为 0.2，口型 B 的幅度为 0.4，而其余口型的幅度均为 0。

在其他实施例中，还可以设计其它表示唇形的方式，只要能够对不同的唇形有着不同的表示即可。例如定义其它的对口型信息的编码方式，以及定义其它的将口型信息和幅度信息结合的方式，例如定义为二维向量，第一行为口型信息的编码向量，第二行为幅度信息对应的值或编码向量。

通过步骤 202-步骤 210 可以确定文本中单个文本单位对应的唇形。通过将一段文本里的多个文本单位对应的唇形按照时间顺序排列，便可得到该段文本对应的出唇形序列。

在一些可能的实施方式中，还包括：

步骤 212，根据多个所述目标唇形组成唇形集合，所述唇形集合包含在时间维度上相邻的第一唇形和第二唇形。

步骤 214，根据所述第一唇形和第二唇形，对所述第一唇形与所述第二唇形之间的空白时间段进行插值操作，得到第一唇形序列。

由于在说话时，在相邻的发音之间会存在静默的空白时间，如果在空白时间直接将唇形设置成闭口状态，那么最后生成的唇形动画会显得极为生硬和突兀，即在有声音的时间突然张口，在没有声音的空白时间突然闭口。所以需要在相邻的文本单位之间的空白时间设置过度唇形，使得最后生成的唇形动画看上去更加自然。

具体地，第一唇形对应着第一文本单位，第二唇形对应第二文本单位，通过预先设置时间的最小时间单位，将空白时间分割成若干个最小时间单位。最小时间单位可以根据实际需要任意设置，例如使用一帧的持续时长作为最小时间单位，或者设置成 0.05 秒、0.02 秒等。然后根据第一唇形和第二唇形，对它们之间的空白时间进行插值，确定空白时间段中任一最小时间单位所对应的唇形，得到第一唇形序列。

在一个实施例中，第一文本单位和第二文本单位位于同一个句子中。不妨假设第一唇形时间顺序在前，第二唇形时间顺序在后。将空白时间分割为若干个最小时间单位，使用线性插值根据第一唇形和第二唇形对空白时间进行插值，得到任一最小时间单位对应的唇

形，将唇形组合起来，得到第一唇形序列。

示例性地，在一个具体的例子中，假设共有 5 种唇形，第一唇形为[0, 1, 0, 0, 0]，第二唇形为[0, 0, 0, 0, 0.8]，最小时间单位为一帧的持续时长，第一唇形和第二唇形之间有 3 帧的空白时间，即[0, 0, 0, 0, 0]，则插值前的唇形序列为：

5 [0, 1, 0, 0, 0],
 [0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0.8]

10 使用线性插值根据第一唇形和第二唇形对空白时间进行插值，得到插值后的第一唇形序列为：

 [0, 1, 0, 0, 0],
 [0, 0.75, 0, 0, 0.2],
 [0, 0.5, 0, 0, 0.4],
15 [0, 0.25, 0, 0, 0.6],
 [0, 0, 0, 0, 0.8]

得到的第一唇形序列的示意图如图 3 所示。

在另一个实施例中，第一文本单位和第二文本单位位于不同的句子中，第一文本单位位于前一个句子的末尾，第二文本单位位于后一个句子的开头。此时，由于句子与句子之间会有一个较长的停顿，应该在停顿的部分时间设置有闭口唇形，如果参照前一实施例将空白时间段的任一最小时间单位都配置上唇形，则会出现第一唇形到第二唇形之间的过渡变换过于缓慢，并且没有闭口唇形的情况。

此时，首先将第一唇形和第二唇形之间的空白时间从中间切分为两段，前半段属于第一文本单位所在句子的结尾空白时间，后半段属于第二文本单位所在句子的开头空白时间。

25 对于第一文本单位之后的结尾空白时间，按照固定速率顺时间顺序进行插值；对于位于句子开头的第二文本单位，按照固定速率逆时间顺序进行插值。速率为一预设值，代表唇形幅度在每个时间单位减少的量。顺时间顺序插值即将某一唇形之后的时间单位的唇形按照时间顺序依次减小；逆时间顺序插值即将某一唇形之前的时间单位的唇形按照时间反向的顺序依次减小。

30 示例性地，在一个具体的例子中，初始情况下，如图 4 (a) 所示，第一唇形和第二唇形之间有着空白时间，空白时间包含多个时间单位，空白时间从中间被分为结尾空白时间和开头空白时间。假设第一唇形为[0, 0.6, 0, 0, 0]，第一唇形之后的结尾空白时间有 4 帧，

速率为 0.2/帧，则顺时间顺序插值之后的唇形序列为：

[0, 0.6, 0, 0, 0],
[0, 0.4, 0, 0, 0],
[0, 0.2, 0, 0, 0],
5 [0, 0, 0, 0, 0],
[0, 0, 0, 0, 0]

假设第二唇形为[0, 0.4, 0, 0, 0]，第二唇形之前的开头空白时间有 4 帧，则逆时间顺序插值之后的唇形序列为：

[0, 0, 0, 0, 0],
10 [0, 0, 0, 0, 0],
[0, 0, 0, 0, 0],
[0, 0.2, 0, 0, 0],
[0, 0.4, 0, 0, 0]

将结尾空白时间和开头空白时间的唇形序列按照时间顺序拼接，得到第一唇形序列。

15 得到的第一唇形序列的示意图如图 4 (b) 所示。

在其他实施例中，还可以使用其他的插值方法，例如二阶插值、三阶插值、样条插值等，只要能合理地设置空白时间的过度唇形即可。

通过步骤 212-步骤 214 可以在唇形之间的空白时间设置合理的过度唇形，使得根据整体的唇形序列生成的唇形动画相较于没有过度唇形的情况更加自然。

20 在一些可能的实施方式中，还包括：

步骤 216，对所述第一唇形、第二唇形和第一唇形序列所组成的目标唇形序列在时间维度上进行平滑操作，得到目标平滑唇形序列。

在通过插值操作设置过度唇形后，可能仍然会存在相邻时间单位的唇形变化不够自然的情况，通过使用平滑操作来调整唇形的变换轨迹，使其更加平滑流畅，避免突兀的变化。

25 具体地，目标唇形序列为二维向量，纵向为时间维度，横向为幅度维度，向量在任一位置上的值代表该唇形序列在该时间点上的该维度对应的口型的幅度。对目标唇形序列的任一幅度维度在多个时间点上形成的一维幅度序列进行平滑操作，综合所有幅度维度的平滑结果，得到目标平滑唇形序列。使用的平滑操作为指数平滑、一维高斯滤波平滑或其他任意平滑方法。

30 示例性的，在一个具体的例子中，对目标唇形序列在第 2 列在连续 12 个时间单位上的幅度序列[10, 12, 13, 15, 17, 20, 19, 18, 16, 14, 12, 11]进行指数平滑，得到平滑后的序列[10.00, 10.30, 10.91, 12.13, 13.29, 15.01, 16.20, 16.94, 16.46, 15.12, 13.88, 12.82]。通过相似的

方法计算出其他列的平滑后序列，再按照原有的顺序将其还原成二位向量，即得到目标平滑唇形序列。

本说明书一个或多个实施例只需要获取音频和文本对齐后的时间戳，针对实时播报与离线生产这两类数字人动画场景，构造了不同的唇形驱动方案。本说明书一个或多个实施例不涉及有监督数据的收集，因此几乎没有数据成本；同时，单实例推理时间短，需要的存储空间少，方案的并发量更高，故能够大大降低部署成本。通过在特定时间区间内编排唇形，确保了在每一个时间点都有正确的唇形出现，因而可以极大降低错误率。另外，本说明书一个或多个实施例基于时间戳和唇形规则的匹配，能够几乎零成本的迁移到不同的数字人，并适配不同的唇形方案，摆脱了数据和特定虚拟人带来的限制，跳过了漫长、高成本的数据收集环节，因此有着很高的兼容性。

根据另一方面的实施例，还提供一种数字人唇形生成装置。图5示出根据一个实施例的数字人唇形生成装置的示意性框图，该装置可以部署在任何具有计算、处理能力的设备、平台或设备集群中。如图5所示，该装置500包括：

获取单元501，配置为，获取文本数据与对应的音频数据；
15 对齐单元502，配置为，将所述文本数据与音频数据在时间维度上对齐，得到所述文本数据对应的文本单位集合，以及所述文本单位集合中任一文本单位所对应的音频片段和时间段；

口型确定单元503，配置为，对于在目标时间段上的目标文本单位，根据第一字典，确定所述目标文本单位所对应的目标口型信息，其中，所述第一字典中包含任一文本单位所对应的口型信息；
20

幅度确定单元504，配置为，根据所述目标文本单位所对应的目标音频片段，确定所述目标口型信息对应的目标幅度信息；

唇形确定单元505，配置为，根据所述目标口型信息和目标幅度信息，确定在所述目标时间段上的目标唇形。

25 在一些可能的实施方式中，装置500还包括：

集合确定单元506，配置为，根据多个所述目标唇形组成唇形集合，所述唇形集合包含在时间维度上相邻的第一唇形和第二唇形；

插值单元507，配置为，根据所述第一唇形和第二唇形，对所述第一唇形与所述第二唇形之间的空白时间段进行插值操作，得到第一唇形序列。

30 在一些可能的实施方式中，装置500还包括：

平滑单元 508，配置为，对所述第一唇形、第二唇形和第一唇形序列所组成的目标唇形序列在时间维度上进行平滑操作，得到目标平滑唇形序列。

根据另一方面的实施例，还提供一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行上述任一实施例所描述的方法。

5 根据再一方面的实施例，还提供一种计算设备，包括存储器和处理器，其中，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现上述任一实施例所描述的方法。

本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于装置实施例而言，由于其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。

需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者装置不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者装置所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括要素的过程、方法、物品或者装置中还存在另外的相同要素。

本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也可以通过程序来指令相关的硬件完成，程序可以存储于一种计算机可读存储介质中，上述提到的存储介质可以是只读存储器，磁盘或光盘等。

以上所述的具体实施方式，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施方式而已，并不用于限定本发明的保护范围，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

权 利 要 求 书

1. 一种数字人唇形生成方法，包括：

获取文本数据与对应的音频数据；

将所述文本数据与音频数据在时间维度上对齐，得到所述文本数据对应的文本单位集合，以及所述文本单位集合中任一文本单位所对应的音频片段和时间段；

对于在目标时间段上的目标文本单位，根据第一字典，确定所述目标文本单位所对应的目标口型信息，其中，所述第一字典中包含任一文本单位所对应的口型信息；

根据所述目标文本单位所对应的目标音频片段，确定所述目标口型信息对应的目标幅度信息；

根据所述目标口型信息和目标幅度信息，确定在所述目标时间段上的目标唇形。

2. 根据权利要求1所述的方法，其中，所述音频数据为预先录制音频，所述文本单位为字；将所述文本数据与音频数据在时间维度上对齐，得到所述文本数据对应的文本单位集合，以及所述文本单位集合中任一文本单位所对应的音频片段和时间段，包括：

根据所述文本数据与音频数据，使用强制对齐模型生成字集合，以及所述字集合中任一字所对应的音频片段和时间段。

3. 根据权利要求1所述的方法，其中，所述音频数据为预先录制音频，所述文本单位为音素；将所述文本数据与音频数据在时间维度上对齐，得到所述文本数据对应的文本单位集合，以及所述文本单位集合中任一文本单位所对应的音频片段和时间段，包括：

根据所述文本数据中的任一字，确定所述字对应的音素，得到音素集合；

根据所述音素集合与音频数据，使用强制对齐模型生成所述音素集合中任一音素所对应的音频片段和时间段。

4. 根据权利要求1所述的方法，其中，所述音频数据为使用文本转语音模型根据所述文本数据生成的，所述文本转语音模型在生成所述音频数据的同时，还生成了所述文本数据对应的文本单位集合以及所述文本单位集合中任一文本单位所对应的音频片段和时间段；

将所述文本数据与音频数据在时间维度上对齐，得到所述文本数据对应的文本单位集合，以及所述文本单位集合中任一文本单位所对应的音频片段和时间段，包括：

从所述文本转语音模型的生成结果中获取所述文本单位集合以及所述文本单位集合中任一文本单位所对应的音频片段和时间段。

5. 根据权利要求1所述的方法，其中，根据所述目标文本单位所对应的目标音频片段，

确定所述目标口型信息对应的目标幅度信息，包括：

根据所述目标音频片段的音频特征，确定所述目标幅度信息，所述音频特征包括音量特征、基频特征和共振峰特征中的至少一种。

6. 根据权利要求5所述的方法，其中，所述音频特征为音量特征；根据所述目标音频片段的音频特征，确定所述目标幅度信息，包括：

根据所述目标文本单位所在的特定文本范围中的最大音量值和最小音量值，确定所述目标幅度信息。

7. 根据权利要求6所述的方法，其中，所述目标文本单位所在的特定文本范围为所述目标文本单位所在句子或段落。

8. 根据权利要求1所述的方法，还包括：

根据多个所述目标唇形组成唇形集合，所述唇形集合包含在时间维度上相邻的第一唇形和第二唇形；

根据所述第一唇形和第二唇形，对所述第一唇形与所述第二唇形之间的空白时间段进行插值操作，得到第一唇形序列。

9. 根据权利要求8所述的方法，其中，所述插值操作为线性插值。

10. 根据权利要求8所述的方法，其中，根据所述第一唇形和第二唇形，对所述第一唇形与所述第二唇形之间的空白时间段进行插值操作，得到第一唇形序列，包括：

根据所述第一唇形和第二唇形，确定所述空白时间段中任一最小时间单位所对应的唇形，得到第一唇形序列。

11. 根据权利要求8所述的方法，还包括：

对所述第一唇形、第二唇形和第一唇形序列所组成的目标唇形序列在时间维度上进行平滑操作，得到目标平滑唇形序列。

12. 根据权利要求11所述的方法，其中，所述平滑操作为指数平滑或一维高斯滤波平滑。

13. 一种数字人唇形生成装置，包括：

获取单元，配置为，获取文本数据与对应的音频数据；

对齐单元，配置为，将所述文本数据与音频数据在时间维度上对齐，得到所述文本数据对应的文本单位集合，以及所述文本单位集合中任一文本单位所对应的音频片段和时间段；

口型确定单元，配置为，对于在目标时间段上的目标文本单位，根据第一字典，确定

所述目标文本单位所对应的目标口型信息，其中，所述第一字典中包含任一文本单位所对应的口型信息；

幅度确定单元，配置为，根据所述目标文本单位所对应的目标音频片段，确定所述目标口型信息对应的目标幅度信息；

唇形确定单元，配置为，根据所述目标口型信息和目标幅度信息，确定在所述目标时间段上的目标唇形。

14. 根据权利要求 13 所述的装置，还包括：

集合确定单元，配置为，根据多个所述目标唇形组成唇形集合，所述唇形集合包含在时间维度上相邻的第一唇形和第二唇形；

插值单元，配置为，根据所述第一唇形和第二唇形，对所述第一唇形与所述第二唇形之间的空白时间段进行插值操作，得到第一唇形序列。

15. 根据权利要求 14 所述的装置，还包括：

平滑单元，配置为，对所述第一唇形、第二唇形和第一唇形序列所组成的目标唇形序列在时间维度上进行平滑操作，得到目标平滑唇形序列。

16. 一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行权利要求 1-12 中任一项所述的方法。

17. 一种计算设备，包括存储器和处理器，其中，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现权利要求 1-12 中任一项所述的方法。