

语音强制对齐模型评价方法、装置、电子设备及存储介质

本公开要求于 2020 年 9 月 7 日提交中国专利局、申请号为 202010925650.2、发明名称为“语音强制对齐模型评价方法、装置、电子设备及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本公
5 开中。

技术领域

本公开实施例涉及计算机领域，尤其涉及一种语音强制对齐模型评价方法、装置、电子设备及存储介质。

10 背景技术

随着计算机技术和深度学习技术的发展，语音合成技术得到广泛应用，例如：语音播报、语音导航以及智能音箱等。

在语音合成中，需要对语音合成模型进行训练，以提高语音合成的性能。为了实现对于语音合成模型的训练，需要得到训练语音的音素时间点。

15 为了得到音素时间点，通常利用语音强制对齐技术（即机器标注）获取，语音强制对齐技术是通过强制对齐模型确定音素时间点的技术。然而，在相关技术中，通过强制对齐模型得到的音素时间点的准确性不高。

发明内容

本公开实施例提供一种语音强制对齐模型评价方法、装置、电子设备及存储介质，以在较低成本的基础上，实现对语音强制对齐模型的准确性
20 评价。

为解决上述问题，本公开实施例提供一种语音强制对齐模型评价方法，包括：

25 利用待评价语音强制对齐模型，根据测试集中各段音频和与所述各段音频对应的文本，获取每段音频对应的音素序列以及该音素序列中各个音素的预测起止时间；

针对各个音素，根据该音素的预测起止时间和预先确定的该音素的基础起止时间，获取该音素的时间准确性得分，其中，所述时间准确性得分

用于表征该音素的所述预测起止时间与所述基准起止时间的接近程度；以及

根据各个音素的时间准确性得分，确定所述待评价语音强制对齐模型的时间准确性得分。

- 5 为解决上述问题，本公开实施例提供一种语音强制对齐模型评价装置，包括：

第一获取单元，配置为利用待评价语音强制对齐模型，根据测试集中各段音频和与所述各段音频对应的文本，获取每段音频对应的音素序列以及该音素序列中各个音素的预测起止时间；

- 10 第二获取单元，配置为针对各个音素，根据该音素的预测起止时间和预先确定的该音素的基准起止时间，获取该音素的时间准确性得分，其中，所述时间准确性得分用于表征该音素的所述预测起止时间与所述基准起止时间的接近程度；以及

- 15 第三获取单元，配置为根据各个音素的时间准确性得分，确定所述待评价语音强制对齐模型的时间准确性得分。

为解决上述问题，本公开实施例提供一种存储介质，所述存储介质存储有适于语音强制对齐模型评价的程序，以实现如前述任一项所述的语音强制对齐模型评价方法。

- 20 为解决上述问题，本公开实施例提供一种电子设备，包括至少一个存储器和至少一个处理器，其中，所述存储器存储有程序，所述处理器调用所述程序，以执行如前述任一项所述的语音强制对齐模型评价方法。

与现有技术相比，本公开的技术方案具有以下优点：

- 25 本公开实施例所提供的语音强制对齐模型评价方法、装置、电子设备及存储介质，其中，语音强制对齐模型评价方法，包括首先将测试集的各段音频和与音频对应的文本输入待评价语音强制对齐模型，利用待评价语音强制对齐模型获取每段音频对应的音素序列，以及各个音素序列的各个音素的预测起止时间，然后根据预测起止时间和预先已知的对应音素的基准起止时间，获取各个所述音素的时间准确性得分，基于各个音素的时间准确性得分，获取所述待评价语音强制对齐模型的时间准确性得分，实现
30 对待评价语音强制对齐模型的评价。可以看出，本公开实施例所提供的语

音强制对齐模型评价方法，在对待评价语音强制对齐模型的评价时，基于各个音素的预测起止时间和基准起止时间的接近程度，就可以得到各个音素的时间准确性得分，进而得到待评价语音强制对齐模型的时间准确性得分，无需在每次通过语音强制对齐模型获取到预测起止时间时，再利用人工方式进行复验，或者通过后续的语音合成得到的语音进行验证，可以简化对强制对齐模型的准确性评价的难度，同时还可以降低对强制对齐模型的准确性评价所需的人工成本和时间成本，提高效率。

可选方案中，本公开实施例所提供的语音强制对齐模型评价方法，还对各个音素执行首先确定当前音素，并构建当前音素的音素组合，得到各个音素的音素组合，各个音素的音素组合的组合方式相同，然后，在获取待评价语音强制对齐模型的时间准确性得分时，根据当前音素中音素组合的各个音素的时间准确性得分，获取当前音素的时间准确性修正得分，得到所述音素序列的各个音素的时间准确性修正得分，根据音素序列的各个音素的时间准确性修正得分获取待评价语音强制对齐模型的时间准确性得分。这样，本公开实施例所提供的语音强制对齐模型评价方法，利用与当前音素相邻的至少一个音素的时间准确性得分对当前音素的时间准确性得分进行修正，利用了当前音素的上下文信息，将当前音素受与其相邻的音素的影响考虑在内，使得所得到的当前音素的时间准确性得分得以修正，从而具有更高的准确性。

可选方案中，本公开实施例所提供的语音强制对齐模型评价方法，为了获取各个所述音素的时间准确性得分，首先获取同一音素的预测起止时间和基准起止时间的起止时间交集和起止时间并集，然后通过起止时间交集和起止时间并集的比值，获取对应音素的时间准确性得分。这样，起止时间交集可以表示预测起止时间和基准起止时间的重合量，起止时间并集可以表示预测起止时间和基准起止时间的最大整体量，利用起止时间交集和起止时间并集的比值可以将预测起止时间的重和程度准确地表示，从而实现对音素时间准确性得分的获取，并且音素时间准确性得分能够准确地表示预测起止时间和基准起止时间的接近程度。

30 **附图说明**

图 1 是本公开实施例所提供的语音强制对齐模型评价方法的一流程示意图；

图 2 为本公开实施例所提供的语音强制对齐模型评价方法的获取各个音素的时间准确性得分步骤的一流程示意图；

5 图 3 为本公开实施例所提供的语音强制对齐模型评价方法的另一流程示意图；

图 4 为本公开实施例所提供的语音强制对齐模型评价方法的又一流程示意图；

10 图 5 为本公开实施例所提供的待评价语音强制对齐模型的时间准确性得分的获取步骤的一流程示意图；

图 6 是本公开实施例所提供的语音强制对齐模型评价装置的一框图；

图 7 是本公开实施例提供的电子设备一种可选硬件设备架构。

具体实施方式

15 在相关技术中，通过人工方式对语音强制对齐模型进行评价的方式较费时费力，且评价结果也易受主观影响。

对此，本公开提供一种语音强制对齐模型评价方法，能够自动实现对语音强制对齐模型的准确性评价。本公开实施例提供了一种语音强制对齐模型评价方法，包括：

20 利用待评价语音强制对齐模型，根据测试集中各段音频和与所述各段音频对应的文本，获取每段音频对应的音素序列以及该音素序列中各个音素的预测起止时间；

针对各个音素，根据该音素的预测起止时间和预先确定的该音素的基准起止时间，获取该音素的时间准确性得分，其中，所述时间准确性得分
25 用于表征该音素的所述预测起止时间与所述基准起止时间的接近程度；以及

根据各个音素的时间准确性得分，确定所述待评价语音强制对齐模型的时间准确性得分。

从而，本公开实施例所提供的语音强制对齐模型评价方法，包括首先

将测试集的各段音频和与音频对应的文本输入待评价语音强制对齐模型，利用待评价语音强制对齐模型获取每段音频对应的音素序列，以及各个音素序列的各个音素的预测起止时间，然后根据预测起止时间和已知的对应音素的基准起止时间，获取各个所述音素的时间准确性得分，基于各个音素的时间准确性得分，获取所述待评价语音强制对齐模型的时间准确性得分，实现对待评价语音强制对齐模型的评价。

可以看出，本公开实施例所提供的语音强制对齐模型评价方法，在对待评价语音强制对齐模型的评价时，基于各个音素的预测起止时间和基准起止时间的接近程度，就可以得到各个音素的时间准确性得分，进而得到待评价语音强制对齐模型的时间准确性得分，无需在每次通过语音强制对齐模型获取到预测起止时间时，再利用人工方式进行复验，或者通过后续的语音合成得到的语音进行验证，可以简化对强制对齐模型的准确性评价的难度，同时还可以降低对强制对齐模型的准确性评价所需的人工成本和时间成本，提高效率。

下面将结合本公开实施例中的附图，对本公开实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本公开一部分实施例，而不是全部的实施例。基于本公开中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本公开保护的范围。

请参考图 1，图 1 是本公开实施例所提供的语音强制对齐模型评价方法的一流程示意图。

如图中所示，本公开实施例所提供的语音强制对齐模型评价方法，包括以下步骤：

步骤 S10：利用待评价语音强制对齐模型，根据测试集中各段音频和与所述各段音频对应的文本，获取每段音频对应的音素序列以及该音素序列中各个音素的预测起止时间。

容易理解的是，本公开实施例所提供的语音强制对齐模型评价方法，用于对待评价语音强制对齐模型的语音强制对齐效果进行评价，因此需要首先建立需要评价的语音强制对齐模型或者获取已经建立的语音强制对齐模型，即待评价语音强制对齐模型。

将测试集的各段音频和与各段音频对应的文本输入待评价语音强制对齐模型，从而得到分别与每段音频对应的音素序列，以及每段音素序列的各个音素的预测起止时间。

当然，预测起止时间可以包括从预测起始时刻到预测结束时刻的时间
5 跨度。

具体地，待评价语音强制对齐模型可以包括 GMM 模型（高斯混合模型）和维特比（viterbi）解码模型，将测试集的各段音频和与各段音频对应的文本输入 GMM 模型，得到未解码的音素序列和预测起止时间，然后经过维特比解码模型进行解码，得到解码后的音素序列和预测起止时间。

10 步骤 S11：针对各个音素，根据该音素的预测起止时间和预先确定的该音素的基准起止时间，获取该音素的时间准确性得分。

可以理解的是，所述时间准确性得分用于表征该音素的所述预测起止时间与所述基准起止时间的接近程度。

其中，基准起止时间是指用于作为评价基准的音素起止时间，可以通
15 过人工标注的方式获取。

通过比较同一音素的预测起止时间和基准起止时间接近程度，得到该音素的时间准确性得分，直至得到各个音素的时间准确性得分。

在一种具体实施方式中，为了方便各个音素的时间准确性得分的获取，请参考图 2，图 2 为本公开实施例所提供的语音强制对齐模型评价方法的
20 获取各个音素的时间准确性得分步骤的一流程示意图。

如图中所示，本公开实施例所提供的语音强制对齐模型评价方法，可以通过以下步骤获取各个音素的时间准确性得分。

步骤 S110：针对各个音素，根据该音素的预测起止时间和预先确定的该音素的基准起止时间，获取该音素的所述预测起止时间和所述基准起止
25 时间的起止时间交集和起止时间并集。

容易理解的是，所述音素的所述预测起止时间和所述基准起止时间的起止时间交集是指同一音素的预测起止时间和基准起止时间的重叠时间，所述音素的所述预测起止时间和所述基准起止时间的起止时间并集是指同一音素的预测起止时间和基准起止时间的整体时间。

30 示例性地，对于音素“b”，假设预测起止时间是从第 3ms 到第 5ms，基

准起止时间是从第 4ms 到第 6ms,那么起止时间交集为从第 4ms 到第 5ms,起止时间并集为从第 3ms 到第 6ms。

步骤 S111: 根据各个音素的起止时间交集与起止时间并集的比值,得到各个音素的时间准确性得分。

- 5 得到各个音素的起止时间交集和起止时间并集后,进一步获取二者的比值,得到各个音素的时间准确性得分。

如前述例子,音素“b”的时间准确性得分即为:第 4ms 到第 5ms/第 3ms 到第 6ms,为 1/3。

- 可以理解的是,某个音素的起止时间交集与起止时间并集的比值分数
10 越大,那么待评价语音强制对齐模型在该音素的准确性越高。

- 这样,起止时间交集可以表示预测起止时间和基准起止时间的重合量,起止时间并集可以表示预测起止时间和基准起止时间的最大整体量,利用起止时间交集和起止时间并集的比值可以将预测起止时间的重和程度准确地表示,从而实现对音素时间准确性得分的获取,并且音素时间准确性得分能够准确地表示预测起止时间和基准起止时间的接近程度。
15

步骤 S12: 根据各个音素的时间准确性得分,确定所述待评价语音强制对齐模型的时间准确性得分。

得到测试集的各个音素的时间准确性得分后,进一步通过各个音素的时间准确性得分就可以获取待评价语音强制对齐模型的时间准确性得分。

- 20 在一种具体实施方式中,可以直接将测试集的各个音素的时间准确性得分进行相加,获取待评价语音强制对齐模型的时间准确性得分。

- 容易理解的是,各个音素的时间准确性得分越高,待评价语音强制对齐模型的时间准确性得分也就越高,待评价语音强制对齐模型的强制对齐效果也就越好,从而实现对于不同的语音强制对齐模型的对齐效果的评价,或者对于参数调整前后的语音强制对齐模型的对齐效果的评价。
25

- 可以看出,本公开实施例所提供的语音强制对齐模型评价方法,在对待评价语音强制对齐模型的评价时,基于各个音素的预测起止时间和基准起止时间的接近程度,就可以得到各个音素的时间准确性得分,进而得到待评价语音强制对齐模型的时间准确性得分,无需在每次通过语音强制对齐模型获取到预测起止时间时,再利用人工方式进行复验,或者通过后续
30

的语音合成得到的语音进行验证，可以简化对强制对齐模型的准确性评价的难度，同时还可以降低对强制对齐模型的准确性评价所需的人工成本和时间成本，提高效率。

5 为了提高对语音强制对齐模型评价的准确性，本公开实施例还提供另一种语音强制对齐模型评价方法，请参考图 3，图 3 为本公开实施例所提供的语音强制对齐模型评价方法的另一流程示意图。

本公开实施例所提供的语音强制对齐模型评价方法包括：

10 步骤 S20：利用待评价语音强制对齐模型，根据测试集中各段音频与所述各段音频对应的文本，获取每段音频对应的音素序列以及该音素序列中各个音素的预测起止时间。

步骤 S20 的具体内容请参考图 1 关于步骤 S10 的描述，在此不再赘述。

步骤 S21：针对各个音素，根据该音素的预测起止时间和预先确定的该音素的基准起止时间，获取该音素的时间准确性得分。

步骤 S21 的具体内容请参考图 1 关于步骤 S11 的描述，在此不再赘述。

15 步骤 S22：确定当前音素，构建所述当前音素的音素组合，以获取各个音素的音素组合。

当然，音素组合包括所述当前音素和与当前音素临近的至少一个音素，并且各个音素的音素组合的组合方式相同。

20 得到测试集的各段音频的音素序列后，确定所述音素序列中一音素为当前音素，然后确定当前音素临近的至少一个音素，与当前音素共同组成音素组合，从而得到所述音素序列中所述当前音素对应的音素组合，逐一确定所述音素序列中各个音素为当前音素，从而得到所述音素序列中各个音素对应的音素组合。

25 可以理解的是，如果音素组合由 2 个音素构建组成，那么音素序列的每个音素都会构建由 2 个音素组成的音素组合，并且组合方式也相同，可以确定位于当前音素前的相邻音素和当前音素组成音素组合，当然也可以确定位于当前音素后的相邻音素和当前音素组成音素组合，如果音素组合由 3 个音素构建组成，那么音素序列的每个音素都会构建由 3 个音素组成的音素组合，并且组合方式也相同，可以则确定当前音素前后相邻的音素
30 和当前音素组成音素组合，如果音素组合由 4 个音素构建组成，那么音素

序列的每个音素都会构建由 4 个音素组成的音素组合，并且组合方式也相同，可以则确定当前音素前的 2 个音素和当前音素后的一个音素和当前音素组成音素组合，当然也可以选择当前音素前的 1 个音素和当前音素后的 2 个音素。

- 5 例如，对于“jintian”这样一组音素序列，确定“t”为当前音素时，如果音素组合由 2 个音素构建组成，则当前音素“t”的音素组合可以为“int”或“tian”，可以任选其一为当前音素“t”的一个音素组合，也可以都作为当前音素“t”的音素组合；如果音素组合由 3 个音素构建组成，则当前音素“t”的音素组合可以为“intian”；如果音素组合由 4 个音素构建组成，则当前音素“t”的音素组合可以为“jintian”或“intian+silence”，任选其一为当前音素“t”的一个音素组合，也可以都作为当前音素“t”的音素组合。

当然，还可以将由 2 个音素构建组成音素组合、由 3 个音素构建组成的音素组合、由 4 个音素构建组成的音素组合都作为同一个音素的音素组合。

- 15 由于每个音素的起止时间会受与其相邻的音素的影响，将当前音素，与临近的音素考虑在内，形成音素组合，可以为后续对当前音素的时间准确性得分提供修正。

- 步骤 S23：根据各个音素组合中各个音素的时间准确性得分，获取各个音素组合中当前音素的时间准确性修正得分，以得到所述音素序列中各个音素的时间准确性修正得分。

得到各个音素的音素组合后，利用当前音素对应的音素组合中的各个音素的时间准确性得分，获取当前音素的时间准确性修正得分。

如前所示事例，音素组合由 3 个音素构建组成，则当前音素“t”的音素组合为“intian”为例，当前音素 t 的时间准确性修正得分可以为：

- 25 $Score(t) = (Score(in) + Score(t) + Score(ian)) / 3$ 。

步骤 S24：根据所述音素序列中各个音素的时间准确性修正得分，获取所述待评价语音强制对齐模型的时间准确性得分。

- 步骤 S24 的具体内容可以参考图 1 所示的步骤 S12 的内容，只是利用各个音素的时间准确性修正得分替换各个音素的时间准确性得分，其他内容不再赘述。

本公开实施例所提供的语音强制对齐模型评价方法，利用与当前音素相邻的至少一个音素的时间准确性得分对当前音素的时间准确性得分进行修正，利用了当前音素的上下文信息，将当前音素受与其相邻的音素的影响考虑在内，使得所得到的当前音素的时间准确性得分得以修正，从而具有更高的准确性。

为了进一步提高评价的准确性，本公开实施例还提供另一种语音强制对齐模型评价方法，请参考图 4，图 4 为本公开实施例所提供的语音强制对齐模型评价方法的又一流程示意图。

如图中所示，本公开实施例所提供的语音强制对齐模型评价方法包括：

10 步骤 S30：利用待评价语音强制对齐模型，根据测试集中各段音频和与所述各段音频对应的文本，获取每段音频对应的音素序列以及该音素序列中各个音素的预测起止时间。

步骤 S30 的具体内容请参考图 1 关于步骤 S10 的描述，在此不再赘述。

15 步骤 S31：针对各个音素，根据该音素的预测起止时间和预先确定的该音素的基准起止时间，获取该音素的时间准确性得分。

步骤 S31 的具体内容请参考图 1 关于步骤 S11 的描述，在此不再赘述。

步骤 S32：确定当前音素，构建所述当前音素的音素组合，以获取各个音素的音素组合。

步骤 S32 的具体内容请参考图 3 关于步骤 S22 的描述，在此不再赘述。

20 步骤 S33：根据所述音素组合中各音素的发音方式对所述音素组合进行分类，得到所述音素组合的组合类别；根据各个音素组合的组合类别，确定同一组合类别的音素组合的数量以及对应的组合权重。

在得到各个当前音素的音素组合后，根据音素组合的各个音素的发音方式进行分类。相邻音素的不同发音方式对于当前音素的参数会有一定的影响，因此可以根据音素组合的各个音素的发音方式进行分类，确定各个音素组合的组合类别，然后根据音素组合的组合类别，可以确定同类音素组合的数量，进而获取某类别音素组合的组合权重，进一步根据组合权重可以获取每个音素的权重得分，降低由于基于测试集所得到的音素数量的不同而导致的待评价语音强制对齐模型的时间准确性得分的差别，提高本公开实施例所提供的语音强制对齐模型评价方法的评价准确性。

具体地，所述发音方式可以根据声母和韵母分别进行划分，包括声母发音方式和韵母发音方式，其中，所述声母发音方式包括根据发音部位分类的部位发音方式和根据发音方法分类的方法发音方式，所述韵母发音方式包括根据结构分类的结构发音方式和根据口型分类的口型发音方式。

5 其中声母发音方式的具体分类可以参考表 1：

表 1 声母发音方式

发音方法 发音部位		塞音 (清音)		塞擦音 (清音)		擦音		鼻音	边音
		不送气	送气	不送气	送气	清音	浊音	浊音	浊音
音唇	双唇音	b	p					m	
	唇齿音					f			
舌尖中音		d	t					n	l
舌根音			k			h			
舌面音				j	q	x			
舌尖后音				zh	ch	sh	r		
舌尖前音				z	c	s			

韵母发音方式的具体分类可以参考表 2：

表 2 韵母发音方式

韵母 按结构分	按口型分	开口呼	齐齿呼	合口呼	撮口呼

单韵母	-i (前)、-i (后)	i	u	ü
	a	ia	ua	
	o		uo	
	e			üe
	ê	ie		
	er			
复韵母	ai		uai	
	ei		uei	
	ao	iao		
	ou	iou		
鼻韵母	an	ian	uan	üan
	en	in	uen	ün
	ang	iang	uang	

当然，所述发音方式可以根据其他语言的发音进行划分，例如：英语。

当基于拼音发音方式进行划分，具体进行分组时，可以将声母和韵母的发音方式进行组合，得到具体分类类别，例如：二音素组合：双唇音+鼻韵母、鼻韵母+唇齿音；三音素组合：双唇音+鼻韵母+唇齿音、单韵母+双唇音+单韵母、或者单韵母开口呼+双唇音塞音+单韵母齐齿呼；四音素组合：单韵母+双唇音+单韵母+双唇音。

这样，将发音方式分类与声母和韵母的发音方式相结合，可以更为方便地实现发音方式分类，降低发音方式分类的难度。得到各个组合类别后，进一步获取各个音素组合的组合权重，具体地，组合权重为同一组合类别的音素组合的数量与所述音素序列中音素总量的比值。

为方便理解，现进行举例说明，当某个音素序列包括 100 个音素时，

如果每个音素形成一个音素组合，那么就会形成 100 个音素组合，可以根据每个音素组合的各个音素的发音方式确定组合类别，然后对每个音素组合进行分类，假设可以共形成 3 个组合类别。

5 然后可以统计各个组合类别中的音素组合的数量，假设其中第一个组合类别有 20 个，第二个组合类别有 45 个，第三个音素组合有 35 个，进而可以根据各个组合类别中各个音素组合的数量确定组合权重，例如：第一类组合类别的组合权重可以为 $20/100=0.2$ ，第二类组合类别的组合权重可以为 $45/100=0.45$ ，第三类组合类别的组合权重可以为 $35/100=0.35$ 。

10 步骤 S34：根据所述当前音素的音素组合中的各个音素的时间准确性得分，获取所述当前音素的时间准确性修正得分。

步骤 S34 的具体内容，请参考图 3 所示的步骤 S23 的描述，在此不再赘述。

并且步骤 S33 和步骤 S34 的执行顺序不做限制，也可以先获取时间准确性修正得分再获取组合权重。

15 步骤 S35：针对各个音素，根据该音素的时间准确性修正得分和与该音素对应的音素组合的组合权重，获取该音素的权重得分。

基于步骤 S33 得到的组合权重和步骤 S34 得到的时间准确性修正得分，获取音素的权重得分。

20 当然组合权重和时间准确性修正得分为基于相同的音素的相同的音素组合获取，二者之间具有相互对应关系。

具体地，通过将组合权重与时间准确性修正得分相乘，获取各个所述音素的权重得分。

步骤 S36：根据所述音素序列的各个音素的权重得分，获取所述待评价语音强制对齐模型的时间准确性得分。

25 得到各个音素的权重得分后，进而可以通过各个音素的权重得分获取待评价语音强制对齐模型的时间准确性得分。

具体地，待评价语音强制对齐模型的时间准确性得分通过以下公式获取：

$$\text{Score 模型} = W1 * \text{Score1} + W2 * \text{Score2} + \dots + Wn * \text{Scoren}$$

30 其中：Score 模型为待评价语音强制对齐模型的时间准确性得分，

W_n 为第 n 个音素的组合权重，
Score_n 为第 n 个音素的时间准确性修正得分。

权重得分的获取，可以降低由于通过不同的待评价语音强制对齐模型所预测的音素序列的音素数量的不同而造成的待评价语音强制对齐模型的时间准确性得分的影响，进一步提高评价的准确性。

在另一种实施例中，为进一步提高对当前音素的时间准确性得分的修正，可以为同一个音素构建多个音素组合，具体地，各个音素的音素组合可以包括 2 个音素组成的二音素组合和 3 个音素组成的三音素组合，当然二音素组合包括当前音素和与当前音素直接相邻的一个音素，三音素组合包括当前音素和与当前音素直接相邻的两个音素，则分别计算各个音素组合的当前音素的时间准确性修正得分，从而得到同一个音素的多个时间准确性修正得分，包括二音素时间准确性修正得分和三音素时间准确性修正得分，并分别获取该音素的二音素组合类别和三音素组合类别，以及二音素组合权重和三音素组合权重，并获取二音素权重得分和三音素权重得分。

图 5 为本公开实施例所提供的待评价语音强制对齐模型的时间准确性得分的获取步骤的一流程示意图，获取所述待评价语音强制对齐模型的时间准确性得分的步骤可以包括：

步骤 S361：根据当前音素的二音素权重得分和三音素权重得分获取所述当前音素的融合权重得分。

在一种具体实施方式中，所述融合权重得分可以通过以下公式获取：

$$\text{score} = v_2 * \text{score}'' + v_3 * \text{score}''' ;$$

其中：v₂+v₃=1，且 v₃>v₂，

score 为融合权重得分，

score'' 为二音素权重得分，

v₂ 为二音素融合因子，

score''' 为三音素权重得分，

v₃ 为三音素融合因子。

这样可以简单地实现同一个音素不同权重得分的融合，并且三音素融合因子大于二音素融合因子可以凸出三音素组合的影响，进一步提高准确性。

步骤 S362: 根据所述音素序列中各个音素的融合权重得分, 获取所述待评价语音强制对齐模型的时间准确性得分。

得到融合权重得分, 就可以获取待评价语音强制对齐模型的时间准确性得分, 具体内容请参考图 1 步骤 S12 的描述, 在此不再赘述。

5 当然, 在另一种具体实施方式中, 每个音素也可以有 3 个音素组合, 除了 2 个音素组成的二音素组合和 3 个音素组成的三音素组合外, 还包括 4 个音素组成的四音素组合, 那么还获取该音素的四音素组合类别和四音素组合权重, 以及四音素权重得分, 所述根据所述音素序列的各个音素的权重得分, 获取所述待评价语音强制对齐模型的时间准确性得分的步骤可以包括:

10

根据所述当前音素的所述二音素权重得分、所述三音素权重得分和所述四音素权重得分获取所述当前音素的融合权重得分; 以及

根据所述音素序列的各个音素的融合权重得分, 获取所述待评价语音强制对齐模型的时间准确性得分。

15 在一种具体实施方式中, 所述融合权重得分可以通过以下公式获取:
$$\text{score} = v2 * \text{score}'' + v3 * \text{score}''' + v4 * \text{score}'''';$$

其中: $v2 + v3 + v4 = 1$, 且 $v3 > v2$, $v3 > v4$,

score 为融合权重得分,

score'' 为二音素权重得分,

20 v2 为二音素融合因子,

score''' 为三音素权重得分,

v3 为三音素融合因子,

score'''' 为四音素权重得分,

v4 为四音素融合因子。

25 这样可以简单地实现同一个音素不同权重得分的融合, 并且三音素融合因子大于二音素融合因子, 且三音素融合因子大于四音素融合因子可以凸出三音素组合的影响, 进一步提高准确性。

下面对本公开实施例提供的语音强制对齐模型评价装置进行介绍, 下文描述的语音强制对齐模型评价装置可以认为是, 电子设备(如: PC)为
30 分别实现本公开实施例提供的语音强制对齐模型评价方法所需设置的功能

模块架构。下文描述的语音强制对齐模型评价装置的内容，可分别与上文描述的语音强制对齐模型评价方法的内容相互对应参照。

图 6 是本公开实施例所提供的语音强制对齐模型评价装置的一框图，该语音强制对齐模型评价装置即可应用于客户端，也可应用于服务器端，

5 参考图 6，该语音强制对齐模型评价装置可以包括：

第一获取单元 100，配置为利用待评价语音强制对齐模型，根据测试集中各段音频和与所述各段音频对应的文本，获取每段音频对应的音素序列以及该音素序列中各个音素的预测起止时间；

10 第二获取单元 110，配置为针对各个音素，根据该音素的预测起止时间和预先确定的该音素的基准起止时间，获取该音素的时间准确性得分，其中，所述时间准确性得分用于表征该音素的所述预测起止时间与所述基准起止时间的接近程度；以及

第三获取单元 120，配置为根据各个音素的时间准确性得分，确定所述待评价语音强制对齐模型的时间准确性得分。

15 容易理解的是，本公开实施例所提供的语音强制对齐模型评价装置，将测试集中的各段音频和与各段音频对应的文本输入待评价语音强制对齐模型，从而得到分别与每段音频对应的音素序列，以及每段音素序列的各个音素的预测起止时间。

20 当然，预测起止时间可以包括从预测起始时刻到预测结束时刻的时间跨度。

具体地，待评价语音强制对齐模型可以包括 GMM 模型（高斯混合模型）和维特比（viterbi）解码模型，将测试集的各段音频和与各段音频对应的文本输入 GMM 模型，得到未解码的音素序列和预测起止时间，然后经过维特比解码模型进行解码，得到解码后的音素序列和预测起止时间。

25 可以理解的是，时间准确性得分即为各个所述音素对应的所述预测起止时间与对应的所述基准起止时间的接近程度。

其中，基准起止时间是指用于作为评价基准的音素起止时间，可以通过人工标注的方式获取。

30 通过比较同一音素的预测起止时间和基准起止时间接近程度，得到该音素的时间准确性得分，直至得到各个音素的时间准确性得分。

所述第二获取单元 110 包括：

第三获取子单元，配置为根据各个音素的预测起止时间和基准起止时间，获取各个音素的预测起止时间和基准起止时间的起止时间交集和起止时间并集；以及

- 5 第四获取子单元，配置为获取各个音素的起止时间交集与起止时间并集的比值，得到各个音素的时间准确性得分。

容易理解的是，所述音素的所述预测起止时间和所述基准起止时间的起止时间交集是指同一音素的预测起止时间和基准起止时间的重叠时间，所述音素的所述预测起止时间和所述基准起止时间的起止时间并集是指同
10 一音素的预测起止时间和基准起止时间的整体时间。

得到各个音素的起止时间交集和起止时间并集后，进一步获取二者的比值，得到各个音素的时间准确性得分。

可以理解的是，某个音素的起止时间交集与起止时间并集的比值分数越大，那么待评价语音强制对齐模型在该音素的准确性越高。

- 15 这样，起止时间交集可以表示预测起止时间和基准起止时间的重合量，起止时间并集可以表示预测起止时间和基准起止时间的最大整体量，利用起止时间交集和起止时间并集的比值可以将预测起止时间的重合程度准确地表示，从而实现对音素时间准确性得分的获取，并且音素时间准确性得分能够准确地表示预测起止时间和基准起止时间的接近程度。

- 20 得到测试集的各个音素的时间准确性得分后，第三获取单元 120，通过各个音素的时间准确性得分就可以获取待评价语音强制对齐模型的时间准确性得分。

在一种具体实施方式中，可以直接将测试集的各个音素的时间准确性得分进行相加，获取待评价语音强制对齐模型的时间准确性得分。

- 25 容易理解的是，各个音素的时间准确性得分越高，待评价语音强制对齐模型的时间准确性得分也就越高，待评价语音强制对齐模型的强制对齐效果也就越好，从而实现对于不同的语音强制对齐模型的对齐效果的评价，或者对于参数调整前后的语音强制对齐模型的对齐效果的评价。

- 30 可以看出，本公开实施例所提供的语音强制对齐模型评价装置，在对待评价语音强制对齐模型的评价时，基于各个音素的预测起止时间和基准

起止时间的接近程度，就可以得到各个音素的时间准确性得分，进而得到待评价语音强制对齐模型的时间准确性得分，无需在每次通过语音强制对齐模型获取到预测起止时间时，再利用人工方式进行复验，或者通过后续的语音合成得到的语音进行验证，可以简化对强制对齐模型的准确性评价的难度，同时还可以降低对强制对齐模型的准确性评价所需的人工成本和时间成本，提高效率。

为了进一步提高对语音强制对齐模型评价的准确性，本公开实施例还提供一种语音强制对齐模型评价装置。

如图 6 所示，本公开实施例所提供的语音强制对齐模型评价装置还包括：

第四获取单元 130，配置为确定当前音素，构建所述当前音素的音素组合，以获取各个音素的音素组合。

音素组合包括所述当前音素和与所述当前音素临近的至少一个音素，各个音素的音素组合的组合方式相同。

得到测试集的各段音频的音素序列后，确定所述音素序列中一音素为当前音素，然后确定当前音素临近的至少一个音素，与当前音素共同组成音素组合，从而得到所述音素序列中所述当前音素对应的音素组合，逐一确定所述音素序列中各个音素为当前音素，从而得到所述音素序列中各个音素对应的音素组合。

可以理解的是，如果音素组合由 2 个音素组成，可以则确定位于当前音素前的相邻音素和当前音素组成音素组合，当然也可以确定位于当前音素后的相邻音素和当前音素组成音素组合，如果音素组合由 3 个音素组成，可以则确定当前音素前后相邻的音素和当前音素组成音素组合，如果音素组合由 4 个音素组成，可以则确定当前音素前的 2 个音素和当前音素后的一个音素和当前音素组成音素组合，当然也可以选择当前音素前的 1 个音素和当前音素后的 2 个音素。

由于每个音素的起止时间会受与其相邻的音素的影响，将当前音素，与临近的音素考虑在内，形成音素组合，可以为后续对当前音素的时间准确性得分提供修正。

所述第三获取单元 120 包括：

第一获取子单元，配置为根据各个音素组合中各个音素的时间准确性得分，获取各个音素组合中当前音素的时间准确性修正得分，以得到所述音素序列中各个音素的时间准确性修正得分；以及

第二获取子单元，配置为根据所述音素序列中各个音素的时间准确性修正得分，获取所述待评价语音强制对齐模型的时间准确性得分。

得到各个音素的音素组合后，利用每个音素构建有 1 个音素组合时，利用当前音素对应的音素组合中的各个音素的时间准确性得分，获取当前音素的时间准确性修正得分。

如音素组合包括 3 个音素，当前音素“t”的音素组合为“intian”，当前音素 t 的时间准确性修正得分可以为：

$$\text{Score}(t)' = (\text{Score}(\text{in}) + \text{Score}(t) + \text{Score}(\text{ian})) / 3。$$

然后，利用各个音素的时间准确性修正得分获取待评价语音强制对齐模型的时间准确性得分。

这样，本公开实施例所提供的语音强制对齐模型评价装置，利用与当前音素相邻的至少一个音素的时间准确性得分对当前音素的时间准确性得分进行修正，利用了当前音素的上下文信息，将当前音素受与其相邻的音素的影响考虑在内，使得所得到的当前音素的时间准确性得分得以修正，从而具有更高的准确性。

为了进一步提高评价的准确性，本公开实施例所提供的语音强制对齐模型评价装置，还包括：

第五获取单元 140，配置为根据所述音素组合中各音素的发音方式对所述音素组合进行分类，得到所述音素组合的组合类别，并根据各个音素组合的组合类别，确定同一组合类别的音素组合的数量以及对应的组合权重。

所述第三获取单元 120 所包括的所述第二获取子单元包括：

第一获取模块，配置为针对各个音素，根据该音素的时间准确性修正得分和该音素对应的音素组合的组合权重，获取该音素的权重得分；以及

第二获取模块，配置为根据所述音素序列的各个音素的权重得分，获取所述待评价语音强制对齐模型的时间准确性得分。

在得到各个当前音素的音素组合后，根据音素组合的各个音素的发音

方式进行分类。相邻音素的不同发音方式对于当前音素的参数会有一定的影响，因此可以根据音素组合的各个音素的发音方式进行分类，确定各个音素组合的组合类别，然后根据音素组合的组合类别，可以确定同类音素组合的数量，进而获取某类别音素组合的组合权重，进一步根据组合权重

5 可以获取每个音素的权重得分，降低由于基于测试集所得到的音素数量的不同而导致的待评价语音强制对齐模型的时间准确性得分的差别，提高本公开实施例所提供的语音强制对齐模型评价方法的评价准确性。具体地，所述发音方式可以根据声母和韵母分别进行划分，包括声母发音方式和韵母发音方式，其中，所述声母发音方式包括根据发音部位分类的部位发音方式和根据发音方法分类的方法发音方式，所述韵母发音方式包括根据结构分类的结构发音方式和根据口型分类的口型发音方式。

10

这样，将发音方式分类与声母和韵母的发音方式相结合，可以更为方便地实现发音方式分类，降低发音方式分类的难度。

得到各个组合类别后，进一步获取各个音素组合的组合权重，具体地，

15 所述组合权重为同一组合类别的音素组合的数量与所述音素序列中音素总量的比值。

为方便理解，现进行举例说明，当某个音素序列包括 100 个音素时，如果每个音素形成一个音素组合，那么就会形成 100 个音素组合，可以根据每个音素组合的各个音素的发音方式确定组合类别，然后对每个音素组合进行分类，假设可以共形成 3 个组合类别。

20

然后可以统计各个组合类别中的音素组合的数量，假设其中第一个组合类别有 20 个，第二个组合类别有 45 个，第三个音素组合有 35 个，进而可以根据各个组合类别中各个音素组合的数量确定组合权重，例如：第一类组合类别的组合权重可以为 $20/100=0.2$ ，第二类组合类别的组合权重可以

25 以为 $45/100=0.45$ ，第二类组合类别的组合权重可以为 $35/100=0.35$ 。

然后，基于组合权重和时间准确性修正得分，获取音素的权重得分。

当然组合权重和时间准确性修正得分为基于相同的音素的相同的音素组合获取，二者之间具有相互对应关系。

具体地，通过将组合权重与时间准确性修正得分相乘，获取各个所述

30 音素的权重得分。

得到各个音素的权重得分后，进而可以通过各个音素的权重得分获取待评价语音强制对齐模型的时间准确性得分。

具体地，待评价语音强制对齐模型的时间准确性得分通过以下公式获取：

5
$$\text{Score 模型} = W1 * \text{Score1} + W2 * \text{Score2} + \dots + Wn * \text{Scoren}$$

其中：Score 模型为待评价语音强制对齐模型的时间准确性得分，

Wn 为第 n 个音素的组合权重，

Scoren 为第 n 个音素的时间准确性修正得分。

权重得分的获取，可以降低由于通过不同的待评价语音强制对齐模型
10 所预测的音素序列的音素数量的不同而造成的待评价语音强制对齐模型的时间准确性得分的影响，进一步提高评价的准确性。

在另一种实施例中，为了提高评价的准确性，还可以构建同一个音素的多个音素组合，各个音素的音素组合可以包括 2 个音素组成的二音素组合和 3 个音素组成的三音素组合，当然二音素组合包括当前音素和与当前
15 音素直接相邻的一个音素，三音素组合包括当前音素和与当前音素直接相邻的两个音素。

为同一个音素构建多个音素组合，可以利用多个音素组合进一步提高对当前音素的时间准确性得分的修正。

当同一个音素包括多个音素组合时，则需分别计算各个音素组合的当前音素的时间准确性修正得分，从而得到同一个音素的多个时间准确性修正得分。
20

当同一个音素同时具有至少两个音素组合时，例如：二音素组合和三音素组合时，则分别获取该音素的二音素组合类别和三音素组合类别，以及二音素组合权重和三音素组合权重。

25 当同一个音素同时构建有二音素组合、三音素组合时，组合权重包括二音素组合权重和三音素组合权重，时间准确性修正得分包括二音素时间准确性修正得分和三音素时间准确性修正得分，所得到的权重得分则包括二音素权重得分和三音素权重得分。

容易理解的是，当同一个音素的权重得分包括二音素权重得分和三音素权重时，为了保证待评价语音强制对齐模型的时间准确性得分的获取，
30

本公开实施例所提供的语音强制对齐模型的评价装置的第三获取单元 120 所包括的所述第二获取子单元中的所述第二获取模块, 包括:

第一获取子模块, 配置为根据当前音素的二音素权重得分和三音素权重得分获取所述当前音素的融合权重得分; 以及

- 5 第二获取子模块, 配置为根据音素序列的各个音素的融合权重得分, 获取所述待评价语音强制对齐模型的时间准确性得分。

在一种具体实施方式中, 所述融合权重得分可以通过以下公式获取:

$$\text{score} = v2 * \text{score}'' + v3 * \text{score}''';$$

其中: $v2 + v3 = 1$, 且 $v3 > v2$;

- 10 score 为融合权重得分;
 score'' 为二音素权重得分;
 v2 为二音素融合因子;
 score''' 为三音素权重得分;
 v3 为三音素融合因子。

- 15 这样可以简单地实现同一个音素不同权重得分的融合, 并且三音素融合因子大于二音素融合因子可以凸出三音素组合的影响, 进一步提高准确性。

得到融合权重得分, 再进一步取得待评价语音强制对齐模型的时间准确性得分。

- 20 当然, 在另一种具体实施方式中, 为提高准确性, 第四获取单元 130 还可以对每个音素构建有 3 个音素组合, 除了 2 个音素组成的二音素组合和 3 个音素组成的三音素组合外, 还构建包括 4 个音素组成的四音素组合。第五获取单元 140, 还配置为获取该音素的四音素组合类别和四音素组合权重。第三获取单元 120 所包括的所述第二获取子单元中的所述第一获取
25 模块, 获取四音素权重得分。所述第二获取子单元中的所述第二获取模块包括:

第三获取子模块, 配置为根据当前音素的二音素权重得分、三音素权重得分和四音素权重得分获取所述当前音素的融合权重得分; 以及

- 30 第四获取子模块, 配置为根据所述音素序列中各个音素的融合权重得分, 获取所述待评价语音强制对齐模型的时间准确性得分。

在一种具体实施方式中，所述融合权重得分可以通过以下公式获取：

$$\text{score} = v2 * \text{score}'' + v3 * \text{score}''' + v4 * \text{score}'''';$$

其中： $v2 + v3 + v4 = 1$ ，且 $v3 > v2$ ， $v3 > v4$ ；

score 为融合权重得分；

5 score'' 为二音素权重得分；

v2 为二音素融合因子；

score''' 为三音素权重得分；

v3 为三音素融合因子；

score'''' 为四音素权重得分；

10 v4 为四音素融合因子。

这样可以简单地实现同一个音素不同权重得分的融合，并且三音素融合因子大于二音素融合因子，且三音素融合因子大于四音素融合因子可以凸出三音素组合的影响，进一步提高准确性。

当然，本公开实施例还提供一种电子设备，本公开实施例提供的电子设备可以通过程序形式装载上述所述的程序模块架构，以实现本公开实施例提供的语音强制对齐模型评价方法；该硬件电子设备可以应用于具体数据处理能力的电子设备，该电子设备可以为：例如终端设备或者服务器设备。

20 可选的，图 7 示出了本公开实施例提供的一种可选硬件设备架构，可以包括：至少一个存储器 3 和至少一个处理器 1；所述存储器存储有程序，所述处理器调用所述程序，以执行前述的语音强制对齐模型评价方法，另外，至少一个通信接口 2 和至少一个通信总线 4；处理器 1 和存储器 3 可以位于同一电子设备，例如处理器 1 和存储器 3 可以位于服务器设备或者终端设备；处理器 1 和存储器 3 也可以位于不同的电子设备。

25 作为本公开实施例公开内容的一种可选实现，存储器 3 可以存储程序，处理器 1 可调用所述程序，以执行本公开上述实施例提供的语音强制对齐模型评价方法。

本公开实施例中，电子设备可以是能够进行语音强制对齐模型评价的平板电脑、笔记本电脑等设备。

30 在本公开实施例中，处理器 1、通信接口 2、存储器 3、通信总线 4 的

数量为至少一个，且处理器 1、通信接口 2、存储器 3 通过通信总线 4 完成相互间的通信；显然，图 7 所示的处理器 1、通信接口 2、存储器 3 和通信总线 4 的通信连接示意仅是可选的一种方式；

可选的，通信接口 2 可以为通信模块的接口，如 GSM 模块的接口；

- 5 处理器 1 可能是中央处理器 CPU，或者是特定集成电路 ASIC (Application Specific Integrated Circuit)，或者是被配置成实施本公开实施例的一个或多个集成电路。

存储器 3 可能包含高速 RAM 存储器，也可能还包括非易失性存储器 (non-volatile memory)，例如至少一个磁盘存储器。

- 10 需要说明的是，上述的设备还可以包括与本公开实施例公开内容可能并不是必需的其他器件 (未示出)；鉴于这些其他器件对于理解本公开实施例公开内容可能并不是必需，本公开实施例对此不进行逐一介绍。

本公开实施例还提供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可执行指令，当该指令被处理器执行时可以实现如上所述语音强制对齐模型评价方法。

- 15 本公开实施例所提供的存储介质所存储的计算机可执行指令，在对待评价语音强制对齐模型的评价时，基于各个音素的预测起止时间和基准起止时间的接近程度，就可以得到各个音素的时间准确性得分，进而得到待评价语音强制对齐模型的时间准确性得分，无需在每次通过语音强制对齐模型获取到预测起止时间时，再利用人工方式进行复验，或者通过后续的语音合成得到的语音进行验证，可以简化对强制对齐模型的准确性评价的难度，同时还可以降低对强制对齐模型的准确性评价所需的人工成本和时间成本，提高效率。

- 25 上述本公开的实施方式是本公开的元件和特征的组合。除非另外提及，否则所述元件或特征可被视为选择性的。各个元件或特征可在不与其它元件或特征组合的情况下实践。另外，本公开的实施方式可通过组合部分元件和/或特征来构造。本公开的实施方式中所描述的操作顺序可重新排列。任一实施方式的一些构造可被包括在另一实施方式中，并且可用另一实施方式的对应构造代替。对于本领域技术人员而言明显的是，所附权利要求中彼此没有明确引用关系的权利要求可组合成本公开的实施方式，或者可
- 30

在提交本公开之后的修改中作为新的权利要求包括。

本公开的实施方式可通过例如硬件、固件、软件或其组合的各种手段来实现。在硬件配置方式中，根据本公开示例性实施方式的方法可通过一个或更多个专用集成电路(ASIC)、数字信号处理器(DSP)、数字信号处理器
5 件(DSPD)、可编程逻辑器件(PLD)、现场可编程门阵列(FPGA)、处理器、
控制器、微控制器、微处理器等来实现。

在固件或软件配置方式中，本公开的实施方式可以模块、过程、功能等形式实现。软件代码可存储在存储器单元中并由处理器执行。存储器单元位于处理器的内部或外部，并可经由各种已知手段向处理器发送数据以及
10 从处理器接收数据。

对所公开的实施例的上述说明，使本领域技术人员能够实现或使用本公开。对这些实施例的多种修改对本领域技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本公开的精神或范围的情况下，在其他实施例中实现。因此，本公开将不会被限制于本文所示的这些实施例，
15 而是符合与本文所公开的原理和新颖特点相一致的最宽的范围。

虽然本公开实施例披露如上，但本公开并非限定于此。任何本领域技术人员，在不脱离本公开的精神和范围内，均可作各种变动与修改，因此本公开的保护范围应当以权利要求所限定的范围为准。

权 利 要 求

1. 一种语音强制对齐模型评价方法，包括：

5 利用待评价语音强制对齐模型，根据测试集中各段音频和与所述各段音频对应的文本，获取每段音频对应的音素序列以及该音素序列中各个音素的预测起止时间；

针对各个音素，根据该音素的预测起止时间和预先确定的该音素的基准起止时间，获取该音素的时间准确性得分，其中，所述时间准确性得分用于表征该音素的所述预测起止时间与所述基准起止时间的接近程度；以及
10 根据各个音素的时间准确性得分，确定所述待评价语音强制对齐模型的时间准确性得分。

2. 如权利要求 1 所述的方法，其中，所述根据各个音素的时间准确性得分，确定所述待评价语音强制对齐模型的时间准确性得分之前，还包括：

确定当前音素，构建所述当前音素的音素组合，以获取各个音素的音素组合，其中，所述当前音素的音素组合包括所述当前音素和与所述当前音素临近的至少一个音素，并且其中，各个音素的音素组合的组合方式相同；
15

并且其中，所述根据各个音素的时间准确性得分，确定所述待评价语音强制对齐模型的时间准确性得分包括：

根据各个音素组合中各个音素的时间准确性得分，获取各个音素组合中当前音素的时间准确性修正得分，以得到所述音素序列中各个音素的时间准确性修正得分；以及
20

根据所述音素序列中各个音素的时间准确性修正得分，获取所述待评价语音强制对齐模型的时间准确性得分。

3. 如权利要求 2 所述的方法，还包括：

25 根据所述音素组合中各音素的发音方式对所述音素组合进行分类，得到所述音素组合的组合类别；以及

根据各个音素组合的组合类别，确定同一组合类别的音素组合的数量以及对应的组合权重，其中，所述组合权重为同一组合类别的音素组合的数量与所述音素序列中音素的数量比值；

30 并且其中，所述根据所述音素序列中各个音素的时间准确性修正得分，

获取所述待评价语音强制对齐模型的时间准确性得分包括：

针对各个音素，根据该音素的时间准确性修正得分和该音素对应的音素组合的组合权重，获取该音素的权重得分；以及

5 根据所述音素序列中各个音素的权重得分，获取所述待评价语音强制对齐模型的时间准确性得分。

4. 如权利要求 3 所述的方法，其中，所述当前音素的音素组合包括二音素组合和三音素组合，所述二音素组合包括所述当前音素和与所述当前音素直接相邻的一个音素，所述三音素组合包括所述当前音素和与所述当前音素直接相邻的两个音素；

10 所述组合类别包括各个二音素组合类别和各个三音素组合类别，所述组合权重包括与各个所述二音素组合类别对应的二音素组合权重和与各个所述三音素组合类别对应的三音素组合权重，所述时间准确性修正得分包括所述当前音素的二音素时间准确性修正得分和三音素时间准确性修正得分，所述权重得分包括所述当前音素的二音素权重得分和三音素权重得分；

15 并且其中，所述根据所述音素序列中各个音素的权重得分，获取所述待评价语音强制对齐模型的时间准确性得分包括：

根据所述当前音素的二音素权重得分和三音素权重得分获取所述当前音素的融合权重得分；以及

20 根据所述音素序列中各个音素的融合权重得分，获取所述待评价语音强制对齐模型的时间准确性得分。

5. 如权利要求 4 所述的方法，其中，所述二音素组合包括所述当前音素和所述当前音素前的音素。

6. 如权利要求 4 所述的方法，其中，所述融合权重得分通过以下公式获取：

25 $score = v_2 * score'' + v_3 * score'''$ ；

其中： $v_2 + v_3 = 1$ ，且 $v_3 > v_2$ ， $score$ 为融合权重得分， $score''$ 为二音素权重得分， v_2 为二音素融合因子， $score'''$ 为三音素权重得分， v_3 为三音素融合因子。

7. 如权利要求 4 所述的方法，其中，所述当前音素的音素组合还包括
30 四音素组合，所述四音素组合包括所述当前音素和与所述当前音素临近的

三个音素；

所述组合类别还包括各个四音素组合类别，所述组合权重还包括与各个四音素组合类别对应的四音素组合权重，所述时间准确性修正得分还包括所述当前音素的四音素时间准确性修正得分，所述权重得分还包括所述当前音素的四音素权重得分；

并且其中，所述根据所述音素序列中各个音素的权重得分，获取所述待评价语音强制对齐模型的时间准确性得分包括：

根据所述当前音素的所述二音素权重得分、所述三音素权重得分和所述四音素权重得分获取所述当前音素的融合权重得分；以及

10 根据所述音素序列中各个音素的融合权重得分，获取所述待评价语音强制对齐模型的时间准确性得分。

8. 如权利要求 7 所述的方法，其中，所述融合权重得分通过以下公式获取：

$$\text{score} = v2 * \text{score}'' + v3 * \text{score}''' + v4 * \text{score}'''';$$

15 其中： $v2 + v3 + v4 = 1$ ，且 $v3 > v2$ ， $v3 > v4$ ，score 为融合权重得分， score'' 为二音素权重得分， $v2$ 为二音素融合因子， score''' 为三音素权重得分， $v3$ 为三音素融合因子， score'''' 为四音素权重得分， $v4$ 为四音素融合因子。

9. 如权利要求 3-8 中任一项所述的方法，其中，所述待评价语音强制对齐模型的时间准确性得分通过以下公式获取：

20 $\text{Score 模型} = W1 * \text{Score}1 + W2 * \text{Score}2 + \dots + Wn * \text{Score}n$,

其中，Score 模型为所述待评价语音强制对齐模型的时间准确性得分， Wn 为第 n 个音素的组合权重， $\text{Score}n$ 为第 n 个音素的时间准确性修正得分。

25 10. 如权利要求 3-8 中任一项所述的方法，其中，所述发音方式包括声母发音方式和韵母发音方式，所述声母发音方式包括根据发音部位分类的部位发音方式和根据发音方法分类的方法发音方式，所述韵母发音方式包括根据结构分类的结构发音方式和根据口型分类的口型发音方式。

11. 如权利要求 1-8 中任一项所述的方法，其中，针对各个音素，所述根据该音素的预测起止时间和预先确定的该音素的基准起止时间，获取该音素的时间准确性得分包括：

获取各个音素的预测起止时间和基准起止时间的起止时间交集和起止时间并集；以及

根据各个音素的起止时间交集与起止时间并集的比值，得到各个音素的时间准确性得分。

5 12. 一种语音强制对齐模型评价装置，包括：

第一获取单元，配置为利用待评价语音强制对齐模型，根据测试集中各段音频和与所述各段音频对应的文本，获取每段音频对应的音素序列以及该音素序列中各个音素的预测起止时间；

10 第二获取单元，配置为针对各个音素，根据该音素的预测起止时间和预先确定的该音素的基准起止时间，获取该音素的时间准确性得分，其中，所述时间准确性得分用于表征该音素的所述预测起止时间与所述基准起止时间的接近程度；以及

第三获取单元，配置为根据各个音素的时间准确性得分，确定所述待评价语音强制对齐模型的时间准确性得分。

15 13. 根据权利要求 12 所述的装置，还包括第四获取单元，配置为：

确定当前音素，构建所述当前音素的音素组合，以获取各个音素的音素组合，其中，所述当前音素的音素组合包括所述当前音素和与所述当前音素临近的至少一个音素，并且其中，各个音素的音素组合的组合方式相同；

并且其中，所述第三获取单元包括：

20 第一获取子单元，配置为根据各个音素组合中各个音素的时间准确性得分，获取各个音素组合中当前音素的时间准确性修正得分，以得到所述音素序列中各个音素的时间准确性修正得分；以及

第二获取子单元，配置为根据所述音素序列中各个音素的时间准确性修正得分，获取所述待评价语音强制对齐模型的时间准确性得分。

25 14. 如权利要求 13 所述的装置，还包括：

第五获取单元，配置为根据所述音素组合中各音素的发音方式对所述音素组合进行分类，得到所述音素组合的组合类别；以及

30 根据各个音素组合的组合类别，确定同一组合类别的音素组合的数量以及对应的组合权重，其中，所述组合权重为同一组合类别的音素组合的数量与所述音素序列中音素的数量比值；

并且其中，所述第二获取子单元包括：

第一获取模块，配置为针对各个音素，根据该音素的时间准确性修正得分和该音素对应的音素组合的组合权重，获取该音素的权重得分；以及

第二获取模块，配置为根据所述音素序列中各个音素的权重得分，获取
5 所述待评价语音强制对齐模型的时间准确性得分。

15. 如权利要求 14 所述的装置，其中，所述当前音素的音素组合包括二音素组合和三音素组合，所述二音素组合包括所述当前音素和与所述当前音素直接相邻的一个音素，所述三音素组合包括所述当前音素和与所述当前音素直接相邻的两个音素；

10 所述组合类别包括各个二音素组合类别和各个三音素组合类别，所述组合权重包括与各个所述二音素组合类别对应的二音素组合权重和与各个所述三音素组合类别对应的三音素组合权重，所述时间准确性修正得分包括所述当前音素的二音素时间准确性修正得分和三音素时间准确性修正得分，所述权重得分包括所述当前音素的二音素权重得分和三音素权重得分；

15 并且其中，所述第二获取模块包括：

第一获取子模块，配置为根据所述当前音素的二音素权重得分和三音素权重得分获取所述当前音素的融合权重得分；以及

第二获取子模块，配置为根据所述音素序列中各个音素的融合权重得分，获取所述待评价语音强制对齐模型的时间准确性得分。

20 16. 如权利要求 15 所述的装置，其中，所述二音素组合包括所述当前音素和所述当前音素前的音素。

17. 如权利要求 15 所述的装置，其中，所述融合权重得分通过以下公式获取：

$$\text{score} = v_2 * \text{score}'' + v_3 * \text{score}''' ;$$

25 其中： $v_2 + v_3 = 1$ ，且 $v_3 > v_2$ ，score 为融合权重得分，score'' 为二音素权重得分， v_2 为二音素融合因子，score''' 为三音素权重得分， v_3 为三音素融合因子。

18. 如权利要求 15 所述的装置，其中，所述当前音素的音素组合还包括四音素组合，所述四音素组合包括所述当前音素和与所述当前音素临近
30 的三个音素；

所述组合类别还包括各个四音素组合类别,所述组合权重还包括与各个四音素组合类别对应的四音素组合权重,所述时间准确性修正得分还包括所述当前音素的四音素时间准确性修正得分,所述权重得分还包括所述当前音素的四音素权重得分;

5 并且其中,所述第二获取模块包括:

第三获取子模块,配置为根据所述当前音素的所述二音素权重得分、所述三音素权重得分和所述四音素权重得分获取所述当前音素的融合权重得分;以及

10 第四获取子模块,配置为根据所述音素序列中各个音素的融合权重得分,获取所述待评价语音强制对齐模型的时间准确性得分。

19. 如权利要求 18 所述的装置,其中,所述融合权重得分通过以下公式获取:

$$\text{score} = v2 * \text{score}'' + v3 * \text{score}''' + v4 * \text{score}'''';$$

15 其中: $v2 + v3 + v4 = 1$, 且 $v3 > v2$, $v3 > v4$, score 为融合权重得分, score'' 为二音素权重得分, $v2$ 为二音素融合因子, score''' 为三音素权重得分, $v3$ 为三音素融合因子, score'''' 为四音素权重得分, $v4$ 为四音素融合因子。

20. 如权利要求 14-19 中任一项所述的装置,其中,所述待评价语音强制对齐模型的时间准确性得分通过以下公式获取:

$$\text{Score 模型} = W1 * \text{Score}1 + W2 * \text{Score}2 + \dots + Wn * \text{Score}n,$$

20 其中, Score 模型 为所述待评价语音强制对齐模型的时间准确性得分, Wn 为第 n 个音素的组合权重, $\text{Score}n$ 为第 n 个音素的时间准确性修正得分。

21. 如权利要求 14-19 中任一项所述的装置,其中,所述发音方式包括声母发音方式和韵母发音方式,所述声母发音方式包括根据发音部位分类的部位发音方式和根据发音方法分类的方法发音方式,所述韵母发音方式
25 包括根据结构分类的结构发音方式和根据口型分类的口型发音方式。

22. 如权利要求 12-19 中任一项所述的装置,其中,所述第二获取单元包括:

30 第三获取子单元,配置为获取各个音素的预测起止时间和基准起止时间的起止时间交集和起止时间并集;以及

第四获取子单元,配置为根据各个音素的起止时间交集与起止时间并集的比值,得到各个音素的时间准确性得分。

23. 一种存储介质,其中,所述存储介质存储有适于语音强制对齐模型评价的程序,以实现如权利要求 1-11 中任一项所述的语音强制对齐模型评

5 价方法。

24. 一种电子设备,包括:

至少一个存储器; 以及

至少一个处理器,

其中,所述存储器存储有程序,所述处理器调用所述程序,以执行如权

10 利要求 1-11 中任一项所述的语音强制对齐模型评价方法。