



(12) 发明专利申请

(10) 申请公布号 CN 112802444 A

(43) 申请公布日 2021.05.14

(21) 申请号 202011607966.3

(22) 申请日 2020.12.30

(71) 申请人 科大讯飞股份有限公司

地址 230088 安徽省合肥市高新区望江西
路666号

(72) 发明人 周良 孟廷 侯秋侠 刘丹 江源
胡亚军

(74) 专利代理机构 北京集佳知识产权代理有限
公司 11227

代理人 付丽

(51) Int.Cl.

G10L 13/02 (2013.01)

G10L 13/08 (2013.01)

G10L 15/16 (2006.01)

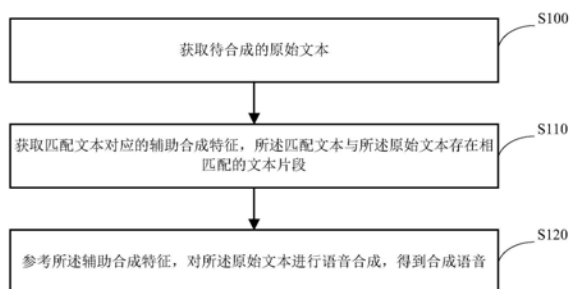
权利要求书4页 说明书19页 附图4页

(54) 发明名称

语音合成方法、装置、设备及存储介质

(57) 摘要

本申请公开了一种语音合成方法、装置、设备及存储介质,本申请在对待合成的原始文本进行语音合成的过程中,参考了与原始文本存在相匹配的文本片段的匹配文本对应的辅助合成特征,该辅助合成特征为基于匹配文本对应的发音音频所确定的用于辅助语音合成的特征,本申请通过参考匹配文本对应的辅助合成特征,可以利用匹配文本对应的发音音频中发音信息辅助对原始文本进行语音合成,丰富了原始文本语音合成时所参考的信息,进而提高了原始文本的语音合成质量。本申请方案可以适用于带前端预处理和不带前端预处理两种类型的语音合成系统,辅助合成特征可以作为前端文本分析结果也可以直接辅助语音合成系统进行语音合成,均能够提高合成语音的质量。



1. 一种语音合成方法,其特征在于,包括:

获取待合成的原始文本;

获取匹配文本对应的辅助合成特征,所述匹配文本与所述原始文本存在相匹配的文本片段,所述辅助合成特征为基于所述匹配文本对应的发音音频所确定的用于辅助语音合成的特征;

参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音。

2. 根据权利要求1所述的方法,其特征在于,所述获取匹配文本对应的辅助合成特征,包括:

获取与所述原始文本存在相匹配的文本片段的匹配文本;

获取基于所述匹配文本对应的发音音频所确定的辅助合成特征。

3. 根据权利要求1或2所述的方法,其特征在于,所述辅助合成特征包括:

基于所述匹配文本对应的发音音频所确定的,所述匹配文本对应的音素序列;

和/或,

基于所述匹配文本对应的发音音频所确定的,所述匹配文本对应的韵律信息;

和/或,

基于所述匹配文本对应的发音音频所确定的,所述匹配文本对应的音素级韵律编码;

和/或,

所述匹配文本对应的发音音频的声学特征。

4. 根据权利要求2所述的方法,其特征在于,所述获取与所述原始文本存在相匹配的文本片段的匹配文本,包括:

在预配置的模板文本中,确定与所述原始文本内的文本片段相匹配的匹配文本。

5. 根据权利要求2所述的方法,其特征在于,所述获取与所述原始文本存在相匹配的文本片段的匹配文本,包括:

获取上传数据中的上传文本,作为所述匹配文本,所述上传数据还包括所述上传文本对应的发音音频,所述上传文本与所述原始文本存在相匹配的文本片段。

6. 根据权利要求4所述的方法,其特征在于,所述预配置的模板文本包括:

各个预配置的资源包中的模板文本,其中每一资源包包含一模板文本,及基于所述模板文本对应的发音音频所确定的,与所述模板文本对应的辅助合成特征。

7. 根据权利要求6所述的方法,其特征在于,所述在预配置的模板文本中,确定与所述原始文本内的文本片段相匹配的匹配文本,包括:

分别将所述原始文本与预配置的每个资源包中的模板文本进行匹配计算;

在匹配度最高的资源包所包含的模板文本中,确定与所述原始文本内的文本片段相匹配的匹配文本。

8. 根据权利要求7所述的方法,其特征在于,所述获取基于所述匹配文本对应的发音音频所确定的辅助合成特征,包括:

获取所述匹配度最高的资源包中包含的,与所述匹配文本对应的辅助合成特征。

9. 根据权利要求6-8任一项所述的方法,其特征在于,预配置的资源包的确定过程,包括:

获取预配置的模板文本及对应的发音音频;

基于所述发音音频,确定所述模板文本对应的音素序列及韵律信息;

将所述音素序列及韵律信息作为所述模板文本对应的辅助合成特征,并将所述辅助合成特征与所述模板文本整理成一个资源包。

10. 根据权利要求9所述的方法,其特征在于,预配置的资源包的确定过程,还包括:

基于所述模板文本及对应的发音音频,确定所述模板文本对应的音素级韵律编码;

将所述音素级韵律编码合并入所述资源包中。

11. 根据权利要求10所述的方法,其特征在于,所述基于所述模板文本及对应的发音音频,确定所述模板文本对应的音素级韵律编码,包括:

基于所述模板文本及对应的发音音频,提取出音素级的韵律信息;

将所述模板文本及所述音素级的韵律信息输入编码预测网络,得到预测的音素级的韵律编码;

将所述预测的音素级的韵律编码及所述模板文本输入生成网络,得到生成的音素级的韵律信息;

以生成的音素级的韵律信息趋近于提取出的所述音素级的韵律信息为目标训练所述编码预测网络及所述生成网络,直至训练结束时,得到训练后的编码预测网络所预测的音素级的韵律编码。

12. 根据权利要求5所述的方法,其特征在于,在所述获取上传数据中的上传文本之前,该方法还包括:

获取并输出所述原始文本的初始合成语音;

则所述上传文本为,所述初始合成语音中合成错误的文本片段,所述上传文本对应的发音音频为,所述合成错误的文本片段对应的正确发音;

或,所述上传文本为,包含所述初始合成语音中合成错误的文本片段的扩展文本,所述上传文本对应的发音音频为,所述扩展文本对应的正确发音。

13. 根据权利要求5或12所述的方法,其特征在于,所述获取基于所述匹配文本对应的发音音频所确定的辅助合成特征,包括:

基于所述上传数据中所述匹配文本对应的发音音频,确定匹配文本对应的辅助合成特征。

14. 根据权利要求3所述的方法,其特征在于,所述参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音,包括:

基于所述匹配文本对应的音素序列,确定所述原始文本的音素序列;

和/或,

基于所述匹配文本对应的韵律信息,确定所述原始文本的韵律信息;

基于所述原始文本的音素序列和/或韵律信息,对所述原始文本进行语音合成,得到合成语音。

15. 根据权利要求14所述的方法,其特征在于,所述参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音,还包括:

基于所述匹配文本对应的音素级韵律编码,获取所述匹配文本与所述原始文本中相同文本片段对应的音素级韵律编码;

在对所述原始文本进行语音合成过程中,将所述相同文本片段对应的音素级韵律编码

作为语音合成模型的补充输入,得到合成语音。

16. 根据权利要求14所述的方法,其特征在于,所述基于所述匹配文本对应的音素序列,确定所述原始文本的音素序列,包括:

基于所述匹配文本对应的音素序列,获取所述匹配文本与所述原始文本中相同文本片段对应的音素序列;

查询发音词典,确定所述原始文本中除所述相同文本片段外的其余文本片段的音素序列,并与所述相同文本片段对应的音素序列合成,得到原始文本的音素序列。

17. 根据权利要求3所述的方法,其特征在于,所述参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音,包括:

基于语音合成模型处理所述原始文本,得到预测当前语音帧的上下文信息;

基于所述上下文信息、所述匹配文本及所述发音音频的声学特征,确定预测当前语音帧所需的目标声学特征;

基于所述上下文信息,及确定的所述目标声学特征,预测当前语音帧,在预测得到所有语音帧后,由预测的各语音帧组成合成语音。

18. 根据权利要求17所述的方法,其特征在于,所述基于所述上下文信息、所述匹配文本及所述发音音频的声学特征,确定预测当前语音帧所需的目标声学特征,包括:

基于所述上下文信息、所述匹配文本及所述发音音频的声学特征,获取所述上下文信息与所述发音音频的声学特征中,每一帧声学特征的关联度;

基于所述关联度,确定预测当前语音帧所需的目标声学特征。

19. 根据权利要求18所述的方法,其特征在于,所述获取所述上下文信息与所述发音音频的声学特征中,每一帧声学特征的关联度,包括:

获取所述发音音频的声学特征对所述匹配文本的第一注意力权重矩阵,所述第一注意力权重矩阵包括每一帧声学特征对所述匹配文本中各文本单元的注意力权重;

获取所述上下文信息对所述匹配文本的第二注意力权重矩阵,所述第二注意力权重矩阵包括所述上下文信息对所述匹配文本中各文本单元的注意力权重;

基于所述第一注意力权重及所述第二注意力权重矩阵,得到所述上下文信息对所述声学特征的第三注意力权重矩阵,所述第三注意力权重矩阵包括所述上下文信息对每一帧声学特征的注意力权重,作为所述上下文信息与每一帧声学特征的关联度。

20. 根据权利要求18所述的方法,其特征在于,所述基于所述关联度,确定预测当前语音帧所需的目标声学特征,包括:

对各个所述关联度进行归一化,并以归一化后的各个关联度作为权重,对所述发音音频的各帧声学特征进行加权相加,得到目标声学特征。

21. 根据权利要求17-20中任一项所述的方法,其特征在于,所述基于所述上下文信息,及确定的所述目标声学特征,预测当前语音帧,包括:

基于语音合成模型的解码端当前的隐层矢量及所述目标声学特征,确定预测当前语音帧时所述目标声学特征的融合系数;

参考所述融合系数,对所述目标声学特征及所述上下文信息进行融合,并基于融合结果预测当前语音帧。

22. 一种语音合成装置,其特征在于,包括:

原始文本获取单元,用于获取待合成的原始文本;

辅助合成特征获取单元,用于获取匹配文本对应的辅助合成特征,所述匹配文本与所述原始文本存在相匹配的文本片段,所述辅助合成特征为基于所述匹配文本对应的发音音频所确定的用于辅助语音合成的特征;

辅助语音合成单元,用于参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音。

23. 一种语音合成设备,其特征在于,包括:存储器和处理器;

所述存储器,用于存储程序;

所述处理器,用于执行所述程序,实现如权利要求1~21中任一项所述的语音合成方法的各个步骤。

24. 一种存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时,实现如权利要求1~21中任一项所述的语音合成方法的各个步骤。

语音合成方法、装置、设备及存储介质

技术领域

[0001] 本申请涉及语音处理技术领域,更具体的说,是涉及一种语音合成方法、装置、设备及存储介质。

背景技术

[0002] 近些年,随着信息的发展及人工智能的兴起,人机交互变得越来越重要。其中语音合成是国内外人机交互研究的热点。语音合成即将输入的待合成的原始文本合成为语音输出的过程。

[0003] 传统的语音合成模型一般为基于端到端的语音合成方案,即直接使用训练文本及对应的语音数据或波形数据来训练语音合成模型,训练完毕的语音合成模型,基于输入的待合成的原始文本,即可输出合成的语音,或者输出波形数据,进而基于波形数据得到对应的合成语音。

[0004] 现有语音合成方案仅参考了原始文本进行语音合成,导致合成语音容易出错,合成效果不佳。

发明内容

[0005] 鉴于上述问题,提出了本申请以便提供一种语音合成方法、装置、设备及存储介质,以提高合成语音的质量。具体方案如下:

[0006] 在本申请的第一方面,提供了一种语音合成方法,包括:

[0007] 获取待合成的原始文本;

[0008] 获取匹配文本对应的辅助合成特征,所述匹配文本与所述原始文本存在相匹配的文本片段,所述辅助合成特征为基于所述匹配文本对应的发音音频所确定的用于辅助语音合成的特征;

[0009] 参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音。

[0010] 优选地,根据权利要求1所述的方法,其特征在于,所述获取匹配文本对应的辅助合成特征,包括:

[0011] 获取与所述原始文本存在相匹配的文本片段的匹配文本;

[0012] 获取基于所述匹配文本对应的发音音频所确定的辅助合成特征。

[0013] 优选地,所述辅助合成特征包括:

[0014] 基于所述匹配文本对应的发音音频所确定的,所述匹配文本对应的音素序列;

[0015] 和/或,

[0016] 基于所述匹配文本对应的发音音频所确定的,所述匹配文本对应的韵律信息;

[0017] 和/或,

[0018] 基于所述匹配文本对应的发音音频所确定的,所述匹配文本对应的音素级韵律编码;

[0019] 和/或,

- [0020] 所述匹配文本对应的发音音频的声学特征。
- [0021] 优选地,所述获取与所述原始文本存在相匹配的文本片段的匹配文本,包括:
- [0022] 在预配置的模板文本中,确定与所述原始文本内的文本片段相匹配的匹配文本。
- [0023] 优选地,所述获取与所述原始文本存在相匹配的文本片段的匹配文本,包括:
- [0024] 获取上传数据中的上传文本,作为所述匹配文本,所述上传数据还包括所述上传文本对应的发音音频,所述上传文本与所述原始文本存在相匹配的文本片段。
- [0025] 优选地,所述预配置的模板文本包括:
- [0026] 各个预配置的资源包中的模板文本,其中每一资源包包含一模板文本,及基于所述模板文本对应的发音音频所确定的,与所述模板文本对应的辅助合成特征。
- [0027] 优选地,所述在预配置的模板文本中,确定与所述原始文本内的文本片段相匹配的匹配文本,包括:
- [0028] 分别将所述原始文本与预配置的每个资源包中的模板文本进行匹配计算;
- [0029] 在匹配度最高的资源包所包含的模板文本中,确定与所述原始文本内的文本片段相匹配的匹配文本。
- [0030] 优选地,所述获取基于所述匹配文本对应的发音音频所确定的辅助合成特征,包括:
- [0031] 获取所述匹配度最高的资源包中包含的,与所述匹配文本对应的辅助合成特征。
- [0032] 优选地,预配置的资源包的确定过程,包括:
- [0033] 获取预配置的模板文本及对应的发音音频;
- [0034] 基于所述发音音频,确定所述模板文本对应的音素序列及韵律信息;
- [0035] 将所述音素序列及韵律信息作为所述模板文本对应的辅助合成特征,并将所述辅助合成特征与所述模板文本整理成一个资源包。
- [0036] 优选地,预配置的资源包的确定过程,还包括:
- [0037] 基于所述模板文本及对应的发音音频,确定所述模板文本对应的音素级韵律编码;
- [0038] 将所述音素级韵律编码合并入所述资源包中。
- [0039] 优选地,所述基于所述模板文本及对应的发音音频,确定所述模板文本对应的音素级韵律编码,包括:
- [0040] 基于所述模板文本及对应的发音音频,提取出音素级的韵律信息;
- [0041] 将所述模板文本及所述音素级的韵律信息输入编码预测网络,得到预测的音素级的韵律编码;
- [0042] 将所述预测的音素级的韵律编码及所述模板文本输入生成网络,得到生成的音素级的韵律信息;
- [0043] 以生成的音素级的韵律信息趋近于提取出的所述音素级的韵律信息为目标训练所述编码预测网络及所述生成网络,直至训练结束时,得到训练后的编码预测网络所预测的音素级的韵律编码。
- [0044] 优选地,在所述获取上传数据中的上传文本之前,该方法还包括:
- [0045] 获取并输出所述原始文本的初始合成语音;
- [0046] 则所述上传文本为,所述初始合成语音中合成错误的文本片段,所述上传文本对

应的发音音频为,所述合成错误的文本片段对应的正确发音;

[0047] 或,所述上传文本为,包含所述初始合成语音中合成错误的文本片段的扩展文本,所述上传文本对应的发音音频为,所述扩展文本对应的正确发音。

[0048] 优选地,所述获取基于所述匹配文本对应的发音音频所确定的辅助合成特征,包括:

[0049] 基于所述上传数据中所述匹配文本对应的发音音频,确定匹配文本对应的辅助合成特征。

[0050] 优选地,所述参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音,包括:

[0051] 基于所述匹配文本对应的音素序列,确定所述原始文本的音素序列;

[0052] 和/或,

[0053] 基于所述匹配文本对应的韵律信息,确定所述原始文本的韵律信息;

[0054] 基于所述原始文本的音素序列和/或韵律信息,对所述原始文本进行语音合成,得到合成语音。

[0055] 优选地,所述参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音,还包括:

[0056] 基于所述匹配文本对应的音素级韵律编码,获取所述匹配文本与所述原始文本中相同文本片段对应的音素级韵律编码;

[0057] 在对所述原始文本进行语音合成过程中,将所述相同文本片段对应的音素级韵律编码作为语音合成模型的补充输入,得到合成语音。

[0058] 优选地,所述基于所述匹配文本对应的音素序列,确定所述原始文本的音素序列,包括:

[0059] 基于所述匹配文本对应的音素序列,获取所述匹配文本与所述原始文本中相同文本片段对应的音素序列;

[0060] 查询发音词典,确定所述原始文本中除所述相同文本片段外的其余文本片段的音素序列,并与所述相同文本片段对应的音素序列合成,得到原始文本的音素序列。

[0061] 优选地,所述参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音,包括:

[0062] 基于语音合成模型处理所述原始文本,得到预测当前语音帧的上下文信息;

[0063] 基于所述上下文信息、所述匹配文本及所述发音音频的声学特征,确定预测当前语音帧所需的目标声学特征;

[0064] 基于所述上下文信息,及确定的所述目标声学特征,预测当前语音帧,在预测得到所有语音帧后,由预测的各语音帧组成合成语音。

[0065] 优选地,所述基于所述上下文信息、所述匹配文本及所述发音音频的声学特征,确定预测当前语音帧所需的目标声学特征,包括:

[0066] 基于所述上下文信息、所述匹配文本及所述发音音频的声学特征,获取所述上下文信息与所述发音音频的声学特征中,每一帧声学特征的关联度;

[0067] 基于所述关联度,确定预测当前语音帧所需的目标声学特征。

[0068] 优选地,所述获取所述上下文信息与所述发音音频的声学特征中,每一帧声学特

征的关联度,包括:

[0069] 获取所述发音音频的声学特征对所述匹配文本的第一注意力权重矩阵,所述第一注意力权重矩阵包括每一帧声学特征对所述匹配文本中各文本单元的注意力权重;

[0070] 获取所述上下文信息对所述匹配文本的第二注意力权重矩阵,所述第二注意力权重矩阵包括所述上下文信息对所述匹配文本中各文本单元的注意力权重;

[0071] 基于所述第一注意力权重及所述第二注意力权重矩阵,得到所述上下文信息对所述声学特征的第三注意力权重矩阵,所述第三注意力权重矩阵包括所述上下文信息对每一帧声学特征的注意力权重,作为所述上下文信息与每一帧声学特征的关联度。

[0072] 优选地,所述基于所述关联度,确定预测当前语音帧所需的目标声学特征,包括:

[0073] 对各个所述关联度进行归一化,并以归一化后的各个关联度作为权重,对所述发音音频的各帧声学特征进行加权相加,得到目标声学特征。

[0074] 优选地,所述基于所述上下文信息,及确定的所述目标声学特征,预测当前语音帧,包括:

[0075] 基于语音合成模型的解码端当前的隐层矢量及所述目标声学特征,确定预测当前语音帧时所述目标声学特征的融合系数;

[0076] 参考所述融合系数,对所述目标声学特征及所述上下文信息进行融合,并基于融合结果预测当前语音帧。

[0077] 在本申请的第二方面,提供了一种语音合成装置,包括:

[0078] 原始文本获取单元,用于获取待合成的原始文本;

[0079] 辅助合成特征获取单元,用于获取匹配文本对应的辅助合成特征,所述匹配文本与所述原始文本存在相匹配的文本片段,所述辅助合成特征为基于所述匹配文本对应的发音音频所确定的用于辅助语音合成的特征;

[0080] 辅助语音合成单元,用于参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音。

[0081] 在本申请的第三方面,提供了一种语音合成设备,包括:存储器和处理器;

[0082] 所述存储器,用于存储程序;

[0083] 所述处理器,用于执行所述程序,实现如上所述的语音合成方法的各个步骤。

[0084] 在本申请的第四方面,提供了一种存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时,实现如上所述的语音合成方法的各个步骤。

[0085] 在本申请的第五方面,提供了一种计算机程序产品,所述计算机程序产品在终端设备上运行时,使得所述终端设备执行上述的语音合成方法的各个步骤。

[0086] 借由上述技术方案,本申请的语音合成方法,在对待合成的原始文本进行语音合成的过程中,参考了与原始文本存在相匹配的文本片段的匹配文本对应的辅助合成特征,该辅助合成特征为基于匹配文本对应的发音音频所确定的用于辅助语音合成的特征,由此可知,本申请通过参考匹配文本对应的辅助合成特征,可以利用匹配文本对应的发音音频中发音信息辅助对原始文本进行语音合成,丰富了原始文本语音合成时所参考的信息,进而提高了原始文本的语音合成质量。

[0087] 可以理解的是,语音合成系统可以划分为带前端预处理和不带前端预处理两种类型,本申请方案可以同时适用于该两种类型的语音合成系统,对于带前端预处理的语音合

成系统,上述匹配文本对应的辅助合成特征可以作为语音合成前端的分析结果或辅助修正语音合成前端的分析结果,进而将分析结果送入语音合成后端辅助对原始文本进行语音合成,对于不带前端预处理的语音合成系统,匹配文本对应的辅助合成特征可以直接作为语音合成系统对原始文本进行合成时的参考信息。对于两种类型的语音合成系统,参考本申请的辅助合成特征进行原始文本的语音合成,能够丰富语音合成时的参考信息,进而能够提高合成语音的质量。

附图说明

[0088] 通过阅读下文优选实施方式的详细描述,各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的,而并不认为是对本申请的限制。而且在整个附图中,用相同的参考符号表示相同的部件。在附图中:

[0089] 图1为本申请实施例提供的语音合成方法的一流程示意图;

[0090] 图2示例了一种音素序列提取模型架构示意图;

[0091] 图3示例了一种语音合成后端的合成流程示意图;

[0092] 图4示例了一种语音合成系统架构示意图;

[0093] 图5示例了一种预测-生成网络确定音素级韵律编码的过程示意图;

[0094] 图6为本申请实施例提供的一种语音合成装置结构示意图;

[0095] 图7为本申请实施例提供的语音合成设备的结构示意图。

具体实施方式

[0096] 下面将结合本申请实施例中的附图,对本申请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申请一部分实施例,而不是全部的实施例。基于本申请中的实施例,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,都属于本申请保护的范围。

[0097] 本申请提供了一种语音合成方案,可以适用于各种语音合成任务。本申请的语音合成方案可以适用于人机交互场景下的语音合成工作,以及其它各种需要进行语音合成的场景。

[0098] 本申请方案可以基于具备数据处理能力的终端实现,该终端可以是手机、电脑、服务器、云端等。

[0099] 接下来,结合图1所述,本申请的语音合成方法可以包括如下步骤:

[0100] 步骤S100、获取待合成的原始文本。

[0101] 具体的,原始文本为待合成语音的文本。原始文本可以是用户提供的,也可以是其它设备、应用提供的需要进行语音合成的文本。

[0102] 步骤S110、获取匹配文本对应的辅助合成特征,所述匹配文本与所述原始文本存在相匹配的文本片段。

[0103] 其中,匹配文本可以是与原始文本或原始文本内的文本片段相匹配的文本,示例如,原始文本为“这条裤子不打折”,匹配文本可以是“这条裤子不打折”或“打折”。除此之外,匹配文本还可以是包含与原始文本内文本片段相匹配文本片段的文本。仍以上述原始文本为例,匹配文本可以是“你这件衣服打折吗”,也即匹配文本包含与原始文本相匹配的

文本片段“打折”。

[0104] 匹配文本可以是本申请预先配置存储的文本, 示例如, 在客服、交互等场景下, 可以预先将固定的话术文本进行录音, 将话术文本进行存储。则在存储的话术文本中查找与原始文本匹配的话术文本, 作为匹配文本。以客服及交互场景为例, 其存在一些固定的话术文本, 如智能客服或终端需要向用户进行信息提示的提示内容文本, 示例如“请问您需要查询什么内容”、“您好, 请问有什么可以帮助您的吗”“查询话费请按1, 查询流量请按2”等等。对应的, 这些固定的话术文本可以预先进行录音, 作为提示音与话术文本共同进行存储。

[0105] 除此之外, 匹配文本还可以是用户上传的文本。示例如, 用户在上传待合成的原始文本的同时, 将原始文本中容易合成出错的文本作为匹配文本一并上传, 同时还可以上传匹配文本对应的发音音频。再比如, 用户上传待合成的原始文本后, 合成系统输出合成的初始合成语音。用户可以确定出初始合成语音中合成错误的文本, 进而录制该合成错误的文本对应的发音音频, 将合成错误的文本及对应的发音音频上传至语音合成系统。或者是, 用户上传包含该合成错误的文本的扩展文本, 以及扩展文本对应的发音音频。

[0106] 上述匹配文本对应的辅助合成特征可以是, 基于所述匹配文本对应的发音音频所确定的用于辅助语音合成的特征。辅助合成特征包含了匹配文本对应的发音音频的发音信息, 示例如发音的音素序列、停顿信息、重读、韵律、情感等发音信息, 该发音信息能够辅助对原始文本的语音合成, 提升原始文本的语音合成质量。

[0107] 匹配文本对应的发音音频中, 对于匹配文本与原始文本匹配的文本片段的发音, 为该文本片段在原始文本中的标准发音。示例如, 原始文本为“这条裤子不打折”。匹配文本为“打折”, 则匹配文本对应的发音音频为“da zhe”对应的音频, 而非“dashe”等其它发音音频。基于此, 可以基于匹配文本对应发音音频确定辅助合成特征, 以辅助对原始文本的语音合成。

[0108] 可以理解的是, 若匹配文本对应的发音音频可以在对原始文本进行语音合成之前获取到, 则可以预先基于匹配文本对应的发音音频确定辅助合成特征并存储在本地或第三方设备。则本步骤中获取匹配文本对应的辅助合成特征的过程可以是, 在本地或第三方存储中查找预存储的匹配文本对应的辅助合成特征。

[0109] 除此之外, 若匹配文本对应的发音音频为对原始文本语音合成过程临时获取的, 则本步骤中获取匹配文本对应的辅助合成特征的过程可以是, 在获取到匹配文本对应的发音音频后, 基于该发音音频确定辅助合成特征。

[0110] 步骤S120、参考所述辅助合成特征, 对所述原始文本进行语音合成, 得到合成语音。

[0111] 具体的, 本步骤中语音合成系统在对原始文本进行语音合成时, 除了参考原始文本之外还可以进一步参考匹配文本对应的辅助合成特征, 也即丰富了原始文本语音合成过程所参考的信息。同时, 由于该辅助合成特征包含了匹配文本对应的发音音频的发音信息, 该发音信息能够辅助对原始文本的语音合成, 提升原始文本的语音合成质量。

[0112] 本申请实施例提供的语音合成方法, 在对待合成的原始文本进行语音合成的过程中, 参考了与原始文本存在相匹配的文本片段的匹配文本对应的辅助合成特征, 该辅助合成特征为基于匹配文本对应的发音音频所确定的用于辅助语音合成的特征, 由此可知, 本申请通过参考匹配文本对应的辅助合成特征, 可以利用匹配文本对应的发音音频中发音信

息辅助对原始文本进行语音合成,丰富了原始文本语音合成时所参考的信息,进而提高了原始文本的语音合成质量。

[0113] 可以理解的是,语音合成系统可以划分为带前端预处理和不带前端预处理两种类型,本申请方案可以同时适用于该两种类型的语音合成系统,对于带前端预处理的语音合成系统,上述匹配文本对应的辅助合成特征可以作为语音合成前端的分析结果或辅助修正语音合成前端的分析结果,进而将分析结果送入语音合成后端辅助对原始文本进行语音合成,对于不带前端预处理的语音合成系统,匹配文本对应的辅助合成特征可以直接作为语音合成系统对原始文本进行合成时的参考信息。对于两种类型的语音合成系统,参考本申请的辅助合成特征进行原始文本的语音合成,能够丰富语音合成时的参考信息,进而能够提高合成语音的质量。

[0114] 在本申请的一些实施例中,对前文中提及的匹配文本对应的辅助合成特征,及参考辅助合成特征对原始文本进行语音合成的过程进行说明。

[0115] 辅助合成特征为基于匹配文本对应的发音音频所确定的用于辅助语音合成的特征,辅助合成特征包含了匹配文本对应的发音音频的发音信息,该发音信息能够辅助对原始文本的语音合成,提升原始文本的语音合成质量。

[0116] 本实施例中提供了几种可选形式的辅助合成特征,如下实施例中分别进行介绍:

[0117] 1)、辅助合成特征为匹配文本对应的音素序列。

[0118] 具体的,语音合成系统可以分为带前端预处理和不带前端预处理两种类型。其中,带前端预处理的语音合成系统在对原始文本进行语音合成之前,首先对原始文本进行前端分析,如通过查询发音词典预测原始文本对应的音素序列,由语音合成后端基于原始文本及音素序列进行语音合成。

[0119] 这种处理方式,能够一定程度上提升语音合成的质量,但是,当预先构建的发音词典存在错误时,则会导致后端合成语音出错。

[0120] 为此,本实施例中可以基于匹配文本对应的发音音频,确定匹配文本对应的音素序列,作为辅助合成特征。

[0121] 可以理解的是,匹配文本对应的发音音频为正确发音,因此可以从发音音频中提取出匹配文本对应的正确音素序列。该正确的音素序列可以作为辅助合成特征,参与到对原始文本的语音合成过程。

[0122] 本实施例中提供了一种从匹配文本对应的发音音频中,提取音素序列的实现方式。

[0123] 如图2所示,其示例了一种音素序列提取模型架构示意图。

[0124] 本申请可以预先训练音素序列提取模型,用于从发音音频中提取音素序列。

[0125] 音素序列提取模型可以采用LSTM(long short term memory,长短时记忆网络)网络架构或HMM、CNN等其它可选的网络架构。如图2所示,其示例了一种采用编码-注意力-解码架构的音素序列提取模型。

[0126] 编码端使用LSTM网络对发音音频的音频特征序列 $(x_1, x_2, \dots, x_n)$ 进行编码得到隐层编码序列 $(h_1, h_2, \dots, h_n)$,解码端同样采用LSTM网络,在解码时刻 t 通过输入 $t-1$ 时刻的隐层状态 h_{t-1} 和由注意力模块计算出的上下文向量 c_{t-1} 共同计算得到解码端隐层向量 s_t ,然后通过投影得到 t 时刻的音素 y_t 。当解码出特殊符号结束标记时解码停止,得到音素序列 $(y_1,$

$y_2, \dots, y_t)$ 。

[0127] 示例性说明如：

[0128] 当匹配文本为“这件衣服不打折”时，从匹配文本对应的发音音频中提取的音素序列为：[zh e4 j ian4 i1 f u7 b u4 d a3 zh e2]。

[0129] 当辅助合成特征为音素序列时，上述步骤S120，参考所述辅助合成特征，对所述原始文本进行语音合成的过程，可以包括：

[0130] S1、基于所述匹配文本对应的音素序列，确定所述原始文本的音素序列。

[0131] 具体的，可以基于所述匹配文本对应的音素序列，获取所述匹配文本与所述原始文本中相同文本片段对应的音素序列。

[0132] 示例如，确定所述匹配文本与所述原始文本中相同文本片段，进而在所述匹配文本对应的音素序列中，提取所述相同文本片段对应的音素序列。

[0133] 进一步，查询发音词典，确定所述原始文本中除所述相同文本片段外的其余文本片段的音素序列，并与所述相同文本片段对应的音素序列组合，得到原始文本的音素序列。

[0134] 当然，还可以通过查询发音词典，确定所述原始文本对应的初始音素序列，并利用在匹配文本对应的音素序列中提取的所述相同文本片段对应的音素序列，替换掉所述初始音素序列中所述相同文本片段对应的音素序列，得到原始文本对应的替换后的音素序列。

[0135] S2、基于所述原始文本的音素序列，对所述原始文本进行语音合成，得到合成语音。

[0136] 具体的，可以将原始文本的音素序列作为语音合成前端的文本分析结果，送入语音合成后端辅助对原始文本进行语音合成。

[0137] 由于本实施例中得到的原始文本的音素序列中包含了匹配文本对应的音素序列，该部分音素序列是基于匹配文本对应的正确发音音频所确定的，从而以原始文本的音素序列辅助进行语音合成时，能够提升合成语音的准确度，尤其是对于一些多音字、易错字，其合成语音准确度大大提高。

[0138] 2)、辅助合成特征为匹配文本对应的韵律信息。

[0139] 结合前文介绍，语音合成前端可以对原始文本进行文本分析，该文本分析的过程还可以预测原始文本的韵律信息，进而由合成后端基于原始文本、韵律信息进行语音合成。通过考虑韵律信息，可以提升合成语音的自然度。

[0140] 可以理解的是，对原始文本所预测的韵律信息也可能出错，进而导致后端合成语音的韵律出错，影响合成语音的质量。

[0141] 为此，本实施例中可以基于匹配文本对应的发音音频，确定匹配文本对应的韵律信息，作为辅助合成特征。这里，匹配文本对应的韵律信息，可以是音素级韵律信息，其包括匹配文本对应的音素序列中，每一音素单元的韵律信息。

[0142] 可以理解的是，匹配文本对应的发音音频为正确发音，因此可以从发音音频中提取出匹配文本对应的正确的韵律信息。该正确的韵律信息可以作为辅助合成特征，参与到对原始文本的语音合成过程。示例如，基于该正确的韵律信息确定原始文本的修正后韵律信息，进而送入合成后端进行语音合成。

[0143] 当辅助合成特征为韵律信息时，上述步骤S120，参考所述辅助合成特征，对所述原始文本进行语音合成的过程，可以包括：

[0144] S1、基于所述匹配文本对应的韵律信息,确定所述原始文本的韵律信息。

[0145] 具体的,可以基于所述匹配文本对应的韵律信息,获取所述匹配文本与所述原始文本中相同文本片段对应的韵律信息。

[0146] 进一步,可以采用韵律预测技术预测原始文本中除所述相同文本片段的其余文本片段的韵律信息,并与所述相同文本片段对应的韵律信息组合,得到原始文本的韵律信息。

[0147] S2、基于所述原始文本的韵律信息,对所述原始文本进行语音合成,得到合成语音。

[0148] 另一种情况下,当辅助合成特征同时包含音素序列和韵律信息时,上述步骤S120,参考所述辅助合成特征,对所述原始文本进行语音合成的过程,可以包括:

[0149] S1、基于所述匹配文本对应的音素序列和韵律信息,确定所述原始文本的音素序列和韵律信息。

[0150] S2、基于所述原始文本的音素序列和韵律信息,对所述原始文本进行语音合成,得到合成语音。

[0151] 3)、辅助合成特征为匹配文本对应的音素级韵律编码。

[0152] 具体的,匹配文本对应的音素级韵律编码包含了匹配文本对应的发音音频的一些发音信息,如发音时长、重读强调等韵律特征。

[0153] 语音合成后端在进行语音合成时,可以对原始文本的韵律信息进行建模,进而提升合成语音的自然度。本实施例中,为了提升语音合成后端对原始文本韵律信息的建模准确度,可以将匹配文本对应的音素级韵律编码作为辅助合成特征,送入语音合成后端辅助进行语音合成。

[0154] 可以理解的是,匹配文本对应的音素级韵律编码包含了匹配文本对应的正确发音信息,语音合成后端基于匹配文本对应的音素级韵律编码进行语音合成时,对于原始文本与匹配文本共同包含的相同文本片段,可以合成出与匹配文本的发音音频一致的语音。

[0155] 同时,语音合成后端对原始文本进行卷积等运算处理过程,对于原始文本中除相同文本片段外的其余文本片段,其在处理过程也会参考相同文本片段对应的音素级韵律编码,从而利用相同文本片段的音素级韵律编码辅助提升原始文本中其余文本片段的语音合成质量。

[0156] 此外,某些现有技术中,仅对原始文本中非相同文本片段进行语音合成,进而将非相同文本片段的合成语音,与预配置的相同文本片段的语音进行拼接,得到原始文本对应的整体合成语音。这种处理方式会导致原始文本的整体合成语音出现音色不一致的问题,降低了合成语音的质量。

[0157] 而本申请的语音合成系统仍然是一个完整的合成系统,通过对原始文本进行整体的语音合成,可以保障合成语音的音色是一致的。

[0158] 进一步的,基于语音合成后端对韵律信息建模的不同形式,本实施例中音素级韵律编码也可以不同。

[0159] 如图3,其示例了一种语音合成后端的合成流程示意图。

[0160] 由图3可知,语音合成后端包括时长模型和声学模型,通过时长模型和声学模型分别对时长韵律信息和声学参数韵律信息进行建模。

[0161] 则为了适配图3所示的语音合成后端的模型结构,本申请实施例中匹配文本对应

的音素级韵律编码可以包括时长编码和声学参数编码。

[0162] 则匹配文本对应的韵律编码在送入语音合成后端辅助进行语音合成时,可以将时长编码送入时长模型辅助进行音素级时长建模,将声学参数编码送入声学模型辅助进行音素级声学参数建模。

[0163] 其中,声学参数编码可以包括一种或多种不同的声学参数编码,示例如基频声学参数编码或其它声学参数编码等。

[0164] 在前述示例的辅助合成特征包括音素序列、韵律信息的基础上,进一步的当辅助合成特征还包括音素级韵律编码时,上述步骤S120,参考所述辅助合成特征,对所述原始文本进行语音合成的过程,还可以进一步包括:

[0165] S3、基于所述匹配文本对应的音素级韵律编码,获取所述匹配文本与所述原始文本中相同文本片段对应的音素级韵律编码。

[0166] 具体的,可以确定所述匹配文本与所述原始文本中相同文本片段,进而在所述匹配文本对应的音素级韵律编码中,提取所述相同文本片段对应的音素级韵律编码。

[0167] S4、在对所述原始文本进行语音合成过程中,将所述相同文本片段对应的音素级韵律编码作为语音合成模型的补充输入,得到合成语音。

[0168] 仍以图3为例进行说明:

[0169] 音素级韵律编码包括时长编码和声学参数编码。

[0170] 则语音合成后端在对原始文本进行语音合成过程中,可以将所述相同文本片段对应的时长编码送入时长模型进行音素级时长建模,以及,将所述相同文本片段对应的声学参数编码送入声学模型进行音素级声学参数建模,最终由语音合成后端得到合成语音。

[0171] 4)、辅助合成特征为匹配文本对应的发音音频的声学特征。

[0172] 如前文介绍所述,语音合成系统可以分为带前端预处理和不带前端预处理两种类型。其中,不带前端预处理的语音合成系统不会对原始文本进行前端分析,而是直接对原始文本进行语音合成。为了提升对原始文本的合成语音的质量把控,本实施例中可以将匹配文本对应的发音音频的声学特征作为辅助合成特征,送入语音合成系统辅助对原始文本进行语音合成。该声学特征包含了匹配文本的发音音频的发音信息,则语音合成系统在对原始文本逐帧进行语音合成时,可以从声学特征中提取与每一帧相关联的声学特征,以辅助合成每一帧,可以达到修正发音错误的问题,如纠正比较容易出错的罕见字、特殊符号、多音字以及外来词汇等的发音错误,最终得到质量较高的合成语音。

[0173] 其中,声学特征包括但不限于发音音频的倒谱特征。

[0174] 当辅助合成特征为匹配文本对应的发音音频的声学特征时,上述步骤S120,参考所述辅助合成特征,对所述原始文本进行语音合成的过程,可以包括:

[0175] S1、基于语音合成模型处理所述原始文本,得到预测当前语音帧的上下文信息。

[0176] 具体的,语音合成模型可以采用encoder-decoder的编解码架构,进一步可以在编码和解码层之间通过注意力模块连接。则原始文本经过encoder-decoder编解码架构以及注意力模块,可以得到合成当前语音帧 y_t 时所需的上下文信息 C_t 。该上下文信息 C_t 表明了合成当前语音帧 y_t 所需要的原始文本中的文本信息。

[0177] S2、基于所述上下文信息、所述匹配文本及所述发音音频的声学特征,确定预测当前语音帧所需的目标声学特征。

[0178] 一种可选的实现方式中,步骤S2可以包括:

[0179] S21、基于所述上下文信息、所述匹配文本及所述发音音频的声学特征,获取所述上下文信息与所述发音音频的声学特征中,每一帧声学特征的关联度。

[0180] 具体的,上下文信息可以通过注意力机制得到与匹配文本的相似程度,以及通过发音音频的声学特征对匹配文本的注意力矩阵,获取每一帧声学特征与匹配文本的关联度,在此基础上,基于上下文信息与匹配文本的相似程度,以及每一帧声学特征与匹配文本的关联度,可以获取到上下文信息与每一帧声学特征的关联度,该关联度表明了上下文信息和每一帧声学特征的接近程度。可以理解的是,当上下文信息与一目标帧声学特征的关联度较高时,表明该上下文信息对应的文本的发音与目标帧声学特征强相关。

[0181] 接下来,介绍步骤S21的一种可选实现方式,可以包括如下步骤:

[0182] S211、获取所述发音音频的声学特征对所述匹配文本的第一注意力权重矩阵 W_{mx} 。

[0183] 其中,第一注意力权重矩阵 W_{mx} 包括每一帧声学特征对所述匹配文本中各文本单元的注意力权重。矩阵 W_{mx} 的大小为 $T_{my} * T_{mx}$,其中 T_{my} 表示发音音频对应的声学特征的帧长, T_{mx} 表示匹配文本的长度。

[0184] S212、获取所述上下文信息 C_t 对所述匹配文本的第二注意力权重矩阵 W_{cmx} 。

[0185] 其中,所述第二注意力权重矩阵 W_{cmx} 包括所述上下文信息 C_t 对所述匹配文本中各文本单元的注意力权重。矩阵 W_{cmx} 的大小为 $1 * T_{mx}$ 。

[0186] S213、基于所述第一注意力权重 W_{mx} 及所述第二注意力权重矩阵 W_{cmx} ,得到所述上下文信息 C_t 对所述声学特征的第三注意力权重矩阵 W_{cmy} 。

[0187] 其中,所述第三注意力权重矩阵 W_{cmy} 包括所述上下文信息 C_t 对每一帧声学特征的注意力权重,作为所述上下文信息与每一帧声学特征的关联度。矩阵 W_{cmy} 的大小为 $1 * T_{my}$ 。矩阵 W_{cmy} 可以表示为:

$$[0188] \quad W_{cmy} = W_{cmx} * W_{mx}'$$

[0189] 其中, W_{mx}' 表示矩阵 W_{mx} 的转置。

[0190] S22、基于所述关联度,确定预测当前语音帧所需的目标声学特征。

[0191] 具体的,在上述步骤中得到上下文信息与发音音频的声学特征中,每一帧声学特征的关联度之后,可以首先对各个关联度进行归一化处理,并以归一化后的各个关联度作为权重,对所述发音音频的各帧声学特征进行加权相加,得到预测当前语音帧所需的目标声学特征。目标声学特征可以表示为 C_{mt} 。

[0192] S3、基于所述上下文信息,及确定的所述目标声学特征,预测当前语音帧,并在预测得到所有语音帧后,由预测的各语音帧组成合成语音。

[0193] 可以理解的是,由于原始文本和匹配文本存在匹配的文本片段,但原始文本可能未必与匹配文本完全相同,这就导致对于上述步骤得到的预测当前语音帧所需的目标声学特征 C_{mt} ,除了对原始文本中与匹配文本相同的文本片段进行语音合成时能够使用到之外,其余文本片段的合成过程不需要使用该目标声学特征 C_{mt} 。为此,本实施例提供了一种解决方案,使得在对原始文本进行语音合成时,对于待预测的不同语音帧,可以控制所参考的目标声学特征 C_{mt} 的信息量。具体实现过程可以包括:

[0194] S31、基于语音合成模型的解码端当前的隐层矢量及所述目标声学特征 C_{mt} ,确定预测当前语音帧时所述目标声学特征 C_{mt} 的融合系数 a_{gate} 。

[0195] 具体的,本实施例中可以采用门限机制或其它策略来决定预测当前语音帧时所述目标声学特征 C_{mt} 的融合系数 a_{gate} 。以门限机制为例, a_{gate} 可以表示为:

[0196] $a_{gate} = \text{sigmoid}(g_g(C_{mt}, s_t))$

[0197] 其中, s_t 表示解码端当前的隐层矢量, $g_g()$ 表示设定函数关系。

[0198] S32、参考所述融合系数 a_{gate} ,对所述目标声学特征 C_{mt} 及所述上下文信息 C_t 进行融合,并基于融合结果预测当前语音帧。

[0199] 具体的,当前语音帧 y_t 可以表示为:

[0200] $y_t = g(y_{t-1}, s_t, (1 - a_{gate}) * C_t + a_{gate} * C_{mt})$

[0201] 其中, $g()$ 表示设定的函数关系。

[0202] 参考图4,其示例了一种语音合成系统架构示意图。

[0203] 图4示例的语音合成系统采用编解码加注意力机制的端到端合成流程。

[0204] 原始文本经过编码端编码得到原始文本的编码向量,经过第一注意力模块可以获得预测当前语音帧 y_t 时所需的上下文信息 C_t 。

[0205] 匹配文本经过编码端编码得到匹配文本的编码向量。进一步,上下文信息 C_t 可以通过第二注意力模块得到对匹配文本中各文本单元的注意力权重,组成第二注意力权重矩阵。

[0206] 同时,本实施例中可以获取到匹配文本的发音音频的声学特征对匹配文本的注意力权重,组成第一注意力权重矩阵。进而基于第一注意力权重矩阵和第二注意力权重矩阵,得到上下文信息 C_t 对声学特征的第三注意力权重矩阵。第三注意力权重矩阵包括上下文信息 C_t 与每一帧声学特征的关联度。通过对第三注意力权重矩阵进行softmax规整,并与发音音频的各帧声学特征进行加权相加,得到预测当前语音帧 y_t 所需要的目标声学特征 C_{mt} 。

[0207] 解码端可以基于目标声学特征 C_{mt} 和上下文信息 C_t ,预测当前语音帧 y_t 。

[0208] 解码端预测当前语音帧 y_t 的表达式可以参照前文相关介绍。

[0209] 预测得到的各语音帧通过声码器映射为合成语音。

[0210] 在本申请的一些实施例中,对前述步骤S110,获取匹配文本对应的辅助合成特征的过程进行介绍,具体的,该过程可以包括:

[0211] S1、获取与所述原始文本存在相匹配的文本片段的匹配文本。

[0212] 本实施例中提供了两种不同的实现方式,分别介绍如下:

[0213] 1)、一种可选的实施方式下,本申请可以预先收集语音合成场景下大量的固定话术文本并进行录音,将收集的话术文本作为模板文本,同时存储模板文本及对应的发音音频。或者是,基于模板文本的发音音频确定辅助合成特征,进而将模板文本与辅助合成特征一并进行存储。

[0214] 基于此,步骤S1的实施过程可以包括:

[0215] S11、在预先配置存储的模板文本中,确定与所述原始文本内的文本片段相匹配的匹配文本。

[0216] 可选的,本实施例中可以基于收集的模板文本及对应的发音音频,整理打包成资源包。具体的,每一资源包中包含一模板文本,以及基于该模板文本对应的发音音频所确定的,与所述模板文本对应的辅助合成特征。

[0217] 辅助合成特征可以包括模板文本对应的音素序列及韵律信息。进一步的,辅助合

成特征还可以包括模板文本对应的音素级韵律编码。

[0218] 举例说明如：

[0219] 模板文本为“欢迎使用人工智能语音助手”。

[0220] 基于该模板文本对应的发音音频，可以确定的辅助合成特征可以包括模板文本的音素序列、韵律信息、音素级韵律编码等。进而，可以将模板文本与辅助合成特征打包成一个资源包。

[0221] 以模板文本的韵律信息为例，其示例性格式可以如下：

[0222] “欢[=huan1]迎[=ying2][w1]使[=shi3]用[=yong4][w3]人[=ren2]工[=gong1]智[=zhi4]能[=neng2][w1]语[=yu3]音[=yin1][w1]助[=zhu4]手[=shou3]”。

[0223] 其中，每个字的发音通过[=pinyin]来指定，“[w1]”、“[w3]”表示不同的韵律停顿信息。

[0224] 可以理解的是，上述仅仅是本申请示例的一种韵律信息表示方式，本领域技术人员还可以采用其它不同的标记格式来表示模板文本的韵律信息。

[0225] 对于打包后的资源包，可以编码成一个二进制资源文本，以降低存储空间的占用，同时便于后续语音合成系统的处理识别。

[0226] 结合图5，对模板文本对应的音素级韵律编码的确定过程进行介绍。

[0227] 如图5所示，可以基于编码预测网络及生成网络来确定模板文本对应的音素级韵律编码，具体可以包括如下步骤：

[0228] A1、基于所述模板文本及对应的发音音频，提取出音素级的韵律信息。

[0229] A2、将所述模板文本及所述音素级的韵律信息输入编码预测网络，得到预测的音素级的韵律编码。

[0230] A3、将所述预测的音素级的韵律编码及所述模板文本输入生成网络，得到生成的音素级的韵律信息。

[0231] A4、以生成的音素级的韵律信息趋近于提取出的所述音素级的韵律信息为目标训练所述编码预测网络及所述生成网络，直至训练结束时，得到训练后的编码预测网络所预测的模板文本对应的音素级韵律编码。

[0232] 其中，以生成的音素级的韵律信息趋近于提取出的所述音素级的韵律信息为目标训练所述编码预测网络及所述生成网络的过程，具体可以是计算生成的音素级的韵律信息和提取出的所述音素级的韵律信息的均方误差MSE，通过迭代训练调整网络参数，使得MSE达到预设阈值，则可以结束训练。

[0233] 进一步的，基于上述预先配置的资源包，上述步骤S11，在预先配置存储的模板文本中，确定与所述原始文本内的文本片段相匹配的匹配文本的实现过程，可以包括：

[0234] S111、分别将所述原始文本与预配置的每个资源包中的模板文本进行匹配计算。

[0235] S112、在匹配度最高的资源包所包含的模板文本中，确定与所述原始文本内的文本片段相匹配的匹配文本。

[0236] 具体的，上述匹配计算过程可以先确定是否存在与原始文本完全匹配的模板文本，若存在，则将完全匹配的模板文本确定为匹配文本。若不存在，则可以进行部分匹配，例如，从原始文本的一端或两端开始，在每个资源包的模板文本中查找最大长度匹配文本，作为匹配文本。

[0237] 示例如,原始文本为“请问您是王宁吗?”,在与资源包中的模板文本进行匹配时,未匹配到完全相同的模板文本,但是匹配到模板文本“请问您是刘武吗?”,将原始文本与上述模板文本进行最大长度匹配,可以得到匹配文本:“请问您是”和“吗?”。

[0238] 2)、另一种可选的实施方式下,本申请可以获取用户上传数据。上传数据中包含上传文本及上传文本对应的发音音频。该上传文本与原始文本存在相匹配的文本片段。进而可以将上传文本作为匹配文本。

[0239] 一种可选的场景如,在上述步骤S100获取待合成的原始文本之后,可以进行初始语音合成,并输出原始文本的初始合成语音。对原始文本进行初始语音合成的过程,可以使用现有或未来可能出现的各种语音合成方案。用户在收到初始合成语音后可以确定出初始合成语音中合成错误的文本片段,并确定该合成错误的文本片段对应的正确发音,进而可以将合成错误的文本片段作为上传文本,将该合成错误的文本片段对应的正确发音作为上传文本对应的发音音频,一并作为上传数据进行上传。或者是,用户可以获取到包含所述初始合成语音中合成错误的文本片段的扩展文本,以及获取到扩展文本对应的正确发音,将扩展文本作为上传文本,扩展文本对应的正确发音作为上传文本对应的发音音频,一并作为上传数据进行上传。

[0240] S2、获取基于所述匹配文本对应的发音音频所确定的辅助合成特征。

[0241] 具体的,参考前文相关介绍可知,若匹配文本对应的发音音频可以在对原始文本进行语音合成之前获取到,则可以预先基于匹配文本对应的发音音频确定辅助合成特征并存储在本地或第三方设备。则本步骤中获取匹配文本对应的辅助合成特征的过程可以是,在本地或第三方存储中查找预存储的匹配文本对应的辅助合成特征。

[0242] 此外,若匹配文本对应的发音音频为对原始文本语音合成过程临时获取的,则本步骤中获取匹配文本对应的辅助合成特征的过程可以是,在获取到匹配文本对应的发音音频后,基于该发音音频确定辅助合成特征。

[0243] 需要说明的是,若上述步骤S1获取匹配文本的方式通过第一种1)方式实现,即分别将所述原始文本与预配置的每个资源包中的模板文本进行匹配计算,并在匹配度最高的资源包所包含的模板文本中,确定与所述原始文本内的文本片段相匹配的匹配文本,则上述步骤S2的实现过程具体可以包括:

[0244] S21、获取所述匹配度最高的资源包中包含的,与所述匹配文本对应的辅助合成特征。

[0245] 可以理解的是,资源包中包含模板文本对应的辅助合成特征,如音素序列、韵律信息、音素级韵律编码等。而匹配文本与模板文本相同或属于模板文本中的部分文本片段,因此可以从模板文本对应的辅助合成特征中,提取出匹配文本对应的辅助合成特征。

[0246] 进一步,若上述步骤S1获取匹配文本的方式通过第二种2)方式实现,即将用户上传数据中的上传文本作为匹配文本,则上述步骤S2的实现过程具体可以包括:

[0247] 基于所述上传数据中所述匹配文本对应的发音音频,确定匹配文本对应的辅助合成特征。

[0248] 下面对本申请实施例提供的语音合成装置进行描述,下文描述的语音合成装置与上文描述的语音合成方法可相互对应参照。

[0249] 参见图6,图6为本申请实施例公开的一种语音合成装置结构示意图。

- [0250] 如图6所示,该装置可以包括:
- [0251] 原始文本获取单元11,用于获取待合成的原始文本;
- [0252] 辅助合成特征获取单元12,用于获取匹配文本对应的辅助合成特征,所述匹配文本与所述原始文本存在相匹配的文本片段,所述辅助合成特征为基于所述匹配文本对应的发音音频所确定的用于辅助语音合成的特征;
- [0253] 辅助语音合成单元13,用于参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音。
- [0254] 可选的,上述辅助合成特征获取单元获取匹配文本对应的辅助合成特征的过程,可以包括:
- [0255] 获取与所述原始文本存在相匹配的文本片段的匹配文本;
- [0256] 获取基于所述匹配文本对应的发音音频所确定的辅助合成特征。
- [0257] 可选的,上述辅助合成特征可以包括:
- [0258] 基于所述匹配文本对应的发音音频所确定的,所述匹配文本对应的音素序列;
- [0259] 和/或,
- [0260] 基于所述匹配文本对应的发音音频所确定的,所述匹配文本对应的韵律信息;
- [0261] 和/或,
- [0262] 基于所述匹配文本对应的发音音频所确定的,所述匹配文本对应的音素级韵律编码;
- [0263] 和/或,
- [0264] 所述匹配文本对应的发音音频的声学特征。
- [0265] 可选的,上述辅助合成特征获取单元获取与所述原始文本存在相匹配的文本片段的匹配文本的过程,可以包括:
- [0266] 在预配置的模板文本中,确定与所述原始文本内的文本片段相匹配的匹配文本。
- [0267] 可选的,上述预配置的模板文本可以包括:
- [0268] 各个预配置的资源包中的模板文本,其中每一资源包包含一模板文本,及基于所述模板文本对应的发音音频所确定的,与所述模板文本对应的辅助合成特征。
- [0269] 可选的,上述辅助合成特征获取单元在预配置的模板文本中,确定与所述原始文本内的文本片段相匹配的匹配文本的过程,可以包括:
- [0270] 分别将所述原始文本与预配置的每个资源包中的模板文本进行匹配计算;
- [0271] 在匹配度最高的资源包所包含的模板文本中,确定与所述原始文本内的文本片段相匹配的匹配文本。
- [0272] 可选的,上述辅助合成特征获取单元获取基于所述匹配文本对应的发音音频所确定的辅助合成特征的过程,可以包括:
- [0273] 获取所述匹配度最高的资源包中包含的,与所述匹配文本对应的辅助合成特征。
- [0274] 可选的,本申请的装置还可以包括:资源包配置单元,用于配置资源包,该过程可以包括:
- [0275] 获取预配置的模板文本及对应的发音音频;
- [0276] 基于所述发音音频,确定所述模板文本对应的音素序列及韵律信息;
- [0277] 将所述音素序列及韵律信息作为所述模板文本对应的辅助合成特征,并将所述辅

助合成特征与所述模板文本整理成一个资源包。

[0278] 可选的,上述资源包配置单元配置资源包的过程还可以包括:

[0279] 基于所述模板文本及对应的发音音频,确定所述模板文本对应的音素级韵律编码;

[0280] 将所述音素级韵律编码合并入所述资源包中。

[0281] 可选的,上述资源包配置单元基于所述模板文本及对应的发音音频,确定所述模板文本对应的音素级韵律编码的过程,可以包括:

[0282] 基于所述模板文本及对应的发音音频,提取出音素级的韵律信息;

[0283] 将所述模板文本及所述音素级的韵律信息输入编码预测网络,得到预测的音素级的韵律编码;

[0284] 将所述预测的音素级的韵律编码及所述模板文本输入生成网络,得到生成的音素级的韵律信息;

[0285] 以生成的音素级的韵律信息趋近于提取出的所述音素级的韵律信息为目标训练所述编码预测网络及所述生成网络,直至训练结束时,得到训练后的编码预测网络所预测的音素级的韵律编码。

[0286] 另一种可选的情况下,上述辅助合成特征获取单元获取与所述原始文本存在相匹配的文本片段的匹配文本的过程,可以包括:

[0287] 获取上传数据中的上传文本,作为所述匹配文本,所述上传数据还包括所述上传文本对应的发音音频,所述上传文本与所述原始文本存在相匹配的文本片段。在此基础上,辅助合成特征获取单元获取基于所述匹配文本对应的发音音频所确定的辅助合成特征的过程,可以包括:

[0288] 基于所述上传数据中所述匹配文本对应的发音音频,确定匹配文本对应的辅助合成特征。

[0289] 可选的,本申请的装置还可以包括:初始合成语音输出单元,用于在所述获取上传数据中的上传文本之前,输出所述原始文本的初始合成语音。在此基础上,所述上传文本为,所述初始合成语音中合成错误的文本片段,所述上传文本对应的发音音频为,所述合成错误的文本片段对应的正确发音;或,所述上传文本为,包含所述初始合成语音中合成错误的文本片段的扩展文本,所述上传文本对应的发音音频为,所述扩展文本对应的正确发音。

[0290] 可选的,当辅助合成特征包括匹配文本对应的音素序列和/或韵律信息时,上述辅助语音合成单元参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音的过程,可以包括:

[0291] 基于所述匹配文本对应的音素序列,确定所述原始文本的音素序列;

[0292] 和/或,

[0293] 基于所述匹配文本对应的韵律信息,确定所述原始文本的韵律信息;

[0294] 基于所述原始文本的音素序列和/或韵律信息,对所述原始文本进行语音合成,得到合成语音。

[0295] 进一步可选的,当辅助合成特征进一步还包括匹配文本对应的音素级韵律编码时,上述辅助语音合成单元参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音的过程,还可以包括:

[0296] 基于所述匹配文本对应的音素级韵律编码,获取所述匹配文本与所述原始文本中相同文本片段对应的音素级韵律编码;

[0297] 在对所述原始文本进行语音合成过程中,将所述相同文本片段对应的音素级韵律编码作为语音合成模型的补充输入,得到合成语音。

[0298] 可选的,上述辅助语音合成单元基于所述匹配文本对应的音素序列,确定所述原始文本的音素序列的过程,可以包括:

[0299] 基于所述匹配文本对应的音素序列,获取所述匹配文本与所述原始文本中相同文本片段对应的音素序列;

[0300] 查询发音词典,确定所述原始文本中除所述相同文本片段外的其余文本片段的音素序列,并与所述相同文本片段对应的音素序列合成,得到原始文本的音素序列。

[0301] 可选的,当辅助合成特征包括匹配文本对应的发音音频的声学特征时,上述辅助语音合成单元参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音的过程,可以包括:

[0302] 基于语音合成模型处理所述原始文本,得到预测当前语音帧的上下文信息;

[0303] 基于所述上下文信息、所述匹配文本及所述发音音频的声学特征,确定预测当前语音帧所需的目标声学特征;

[0304] 基于所述上下文信息,及确定的所述目标声学特征,预测当前语音帧,在预测得到所有语音帧后,由预测的各语音帧组成合成语音。

[0305] 可选的,上述辅助语音合成单元基于所述上下文信息、所述匹配文本及所述发音音频的声学特征,确定预测当前语音帧所需的目标声学特征的过程,可以包括:

[0306] 基于所述上下文信息、所述匹配文本及所述发音音频的声学特征,获取所述上下文信息与所述发音音频的声学特征中,每一帧声学特征的关联度;

[0307] 基于所述关联度,确定预测当前语音帧所需的目标声学特征。

[0308] 可选的,上述辅助语音合成单元获取所述上下文信息与所述发音音频的声学特征中,每一帧声学特征的关联度的过程,可以包括:

[0309] 获取所述发音音频的声学特征对所述匹配文本的第一注意力权重矩阵,所述第一注意力权重矩阵包括每一帧声学特征对所述匹配文本中各文本单元的注意力权重;

[0310] 获取所述上下文信息对所述匹配文本的第二注意力权重矩阵,所述第二注意力权重矩阵包括所述上下文信息对所述匹配文本中各文本单元的注意力权重;

[0311] 基于所述第一注意力权重及所述第二注意力权重矩阵,得到所述上下文信息对所述声学特征的第三注意力权重矩阵,所述第三注意力权重矩阵包括所述上下文信息对每一帧声学特征的注意力权重,作为所述上下文信息与每一帧声学特征的关联度。

[0312] 可选的,上述辅助语音合成单元基于所述关联度,确定预测当前语音帧所需的目标声学特征的过程,可以包括:

[0313] 对各个所述关联度进行归一化,并以归一化后的各个关联度作为权重,对所述发音音频的各帧声学特征进行加权相加,得到目标声学特征。

[0314] 可选的,上述辅助语音合成单元基于所述上下文信息,及确定的所述目标声学特征,预测当前语音帧的过程,可以包括:

[0315] 基于语音合成模型的解码端当前的隐层矢量及所述目标声学特征,确定预测当前

语音帧时所述目标声学特征的融合系数；

[0316] 参考所述融合系数,对所述目标声学特征及所述上下文信息进行融合,并基于融合结果预测当前语音帧。

[0317] 本申请实施例提供的语音合成装置可应用于语音合成设备,如终端:手机、电脑等。可选的,图7示出了语音合成设备的硬件结构框图,参照图7,语音合成设备的硬件结构可以包括:至少一个处理器1,至少一个通信接口2,至少一个存储器3和至少一个通信总线4;

[0318] 在本申请实施例中,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;

[0319] 处理器1可能是一个中央处理器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发明实施例的一个或多个集成电路等;

[0320] 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)等,例如至少一个磁盘存储器;

[0321] 其中,存储器存储有程序,处理器可调用存储器存储的程序,所述程序用于:

[0322] 获取待合成的原始文本;

[0323] 获取匹配文本对应的辅助合成特征,所述匹配文本与所述原始文本存在相匹配的文本片段,所述辅助合成特征为基于所述匹配文本对应的发音音频所确定的用于辅助语音合成的特征;

[0324] 参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音。

[0325] 可选的,所述程序的细化功能和扩展功能可参照上文描述。

[0326] 本申请实施例还提供一种存储介质,该存储介质可存储有适于处理器执行的程序,所述程序用于:

[0327] 获取待合成的原始文本;

[0328] 获取匹配文本对应的辅助合成特征,所述匹配文本与所述原始文本存在相匹配的文本片段,所述辅助合成特征为基于所述匹配文本对应的发音音频所确定的用于辅助语音合成的特征;

[0329] 参考所述辅助合成特征,对所述原始文本进行语音合成,得到合成语音。

[0330] 可选的,所述程序的细化功能和扩展功能可参照上文描述。

[0331] 进一步地,本申请实施例还提供了一种计算机程序产品,所述计算机程序产品在终端设备上运行时,使得所述终端设备执行上述语音合成方法中的任意一种实现方式。

[0332] 最后,还需要说明的是,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、物品或者设备不仅包括那些要素,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、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0333] 本说明书中各个实施例采用递进的方式描述,每个实施例重点说明的都是与其他

实施例的不同之处,各个实施例之间可以根据需要进行组合,且相同相似部分互相参见即可。

[0334] 对所公开的实施例的上述说明,使本领域专业技术人员能够实现或使用本申请。对这些实施例的多种修改对本领域的专业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申请的精神或范围的情况下,在其它实施例中实现。因此,本申请将不会被限制于本文所示的这些实施例,而是要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。

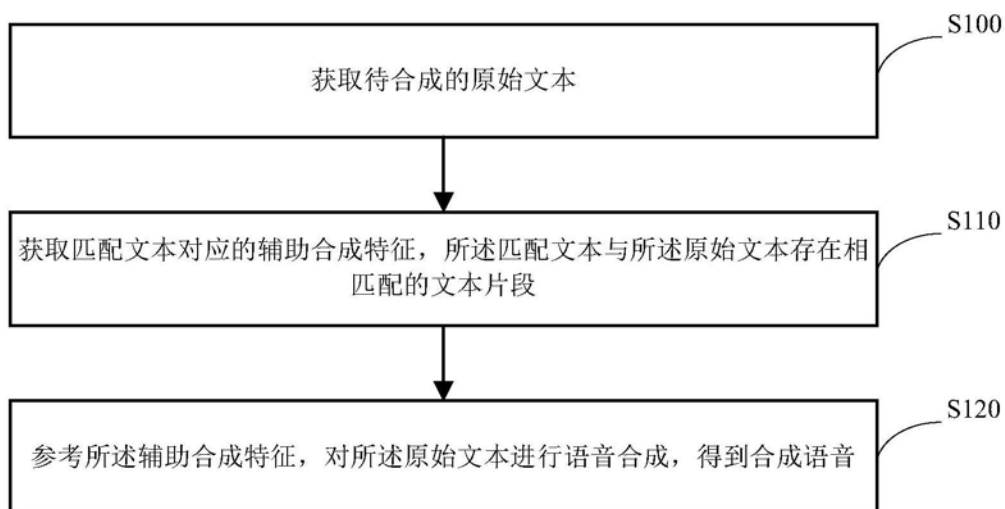


图1

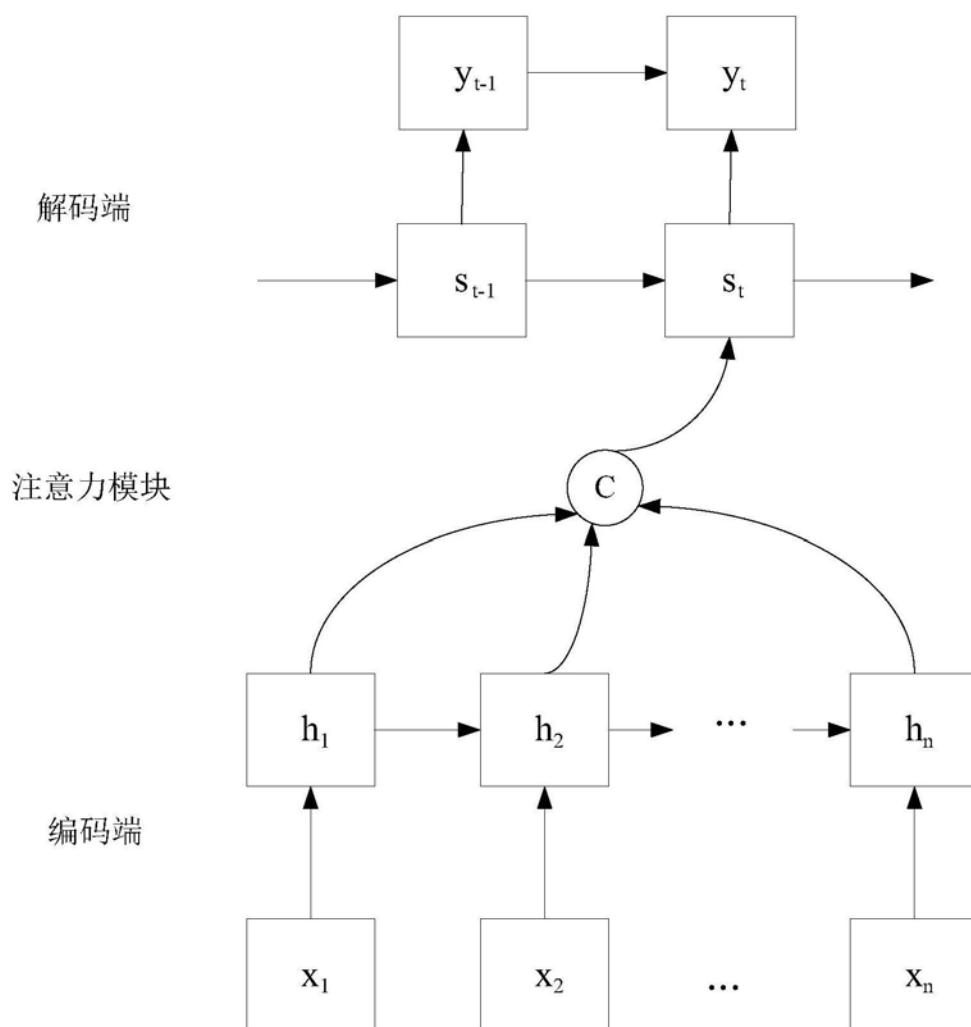


图2

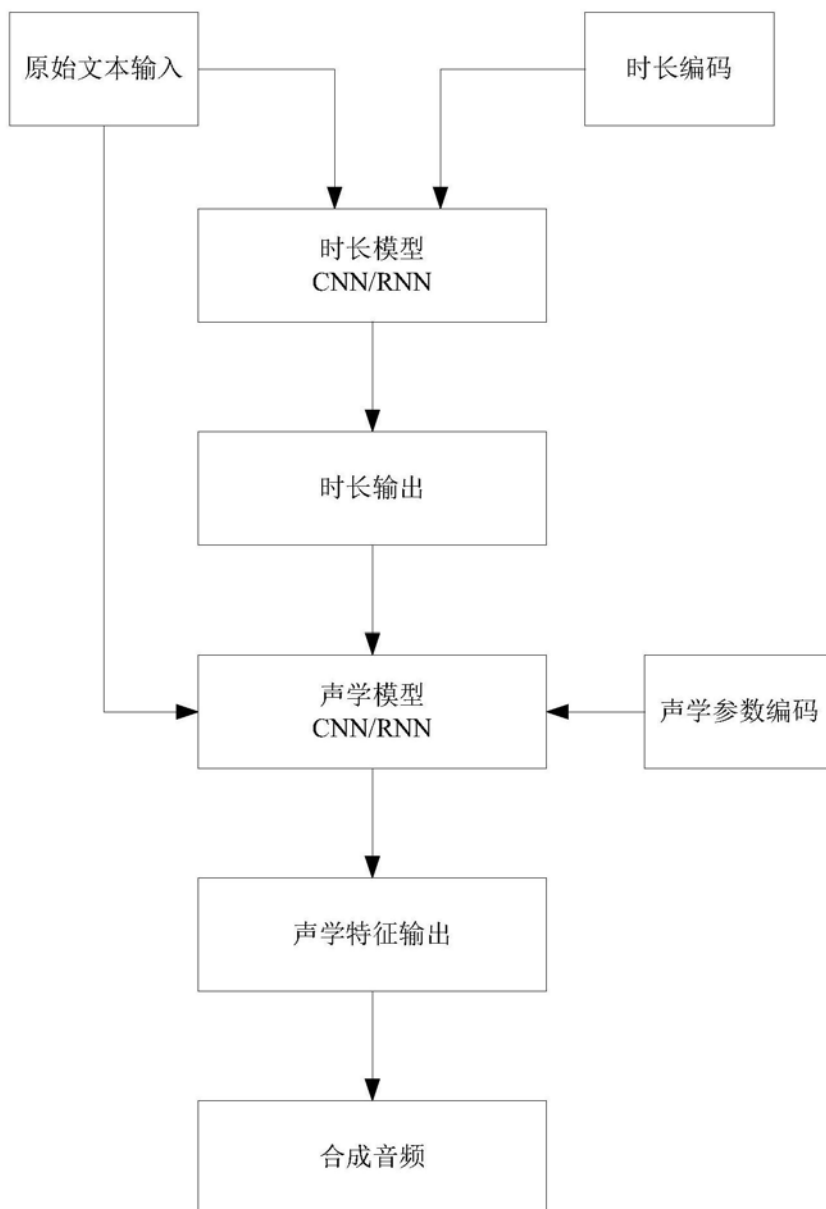


图3

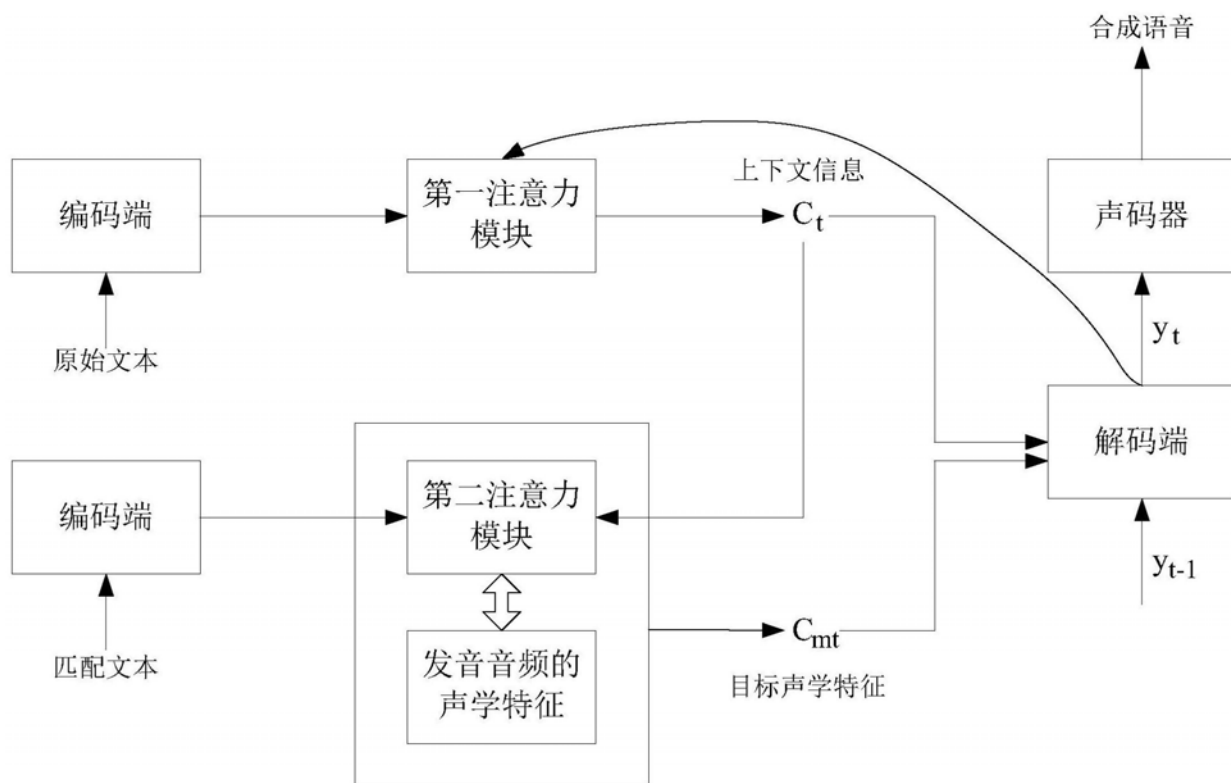


图4

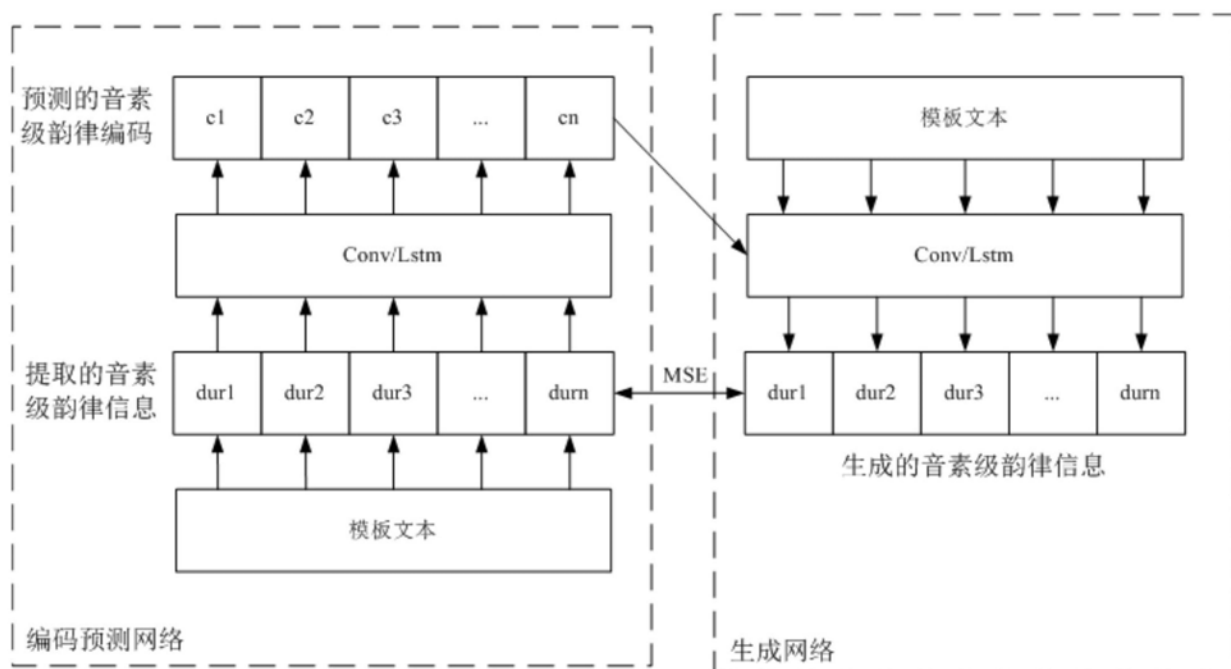


图5

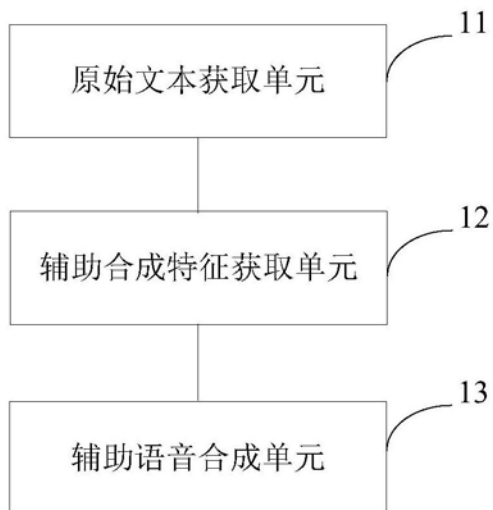


图6

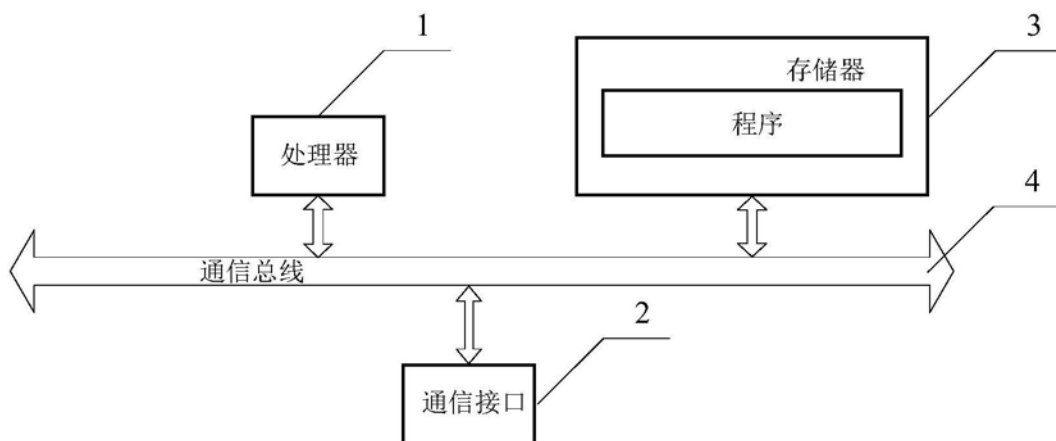


图7