(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114203150 A
(43)申请公布日 2022.03.18

(21)申请号 202111420017.9

(22)申请日 2021.11.26

(71)申请人 南京星云数字技术有限公司
地址 211800 江苏省南京市江北新区研创园团结路99号孵鹰大厦834室

(72)发明人 吴少铎 戴治波 王瑞 吴晨捷 丁进飞

(74)专利代理机构 北京市万慧达律师事务所 11111
代理人 刘艳丽

(51)Int.Cl.
G10L 13/027(2013.01)

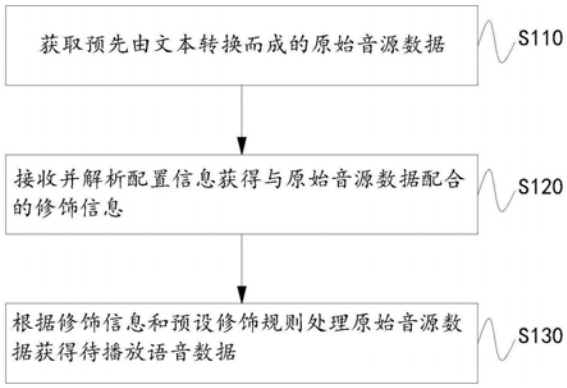
权利要求书2页 说明书9页 附图1页

(54)发明名称

语音数据处理方法及装置

(57)摘要

本发明公开了一种语音数据处理方法及装置，方法包括：获取预先由文本转换而成的原始音源数据；接收并解析配置信息获得与原始音源数据配合的修饰信息；根据修饰信息处理原始音源数据获得待播放语音数据；通过修饰信息调校处理文本转换而成的原始音源数据，对原始音源数据进行二次创作从而获得定制化的更具有情感色彩的带播放语音，使得听者聆听时得到更拟人化更具娱乐性的朗读体验。



CN 114203150 A

1.一种语音数据处理方法,其特征在于,所述方法包括:

获取预先由文本转换而成的原始音源数据;

接收并解析配置信息获得与所述原始音源数据配合的修饰信息;

根据所述修饰信息和预设修饰规则处理所述原始音源数据获得待播放语音数据。

2.根据权利要求1所述的语音数据处理方法,其特征在于,所述修饰信息包括与所述原始音源数据配合的控制指令,所述控制指令至少包括:停顿指令、重音指令、语速调节指令、句调调节指令以及添加口癖指令中的至少一种。

3.根据权利要求2所述的语音数据处理方法,其特征在于,所述控制指令包括所述添加口癖指令和/或所述句调调节指令时,所述修饰信息还包括与所述控制指令匹配的辅助音源数据。

4.根据权利要求3所述的语音数据处理方法,其特征在于,所述控制指令包括停顿指令时,所述根据所述修饰信息和预设修饰规则处理所述原始音源数据获得待播放语音数据包括:

基于所述停顿指令和所述预设修饰规则对所述原始音源数据中对应位置进行停顿标识以获得待播放语音数据。

5.根据权利要求3所述的语音数据处理方法,其特征在于,所述控制指令包括重音指令时,所述根据所述修饰信息和预设修饰规则处理所述原始音源数据获得待播放语音数据包括:

基于所述重音指令和所述预设修饰规则调整所述原始音源数据对应位置的振幅,或基于所述重音指令和所述预设修饰规则调整所述原始音源数据对应位置的振幅和频率以获得待播放语音数据。

6.根据权利要求3所述的语音数据处理方法,其特征在于,所述控制指令包括语速调节指令时,所述根据所述修饰信息和预设修饰规则处理所述原始音源数据获得待播放语音数据包括:

基于所述语速调节指令和所述预设修饰规则调整所述原始音源数据对应位置的播放帧速以获得待播放语音数据。

7.根据权利要求3所述的语音数据处理方法,其特征在于,所述控制指令包括句调调节指令时,所述根据所述修饰信息和预设修饰规则处理所述原始音源数据获得待播放语音数据包括:

获取所述句调调节指令关联的辅助音源数据;

基于所述句调调节指令截取所述辅助音源数据中的音频帧并以截取的所述辅助音源数据中的音频帧替换所述原始音源数据中的对应音频帧以获得待播放语音数据。

8.根据权利要求3所述的语音数据处理方法,其特征在于,所述控制指令包括添加口癖指令时,所述根据所述修饰信息和预设修饰规则处理所述原始音源数据获得待播放语音数据包括:

基于所述添加口癖指令以所述原始音源数据为模板进行二次合成以获得待播放语音数据。

9.根据权利要求8所述的语音数据处理方法,其特征在于,所述基于所述添加口癖指令以所述原始音源数据为模板进行二次合成以获得待播放语音数据包括:

基于所述添加口癖指令获取所述辅助音源数据中对应的第一音频帧；

基于所述添加口癖指令将第一音频帧插入所述原始音源数据中指定位置并对所述原始音源数据中位于所述第一音频帧之后的音频帧进行位移处理以消除插入所述第一音频帧造成的时间差以获得待播放语音数据；或：

基于所述添加口癖指令获取所述辅助音源数据中对应的第一音频帧并定位所述原始音源数据中对应的第二音频帧；

基于所述添加口癖指令以第一音频帧替换所述第二音频帧，并对所述原始音源数据中位于所述第一音频帧之后的音频帧进行位移处理以消除插入所述第一音频帧造成的时间差以获得待播放语音数据。

10.一种语音数据处理装置，其特征在于，所述装置包括：

获取模块，用于获取预先由文本转换而成的原始音源数据；

接收解析模块，用于接收并解析配置信息获得与所述原始音源数据配合的修饰信息；

处理模块，用于根据所述修饰信息和预设修饰规则处理所述原始音源数据获得待播放语音数据。

语音数据处理方法及装置

技术领域

[0001] 本发明涉及计算机数据处理领域，具体涉及一种语音数据处理方法及装置。

背景技术

[0002] 语言是一门伟大的艺术，人类的语言与其文明历程关系密切，反映着时代下的文化特色社会形态，在交流过程中有着丰富的感情色彩。文字作为其载体，在表达过程中往往需要采用诸多技巧结合语境，才能准确表达出其中的感情色彩，而由文字转换成语音时，由于进行转换的机器是不能理解感情色彩的，因此现行技术下的人工智能技术能够将文本转换生成语音，但难以模拟出其中的感情色彩，语音的感情色彩主要表现为播放时的朗读技巧，包括重音、停顿、语气、语速、语调等。

[0003] 为解决上述问题，目前通常采用机器深度学习的方法来进行语音数据处理，但只能局限于部分领域的文本转换，并且转换成的语音播放效果距离人工朗读文本的效果差距仍然遥远。

发明内容

[0004] 本发明目的是：提供一种能丰富文本转换成的语音中的感情色彩，从而能接近人工朗读文本效果的语音数据处理方法及装置。

[0005] 本发明的技术方案是：第一方面，本发明提供一种语音数据处理方法，所述方法包括：

[0006] 获取预先由文本转换而成的原始音源数据；

[0007] 接收并解析配置信息获得与所述原始音源数据配合的修饰信息；

[0008] 根据所述修饰信息和预设修饰规则处理所述原始音源数据获得待播放语音数据。

[0009] 在一种较佳的实施方式中，所述修饰信息包括与所述原始音源数据配合的控制指令，所述控制指令至少包括：停顿指令、重音指令、语速调节指令、句调调节指令以及添加口癖指令中的至少一种。

[0010] 在一种较佳的实施方式中，所述控制指令包括所述添加口癖指令和/或所述句调调节指令时，所述修饰信息还包括与所述控制指令匹配的辅助音源数据。

[0011] 在一种较佳的实施方式中，所述控制指令包括停顿指令时，所述根据所述修饰信息和预设修饰规则处理所述原始音源数据获得待播放语音数据包括：

[0012] 基于所述停顿指令和所述预设修饰规则对所述原始音源数据中对应位置进行停顿标识以获得待播放语音数据。

[0013] 在一种较佳的实施方式中，所述控制指令包括重音指令时，所述根据所述修饰信息和预设修饰规则处理所述原始音源数据获得待播放语音数据包括：

[0014] 基于所述重音指令和所述预设修饰规则调整所述原始音源数据对应位置的振幅，或基于所述重音指令和所述预设修饰规则调整所述原始音源数据对应位置的振幅和频率以获得待播放语音数据。

[0015] 在一种较佳的实施方式中，所述控制指令包括语速调节指令时，所述根据所述修饰信息和预设修饰规则处理所述原始音源数据获得待播放语音数据包括：

[0016] 基于所述语速调节指令和所述预设修饰规则调整所述原始音源数据对应位置的播放帧速以获得待播放语音数据。

[0017] 在一种较佳的实施方式中，所述控制指令包括句调调节指令时，所述根据所述修饰信息和预设修饰规则处理所述原始音源数据获得待播放语音数据包括：

[0018] 获取所述句调调节指令关联的辅助音源数据，

[0019] 基于所述句调调节指令截取所述辅助音源数据中的音频帧并以截取的所述辅助音源数据中的音频帧替换所述原始音源数据中的对应音频帧以获得待播放语音数据。

[0020] 在一种较佳的实施方式中，所述控制指令包括添加口癖指令时，所述根据所述修饰信息和预设修饰规则处理所述原始音源数据获得待播放语音数据包括：

[0021] 基于所述添加口癖指令以所述原始音源数据为模板进行二次合成以获得待播放语音数据。

[0022] 在一种较佳的实施方式中，所述基于所述添加口癖指令以所述原始音源数据为模板进行二次合成以获得待播放语音数据包括：

[0023] 基于所述添加口癖指令获取所述辅助音源数据中对应的第一音频帧；

[0024] 基于所述添加口癖指令将第一音频帧插入所述原始音源数据中指定位置并对所述原始音源数据中位于所述第一音频帧之后的音频帧进行位移处理以消除插入所述第一音频帧造成的时间差以获得待播放语音数据；或：

[0025] 基于所述添加口癖指令获取所述辅助音源数据中对应的第一音频帧并定位所述原始音源数据中对应的第二音频帧；

[0026] 基于所述添加口癖指令以第一音频帧替换所述第二音频帧，并对所述原始音源数据中位于所述第一音频帧之后的音频帧进行位移处理以消除插入所述第一音频帧造成的时间差以获得待播放语音数据。

[0027] 第二方面，本发明提供一种语音数据处理装置，所述装置包括：

[0028] 获取模块，用于获取预先由文本转换而成的原始音源数据；

[0029] 接收解析模块，用于并解析配置信息获得接收与所述原始音源数据配合的修饰信息；

[0030] 处理模块，用于根据所述修饰信息和预设修饰规则处理所述原始音源数据获得待播放语音数据。

[0031] 与现有技术相比，本发明的优点是：提供一种语音数据处理方法及装置，方法包括：获取预先由文本转换而成的原始音源数据；接收并解析配置信息获得与原始音源数据配合的修饰信息；根据修饰信息处理原始音源数据获得待播放语音数据；通过修饰信息调校处理文本转换而成的原始音源数据，对原始音源数据进行二次创作从而获得定制化的更具有情感色彩的带播放语音，使得听者聆听时得到更拟人化更具娱乐性的朗读体验。

附图说明

[0032] 为了更清楚地说明本申请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申请的一些实施例，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。

[0033] 图1为本发明实施例一所提供的语音数据处理方法的流程图；

[0034] 图2为本发明实施例二所提供的语音数据处理装置的结构图。

具体实施方式

[0035] 为使本申请的目的、技术方案和优点更加清楚，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。

[0036] 如背景技术中所述，目前的文本被转换成语音后，语音中的每个字句都是机械的读音，其播放时缺乏人工朗读所具有的感情色彩，亦或者由于程序难以识别文字中的反语等语境，程序理解的感情色彩出现偏差，听者的体验较差。

[0037] 为解决上述问题，本发明提供一种语音数据处理方法及装置，播放系统获取文本转换而成的原始音源数据和预先配置的与原始音源数据配合的修饰信息，采用修饰信息对原始音源数据进行信息补充，通过抽象建立控制模型，尽可能补充朗读控制因子，丰富原始音源数据，从而使得TTS(Text To Speech的缩写，即"从文本到语音"，是人机对话的一部分，让机器能够说话)能够表达出情感元素。

[0038] 下面将结合附图和各个实施例，对本申请的方案进行详细介绍。

[0039] 实施例一：本实施例提供一种语音数据处理方法，参照图1所示，该方法包括：

[0040] S110、获取预先由文本转换而成的原始音源数据。

[0041] 具体的，在获取文本后识别文本中的所有文本串，使用包含韵律匹配模板的树结构识别每个文本串的韵律信息，树结构基于重音模式使得树结构的每个节点提供与文本串的音节部分管理的重音级，利用韵律信息将文本串转变成可听的语音。韵律指的是说话的节奏和声调。或者，使用文本转语音TTS转换器将接收的文本数据转换成音频语音信号。

[0042] 当然，还可以使用其他方法来将文本转换成语音以获得原始音源数据，本实施例对此不作限制。

[0043] S120、接收并解析配置信息获得与原始音源数据配合的修饰信息。

[0044] 具体的，接收预先配置的与原始音源数据配合以丰富原始音源数据中感情色彩的修饰信息。由于语音中的感情色彩主要通过重音、停顿、语气、语速、语调等朗读技巧体现，因而优选的，修饰信息包括与原始音源数据配合的控制指令，控制指令至少包括：停顿指令、重音指令、语速调节指令、句调调节指令以及添加口癖指令中的至少一种。即：控制指令可以是停顿指令，或者是重音指令，或者是语速调节指令，或者是句调调节指令，或者是添加口癖指令单独一种指令，控制指令也可以是上述指令中任意两种指令的结合，也可以是上述指令中任意三种指令的结合，也可以是上述指令中任意四种的结合，还可以是上述指令中五种指令的结合。

[0045] 更具体的，停顿指令中还包括停顿节点位置信息和停顿时长信息，重音指令中还包括重音节点位置信息和重音级别信息，语速调节指令还包括语速调节节点位置信息和语速速度信息，句调调节指令还包括句调调节节点位置信息和目标句调信息，添加口癖指令

还包括添加节点位置信息和口癖内容信息。

[0046] 在一种较佳的实施方式中，控制指令包括添加口癖指令和/或句调调节指令时，修饰信息还包括与控制指令匹配的辅助音源数据。

[0047] 具体的，在控制指令中包含有句调调节指令，或者包含有添加口癖指令，或者包含有句调调节指令与添加口癖指令时，由于对原始音源数据进行句调调节处理时需要在插入不同句调的音频帧，在进行添加口癖处理时需要插入口癖内容音频帧，因此在控制指令中包含句调调节指令与添加口癖指令中至少一种时，修饰信息中需要配置与控制指令匹配的辅助音源数据，即在控制指令中包含句调调节指令时，修饰信息中配置与句调调节指令匹配的与原始音源不同句调的句调音源，在控制指令中包括添加口癖指令时，修饰信息中配置与添加口癖指令匹配的口癖音源，从而便于后续对原始音源数据作相应处理。

[0048] S130、根据修饰信息和预设修饰规则处理原始音源数据获得待播放语音数据。

[0049] 具体的，播放系统解析接收到的原始音源数据获得原始音频数据，再根据对应的控制指令和预设修饰规则对原始音频数据中对应的数据帧也就是音频帧进行修饰，修饰后的音频数据输出给最终播放器进行播放。关于播放系统采用解码器解析原始音源数据的解码过程以及最终播放器播放修饰后的音频数据的播放过程为解码器与播放器现有功能，本实施例在此不作赘述。

[0050] 在一种实施方式中，控制指令包括停顿指令时，根据修饰信息和预设修饰规则处理原始音源数据获得待播放语音数据包括：

[0051] S130-1、基于停顿指令和预设修饰规则对所述原始音源数据中对应位置进行停顿标识以获得待播放语音数据。

[0052] 停顿的目的有模拟生理换气需要、语法句法结构、表强调，表达感情或给听者一个领略和思考、理解和接受的余地，帮助加深印象或给出反馈间隙，比如赞叹鼓掌，防备后续朗读内容丢失等。停顿控制帧插入到指定的时间位置，以及停顿时长。即插入静音，将原有音波右相移。

[0053] 播放系统接收的停顿指令后，根据预设修饰规则以及停顿指令中的停顿节点位置信息和停顿时长信息对所述原始音源数据中对应位置进行停顿标识。示例性的，停顿指令包括第一停顿节点位置3s、第一停顿时长1s、第二停顿节点位置1min20s，第二停顿时长2s，则播放系统接收到停顿指令后，根据上述停顿指令在原始音频数据播放时间轴3s处插入第一停顿控制帧，在原始音源数据原1min20s处插入第二停顿控制帧，播放器播放处理后的音频时播放到第一停顿控制帧时识别停顿标识停顿1s，播放到第二停顿控制帧时识别停顿标识停顿2s。

[0054] 在一种实施方式中，控制指令包括重音指令时，根据修饰信息和预设修饰规则处理原始音源数据获得待播放语音数据包括：

[0055] S130-2、基于重音指令和预设修饰规则调整原始音源数据对应位置的振幅，或基于重音指令调整原始音源数据对应位置的振幅和频率以获得待播放语音数据。

[0056] 具体的，对于文本关键字句进行重读，是为了让听者了解表达的关注点，在不结合语境上下文的前提下，光是直述很难表达原作者的意图，这是目前很难做到的——因为受限于资源条件、时间要求，目前往往将文本内容进行分词处理，而不去对全篇论述中心思想进行“思考”。

[0057] 重音控制指定的到文字或者音素，进行音调补强。一般情况下增加音源对应位置的振幅即可，当然，特别情况下还需要补充频率。即增加该处的输出能量。

[0058] 播放系统在接收到重音指令后，根据重音指令中的重音节点位置信息和重音级别信息调节原始音源数据中指定位置音频帧的振幅或者调节原始音源数据中指定位置音频帧的振幅与频率以获得待播放语音数据。

[0059] 示例性的，播放系统接收到重音指令，重音指令中包括第一重音节点位置1min处和第二重音节点位置1min40s处，以及第一重音级别一级和第二重音级别二级，由于预设修饰规则中规定一级重音增加第一预设振幅，二级重音增加第二预设振幅并补充预设频率，则播放系统接收到重音指令后，根据上述重音指令和预设修饰规则，对原始音源数据的1min位置处的音频帧进行增加第一预设振幅处理，对原始音源数据的1min40s处的音频帧进行增加第二预设振幅并补充预设频率处理，从而在播放处理后的音频时在第一重音节点位置处获得一级重音效果，在第二重音节点位置处获得二级重音效果。

[0060] 在一种实施方式中，控制指令包括语速调节指令时，根据修饰信息和预设修饰规则处理原始音源数据获得待播放语音数据包括：

[0061] S130-3、基于语速调节指令和预设修饰规则调整原始音源数据对应位置的播放帧速以获得待播放语音数据。

[0062] 具体的，语速在朗读中用于表达说话人的感情，通过调节语速也就是控制音节的长短与松紧，表达热烈、欢快、兴奋、紧张的情绪时加快，表达平静、庄重、悲伤、沉重、追忆内容时放缓，表达一般叙述、说明、议论时中速。当然，听者也会根据接受度与习惯临时变更语速。

[0063] 由于语速控制一般需要指定文字的起始与终止位置，附带速度讯息，因此语速调节指令包括语速调节节点位置信息和语速速度信息。假设1.0是正常，0到1之间是慢速，1.0以上是倍速。

[0064] 示例性的，播放系统接收到语速调节指令，语速调节指令包括开始时间节点11min15s、结束时间节点11min20s，语速速度0.8倍，则对原始音频数据中对应的11min15s-11min20s的音频帧进行播放速度调节至0.8倍。

[0065] 句调是贯穿整个句干的高低升降，表示语气或态度，分为升、降、平、曲。用以表达讽刺、厌恶、反语、意外、笑语、颤音、泣诉等等。

[0066] 在一种实施方式中，控制指令包括句调调节指令时，根据修饰信息和预设修饰规则处理原始音源数据获得待播放语音数据包括：

[0067] S130-4A、获取句调调节指令关联的辅助音源数据。

[0068] 句调调节指令关联的辅助音源数据与原始音源数据由相同的文本转换而成，但句调调节指令关联的辅助音源数据中的句调与原始音源数据中的句调不同，句调调节指令关联的辅助音源数据的数量可以大于或等于2个，每个句调调节指令关联的辅助音源数据的句调均不相同，以便将原始音源数据修饰成不同的句调。示例性的，原始音源数据中采用的是平调句调，第一句调调节指令关联的辅助音源数据中均采用升调句调，第二句调调节指令关联的辅助音源数据中均采用降低句调，第三句调调节指令关联的辅助音源数据中均采用曲调句调。后续根据句调调节指令从第一、第二、第三句调调节指令关联的辅助音源数据中获取指定位置的音频帧对原始音源数据进行修饰。

[0069] S130-4B、基于句调调节指令和预设修饰规则截取句调调节指令关联的辅助音源数据中的音频帧并以句调调节指令关联的辅助音源数据中的音频帧替换原始音源数据中的对应音频帧以获得待播放语音数据。

[0070] 本实施例直接发送句调调节指令关联的辅助音源数据中对应的部分重新进行替换，而不直接使用合成算法，操作简单不易出错。替换时，判断时长位置删除原有的音频帧，再引入新的音频帧进行合成，同时后续的音频帧根据前后时间差值进行相位移。

[0071] 示例性的，播放系统接收到句调调节指令，句调调节指令包括第一句调调节起点11min15s、第一句调调节终点11min20s、第一目标句调1，第二句调调节起点11min40s、第二句调调节终点11min55s、第二目标句调2、第三句调调节起点12min40s、第三句调调节终点12min55s、第三目标句调3，由于修饰规则中规定句调1对应升调，句调2对应降调，句调3对应曲调，而辅助音源数据中第一句调调节指令关联的辅助音源数据中均采用升调句调，第二句调调节指令关联的辅助音源数据中均采用降低句调，第三句调调节指令关联的辅助音源数据中均采用曲调句调，则截取第一句调调节指令关联的辅助音源数据中播放时间轴11min15s-11min20s对应的音频帧并用从第一句调调节指令关联的辅助音源数据中截取的音频帧替换原始音源数据中播放时间轴11min15s-11min20s对应的音频帧，截取第二句调调节指令关联的辅助音源数据中播放时间轴11min40s-11min55s对应的音频帧并用从第二句调调节指令关联的辅助音源数据中截取的音频帧替换原始音源数据中播放时间轴11min40s-11min55s对应的音频帧，截取第三句调调节指令关联的辅助音源数据中播放时间轴12min40s-12min55s对应的音频帧并用从第三句调调节指令关联的辅助音源数据中截取的音频帧替换原始音源数据中播放时间轴12min40s-12min55s对应的音频帧，从而使得原始音源数据被替换修饰后获得的带播放音源播放时11min15s-11min20s句调为升调，11min40s-11min55s句调为降调，12min40s-12min55s为曲调。

[0072] 在一种实施方式中，控制指令包括添加口癖指令时，根据修饰信息和预设修饰规则处理原始音源数据获得待播放语音数据包括：

[0073] S130-5、基于添加口癖指令以原始音源数据为模板进行二次合成以获得待播放语音数据。

[0074] 具体的，口癖是展示个人独特风格的标志性产物，包括儿话音、对某个特定字的特殊发音声调或在句首添加口头禅、句尾添加句末助词等。

[0075] 优选的，S130-5具体包括：

[0076] S130-5A、基于添加口癖指令获取辅助音源数据中对应的第一音频帧。

[0077] 第一音频帧为与添加口癖指令中所包含的口癖内容信息对应的音频帧，具体的，当添加口癖指令中所包含的口癖内容信息为儿化音时，第一音频帧包括对应的内容为儿化音的音频帧；当添加口癖指令中所包含的口癖内容信息为特定字特殊发音时，第一音频帧包括对应的内容为该特定字特殊发音的音频帧；当添加口癖指令中所包含的口癖内容信息为口头禅时，第一音频帧包括对应的内容为口头禅的音频帧；当添加口癖指令中所包含的口癖内容信息为句尾添加句末助词时，第一音频帧包括对应的内容为句尾添加句末助词的音频帧。根据添加口癖指令中的口癖内容信息的具体内容去获取对应的第一音频帧。

[0078] S130-5B、基于添加口癖指令将第一音频帧插入原始音源数据中指定位置并对原始音源数据中位于第一音频帧之后的音频帧进行位移处理以消除插入所述第一音频帧造

成的时间差以获得待播放语音数据。

[0079] 具体的，当添加口癖指令中所包含的口癖内容信息为儿化音时，根据添加口癖指令中的添加节点位置信息确定原始音源数据中的指定插入位置，将获取的内容为对应的儿化音的第一音频帧插入指定插入位置，并少量删除词组末尾音素，后续音频右相位移。其中，根据添加口癖指令中的添加节点位置信息可以是预设特定词组，根据添加口癖指令中的添加节点位置信息确定原始音源数据中的指定插入位置包括在原始音源数据中查找定位预设特定词组位置，查找定位到原始音源数据中的预设特定词组后，在预设特定词组后插入内容为er(轻声)的第一音频帧，并少量删除词组末尾音素，后续音频右相位移len(儿话音)-offset(删除音素长度)。

[0080] 当添加口癖指令中所包含的口癖内容信息为口头禅时，第一音频帧包括对应的内容为口头禅的音频帧，添加口癖指令中所包含的添加节点位置信息为目标语句的开头，根据添加节点位置信息判断原始音源数据中目标语句的开头位置，在目标语句的开头位置插入内容为口头禅的第一音频帧，后续音频右相位移。示例性的，播放系统接的控制指令中包括添加口癖指令，添加口癖指令包括口癖内容信息——“阿弥陀佛”和目标语句的开头——“每句语句的开头”，则播放系统从辅助音源数据中获取对应的内容为“阿弥陀佛”的第一音频帧，并遍历原始音源数据定位每句语句的起始位置，在每句语句的起始位置插入从辅助音源数据中获得的内容为“阿弥陀佛”的第一音频帧。当然，目标语句还可以是第二句、第八句等其他语句位置。

[0081] 当添加口癖指令中所包含的口癖内容信息为句尾添加句末助词时，第一音频帧包括对应的内容为句尾添加句末助词的音频帧，添加口癖指令中所包含的添加节点位置信息为目标语句句尾，根据添加节点位置信息判断原始音源数据中每句话的句尾位置，在每句话的句尾位置插入内容为句尾添加句末助词的第一音频帧，后续音频右相位移。示例性的，播放系统接的控制指令中包括添加口癖指令，添加口癖指令包括第一句尾添加句末助词——“喵呜”、第一添加节点位置信息——“第一句句尾、第三句句尾”、第二句尾添加句末助词“的说”、第二添加节点位置信息——“第二句句尾、第四句句尾”，播放系统从辅助音源数据中获取对应的内容为“喵呜”的第一音频帧A和对应的内容为“的说”的第一音频帧B，并遍历原始音源数据定位第一句句尾和第三句尾，在第一句句尾和第三句句尾处插入从辅助音源数据中获得的内容为“喵呜”的第一音频帧A，定位第二句句尾和第四句句尾，在第二句句尾和第四句句尾处插入从辅助音源数据中获得的内容为“的说”的第一音频帧B，后续音频右相位移。

[0082] 而对于口癖内容信息为特定字特殊发音的添加口癖指令，与句调处理类似的，采用合成方法处理音源容易出错，因此对于口癖内容信息为特定字特殊发音的添加口癖指令本步骤包括：

[0083] S130-5a、基于添加口癖指令获取辅助音源数据中对应的第一音频帧并定位原始音源数据中对应的第二音频帧。

[0084] 具体的，当播放系统接收的添加口癖指令中所包含的口癖内容信息为特定字特殊发音时，从辅助音源数据中获取的第一音频帧包括对应的内容为该特定字特殊发音的音频帧。

[0085] S130-5b、基于添加口癖指令以第一音频帧替换第二音频帧，并对原始音源数据中

位于所述第一音频帧之后的音频帧进行位移处理以消除插入第一音频帧造成的时间差以获得待播放语音数据。

[0086] 具体的，整句遍历原始音源数据查找定位特定字所有所在位置，或者在修饰信息中包含预先配置的特定字所在位置信息，根据修饰信息中所包含的预先配置的特定字所在位置信息获得指定帧位置，原始音源数据中特定字所在的音频帧即为第二音频帧，将特定字所在位置的原始音源数据中的音频帧全部替换为内容为特定字特殊发音的第一音频帧，后续音频相位移替换前后时间差值。上述将所有特定字所在的第二音频帧全部替换为第一音频帧是在添加口癖指令中添加节点位置信息为“每个特定字位置”的情况下，事实上，本实施例中添加节点位置信息也可以是其他的指定特定字位置，对于特定字的具体内容本实施例也不作限定。在第一个示例中，添加口癖指令包括添加口癖内容——特定字“我”，添加节点位置信息为“第二个”，则以获取的内容为“我”特殊升调发音的第一音频帧替换原始音源数据中第二个特定字“我”所在的音频帧，并对原始音源数据中位于替换后的内容为“我”特殊升调发音的音频帧之后的音频帧相位移替换前后时间差值。

[0087] 播放系统在解码播放过程中进行动态调校，即为二次合成。

[0088] 本实施例所提供的语音数据处理方法及装置，方法包括：获取预先由文本转换而成的原始音源数据；接收与原始音源数据配合的修饰信息；根据修饰信息处理原始音源数据获得待播放语音数据；通过修饰信息调校处理文本转换而成的原始音源数据，对原始音源数据进行二次创作从而获得定制化的更具有情感色彩的带播放语音，使得听者聆听时得到更拟人化更具娱乐性的朗读体验。

[0089] 进一步的，控制指令包括停顿指令、重音指令、语速调节指令、句调调节指令以及添加口癖指令中的至少一种，根据控制指令对原始音源数据进行对应的停顿和/或重音和/或语速调节和/或句调调节和/或添加口癖处理，从而使得处理后获得的带播放音源播放时能够通过停顿、重音、语速、句调、口癖的朗读技巧体现出感情色彩，播放时的语音更生动人性化，使听众获得更好的听觉体验。

[0090] 实施例二：本实施例提供一种语音数据处理装置，该装置包括：

[0091] 获取模块210，用于获取预先由文本转换而成的原始音源数据；

[0092] 接收解析模块220，用于接收并解析配置信息获得与原始音源数据配合的修饰信息；

[0093] 处理模块230，用于根据修饰信息和预设修饰规则处理原始音源数据获得待播放语音数据。

[0094] 在一种较佳的实施方式中，修饰信息包括与原始音源数据配合的控制指令，控制指令至少包括：停顿指令、重音指令、语速调节指令、句调调节指令以及添加口癖指令中的至少一种。

[0095] 控制指令包括停顿指令和/或添加口癖指令和/或句调调节指令时，修饰信息还包括与控制指令匹配的辅助音源数据。

[0096] 控制指令包括停顿指令时，辅助音源数据包括静音音频帧，处理模块230具体用于：

[0097] 基于停顿指令和预设修饰规则将静音音频帧按指定位置和指定时长插入原始音源数据中。

[0098] 控制指令包括重音指令时，处理模块230具体用于：

[0099] 基于重音指令和预设修饰规则调整原始音源数据对应位置的振幅，或基于所述重音指令和所述预设修饰规则调整所述原始音源数据对应位置的振幅和频率。

[0100] 控制指令包括语速调节指令时，处理模块230具体用于：

[0101] 基于语速调节指令和预设修饰规则调整原始音源数据对应位置的播放帧速。

[0102] 控制指令包括句调调节指令时，处理模块230包括：

[0103] 第一获取单元231，用于获取原始音源数据对应的句调调节指令关联的辅助音源数据，

[0104] 截取替换单元232，用于基于句调调节指令截取句调调节指令关联的辅助音源数据中的音频帧并以截取的句调调节指令关联的辅助音源数据中的音频帧替换原始音源数据中的对应音频帧。

[0105] 控制指令包括添加口癖指令时，处理模块230具体用于：

[0106] 基于添加口癖指令以原始音源数据为模板进行二次合成。

[0107] 更优选的，处理模块230具体用于基于添加口癖指令以原始音源数据为模板进行二次合成时包括：

[0108] 第二获取单元233，用于基于添加口癖指令获取辅助音源数据中对应的第一音频帧；

[0109] 插入处理单元234，用于基于添加口癖指令将第一音频帧插入原始音源数据中指定位置并对原始音源数据中位于第一音频帧之后的音频帧进行位移处理以消除插入第一音频帧造成的时间差；或者包括：

[0110] 获取定位单元235，用于基于添加口癖指令获取辅助音源数据中对应的第一音频帧并定位原始音源数据中对应的第二音频帧；

[0111] 替换处理单元236，用于基于添加口癖指令以第一音频帧替换第二音频帧，并对原始音源数据中位于第一音频帧之后的音频帧进行位移处理以消除插入第一音频帧造成的时间差。

[0112] 本实施例提供的语音数据处理装置用于实现实施例一中所提供的语音数据处理方法，其有益效果与实施例一中所提供的语音数据处理方法的有益效果相同，在此不做赘述。

[0113] 需要说明的是：上述实施例提供的语音数据处理装置在执行语音数据处理方法时，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将装置的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的语音数据处理装置与语音数据处理方法实施例属于同一构思，其具体实现过程详见方法实施例，这里不再赘述。

[0114] 另外还需要说明的是：本发明中术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。

[0115] 当然上述实施例只为说明本发明的技术构思及特点，其目的在于让熟悉此项技术的人能够了解本发明的内容并据以实施，并不能以此限制本发明的保护范围。凡根据本发明主要技术方案的精神实质所做的修饰，都应涵盖在本发明的保护范围之内。

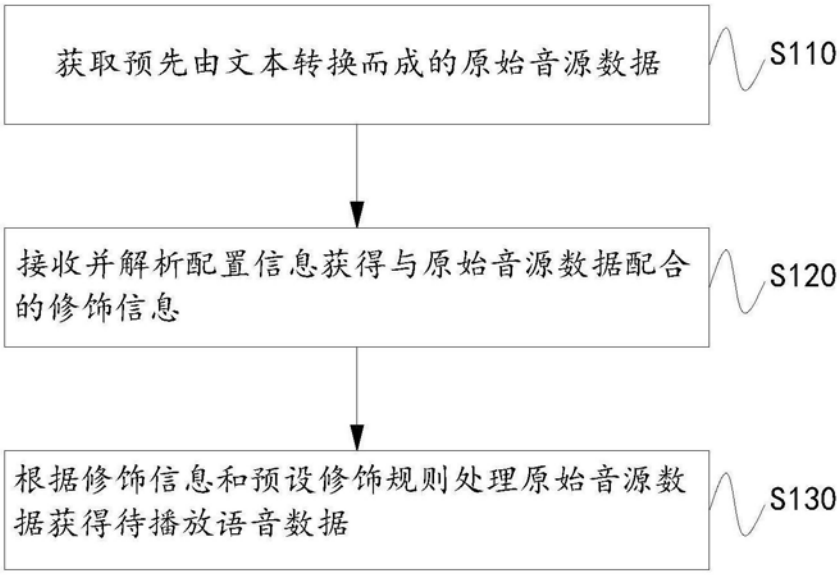

获取预先由文本转换而成的原始音源数据
S110
接收并解析配置信息获得与原始音源数据配合的修饰信息
S120
根据修饰信息和预设修饰规则处理原始音源数据获得待播放语音数据
S130

图1


语音数据处理装置
获取模块210
接收解析模块220
处理模块230

图2