(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116711002 A
(43)申请公布日 2023.09.05

(21)申请号 202180091481.X
(22)申请日 2021.12.08
(30)优先权数据
17/154,372 2021.01.21 US
(85)PCT国际申请进入国家阶段日
2023.07.20
(86)PCT国际申请的申请数据
PCT/US2021/072800 2021.12.08
(87)PCT国际申请的公布数据
WO2022/159256 EN 2022.07.28
(71)申请人 高通股份有限公司
地址 美国加利福尼亚州
(72)发明人 K·白 S·文 S·张
V·蒙塔泽里 L-H·金 E·维瑟
(74)专利代理机构 北京市柳沈律师事务所 11105
专利代理师 安之斐
(51)Int.Cl.
G10L 13/033(2006.01)

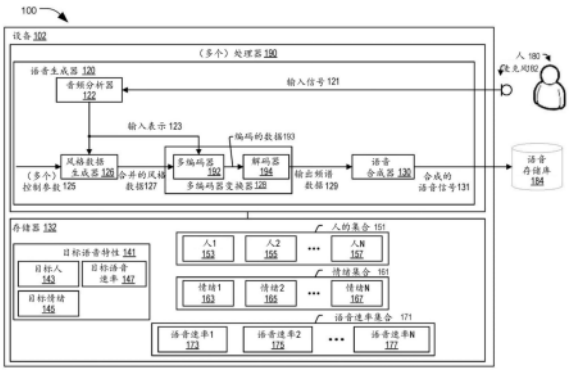
权利要求书4页 说明书23页 附图15页

(54)发明名称

合成语音生成

(57)摘要

一种用于语音生成的设备包括被配置为接收指示目标语音特性的一个或多个控制参数的一个或多个处理器。一个或多个处理器还被配置为使用多编码器基于一个或多个控制参数处理语音的输入表示,以生成与表示基于目标语音特性的语音版本的音频信号对应的编码的数据。



CN 116711002 A

1.一种用于语音生成的设备,包括:

一个或多个处理器被配置为:

接收指示目标语音特性的一个或多个控制参数;以及

使用多编码器,基于所述一个或多个控制参数处理语音的输入表示,以生成与表示基于目标语音特性的语音版本的音频信号对应的编码的数据。

2.根据权利要求1所述的设备,其中,所述控制参数指示要使用其语音特性的目标人、目标情绪、目标语音速率或其组合。

3.根据权利要求1所述的设备,其中,所述一个或多个处理器还被配置为基于一个或多个控制参数生成合并的风格数据,并且其中,所述合并的风格数据在输入表示的处理期间由多编码器使用。

4.根据权利要求1所述的设备,其中,所述多编码器包括:

第一编码器,被配置为独立于一个或多个控制参数对输入表示进行编码以生成第一编码的数据;以及

一个或多个第二编码器,被配置为基于一个或多个控制参数对输入表示进行编码以生成第二编码的数据,

其中,所述编码的数据包括第一编码的数据和第二编码的数据。

5.根据权利要求4所述的设备,其中,所述一个或多个处理器还被配置为:

在语音特性编码器处,基于一个或多个控制参数中的至少一个来处理输入表示以生成编码的输入语音表示;

在编码器前置网络处,至少部分地基于编码的输入语音表示生成合并的风格数据;

将输入表示提供给第一编码器以生成第一编码的数据;以及

将合并的风格数据提供给一个或多个第二编码器以生成第二编码的数据。

6.根据权利要求4所述的设备,还包括包含多编码器和解码器的多编码器变换器,其中,第一编码器包括第一注意力网络,其中,所述一个或多个第二编码器中的每一个包括第二注意力网络,并且其中,所述解码器包括解码器注意力网络。

7.根据权利要求6所述的设备,其中:

第一编码器包括:

第一层,包括第一注意力网络,其中,第一注意力网络对应于第一多头注意力网络;以及

第二层,包括第一神经网络,并且

一个或多个第二编码器中的每一个包括:

第一层,包括第二注意力网络,其中,第二注意力网络对应于第二多头注意力网络;以及

第二层,包括第二神经网络。

8.根据权利要求4所述的设备,还包括:

解码器,耦合到多编码器,所述解码器包括被配置为基于第一编码的数据和第二编码的数据生成输出频谱数据的解码器网络;以及

语音合成器,被配置为基于输出频谱数据生成表示基于目标语音特性的语音版本的音频信号。

9.根据权利要求8所述的设备，其中，所述解码器网络包括解码器注意力网络，所述解码器注意力网络包括：

第一多头注意力网络，被配置为处理第一编码的数据；

一个或多个第二多头注意力网络，被配置为处理第二编码的数据；以及

组合器，被配置为组合第一多头注意力网络和一个或多个第二多头注意力网络的输出。

10.根据权利要求9所述的设备，其中，所述解码器还包括：

掩码的多头注意力网络，耦合到解码器注意力网络的输入；以及

解码器神经网络，耦合到解码器注意力网络的输出。

11.根据权利要求1所述的设备，其中，所述一个或处理器还被配置为：

从所述音频信号生成一个或多个估计的控制参数；以及

基于一个或多个控制参数与一个或多个估计的控制参数的比较，训练多编码器、一个或多个语音特性编码器、编码器前置网络、解码器网络，或其组合的一个或多个神经网络权重。

12.根据权利要求1所述的设备，其中，所述一个或多个处理器还被配置为：

接收输入语音信号；以及

基于输入语音信号生成输入表示。

13.根据权利要求1所述的设备，其中，所述一个或多个处理器还被配置为接收输入表示。

14.根据权利要求1所述的设备，其中，所述输入表示包括文本、梅尔标度频谱图、基频(F0)特征或其组合。

15.一种语音生成方法，包括：

在设备处接收指示目标语音特性的一个或多个控制参数；以及

使用多编码器基于一个或多个控制参数处理语音的输入表示，以生成与表示基于所述目标语音特性的语音版本的音频信号对应的编码的数据。

16.根据权利要求15所述的方法，其中，所述控制参数指示要使用其语音特性的目标人、目标情绪、目标语音速率或其组合。

17.根据权利要求15所述的方法，还包括在所述设备处基于一个或多个控制参数生成合并的风格数据，其中，所述合并的风格数据在输入表示的处理期间由所述多编码器使用。

18.根据权利要求15所述的方法，还包括：

在多编码器的第一编码器处独立于一个或多个控制参数对输入表示进行编码以生成第一编码的数据；以及

在多编码器的一个或多个第二编码器处，基于一个或多个控制参数对所述输入表示进行编码以生成第二编码的数据，

其中，所述编码的数据包括第一编码的数据和第二编码的数据。

19.根据权利要求18所述的方法，还包括：

在语音特性编码器处，基于一个或多个控制参数中的至少一个来处理所述输入表示以生成编码的输入语音表示；

在编码器前置网络处至少部分地基于编码的输入语音表示来生成合并的风格数据；

将输入表示提供给第一编码器以生成第一编码的数据;以及

将合并的风格数据提供给一个或多个第二编码器以生成所述第二编码的数据。

20.根据权利要求18所述的方法,其中,所述多编码器和解码器被包括在多编码器变换器中,其中,第一编码器包括第一注意力网络,其中,一个或多个第二编码器中的每一个包括第二注意力网络,并且其中,所述解码器包括解码器注意力网络。

21.根据权利要求20所述的方法,其中:

第一编码器包括:

第一层,包括所述第一注意力网络,其中,第一注意力网络对应于第一多头注意力网络;以及

第二层,包括第一神经网络,并且

一个或多个第二编码器中的每一个包括:

第一层,包括所述第二注意力网络,其中,所述第二注意力网络对应于第二多头注意力网络;以及

第二层,包括第二神经网络。

22.根据权利要求18所述的方法,还包括:

在解码器的解码器网络处,基于第一编码的数据和第二编码的数据生成输出频谱数据;以及

在语音合成器处并基于输出频谱数据生成表示基于目标语音特性的语音版本的音频信号。

23.根据权利要求22所述的方法,还包括:

在解码器注意力网络的第一多头注意力网络处处理第一编码的数据,其中,所述解码器网络包括解码器注意力网络;

在解码器注意力网络的一个或多个第二多头注意力网络处处理第二编码的数据;以及

在组合器处组合第一多头注意力网络和一个或多个第二多头注意力网络的输出。

24.根据权利要求15所述的方法,还包括:

在所述设备处从音频信号生成一个或多个估计的控制参数;以及

基于一个或多个控制参数与一个或多个估计的控制参数的比较,训练多编码器、一个或多个语音特性编码器、编码器前置网络、解码器网络,或其组合的一个或多个神经网络权重。

25.根据权利要求15所述的方法,还包括:

在所述设备处接收输入语音信号;以及

在所述设备处基于输入语音信号生成输入表示。

26.根据权利要求15所述的方法,还包括在所述设备处接收输入表示。

27.一种存储指令的非暂时性计算机可读存储介质,当所述指令由一个或多个处理器执行时,使所述一个或多个处理器:

接收指示目标语音特性的一个或多个控制参数;以及

使用多编码器,基于一个或多个控制参数处理语音的输入表示,以生成与表示基于目标语音特性的语音版本的音频信号对应的编码的数据。

28.根据权利要求27所述的非暂时性计算机可读存储介质,其中,所述输入表示包括文

本、梅尔标度频谱图、基频(F0)特征或其组合。

29.一种装置,包括:

用于接收指示目标语音特性的一个或多个控制参数的部件;以及

用于使用多编码器基于一个或多个控制参数处理语音的输入表示以生成与表示基于目标语音特性的语音版本的音频信号对应的编码的数据的部件。

30.根据权利要求29所述的装置,其中,用于接收的所述部件和用于处理的所述部件被集成到虚拟助手、家用电器、智能设备、物联网(IoT)设备、通信设备、耳机、载具、计算机、显示器设备、电视机、游戏机、音乐播放器、收音机、视频播放器、娱乐单元、个人媒体播放器、数字视频播放器、相机或导航设备中的至少一个中。

合成语音生成

[0001] 优先权要求

[0002] 本申请要求于2021年1月21日提交的共同拥有的美国非临时专利申请第17/154,372号的优先权的权益,其内容通过引用其整体明确并入本文。

技术领域

[0003] 本公开总体上涉及合成语音生成。

背景技术

[0004] 技术的进步导致了更小但更强大的计算设备。例如,目前存在各种便携式个人计算设备,包括诸如移动电话和智能电话的无线电话、平板电脑和膝上型计算机,它们体积小、重量轻并且易于用户携带。这些设备可以通过无线网络通信语音和数据分组。此外,许多这样的设备结合了附加功能,诸如数字静态相机、数字摄像机、数字录音机和音频文件播放器。此外,这样的设备可以处理可执行指令,包括可以用于访问互联网的软件应用,诸如网络浏览器应用。因此,这些设备可以包括重要的计算能力。

[0005] 这样的计算设备通常结合从一个或多个麦克风接收音频信号的功能。例如,音频信号可以表示由麦克风捕获的用户语音、由麦克风捕获的外部声音或其组合。这样的设备可能包括使用语音模型的应用。例如,语音激活的应用识别用户语音命令。基于语音的应用使用音频语音样本进行训练。使用更大、更不同的音频语音样本数据集进行训练导致更鲁棒的语音模型。然而,从人们获得大量不同的音频语音样本可能费时并且不方便。

发明内容

[0006] 根据本公开的一个实施方式,一种用于语音生成的设备包括被配置为接收指示目标语音特性的一个或多个控制参数的一个或多个处理器。一个或多个处理器还被配置为使用多编码器基于一个或多个控制参数处理语音的输入表示,以生成与表示基于目标语音特性的语音版本的音频信号对应的编码的数据。

[0007] 根据本公开的另一实施方式,一种语音生成方法包括在设备处接收指示目标语音特性的一个或多个控制参数。该方法还包括使用多编码器基于一个或多个控制参数处理语音的输入表示以生成与表示基于目标语音特性的语音版本的音频信号对应的编码的数据。

[0008] 根据本公开的另一实施方式,一种非暂时性计算机可读介质存储指令,当指令由一个或多个处理器执行时,使一个或多个处理器接收指示目标语音特性的一个或多个控制参数。当指令由一个或多个处理器执行时,还使一个或多个处理器使用多编码器基于一个或多个控制参数处理语音的输入表示,以生成与表示基于目标语音特性的语音版本的音频信号对应的编码的数据。

[0009] 根据本公开的另一实施方式,一种装置包括用于接收指示目标语音特性的一个或多个控制参数的部件。该装置还包括用于使用多编码器基于一个或多个控制参数处理语音的输入表示以生成与表示基于目标语音特性的语音版本的音频信号对应的编码的数据的

部件。

[0010] 在审阅包括以下部分的整个申请之后，本公开的其他方面、优点和特征将变得清晰：附图的简要描述、详细描述和权利要求。

附图说明

[0011] 图1是根据本公开的一些示例的可操作以生成合成语音的系统的特定说明性方面的框图。

[0012] 图2是根据本公开的一些示例的图1的系统的语音生成器的特定实施方式的示意图。

[0013] 图3是根据本公开的一些示例的图1的语音生成器的另一特定实施方式的示意图。

[0014] 图4A是根据本公开的一些示例的图1的语音生成器的另一特定实施方式的示意图。

[0015] 图4B是根据本公开的一些示例的图1的语音生成器的另一特定实施方式的示意图。

[0016] 图5是根据本公开的一些示例的图1的语音生成器的编码器的说明性方面的示意图。

[0017] 图6是根据本公开的一些示例的图1的系统的组件的操作的说明性方面的示意图。

[0018] 图7是根据本公开的一些示例的图1的系统的组件的操作的说明性方面的示意图。

[0019] 图8示出了根据本公开的一些示例的可操作以生成合成语音的集成电路的示例。

[0020] 图9是根据本公开的一些示例的可操作以生成合成语音的移动设备的示意图。

[0021] 图10是根据本公开的一些示例的可操作以生成合成语音的耳机的示意图。

[0022] 图11是根据本公开的一些示例的可操作以生成合成语音的可穿戴电子设备的示意图。

[0023] 图12是根据本公开的一些示例的可操作以生成合成语音的声控扬声器系统的示意图。

[0024] 图13是根据本公开的一些示例的可操作以生成合成语音的相机的示意图。

[0025] 图14是根据本公开的一些示例的可操作以生成合成语音的诸如虚拟现实或增强现实耳机的耳机的示意图。

[0026] 图15是根据本公开的一些示例的可操作以生成合成语音的载具的第一示例的示意图。

[0027] 图16是根据本公开的一些示例的可操作以生成合成语音的载具的第二示例的示意图。

[0028] 图17是根据本公开的一些示例的可以由图1的设备执行的合成语音生成的方法的特定实施方式的示意图。

[0029] 图18是根据本公开的一些示例的可以由图1的设备执行的合成语音生成方法的另一特定实施方式的示意图。

[0030] 图19是根据本公开的一些示例的可操作以生成合成语音的设备的特定说明性示例的框图。

具体实施方式

[0031] 使用更大、更不同的音频语音样本数据集来训练导致更鲁棒的语音模型。为语音训练提供大量音频语音样本对用户来说可能是不方便和费时的。公开了合成语音生成的系统和方法。例如,语音生成器接收(或访问)语音的输入表示(例如,音频信号、视频信号、文本等)。语音生成器接收指示目标语音特性的控制参数,诸如目标情绪(例如,中性、快乐、悲伤等)、目标语音速率(例如,慢、中或快)、要使用的语音特性的目标人(例如,用户、名人、具有特定口音的儿童等)或其组合。语音生成器包括诸如在多编码器变换器中的多编码器,多编码器基于控制参数处理输入表示以生成基于目标语音特性的编码的数据,以及解码器,解码器通过解码编码的数据生成音频信号(例如,合成语音信号)。语音生成器可以通过将各种控制参数集应用于相同的输入表示来生成各种合成音频样本集。

[0032] 下面参考附图描述本公开的特定方面。在描述中,共同的特征由共同的附图标记指定。如本文所使用的,各种术语仅用于描述特定实施方式的目的而不旨在限制实施方式。例如,单数形式“一(a)”、“一(a)”和“所述、该(the)”旨在也包括复数形式,除非上下文清楚地另有说明。此外,本文描述的一些特征在一些实施方式中是单数的,而在其他实施方式中是复数的。为了说明,图1描绘了包括一个或多个处理器(图1的“(多个)处理器”190)的设备102,其指示在一些实施方式中设备102包括单个处理器190并且在其他实施方式中设备102包括多个处理器190。为了在本文中便于引用,这样的特征通常被引入为“一个或多个”特征并且随后以单数形式被提及,除非正在描述与多个特征相关的方面。

[0033] 如本文所用,术语“包含(comprise)”、“包含(comprises)”和“包含(comprising)”可与“包括(include)”、“包括(includes)”或“包括(including)”互换使用。此外,术语“其中(wherein)”可与“在该处(where)”互换使用。如本文所用,“示例性”指示示例、实施方式和/或方面,并且不应被解释为限制或指示优选或优选的实施方式。如本文所用,用于修饰诸如结构、组件、操作等元素的序数术语(例如,“第一”、“第二”、“第三”等)本身并不指示元素相对于另一元素的任何优先级或顺序,而仅仅是将该元素与另一具有相同名称的元素区分开来(但使用序数术语)。如本文所用,术语“集合”指一个或多个特定元素,术语“多个”指多个(例如,两个或更多个)特定元素。

[0034] 如本文所用,“耦合”可以包括“通信地耦合”、“电耦合”或“物理耦合”,并且还可以(或可替代地)包括它们的任何组合。两个设备(或组件)可以直接地或者经由一个或多个其他设备、组件、线、总线、网络(例如,有线网络、无线网络、或其组合)等间接地耦合(通信地耦合、电耦合或物理耦合)。电耦合的两个设备(或组件)可以被包括在同一设备或不同设备中,并且可以经由电子设备、一个或多个连接器或电感耦合连接,如说明性的、非限制示例所示。在一些实施方式中,诸如在电气通信中通信耦合的两个设备(或组件)可以经由一个或多个线、总线、网络等直接或间接地发送和接收信号(例如,数字信号或模拟信号)。如本文所用,“直接耦合”可以包括在没有中间组件的情况下耦合(例如,通信地耦合、电耦合或物理耦合)的两个设备。

[0035] 在本公开中,可以使用诸如“确定”、“计算”、“估计”、“移位”、“调整”等术语来描述如何执行一个或多个操作。应当注意,这样的术语不应被解释为限制性的,并且可以利用其他技术来执行类似的操作。此外,如本文所指,“生成”、“计算”、“估计”、“使用”、“选择”、“访问”和“确定”可以互换使用。例如,“生成”、“计算”、“估计”或“确定”参数(或信号)可指主动

生成、估计、计算或确定参数(或信号)或可以指使用、选择或访问已经生成的参数(或信号),诸如由另一组件或设备生成的。

[0036] 参考图1,公开了被配置为生成合成语音的系统的特定说明性方面并且总体上指定为100。系统100包括被配置为使用语音生成器120生成合成语音的设备102。设备102耦合到存储器132的一个或多个处理器190。一个或多个处理器190包括语音生成器120。在一些实施方式中,设备102耦合到麦克风182、语音存储库(repository)184或两者。在特定实施方式中,麦克风182、语音存储库184或两者在设备102外部。在可替代实施方式中,麦克风182、语音存储库184或两者集成在设备102中。

[0037] 语音生成器120包括音频分析器122、风格数据生成器126、多编码器变换器128、语音合成器130或其组合。在特定方面,音频分析器122耦合到风格数据生成器126、多编码器变换器128或两者。风格数据生成器126经由多编码器变换器128耦合到语音合成器130。在特定方面,语音合成器130耦合到语音储存库184。

[0038] 音频分析器122被配置为处理输入信号121以生成语音的输入表示123。在特定示例中,输入信号121包括音频信号,例如,经由麦克风182接收的或从存储设备检索的。在另一示例中,输入信号121包括指示音频信息和图像信息的视频信号。输入表示123包括基于输入信号121的文本、梅尔标度频谱图、基频(F0)特征、语音的一个或多个其他表示或其组合。在特定示例中,音频分析器122被配置为执行特征提取以生成梅尔标度频谱图、基频(F0)特征或其组合,作为输入表示123。在特定示例中,音频分析器122被配置为对输入信号121执行语音到文本的转换以生成文本作为输入表示123。

[0039] 风格数据生成器126被配置为基于输入表示123和一个或多个控制参数125生成合并的风格数据127,如参考图2进一步描述的。一个或多个控制参数125指示一个或多个目标语音特性141。例如,目标语音特性141指示目标人143(例如,人180)、目标情绪145(例如,快乐、愤怒、兴奋、悲伤或中性情绪)、目标语音速率147(例如,中等语音速率)或其组合。在特定方面,合并的风格数据127对应于表示输入表示123和一个或多个控制参数125的嵌入(例如,嵌入向量)。

[0040] 多编码器变换器128被配置为处理输入表示123和合并的风格数据127以生成输出频谱数据129,如参考图4A-图5进一步描述的。多编码器变换器128包括多编码器192和解码器194。多编码器192被配置为基于一个或多个控制参数125处理输入表示123以生成对应于合成的语音信号131的编码的数据193。例如,合成的语音信号131基于编码的数据193。在特定方面,多编码器192包括第一编码器,被配置为独立于一个或多个控制参数125(例如,独立于目标语音特性141)处理输入表示123以生成第一编码的数据,如参考图4A进一步描述的。多编码器192包括一个或多个第二编码器,其被配置为处理合并的风格数据127(即基于一个或多个控制参数125)以生成第二编码的数据,如参考图4A-图4B进一步描述的。第二编码的数据因此基于一个或多个控制参数125(例如,目标语音特性141)。解码器194被配置为基于编码的数据193(例如,第一编码的数据和第二编码的数据)生成输出频谱数据129,如参考图4A-图4B进一步描述的。语音合成器130被配置为基于目标语音特性141对输出频谱数据129执行语音合成以生成表示语音版本(例如,由输入表示123指示的)的合成的语音信号131。

[0041] 存储器132被配置为存储由语音生成器120使用或生成的数据。例如,存储器132被

配置为存储输入信号121的至少一部分、输入表示123、一个或多个控制参数125、合并的风格数据127、编码的数据193、输出频谱数据129、合成的语音信号131的至少一部分或其组合。在特定方面，存储器132存储指示一个或多个控制参数125与目标语音特性141之间的映射的映射数据。例如，映射数据指示一个或多个控制参数125的第一控制参数的值与人的集合151之间的第一映射，一个或多个控制参数125的第二控制参数的值与情绪集合161之间的第二映射，一个或多个控制参数125的第三控制参数的值与语音速率集合171之间的第三映射，或其组合。响应于确定映射数据指示由一个或多个控制参数125指示的第一控制参数的值映射到人的集合151的目标人143，风格数据生成器126确定一个或多个控制参数125指示目标人143。类似地，响应于确定映射数据指示由一个或多个控制参数125指示的第二控制参数的值映射到情绪集合161的目标情绪145，风格数据生成器126确定一个或多个控制参数125指示目标情绪145。响应于确定映射数据指示由一个或多个控制参数125指示的第三控制参数的值映射到语音速率集合171的目标语音速率147，风格数据生成器126确定一个或多个控制参数125指示目标语音速率147。

[0042] 在一些实施方式中，设备102对应于或被包括在一个或多个类型的设备中。在说明性示例中，一个或多个处理器190被集成在包括麦克风182的耳机设备中，诸如参考图10进一步描述的。在其他示例中，一个或多个处理器190被集成在以下各项的至少一个中：如参考图9所描述的移动电话或平板计算机设备、如参考图11所描述的可穿戴电子设备、如参考图12所描述的语音控制的扬声器系统、如参考图13所描述的相机设备、或者如参考图14所描述的虚拟现实、增强现实或混合现实耳机。在另一说明性示例中，一个或多个处理器190被集成到还包括麦克风182的载具中，诸如参考图15和图16进一步描述的。

[0043] 在操作期间，音频分析器122经由麦克风182接收输入信号121。例如，麦克风182捕获人180的语音(例如，“将门解锁”)并将表示语音的输入信号121提供给音频分析器122。由输入信号121表示的语音(例如，“将门解锁”)具有特定的语音速率(例如，快速的语音速率)和特定的情绪(例如，焦虑)。

[0044] 音频分析器122生成输入信号121的输入表示123。例如，音频分析器122对输入信号121执行语音到文本的转换以生成文本(例如，“将门解锁”)。作为另一示例，音频分析器122对输入信号121执行音频特征提取以生成梅尔标度频谱图、基频(F0)特征或其组合。输入表示123包括文本(例如，“将门解锁”)、梅尔标度频谱图、基频(F0)特征或其组合。

[0045] 设备102接收或访问一个或多个控制参数125。在特定方面，一个或多个控制参数125基于默认数据、配置设置、用户输入或其组合。一个或多个控制参数125指示要由语音生成器120生成的合成语音的目标语音特性141。

[0046] 在特定方面，目标语音特性141指示人的集合151中的目标人143(例如，人180)。人的集合151包括第一人(人1)153、第二人(人2))155，一个或多个另外的人，诸如第N人(人N)157，或其组合。在说明性非限制性示例中，人的集合151包括特定名人、特定用户、具有特定口音的人、特定年龄的人、特定性别的人或其组合。目标人143的语音特性将被语音生成器120用来生成合成语音。在特定示例中，目标人143指示“奥巴马总统”。在另一示例中，目标人143指示“具有德克萨斯口音的中年男性”。在特定示例中，目标人143指示人180。

[0047] 在特定方面，目标语音特性141指示从情绪集合161中选择的目标情绪145(例如，中性的)。情绪集合161包括第一情绪(情绪1)163、第二情绪(情绪2)165、一个或多个另外的

情绪，诸如第N情绪(情绪N)167，或其组合。在说明性、非限制性示例中，情绪集合161包括快乐、恐惧、焦虑、兴奋、愤怒、平静、中性或其组合。

[0048] 在特定方面，目标语音特性141指示从语音速率集合171中选择的目标语音速率147(例如，中等语音速率)。语音速率集合171包括第一语音速率(语音速率1)173、第二语音速率(语音速率2)175、一个或多个另外的语音速率，诸如第N语音速率(语音速率N)177，或其组合。在说明性、非限制性示例中，语音速率集合171包括慢速语音速率、中等语音速率、快速语音速率或其组合。

[0049] 风格数据生成器126基于输入表示123和一个或多个控制参数125生成合并的风格数据127，如参考图2进一步描述的。例如，合并的风格数据127对应于表示输入表示123和一个或多个控制参数125的嵌入(例如，嵌入向量)。

[0050] 多编码器变换器128处理输入表示123和合并的风格数据127以生成输出频谱数据129，如参考图4A-图4B进一步描述的。例如，多编码器192的第一编码器独立于一个或多个控制参数125(例如，独立于目标语音特性141)处理输入表示123以生成第一编码的数据，如参考图4A进一步描述的。多编码器192的一个或多个第二编码器处理合并的风格数据127以生成第二编码的数据，如参考图4A-图4B进一步描述的。

[0051] 在特定方面，合并的风格数据127在输入表示123的处理期间由多编码器192使用。例如，在与第一编码器处理输入表示123的同时，一个或多个第二编码器处理合并的风格数据127。多编码器192向解码器194提供编码的数据193(例如，第一编码的数据、第二编码的数据或其组合)。在特定方面，编码的数据193对应于多编码器192的输出嵌入(例如，嵌入向量)。

[0052] 解码器194处理编码的数据193(例如，第一编码的数据和第二编码的数据)以生成输出频谱数据129，如参考图4A进一步描述的。语音合成器130对输出频谱数据129执行语音合成以生成合成的语音信号131。合成的语音信号131对应于由基于目标语音特性141的输入表示123指示的语音版本。

[0053] 作为说明性示例，多编码器192和解码器194被描述为包括在多编码器变换器128中。在一些示例中，多编码器192和解码器194具有不同于变换器的配置。

[0054] 在多编码器192中具有用于一个或多个控制参数125的一个或多个第二编码器(并且与独立于一个或多个控制参数125的第一编码器分离)使得第二编码器能够被专门训练(例如，优化)，以用于基于一个或多个控制参数125中的每一个对输入表示123进行编码，如参考图3进一步描述的。

[0055] 在特定示例中，合成的语音信号131具有目标语音特性141。在特定示例中，合成的语音信号131更接近(与输入信号121相比)具有目标语音特性141。在特定方面，语音生成器120包括调整器，调整器被配置为基于一个或多个控制参数125与对应于合成语音信号131的估计的控制参数的比较来调整多编码器变换器128的配置参数(例如，神经网络权重)，如参考图3进一步描述的。语音生成器120因此可以动态地训练多编码器变换器128以提高合成语音生成的准确性。

[0056] 作为说明性示例，设备102被描述为接收输入信号121并且基于输入信号121生成输入表示123。例如，风格数据生成器126和多编码器变换器128中的每一个都从音频分析器122接收输入表示123。在一些示例中，设备102接收输入表示123(例如，而不是输入信号

121)。例如,设备102(例如,多编码器转换器128和风格数据生成器126中的每一个)接收输入表示123作为来自人180、另一设备或两者的输入。

[0057] 在说明性示例中,输入表示123对应于具有特定情绪(例如,焦虑)和特定语音速率(例如,高语音速率)的人180的语音。合成的语音信号131对应于具有目标情绪145(例如,中性)和目标语音速率147(例如,中等语音速率)的目标人143(例如,人180)的语音。在这个示例中,与对输入信号121(例如,具有高语音速率的焦虑语音)执行语音识别相比,对合成的语音信号131(例如,具有中等语音速率的中性语音)执行语音识别提高了识别准确性。

[0058] 在特定方面,语音生成器120将合成的语音信号131添加到语音存储库184。例如,语音生成器120基于输入信号121生成(或更新)训练数据集,如参考图6进一步描述的。为了说明,语音生成器120基于一个或多个控制参数125的第一集合生成第一合成的语音信号131,基于一个或多个控制参数125的第二集合生成第二合成的语音信号131,基于一个或多个控制参数125的一个或多个附加集合生成一个或多个附加合成的语音信号131,或其组合。语音生成器120将输入信号121、第一合成的语音信号131、第二合成的语音信号131、一个或多个附加合成的语音信号131或其组合添加到语音存储库184中的训练数据集。

[0059] 在特定示例中,语音生成器120基于同一输入信号121生成对应于各种情绪和语音速率的多个合成的语音信号。在该示例中,合成的语音信号可以用于训练鲁棒语音模型(例如,神经网络),鲁棒语音模型使用人180说的几个词来表示人180的语音。为了说明,在注册期间,具有特定的语音特性(例如,具有中等语音速率的中性语音)的人180的语音(例如,注册语音)被接收作为用于训练神经网络的输入信号121。基于对应于各种语音特性的合成的语音训练神经网络使得神经网络能够识别具有与注册语音的语音特性(例如,具有中等语音速率的中性语音)不同的语音特性的语音(例如具有快速语音速率的焦虑语音)。

[0060] 在特定示例中,语音生成器120基于相同的输入信号121生成对应于各种情绪、各种语音速率和各种人的语音特性的多个合成的语音信号。在该示例中,大量且不同的合成的语音数据集可以使用人180所说的几个词生成。

[0061] 尽管麦克风182被示为耦合到设备102,但是在其他实施方式中麦克风182可以集成在设备102中。尽管示出了一个麦克风182,但是在其他实施方式中可以包括配置为捕获用户语音的一个或多个附加麦克风。在一些实施方式中,麦克风182被省略,诸如当语音生成器120从存储装置中检索输入信号121或输入表示123时(例如,作为从存储器132检索或从另一设备接收的文本数据)。尽管系统100被示为包括设备102中的语音生成器120的所有组件,但是在其他实施方式中可以添加第二设备,并且第二设备可以执行描述为在设备102处执行的一些操作。

[0062] 参考图2,示出了语音生成器120的特定实施方式。风格数据生成器126包括多个语音特征编码器,诸如人编码器206、情绪编码器208、语音速率编码器210、一个或多个附加语音特性编码器或其组合。

[0063] 在特定方面,音频分析器122耦合到多个语音特性编码器中的每一个。多个语音特性编码器中的每一个耦合到编码器前置网络212。多个语音特性编码器中的每一个被配置为基于由一个或多个控制参数125指示的对应语音特性生成输入表示123的编码的版本。例如,人编码器206被配置为处理输入表示123和一个或多个控制参数125中的人控制参数(例如,指示目标人143)以生成人编码的输入表示207。为了说明,人编码器206包括神经网络

(例如,卷积神经网络(CNN)),其被配置为接收输入表示123和人控制参数(例如,指示目标人143)作为输入并且输出表示输入表示123和人控制参数的人编码的输入表示207(例如,嵌入向量)。

[0064] 类似地,情绪编码器208被配置为处理输入表示123和一个或多个控制参数125的情绪控制参数(例如,指示目标情绪145)以生成情绪编码的输入表示209。在特定示例中,语音速率编码器210被配置为处理输入表示123和一个或多个控制参数125的语音速率参数(例如,指示目标语音速率147)以生成语音速率编码的输入表示211。

[0065] 编码器前置网络212包括神经网络(例如,CNN)、线性投影层或两者。在特定示例中,人编码的输入表示207、情绪编码的输入表示209、语音速率编码的输入表示211或其组合对应于神经网络(例如,CNN)的输入嵌入。编码器前置网络212的神经网络处理输入嵌入(例如,编码的输入表示)以生成输出嵌入。神经网络向线性投影层提供输出嵌入,并且线性投影层通过对神经网络的输出嵌入执行线性投影来生成合并的风格数据127。在特定方面,多编码器变换器128不包括任何卷积,并且经由使用编码器前置网络212执行卷积以处理人编码的输入表示207、情绪编码的输入表示209、语音速率编码的输入表示211,或其组合。编码器前置网络212将合并的风格数据127提供给多编码器192。

[0066] 多编码器192接收来自音频分析器122的输入表示123和来自风格数据生成器126的合并的风格数据127。多编码器192处理输入表示123和合并的风格数据127以生成编码的数据193。例如,多编码器192处理输入表示123以生成独立于图1的目标语音特性141的第一编码的数据,并且处理合并的风格数据127以生成对应于图1的目标语音特性141的第二编码的数据,如参考图4A-图4B进一步描述的。在特定方面,编码器前置网络212还向解码器194提供合并的风格数据127。在这个方面,解码器194基于合并的风格数据127处理编码的数据193以生成输出频谱数据129。

[0067] 参考图3,示出了语音生成器120的特定实施方式,其中语音合成器130经由控制参数估计器310和损失计算器312耦合到调整器314。

[0068] 控制参数估计器310处理合成的语音信号131以生成一个或多个估计的控制参数311。例如,控制参数估计器310确定对应于合成的语音信号131的估计的人、估计的情绪、估计的语音速率或其组合。

[0069] 在特定示例中,控制参数估计器310通过将合成的语音信号131与和图1的人的集合151中的每个人相关联的语音模型进行比较来生成模型匹配分数。控制参数估计器310从模型匹配分数中选择对应于最紧密匹配合成的语音信号131的特定语音模型的模型匹配分数(例如,最高匹配分数)。控制参数估计器310响应于确定特定语音模型表示特定人的语音,生成将特定人指示为估计的人的一个或多个估计的控制参数311。

[0070] 在特定方面,控制参数估计器310使用情绪检测技术来检测由合成的语音信号131指示的估计的情绪。例如,控制参数估计器310包括神经网络,神经网络被配置为处理表示合成的语音信号131的输入嵌入以生成表示估计的情绪的输出(例如,分类器输出或嵌入)。

[0071] 在特定方面,控制参数估计器310使用音频分析技术来确定在合成的语音信号131中检测的语音的平均(例如,均值、中值或众数)速率,并生成指示作为估计的语音速率的平均语音速率的一个或多个估计的控制参数311。控制参数估计器310将一个或多个估计的控制参数311提供给损失计算器312。

[0072] 损失计算器312基于一个或多个控制参数125与估计的控制参数311的比较生成损失度量313。例如，损失度量313包括人损失度量、情绪损失度量、语音速率损失度量或其组合。

[0073] 在特定方面，目标人143由向量空间中的第一人嵌入向量表示，并且估计的人由向量空间中的第二人嵌入向量表示。例如，向量空间的维度表示各种语音特性(例如，音高、音调等)。控制参数估计器310基于第一人嵌入向量和第二人嵌入向量之间的向量距离来确定人损失度量。

[0074] 在特定方面，目标情绪145由向量空间中的第一情绪嵌入向量表示，并且估计的情绪由向量空间中的第二情绪嵌入向量表示。例如，向量空间表示情绪图表。控制参数估计器310基于第一情绪嵌入向量和第二情绪嵌入向量之间的向量距离来确定情绪损失度量。

[0075] 在特定方面，控制参数估计器310基于目标语音速率147和估计的语音速率之间的差异来确定速度速率损失度量(例如，速度速率损失度量＝目标语音速率147减去估计的语音速率)。

[0076] 损失计算器312向调整器314提供损失度量313(例如，人损失度量、情绪损失度量、语音速率损失度量或其组合)。调整器314响应于确定损失度量313指示大于损失阈值的损失，使用机器学习(例如，神经网络训练)技术来训练多编码器192、人编码器206、情绪编码器208、语音速率编码器210、编码器前置网络212、解码器194的解码器网络或其组合的一个或多个配置参数(例如，神经网络权重)，以减少用于输入表示的后续处理的损失度量313。

[0077] 在特定方面，调整器314响应于确定人损失度量指示损失大于人损失阈值，训练人编码器206、编码器前置网络212、图4A的人码器420、掩码的多头注意力466、多头注意力478或其组合的配置参数(例如，神经网络权重)。

[0078] 在特定方面，调整器314响应于确定情绪损失度量指示大于情绪损失阈值的损失，训练情绪编码器208、编码器前置网络212、图4A的情绪编码器430、掩码的多头注意力466、多头注意力480或其组合的配置参数(例如，神经网络权重)。

[0079] 在特定方面，调整器314响应于确定语音速率损失度量指示大于语音速率损失阈值的损失，训练语音速率编码器210、编码器前置网络212、图4A的速率编码器440、掩码的多头注意力466、多头注意力482或其组合的配置参数(例如，神经网络权重)。

[0080] 调整器314因此能够训练语音生成器120以提高合成语音生成在匹配目标语音特性141方面的准确性。例如，语音生成器120的配置参数(例如，神经网络权重)可以由调整器314独立于用户输入自动调谐，以提高合成语音生成的准确性。

[0081] 参考图4A，示出了语音生成器120的特定实施方式。多编码器192包括输入表示编码器450和用于每个语音特性的语音特性编码器。例如，多编码器192包括人编码器420、情绪编码器430、速率编码器440或其组合。

[0082] 多编码器192的每个编码器被配置为通过独立于多编码器192的其他编码器处理输入数据来生成编码的数据，如参考图5进一步描述的。例如，输入表示编码器450被配置为处理输入表示123以生成第一编码的数据458。人编码器420被配置为处理合并的风格数据127生成人编码的数据428。情绪编码器430被配置为处理合并的风格数据127以生成情绪编码的数据438。速率编码器440被配置为处理合并的风格数据127以生成语音速率编码的数据448。由于多编码器192的每个编码器独立于多编码器192的其他编码器操作，调整器314

可以基于对应的语音特性损失度量来专门调谐每个编码器以提高合成的语音生成的准确性，如参考图3描述的。

[0083] 多编码器192的每个编码器包括输入级、一个或多个编码层和归一化层(层归一化)，如参考图5进一步描述的。例如，输入表示编码器450包括输入级452、一个或多个编码层454和层归一化456。人编码器420包括输入级422、一个或多个编码层424和层归一化426。情绪编码器430包括输入级432、一个或多个编码层434和层归一化436。速率编码器440包括输入级442、一个或多个编码层444和层归一化446。

[0084] 在一些实施方式中，编码器420、430、440和450中的每一个都如参考图5中描绘的编码器500所描述的那样配置和操作。在图5中，编码器500包括输入级522、一个或多个编码层554和层归一化526。一个或多个编码层554的每个编码层包括注意力层和前馈层。注意力层包括注意力网络，诸如多头注意力564。前馈层包括前馈神经网络，诸如前馈570(例如，全连接的前馈神经网络)。在特定示例中，注意力层包括经由多头注意力564耦合到组合器566的层归一化562。前馈层包括经由前馈570耦合到组合器572的层归一化568。注意力层耦合到前馈层。例如，组合器566耦合到层归一化568和组合器572。

[0085] 包括单个编码层的一个或多个编码层554被提供作为说明性示例。在其他示例中，一个或多个编码层554包括多个编码层，其中输入级522的输出耦合到初始编码层的注意力层(例如，层归一化562)、每个先前编码层的前馈层(例如，组合器572)耦合到后续编码层的注意力层(例如，层归一化562)，以及最后的编码层的前馈层(例如，组合器572)耦合到层归一化526。

[0086] 输入级522接收输入数据521。在特定示例中，编码器500对应于图4的输入表示编码器450。例如，输入数据521对应输入表示123，输入级522对应输入级452，一个或多个编码层554对应一个或多个编码层454，层归一化526对应层归一化456。在特定实施方式中，输入级452包括处理输入表示123以生成输出的编码器前置网络(例如，CNN、线性投影层或两者)。例如，输入表示编码器450的编码器前置网络用于添加卷积。输入级452被配置为处理编码器前置网络的输出。例如，输入级452被配置为通过将编码器前置网络的输出与位置编码数据组合来生成输出。

[0087] 在特定方面，编码器500对应于人编码器420、情绪编码器430或速率编码器440，输入数据521对应于合并的风格数据127，并且输入级522被配置为通过处理合并的风格数据127生成输出。例如，输入级522被配置为通过将合并的风格数据127与位置编码数据组合来生成输出。

[0088] 在特定示例中，编码器500对应于人编码器420。为了说明，输入级522对应于输入级422，一个或多个编码层554对应于一个或多个编码层424，并且层归一化526对应于层归一化426。在特定示例中，编码器500对应于情绪编码器430。为了说明，输入级522对应于输入级432，一个或多个编码层554对应于一个或多个编码层434，层归一化526对应于层归一化436。在特定示例中，编码器500对应于速率编码器440。为了说明，输入级522对应于输入级442，一个或多个编码器层554对应于一个或多个编码层444，层归一化526对应于层归一化446。

[0089] 输入级522的输出在层归一化562归一化之后提供给多头注意力564。多头注意力564可以使用不同的注意力头从不同方面构建上下文向量。例如，多头注意力564包括并行

处理多头注意力564的输入的注意力头。为了说明，将输入乘以第一矩阵、第二矩阵和第三矩阵以分别生成第一查询(Query)向量、第一键(Key)向量和第一值(Value)向量。第一查询向量、第一键向量和第一值向量由第一注意力头处理。输入与第四矩阵、第五矩阵和第六矩阵相乘以分别生成第二查询向量、第二键向量和第二值向量。第二查询向量、第二键向量和第二值向量由第二注意力头与处理第一查询向量、第一键向量和第一值向量的第一注意力头并行或同时处理。

[0090] 在特定方面，注意力头的输出对应于以下等式：

[0091] $$Z = softmax\left(\frac{QxK^T}{\sqrt{d_k}}\right)V,$$

[0092] 其中Z对应注意力头的输出，Q对应查询向量，x对应乘法运算符，K对应键向量，V对应值向量，dk对应键向量的维度，而softmax对应归一化操作。

[0093] 注意力头的独立输出被级联并线性变换以生成多头注意力564的输出。组合器566通过组合层归一化562的输入和多头注意力564的输出来生成输出。

[0094] 在通过层归一化568归一化之后，组合器566的输出被提供给前馈570(例如，全连接的前馈神经网络)。在特定示例中，前馈570包括第一线性变换层，该第一线性变换层经由整流线性单元(ReLU)层耦合到第二线性变换层。前馈570通过处理组合器566的输出来生成输出。

[0095] 组合器572通过将组合器566的输出与前馈570的输出组合来生成输出。在特定方面，层归一化526通过对组合器572的输出(例如，一个或多个编码层554的最后的编码层)应用归一化来生成编码的数据528。例如，应用归一化包括将组合器572的输出(例如，输出向量)的每个值调整到特定范围内。

[0096] 回到图4A，输入表示编码器450被配置为独立于一个或多个控制参数125对输入表示123进行编码以生成第一编码的数据458。人编码器420、情绪编码器430和速率编码器440中的每一个被配置为基于一个或多个控制参数125对输入表示123进行编码以生成编码的数据。例如，基于一个或多个控制参数125对输入表示123进行编码包括对合并的风格数据127进行编码。为了说明，人编码器420通过对合并的风格数据127进行编码来生成人编码的数据428。情绪编码器430通过对合并的风格数据127进行编码来生成情绪编码的数据438。速率编码器440通过对合并的风格数据127进行编码来生成语音速率编码的数据448。

[0097] 在特定方面，人编码器420、情绪编码器430和速率编码器440在对应的多头注意力564使用的权重或矩阵方面彼此不同。例如，人编码器420的多头注意力564使用指示权重的第一集合的矩阵的第一集合来生成查询向量、键向量和值向量。情绪编码器430的多头注意力564使用指示权重的第二集合的矩阵的第二集合来生成查询向量、键向量和值向量。类似地，速率编码器440的多头注意力564使用指示权重的第三集合的矩阵的第三集合来生成查询向量、键向量和值向量。

[0098] 多编码器192向解码器194提供编码的数据193。例如，编码的数据193包括独立于一个或多个控制参数125的第一编码的数据458。在特定方面，编码的数据193还包括基于一个或多个控制参数125的第二编码的数据(例如，人编码的数据428、情绪编码的数据438、语音速率编码的数据448或其组合)。

[0099] 解码器194包括解码器前置网络460、组合器462、一个或多个解码层498、层归一化

492、线性变换层(线性)494或其组合。一个或多个解码层498中的每个解码层包括掩码的注意力层、注意力层和前馈层。例如,掩码的注意力层包括经由掩码的多头注意力466(例如,掩码的解码器注意力网络)耦合到多个归一化层(层归一化)(诸如层归一化468,层归一化470、层归一化472、层归一化474或其组合)的层归一化464。注意力层包括经由组合器486耦合到层归一化488的多个解码器注意力网络。例如,多个解码器注意力网络(例如,解码器多头注意力网络)包括多头注意力478、多头注意力480、多头注意力482、多头注意力484或其组合。前馈层包括前馈490(例如,全连接的前馈神经网络)。

[0100] 掩码的注意力层耦合到注意力层。例如,掩码的注意力层的多个层归一化中的每一个被耦合到注意力层的对应多头注意力网络。例如层归一化468、层归一化470、层归一化472、层归一化474分别耦合到多头注意力478、多头注意力480、多头注意力482、多头注意力484。掩码的多头注意力466经由组合器486耦合到层归一化488。注意力层耦合到前馈层。例如,层归一化488耦合到前馈490。

[0101] 组合器462耦合到一个或多个解码层498的解码层。例如,组合器462耦合到解码层的层归一化464。一个或多个解码层498中的解码层耦合到层归一化492。例如,解码层的前馈490耦合到层归一化492。

[0102] 在特定方面,解码器前置网络460接收对应于解码器194为先前时间步长生成的输出频谱数据129的反馈输出频谱数据429。解码器前置网络460包括处理反馈输出频谱数据429以生成输出的神经网络(例如,CNN)。在特定方面,解码器前置网络460添加卷积。在特定方面,组合器462通过组合解码器前置网络460和位置编码数据461的输出来生成输出。

[0103] 组合器462的输出由一个或多个解码层498的解码层处理。例如,组合器462的输出在解码层的层归一化464归一化之后提供给掩码的多头部注意力466。在特定方面,掩码的多头注意力466掩蔽(mask)到掩码的多头注意力466的输入中的未来位置。掩码的多头注意力466从到掩码的多头注意力466的输入的掩码的版本中生成查询向量、键向量和值向量。掩码的多头注意力466的每个注意力头处理查询向量、键向量和值向量以生成输出。将掩码的多头注意力466的注意力头的独立输出级联起来并进行线性变换,以生成掩码的多头注意力466的输出。

[0104] 掩码的多头注意力466的输出被提供给组合器486,并且在归一化之后提供给解码层的多个多头注意力。例如,掩码的多头注意力466的输出在由层归一化468、层归一化470、层归一化472和层归一化474归一化之后被分别提供给多头注意力478、多头部注意力480、多头注意力482和多头注意力484。

[0105] 解码层的多个多头注意力中的每一个被配置为处理掩码的多头注意力466的输出的归一化版本和从多编码器192的对应编码器接收的编码的数据。例如,解码层的多个多头注意力中的每一个处理(例如,生成和处理)基于掩码的多头注意力466的输出的归一化版本的查询向量以及处理(例如,生成和处理)基于从多编码器192的对应编码器接收的编码的数据的键向量和值向量。为了说明,多头注意力484被配置为处理基于层归一化474的输出的查询向量并处理基于第一编码的数据458的键向量和值向量。多头注意力478、多头注意里480和多头注意力482被配置为分别处理人编码的数据428、情绪编码的数据438和语音速率编码的数据448,以及层归一化468、层归一化470和层归一化472的输出。例如,多头注意力478处理基于层归一化468输出的查询向量,并处理基于人编码的数据428的键向量和

值向量。

[0106] 组合器486通过将掩码的多头注意力466的输出与多头注意力478的输出、多头注意力480的输出、多头注意力482的输出、多头注意力484的输出，或其组合进行组合来生成输出。在通过层归一化488归一化之后，组合器486的输出被提供给解码层的前馈490。

[0107] 一个或多个解码层498的特定解码层的前馈490的输出，在层归一化492的归一化和层归一化492的线性变换之后，作为输出频谱数据129提供给语音合成器130。在特定方面，语音合成器130基于输出频谱数据129、编码的数据193或其组合执行语音合成，以生成合成的语音信号131。输出频谱数据129被提供给解码器前置网络460作为后续时间步长的反馈输出频谱数据429。

[0108] 包括单个解码层的一个或多个解码层498被提供作为说明性示例。在其他示例中，一个或多个解码层498包括多个解码层，其中，组合器462的输出耦合到初始解码层的层归一化464，每个先前解码层的前馈490耦合到后续编码层的层归一化464，以及最后的解码层的前馈490耦合到层归一化492。多编码器192的最后的编码层将编码的数据193提供给一个或多个解码层498的每个的多个多头注意力。

[0109] 在特定方面，图3的调整器314基于损失度量313调整多编码器192和解码器194的配置参数(例如，神经网络权重)。例如，调整器314响应于确定损失度量313的人损失度量不满足人损失阈值，训练人编码器420的一个或多个编码层424、多头注意力478、掩码的多头注意力466、一个或多个解码层498的前馈490、解码器前置网络460或其组合的神经网络权重。

[0110] 参考图4B，示出了语音生成器120的特定实施方式。多编码器192包括用于语音特性的特性编码器451，而不是用于每个语音特性的单独的语音特性编码器。例如，特性编码器451包括输入级453、一个或多个编码层455和层归一化457。在特定方面，图5的编码器500是特性编码器451的实施方式的示例。例如，输入级522对应输入级453，一个或多个编码层554对应一个或多个编码层455，层归一化526对应层归一化457。

[0111] 特性编码器451从风格数据生成器126接收合并的风格数据127(例如，输入嵌入向量)。在特定方面，输入级453通过将合并的风格数据127与位置编码数据组合来生成输出。输入级453的输出由一个或多个编码层455处理，如参考图5的一个或多个编码层554所描述的。一个或多个编码层455的输出在通过层归一化457归一化之后作为第二编码的数据459提供给解码器194。例如，编码的数据193包括第一编码的数据458和第二编码的数据459。

[0112] 一个或多个解码层498包括经由归一化层(层归一化)耦合到两个多头注意力的掩码的多头注意力466。例如，掩码的多头注意力466经由层归一化475耦合到多头注意力485，并经由层归一化474耦合到多头注意力484。组合器486耦合到掩码的多头注意力466，多头注意力485，多头注意力484。

[0113] 在通过层归一化475归一化之后，将掩码的多头注意力466的输出提供给多头注意力485。特性编码器451将第二编码的数据459提供给多头注意力485。多头注意力485基于掩码的多头注意力466的输出生成查询向量并且基于第二编码的数据459生成键向量和值向量。多头注意力485的每个注意力头处理查询向量，键向量和值向量以生成输出。多头注意力485的每个注意力头的输出被级联并线性变换以生成多头注意力485的输出。组合器486通过将掩码的多头注意力466的输出与多头注意力485的输出、多头注意力484的输出、或两

者进行组合以生成输出。

[0114] 在特定方面，图3的调整器314基于损失度量313调整多编码器192和解码器194的配置参数(例如，神经网络权重)。例如，调整器314响应于确定损失度量313中的人损失度量、情绪损失度量、或语音速率损失度量中的任何一个语音速率不满足对应的损失阈值，训练特性编码器451的一个或多个编码层455、多头注意力485，掩码的多头注意力466、一个或多个解码层498的前馈490、解码器前置网络460或其组合的神经网络权重。

[0115] 参考图6，示出了图1的系统100的组件的操作的说明性方面并且总体上被指定为600。输入信号121包括多个音频样本602，例如音频样本602A、音频样本602B、另外的音频样本的一个或多个集合，诸如第N音频样本602N(其中N是大于二的整数)或其组合。音频样本602表示特定人(例如，人180)的语音、特定情绪(例如，中性或无聊)、特定语音速率(例如，中速)或其组合。

[0116] 语音生成器120接收(例如，访问)一个或多个控制参数621。一个或多个控制参数621指示目标人623(例如，人180)、目标情绪625(例如，兴奋)、目标语音速率627(例如，快速语音速率)或其组合。一个或多个控制参数621指示将由语音生成器120基于输入信号121生成的合成的语音信号661的语音特性。在特定示例中，目标人623、目标情绪625、以及目标语音速率627中的至少一个分别不同于特定的人、特定的情绪和特定的语音速率。如参考图1所描述的，语音生成器120基于一个或多个控制参数621来处理输入信号121，以生成合成的语音信号661。合成的语音信号661包括基于目标人623、目标情绪625、目标语音速率627或其组合的合成的音频样本629。例如，合成的音频样本629表示具有与目标人623(例如，人180)相似的语音特性、与目标情绪625(例如，兴奋)、目标语音速率627(例如，快速语音速率)或其组合相关联的相似的语音特性(例如，音高或音调)的语音。在特定方面，语音生成器120(或设备102的另一组件)将音频样本602和合成的音频样本629添加到训练数据集665。

[0117] 在特定方面，语音生成器120从同一输入信号121生成具有对应于各种控制参数的各种语音特性的音频样本的多个集合。例如，语音生成器120接收(例如，访问)指示目标人633(例如，人180)、目标情绪635(例如，疲倦)、目标语音速率637(例如，缓慢的语音速率)或它们的组合的一个或多个控制参数631。如关于图1所描述的，语音生成器120基于一个或多个控制参数631来处理输入信号121，以生成合成的语音信号663。合成的语音信号663包括基于目标人633(例如人180)、目标情绪635(例如疲倦)、目标语音速率637(例如，缓慢语音速率)或其组合的合成的音频样本639。在特定方面，语音生成器120(或设备102的另一组件)将合成的音频样本639添加到训练数据集665。

[0118] 网络训练器670使用机器学习技术基于训练数据集665训练语音识别网络672。在训练数据集665中包括合成的音频样本629和合成的音频样本639(除了音频样本602之外)提高了语音识别网络672在识别语音方面的准确性。例如，语音识别网络672(基于合成的音频样本629和合成的音频样本639训练的)在识别与输入信号121中由人180提供的语音的特性(例如，中性)具有不同特性(例如，快或慢和疲倦或兴奋)的人180的语音方面更加鲁棒。

[0119] 图7是图1的系统的组件的操作的说明性方面的示意图。音频分析器122被配置为接收音频数据样本的序列710，诸如输入信号121的连续捕获的帧序列，示为第一帧(F1)712、第二帧(F2)714、以及包括第N帧(FN)716(其中N是大于二的整数)的一个或多个附加

帧。音频分析器122被配置为输出输入表示数据集合的序列720，诸如输入表示123的集合序列，包括第一集合722、第二集合724和包括第N集合726的一个或多个附加集合。

[0120] 风格数据生成器126被配置为接收输入表示数据集的序列720并且基于序列720和一个或多个控制参数125生成合并的风格数据集的序列730。序列730包括第一集合(M1)732、第二集合(M2)734和包括第N集合(MN)736的一个或多个附加集合。多编码器192被配置为接收序列720和序列730并生成数据的编码集合的序列740，诸如第一编码集合(E1)742、第二编码集合(E2)744，以及包括第N集合(EN)746的一个或多个附加编码集合。解码器194被配置为接收编码集合的序列740并生成频谱数据序列750，频谱数据序列750包括第一频谱数据(S1)752、第二频谱数据(S2)754，以及包括第N频谱数据(SN)756的一个或多个另外的频谱数据集合。

[0121] 在操作期间，音频分析器122处理第一帧(F1)712以生成输入表示123的第一集合722并且风格数据生成器126处理第一集合722以生成合并的风格数据127的第一集合(M1)732。多编码器192处理输入表示123的第一集合722和合并的风格数据127的第一集合(M1)732以生成编码的数据193的第一编码集合(E1)。解码器194处理编码的数据193的第一编码集合(E1)以生成输出频谱数据129的第一频谱数据(S1)752。随后的编码集合(例如，第二编码集合(E2))由解码器194至少部分地基于第一频谱数据(S1)752处理。这样的处理继续，包括音频分析器122处理第N帧716以生成输入表示123的第N集合726，风格数据生成器126处理输入表示123的第N集合726以生成合并的风格数据127的第N集合(MN)736，多编码器192处理输入表示123的第N集合726和合并的风格数据127的第N集合(MN)736以生成编码的数据193的第N编码集合(EN)746，解码器194处理编码的数据193的第N编码集合(EN)746以生成第N频谱数据(SN)756。第N频谱数据(SN)756基于特征数据的第N集合726并且至少部分地基于序列720的先前特征数据集合中的一个或多个。通过基于频谱数据的一个或多个先前集合进行解码，对于可能跨多个音频数据帧的语音信号，可以改进解码器194的合成的语音生成的准确性。

[0122] 图8描绘了作为包括一个或多个处理器190的集成电路802的设备102的实施方式800。集成电路802还包括音频输入804，诸如一个或多个总线接口，以使输入信号121能够被接收以用于处理。集成电路802还包括诸如总线接口的音频输出806，以能够发送输出信号，诸如合成的语音信号131。集成电路802能够将合成的语音生成实现作为系统中的组件，该组件包括麦克风，诸如如图9中描绘的移动电话或平板电脑、如图10所示的耳机、如图11所示的可穿戴电子设备、如图12所示的语音控制扬声器系统、如图13图所示的相机、如图14所示的虚拟现实耳机或增强现实耳机或如图15或图16所示的载具。

[0123] 图9描绘了作为说明性非限制性示例的实施方式900，其中设备102包括诸如电话或平板电脑的移动设备902。移动设备902包括麦克风182和显示屏904。包括语音生成器120的处理器190的组件集成在移动设备902中，并使用虚线示出，以指示通常对移动设备902的用户不可见的内部组件。在特定示例中，语音生成器120操作以生成合成的语音(或图6的语音识别网络672操作以识别语音)，然后对其进行处理以在操作移动设备902处执行一个或多个操作，诸如启动图形用户界面或以其他方式在显示屏904处(例如，经由集成的“智能助手”应用)显示与用户的语音相关联的其他信息。

[0124] 图10描绘了设备102包括耳机设备1002的实施方式1000。耳机设备1002包括麦克

风182。包括语音生成器120的处理器190的组件集成在耳机设备1002中。在特定示例中，语音生成器120操作以生成合成的语音(或者图6的语音识别网络672操作以识别语音)，这可以使耳机设备1002在耳机设备1002处执行一个或多个操作，以向第二个设备(未示出)发送对应于合成的语音的音频数据以进行进一步处理，或其组合。

[0125] 图11描绘了设备102包括可穿戴电子设备1102，被示为“智能手表”的实施方式1100。语音生成器120和麦克风182被集成到可穿戴电子设备1102中。在特定示例中，语音生成器120操作以生成合成的语音(或图6的语音识别网络672操作以识别语音)，其然后被处理以在可穿戴电子设备1102处执行一个或多个操作，诸如启动图形用户界面或以其他方式在可穿戴电子设备1102的显示屏1104处显示与用户的语音相关联的其他信息。为了说明，可穿戴电子设备1102可以包括显示屏，该显示屏被配置为基于可穿戴电子设备1102检测的用户语音显示通知。在特定示例中，可穿戴电子设备1102包括响应用户语音活动的检测或合成的语音的生成提供触觉通知(例如，振动)的触觉设备。例如，触觉通知可以使用户注视可穿戴电子设备1102以看到指示检测到用户说出的关键词的显示通知。可穿戴电子设备1102因此可以警告有听力障碍的用户或佩戴耳机的用户检测到用户的语音活动。

[0126] 图12是设备102包括无线扬声器和语音激活设备1202的实施方式1200。无线扬声器和语音激活设备1202可以具有无线网络连接性并且被配置为执行辅助操作。包括语音发生器120、麦克风182或其组合的处理器190被包括在无线扬声器和语音激活设备1202中。无线扬声器和语音激活设备1202还包括扬声器1204。在操作期间，响应于接收口头命令并经由语音生成器120的操作生成合成的语音(或经由图6的语音识别网络672的操作来识别语音)，无线扬声器和语音激活设备1202可以执行辅助操作，诸如经由执行集成的辅助应用。辅助操作可以包括调整温度、播放音乐、开灯等。例如，响应于在关键词或关键短语(例如，“你好助手”)之后接收到命令来执行辅助操作。

[0127] 图13描绘了设备102包括对应于相机设备1302的便携式电子设备的实施方式1300。语音生成器120、麦克风182或其组合被包括在相机设备1302中。在操作期间，响应经由语音生成器120的操作接收口头命令并生成合成的语音(或经由图6的语音识别网络672的操作识别语音)，相机设备1302可以响应于说出的用户命令执行操作，诸如作为说明性示例的调整图像或视频捕获设置、图像或视频播放设置或图像或视频捕获指令。

[0128] 图14描绘了设备102包括对应于虚拟现实、增强现实或混合现实耳机1402的便携式电子设备的实施方式1400。语音生成器120、麦克风182或其组合被集成到耳机1402中。在特定方面，耳机1402包括麦克风182，麦克风182定位成主要捕获用户的语音。可以基于从耳机1402的麦克风182接收的音频信号来执行合成的语音生成。视觉界面设备定位在用户眼前，以便能够在佩戴耳机1402时向用户显示增强现实或虚拟现实图像或场景。在特定示例中，视觉界面设备被配置为显示指示在音频信号中检测的用户语音的通知。

[0129] 图15描绘了设备102对应于载具1502或集成在载具1502内实施方式1500，载具1502被示为有人驾驶或无人驾驶的空中设备(例如，包裹递送无人机)。语音生成器120、麦克风182或其组合被集成到载具1502中。可以基于从载具1502的麦克风182接收的音频信号执行合成的语音生成，诸如用于传递来自载具1502的授权用户的指令。

[0130] 图16描绘了设备102对应于或集成在载具1602内的另一实施方式1600，载具1602被示为汽车。载具1602包括处理器190，处理器190包括语音生成器120。载具1602还包括麦

克风182。麦克风182被定位以捕获载具1602的操作员的话语。合成的语音生成可以基于从载具1602的麦克风182接收的语音信号来执行。在一些实施方式中，可以基于从内部麦克风(例如，麦克风182)接收的音频信号来执行合成的语音生成，诸如来自授权乘客的语音命令。例如，合成的语音生成可以用于检测来自载具1602的操作员的语音命令(例如，将音量设置为5或设置自动驾驶载具的目的地)。在一些实施方式中，可以基于从诸如载具的授权用户的外部麦克风(例如，麦克风182)接收的音频信号来执行合成的语音生成。在特定实施方式中，响应于接收口头命令并经由语音生成器120的操作生成合成的语音，语音激活系统162基于在合成的语音信号131中检测的一个或多个关键词(例如，“解锁”、“开启引擎”、“播放音乐”、“播放天气预报”或其他语音命令)，诸如经由显示器1620或一个或多个扬声器(例如，扬声器1610)提供反馈或信息来发起载具1602的一个或多个操作。

[0131] 参考图17，示出了合成的语音生成方法1700的特定实施方式。在特定方面，方法1700的一个或多个操作由以下各项中的至少一个来执行：风格数据生成器126、多编码器192、解码器194、多编码器变换器128、语音生成器120、一个或多个处理器190、设备102、图1的系统100、人编码器206、情绪编码器208、语音速率编码器210、编码器前置网络212、人编码器420、情绪编码器430、速率编码器440、图4A的输入表示编码器450、图4B的特性编码器451、图5的编码器500或其组合。

[0132] 方法1700包括在1702处接收指示目标语音特性的一个或多个控制参数。例如，图1的风格数据生成器126接收指示目标语音特性141的一个或多个控制参数125，如参考图1所描述的。

[0133] 方法1700还包括在1704处使用多编码器基于一个或多个控制参数处理语音的输入表示以生成与表示基于目标语音特性的语音版本的音频信号对应的编码的数据。例如，多编码器192基于一个或多个控制参数125处理输入表示123以生成对应于表示基于目标语音特性141的语音版本的合成的语音信号131的编码的数据193，如参考图1所描述的。

[0134] 方法1700能够基于相同的输入信号121生成对应于各种情绪、各种语音速率和各种人的语音特性的多个合成的语音信号。可以使用相对较少的说出的词来生成大量且多样的合成的语音数据集。

[0135] 图17的方法1700可以通过现场可编程门阵列(FPGA)设备、专用集成电路(ASIC)、诸如中央处理器(CPU)的处理单元、DSP、控制器、另一硬件设备、固件设备或其任何组合来实现。作为示例，图17的方法1700可以由执行指令的处理器来执行，诸如参考图19所描述的。

[0136] 参考图18，示出了合成的语音生成方法1800的特定实施方式。在特定方面，方法1800的一个或多个操作由以下各项中的至少一个来执行：风格数据生成器126、多编码器192、解码器194、多编码器变换器128、语音生成器120、语音合成器130、一个或多个处理器190、设备102、图1的系统100、人编码器206、情绪编码器208、语音速率编码器210、编码器前置网络212、人编码器420、情绪编码器430、速率编码器440、图4A的输入表示编码器450、图4B的特性编码器451、图5的编码器500或其组合。

[0137] 方法1800包括在1802处接收指示目标语音特性的一个或多个控制参数。例如，图1的风格数据生成器126接收指示目标语音特性141的一个或多个控制参数125，如参考图1所描述的。

[0138] 方法1800还包括在1804处基于一个或多个控制参数生成合并的风格数据。例如，图1的风格数据生成器126基于一个或多个控制参数125生成合并的风格数据127，如参考图1所描述的。

[0139] 方法1800还包括在1806处在多编码器的第一编码器处独立于一个或多个控制参数对输入表示进行编码以生成第一编码的数据。例如，多编码器192的输入表示编码器450独立于一个或多个控制参数125对输入表示123进行编码以生成第一编码的数据458，如参考图4A所描述的。

[0140] 方法1800还包括在1808处，在多编码器的一个或多个第二编码器处基于一个或多个控制参数对输入表示进行编码以生成第二编码的数据。例如，图4A的人编码器420、情绪编码器430，以及速率编码器440基于一个或多个控制参数125对输入表示123进行编码以生成第二编码的数据(例如，人编码的数据428、情绪编码的数据438和语音速率编码的数据448)，如参考图4A所描述的。作为另一示例，图4B的特性编码器451基于一个或多个控制参数125对输入表示123进行编码以生成第二编码的数据459，如参考图4B所描述的。

[0141] 方法1800还包括在1810处，在解码器的解码器网络处基于第一编码的数据和第二编码的数据生成输出频谱数据。例如，解码器194(例如，包括多个解码器网络，诸如多头注意力478、多头注意力480、多头注意力482、图4A的多头注意力484和图4B的多头注意力485)基于编码的数据193生成输出频谱数据129，如参考图1和图4A-图4B所描述的。

[0142] 方法1800还包括在1812处在语音合成器处并基于输出频谱数据生成表示基于目标语音特性的语音版本的音频信号。例如，图1的语音合成器130基于输出频谱数据129，生成表示基于目标语音特性141的语音版本的合成语音信号131，如参考图1所描述的。

[0143] 方法1800能够基于相同的输入信号121生成对应于各种情绪、各种语音速率和各种人的语音特性的多个合成的语音信号。可以使用相对较少的说出的词来生成大量且多样的合成的语音数据集。

[0144] 图18的方法1800可以通过现场可编程门阵列(FPGA)设备、专用集成电路(ASIC)、诸如中央处理器(CPU)的处理单元、DSP、控制器、另一硬件设备、固件设备来实现，或其任何组合来实现。作为示例，图18的方法1800可以由执行指令的处理器来执行，诸如参考图19所描述的。

[0145] 参考图19，描绘了设备的特定说明性实施方式的框图并且通常指定为1900。在各种实施方式中，设备1900可具有比图19中所说明的更多或更少的组件。在说明性实施方式中，设备1900可以对应于设备102。在说明性实施方式中，设备1900可以执行参考图1-图18描述的一个或多个操作。

[0146] 在特定实施方式中，设备1900包括处理器1906(例如，中央处理单元(CPU))。设备1900可以包括一个或多个附加处理器1910(例如，一个或多个DSP)。在特定方面，图1的一个或多个处理器190对应于处理器1906、处理器1910或其组合。处理器1910可以包括语音和音乐编码器-解码器(CODEC)1908，其包括语音编码器(“声码器”)编码器1936、声码器解码器1938、语音生成器120、网络训练器670或其组合。

[0147] 设备1900可包括存储器132和CODEC1934。存储器132可包括指令1956，其可由一个或多个附加处理器1910(或处理器1906)执行以实现参考语音生成器120、网络训练器670，或两者描述的功能。设备1900可以包括经由收发器1950耦合到天线1952的调制解调器

1970。

[0148] 设备1900可以包括耦合到显示器控制器1926的显示器1928。扬声器1992、麦克风182或两者可耦合到CODEC 1934。CODEC 1934可以包括数模转换器(DAC)1902，模数转换器(ADC)1904，或两者。在特定实施方式中，CODEC 1934可以从麦克风182接收模拟信号，使用模数转换器1904将模拟信号转换成数字信号，并且将数字信号提供给语音和音乐编解码器1908。语音和音乐编解码器1908可以处理数字信号，并且数字信号还可以由语音生成器120处理。在特定实施方式中，语音和音乐编解码器1908可以向CODEC 1934提供数字信号。CODEC 1934可以使用数模转换器1902将数字信号转换为模拟信号，并且可以将模拟信号提供给扬声器1992。

[0149] 在特定实施方式中，设备1900可以被包括在系统级封装或片上系统设备1922中。在特定实施方式中，存储器132、处理器1906、处理器1910、显示控制器1926、CODEC 1934和调制解调器1970被包括在系统级封装或片上系统设备1922中。在特定实施方式中，输入设备1930和电源1944被耦合到片上系统设备1922。进一步地，在具体实施方式中，如图19所示，显示器1928、输入设备1930、扬声器1992、麦克风182、天线1952和电源1944在片上系统设备1922的外部。在特定实施方式中，显示器1928、输入设备1930、扬声器1992、麦克风182、天线1952和电源1944中的每一个可以被耦合到片上系统设备1922的组件，诸如接口或控制器。

[0150] 设备1900可以包括虚拟助手、家用电器、智能设备、物联网(IoT)设备、通信设备、耳机、载具、计算机、显示器设备、电视机、游戏机、音乐播放器、收音机、视频播放器、娱乐单元、个人媒体播放器、数字视频播放器、相机、导航设备、智能扬声器、条形音箱、移动通信设备、智能电话、蜂窝电话、笔记本电脑、平板电脑、个人数字助理、数字视频光盘(DVD)播放器、调谐器、增强现实耳机、虚拟现实耳机、飞行器、家庭自动化系统、声音激活设备、无线扬声器和语音激活设备、便携式电子设备、汽车、载具、计算设备、虚拟现实(VR)设备、基站、移动设备或其任何组合。

[0151] 结合所描述的实施方式，一种装置包括用于接收指示目标语音特性的一个或多个控制参数的部件。例如，用于接收的部件包括风格数据生成器126、语音生成器120、一个或多个处理器190、图1的设备102、收发器1950、调制解调器1970、输入设备1930、设备1900的输入接口、被配置为接收一个或多个控制参数的一个或多个其他电路或组件，或其任何组合。

[0152] 该装置还包括用于使用多编码器基于一个或多个控制参数处理语音输入表示以生成与表示基于目标语音特性的语音版本的音频信号对应的编码的数据的部件。例如，用于处理的部件包括多编码器192、多编码器变换器128、语音生成器120、一个或多个处理器190、图1的设备102、语音和音乐CODEC1908、处理器1910、处理器1906、设备1900、被配置为基于一个或多个控制参数来处理语音的输入表示的一个或多个其他电路或组件，或其任何组合。

[0153] 在一些实施方式中，非暂时性计算机可读介质(例如，计算机可读存储设备，诸如存储器132)包括指令(例如，指令1956)，当由一个或多个处理器(例如，一个或多个处理器1910或处理器1906)执行时，使一个或多个处理器接收指示目标语音特性(例如，目标语音特性141)的一个或多个控制参数(例如，一个或多个控制参数125)。当指令由一个或多个处

理器执行时，还使一个或多个处理器使用多编码器(例如，多编码器192)基于一个或多个控制参数(例如，一个或多个控制参数125)来处理语音的输入表示(例如，输入表示123)以生成对应于表示基于目标语音特性(例如，目标语音特性141)的语音版本的音频信号(例如，合成的语音信号131)的编码的数据(例如，编码的数据193)。

[0154] 下面相互关联的条款第一集合中描述了本公开的特定方面：

[0155] 根据条款1，一种用于语音生成的设备包括一个或多个处理器，被配置为：接收指示目标语音特性的一个或多个控制参数；使用多编码器，基于一个或多个控制参数处理语音的输入表示，以生成与表示基于目标语音特性的语音版本的音频信号对应的编码的数据。

[0156] 条款2包括条款1所述的设备，其中，控制参数指示要使用其语音特性的目标人、目标情绪、目标语音速率或其组合。

[0157] 条款3包括条款1或条款2所述的设备，其中，一个或多个处理器还被配置为基于一个或多个控制参数生成合并的风格数据，并且其中，合并的风格数据在处理输入表示期间由多编码器使用。

[0158] 条款4包括条款1至3中任一项所述的设备，其中，多编码器包括：被配置为独立于一个或多个控制参数对输入表示进行编码以生成第一编码的数据的第一编码器；以及被配置为基于一个或多个控制参数对输入表示进行编码以生成第二编码的数据的一个或多个第二编码器，其中编码的数据包括第一编码的数据和第二编码的数据。

[0159] 条款5包括条款4所述的设备，其中，一个或多个处理器还被配置为：在语音特性编码器处，基于一个或多个控制参数中的至少一个来处理输入表示以生成编码的输入语音表示；在编码器前置网络处，至少部分地基于编码的输入语音表示生成合并的风格数据；向第一编码器提供输入表示以生成第一编码的数据；将输入表示和合并的风格数据提供给一个或多个第二编码器以生成目标第二编码的数据。

[0160] 条款6包括条款4或条款5所述的设备，还包括包含多编码器和解码器的多编码器变换器，其中第一编码器包括第一注意力网络，其中一个或多个第二编码器中的每一个包括第二注意力网络，并且其中解码器包括解码器注意力网络。

[0161] 条款7包括条款6所述的设备，其中：第一编码器包括：包括第一注意力网络的第一层，其中，第一注意力网络对应于第一多头注意力网络；以及包括第一神经网络的第二层，并且一个或多个第二编码器中的每一个包括：包括第二注意力网络的第一层，其中，第二注意力网络对应于第二多头注意力网络；并且第二层包括第二神经网络。

[0162] 条款8包括条款4至7中任一项所述的设备，还包括：耦合到多编码器的解码器，解码器包括被配置为基于第一编码的数据和目标第二编码的数据生成输出频谱数据的解码器网络；以及语音合成器，被配置为基于输出频谱数据生成表示基于目标语音特性的语音版本的音频信号。

[0163] 条款9包括条款8所述的设备，其中，解码器网络包括解码器注意力网络，解码器注意力网络包括：被配置为处理第一编码的数据的第一多头注意力网络；被配置为处理目标第二编码的数据的一个或多个第二多头注意力网络；以及被配置为组合第一多头注意力网络和一个或多个第二多头注意力网络的输出的组合器。

[0164] 条款10包括条款9所述的设备，其中解码器还包括：被耦合到解码器注意力网络的

输入的掩码的多头注意力网络;以及被耦合到解码器注意力网络的输出的解码器神经网络。

[0165] 条款11包括条款1至10中任一项所述的设备,其中,一个或处理器还被配置为:从音频信号生成一个或多个估计的控制参数;以及基于一个或多个控制参数与一个或多个估计的控制参数的比较,训练多编码器、一个或多个语音修改特性编码器、编码器前置网络、解码器网络,或其组合的一个或多个神经网络权重。

[0166] 条款12包括条款1至11中任一项所述的设备,其中,一个或多个处理器还被配置为:接收输入语音信号;并且基于输入语音信号生成输入表示。

[0167] 条款13包括条款1至11中任一项所述的设备,其中一个或多个处理器还被配置为接收输入表示。

[0168] 条款14包括条款1至13中任一项所述的设备,其中,输入表示包括文本、梅尔标度频谱图、基频(F0)特征或其组合。

[0169] 下面相互关联的条款的第二集合中描述了本公开的特定方面:

[0170] 根据条款15,一种语音生成方法包括:在设备处接收指示目标语音特性的一个或多个控制参数;使用多编码器,基于一个或多个控制参数处理语音的输入表示,以生成与表示基于目标语音特性的语音版本的音频信号对应的编码的数据。

[0171] 条款16包括条款15所述的方法,其中,控制参数指示要使用其语音特性的目标人、目标情绪、目标语音速率或其组合。

[0172] 条款17包括条款15或条款16所述的方法,还包括在设备处基于一个或多个控制参数生成合并的风格数据,其中,合并的风格数据在输入表示的处理期间由多编码器使用。

[0173] 条款18包括条款15至17中任一项所述的方法,还包括:在多编码器的第一编码器处,独立于一个或多个控制参数对输入表示进行编码,以生成第一编码的数据;在多编码器的一个或多个第二编码器处,基于一个或多个控制参数对输入表示进行编码以生成目标第二编码的数据,其中,音频信号编码的数据包括基于第一编码的数据和目标第二编码的数据。

[0174] 条款19包括条款18所述的方法,还包括:在语音特性编码器处,基于一个或多个控制参数中的至少一个来处理输入表示,以生成编码的输入语音表示;在编码器前置网络处至少部分地基于编码的输入语音表示生成合并的风格数据;向第一编码器提供输入表示以生成第一编码的数据;将输入表示和合并的风格数据提供给一个或多个第二编码器以生成目标第二编码的数据。

[0175] 条款20包括条款18所述的方法,其中多编码器和解码器被包括在多编码器变换器中,其中第一编码器包括第一注意力网络,其中,一个或多个第二编码器中的每一个包括第二注意力网络,并且其中,解码器包括解码器注意力网络。

[0176] 条款21包括条款20所述的方法,其中:第一编码器包括:包括第一注意力网络的第一层,其中,第一注意力网络对应于第一多头注意力网络;包括第一神经网络的第二层,并且一个或多个第二编码器中的每一个包括:包括第二注意力网络的第一层,其中,第二注意力网络对应于第二多头注意力网络;以及包括第二神经网络的第二层。

[0177] 条款22包括条款18所述的方法,还包括:在解码器的解码器网络处,基于第一编码的数据和目标第二编码的数据生成输出频谱数据;以及在语音合成器处,基于输出的频谱

数据,生成表示基于目标语音特性的语音版本的音频信号。

[0178] 条款23包括条款22所述的方法,还包括:在解码器注意力网络的第一多头注意力网络处处理第一编码的数据,其中,解码器网络包括解码器注意力网络;在解码器注意力网络的一个或多个第二多头注意力网络处处理目标第二编码的数据;以及在组合器处组合第一多头注意力网络和一个或多个第二多头注意力网络的输出。

[0179] 条款24包括条款15至23中任一项所述的方法,还包括:在设备处从音频信号生成一个或多个估计的控制参数;以及基于一个或多个控制参数以及一个或多个估计的控制参数的比较,训练多编码器、一个或多个语音修改特性编码器、编码器前置网络、解码器网络,或其组合的一个或多个神经网络权重。

[0180] 条款25包括条款15至24中任一项所述的方法,还包括:在设备处接收输入语音信号;以及在设备处基于输入语音信号生成输入表示。

[0181] 条款26包括条款15至24中任一项所述的方法,还包括在设备处接收输入表示。

[0182] 下面相互关联的条款的第三集合中描述了本公开的特定方面:

[0183] 根据条款27,一种存储指令的非暂时性计算机可读存储介质,当指令由一个或多个处理器执行时,使一个或多个处理器:接收指示目标语音特性的一个或多个控制参数;以及使用多编码器,基于一个或多个控制参数处理语音的输入表示,以生成与表示基于目标语音特性的语音版本的音频信号对应的编码的数据。

[0184] 条款28包括条款27所述的非暂时性计算机可读存储介质,其中,输入表示包括文本、梅尔标度频谱图、基频(F0)特征或其组合。

[0185] 下面相互关联的条款的第四集合中描述了本公开的特定方面:

[0186] 根据条款29,一种装置包括:用于接收指示目标语音特性的一个或多个控制参数的部件;以及用于使用多编码器基于一个或多个控制参数处理语音输入表示以生成与表示基于目标语音特性的语音版本的音频信号对应的编码的数据的部件。

[0187] 条款30包括条款29所述的装置,其中用于接收的部件和用于处理的部件被集成到虚拟助手、家用电器、智能设备、物联网(IoT)设备、通信设备、耳机、载具、计算机、显示器设备、电视机、游戏机、音乐播放器、收音机、视频播放器、娱乐单元、个人媒体播放器、数字视频播放器、相机,或导航设备中的至少一个中。

[0188] 本领域技术人将进一步理解,结合本文公开的实施方式描述的各种说明性逻辑块、配置、模块、电路和算法步骤可以实现为电子硬件、由处理器执行的计算机软件或两者的组合。各种说明性组件、块、配置、模块、电路和步骤已在上文中就它们的功能进行了一般性描述。这样的功能是作为硬件还是作为处理器可执行指令来实现取决于特定的应用和对整个系统施加的设计约束。技术人可以针对每个特定应用以不同的方式实现所描述的功能,这样的实现决定不应被解释为导致偏离本公开的范围。

[0189] 结合本文所公开的实施方式描述的方法或算法的步骤可以直接体现在硬件中、由处理器执行的软件模块中或两者的组合中。软件模块可以驻留在随机存取存储器(RAM)、闪存、只读存储器(ROM)、可编程只读存储器(PROM)、可擦可编程只读存储器(EPROM)、电可擦可编程只读存储器中(EEPROM)、寄存器、硬盘、可移动磁盘、光盘只读存储器(CD-ROM)或本领域已知的任何其他形式的非瞬态存储介质。示例性存储介质耦合到处理器,使得处理器可以从存储介质读取信息和向存储介质写入信息。在可替代方案中,存储介质可以被集成

到处理器中。处理器和存储介质可以驻留在专用集成电路(ASIC)中。ASIC可以驻留在计算设备或用户终端中。在可替代方案中,处理器和存储介质可以作为分立组件驻留在计算设备或用户终端中。

[0190] 提供所公开方面的先前描述以使本领域技术人能够制造或使用所公开方面。对这些方面的各种修改对于本领域技术人来说将是清晰的,并且在不脱离本公开的范围的情况下,本文定义的原理可以应用于其他方面。因此,本公开不旨在限于本文所示的方面,而是要符合与所附权利要求所定义的原理和新颖特征一致的可能的最宽范围。

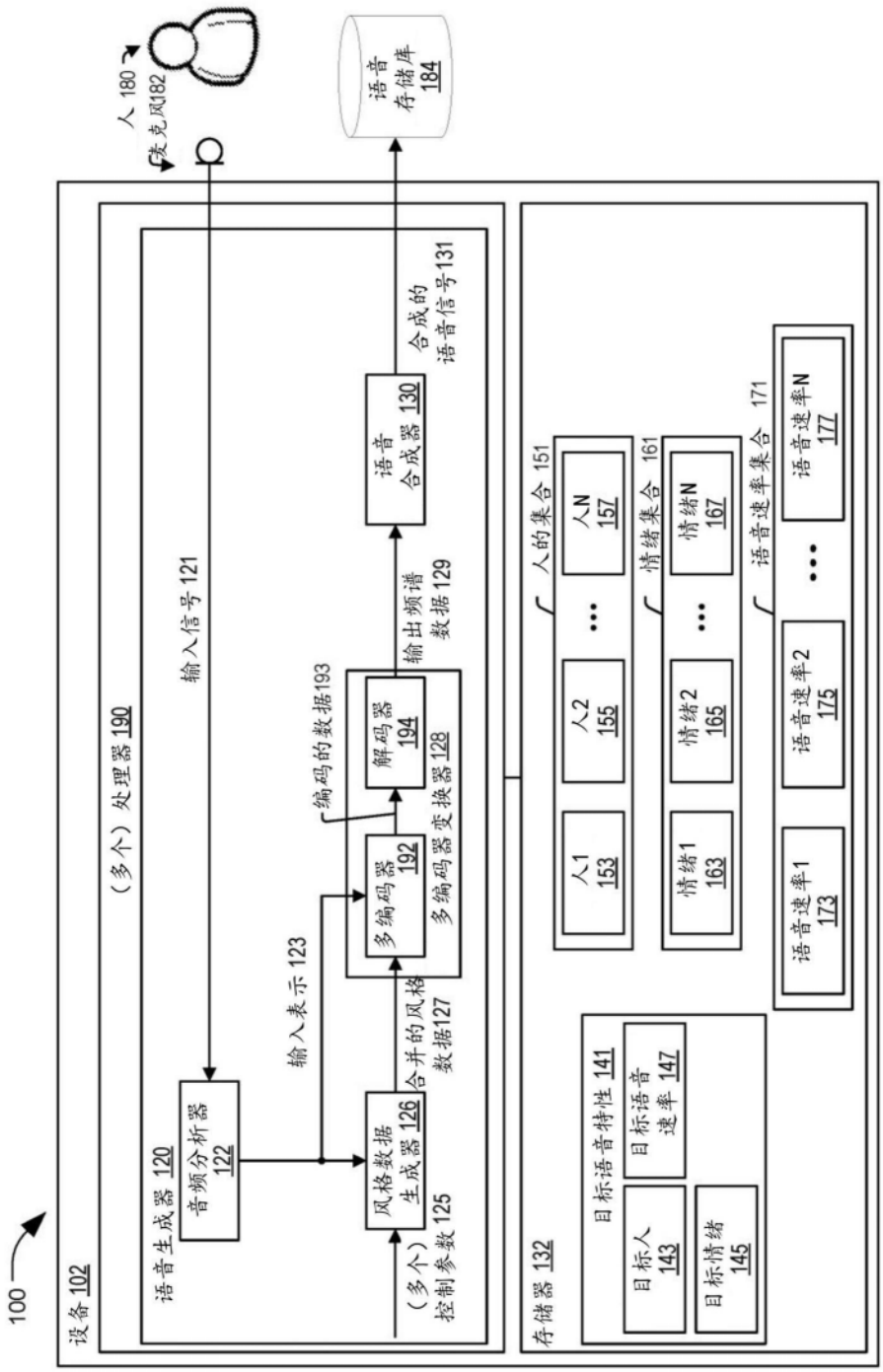

100
设备 102
人 180
麦克风182
语音存储库 184
（多个）处理器 190
语音生成器 120
音频分析器 122
输入信号 121
输入表示 123
风格数据生成器 126
（多个）控制参数 125
合并的风格数据127
多编码器 192
多编码器变换器128
编码的数据193
解码器 194
输出频谱数据 129
语音合成器 130
合成的语音信号131
存储器 132
目标语音特性 141
目标人 143
目标语音速率 147
目标情绪 145
人的集合 151
人1 153
人2 155
人N 157
情绪集合 161
情绪1 163
情绪2 165
情绪N 167
语音速率集合 171
语音速率1 173
语音速率2 175
语音速率N 177

图1

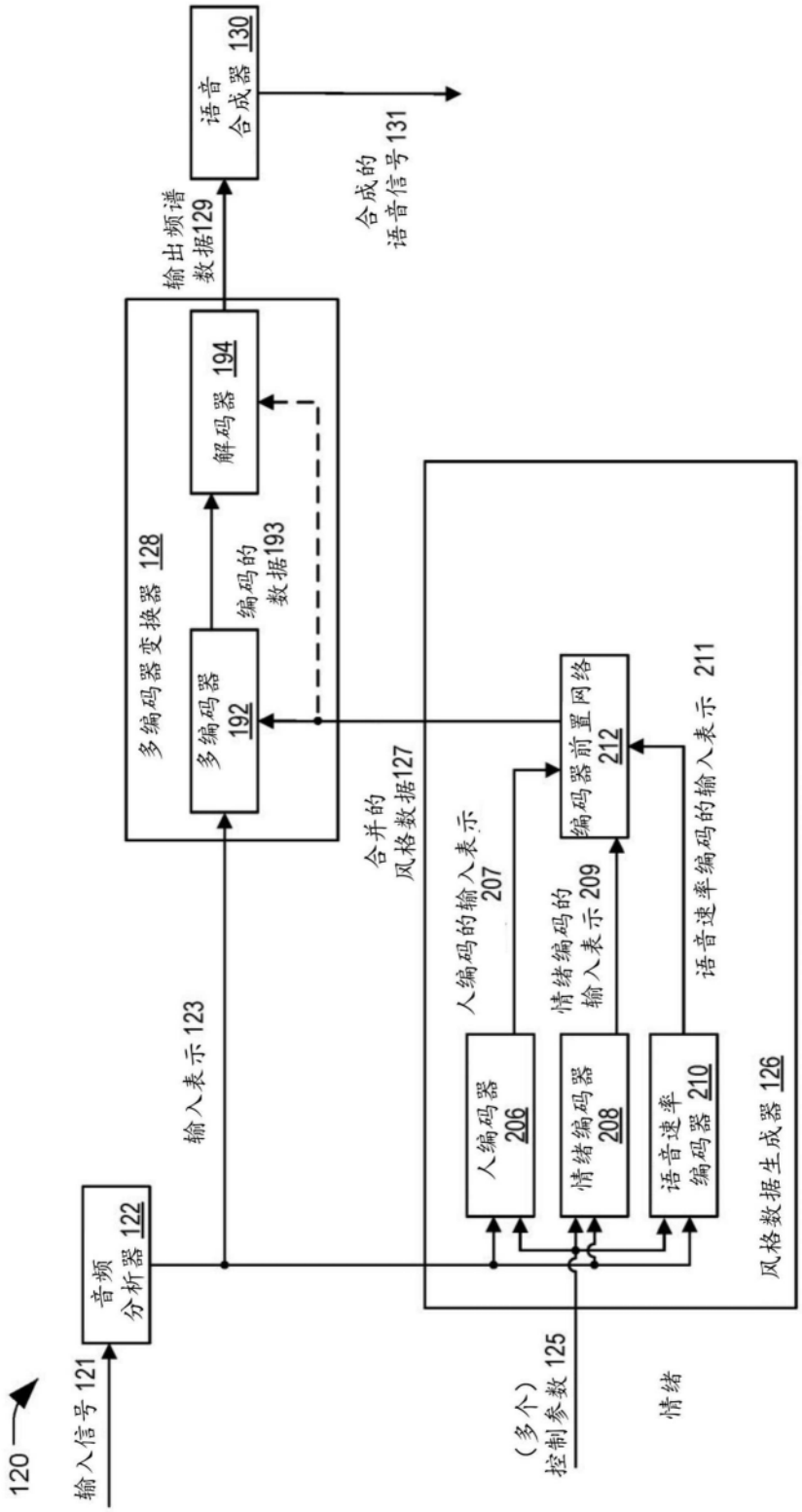

120
输入信号 121
音频分析器 122
输入表示 123
多编码器变换器 128
多编码器 192
编码的数据193
解码器 194
输出频谱数据129
语音合成器 130
合成的语音信号 131
合并的风格数据127
人编码的输入表示 207
情绪编码的输入表示 209
语音速率编码的输入表示 211
编码器前置网络 212
人编码器 206
情绪编码器 208
语音速率编码器 210
风格数据生成器 126
(多个)控制参数 125
情绪

图2

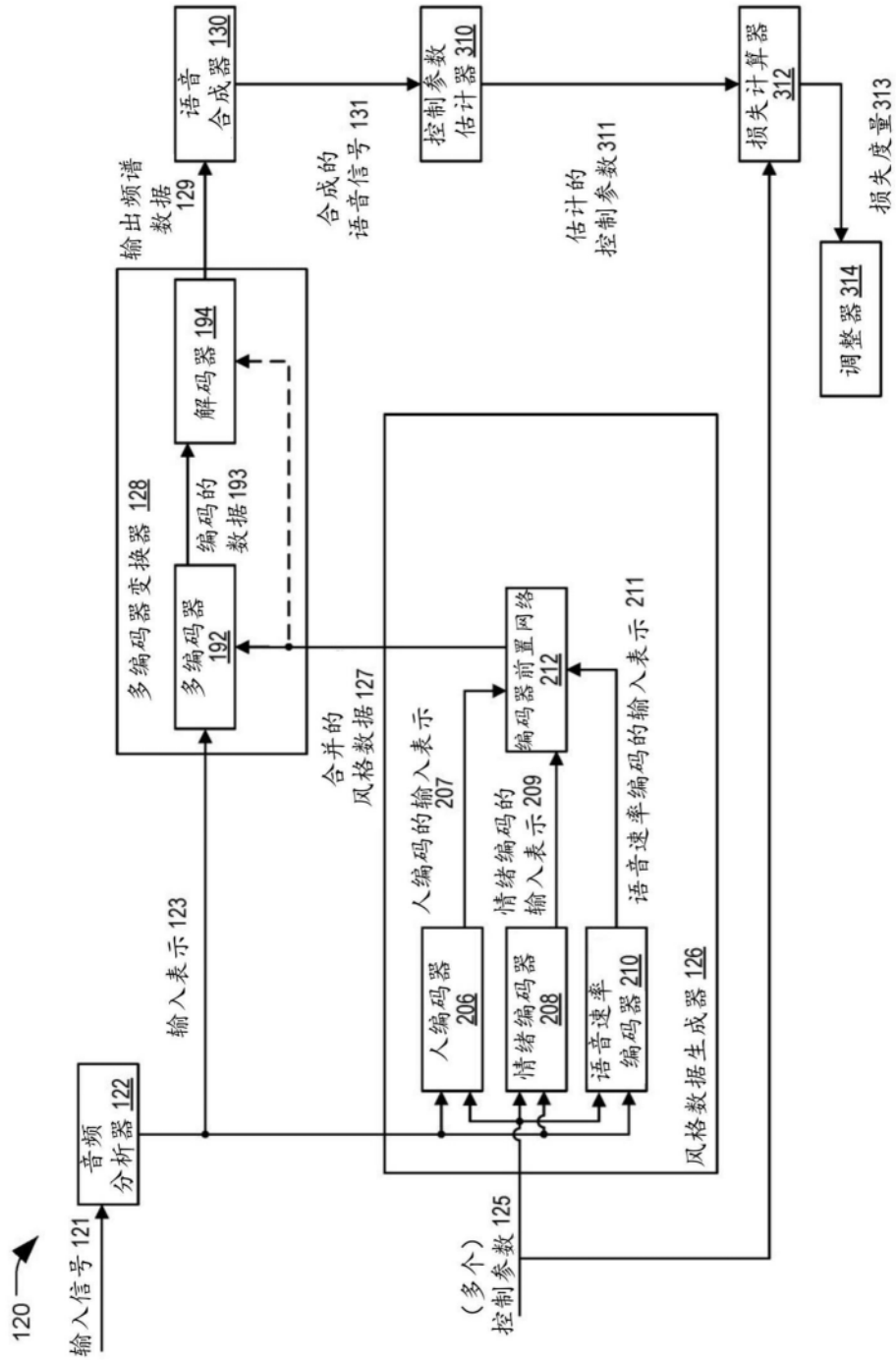

120
输入信号 121
音频分析器 122
输入表示 123
（多个）控制参数 125
风格数据生成器 126
人编码器 206
情绪编码器 208
语音速率编码器 210
人编码的输入表示 207
情绪编码的输入表示 209
语音速率编码的输入表示 211
编码器前置网络 212
合并的风格数据 127
多编码器变换器 128
多编码器 192
编码的数据 193
解码器 194
输出频谱数据 129
语音合成器 130
合成的语音信号 131
控制参数估计器 310
估计的控制参数 311
损失计算器 312
损失度量 313
调整器 314

图3

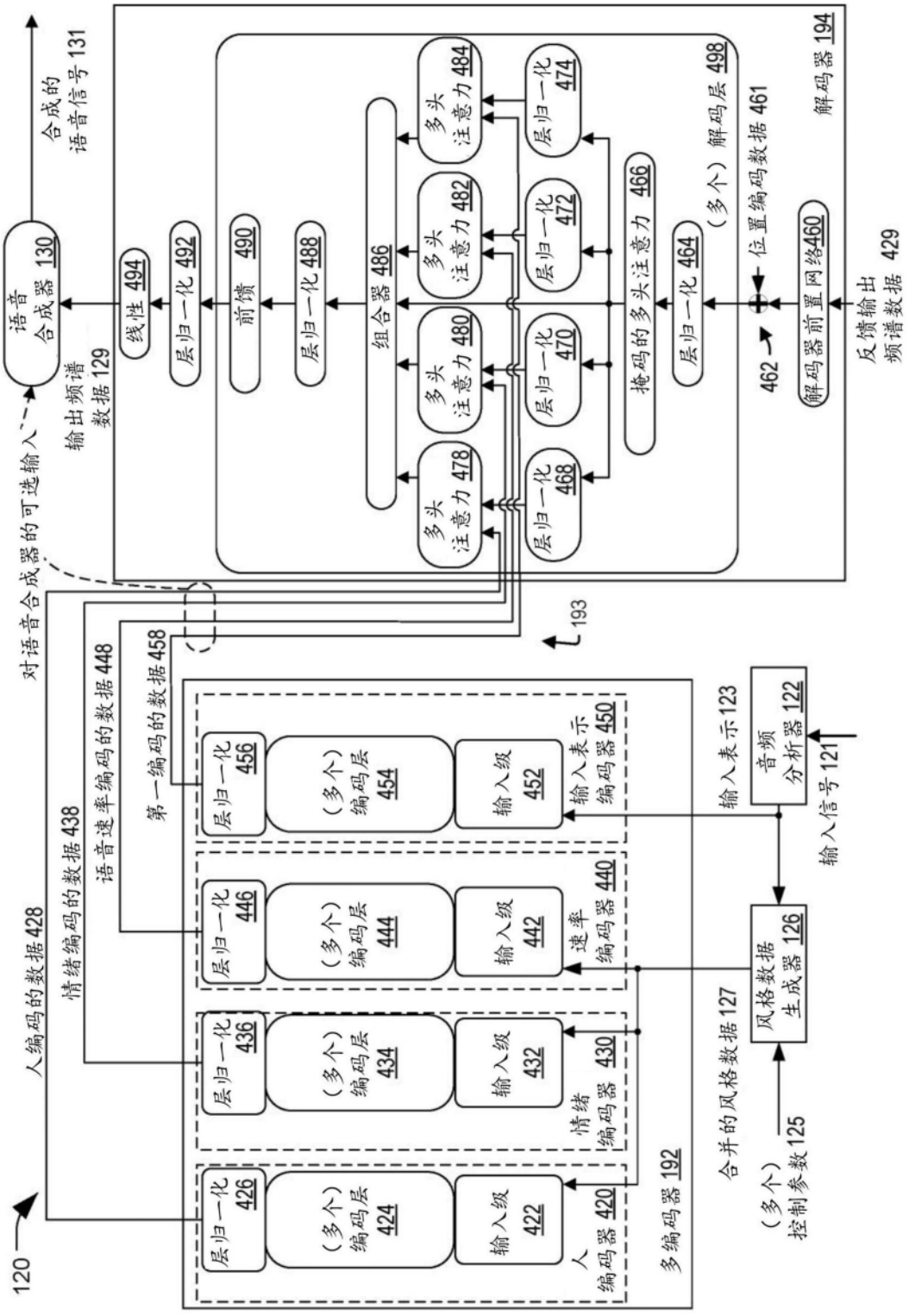

120
人编码的数据 428
对语音合成器的可选输入
情绪编码的数据 438
语音速率编码的数据 448
第一编码的数据 458
输出频谱数据 129
语音合成器 130
合成的语音信号 131
线性 494
层归一化 492
前馈 490
层归一化 488
组合器 486
多头注意力 478
多头注意力 480
多头注意力 482
多头注意力 484
层归一化 468
层归一化 470
层归一化 472
层归一化 474
掩码的多头注意力 466
层归一化 464
(多个) 解码层 498
位置编码数据 461
462
解码器前置网络 460
解码器 194
反馈输出频谱数据 429
193
层归一化 426
(多个) 编码层 424
输入级 422
人编码器 420
层归一化 436
(多个) 编码层 434
输入级 432
情绪编码器 430
层归一化 446
(多个) 编码层 444
输入级 442
速率编码器 440
层归一化 456
(多个) 编码层 454
输入级 452
输入表示编码器 450
多编码器 192
合并的风格数据 127
风格数据生成器 126
输入表示 123
音频分析器 122
输入信号 121
(多个) 控制参数 125

图4A

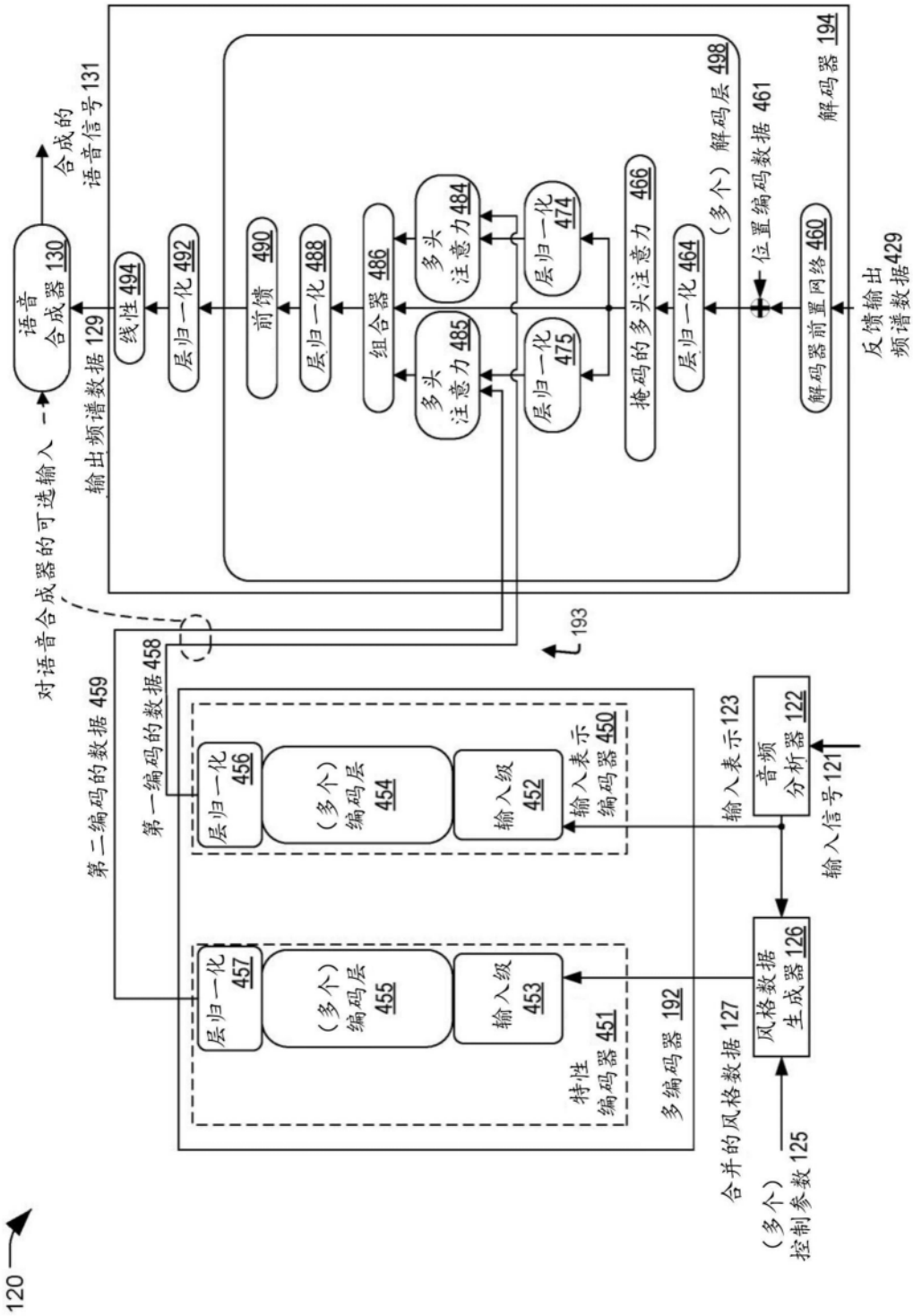

120
(多个)
控制参数 125
合并的风格数据 127
多编码器 192
特性
编码器 451
层归一化
457
(多个)
编码层
455
输入级
453
第二编码的数据 459
第一编码的数据 458
输入表示
编码器 450
层归一化
456
(多个)
编码层
454
输入级
452
对语音合成器的可选输入
193
风格数据
生成器 126
输入表示 123
音频
分析器 122
输入信号 121
语音
合成器 130
合成的
语音信号 131
输出频谱数据 129
线性 494
层归一化 492
前馈 490
层归一化 488
组合器 486
多头
注意力 485
多头
注意力 484
层归一化
475
层归一化
474
掩码的多头注意力 466
层归一化 464
(多个) 解码层 498
位置编码数据 461
解码器前置网络 460
解码器 194
反馈输出
频谱数据 429

图4B

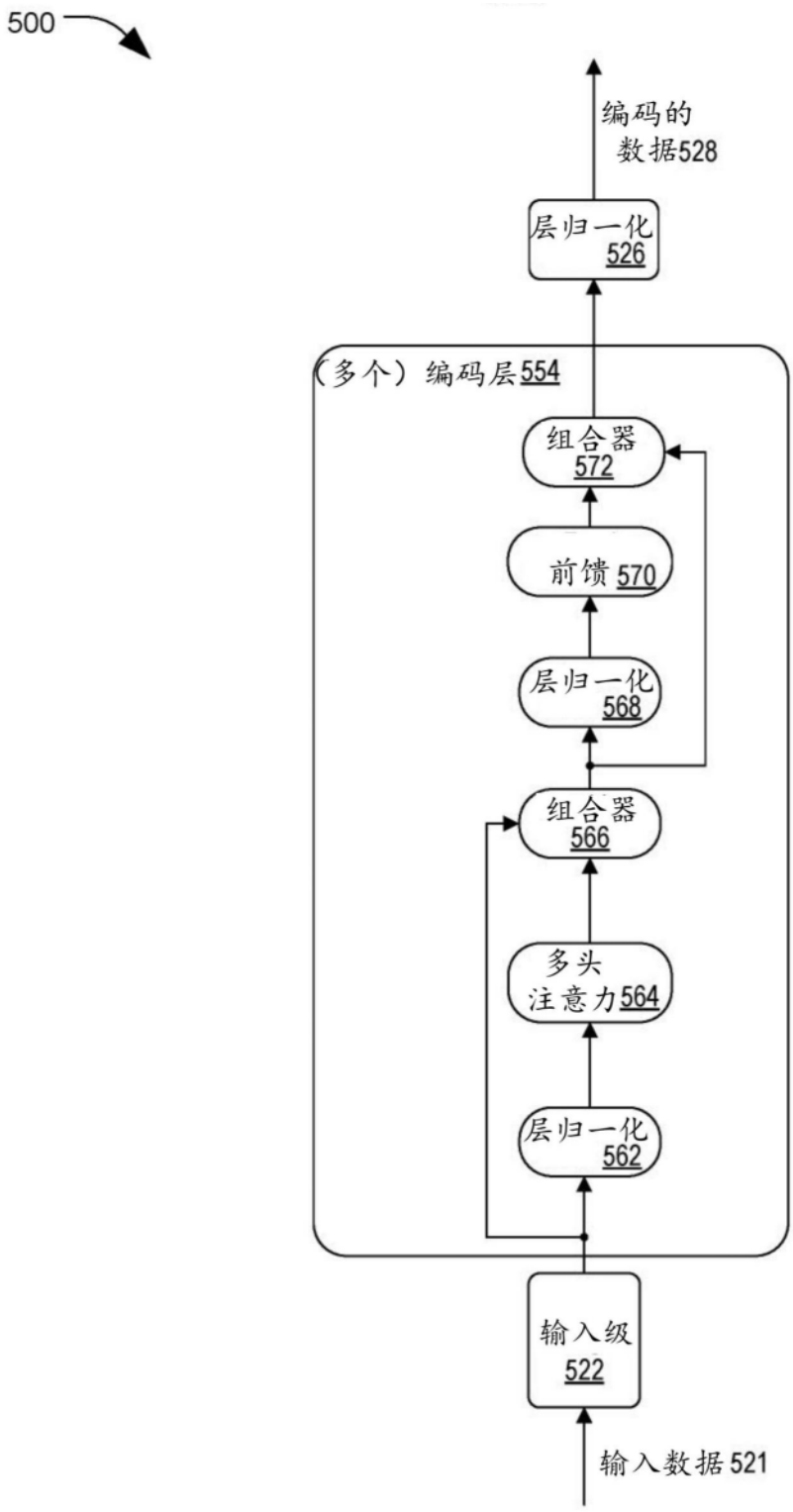

500
编码的
数据528
层归一化
526
(多个)编码层554
组合器
572
前馈570
层归一化
568
组合器
566
多头
注意力564
层归一化
562
输入级
522
输入数据521

图5

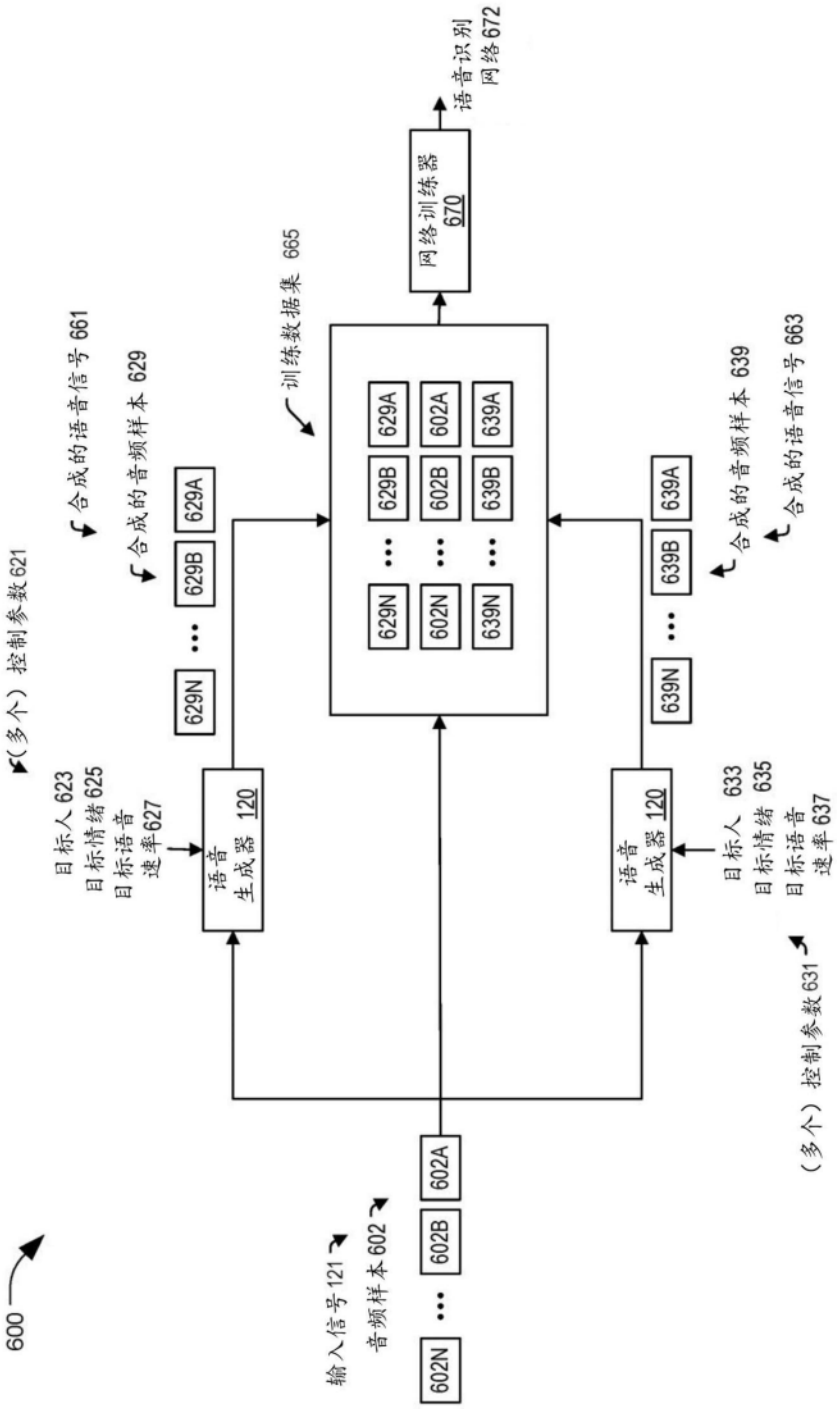

600
输入信号 121
音频样本 602
602N
602B
602A
（多个）控制参数 621
目标人 623
目标情绪 625
目标语音速率 627
语音生成器 120
629N
629B
629A
合成的音频样本 629
合成的语音信号 661
训练数据集 665
629N
629B
629A
602N
602B
602A
639N
639B
639A
网络训练器 670
语音识别网络 672
语音生成器 120
639N
639B
639A
合成的音频样本 639
合成的语音信号 663
目标人 633
目标情绪 635
目标语音速率 637
（多个）控制参数 631

图6

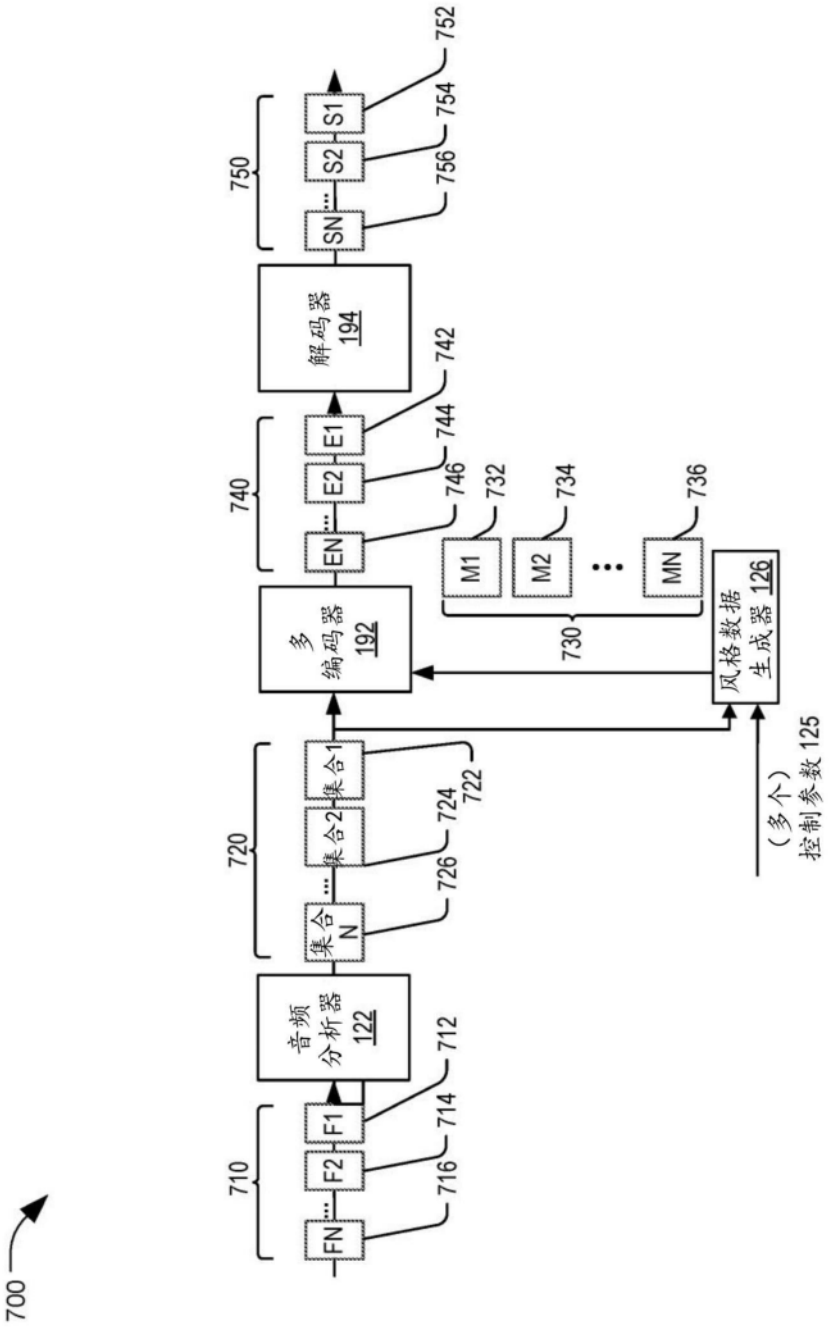

700
710
FN
F2
F1
712
714
716
音频分析器 122
720
集合 N
集合2
集合1
722
724
726
多编码器 192
740
EN
E2
E1
742
744
746
解码器 194
750
SN
S2
S1
752
754
756
732
734
736
M1
M2
MN
730
风格数据生成器 126
（多个）控制参数 125

图7

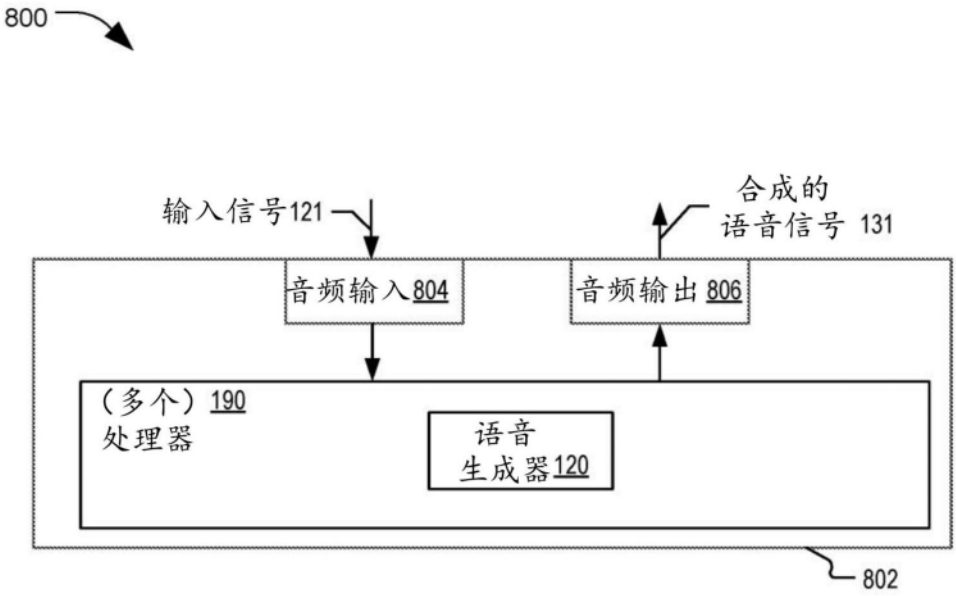

800
输入信号121
合成的
语音信号 131
音频输入804
音频输出 806
(多个) 190
处理器
语音
生成器120
802

图8

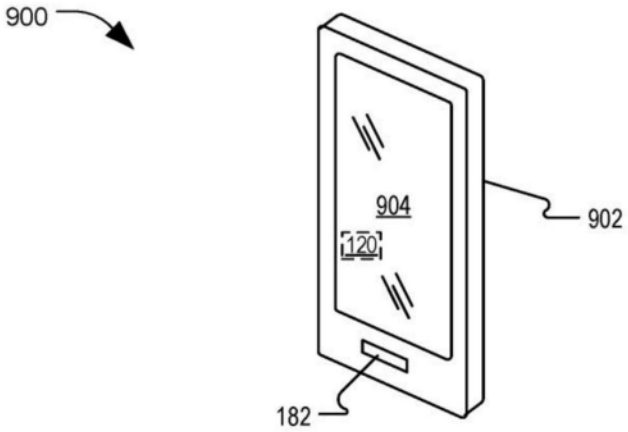

900
904
120
902
182

图9

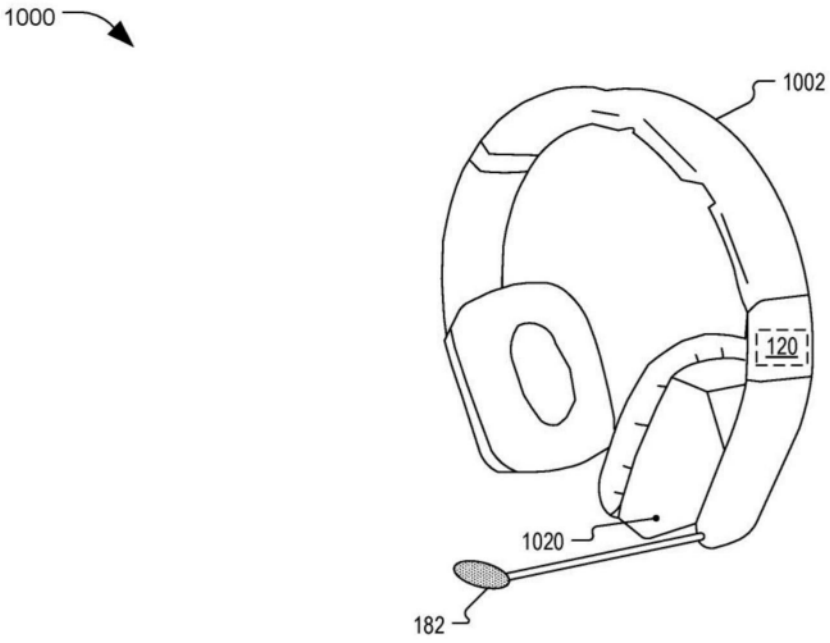


图10

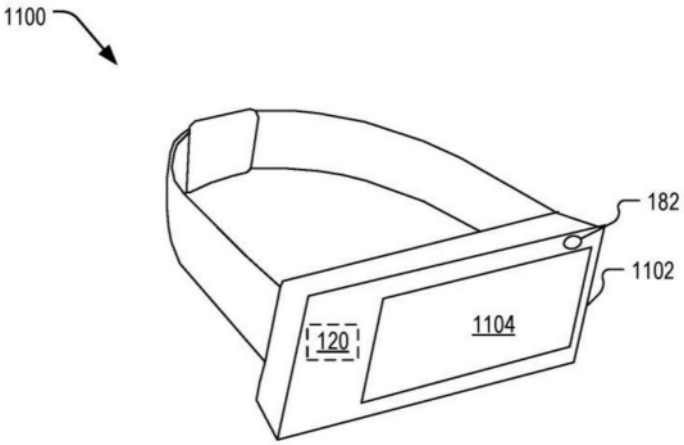


图11

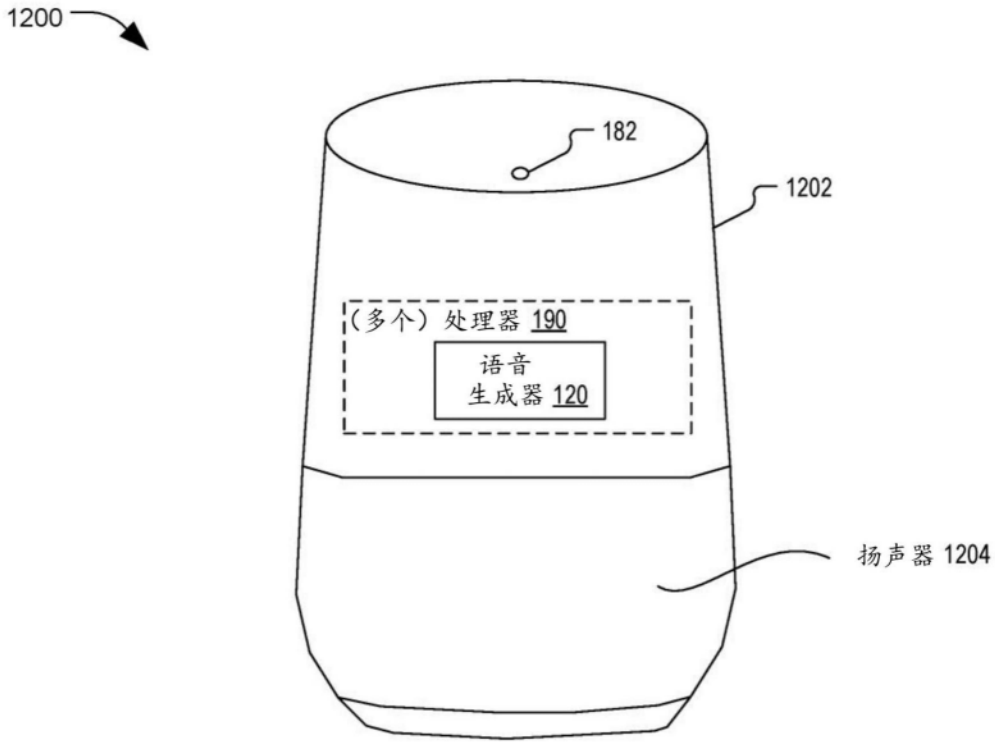

1200
182
1202
(多个) 处理器 190
语音
生成器 120
扬声器 1204

图12

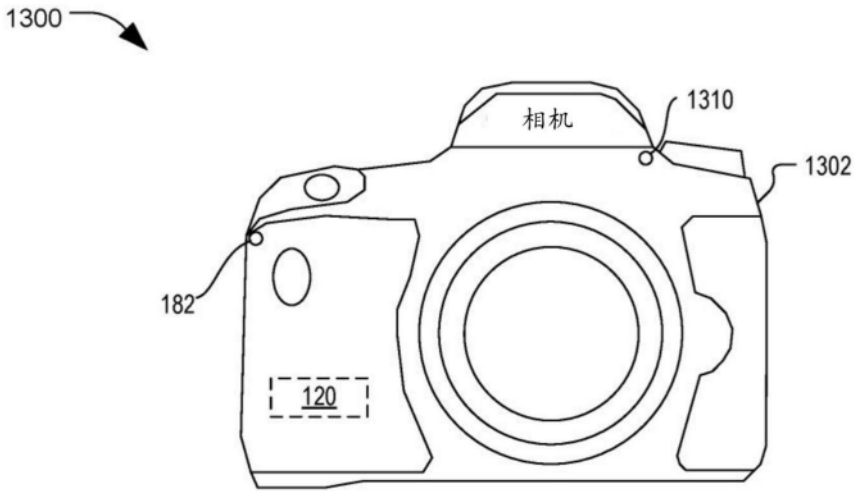

1300
相机
1310
1302
182
120

图13

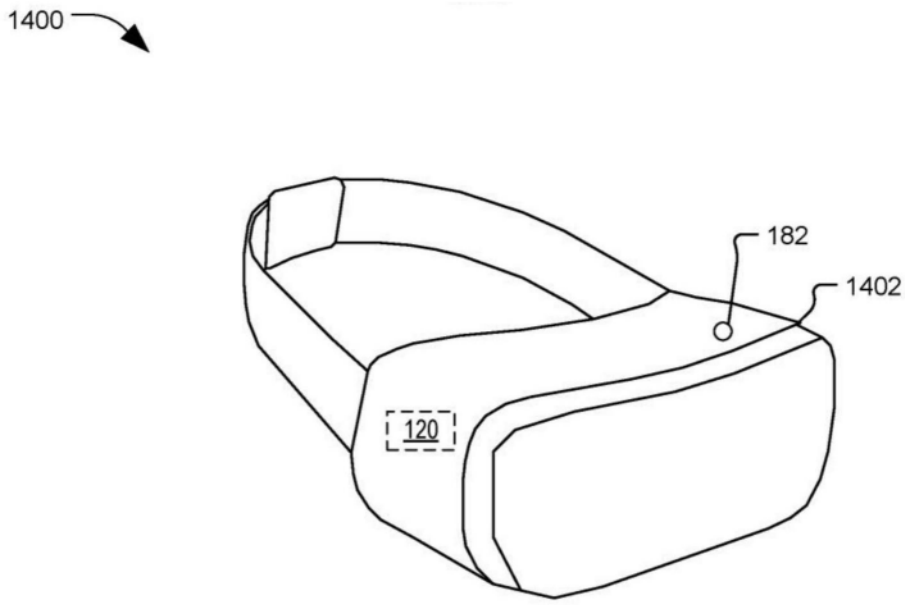


图14

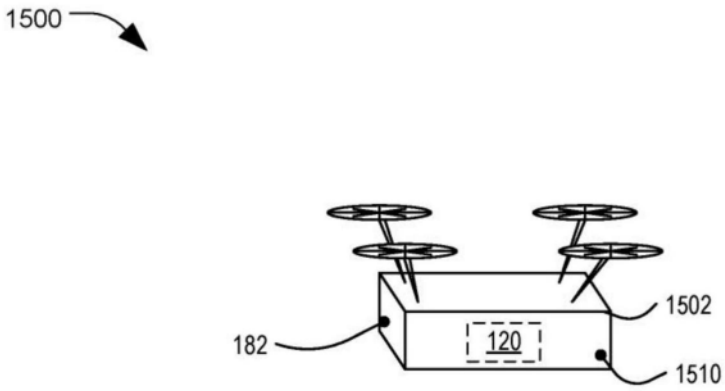


图15

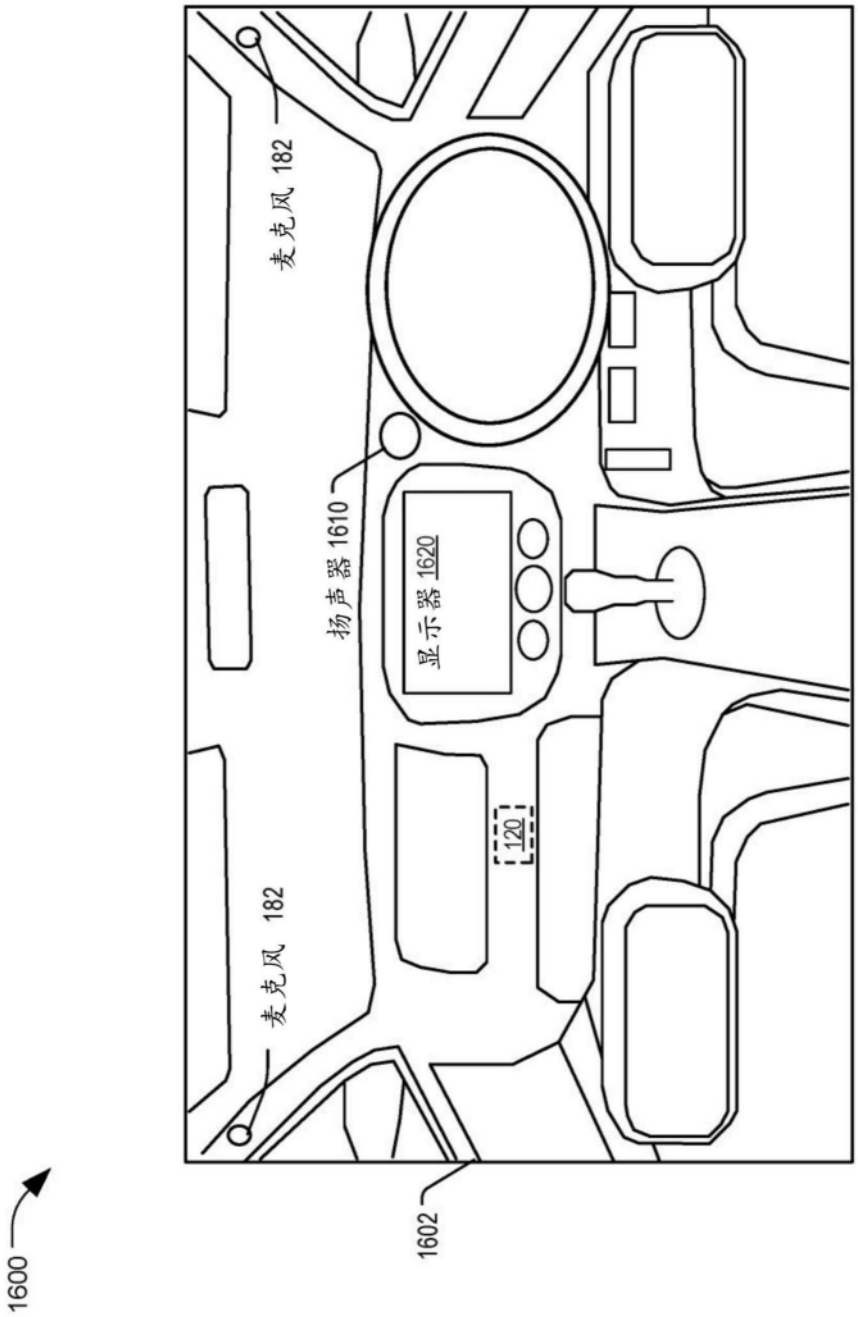


图16

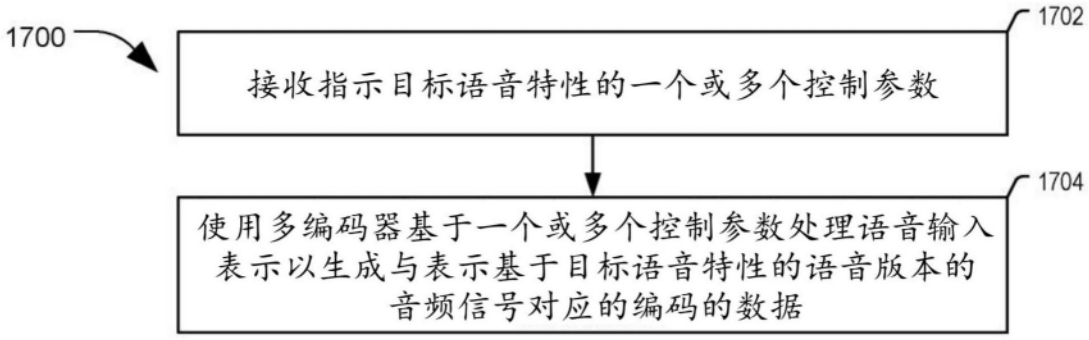

1700
1702
接收指示目标语音特性的一个或多个控制参数
1704
使用多编码器基于一个或多个控制参数处理语音输入表示以生成与表示基于目标语音特性的语音版本的音频信号对应的编码的数据

图17

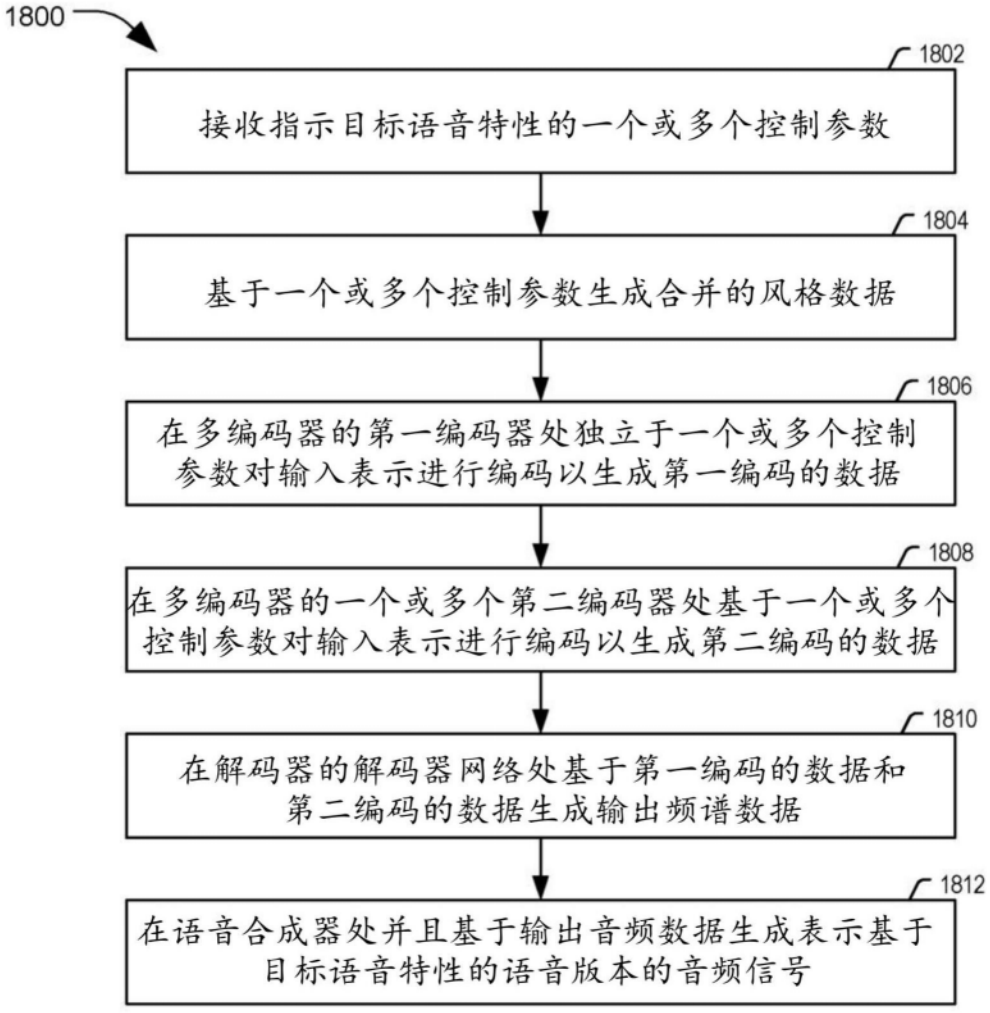

1800
1802
接收指示目标语音特性的一个或多个控制参数
1804
基于一个或多个控制参数生成合并的风格数据
1806
在多编码器的第一编码器处独立于一个或多个控制参数对输入表示进行编码以生成第一编码的数据
1808
在多编码器的一个或多个第二编码器处基于一个或多个控制参数对输入表示进行编码以生成第二编码的数据
1810
在解码器的解码器网络处基于第一编码的数据和第二编码的数据生成输出频谱数据
1812
在语音合成器处并且基于输出音频数据生成表示基于目标语音特性的语音版本的音频信号

图18

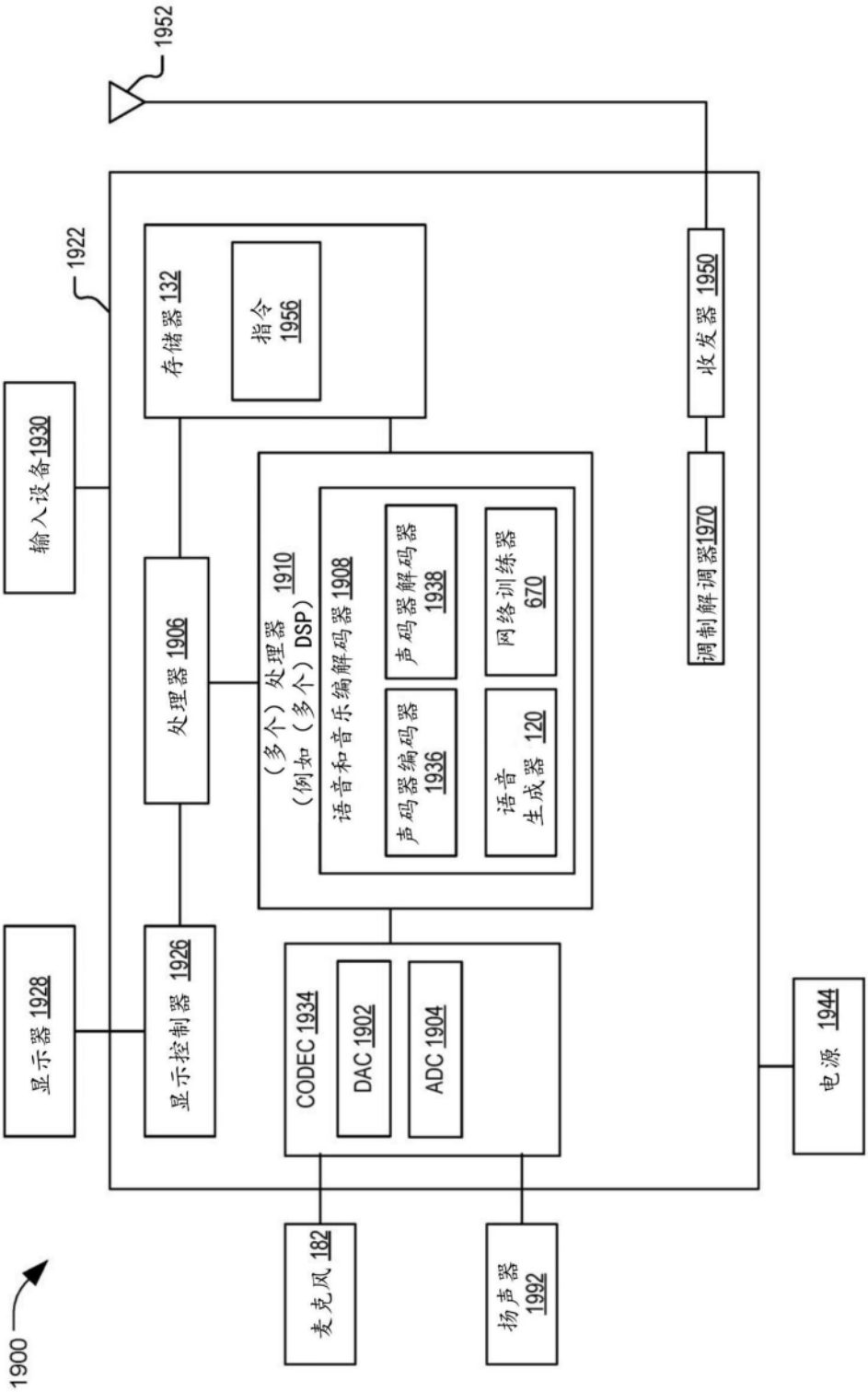

1952
1922
存储器 132
指令 1956
输入设备 1930
处理器 1906
(多个)处理器 1910
(例如(多个)DSP)
语音和音乐编解码器 1908
声码器解码器 1938
网络训练器 670
声码器编码器 1936
语音生成器 120
收发器 1950
调制解调器 1970
显示器 1928
显示控制器 1926
CODEC 1934
DAC 1902
ADC 1904
电源 1944
麦克风 182
扬声器 1992
1900

图19