(12) 按照专利合作条约所公布的国际申请

更正本

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2024 年 3 月 21 日 (21.03.2024)

WIPO | PCT

(10) 国际公布号
WO 2024/055752 A9

(51) 国际专利分类号:
*G10L 13/02* (2013.01)

(21) 国际申请号: PCT/CN2023/108845

(22) 国际申请日: 2023 年 7 月 24 日 (24.07.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202211121568.X 2022年9月15日 (15.09.2022) CN

(71) 申请人: 腾 讯 科 技 （深 圳） 有 限 公 司 **(TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED)** [CN/CN]; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(72) 发明人:宋堃**(SONG, Kun)**; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 杨兵**(YANG, Bing)**; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 张 雄 **(ZHANG, Xiong)**; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(74) 代理人:深圳市深佳知识产权代理事务所（普通合伙）**(SHENPAT INTELLECTUAL PROPERTY AGENCY)**; 中国广东省深圳市罗湖区南湖街

(54) **Title:** SPEECH SYNTHESIS MODEL TRAINING METHOD, SPEECH SYNTHESIS METHOD, AND RELATED APPARATUSES

(54) 发明名称: 语音合成模型的训练方法、语音合成方法和相关装置

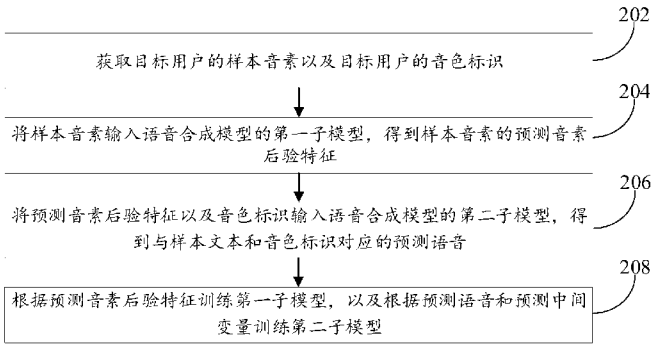



图 2

202 Acquire a sample phoneme of a target user and a timbre identifier of the target user

204 Input the sample phoneme into a first sub-model of a speech synthesis model to obtain a predicted phoneme posterior feature of the sample phoneme

206 Input the predicted phoneme posterior feature and the timbre identifier into a second sub-model of the speech synthesis model to obtain a predicted speech corresponding to a sample text and the timbre identifier

208 Train the first sub-model according to the predicted phoneme posterior feature, and train the second sub-model according to the predicted speech and a predicted intermediate variable

(57) **Abstract:** A speech synthesis model training method and apparatus, a speech synthesis method and apparatus, and a device. The speech synthesis model training method comprises: acquiring a sample phoneme of a target user and a timbre identifier of the target user (202); inputting the sample phoneme into a first sub-model of a speech synthesis model to obtain a predicted phoneme posterior feature of the sample phoneme (204); inputting the predicted phoneme posterior feature and the timbre identifier into a second sub-model of the speech synthesis model to obtain a predicted speech corresponding to a sample text and the timbre identifier, wherein the second

WO 2024/055752 A9

[见续页]

道春风路庐山大厦B座18C2、18D、18E、18E2, Guangdong 518001 (CN)。

**(81)** 指定国(除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

**(84)** 指定国(除另有指明，要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:
— 包括国际检索报告(条约第21条(3))。

**(48)** 更正本的公布日:
2024 年 9 月 12 日 (12.09.2024)

**(15)** 更正内容:
见 2024 年 9 月 12 日 (12.09.2024) 公布的公告

sub-model obtains the predicted speech on the basis of the inverse Fourier transform and by predicting the predicted phoneme posterior feature and a predicted intermediate variable of the predicted speech (206); and training the first sub-model according to the predicted phoneme posterior feature, and training the second sub-model according to the predicted speech and the predicted intermediate variable (208). The method can reduce computing resource consumption of the model.

**(57)** 摘要: 一种语音合成模型的训练方法、语音合成方法、装置及设备，该方法包括：获取目标用户的样本音素以及目标用户的音色标识（202）；将样本音素输入语音合成模型的第一子模型，得到样本音素的预测音素后验特征（204）；将预测音素后验特征以及音色标识输入语音合成模型的第二子模型，得到与样本文本和音色标识对应的预测语音，第二子模型是通过预测预测音素后验特征与预测语音的预测中间变量，基于逆傅里叶变换得到预测语音的（206）；根据预测音素后验特征训练第一子模型；以及根据预测语音和预测中间变量训练第二子模型（208）。该方法能够减少模型的计算消耗资源。

# 说明书

# 发明名称: 语音合成模型的训练方法、语音合成方法和相关装置

[0001] 本申请要求于2022年09月15日提交中国专利局、申请号为202211121568.X、申请名称为“语音合成模型的训练方法、语音合成方法、装置及设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

## 技术领域

[0002] 本申请涉及语音合成领域，特别涉及语音合成。

## 背景技术

[0003] 语音合成是指根据用户针对部分文字录制的语音，来合成符合该用户音色的其它文字的语音。

[0004] 相关技术中，通常会预先训练一个多用户声学模型和一个声码器，声学模型用于将文本转化为符合某个用户音色的频谱特征，声码器用于将频谱特征转化为语音信号。其中，声学模型中的编码器用于建模文本信息，声学模型中的解码器用于建模声学信息。通过使用目标用户的录音，可在编码器的输入上引入该目标用户的信息，从而在声学模型上进行微调，进而可得到符合该目标用户音色并与文本对应的频谱特征。之后，通过声码器基于上采样结构可合成语音信号，由此得到符合该目标用户音色并与文本对应的合成语音。示例地，上述声学模型是快速语音(Fastspeech)模型，声码器是高保真生成对抗网络(HifiGAN)。

[0005] 通过上述模型进行语音合成，由于模型参数较多，计算复杂度较大，在低计算资源场景如在终端中合成语音的情况下，存在计算消耗资源较多导致模型难以部署的情况。

[0006] 发明内容

[0007] 本申请提供了一种语音合成模型的训练方法、语音合成方法、装置及设备，可以减少模型的计算消耗资源，实现在低计算资源设备中部署模型。所述技术方案如下：

[0008] 根据本申请的一方面，提供了一种语音合成模型的训练方法，所述方法由计算机设备执行，所述方法包括：

[0009] 获取目标用户的样本音素以及所述目标用户的音色标识，所述样本音素基于所述目标用户的样本语音对应的样本文本确定，所述音色标识用于标识所述目标用户的音色；

[0010] 将所述样本音素输入所述语音合成模型的第一子模型，得到所述样本音素的预测音素后验特征，所述预测音素后验特征用于反映所述样本音素中各音素的特征以及所述样本音素中各音素的发音时长特征；

[0011] 将所述预测音素后验特征以及所述音色标识输入所述语音合成模型的第二子模型，得到与所述样本文本和所述音色标识对应的预测语音，所述第二子模型是通过预测所述预测音素后验特征与所述预测语音的预测中间变量得到所述预测语音的，所述预测中间变量用于反映所述预测语音的频域特征；

[0012] 根据所述预测音素后验特征训练所述第一子模型；以及根据所述预测语音和所述预测中间变量训练所述第二子模型。

[0013] 根据本申请的另一方面，提供了一种语音合成方法，所述方法由计算机设备执行，所述计算机设备中包括通过如上方面所述的方法训练得到的所述语音合成模型，所述方法包括：

[0014] 获取目标用户的目标音素以及所述目标用户的音色标识，所述目标音素基于目标文本确定，所述音色标识用于标识所述目标用户的音色；

[0015] 将所述目标音素输入所述语音合成模型的第一子模型，得到所述目标音素的音素后验特征；

[0016] 将所述音素后验特征以及所述音色标识输入所述语音合成模型的第二子模型，得到与所述目标文本和所述音色标识对应的目标语音。

[0017] 根据本申请的另一方面，提供了一种语音合成模型的训练装置，所述装置包括：

[0018] 获取模块，用于获取目标用户的样本音素以及所述目标用户的音色标识，所述样本音素基于所述目标用户的样本语音对应的样本文本确定，所述音色标识用于标识所述目标用户的音色；

[0019] 输入输出模块，用于将所述样本音素输入所述语音合成模型的第一子模型，得到所述样本音素的预测音素后验特征，所述预测音素后验特征用于反映所述样本音素中各音素的特征以及所述样本音素中各音素的发音时长特征；

[0020] 所述输入输出模块，还用于将所述预测音素后验特征以及所述音色标识输入所述语音合成模型的第二子模型，得到与所述样本文本和所述音色标识对应的预测语音，所述第二子模型是通过预测所述预测音素后验特征与所述预测语音的预测中间变量得到所述预测语音的，所述预测中间变量用于反映所述预测语音的频域特征；

[0021] 训练模块，用于根据所述预测音素后验特征训练所述第一子模型；以及根据所述预测语音和所述预测中间变量训练所述第二子模型。

[0022] 根据本申请的另一方面，提供了一种语音合成装置，所述装置中包括通过如上方面所述的装置训练得到的所述语音合成模型，所述装置包括：

[0023] 获取模块，用于获取目标用户的目标音素以及所述目标用户的音色标识，所述目标音素基于目标文本确定，所述音色标识用于标识所述目标用户的音色；

[0024] 输入输出模块，用于将所述目标音素输入所述语音合成模型的第一子模型，得到所述目标音素的音素后验特征；

[0025] 所述输入输出模块，还用于将所述音素后验特征以及所述音色标识输入所述语音合成模型的第二子模型，得到与所述目标文本和所述音色标识对应的目标语音。

[0026] 根据本申请的另一方面，提供了一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由所述处理器加载并执行以实现如上方面所述的语音合成模型的训练方法或语音合成方法。

[0027] 根据本申请的另一方面，提供了一种计算机可读存储介质，所述可读存储介质中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由处理器加载并执行以实现如上方面所述的语音合成模型的训练方法或语音合成方法。

[0028] 根据本申请的另一方面，提供了一种计算机程序产品，该计算机程序产品包括计算机程序，当其在计算机上运行时，使得所述计算机执行上述方面所述的语音合成模型的训练方法或语音合成方法。

[0029] 本申请提供的技术方案带来的有益效果至少包括：

[0030] 通过训练上述语音合成模型，能够通过语音合成模型根据目标用户的音色标识和目标文本，来生成符合目标用户音色的目标语音。合成目标语音的过程，是通过预测的音素后验特征对中间变量进行预测实现的。由于音素后验特征相较于频谱特征包含的信息量较少，因此预测音素后验特征所需的模型参数也较少，能够减少模型的参数，从而减少模型的计算消耗资源，可实现在低计算资源设备中部署模型。

## 附图说明

[0031] 图1是本申请一个示例性实施例提供的语音合成模型的结构示意图；

[0032] 图2是本申请一个示例性实施例提供的语音合成模型的训练方法的流程示意图；

[0033] 图3是本申请一个示例性实施例提供的语音合成模型的训练方法的流程示意图；

[0034] 图4是本申请一个示例性实施例提供的文本编码器的结构示意图；

[0035] 图5是本申请一个示例性实施例提供的一种逆傅里叶变换解码器的结构示意图；

[0036] 图6是本申请一个示例性实施例提供的另一种逆傅里叶变换解码器的结构示意图；

[0037] 图7是本申请一个示例性实施例提供的正则化流层的结构示意图；

[0038] 图8是本申请一个示例性实施例提供的多尺度频谱判别器的结构示意图；

[0039] 图9是本申请一个示例性实施例提供的语音合成方法的流程示意图；

[0040] 图10是本申请一个示例性实施例提供的语音合成模型的训练装置的结构示意图；

[0041] 图11是本申请一个示例性实施例提供的语音合成装置的结构示意图；

[0042] 图12是本申请一个示例性实施例提供的计算机设备的结构示意图。

# 具体实施方式

[0043] 为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

[0044] 首先，对本申请实施例涉及的相关名词进行介绍：

[0045] 人工智能(Artificial Intelligence，AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个综合技术，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。

[0046] 人工智能技术是一门综合学科，涉及领域广泛，既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0047] 语音技术(Speech Technology)的关键技术有自动语音识别技术(Automatic Speech Recognition，ASR)和从文本到语音(Text To Speech，TTS)的语音合成技术，以及声纹识别技术。让计算机能听、能看、能说、能感觉，是未来人机交互的发展方向，其中语音成为未来最被看好的人机交互方式之一。

[0048] 音素(phone)是根据语音的自然属性划分出来的最小语音单位，依据音节里的发音动作来分析，一个动作构成一个音素。音素可分为元音与辅音两大类。示例地，汉语音节啊(ā)只有一个音素，爱(ài)有两个音素。

[0049] 图1是本申请一个示例性实施例提供的语音合成模型的结构示意图。如图1所示，语音合成模型包括第一子模型101和第二子模型102。其中，第一子模型可称为文本至音素后特征(Text2PPG)模型，第二子模型可称为音素后特征至语音(PPG2Wav)模型。

[0050] 模型训练阶段：

[0051] 计算机设备获取目标用户的样本语音、目标用户的音色标识、样本语音对应的样本文本以及风格标识，并根据样本文本确定组成样本文本的样本音素，以及根据样本语音确定样本音素中各音素的真实发音时长特征，以及根据样本语音确定样本音素的真实音素后验特征(Phonetic Posterior Gram，PPG)。之后将样本音素和风格标识输入第一子模型101，通过文本编码器1011根据风格标识对样本音素进行编码，得到与风格标识对应的样本音素的隐层特征。以及，通过时长预测器1012对样本音素的隐层特征进行预测，得到与风格标识对应的样本音素中各音素对应的预测发音时长特征。以及，通过基频预测器1013对样本音素的隐层特征进行预测，得到与风格标识和样本音素对应的基频特征。其中，不同的风格标识具有与其对应的模型参数。之后，计算机设备通过时长规整器1014根据样本音素中各音素对应的真实发音时长特征，对样本音素的隐层特征进行扩帧处理，并通过后处理网络1015对扩帧后的样本音素的隐层特征进行卷积处理，从而得到样本音素的预测音素后验特征。该预测音素后验特征用于反映样本音素中各音素的特征以及样本音素中各音素的发音时长特征。在训练第一子模型101时，计算机设备计算预测音素后验特征与真实音素后验特征的损失函数，训练第一子模型。以及，计算预测发音时长特征和真实发音时长特征的损失函数，训练第一子模型。

[0052] 之后，计算机设备将预测音素后验特征、音色标识和基频特征输入第二子模型102，通过先验编码器中的音素后验编码器1021(也称PPG编码器)，对预测中间变量的先验分布采样均值与方差，得到预测中间变量。该预测中间变量是通过预测音素后验特征合成预测语音的过程中的中间变量，预测中间变量用于反映预测语音的频域特征。计算机设备还会将样本语音(线性谱)输入后验编码器的后验预测器1024，对真实中间变量的后验分布采样均值与方差，从而得到真实中间变量。并通过先验编码器的正则化流层1022对真实中间变量进行仿射耦合处理，从而得到处理后的真实中间变量。之后计算预测中间变量与处理后的真实中间变量的相对熵损失(也称库尔巴克·莱布勒(KL)散度损失)，从而训练第二子模型。另外，先验编码器还包括音素后验预测器1023(PPG预测器)，音素后验预测器1023用于在预训练第二子模型的过程中，根据预训练过程中的预测中间变

量，预测预训练过程中的预测音素后验特征，从而使得计算机设备可在预训练第二子模型的过程中，计算预训练过程中的预测音素后验特征，与预训练过程中的真实音素后验特征的损失函数，从而对第二子模型进行预训练。在得到预测中间变量后，计算机设备通过解码器的逆傅里叶变换解码器1025根据音色标识对预测中间变量进行逆傅里叶变换，从而得到预测语音。之后计算预测语音与样本语音之间的梅尔频谱损失，训练第二子模型。另外，图1所示的模型结构中，解码器中的判别器1026与语音合成模型中除判别器1026以外的部分能够组成生成对抗网络。计算机设备将预测语音输入判别器1026，能够得到预测语音的判别结果。之后根据判别结果与预测语音的真实来源确定生成对抗损失，从而能够训练上述生成对抗网络。

[0053] 语音合成阶段：

[0054] 计算机设备获取目标用户的目标音素、目标用户的音色标识以及目标风格标识。其中，目标音素是根据目标文本确定的。计算机设备通过将目标音素和目标风格标识输入第一子模型101，能够得到与目标风格标识对应的目标音素的音素后验特征以及与目标风格标识和目标音素对应的目标基频特征。之后将音素后验特征、音色标识以及目标基频特征输入第二子模型102，从而得到与目标文本、音色标识以及目标风格标识对应的目标语音。其中，目标文本决定了目标语音的发音内容，音色标识决定了目标语音的音色，目标风格标识决定了目标语音的发音风格，包括各音素的发音时长和基频。

[0055] 需要说明的是，在模型训练阶段，计算机设备向第二子模型102输入的基频特征是通过发音时长特征进行扩帧处理的。在语音合成阶段，计算机设备在第一子模型101进行扩帧使用的发音时长特征是通过时长预测器1012预测的。另外，在语音合成阶段，第二子模型102中的正则化流层1022会将音素后验编码器1021输出的中间变量进行反向流转化(与训练时的数据流转方向相反)，并输入解码器中进行处理，从而得到目标语音。并且，第二子模型102中的音素后验预测器1023不参与语音合成的过程。

[0056] 通过训练上述语音合成模型，能够通过语音合成模型根据目标用户的音色标识和目标文本，来生成符合目标用户音色的目标语音。合成目标语音的过程，是

通过预测的音素后验特征对中间变量进行预测，并通过逆傅里叶变换实现的。由于音素后验特征相较于频谱特征包含的信息量较少，因此预测音素后验特征所需的模型参数也较少，且逆傅里叶变换相较于上采样所需的模型参数也较少，因此能够减少模型的参数，从而减少模型的计算消耗资源，可实现在低计算资源设备中部署模型。

[0057] 图2是本申请一个示例性实施例提供的语音合成模型的训练方法的流程示意图。该方法可以用于计算机设备或计算机设备上的客户端。如图2所示，该方法包括：

[0058] 步骤202：获取目标用户的样本音素以及目标用户的音色标识。

[0059] 该目标用户是需要进行语音合成的用户，计算机设备通过训练得到的语音合成模型能够合成符合该目标用户音色且内容为目标文本的语音。

[0060] 可选地，计算机设备还会获取目标用户的样本语音对应的样本文本，该样本文本包括样本语音对应的文本内容。样本语音以及样本文本可支持不同类型的语言，本申请实施例对此不作限制。示例性地，该样本语音是通过录制目标用户针对少量文本进行的发音从而得到的。该样本音素基于目标用户的样本语音对应的样本文本确定。

[0061] 该目标用户的音色标识是用于标识目标用户的音色。在训练语音合成模型时，使用音色标识能够将模型学习到的模型参数与音色标识建立对应关系，从而在合成语音时通过向模型输入音色标识可实现合成符合该音色标识(音色标识对应的模型参数)对应的音色的语音。

[0062] 步骤204：将样本音素输入语音合成模型的第一子模型，得到样本音素的预测音素后验特征。

[0063] 该预测音素后验特征用于反映样本音素中各音素的特征以及样本音素中各音素的发音时长特征。“音素后验特征”也可称为“PPG”。

[0064] 可选地，计算机设备通过第一子模型提取样本音素的隐层特征，并通过该第一子模型预测样本音素中各音素的发音时长特征或根据样本语音获取样本音素中各音素真实的发音时长特征，从而获取到样本音素中各音素的发音时长特征。之后根据样本音素的隐层特征以及样本音素中各音素的发音时长特征，即可确

定样本音素的预测音素后验特征。可选地，确定的过程是计算机设备通过样本音素中各音素的发音时长特征对样本音素的隐层特征进行扩帧实现的。

[0065] 步骤206：将预测音素后验特征以及音色标识输入语音合成模型的第二子模型，得到与样本文本和音色标识对应的预测语音。

[0066] 该预测语音的内容为样本文本，该预测语音的音色为音色标识对应的音色，即音色标识所标识的目标用户的音色。该第二子模型是通过预测上述预测音素后验特征与预测语音的预测中间变量，并对预测中间变量得到预测语音的。其中，该预测中间变量用于反映预测语音的频域特征。该预测中间变量是第二子模型在确定预测语音的过程中预测出的中间变量。该预测中间变量也可称为预测潜在变量。

[0067] 在一种可能的实现方式中，该预测语音可以基于逆傅里叶变换得到，即语音合成模型的第二子模型通过预测预测音素后验特征与预测语音的预测中间变量，基于逆傅里叶变换得到预测语音。

[0068] 步骤208：根据预测音素后验特征训练第一子模型，以及根据预测语音和预测中间变量训练第二子模型。

[0069] 计算机设备会根据样本语音来确定真实音素后验特征，通过计算预测音素后验特征和真实音素后验特征的损失函数，能够对第一子模型进行训练。计算机设备通过计算样本语音和预测语音的损失函数，能够对第二子模型进行训练。

[0070] 可选地，样本语音和预测语音的损失函数指样本语音和预测语音的梅尔频谱损失。计算机设备通过将预测语音和样本语音分别转换至梅尔频谱，并计算两者的梅尔频谱的L1范数距离，从而可确定损失函数训练第二子模型。

[0071] 计算机设备还会获取样本语音与预测音素后验特征对应的真实中间变量。可选地，该真实中间变量是根据样本语音确定的。计算机设备计算预测中间变量与真实中间变量的损失函数，从而训练第二子模型。可选地，该损失函数指相对熵损失(KL散度损失)。

[0072] 需要说明的是，训练的上述语音合成模型是经过预训练得到的，预训练是采用预训练的数据对语音合成模型进行的训练，可参照上述训练过程。语音合成模型的上述训练过程的目的在于学习目标用户发音的频域特征(可称为音素克隆微

调训练)，即学习目标用户的音色，并建立学习到的模型参数与音色标识的对应关系。在完成语音合成模型的训练后，通过该语音合成模型能够合成符合该目标用户音色(输入音色标识)且内容为目标文本(与样本文本不同)的目标语音。

[0073] 需要说明的是，本实施例提供的方法所训练得到的语音合成模型，能够在低计算资源设备中部署并能运行。低计算资源设备包括用户终端，用户终端包括但不限于手机、电脑、智能语音交互设备、智能家电、车载终端、飞行器等。

[0074] 综上所述，本实施例提供的方法，通过训练上述语音合成模型，能够通过语音合成模型根据目标用户的音色标识和目标文本，来生成符合目标用户音色的目标语音。合成目标语音的过程，是通过预测的音素后验特征对中间变量进行预测，并通过逆傅里叶变换实现的。由于音素后验特征相较于频谱特征包含的信息量较少，因此预测音素后验特征所需的模型参数也较少，且逆傅里叶变换相较于上采样所需的模型参数也较少，因此能够减少模型的参数，从而减少模型的计算消耗资源，可实现在低计算资源设备中部署模型。

[0075] 图3是本申请一个示例性实施例提供的语音合成模型的训练方法的流程示意图。该方法可以用于计算机设备或计算机设备上的客户端。如图3所示，该方法包括：

[0076] 步骤302：获取目标用户的样本音素以及目标用户的音色标识。

[0077] 该目标用户是需要进行语音合成的用户，该目标用户是训练或使用语音合成模型的用户确定的。可选地，计算机设备还会获取目标用户的样本语音对应的样本文本，该样本文本包括样本语音对应的文本内容，该样本音素基于目标用户的样本语音对应的样本文本确定。

[0078] 该目标用户的音色标识是用于标识目标用户的信息。在训练语音合成模型时，使用音色标识能够将模型学习到的模型参数与音色标识建立对应关系，从而在合成语音时通过向模型输入音色标识可实现合成符合该音色标识对应的音色(目标用户的音色)的语音。

[0079] 步骤304：通过语音合成模型的第一子模型的文本编码器对样本音素进行编码，得到样本音素的隐层特征。

[0080] 计算机设备将样本音素输入语音合成模型的第一子模型，能够得到样本音素的预测音素后验特征。该预测音素后验特征用于反映样本音素中各音素的特征以及样本音素中各音素的发音时长特征。

[0081] 该第一子模型可称为Text2PPG模型，用于将输入的音素序列转换为包含更多的发音信息的语言特征。该第一子模型包括文本编码器、时长规整器以及后处理网络。计算机设备通过文本编码器对样本音素进行编码，从而能够得到样本音素的隐层特征。

[0082] 可选地，该文本编码器采用前馈变换器结构(Feed-Forward Transformer，FFT)结构。每个FFT模块由多头自注意力模块和卷积模块组成，在自注意力模块和卷积模块后也加入了残差连接和层归一化结构，从而提高结构的稳定性和性能。

[0083] 示例地，图4是本申请一个示例性实施例提供的文本编码器的结构示意图。如图4所示，文本编码器包括多头注意力层(Multi-Head Attention)401，残差&标准化层(Add&Norm)402以及卷积层(Conv1D)403。其中，多头注意力层401采用线性注意力机制。示例地，线性注意力机制的公式如下：

$\text{Attention} = \phi(Q)(\phi(K)^T V)$；

$$\phi(x) = \varphi(x) = 1 + \text{elu}(x) = \begin{cases} 1 + x, x \geq 0 \\ e^x, x < 0 \end{cases};$$

[0084] 其中，Q、K、V均表示样本音素的隐层表示序列。由于要保证Q与K的内积为正数，使输出概率有意义，线性注意力中使用了指数线性单元(Exponential Linear Unit，ELU)函数，这样可以先对$\phi(K)^T$和V进行矩阵乘法，其计算复杂度为O(N)。另外，在第一子模型中不考虑样本用户或声学相关信息，使得输出只与输入音素序列相关。采用线性注意力机制相较于点积注意力，在保证注意效果的同时能够实现降低计算复杂度。

[0085] 步骤306：通过语音合成模型的第一子模型的时长规整器对样本音素的隐层特征进行扩帧处理。

[0086] 由于音素后验特征能够反映完整的发音信息，其包含每个音素发音持续时间的信息，因此需要确定样本音素中各音素对应的发音时长特征，并对样本音素的隐层特征进行扩帧。

[0087] 计算机设备还会获取样本音素中各音素对应的真实发音时长特征。例如，计算机设备获取样本语音，通过对样本语音进行分析处理能够得到样本音素中各音素对应的真实发音时长特征。例如通过预训练的时长模型根据样本语音能够确定样本音素中各音素对应的真实发音时长特征。计算机设备通过时长规整器根据样本音素中各音素对应的真实发音时长特征，能够对样本音素的隐层特征进行扩帧处理。

[0088] 步骤308：通过语音合成模型的第一子模型的后处理网络对扩帧后的样本音素的隐层特征进行卷积处理，得到样本音素的预测音素后验特征。

[0089] 在对样本音素的隐层特征进行扩帧后，会将该信息输入后处理网络，后处理网络会对输入的信息进行卷积，从而对输入的信息进行平滑处理。由于扩帧后的音素后验特征已包含了音素的特征以及发音时长的特征，因此能够得到样本音素的预测音素后验特征。

[0090] 步骤310：根据预测音素后验特征训练第一子模型。

[0091] 计算机设备还会获取样本音素的真实音素后验特征。例如，通过将样本语音输入语音识别模型能够得到样本音素的真实音素后验特征。计算机设备通过计算预测音素后验特征与真实音素后验特征的损失函数，训练第一子模型。示例地，该损失函数为L2范数损失。

[0092] 可选地，由于在合成语音时无法获取真实的发音时长特征，第一子模型还包括时长预测器。在训练第一子模型时，计算机设备会通过时长预测器对样本音素的隐层特征进行预测，从而得到样本音素中各音素对应的预测发音时长特征。之后，计算机设备会计算预测发音时长特征和真实发音时长特征的损失函数，训练第一子模型。示例地，该损失函数为L2范数损失。

[0093] 在训练完成后，用于输入所述时长规整器的所述真实发音时长特征被替换为所述时长预测器得到的预测发音时长特征。

[0094] 另外，由于样本用户的录音往往较为急促没有情感，对于不同的合成语音的场景需要合成不同风格的语音，该风格包括语音中各音素的发音时长(发音的时长停顿)以及基频(基频变化)，在本申请实施例中，通过风格标识来标识上述语音风格。计算机设备通过使用不同的风格标识能够控制模型生成适应不同场景的预

测音素后验特征。可选地，该方案中，计算机设备还会获取风格标识，并通过文本编码器根据风格标识对样本音素进行编码，得到与风格标识对应的样本音素的隐层特征。

[0095] 需要说明的是，不同的风格标识具有与其对应的模型参数。即输入的风格标识不同，模型的模型参数也不同，这会影响到第一子模型对样本音素进行编码从而得到的隐层特征， 输入不同风格标识能够得到样本音素在不同风格标识所对应的风格下的隐层特征。风格标识对隐层特征的影响会影响到模型后续的输入，例如会影响时长预测器预测的发音时长特征(输出与样本音素和风格标识对应的预测发音时长特征)以及影响基频预测器预测的基频特征(输出与风格标识和样本音素对应的基频特征)。可选地，在预训练第一子模型的过程中，计算机设备会使用不同风格的预训练数据以及对应的风格标识进行训练，从而使得第一子模型能够学习到不同风格对应的模型参数。

[0096] 可选地，第一子模型还包括基频预测器。计算机设备通过基频预测器对样本音素的隐层特征进行预测，能够得到与风格标识和样本音素对应的基频特征。其中，该基频特征用于输入第二子模型从而得到与风格标识对应的预测语音，即得到与风格标识所对应的风格的预测语音。由于语音的风格与基频相关，且增加基频预测能提高预测音素后验特征的预测效果，可选地，计算机设备还会将通过基频预测器预测得到的基频特征拼接在文本编码器的输出上。

[0097] 示例地，第一子模型的结构，可参照图1中的实例。

[0098] 步骤312：将预测音素后验特征和音色标识输入第二子模型的先验编码器，得到预测中间变量。

[0099] 计算机设备将预测音素后验特征以及音色标识输入语音合成模型的第二子模型，能够得到与样本文本和音色标识对应的预测语音。其中，第二子模型是通过预测预测音素后验特征与预测语音的预测中间变量，基于逆傅里叶变换得到预测语音的，预测中间变量用于反映预测语音的频域特征。由于语音合成模型使用音素后验特征作为语言特征，其已经提供了音素持续时间信息，因此第二子模型不需要考虑建模发音时长的信息。

[0100] 该第二子模型包括先验编码器和解码器。计算机设备通过将预测音素后验特征和音色标识输入第二子模型的先验编码器，能够得到上述预测中间变量。

[0101] 可选地，先验编码器包括音素后验编码器(PPG编码器)。计算机设备将预测音素后验特征和音色标识输入音素后验编码器，通过音素后验编码器以包含预测音素后验特征和音色标识的条件信息c为条件，对预测中间变量z的先验分布p(z|c)采样均值与方差，从而能够得到上述预测中间变量。可选地，该音素后验编码器基于FFT结构，并采用线性注意力机制，具体结构可参照如图4的示例。

[0102] 可选地，在第一子模型输出了样本音素和风格标识对应的基频特征的情况下，计算机设备还会获取与样本音素和风格标识对应的基频特征。其中，基频特征是通过第一子模型基于风格标识对样本音素进行特征提取得到的。计算机设备将预测音素后验特征、音色标识和基频特征输入先验编码器，能够得到与风格标识对应的预测中间变量，即该预测中间变量具有与该风格标识对应的风格。通过该预测中间变量能够合成具有与该风格标识对应的风格的预测语音。

[0103] 步骤314：通过第二子模型的解码器对预测中间变量进行逆傅里叶变换，得到预测语音。

[0104] 可选地，解码器包括逆傅里叶变换解码器。计算机设备通过逆傅里叶变换解码器对预测中间变量进行逆傅里叶变换，能够得到预测语音。可选地，在实际应用中，若不将目标用户的风格标识输入解码器，目标用户的音色克隆的相似度(样本语音与预测语音的音色相似度)会显著降低。因此，计算机设备会向逆傅里叶变换解码器输入预测中间变量以及风格标识，从而通过逆傅里叶变换解码器根据风格标识对预测中间变量进行逆傅里叶变换，得到预测语音。

[0105] 可选地，逆傅里叶变换解码器包括多个一维卷积层，最后一个一维卷积层与逆傅里叶变换层连接。示例地，图5是本申请一个示例性实施例提供的一种逆傅里叶变换解码器的结构示意图。如图5所示，z表示预测中间变量，说话人id表示风格标识。计算机设备通过逆傅里叶变换解码器使用多个一维卷积层501能够将输入的维数逐渐增加到(f/2+1)*2，使输出符合实部和虚部的总维数，其中f表示快速傅里叶变换的大小。残差网络502的堆叠跟随每个一维卷积层501以获得相应尺度上的更多信息。由于在频域维度建模，因此没有使用扩张卷积，而是使用较

小的核大小，目的是确保接收场不会太大。在一维卷积层501中通过采用群卷积能够节省计算量。在通过一维卷积层501和残差网络502后，输出被分成实部和虚部，最终的波形(预测语音)可以通过逆傅立叶变换层503产生。可选地，一维卷积层501的数量是根据模型训练的效果确定的。

[0106] 可选地，上述逆傅里叶变换解码器也可以通过上采样结构以及逆傅立叶变换结构的组合进行替换。示例地，图6是本申请一个示例性实施例提供的另一种逆傅里叶变换解码器的结构示意图。如图6所示，z表示预测中间变量，说话人id表示风格标识。在该结构下，逆傅里叶变换解码器包括一维卷积层601、逆傅里叶变换结构，上采样网络结构以及伪正交镜像滤波器组605。其中，逆傅里叶变换结构包括残差网络602、一维卷积层601以及逆傅里叶变换层603。上采样网络结构包括上采样网络604以及残差网络602。

[0107] 由于语音的中高频特征较好还原，谐波主要集中于低频部分，因此可以通过上采样结构还原低频部分并通过逆傅立叶变换层建模中高频。之后通过伪正交镜像滤波器组，可将语音分为多个子频带，由上采样结构建模第一个子频带，其余由图5所示的结构生成。由于上采样结构只建模低频，相当于原始的上采样结构减少了大量的参数和计算复杂度。该方案虽然参数和计算量较图5所示的结构较多，但也会取得模型效果的提升。针对不同的部署场景，可以作为适应该场景计算和存储要求的方案。

[0108] 步骤316：根据预测语音和预测中间变量训练第二子模型。

[0109] 计算机设备还会获取样本语音，并计算预测语音与样本语音之间的梅尔频谱损失，从而训练第二子模型。

[0110] 可选地，第二子模型还包括后验编码器。计算机设备获取样本语音，并将样本语音(线性谱)输入后验编码器，从而得到样本语音与预测音素后验特征的真实中间变量。计算机设备通过计算预测中间变量与真实中间变量的相对熵损失(KL散度损失)，训练第二子模型。

[0111] 可选地，后验编码器包括后验预测器(PPG预测器)。计算机设备将样本语音输入后验预测器，对真实中间变量的后验分布$p(z|y)$采样均值与方差，得到真实中间变量。其中，y表示样本语音。

[0112] 可选地，先验编码器还包括正则化流层。计算机设备通过正则化流层对真实中间变量进行仿射耦合处理，从而得到处理后的真实中间变量。在计算KL散度损失时，计算机设备会计算预测中间变量与处理后的真实中间变量的相对熵损失，从而训练第二子模型。该正则化流能够将中间变量z转换为更复杂的分布，KL散度损失用于使真实中间变量与预测中间变量的分布保持一致。

[0113] 上述正则化流层包括多个仿射耦合层，每个仿射耦合层用于对真实中间变量进行仿射耦合处理。由于使用了多层处理，正则化流层具有大量的参数，本申请实施例通过使不同仿射耦合层共享模型参数，来减少模型复杂度。示例地，图7是本申请一个示例性实施例提供的正则化流层的结构示意图。如图7所示，通过使不同仿射耦合层701共享模型参数，且每个仿射耦合层701对应有嵌入层标识702，能够实现将将正则化流层的参数控制为单层，减少模型参数。嵌入层标识是用于标识仿射耦合层的信息，每个仿射耦合层的嵌入层标识不同。

[0114] 需要说明的是，在训练语音合成模型的过程中，中间变量z通过流(正则化流层)转换为f(z)。在推断过程(语音合成)中，音素后验编码器的输出被反向流转化从而得到中间变量z。示例地，中间变量z的表达式如下：

$$p(z \mid c) = N(f_\theta(z); \mu_\theta(c), \sigma_\theta(c)) \left| \det \frac{\partial f_\theta(z)}{\partial z} \right|;$$

[0115] 其中，$f_\theta$表示分布，$\mu_\theta$表示均值，$\sigma_\theta$表示方差。

[0116] 在缺乏对中间变量的明确约束的情况下，输出的预测语音容易出现错误读音、语调异常等发音错误。本申请实施例提供的方法通过引入音素后验预测器(PPG预测器)来提供发音约束。可选地，先验编码器还包括音素后验预测器，音素后验预测器用于在预训练第二子模型的过程中，根据预训练过程中的预测中间变量预测预训练过程中的预测音素后验特征。在第二子模型的预训练过程中，计算预训练过程中的预测音素后验特征与预训练过程中的真实音素后验特征的损失函数，训练第二子模型。可选地，该损失函数为L1范数损失。示例地，该损失函数的表达式如下：

$L_{PPg} = ||PPG1 - PPG2||_1$;

[0117] 其中，PPG1表示预训练过程中的预测音素后验特征，PPG2表示预训练过程中的真实音素后验特征。

[0118] 需要说明的是，音素后验预测器仅在预训练语音合成模型时进行训练，在音素克隆微调训练时冻结。

[0119] 综上，训练第二子模型过程中的上述损失函数可以归类为条件变分自编码器(Conditional Variational Auto Encoder，CVAE)损失，其损失函数如下：

$L_{cvae}=L_{kl}+\lambda_{recon}\star L_{recon}+\lambda_{ppg}\star L_{ppg}$；

[0120] 其中，$L_{kl}$表示预测中间变量与真实中间变量的KL散度损失，$L_{recon}$表示样本语音与预测语音的梅尔频谱的L1损失，$L_{ppg}$表示PPG预测器的损失。$\lambda_{recon}$和$\lambda_{ppg}$为权重参数。

[0121] 可选地，计算机设备也能够将真实中间变量输入解码器，从而得到与真实中间变量对应的预测语音，之后根据该预测语音与样本语音计算梅尔频谱的L1损失。在该情况下，计 算机设备获取到的预测中间变量主要用于确定预测中间变量与真实中间变量的KL散度损失来训练第二子模型。

[0122] 可选地，解码器还包括判别器，判别器与语音合成模型中除判别器以外的部分组成生成对抗网络(Generative Adversarial Networks，GAN)。计算机设备将预测语音输入判别器，能够得到预测语音的判别结果，该判别结果用于反映预测语音为真实的信息或预测的信息。之后根据判别结果与预测语音的真实来源确定生成对抗损失，从而训练生成对抗网络。

[0123] 可选地，判别器包括多尺度频谱判别器。多尺度频谱判别器在逆傅立叶变换解码器中特别有效，在高频谐波重构方面有明显的增益。计算机设备将预测语音输入多尺度频谱判别器，通过多尺度频谱判别器的多个子判别器分别在幅度谱上对预测语音进行短时傅立叶变换，并通过多个二维卷积层进行二维卷积处理，从而得到预测语音的判别结果。其中，每个子判别器的短时傅立叶变换参数不同。

[0124] 示例地，图8是本申请一个示例性实施例提供的多尺度频谱判别器的结构示意图。如图8所示，多尺度频谱判别器由多个不同短时傅立叶变换参数的子判别器组成，预测语音输入判别器后，经过短时傅立叶变换层801得到实部和虚部后取幅值输出频谱特征并通过多层二维卷积层802，从而可以学习预测语音中的不同分辨率的频域特征，进而可实现判别。

[0125] 可选地，判别器包括多尺度复数谱判别器。多尺度复数谱判别器对语音信号的实部和虚部之间的关系进行建模，这有助于提高相位精度的判别。计算机设备将预测语音输入多尺度复数谱判别器，通过多尺度复数谱判别器的多个子判别器分别在复数谱上对预测语音进行短时傅立叶变换，并通过多个二维复数卷积层进行二维复数卷积处理，从而得到预测语音的判别结果。其中，每个子判别器的短时傅立叶变换参数不同。多尺度复数谱判别器通过短时傅里叶变换在多个尺度上将信号分为实部和虚部，然后对输入进行二维复数卷积，该方法在复值域内有较好的效果。

[0126] 综上，训练第二子模型过程中的损失函数由CVAE损失以及GAN损失组成，总体的损失函数如下：

$L_G = L_{adv}(G) + \lambda_{fm} \star L_{fm}(G) + L_{cvae}$；

$L_D = L_{adv}(D)$；

[0127] 其中，$L_{adv}(G)$为GAN中的生成器(generator)的损失，$L_{adv}(D)$为GAN中的判别器(discriminator)的损失。$L_{fm}(G)$为生成器的特征匹配损失，具体是指生成器中每个网络层输入真实数据和样本数据所得到的输出之间的差异的损失。$\lambda_{fm}$为权重参数。

[0128] 示例地，第二子模型的结构，可参照图1中的实例。

[0129] 综上所述，本实施例提供的方法，通过训练上述语音合成模型，能够通过语音合成模型根据目标用户的音色标识和目标文本，来生成符合目标用户音色的目标语音。合成目标语音的过程，是通过预测的音素后验特征对中间变量进行预测，并通过逆傅里叶变换实现的。由于音素后验特征相较于频谱特征包含的信息量较少，因此预测音素后验特征所需的模型参数也较少，且逆傅里叶变换相较于上采样所需的模型参数也较少，因此能够减少模型的参数，从而减少模型的计算消耗资源，可实现在低计算资源设备中部署模型。

[0130] 本实施例提供的方法，还通过逆傅立叶变换的解码器、使正则化流层的不同仿射耦合层共享模型参数、使用线性注意力机制、提取音素后验特征而非频谱特征实现了有效降低模型参数量和低计算复杂度。

[0131] 在相关技术的两段式模型(Fastspeech)中，由于分为声学模型和声码器，两者间存在误差，会导致合成的语音音质损失，而这一问题在小样本音色克隆中更明显。而目前端到端模型(输入文本直接输出语音的模型)仍存在语音生成不稳定的问题，甚至存在发音错误，这会严重影响听感。本实施例提供的方法，在采用端到端方式建模的基础上，通过引入风格标识、根据时长预测器和基频预测器生成音素后验特征、通过音素后验预测器约束中间变量、通过正则化流约束中间变量的分布以及通过生成对抗网络的方式进行训练，改善了上述问题，提升了语音合成的性能。

[0132] 如何有效的降低参数和计算复杂度是一个直观的挑战。相关技术中由于声学信息和文本信息一起建模，直接降低模型参数会导致模型效果的快速下降，例如蒸馏和量化等模型压缩算法也会导致明显的性能损失。本实施例提供的方法，不仅能够通过上述减少参数的方式实现有效降低模型参数量和低计算复杂度，还通过以端到端为基础构造模型结构，并采用提升性能的上述结构来提升模型的语音合成性能表现，可同时实现减少模型参数以及提升模型性能。

[0133] 相关技术中由于用户录音语速等节奏问题，与应用场景适配度差。如在导航场景中，字正腔圆的朗读风格是用户所追求的，但由于相关技术的模型将音色和内容共同建模，生成语音风格并不合适。本实施例提供的方法，还通过分开建模音色和内容，并引入风格标识和基频来进行特征提取以及语音合成，可以实现针对不同场景合成与其适配的朗读风格的语音。

[0134] 图9是本申请一个示例性实施例提供的语音合成方法的流程示意图。该方法可以用于计算机设备或计算机设备上的客户端。如图9所示，该方法包括：

[0135] 步骤902：获取目标用户的目标音素以及目标用户的音色标识。

[0136] 该目标用户是需要进行语音合成的用户。计算机设备通过目标用户的样本音素以及目标用户的音色标识能够训练语音合成模型。样本音素基于目标用户的样本语音对应的样本文本确定。该目标音素基于目标文本确定。目标文本与样本文本相同、部分相同或不同，目标文本是使用语音合成模型的用户确定的。

[0137] 该目标用户的音色标识是用于标识目标用户的信息。在训练语音合成模型时，使用音色标识能够将模型学习到的模型参数与音色标识建立对应关系，从而在

合成语音时通过向模型输入音色标识可实现合成符合该音色标识对应的音色的语音。

[0138] 步骤904：将目标音素输入语音合成模型的第一子模型，得到目标音素的音素后验特征。

[0139] 上述计算机设备或客户端中部署有语音合成模型，该语音合成模型是通过本申请实施例提供的语音合成模型的训练方法训练得到的。

[0140] 该音素后验特征用于反映目标音素中各音素的特征以及目标音素中各音素的发音时长特征。可选地，计算机设备通过第一子模型提取目标音素的隐层特征，并通过该第一子模型的时长预测器预测目标音素中各音素的发音时长特征，之后根据目标音素的隐层特征以及目标音素中各音素的发音时长特征，即可确定目标音素的预测音素后验特征。

[0141] 步骤906：将音素后验特征以及音色标识输入语音合成模型的第二子模型，得到与目标文本和音色标识对应的目标语音。

[0142] 该目标语音的内容为目标文本，该目标语音的音色为音色标识对应的音色，即音色标识所标识的目标用户的音色。该第二子模型是通过预测上述音素后验特征与目标语音的中间变量，并对中间变量基于逆傅里叶变换从而得到目标语音的。其中，该中间变量用于反映目标语音的频域特征。

[0143] 可选地，计算机设备还会获取目标风格标识。并将目标音素和目标风格标识输入语音合成模型的第一子模型，得到与目标风格标识对应的目标音素的音素后验特征以及与目标风格标识和目标音素对应的目标基频特征。之后将音素后验特征、音色标识以及目标基频特征输入语音合成模型的第二子模型，得到与目标文本、音色标识和目标基频特征对应的目标语音。

[0144] 综上所述，本实施例提供的方法，通过语音合成模型根据目标用户的音色标识和目标文本，能够生成符合目标用户音色的目标语音。合成目标语音的过程，是通过预测的音素后验特征对中间变量进行预测，并通过逆傅里叶变换实现的。由于音素后验特征相较于频谱特征包含的信息量较少，因此预测音素后验特征所需的模型参数也较少，且逆傅里叶变换相较于上采样所需的模型参数也较

少，因此能够减少模型的参数，从而减少模型的计算消耗资源，可实现在低计算资源设备中部署模型。

[0145] 在同等模型效果下，本申请实施例提供的方法大大减少了模型的参数量和计算量，这将有益于减少资源的使用，包括计算资源和存储资源，并更方便的部署在端侧等应用场景中。同时，模型相比相关技术减少了发音错误，使模型更稳定。

[0146] 实验过程和分析如下：

[0147] 使用开放的数据集来对本申请实施例中的模型以及相关技术的模型进行预训练，该开放的数据集包含了1151个说话人大约242小时的语音。为了评估模型在说话人音色克隆任务中的性能，采用不同声学条件的多说话人语料库对预训练模型进行微调。在实际操作中，随机选取5名男性和5名女性作为目标说话人进行音色克隆。每个演讲者随机抽取20句话。另外从每个说话者中随机选择10个额外的句子，得到一个总共有10个说话者100句话的测试集。

[0148] 本申请实施例提供的语音合成模型与用于语音合成带有对抗学习的条件变分自编码器(Variational Inference with adversarial learning for end-to-end Text-to-Speech，VITS)、Fastspeech+HiFiGAN对比。VITS模型采用原论文结构。对于Fastspeech+HiFiGAN，为了比较和控制不同参数量下的情况，采用了两种结构，结构1称为Fastspeech+HiFiGAN v1，结构2称为Fastspeech+HiFiGAN v2。v1采用原论文结构Fastspeech和HiFiGAN v1。相比较于v1，v2采用了更小的结构。v2在编码器和解码器中采用两层FFT，且隐藏层的维度设置为128，HiFiGAN采用v2版本。

[0149] 在客观指标的评估中，每句测试样本由二十名听众评测，参与者对样本的自然度和说话人音色相似度进行打分，最高5分，最低1分。计算复杂度测量采用每秒10亿次浮点运算数(Giga Floating-point Operations Per Second，GFLOPS)作为单位。另外测算了每个系统的词错误率(Word Error Rate，WER)，以测试每个模型的稳定性，特别是在发音和语调方面。实验结果如表1所示：

[0150] 表1

| 模型 | 参数量(M) | 计算量(GFlops) | 自然度 | 音色相似度 | WER(%) |
|---|---|---|---|---|---|
| Fastspeech+HifiGAN v1 | 40.16 | 15.85 | 3.08 | 3.21 | 8.90 |
| Fastspeech+HifiGAN v2 | 8.67 | 0.98 | 2.63 | 3.08 | 10.53 |
| VITS | 29.36 | 15.76 | 2.59 | 3.53 | 15.29 |
| 本申请实施例的模型 | 8.97 | 0.72 | 2.94 | 3.10 | 8.19 |
| 原始录音 | - | - | 3.70 | 3.62 | 4.68 |

[0151] 根据表1可知，与本申请实施例中的模型大小相似的Fastspeech+HifiGAN v2相比，本申请实施例中的模型实现了更好的自然度和更少的计算复杂度。与Fastspeech+HifiGAN v1相比，本申请实施例中的模型在自然度和说话人相似度上仍有差距，但是其实现了更好的WER，保证了模型的稳定性。

[0152] 需要说明的是，本申请实施例提供的方法步骤的先后顺序可以进行适当调整，步骤也可以根据情况进行相应增减，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化的方法，都应涵盖在本申请的保护范围之内，因此不再赘述。

[0153] 以本申请实施例提供的方法应用于AI播报业务中的用户语音包制作的场景为例，例如应用于地图导航的语音包制作。计算机设备会获取目标用户的目标文本的目标音素、目标用户的音色标识以及目标风格标识。目标用户为用户自身或者为明星等，目标文本包括地图导航场景的语音播报的文本。目标风格标识是与地图导航的语音播报所对应的风格的标识，例如该风格为朗读音素时发音缓慢且基频准确。计算机设备通过将目标音素和目标风格标识输入第一子模型，能够得到与目标风格标识对应的目标音素的音素后验特征以及与目标风格标识和目标音素对应的目标基频特征。之后将音素后验特征、音色标识以及目标基频特征输入第二子模型，从而得到与目标文本、音色标识以及目标风格标识对应的目标语音。其中，目标文本决定了目标语音的发音内容，即目标语音的发音内容为地图导航场景的语音播报的文本。音色标识决定了目标语音的音色，即为用户所选择的目标用户的音色。目标风格标识决定了目标语音的发音风格，包括各音素的发音时长和基频。该语音合成模型是通过目标用户的样本语音、音色标识进行音色克隆微调得到的。该样本语音的内容与目标文本不同。

例如样本语音是通过录制目标用户针对少量文字的朗读语音得到的，目标文本可包括与该少量文字不同的大量文字。从而可实现根据用户针对少量文字录制的语音来合成符合该用户音色的针对其它文字的语音。

[0154] 图10是本申请一个示例性实施例提供的语音合成模型的训练装置的结构示意图。如图10所示，该装置包括：

[0155] 获取模块1001，用于获取目标用户的样本音素以及所述目标用户的音色标识，所述样本音素基于所述目标用户的样本语音对应的样本文本确定，所述音色标识用于标识所述目标用户的音色；

[0156] 输入输出模块1002，用于将所述样本音素输入所述语音合成模型的第一子模型，得到所述样本音素的预测音素后验特征，所述预测音素后验特征用于反映所述样本音素中各音素的特征以及所述样本音素中各音素的发音时长特征；

[0157] 所述输入输出模块1002，还用于将所述预测音素后验特征以及所述音色标识输入所述语音合成模型的第二子模型，得到与所述样本文本和所述音色标识对应的预测语音，所述第二子模型是通过预测所述预测音素后验特征与所述预测语音的预测中间变量得到所述预测语音的，所述预测中间变量用于反映所述预测语音的频域特征；

[0158] 训练模块1003，用于根据所述预测音素后验特征训练所述第一子模型；以及根据所述预测语音和所述预测中间变量训练所述第二子模型。

[0159] 在一个可选的设计中，所述第二子模型包括先验编码器和解码器；所述输入输出模块1002，用于：

[0160] 将所述预测音素后验特征和所述音色标识输入所述先验编码器，得到所述预测中间变量；

[0161] 通过所述解码器对所述预测中间变量进行逆傅里叶变换，得到所述预测语音。

[0162] 在一个可选的设计中，所述先验编码器包括音素后验编码器；所述输入输出模块1002，用于：

[0163] 将所述预测音素后验特征和所述音色标识输入所述音素后验编码器，对所述预测中间变量的先验分布采样均值与方差，得到所述预测中间变量。

[0164] 在一个可选的设计中，所述第二子模型还包括后验编码器；所述获取模块1001，用于：

[0165] 获取所述样本语音；

[0166] 所述输入输出模块1002，用于：

[0167] 将所述样本语音输入所述后验编码器，得到真实中间变量；

[0168] 所述训练模块1003，用于：

[0169] 计算所述预测中间变量与所述真实中间变量的相对熵损失，训练所述第二子模型。

[0170] 在一个可选的设计中，所述后验编码器包括后验预测器；所述输入输出模块1002，用于：

[0171] 将所述样本语音输入所述后验预测器，对所述真实中间变量的后验分布采样均值与方差，得到所述真实中间变量。

[0172] 在一个可选的设计中，所述先验编码器还包括正则化流层；所述输入输出模块1002，用于：

[0173] 通过所述正则化流层对所述真实中间变量进行仿射耦合处理，得到处理后的真实中间变量；

[0174] 所述训练模块1003，用于：

[0175] 计算所述预测中间变量与所述处理后的真实中间变量的相对熵损失，训练所述第二子模型。

[0176] 在一个可选的设计中，所述正则化流层包括多个仿射耦合层，每个所述仿射耦合层用于对所述真实中间变量进行仿射耦合处理；不同所述仿射耦合层共享模型参数，且每个所述仿射耦合层对应有嵌入层标识。

[0177] 在一个可选的设计中，所述先验编码器还包括音素后验预测器，所述音素后验预测器用于在预训练所述第二子模型的过程中，根据预训练过程中的预测中间变量预测所述预训练过程中的预测音素后验特征。

[0178] 在一个可选的设计中，所述训练模块1003，用于：

[0179] 在所述第二子模型的所述预训练过程中，计算所述预训练过程中的预测音素后验特征与所述预训练过程中的真实音素后验特征的损失函数，训练所述第二子模型。

[0180] 在一个可选的设计中，所述获取模块1001，用于：

[0181] 获取与所述样本音素和风格标识对应的基频特征，所述基频特征是通过所述第一子模型基于所述风格标识对所述样本音素进行特征提取得到的，所述风格标识用于标识语音风格；

[0182] 所述输入输出模块1002，用于：

[0183] 将所述预测音素后验特征、所述音色标识和所述基频特征输入所述先验编码器，得到与所述风格标识对应的所述预测中间变量。

[0184] 在一个可选的设计中，所述解码器包括逆傅里叶变换解码器；所述输入输出模块1002，用于：

[0185] 通过所述逆傅里叶变换解码器对所述预测中间变量进行逆傅里叶变换，得到所述预测语音。

[0186] 在一个可选的设计中，所述输入输出模块1002，用于：

[0187] 通过所述逆傅里叶变换解码器根据所述风格标识对所述预测中间变量进行逆傅里叶变换，得到所述预测语音。

[0188] 在一个可选的设计中，所述逆傅里叶变换解码器包括多个一维卷积层，最后一个所述一维卷积层与逆傅里叶变换层连接。

[0189] 在一个可选的设计中，所述获取模块1001，用于：

[0190] 获取所述样本语音；

[0191] 所述训练模块1003，用于：

[0192] 计算所述预测语音与所述样本语音之间的梅尔频谱损失，训练所述第二子模型。

[0193] 在一个可选的设计中，所述解码器还包括判别器，所述判别器与所述语音合成模型中除所述判别器以外的部分组成生成对抗网络；所述输入输出模块1002，用于：

[0194] 将所述预测语音输入所述判别器，得到所述预测语音的判别结果，所述判别结果用于反映所述预测语音为真实的信息或预测的信息；

[0195] 所述训练模块1003，用于：

[0196] 根据所述判别结果与所述预测语音的真实来源确定生成对抗损失，训练所述生成对抗网络。

[0197] 在一个可选的设计中，所述判别器包括多尺度频谱判别器；所述输入输出模块1002，用于：

[0198] 将所述预测语音输入所述多尺度频谱判别器，通过所述多尺度频谱判别器的多个子判别器分别在幅度谱上对所述预测语音进行短时傅立叶变换，并通过多个二维卷积层进行二维卷积处理，得到所述预测语音的判别结果；

[0199] 其中，每个所述子判别器的短时傅立叶变换参数不同。

[0200] 在一个可选的设计中，所述判别器包括多尺度复数谱判别器；所述输入输出模块1002，用于：

[0201] 将所述预测语音输入所述多尺度复数谱判别器，通过所述多尺度复数谱判别器的多个子判别器分别在复数谱上对所述预测语音进行短时傅立叶变换，并通过多个二维复数卷积层进行二维复数卷积处理，得到所述预测语音的判别结果；

[0202] 其中，每个所述子判别器的短时傅立叶变换参数不同。

[0203] 在一个可选的设计中，所述第一子模型包括文本编码器、时长规整器以及后处理网络；所述获取模块1001，用于：

[0204] 获取所述样本音素中各音素对应的真实发音时长特征；

[0205] 所述输入输出模块1002，用于：

[0206] 通过所述文本编码器对所述样本音素进行编码，得到所述样本音素的隐层特征；

[0207] 通过所述时长规整器根据所述样本音素中各音素对应的真实发音时长特征，对所述样本音素的隐层特征进行扩帧处理；

[0208] 通过所述后处理网络对扩帧后的所述样本音素的隐层特征进行卷积处理，得到所述样本音素的所述预测音素后验特征。

[0209] 在一个可选的设计中，所述获取模块1001，用于：

[0210] 获取所述样本音素的真实音素后验特征；

[0211] 所述训练模块1003，用于：

[0212] 计算所述预测音素后验特征与所述真实音素后验特征的损失函数，训练所述第一子模型，在训练完成后，用于输入所述时长规整器的所述真实发音时长特征被替换为所述时长预测器得到的预测发音时长特征。

[0213] 在一个可选的设计中，所述第一子模型还包括时长预测器，所述输入输出模块1002，用于：

[0214] 通过所述时长预测器对所述样本音素的隐层特征进行预测，得到所述样本音素中各音素对应的预测发音时长特征；

[0215] 所述训练模块1003，用于：

[0216] 计算所述预测发音时长特征和所述真实发音时长特征的损失函数，训练所述第一子模型。

[0217] 在一个可选的设计中，所述获取模块1001，用于：

[0218] 获取风格标识；

[0219] 所述输入输出模块1002，用于：

[0220] 通过所述文本编码器根据所述风格标识对所述样本音素进行编码，得到与所述风格标识对应的所述样本音素的隐层特征。

[0221] 在一个可选的设计中，所述第一子模型还包括基频预测器；所述输入输出模块1002，用于：

[0222] 通过所述基频预测器对所述样本音素的隐层特征进行预测，得到与所述风格标识和所述样本音素对应的基频特征，所述基频特征用于输入所述第二子模型得到与所述风格标识对应的所述预测语音。

[0223] 图11是本申请一个示例性实施例提供的语音合成装置的结构示意图。该装置中包括通过如图10所述的装置训练得到的语音合成模型。如图11所示，该装置包括：

[0224] 获取模块1101，用于获取目标用户的目标音素以及所述目标用户的音色标识，所述目标音素基于目标文本确定，所述音色标识用于标识所述目标用户的音色；

[0225] 输入输出模块1102，用于将所述目标音素输入所述语音合成模型的第一子模型，得到所述目标音素的音素后验特征；

[0226] 所述输入输出模块1102，还用于将所述音素后验特征以及所述音色标识输入所述语音合成模型的第二子模型，得到与所述目标文本和所述音色标识对应的目标语音。

[0227] 在一个可选的设计中，所述获取模块1101，用于：

[0228] 获取目标风格标识，所述目标风格标识用于标识所述目标语音的语音风格；

[0229] 所述输入输出模块1102，用于：

[0230] 将所述目标音素和所述目标风格标识输入所述语音合成模型的第一子模型，得到与所述目标风格标识对应的所述目标音素的所述音素后验特征以及与所述目标风格标识和所述目标音素对应的目标基频特征；

[0231] 将所述音素后验特征、所述音色标识以及所述目标基频特征输入所述语音合成模型的第二子模型，得到与所述目标文本、所述音色标识和所述目标基频特征对应的所述目标语音。

[0232] 需要说明的是：上述实施例提供的语音合成模型的训练装置，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，即将设备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合成模型的训练装置与语音合成模型的训练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。

[0233] 同理，上述实施例提供的语音合成装置，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，即将设备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合成装置与语音合成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。

[0234] 本申请的实施例还提供了一种计算机设备，该计算机设备包括：处理器和存储器，存储器中存储有至少一条指令、至少一段程序、代码集或指令集，至少一

条指令、至少一段程序、代码集或指令集由处理器加载并执行以实现上述各方法实施例提供的语音合成模型的训练方法或语音合成方法。

[0235] 可选地，该计算机设备为服务器。示例地，图12是本申请一个示例性实施例提供的计算机设备的结构示意图。

[0236] 所述计算机设备1200包括中央处理单元(Central Processing Unit，CPU)1201、包括随机存取存储器(Random Access Memory，RAM)1202和只读存储器(Read-Only Memory，ROM)1203的系统存储器1204，以及连接系统存储器1204和中央处理单元1201的系统总线1205。所述计算机设备1200还包括帮助计算机设备内的各个器件之间传输信息的基本输入/输出系统(Input/Output系统，I/O系统)1206，和用于存储操作系统1213、应用程序1214和其他程序模块1215的大容量存储设备1207。

[0237] 所述基本输入/输出系统1206包括有用于显示信息的显示器1208和用于用户输入信息的诸如鼠标、键盘之类的输入设备1209。其中所述显示器1208和输入设备1209都通过连接到系统总线1205的输入输出控制器1210连接到中央处理单元1201。所述基本输入/输出系统1206还可以包括输入输出控制器1210以用于接收和处理来自键盘、鼠标、或电子触控笔等多个其他设备的输入。类似地，输入输出控制器1210还提供输出到显示屏、打印机或其他类型的输出设备。

[0238] 所述大容量存储设备1207通过连接到系统总线1205的大容量存储控制器(未示出)连接到中央处理单元1201。所述大容量存储设备1207及其相关联的计算机可读存储介质为计算机设备1200提供非易失性存储。也就是说，所述大容量存储设备1207可以包括诸如硬盘或者只读光盘(Compact Disc Read-Only Memory，CD-ROM)驱动器之类的计算机可读存储介质(未示出)。

[0239] 不失一般性，所述计算机可读存储介质可以包括计算机存储介质和通信介质。计算机存储介质包括以用于存储诸如计算机可读存储指令、数据结构、程序模块或其他数据等信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括RAM、ROM、可擦除可编程只读寄存器(Erasable Programmable Read Only Memory，EPROM)、电子抹除式可复写只读存储器(Electrically-Erasable Programmable Read-Only Memory，EEPROM)、闪存

或其他固态存储设备、CD-ROM、数字多功能光盘(Digital Versatile Disc，DVD)或其他光学存储、磁带盒、磁带、磁盘存储或其他磁性存储设备。当然，本领域技术人员可知所述计算机存储介质不局限于上述几种。上述的系统存储器1204和大容量存储设备1207可以统称为存储器。

[0240] 存储器存储有一个或多个程序，一个或多个程序被配置成由一个或多个中央处理单元1201执行，一个或多个程序包含用于实现上述方法实施例的指令，中央处理单元1201执行该一个或多个程序实现上述各个方法实施例提供的方法。

[0241] 根据本申请的各种实施例，所述计算机设备1200还可以通过诸如因特网等网络连接到网络上的远程计算机设备运行。也即计算机设备1200可以通过连接在所述系统总线1205上的网络接口单元1211连接到网络1212，或者说，也可以使用网络接口单元1211来连接到其他类型的网络或远程计算机设备系统(未示出)。

[0242] 所述存储器还包括一个或者一个以上的程序，所述一个或者一个以上程序存储于存储器中，所述一个或者一个以上程序包含用于进行本申请实施例提供的方法中由计算机设备所执行的步骤。

[0243] 本申请实施例中还提供了一种计算机可读存储介质，该可读存储介质中存储有至少一条指令、至少一段程序、代码集或指令集，当该至少一条指令、至少一段程序、代码集或指令集由计算机设备的处理器加载并执行时，实现上述各方法实施例提供的语音合成模型的训练方法或语音合成方法。

[0244] 本申请还提供了一种计算机程序产品，该计算机程序产品包括计算机程序，当其在计算机上运行时，使得该计算机设备执行上述的语音合成模型的训练方法或语音合成方法。

[0245] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也可以通过程序来指令相关的硬件完成，该程序可以存储于一种计算机可读存储介质中，上述提到的可读存储介质可以是只读存储器，磁盘或光盘等。

[0246] 以上所述仅为本申请的可选实施例，并不用以限制本申请，凡在本申请的精神和原则之内，所作的任何修改、等同切换、改进等，均应包含在本申请的保护范围之内。

# 权利要求书

[权利要求 1] 一种语音合成模型的训练方法，所述方法由计算机设备执行，所述方法包括：

获取目标用户的样本音素以及所述目标用户的音色标识，所述样本音素基于所述目标用户的样本语音对应的样本文本确定，所述音色标识用于标识所述目标用户的音色；

将所述样本音素输入所述语音合成模型的第一子模型，得到所述样本音素的预测音素后验特征，所述预测音素后验特征用于反映所述样本音素中各音素的特征以及所述样本音素中各音素的发音时长特征；

将所述预测音素后验特征以及所述音色标识输入所述语音合成模型的第二子模型，得到与所述样本文本和所述音色标识对应的预测语音，所述第二子模型是通过预测所述预测音素后验特征与所述预测语音的预测中间变量得到所述预测语音的，所述预测中间变量用于反映所述预测语音的频域特征；

根据所述预测音素后验特征训练所述第一子模型；以及根据所述预测语音和所述预测中间变量训练所述第二子模型。

[权利要求 2] 根据权利要求1所述的方法，所述第二子模型包括先验编码器和解码器；所述将所述预测音素后验特征以及所述音色标识输入所述语音合成模型的第二子模型，得到与所述样本文本和所述音色标识对应的预测语音，包括：

将所述预测音素后验特征和所述音色标识输入所述先验编码器，得到所述预测中间变量；

通过所述解码器对所述预测中间变量进行逆傅里叶变换，得到所述预测语音。

[权利要求 3] 根据权利要求2所述的方法，所述先验编码器包括音素后验编码器；所述将所述预测音素后验特征和所述音色标识输入所述先验编码器，得到所述预测中间变量，包括：

将所述预测音素后验特征和所述音色标识输入所述音素后验编码器，对所述预测中间变量的先验分布采样均值与方差，得到所述预测中间变量。

[权利要求 4] 根据权利要求2所述的方法，所述第二子模型还包括后验编码器；所述根据所述预测中间变量训练所述第二子模型，包括：

获取所述样本语音；

将所述样本语音输入所述后验编码器，得到真实中间变量；

计算所述预测中间变量与所述真实中间变量的相对熵损失，训练所述第二子模型。

[权利要求 5] 根据权利要求4所述的方法，所述后验编码器包括后验预测器；所述将所述样本语音输入所述后验编码器，得到真实中间变量，包括：

将所述样本语音输入所述后验预测器，对所述真实中间变量的后验分布采样均值与方差，得到所述真实中间变量。

[权利要求 6] 根据权利要求2所述的方法，所述先验编码器还包括音素后验预测器，所述音素后验预测器用于在预训练所述第二子模型的过程中，根据预训练过程中的预测中间变量预测所述预训练过程中的预测音素后验特征；所述方法还包括：

在所述第二子模型的所述预训练过程中，计算所述预训练过程中的预测音素后验特征与所述预训练过程中的真实音素后验特征的损失函数，训练所述第二子模型。

[权利要求 7] 根据权利要求2所述的方法，所述方法还包括：

获取与所述样本音素和风格标识对应的基频特征，所述基频特征是通过所述第一子模型基于所述风格标识对所述样本音素进行特征提取得到的，所述风格标识用于标识语音风格；

所述将所述预测音素后验特征和所述音色标识输入所述先验编码器，得到所述预测中间变量，包括：

将所述预测音素后验特征、所述音色标识和所述基频特征输入所述先验编码器，得到与所述风格标识对应的所述预测中间变量。

[权利要求 8] 根据权利要求2所述的方法，所述解码器包括逆傅里叶变换解码器；所述通过所述解码器对所述预测中间变量进行逆傅里叶变换，得到所述预测语音，包括：

通过所述逆傅里叶变换解码器对所述预测中间变量进行逆傅里叶变换，得到所述预测语音。

[权利要求 9] 根据权利要求8所述的方法，所述逆傅里叶变换解码器包括多个一维卷积层，最后一个所述一维卷积层与逆傅里叶变换层连接。

[权利要求 10] 根据权利要求2至9任一所述的方法，所述解码器还包括判别器，所述判别器与所述语音合成模型中除所述判别器以外的部分组成生成对抗网络；所述方法还包括：

将所述预测语音输入所述判别器，得到所述预测语音的判别结果，所述判别结果用于反映所述预测语音为真实的信息或预测的信息；

根据所述判别结果与所述预测语音的真实来源确定生成对抗损失，训练所述生成对抗网络。

[权利要求 11] 根据权利要求1至9任一所述的方法，所述第一子模型包括文本编码器、时长规整器以及后处理网络；

所述将所述样本音素输入所述语音合成模型的第一子模型，得到所述样本音素的预测音素后验特征，包括：

获取所述样本音素中各音素对应的真实发音时长特征；

通过所述文本编码器对所述样本音素进行编码，得到所述样本音素的隐层特征；

通过所述时长规整器根据所述样本音素中各音素对应的真实发音时长特征，对所述样本音素的隐层特征进行扩帧处理；

通过所述后处理网络对扩帧后的所述样本音素的隐层特征进行卷积处理，得到所述样本音素的所述预测音素后验特征。

[权利要求 12] 根据权利要求11所述的方法，所述第一子模型还包括时长预测器，所述方法还包括：

通过所述时长预测器对所述样本音素的隐层特征进行预测，得到所述样本音素中各音素对应的预测发音时长特征；

计算所述预测发音时长特征和所述真实发音时长特征的损失函数，训练所述第一子模型，在训练完成后，用于输入所述时长规整器的所述真实发音时长特征被替换为所述时长预测器得到的预测发音时长特征。

[权利要求 13] 根据权利要求11所述的方法，所述方法还包括：

获取风格标识；

所述通过所述文本编码器对所述样本音素进行编码，得到所述样本音素的隐层特征，包括：

通过所述文本编码器根据所述风格标识对所述样本音素进行编码，得到与所述风格标识对应的所述样本音素的隐层特征。

[权利要求 14] 根据权利要求13所述的方法，所述第一子模型还包括基频预测器；所述方法还包括：

通过所述基频预测器对所述样本音素的隐层特征进行预测，得到与所述风格标识和所述样本音素对应的基频特征，所述基频特征用于输入所述第二子模型得到与所述风格标识对应的所述预测语音。

[权利要求 15] 一种语音合成方法，所述方法由计算机设备执行，所述计算机设备中包括通过权利要求1至14任一所述的方法训练得到的所述语音合成模型，所述方法包括：

获取目标用户的目标音素以及所述目标用户的音色标识，所述目标音素基于目标文本确定，所述音色标识用于标识所述目标用户的音色；

将所述目标音素输入所述语音合成模型的第一子模型，得到所述目标音素的音素后验特征；

将所述音素后验特征以及所述音色标识输入所述语音合成模型的第二子模型，得到与所述目标文本和所述音色标识对应的目标语音。

[权利要求 16] 一种语音合成模型的训练装置，所述装置包括：

获取模块，用于获取目标用户的样本音素以及所述目标用户的音色标识，所述样本音素基于所述目标用户的样本语音对应的样本文本确定，所述音色标识用于标识所述目标用户的音色；

输入输出模块，用于将所述样本音素输入所述语音合成模型的第一子模型，得到所述样本音素的预测音素后验特征，所述预测音素后验特征用于反映所述样本音素中各音素的特征以及所述样本音素中各音素的发音时长特征；

所述输入输出模块，还用于将所述预测音素后验特征以及所述音色标识输入所述语音合成模型的第二子模型，得到与所述样本文本和所述音色标识对应的预测语音，所述第二子模型是通过预测所述预测音素后验特征与所述预测语音的预测中间变量得到所述预测语音的，所述预测中间变量用于反映所述预测语音的频域特征；

训练模块，用于根据所述预测音素后验特征训练所述第一子模型；以及根据所述预测语音和所述预测中间变量训练所述第二子模型。

[权利要求 17] 一种语音合成装置，所述装置中包括通过权利要求16所述的装置训练得到的所述语音合成模型，所述装置包括：

获取模块，用于获取目标用户的目标音素以及所述目标用户的音色标识，所述目标音素基于目标文本确定，所述音色标识用于标识所述目标用户的音色；

输入输出模块，用于将所述目标音素输入所述语音合成模型的第一子模型，得到所述目标音素的音素后验特征；

所述输入输出模块，还用于将所述音素后验特征以及所述音色标识输入所述语音合成模型的第二子模型，得到与所述目标文本和所述音色标识对应的目标语音。

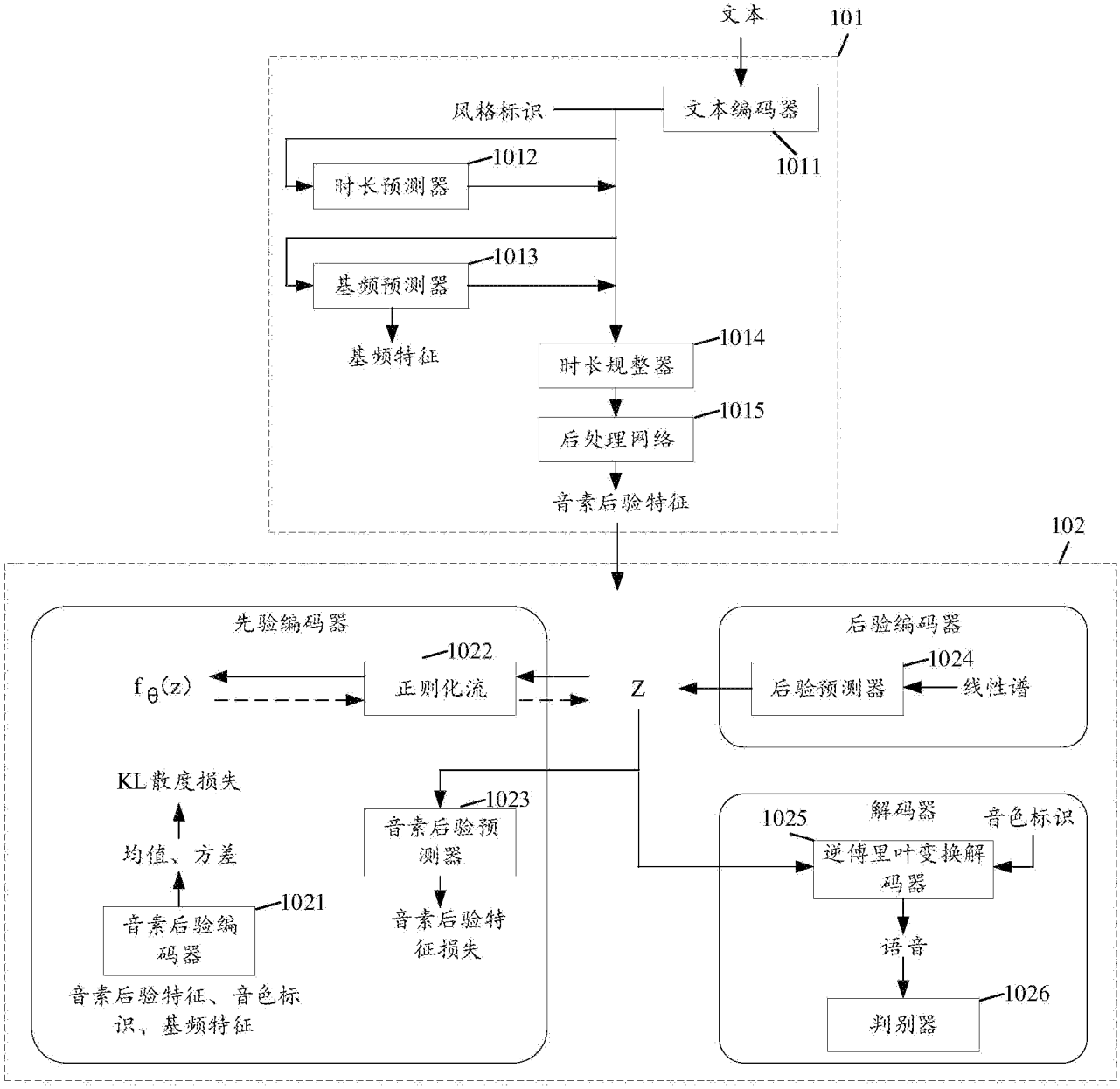
[权利要求 18] 一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或所述指令集由所述处理器加载并执行以实现如权利要求1至14任一所述的语音合成模型的训练方法，或权利要求15所述的语音合成方法。

[权利要求 19] 一种计算机可读存储介质，所述可读存储介质中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由处理器加载并执行以实现如权利要求1至14任一所述的语音合成模型的训练方法，或权利要求15所述的语音合成方法。

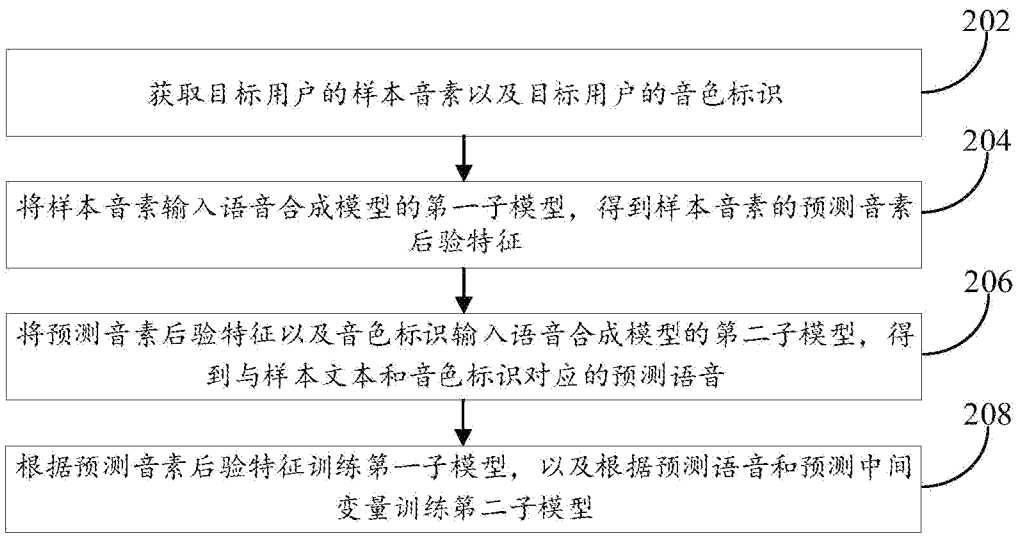
[权利要求 20] 一种计算机程序产品，所述计算机程序产品包括计算机程序，当其在计算机上运行时，使得所述计算机执行如权利要求1至14任一所述的语音合成模型的训练方法，或权利要求15所述的语音合成方法。

WO 2024/055752 PCT/CN2023/108845

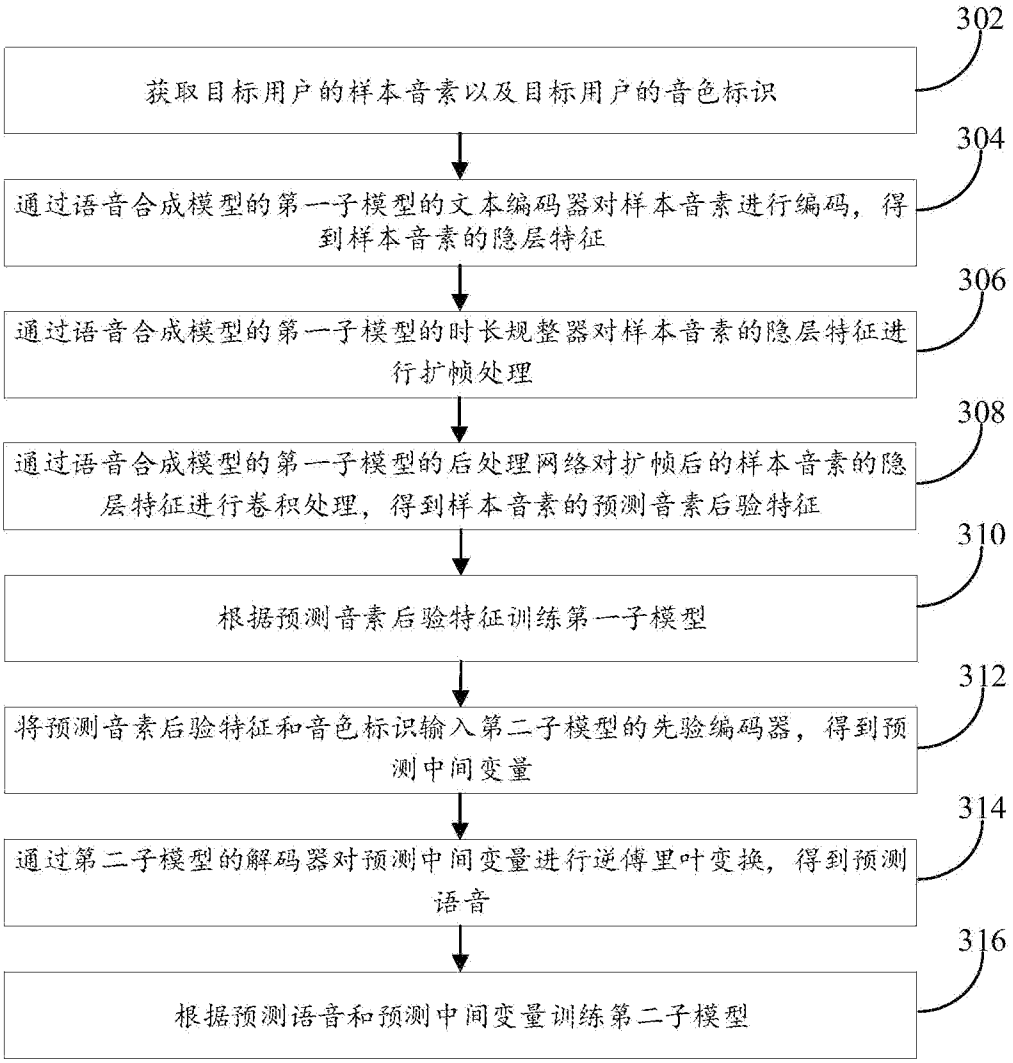
[图 1]



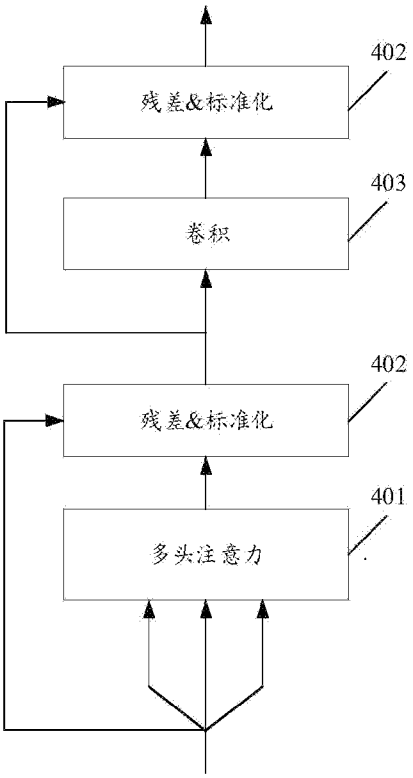

[图 2]



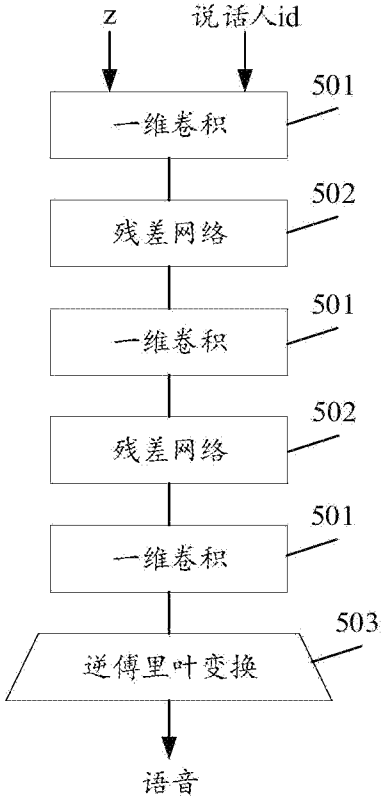
[图 3]


302
获取目标用户的样本音素以及目标用户的音色标识
304
通过语音合成模型的第一子模型的文本编码器对样本音素进行编码，得到样本音素的隐层特征
306
通过语音合成模型的第一子模型的时长规整器对样本音素的隐层特征进行扩帧处理
308
通过语音合成模型的第一子模型的后处理网络对扩帧后的样本音素的隐层特征进行卷积处理，得到样本音素的预测音素后验特征
310
根据预测音素后验特征训练第一子模型
312
将预测音素后验特征和音色标识输入第二子模型的先验编码器，得到预测中间变量
314
通过第二子模型的解码器对预测中间变量进行逆傅里叶变换，得到预测语音
316
根据预测语音和预测中间变量训练第二子模型

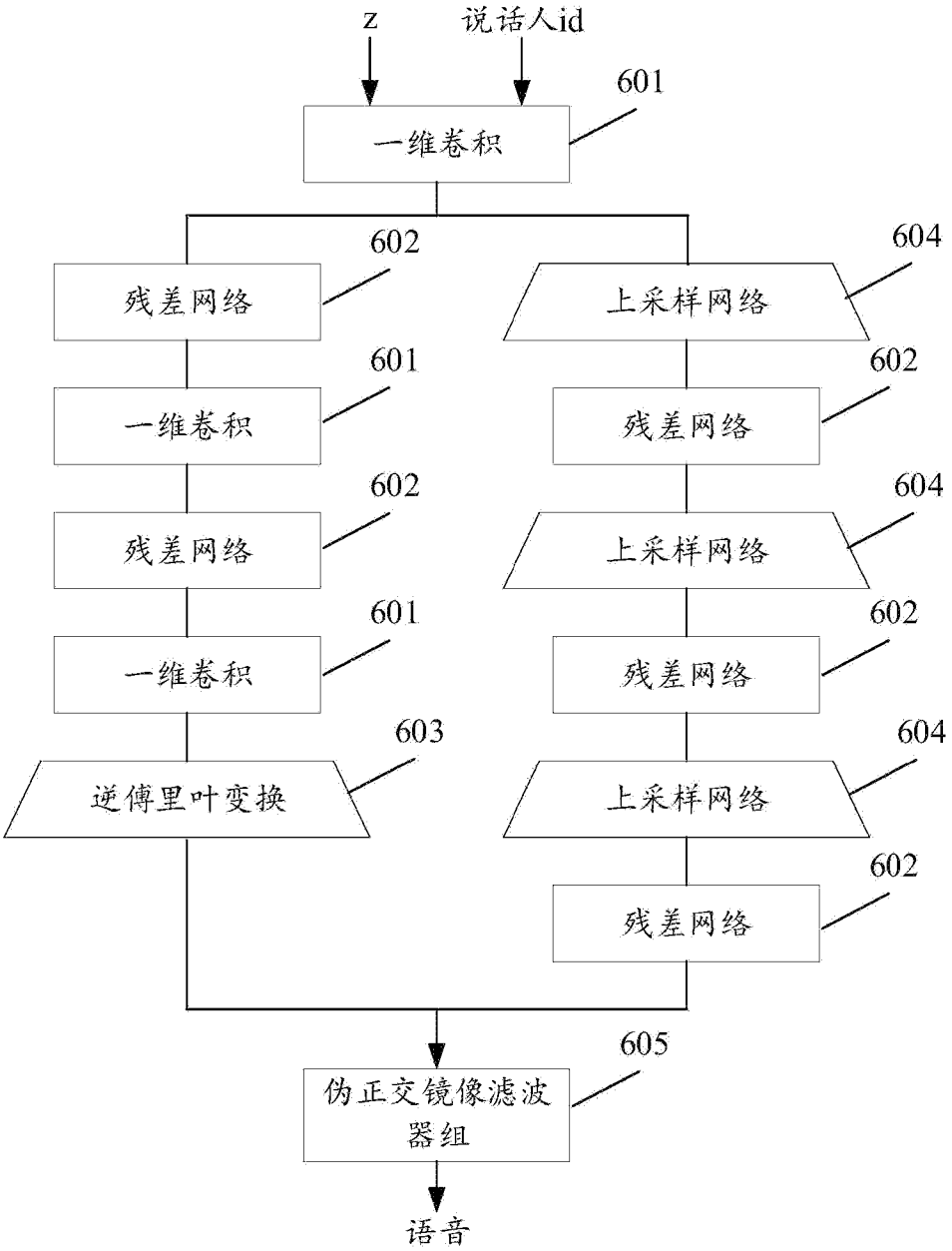
[图 4]


402
残差&标准化
403
卷积
402
残差&标准化
401
多头注意力

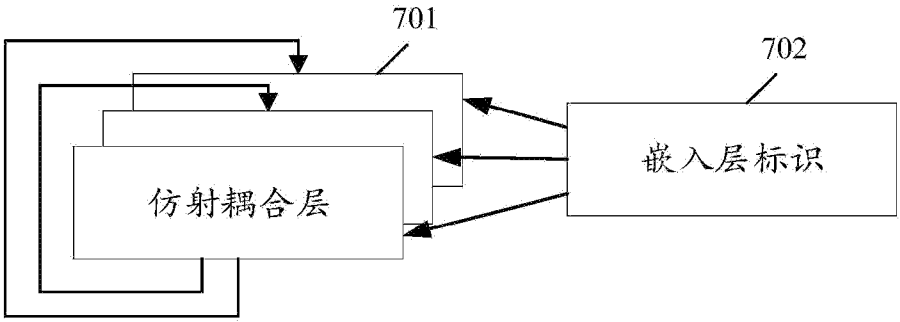
[图 5]


z
说话人id
501
一维卷积
502
残差网络
501
一维卷积
502
残差网络
501
一维卷积
503
逆傅里叶变换
语音

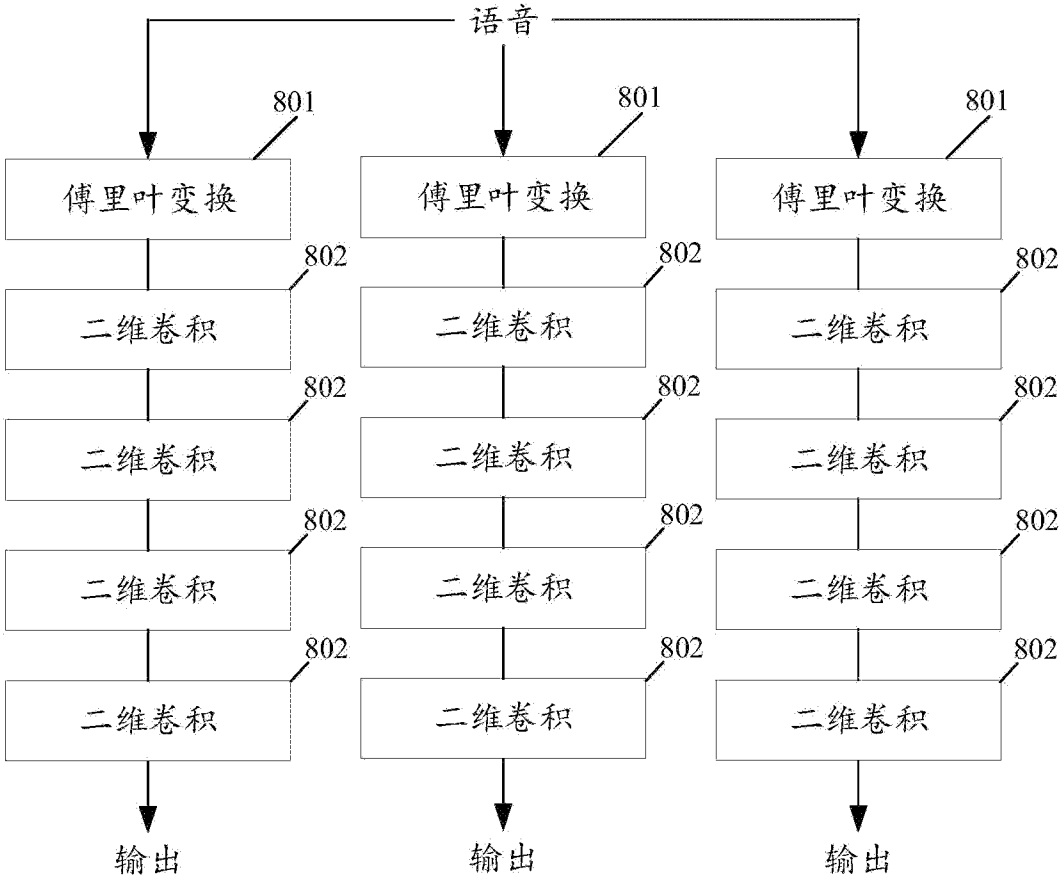
[图 6]


z
说话人id
601
一维卷积
602
残差网络
604
上采样网络
601
一维卷积
602
残差网络
602
残差网络
604
上采样网络
601
一维卷积
602
残差网络
603
逆傅里叶变换
604
上采样网络
602
残差网络
605
伪正交镜像滤波器组
语音

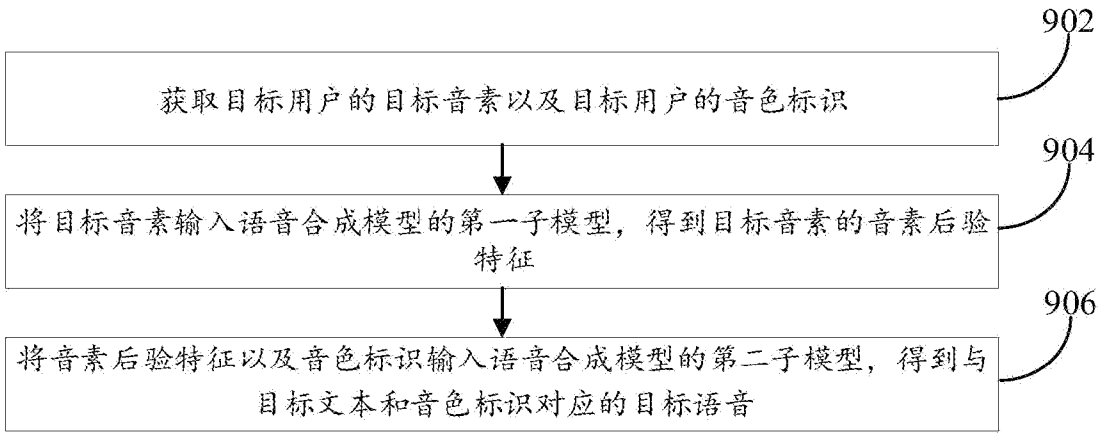
[图 7]


701
702
嵌入层标识
仿射耦合层

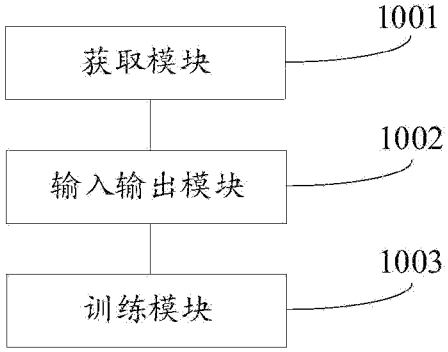
[图 8]


语音
801
傅里叶变换
802
二维卷积
输出

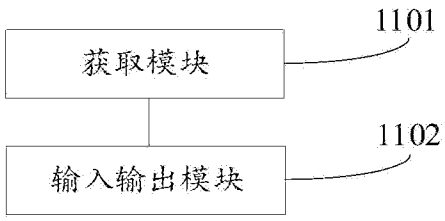
[图 9]


902
获取目标用户的目标音素以及目标用户的音色标识
904
将目标音素输入语音合成模型的第一子模型，得到目标音素的音素后验特征
906
将音素后验特征以及音色标识输入语音合成模型的第二子模型，得到与目标文本和音色标识对应的目标语音

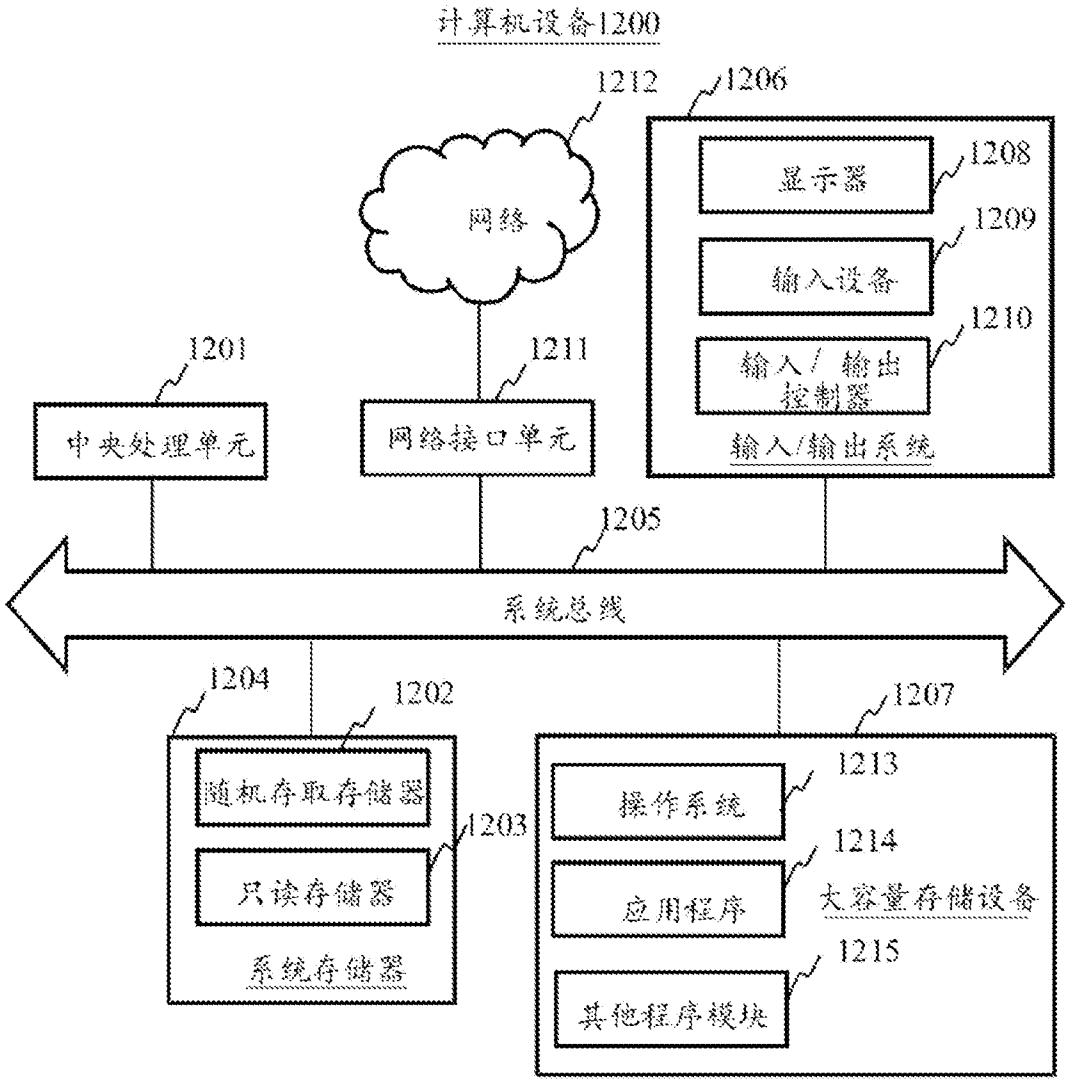
[图 10]


1001
获取模块
1002
输入输出模块
1003
训练模块

[图 11]




[图 12]



# INTERNATIONAL SEARCH REPORT

International application No.
**PCT/CN2023/108845**

## A. CLASSIFICATION OF SUBJECT MATTER

G10L 13/02(2013.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: G10L 13

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, CNTXT, DWPI, CNKI, 百度学术, BAIDU SCHOLAR: 说话人自适应, 目标说话人, 音色, 音素后验特征, speaker adaptation, speech synthesis, PPG, timbre, target speaker

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X | Kun Song et al. "AdaVITS: Tiny VITS for Low Computing Resource Speaker Adaptation" *arXiv:2206.00208v1,* 01 June 2022 (2022-06-01), parts 2-3 | 1-20 |
| A | CN 111462769 A (SPEECHX LTD.) 28 July 2020 (2020-07-28) entire document | 1-20 |
| A | CN 113793591 A (IFLYTEK CO., LTD.) 14 December 2021 (2021-12-14) entire document | 1-20 |
| A | CN 113808576 A (ALIBABA GROUP HOLDING LIMITED) 17 December 2021 (2021-12-17) entire document | 1-20 |
| A | WO 2020145472 A1 (NAVER CORP. et al.) 16 July 2020 (2020-07-16) entire document | 1-20 |

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| **13 November 2023** | **15 November 2023** |

| Name and mailing address of the ISA/CN | Authorized officer |
|---|---|
| **China National Intellectual Property Administration (ISA/CN)**<br>**China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088** | Telephone No. |

**INTERNATIONAL SEARCH REPORT**
**Information on patent family members**

International application No.
**PCT/CN2023/108845**

| Patent document cited in search report | Publication date (day/month/year) | Patent family member(s) | Publication date (day/month/year) |
|---|---|---|---|
| CN 111462769 A | 28 July 2020 | None | |
| CN 113793591 A | 14 December 2021 | None | |
| CN 113808576 A | 17 December 2021 | None | |
| WO 2020145472 A1 | 16 July 2020 | JP 2023089256 A | 27 June 2023 |
| | | JP 7274184 B2 | 16 May 2023 |
| | | KR 20200092500 A | 04 August 2020 |
| | | KR 20200092501 A | 04 August 2020 |
| | | KR 102198597 B1 | 05 January 2021 |
| | | KR 102198598 B1 | 05 January 2021 |

# 国际检索报告

国际申请号
PCT/CN2023/108845

A. 主题的分类

G10L 13/02(2013.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC：G10L 13

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNABS，CNTXT，DWPI，CNKI，百度学术：说话人自适应，目标说话人，音色，音素后验特征，speaker adaptation, speech synthesis, PPG, timbre, target speaker

C. 相关文件

| 类 型* | 引用文件，必要时，指明相关段落 | 相关的权利要求 |
|---|---|---|
| X | Kun Song 等. "AdaVITS: Tiny VITS for Low Computing Resource Speaker Adaptation" arXiv:2206.00208v1, 2022年6月1日 (2022 - 06 - 01)，<br>第2-3部分 | 1-20 |
| A | CN 111462769 A (深圳市声希科技有限公司) 2020年7月28日 (2020 - 07 - 28)<br>全文 | 1-20 |
| A | CN 113793591 A (科大讯飞股份有限公司) 2021年12月14日 (2021 - 12 - 14)<br>全文 | 1-20 |
| A | CN 113808576 A (阿里巴巴集团控股有限公司) 2021年12月17日 (2021 - 12 - 17)<br>全文 | 1-20 |
| A | WO 2020145472 A1 (NAVER CORP. 等) 2020年7月16日 (2020 - 07 - 16)<br>全文 | 1-20 |

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

| 国际检索实际完成的日期 | 国际检索报告邮寄日期 |
|---|---|
| 2023年11月13日 | 2023年11月15日 |
| ISA/CN的名称和邮寄地址<br>中国国家知识产权局<br>中国北京市海淀区蓟门桥西土城路6号 100088 | 受权官员<br>游晓梅<br>电话号码 (+86) 010-62089539 |

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/108845

| 检索报告引用的专利文件 | | | 公布日<br>(年/月/日) | 同族专利 | | | 公布日<br>(年/月/日) |
|---|---|---|---|---|---|---|---|
| CN | 111462769 | A | 2020年7月28日 | 无 | | | |
| CN | 113793591 | A | 2021年12月14日 | 无 | | | |
| CN | 113808576 | A | 2021年12月17日 | 无 | | | |
| WO | 2020145472 | A1 | 2020年7月16日 | JP | 2023089256 | A | 2023年6月27日 |
| | | | | JP | 7274184 | B2 | 2023年5月16日 |
| | | | | KR | 20200092500 | A | 2020年8月4日 |
| | | | | KR | 20200092501 | A | 2020年8月4日 |
| | | | | KR | 102198597 | B1 | 2021年1月5日 |
| | | | | KR | 102198598 | B1 | 2021年1月5日 |