(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119923686 A
(43)申请公布日 2025.05.02

(21)申请号 202380068133.X
(22)申请日 2023.09.07
(30)优先权数据
2022-164477 2022.10.13 JP
(85)PCT国际申请进入国家阶段日
2025.03.21
(86)PCT国际申请的申请数据
PCT/JP2023/032711 2023.09.07
(87)PCT国际申请的公布数据
WO2024/080044 EN 2024.04.18
(71)申请人 索尼集团公司
地址 日本东京
(72)发明人 成田岳 清水惇一 赤间健人 小口真太郎 山本昂平 岸治彦
(74)专利代理机构 北京康信知识产权代理有限责任公司 11240
专利代理师 段影影
(51)Int.Cl.
G10H 1/00(2006.01)
G06N 3/045(2006.01)
G10H 1/18(2006.01)

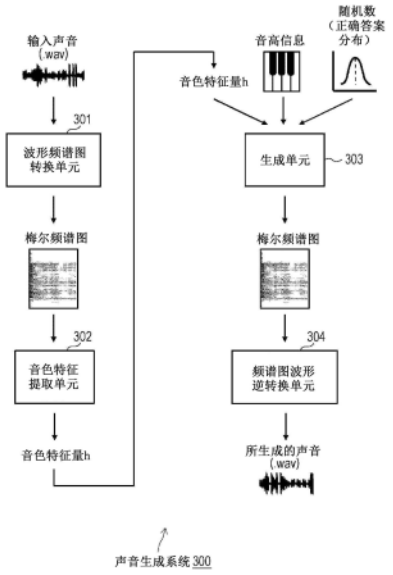
权利要求书2页 说明书25页 附图27页

(54)发明名称

用于生成对手网络音乐合成器的图形用户界面

(57)摘要

信息处理系统,接收输入声音和音高信息;从输入声音提取音色特征量;并且基于音色特征量和音高信息生成具有音高的乐器声音的信息。



CN 119923686 A

1.一种信息处理系统,包括:

电路系统,被配置为接收输入声音和音高信息;

从所述输入声音提取音色特征量;并且

基于所述音色特征量和所述音高信息来生成具有音高的乐器声音的信息。

2.根据权利要求1所述的信息处理系统,其中,

所述电路系统被配置为使用学习模型来生成所述乐器声音的信息。

3.根据权利要求1所述的信息处理系统,其中,

所述电路系统被配置为以通过对所述输入声音进行预处理而生成的信息和所述音高信息作为实例条件使用学习模型来生成所述乐器声音的信息。

4.根据权利要求1所述的信息处理系统,其中,

所述电路系统被配置为提取所述音色特征量以使得不保留音高信息。

5.根据权利要求1所述的信息处理系统,其中,

所述电路系统被配置为使用音色特征提取器来提取所述音色特征,所述音色特征提取器已执行关于音高的对手学习。

6.根据权利要求1所述的信息处理系统,其中,所述电路系统被配置为:

将所述输入声音转换成梅尔频谱图;并且

基于来自所述输入声音的梅尔频谱图提取所述输入声音的音色特征量。

7.根据权利要求6所述的信息处理系统,其中,所述电路系统被配置为:

使用所述音色特征量和所述音高信息生成具有音高的乐器声音的梅尔频谱图,并且

基于所述梅尔频谱图构造音频波形。

8.根据权利要求7所述的信息处理系统,其中,所述电路系统被配置为:

将所述梅尔频谱图转换成线性频谱图中的频率刻度;

恢复所述线性频谱图的相位;并且

在恢复所述线性频谱图的相位之后,对所述线性频谱图执行傅立叶逆转换。

9.根据权利要求8所述的信息处理设备,其中,所述电路系统被配置为:

将校正为非负值的解设置为无非负值的最小二乘法的解作为迭代计算的初始值;并且

根据依照梯度法重复更新并校正为非负值的迭代方法执行频率刻度转换。

10.根据权利要求1所述的信息处理系统,其中,所述电路系统被配置为:

接收多个输入声音的输入;

从每个输入声音提取音色特征量;并且

基于通过混合所述多个输入声音的音色特征量而获得的音色特征量和音高信息来生成乐器声音信息。

11.根据权利要求10所述的信息处理系统,其中,所述电路系统被配置为:

接收关于多个输入声音的混合比例的信息;以及

基于通过混合多个音色特征量获得的音色特征量、所述混合比例和所述音高信息来生成乐器声音信息。

12.根据权利要求1所述的信息处理系统,其中,

所述电路系统被配置为基于用户操作接收所述输入声音和所述音高信息。

13.根据权利要求1所述的信息处理系统,其中,

所述电路系统被配置为输出所述乐器声音的信息。

14.根据权利要求1所述的信息处理系统,其中,

所述电路系统被配置为显示用户界面,所述用户界面被配置为接收与所述输入声音和所述音高信息相对应的用户输入。

15.根据权利要求14所述的信息处理系统,其中,

所述用户界面被配置为接收与第一输入声音相对应的第一输入和与第二输入声音相对应的第二输入,并且

所述用户界面被配置为接收与所述第一输入声音和所述第二输入声音相对应的混合比例。

16.根据权利要求15所述的信息处理系统,其中,

所述图形用户界面包括至少第一图形和第二图形,其中,

所述第一图形被配置为接收与所述第一输入声音相对应的所述第一输入和与所述第二输入声音相对应的所述第二输入,并且所述第二图形被配置为接收所述音色特征量。

17.一种信息处理方法,包括:

接收输入声音和音高信息;

从所述输入声音提取音色特征量;并且

基于所述音色特征量和所述音高信息生成具有音高的乐器声音的信息。

18.一个或多个非瞬态计算机可读介质,当由电路系统执行时,使所述电路系统:

接收输入声音和音高信息;

从所述输入声音提取音色特征量;并且

基于所述音色特征量和所述音高信息生成具有音高的乐器声音的信息。

19.一种声音生成系统,包括:

终端,被配置为请求生成乐器声音;以及

信息处理设备,生成乐器声音,其中,

所述信息终端被配置为:

接收输入声音和音高信息;

从所述输入声音提取音色特征量;并且

基于所述音色特征量和所述音高信息来生成具有音高的乐器声音的所述信息。

20.一种信息终端,包括:

通信接口,被配置为与信息处理系统通信;以及

用户界面,被配置为接收与包括生成输入声音和音高信息的乐器声音相关的指示,其中,

所述通信接口被配置为将用于生成包括所述输入声音和所述音高信息的所述乐器声音的请求发送至所述信息处理系统,并且

从所述信息处理系统接收所述乐器声音的信息。

用于生成对手网络音乐合成器的图形用户界面

[0001] 相关申请

[0002] 本申请要求于2022年10月13日提交的日本在先专利申请JP 2022-164477的权益，通过引用将其全部内容结合在此。

技术领域

[0003] 在本说明书(在下文中，“本公开”)中公开的技术涉及执行与音乐制作有关的信息处理的信息处理设备和信息处理方法、计算机程序、声音生成系统和信息终端。

背景技术

[0004] 人工智能(AI)技术的发展是显著的，并且使用学习模型的图像、语音等的识别技术已经变得普遍。最近，还开发了图像生成技术，其使用生成不良网络(GAN)来生成复杂的图像。此外，还正在寻找利用AI技术进行音乐制作的方法。例如，已经提出使用反映乐器声音的特征的深度神经网络(DNN)来增强声源的音乐声音增强装置(参见专利文献1)、使用利用GAN或变分自动编码器(VAE)生成的学习模型来自动生成各种音乐的信息处理方法(参见专利文献2)等。

[0005] [引用列表]

[0006] [专利文献]

[0007] 专利文献1:JP 2019-78864A

[0008] 专利文献2:JP 2019-78864A

[0009] 非专利文献

[0010] 非专利文献1:Arantxa Casanova,Marl'ene Careil,Jakob Verbeek,Michal Drozdzal和Adririana Romero-Soririano,“Instance-conditioned GAN”,in Advances in Neural Information Processing Systems(NeurIPS),2021中。

[0011] 非专利文献2:Jesse Engel,CiCinjon Resnick,Adam Roberts,Sander Dieleman,Mohammad Norouzi,Douglas Eck,and Karen Simonyan,“Neural audio synthesis of musical notes with wavenet autoencoders”,in International Conference on Machihine Learning.PMLR,2017,第1068-1077页。

[0012] 非专利文献3:Aaron van den Oord,Sander Dieleman,Heiga Zen,Karen Simonyan,Oririol Vinyals,Alex Graves,Nal Kalchbrenner,Andrew Senior,and Koray Kavukcuoglu,“Wavenet:A generative model for raw audio”,arXiv preprirint arXiv:1609.03499,2016。

[0013] 非专利文献4:Jesse Engel,Kumar Kririshna Agrawal,Shuo Chen,Ishaan Gulrajani,Chriris Donahue,and Adam Roberts,“Gansynth:Adversaririal neural audio synthesis”,in International Conference on Learning Representations,2018。

[0014] 非专利文献5:Sean Vasquez and Mike Lewis,“Melnet:A generative model

for audio in the frequency domain",arXiv preprirint arXiv:1906.01083,2019。

发明内容

[0015] 技术问题

[0016] 利用计算机生成音乐的两种途径主要是直接合成包括旋律和伴奏的音乐的声音的方法以及合成单声道乐器声音并播放乐器数字接口(MIDI)的方法。在前者中,虽然可以端对端地生成音乐,但是存在生成的可控制性低的问题。另一方面,后者具有以下优点:MIDI的生成和音色的设计可被独立地控制,并且所生成的声音的质量高。

[0017] 因此,例如,本发明提供一种信息处理设备和信息处理方法、计算机程序、声音生成系统和信息终端,其执行与可用于MIDI播放的乐器声音的生成有关的信息处理。

[0018] 问题的解决方案

[0019] 鉴于以上问题做出本公开,并且其第一方面是一种信息处理系统,包括:电路系统,被配置为接收输入声音和音高信息;从输入声音提取音色特征量;以及基于音色特征量和音高信息生成具有音高的乐器声音的信息。

[0020] 电路系统使用学习模型来生成乐器声音的信息。

[0021] 电路系统使用在对输入声音进行预处理之后的信息和音高信息作为实例条件使用学习模型来生成具有音高的乐器声音的信息。

[0022] 此外,电路系统被配置为提取输入声音的音色特征量,使得不保留音高信息。

[0023] 此外,电路被配置为使用音色特征提取器来提取音色特征,音色特征提取器已执行关于音高的对手学习。

[0024] 本发明的另一方面是针对一种信息处理方法,其包括:接收输入声音及音高信息;从输入声音提取音色特征量;基于音色特征量和音高信息生成具有音高的乐器声音的信息

[0025] 本公开的另一方面针对一个或多个非暂态计算机可读介质,当由电路系统执行时,该非暂态计算机可读介质使电路系统:接收输入声音和音高信息;从输入声音提取音色特征量;基于音色特征量和音高信息来生成具有音高的乐器声音的信息

[0026] 计算机程序可以通过定义以计算机可读格式描述的计算机程序以便在计算机上实现预定处理来获得。计算机程序可以通过以计算机可读格式提供的存储介质、通信介质(例如,诸如光盘、磁盘、半导体存储器等的存储介质)或者诸如网络等的通信介质提供给能够执行各种程序或者代码的计算机。然后,通过经由任何介质在计算机中安装根据本公开的第三方面的计算机程序,在计算机上施加协作动作,并且可以获得与根据本公开的第一方面的信息处理设备的操作和效果类似的操作和效果。

[0027] 此外,本发明的另一方面是针对一种声音生成系统,其包括:终端,被配置为请求生成乐器声音;以及信息处理设备,生成乐器声音,其中,信息终端被配置为接收输入声音和音高信息;从输入声音提取音色特征量;并且基于音色特征量和音高信息来生成具有音高的乐器声音的信息。

[0028] 然而,在此描述的"系统"是指多个设备(或实现特定功能的功能模块)的逻辑组件,并且这些设备或功能模块中的每一个可以在或可以不在单个壳体中。即,包括多个部件或功能模块的一个装置以及多个装置的组件对应于"系统"。

[0029] 此外,本公开的另一方面涉及一种信息终端,包括:通信接口,被配置为与信息处

理系统通信;以及用户界面,被配置为接收与包括生成输入声音和音高信息的乐器声音相关的指示,其中,通信接口被配置为将生成包括输入声音和音高信息的乐器声音的请求发送至信息处理系统,并且从信息处理系统接收乐器声音的信息。

[0030] 本发明的有利效果

[0031] 根据本公开,可以提供一种信息处理设备和信息处理方法、计算机程序、声音生成系统以及信息终端,其执行生成具有反映任意输入声音的特征的音高的乐器声音的信息处理。

[0032] 应注意,在本说明书中描述的效果仅是实施例,并且本公开要带来的效果不限于此。此外,除了上述效果之外,本公开在某些情况下可进一步表现出另外的效果。

[0033] 通过基于以下描述的实施方式和附图的更详细描述,本公开的其他目的、特性和优点将变得显而易见。

附图说明

[0034] 图1是示出本公开的概要的示图。

[0035] 图2是示出用于基于通过混合两个输入声音获得的启示来生成乐器声音的机制的简图。

[0036] 图3是说明生成具有反映输入声音的特性的音高的乐器声音的乐器声音生成系统300的功能配置的示图。

[0037] 图4是示出用于生成具有反映输入声音的特性的音高的乐器声音的处理步骤的流程图。

[0038] 图5是说明在学习音色特征提取器时的工作流的实施例的示图。

[0039] 图6是说明在执行关于音高的对手学习的情况下在音色特征提取器的学习时的工作流的示图。

[0040] 图7是示出在乐器声音生成系统300的生成单元303中使用的发生器被学习的情况下的模型的概要的示图。

[0041] 图8是说明从梅尔频谱图重建音频波形数据的处理流程的示图。

[0042] 图9是示出了通过梯度法重复更新并校正为非负值的迭代法的频率刻度转换处理流程的示图。

[0043] 图10是示出在设置利用无非负值约束的最小二乘法的解的初始值的情况下通过迭代方法的频率刻度转换处理流程的示图。

[0044] 图11是示出包含客户端服务器模型1100的乐器声音生成系统300的配置的示图。

[0045] 图12是示出由客户端1102和服务器1101执行的示范性处理序列的示图。

[0046] 图13是示出根据本公开的实施方式的用于生成和再现具有音高的乐器声音的GUI屏幕的配置实施例的简图。

[0047] 图14是示出图13所示的GUI屏幕上的操作实施例的示图。

[0048] 图15是示出图13所示的GUI屏幕上的操作实施例的示图。

[0049] 图16是示出图13所示的GUI屏幕上的操作实施例的示图。

[0050] 图17是示出图13所示的GUI屏幕上的操作实施例的示图。

[0051] 图18是示出图13所示的GUI屏幕上的操作实施例的示图。

[0052] 图19是示出图13所示的GUI屏幕上的操作实施例的示图。

[0053] 图20是示出图13所示的GUI屏幕上的操作实施例的示图。

[0054] 图21是示出图13所示的GUI屏幕上的操作实施例的示图。

[0055] 图22是示出图13所示的GUI屏幕上的操作实施例的示图。

[0056] 图23是示出图13所示的GUI屏幕上的操作实施例的示图。

[0057] 图24是示出根据本公开的实施方式的用于生成和再现具有音高的乐器声音的GUI屏幕的另一配置实施例的简图。

[0058] 图25是示出图24所示的GUI屏幕的变形例的示图。

[0059] 图26是示出信息处理设备2000的具体硬件配置实施例的示图。

[0060] 图27是示出IC-GAN模型的概要的示图。

具体实施方式

[0061] 在以下描述中,将参考附图按照以下顺序解释本公开。

[0062] A.概述

[0063] B.声音生成系统

[0064] B-1.系统配置

[0065] B-2.系统操作

[0066] B-3.系统特征

[0067] B-4.系统操作

[0068] C.音色特征量的提取

[0069] D.具有输入声音的音高反射特性的乐器声音的生成

[0070] E.音频波形的重构

[0071] F.客户端服务器模型的操作形式

[0072] G.GUI的配置和操作实施例

[0073] G-1.第一实施例

[0074] G-2.第二实施例

[0075] H.信息处理设备的配置

[0076] I.与相关研究的比较

[0077] A.概述

[0078] 本发明是一种用于生成单声道乐器声音的技术。利用基于本公开实施方式生成的乐器声音,通过演奏MIDI的方法来实现在计算机上的音乐制作。根据使用本公开的实施方式的音乐生成方法,能够独立地控制MIDI的生成和音色的设计,具有提高生成的声音的质量的优点。

[0079] 本公开是用于生成单声道乐器声音的技术,但是可基于例如从在人类的居住空间中生成的任意输入声音获得的启示来生成乐器声音。此外,本公开生成的乐器声音不限于使用真实乐器生成的乐器声音。即,通过本公开生成的乐器声音是不太可能被辨别为乐器但不被辨别为由乐器生成的声音的声音,换言之,不被辨别为由除了乐器之外的声音生成的声音的声音。

[0080] 图1示意性地示出了本披露的轮廓。在本公开中,给出启示的任何输入声音是例如

在人类的生活环境中生成的各种声音,并且不仅包括自然声音和环境声音,而且包括预先人工生成的人造声音。例如,噪声可以是谈话的声音、卡拉OK的歌声、诸如狗或鸟的动物的哭声、诸如鹿窥视或雨声、风铃声或风声的环境声音、或用链锯、重型机器等切割或粉碎对象的噪声。为了方便计算机处理,这些可选的输入声音作为音频文件(诸如wav格式文件)被输入。

[0081] 然后,在本公开中,基于从输入声音获得的启示,生成乐器声音。具体地,根据本公开,例如,在大约一秒或几秒的相对短时间内具有指定音高的单声道乐器声音被输出为MIDI数据。本公开生成的乐器声音可以是由现有的乐器(例如,键盘乐器、打击乐器、弦乐器、风乐器或电动或电子乐器)生成的单个声音,但不限于此,并且可生成全新且独特的乐器声音。由本公开生成的独特的乐器声音是不太可能被辨别为乐器但不被辨别为由乐器生成的声音(换言之,由其他乐器生成的声音)的声音。

[0082] 在本公开中,使用深度学习的生成模型生成乐器声音作为输入声音和音高的条件。因此,根据本公开,存在用户可以自由定制乐器并且音乐可以与声音相关联的效果。这里所指的深度学习生成模型是对本公开唯一的学习模型,具体是由GAN的框架(下面也称为“本公开模型”)生成的使用实例条件的GAN(IC-GAN)的概念生成乐器声音的发生器。

[0083] 过去,作为合成乐器声音的方法,存在诸如“合成器”和“采样”的方法,“合成器”调制人工生成的周期性振荡器波形以控制音色,“采样”记录和处理实际乐器声音以表达难以由合成器合成的声学乐器的真实性。采样可直接利用任何声音来生成音乐,但不能生成全新的音色或组合多个声音的特性。

[0084] 另一方面,根据本公开,可以搜索潜在空间,生成各种各样的全新且独特的乐器声音,并通过使用深度学习的生成模型来执行组合多个声音的特性的智能声音合成处理。此外,根据本公开,可以通过在深度学习的生成模型的潜在表示中混合两个或三个或更多个任意输入声音来创建全新且独特的乐器声音。

[0085] 图2示意性示出了根据本公开的作为通过混合两个输入声音获得的启示的用于生成乐器声音的机制。在图2中所说明的实施例中,作为第一输入声音的喇叭的音频波形和作为第二输入声音的狗叫声的音频波形被捕获为wav格式的文件。首先,每一输入声音经受音色特征提取器以分别获得特征向量h_t和h_d。接下来,以由用户等指定的混合比例合成每个特征向量。然后,生成基于从第一输入声音和第二输入声音获得的激励的独特的乐器声音。具体地,由本公开模型生成的学习模型(发生器)在由用户指定的合成音色和音高的特征向量的条件下生成唯一的乐器声音。

[0086] 用户可通过调整多个输入声音的混合比例来控制发出的乐器声音。在图2的上部中所示的实施例中,第一输入声音和第二输入声音的特征向量h_t和h_d以0.5:0.5的比率混合以生成组合的特征向量h_{s1}。然后,通过使用深度学习的生成模型从特征向量h_{s1}生成类似于具有狗的音色的喇叭的唯一乐器声音S_1。此外,在图2的下部中所示的实施例中,第一输入声音和第二输入声音的特征向量h_t和h_d以0.8:0.2的比率混合以生成组合的特征向量h_{s2}。然后,使用深度学习的生成模型生成类似于用喇叭的音色拨打的狗的独特的乐器声音S_2。

[0087] 使用根据本公开的深度学习的生成模型的乐器声音的生成技术具有以下特征(1)至(4)。

[0088] (1)用户的任意声音的启示可被输入至如采样器的生成模型,并且被有效地推广

至各种输入声音。

[0089] (2)可以通过使用深度学习的生成模型经由潜在空间混合多个声音。

[0090] (3)可生成具有准确且一致的音色的宽范围的音高。

[0091] (4)可以在交互时间内生成乐器声音。

[0092] 本公开从使能对模型的输入和提高输入的泛化特性的角度应用采用IC-GAN生成的生成模型。IC-GAN通过数据点(即实施例)的特征量调节发生器和鉴别器,将整个数据的分布表示为实施例附近的局部分布的叠加。IC-GAN是一种新的GAN学习技术,可以实现对模型的输入和模式崩溃的避免。

[0093] B.声音生成系统

[0094] B-1.系统配置

[0095] 图3示意性地说明根据本发明的生成音高反映输入声音的特性的乐器声音的乐器声音生成系统300的功能配置。

[0096] 参见图3,乐器声音生成系统300包含波形频谱图转换单元301、音色特征提取单元302、生成单元303和频谱图波形逆转换单元304。其中,音色特征提取单元302和生成单元303使用DNN实现。乐器声音生成系统300接收包括短时音频波形(wav文件)、音高和随机数的输入声音,并输出具有反映输入声音的特征的音高的乐器声音。输出的乐器声音具有大约一秒或几秒的长度。

[0097] 乐器声音生成系统300接收包括作为wav格式文件的音频波形的输入声音。频谱图转换单元301通过短时傅里叶转换生成输入声音的线性频谱图,且进一步在线性频谱图上执行对数尺度转换以生成梅尔频谱图。在两个或更多个输入声音被输入到乐器声音生成系统300的情况下,频谱图转换单元301将每个输入声音的音频波形转换成梅尔频谱图。

[0098] 在此,频谱图对应于所谓的声纹,其中,通过傅立叶转换从音频波形提取音频信号的频率分量和振幅分量而获得的各个音频数据段(帧)的频谱沿着时间轴布置。在附属于本说明书的附图中,频谱图被图示为二维图,其中用阴影使时间分量和频率分量的每中的信号分量的强度(振幅)可视化。此外,梅尔频谱图是通过以下方式计算的对数梅尔频谱图:将梅尔滤波器组应用于线性频谱图,该梅尔滤波器组在梅尔标度中以相等的间隔仅提取特定的频带,集中于以下事实:人耳没有直接听到实际频率的声音,并且听到接近可听范围的上限的声音低于实际声音。另外,融化标度是基于人类听觉(即,如何听到声音)的标度。

[0099] 音色特征提取单元302从可视化和表达输入声音的音频波形的梅尔频谱图提取音色特征量h。在两个或两个以上输入声音输入到乐器声音生成系统300的情况下,音色特征提取单元302提取音色特征量h_1、h_2、...从针对每个输入声音的梅尔频谱图(图3中未示出)。

[0100] 音色特征提取单元302使用例如由卷积神经网络(CNN)配置且预先学习的学习模型来提取音色特征量以从梅尔频谱图(图像信息)提取音色特征量。另外,音色特征量具体为n维(这里n为正整数)特征向量。

[0101] 生成单元303使用输入声音的音色特征量、音高和随机数作为输入来生成具有反映输入声音的特征的音高的乐器声音的梅尔频谱图。在两个或两个以上输入声音输入到乐器声音生成系统300且音色特征提取单元302提取多个音色特征量(特征向量)h_1、h_2、...的情况下从每一输入声音的梅尔频谱图,将在指定混合比例下的音色特征量的混合输入到生成单元303。

[0102] 生成单元303使用深度学习的生成模型生成具有音高的乐器声音的梅尔频谱图。具体来说,生成单元303以输入声音的音色特征量和音高作为实例条件使用由本发明模型生成的学习模型(发生器)来生成乐器声音的梅尔频谱图,所述乐器声音的梅尔频谱图具有反映输入声音的特征的音高。

[0103] 频谱图波形逆转换单元304对生成单元303生成的梅尔频谱图执行傅立叶逆转换以重建包括例如wav格式文件的音频波形数据。重构的音频波形具有大约一秒或几秒的长度。存在从梅尔频谱图到音频波形的转换处理慢的问题,但是稍后将详细描述这一点。

[0104] B-2.系统操作

[0105] 图4以流程图的形式说明用于生成乐器声音的处理过程,其中音高反映乐器声音生成系统300中的输入声音的特征。

[0106] 首先,由用户指定的目标输入声音被输入到乐器声音生成系统300(步骤S401)。在该步骤中,例如,指定作为输入声音的声源的wav格式文件的文件名。此外,在用户指定两个以上输入声音的情况下,在该步骤中获取每个输入声音的wav格式文件。

[0107] 接下来,频谱图转换单元301将输入声音的音频波形转换成梅尔频谱图(步骤S402)。即,频谱图转换单元301通过傅立叶转换生成输入声音的线性频谱图,且进一步在线性频谱图上执行对数尺度转换以生成梅尔频谱图。在步骤S401中已输入两个以上输入声音的情况下,频谱图转换单元301在步骤S402中针对所有输入声音生成梅尔频谱图。

[0108] 接下来,音色特征提取单元302从输入声音的梅尔频谱图提取音色特征量h(步骤S403)。

[0109] 在步骤S401中已输入两个或两个以上输入声音的情况下(步骤S404中的是),在步骤S403中,音色特征提取单元302从所有输入声音的梅尔频谱图提取音色特征量h_1、h_2、...,且进一步混合相应音色特征量h_1、h_2、...以生成音色特征量h(步骤S405)。

[0110] 在针对每一输入声音指定混合比例的情况下,在步骤S405中,对每一输入声音的音色特征量h_1、h_2、...进行加权平均且根据指定的混合比例进行混合。此外,在不指定混合比例的情况下,可仅对各个输入声音的音色特征量h_1、h_2、...进行平均以执行混合处理。这里,当在步骤S401中输入N个输入声音时,分别从每个输入声音的梅尔频谱图生成音色特征量h_1,h_2,...,h_N,并且指定第i个输入声音的混合比例r_i(其中,$r_1+r_2+...+r_N=1$),在步骤S405中可根据以下表达式(1)生成混合音色特征量h。

[0111] [数学式1]

[0112] $$h = \sum_{i=1}^{N} r_i h_i \Big/ N \qquad \cdots(1)$$

[0113] 接下来,将由用户指定的要生成的乐器声音的音高信息输入到乐器声音生成系统300(步骤S406)。然而,在步骤S401中,可将音高信息与输入声音同时输入到乐器声音生成系统300。

[0114] 然后,生成单元303根据在步骤S403或S405中获得的音色特征量h和在步骤S406中获得的音高信息,生成具有反映输入声音的特征的音高的乐器声音的梅尔频谱图(步骤S407)。具体来说,使用由本发明模型生成的学习模型(发生器),生成单元303以输入声音的音色特征量和音高作为实例条件,根据由乐器声音生成系统300生成的随机数生成反映输

入声音的特性的经音高的乐器声音的梅尔频谱图。

[0115] 然后,频谱图波形逆转换单元304对由生成单元303生成的梅尔频谱图执行傅立叶逆转换,重构包括例如wav格式文件的音频波形数据,并且将该音频波形数据输出为MIDI数据(步骤S408),并且结束本处理。输出的乐器声音具有大约一秒或几秒的长度。

[0116] B-3.系统特征

[0117] 上面已经示意性地描述了声音再现系统300的配置和操作。声音再现系统300具有以下特征。

[0118] (1)使用学习模型,乐器声音生成系统300可生成具有在交互时间内反映输入声音的音色的音高的乐器声音。

[0119] (2)通过使用实例调节,可以提高生成的乐器声音的质量和生成乐器声音的能力。

[0120] (3)通过对音色特征提取器的音高执行对手学习,可改进音高准确性和音色一致性。

[0121] B-4.系统操作

[0122] 图3中所示的乐器声音生成系统300安装在包括例如计算机等的信息处理设备上。使用本公开的模型生成学习模型(发生器)的处理和使用学习模型(发生器)生成具有反映输入声音的特征的音高的乐器声音的处理具有大的计算负荷。因此,客户端服务器模型也被假设为乐器声音生成系统300的一种操作模式。

[0123] 在这种情况下,在客户端,例如,选择输入声音的wav格式文件(在选择多个输入声音的情况下,还指定每个输入声音的混合比例)并且通过用户的图形用户界面(GUI)操作指定音高,并且请求服务器生成具有反映输入声音的特征的音高的乐器声音。另一方面,在服务器侧,以从客户端侧指定的输入声音(其音色特征量)和音高作为实例条件,生成具有反映输入声音的特征的音高的乐器声音,并作为请求源返回到客户端。稍后将描述该操作形式的细节(部分F)。

[0124] C.音色特征量的提取

[0125] 如上文部分B中所描述,音色特征提取单元302从使输入声音的音频波形可视化和表达的梅尔频谱图中提取音色特征量h。具体来说,音色特征提取单元302是使用由CNN配置且预先学习的学习模型来从梅尔频谱图(图像信息)提取音色特征量的特征提取器。

[0126] 重要的是使用高质量特征提取器来学习在以下部分D中描述的实例条件的GAN。获得特征提取器的最简单方式是学习与标记的训练数据一起生活的鉴别器,并且利用紧接在最终全连接层之前的输出作为特征量。

[0127] 然而,当根据上述方法学习的特征提取器应用于音色特征提取单元302时,存在乐器声音生成系统300所生成的声音的音高准确性和音色一致性劣化的问题。例如,存在以下问题:当输入声音的音高是“Re”时,生成的声音的音高是接近于“Mi”的“Re”。以下两点被认为是该问题的原因。

[0128] (a)通过一般方法学习的特征提取器的特征量包括音高信息。

[0129] (b)由于用户指定的音高和特征量中包括的音高信息彼此干扰,(生成单元303使用的)发生器的学习变得不稳定。例如,在由用户指定的音高为C4,而由特征提取器提取的特征量包含G4作为音高信息的情况下,后续阶段的发生器不能确定可生成C4或G4的哪一乐器声音。

[0130] 因此,在本发明中,执行音色特征提取器的学习,使得可从梅尔频谱图提取其中未剩余音高信息的音色特征量。具体来说,在本发明中,对音色特征提取器执行关于音高的对手学习,使得音色特征量不保留在音色特征量中。

[0131] 图5说明学习音色特征提取器时的工作流的实例。在图5中所说明的实施例中,在音色特征提取单元302中使用的音色特征提取器501与乐器鉴别器502一起被学习。如上所述,音色特征提取器501从音频波形的梅尔频谱图提取音色特征量h。另外,乐器鉴别器502从音色特征量h鉴别具有原始音频波形的乐器。接着,比较乐器鉴别器502输出的预测分布C_{pred}与正确答案分布C_{gt},且通过错误反向传播进行音色特征提取器501和乐器鉴别器502的学习。例如,乐器鉴别器502固定且进行音色特征提取器501的学习的学习阶段以及音色特征提取器501固定且进行乐器鉴别器502的学习的学习阶段交替地重复。

[0132] 然而,在图5所示的学习方法中,难以防止音色特征提取器501所提取的特征量h中剩余音高信息。存在难以精确地生成特定音高的乐器声音的问题。这是因为,如下文部分D中所描述,发生器G和鉴别器D输入音色特征量h和音高信息p两者,且因此,如果音色特征量包含音高信息,那么使哪一音高信息是正确的混淆,且适当的学习变得困难。

[0133] 图6说明在执行关于音高的对手学习以使得不保留特征量的音高信息的情况下音色特征提取器的学习时的工作流程。在图6所说明的实施例中,在音色特征提取单元302中使用的音色特征提取器601与乐器鉴别器602和音高鉴别器603一起学习。具体地,在本实施方式中,乐器鉴别器602和音高鉴别器603同时学习,且进行学习使得不能使用由音色特征提取器601提取的音色特征量来鉴别音高,从而避免音高信息保留在由音色特征提取器601提取的音色特征量中。

[0134] 如上所述,音色特征提取器601从音频波形的梅尔频谱图提取音色特征量h。另外,乐器鉴别器602从音色特征量h鉴别具有原始音频波形的乐器。乐器鉴别器602的学习类似于图5中所说明的工作流程的情况,且在此将省略其详细描述。

[0135] 另外,在音色特征提取器601上执行关于音高的对手学习,使得在音色特征量中不保留音高信息。音高鉴别器603从音色特征量h鉴别原始音频波形的音高。

[0136] 首先,音高鉴别器603执行学习,使得可从由音色特征提取器601提取的音色特征量准确地区分原始音频波形的音调。即将音高鉴别器603输出的预测分布$C_{2,pred}$与正确答案分布$C_{2,gt}$进行比较,通过反向传播误差来学习音高鉴别器603。以此方式,在执行音高鉴别器603的学习之后,随后固定音高鉴别器603,且进行音色特征提取器601的学习,使得可生成没有音调信息剩余的音色特征量h。即,从音高鉴别器603输出的预测分布$C_{2,pred}$变为均匀分布$C_{2,uni}$,换句话说,执行音色特征提取器601的学习,使得可生成音高鉴别器603中键无法辨识的音色特征量h。

[0137] 根据图6所示的基于对手学习的音高不变音特征提取器的学习方法,具有避免GAN的学习由于特征量空间中的音色和音高信息的纠缠而不稳定的效果,并且可以提高音高的准确性和音色的一致性。

[0138] 将更具体地描述关于音色的不良学习。以特征量$f_{\varphi}(x)$(=h)作为输入,音高鉴别器603对其进行乐器鉴别和音高鉴别的浅MLP分别由C_i和C_p表示。通过交替优化以下表达式(2)和(3)中所示的损失函数的对手学习,可获得可提取不包含音高信息的音色特征量的音色特征提取器f_{φ}。

[0139] [数学式2]

[0140] $$\min_{f_\phi, C_i} CE\left(i(x), C_i\left(f_\phi(x)\right)\right) + KL\left(\frac{1}{|p|}1p \parallel C_p\left(f_\phi(x)\right)\right) \quad \cdots(2)$$

[0141] [数学式3]

[0142] $$\min_{C_p} CE\left(p(x), C_p\left(f_\phi(x)\right)\right) \quad \cdots(3)$$

[0143] 在以上的表达式(2)和(3)中,i(x)和p(x)分别表示样本x的乐器标记和音高标记。另外,CE为交叉熵,且KL为Kullback-Leibler分歧。

[0144] 上述表达式(2)的第一项更新特征提取器f_φ和乐器辨别C_i,使得能够正确地辨别乐器。另一方面,以上表达式(2)的第二项更新音色特征提取器f_φ,使得其允许使用特征量$f_\varphi(x)$来区分音高,即,关于音高的预测分布接近均匀分布。另一方面,以上表达式(3)更新C_p以便在给定特征量$f_\varphi(x)$的条件下使音高辨别准确度最大化。通过执行此对手学习,可获得能够提取音色特征量以使得不保留音高信息的音色特征提取器。

[0145] 事实上,当经学习音色特征提取器f_φ固定且之后使用上述表达式(3)仅学习鉴别器603的音高辨别时,已证实,在不使用对手学习的情况下可以17%或更大的精度执行音高辨别,而在关于音高的对手学习中音高辨别减小到2.5%。

[0146] D.具有输入声音的音高反射特性的乐器声音的生成

[0147] 如上文部分B中所描述,生成单元303使用输入声音的音色特征量、音高和随机数作为输入以生成具有反映输入声音的特征的音高的乐器声音的梅尔频谱图。具体来说,生成单元303使用由本发明模型生成的学习模型(发生器)来生成乐器声音的梅尔频谱图,所述乐器声音的梅尔频谱图具有反映输入声音的特征的音高,其中输入声音的音色特征量和音高为实例条件。

[0148] 这里,如本领域中已知的,GAN是使两个神经网络、在真实数据与人工数据之间进行区分的鉴别器(D)以及从噪声生成数据的发生器(G)竞争学习的深度学习模型。在GAN中,存在模式崩溃的问题,其中,所生成的样本的质量和多样性受到损害,因为所生成的数据被偏置到训练数据中的一部分。另一方面,IC-GAN是学习GAN的新技术,其通过将鉴别器D和发生器G调节与数据点(实例)对应的特征量并将数据点附近教导给鉴别器作为真实数据来解决模式崩溃的问题。

[0149] 图27示出了IC-GAN模型的概要(例如,参见非专利文献1)。鉴别器D和发生器G分别使用DNN安装。在附图中,作为实施例,输入图像x_i由特征提取器f_φ映射到特征量空间。然后,作为特征提取器f_φ的输出获得的图像x_i的特征量h_i被输入到发生器G和鉴别器D中的每一个。发生器G根据从实例x_i提取的特征量h_i和采样噪声(随机数)z生成图像x_g。此外,鉴别器D基于特征量h_i和作为实际采样的相邻图像x_n鉴别由发生器G生成的图像x_g。然后,发生器G使鉴别器D相互竞争地学习,能够判别生成的图像x_g和发生器G生成的相邻图像x_n,鉴别器D能够使生成的图像x_g和相邻图像x_n难以区分。结果,可获得生成精确图像X_g的发生器G,该精确图像X_g不允许判定鉴别器D中的真实性。

[0150] 在本实施方式中,特征提取器f_φ对应于音色特征提取单元302,发生器G对应于生成单元303。然后,输入图像x_i与可视化输入声音的梅尔频谱图相对应,特征量h_i与音色特征

量相对应,并且所生成的图像x_g与由发生器G生成的梅尔频谱图相对应。

[0151] 本公开模型是由GAN框架学习的生成模型,GAN框架使用IC-GAN的想法来生成乐器声音(如上所述)。图7示出在乐器声音生成系统300的生成单元303中使用的发生器被学习的情况下的本公开模型的概要。在这种情况下,发生器G从输入声音、音高信息p和噪声矢量z生成具有反映输入声音的特征的音高的乐器声音。此外,鉴别器D鉴别发生器G生成的声音的真/假。

[0152] 通过频谱图转换单元301中的短时傅里叶转换将输入声音转换成对数尺度梅尔频谱图x_i。然后,音色特征提取单元302使用上述部分C中描述的特征提取器f_φ将梅尔频谱图x_i映射到音色特征量h_i。发生器G将从标准正态分布采样的音调信息p和噪声向量z的单次向量与音色特征量h_i一起输入,以生成梅尔频谱图x_g。生成的梅尔频谱图由稍后描述的部分E中描述的频谱图波形逆转换单元304重构为音频波形。音频波形是具有反映输入声音的特性的音高的乐器声音。

[0153] 一般类调节将全部数据的分布按类的数量划分成没有重叠的多个分布。另一方面,在IC-GAN中的实例调节(见非专利文献1)试图通过将整个数据的分布分成大量具有重叠的局部分布来获得复杂的数据分布。通过使用实例x_i的特征量$h_i = f_\varphi(x_i)$和音高信息的单稳态向量p来调节发生器G和鉴别器D两者,实例x_i附近的局部分布$P(x|h_i,p)$被建模,并且整个数据的分布P(x)被表示为以下表达式(4)作为其叠加。

[0154] [数学式4]

[0155] $$P(x) = \sum_{x_i} \sum_p p(x|h_i,p) \quad \cdots(4)$$

[0156] 生成模型的学习过程遵循非专利文献1。关于输入x_i,在由学习特征提取器$f_\varphi(\cdot)$定义的特征量空间中L2距离为k的圣基obo的数据集被设置为A_i。此时,如图7所示,基于均匀分布从A_i采样相邻数据点x_j。x_j与作为真实样本的生成样本x_g一起用于鉴别器D的学习。此外,与实际采样x_j相应的基音$p(x_j)$作为条件被输入到发生器G和鉴别器D。在本实施例中,通过以下表达式(5)所示的最小-最大游戏来优化发生器G和鉴别器D。

[0157] [数学式5]

[0158] $$\min_G \max_D \mathbb{E}_{x_i \sim P(x), x_j \sim \mathcal{U}(\mathcal{A}_i)} \left[\log D\left(x_j, p(x_j), h_j\right)\right] + \mathbb{E}_{x_i \sim P(x), z \sim P(z)} \left[\log\left(1 - D\left(G\left(z, p(x_j), h_i\right), p(x_j), h_i\right)\right)\right] \quad \cdots(5)$$

[0159] E.音频波形的重构

[0160] 用于从梅尔频谱图重建音频波形数据的方法主要包括基于学习粘贴优化的两种方法。在文本-语音域中,已经积极地研究了通过学习获得声码器的方法。然而,本发明既定生成各种音色及音高的乐器声音,且不一定容易获得能够应对此生成的声音的通用声码器。另一方面,存在这样的研究结果:通过在频率方向上生成高分辨率梅尔频谱图,即使在使用基于优化的音频反转的情况下,也可以特定水平的声音质量合成包括音乐的各种声音(参见非专利文献5)。因此,在本公开中,通过采用基于优化的方法从梅尔频谱图重构音频波形数据。

[0161] 图8示意性地示出了通过基于优化的方法从梅尔频谱图重建音频波形数据的一般处理流程。

[0162] 梅尔频谱图是通过应用梅尔滤波器组计算的对数梅尔频谱图,该梅尔滤波器组基于人类听觉在线性频谱图的梅尔标度中以相等的间隔仅提取特定的频带。因此,频率刻度转换单元801将由生成单元303生成的梅尔频谱图转换成频率刻度上的线性频谱图。接下来,相位恢复单元802使用例如已知的Griffin-Lim算法来恢复线性频谱图的相位。然后,逆短时傅立叶转换单元(iSTFT)803执行逆傅立叶转换以重建音频波形。该音频波形是具有由乐器声音生成系统300生成的音高的乐器声音的音频波形。

[0163] 在图8中所示的处理流程中,具体地,频率刻度转换单元801执行梅尔频谱图到线性频谱图的频率刻度转换的处理具有计算成本高并且处理变成瓶颈的问题。从梅尔频谱图到线性频谱图的频率刻度转换可公式化为非负值约束的最小二乘问题,如在以下表达式(6)中,但是一般解决方案具有大的计算量。在以下表达式(6)中,F_{mel}是梅尔滤波器组矩阵,x_{mel}是梅尔刻度频谱图,并且x_{lin}是线性刻度频谱图。

[0164] [数学式6]

[0165] $$\min_{x_{lin}} \|F_{mel}x_{lin} - x_{mel}\|^2 \ s.t. x_{lin} \geq 0 \quad \cdots(6)$$

[0166] 此外,可以想到通过重复梯度法更新并校正为非负值的迭代法获得良好解的方法。然而,由于初始值是由随机数设定的,所以为了收敛到良好解,需要足够次数的迭代。

[0167] 图9示出通过重复梯度法更新并校正为非负值的迭代法的频率刻度转换处理流程的概要。在现有技术中,已经提供了用于基于所示出的处理流程执行频率刻度转换的软件库。在这个处理流程中,首先,初始化单元901用随机数(例如,给出时间轴和频率轴上的每个点的强度(振幅)作为随机数)初始化频谱图。然后,更新单元902通过梯度法更新时间轴和频率轴上的各点的强度(振幅),然后,校正单元903将0代入具有负值的变量。重复执行更新单元902和校正单元903的处理,直到计算结果收敛为止。如已经提到的,重复通过梯度法进行的更新和对非负值的校正的迭代方法是有效的,但是收敛速度慢。

[0168] 因此,在本公开中,基本上使用类似的迭代方法,但是使用没有非负值约束的最小二乘法的解而不是用于频谱初始化的随机数。具体而言,在得到高速计算的无限制的最小二乘法的解(见以下表达式(7))之后,将通过将解校正为非负值而得到的解设置为迭代计算的初始值,使得可以收敛到具有少量迭代的良好解。因此,通过实验已经确认该解以大约1/10的迭代次数收敛到相同的准确度。

[0169] [数学式7]

[0170] $$\min_{x_{lin}} \|F_{mel}x_{lin} - x_{mel}\|^2 \quad \cdots(7)$$

[0171] 图10示意性地示出在类似于图9的迭代方法中利用没有非负值约束的最小二乘法的解代替随机数的情况下的频率刻度转换处理流程。首先,初始化单元1001在没有限制的情况下利用最小二乘法的解初始化频谱图,并且此时,初始值校正单元1002将0代入具有负值的变量。然后,更新单元1003通过梯度法更新时间轴和频率轴上的每个点的强度(振幅),然后,校正单元1004将0代入具有负值的变量。重复执行更新单元1003和校正单元1004的处理,直到计算结果收敛为止。

[0172] 根据图10中所示的频率刻度转换方法,通过在最小二乘法的解中用0代替负值而获得的解被设置为迭代计算的初始值,使得可以通过少量迭代收敛到满意的水平。结果,乐

器声音生成系统300可以实现在交互时间内乐器声音的生成。

[0173] F.客户端服务器模型的操作形式

[0174] 乐器声音生成系统300安装在包括例如计算机等的信息处理设备上。使用本公开的模型生成学习模型(发生器)的处理和使用学习模型(发生器)生成具有反映输入声音的特征的音高的乐器声音的处理具有大的计算负荷。因此,客户端服务器模型被假设为乐器声音生成系统300的一个操作模式。

[0175] 图11示意性地说明包含客户端服务器模型1100的乐器声音生成系统300的配置。客户端服务器模型1100包括服务器1101和一个或多个客户端1102,服务器1101提供具有音高的乐器声音的生成服务,客户端1102请求生成具有音高的乐器声音。服务器1101和每个客户端1102经由诸如广域网(WAN)、局域网(LAN)或互联网的网络互连。

[0176] 客户端1102包括例如由用户使用的信息终端(边缘设备),诸如智能电话、平板电脑或个人计算机(PC)。这里提到的用户是例如使用从服务器1101提供的唯一乐器声音编写音乐或执行其他音乐活动的一般用户。在客户端1102侧,例如,经由GUI屏幕执行诸如输入声音的wav格式文件的选择和音高的指定的GUI操作。此时,在选择多个输入声音的情况下,在GUI操作中还包括每个输入声音的混合比例的指定。然后,客户端1102请求服务器1101执行生成具有反映输入声音的特征的音高的乐器声音的过程。

[0177] 服务器1101包括例如诸如计算机的信息处理设备,并且配备有乐器声音生成系统300的主要组件301至304。响应于来自客户端1102的请求,服务器1101使用学习模型(发生器)生成具有反映输入声音的特征的音高的乐器声音,并将具有音高的乐器声音返回至客户端1102作为请求源,所述学习模型是使用本公开模型以指定的输入声音和音高作为实例条件生成的。此外,在从客户端102请求多个输入声音的情况下,服务器1101通过使用通过以由客户端1102指定的混合比例组合各个输入声音的特征向量而获得的特征向量,生成具有音高的乐器声音,并将乐器声音返回至客户端1102。

[0178] 图12示出了由客户端1102和服务器1101执行的示范性处理序列。

[0179] 在客户端1102侧,用户通过GUI操作指定用作用于生成乐器声音的声源的输入声音和将生成的乐器声音的音高(SEQ 1201)。

[0180] 以从预设中指定相应wav格式文件的文件名的形式指定输入声音。例如,在服务器1101侧预先准备的wav格式文件,服务器1101可以在客户端1102侧获取的可以指定的wav格式文件,以及客户端1102可以上传到服务器1101的wav格式文件也可以被指定为输入声音的预设。此外,用户可指定两个或更多个输入声音,并且在指定多个输入声音的情况下,用户可进一步指定每个输入声音的混合比例。

[0181] 然后,客户端1102向服务器1101发送生成具有音高的乐器声音的请求(SEQ 1202)。请求包括由用户指定的输入声音和音高的信息。在用户指定多个输入声音的情况下,请求还包括相应输入声音的混合比例。

[0182] 当从客户端1102接收到请求(SEQ 1203)时,服务器1101首先获取由请求指定的输入声音(SEQ 1204)。服务器1101可从其自身的本地磁盘或经由网络从外部累积装置获取指定的输入声音的wav格式文件。此外,服务器1101可以获取从客户端1102上传的wav格式文件。

[0183] 接下来,服务器1101使用频谱图转换单元301(SEQ 1205)将输入声音的音频波形

转换成梅尔频谱图。在来自客户端1102的请求指定多个输入声音的情况下,服务器1101将所有输入声音的音频波形转换为梅尔频谱图。

[0184] 接下来,服务器1101使用音色特征提取单元302从输入声音的梅尔频谱图提取音色特征量h(SEQ 1206)。在来自客户端1102的请求指定多个输入声音的情况下,以指定混合比例混合从每一输入声音生成的梅尔频谱图,以根据以上表达式(1)计算音色特征量h。

[0185] 接下来,服务器1101使用生成单元303根据从输入声音的梅尔频谱图提取的音色特征量h生成包含由来自客户端1102的请求指定的音高的乐器声音的梅尔频谱图(SEQ 1207)。具体地,使用由本公开模型生成的学习模型(发生器),生成单元303以输入声音的音色特征量和音高作为实例条件根据由乐器声音生成系统300生成的随机数生成具有反映输入声音的特征的音高的乐器声音的梅尔频谱图。

[0186] 接下来,服务器1101使用频谱图波形逆转换单元304从生成的梅尔频谱图重构音频波形(SEQ 1208)。音频波形是具有反映输入声音的特征的音高的乐器声音,并且输出乐器声音具有大约一秒或几秒的长度。它被输出为MIDI数据。

[0187] 然后,服务器1101将生成的具有音高的乐器声音的数据作为请求源返回至客户端1102(SEQ 1209)。应注意,服务器1101可在重构之前返回梅尔频谱图的信息以及重构的音频波形。此外,服务器1101可流传输乐器声音的数据,或者可将文件本身发送到客户端1102。

[0188] 当从服务器1101接收到具有音高的乐器声音的MIDI数据(SEQ 1209)时,客户端1102可根据用户的GUI操作再现(收听用户)并存储乐器声音(SEQ 1210)。此外,用户可通过GUI操作进一步请求生成下一乐器声音。

[0189] G.GUI的配置和操作实施例

[0190] 在该部分G中,将描述GUI屏幕和GUI操作,该GUI屏幕和GUI操作用于请求在客户端1102侧生成具有音高的乐器声音,并指示诸如再现和存储生成的具有音高的乐器声音的处理。在声音再现系统300被安装为客户端服务器模型的情况下,如在部分F中描述的,在客户端的终端上执行在部分G中描述的GUI操作。要注意的是,在声音再现系统300安装在单个信息处理设备上的情况下,使用信息处理设备的控制台执行在部分G中描述的GUI操作。

[0191] G-1.第一实施例

[0192] 图13示出根据本公开的用于生成和再现具有音高的乐器声音的GUI屏幕的配置实例。注意,图13示出了在组合两个输入声音以生成具有音高的乐器声音的情况下使用的GUI屏幕的配置。在本公开中,还可以基于三个或更多个输入声音生成具有音高的乐器声音。稍后将描述在指定三个输入声音以生成具有音高的乐器声音的情况下使用的GUI屏幕的配置和操作。

[0193] 图13所示的GUI屏幕1300包括作为输入/输出域的预设选择单元1301、第一输入声音信息显示单元1302、第二输入声音信息显示单元1303、混合比例指定单元1304、音高信息指定单元1305和生成的乐器声音信息呈现单元1306。

[0194] 预设选择单元1301是经由下拉菜单选择第一输入声音和第二输入声音的声源的GUI部件。下拉菜单包括可被选择作为乐器声音生成系统300(可替换地,服务器1101)(未示出)预先准备的输入声音的预设的列表(wav格式文件的文件名)。在该下拉菜单上由用户选择的wav格式文件被顺序地指定为第一输入声音和第二输入声音。

[0195] 第一输入声音信息显示单元1302和第二输入声音信息显示单元1303显示通过预设选择单元1301指定的第一输入声音和第二输入声音的wav格式文件的文件名。此外,第一输入声音信息显示单元1302和第二输入声音信息显示单元1303中的每个具有用于改变输入声音的下拉菜单。第一输入声音信息显示单元1302和第二输入声音信息显示单元1303的下拉菜单中的每一个不是由乐器声音生成系统300(可替换地,服务器1101)预先准备的预设菜单,而是在客户端(可替代地,信息处理设备)中可被独立地选择为输入声音的wav格式文件的文件名的列表。用户还可使用第一输入声音信息显示单元1302和第二输入声音信息显示单元1303的相应的下拉菜单来选择第一输入声音和第二输入声音。

[0196] 第一输入声音信息显示单元1302和第二输入声音信息显示单元1303分别包括用于指示被指定作为输入声音的wav格式文件的再现的播放按钮1302-1和1303-1。利用这些播放按钮1302-1和1303-1,用户可以在请求生成乐器声音之前再现和收听被指定为输入声音的每个wav格式文件以确认声源是否是用户所期望的声源。

[0197] 混合比例指定单元1304是供用户指定在第一输入声音信息显示单元1302和第二输入声音信息显示单元1303中设置的两个输入声音的混合比例的输入域。在图13中所示的实施例中,混合比例指定单元1304包括用于选择性地指定第二输入声音与第一输入声音的混合比例0.0、0.1、0.2、...、0.8、0.9以及1.0的单选按钮。较接近于0.0的混合比例可请求生成更多俘获第一输入声音的特征的乐器声音,且较接近于1.0的混合比例可请求生成更多俘获第二输入声音的特征的乐器声音。混合比例0.0意味着仅指定第一输入声音的单个声音,混合比例1.0意味着仅指定第二输入声音的单个声音,并且可从单个声音的特征请求乐器声音生成。

[0198] 在图13所示的GUI屏幕配置实例中,球场信息指定单元1305被布置在屏幕的底部。音高信息指定单元1305包括使用钢琴键的布局的设计(在下文中,简称为“键盘”)1305-1。用户可通过点击或触摸键盘1305-1中的键来指定要生成的乐器声音的音高。由于指示要生成的音高的文本1305-2(在图13中所示的实施例中,文本“生成目标MIDI音符/音高:60”被显示)被显示在键盘1305-1的顶部附近,用户可以视觉地确认该文本。

[0199] 一对正和负按钮1305-3设置在键盘1305-1的左下端附近。用户可通过选择“+”按钮和“-”按钮来指示倍频程的向上或向下。因此,能够指定从A-1到C7的88个音高的音高。

[0200] 此外,用于在“低声音质量”的接通和断开两种状态之间切换的拨动开关1305-4被布置在键盘1305-1的下部中心部分中。当拨动开关1305-4被用于拨动至“低声音质量”的接通状态时,以低声音质量生成反映输入声音的特征的乐器声音。另一方面,当拨动开关1305-4被用于拨动至“低声音质量”的断开状态时,生成具有高声音质量的反映输入声音的特性的乐器声音。

[0201] 此外,投球信息指定单元1305包括基本上在键盘1305-1上方的中心处的“更新”按钮1305-5。在用户希望生成具有另一音高的相同特征的乐器声音的情况下,在通过点击或触摸从键盘1305-1指定对应于另一期望音高的音高之后,用户可通过选择更新按钮1305-5来指示再生具有另一音高的相同乐器声音。

[0202] 当用户在音高信息指定单元1305上选择键盘1305-1上的键之一时,用于生成具有音高的乐器声音的请求被输出到服务器1101(可选地,在信息处理设备中激活生成具有音高的乐器声音的处理)。

[0203] 在服务器1101侧(可替代地,在信息处理设备中),第一输入声音和第二输入声音的各个输入声音的音频波形被转换成梅尔频谱图,从各个梅尔频谱图中提取音色特征量,并以指定的混合比例进行混合,然后使用音色特征量和音高信息作为实例条件,生成具有指定声音质量(高声音质量或低声音质量)的音调的乐器声音。另一方面,在客户端1102侧,在服务器1101侧生成的具有音高的乐器声音被流传输和再现(可替换地,其被下载和再现和输出)(然而,在信息处理设备内部生成具有音高的乐器声音的情况下,信息处理设备再现并输出所生成的声音)。

[0204] 生成的乐器声音信息呈现单元1306包括呈现域1306-1,呈现域1306-1以音高呈现关于所生成的乐器声音的信息。要呈现的“关于具有音高的乐器声音的信息”没有特别限制。例如,可在呈现域1306-1(稍后描述)中显示在视觉上表达乐器声音的音频波形的特性的信息,例如梅尔频谱图(或者,频率频谱图)。当然,代替频谱图,乐器声音的音频波形可显示在呈现域1306-1中。

[0205] 另外,所生成的乐器声音信息呈现单元1306包括播放按钮1306-2。用户可通过使用播放按钮1306-2以音高再现和收听生成的乐器声音,并检查音高和乐器声音是否如期望的那样反映指定的输入声音的特征。

[0206] 在下文中,将参考图14至图23描述在图13中示出的GUI屏幕上的操作实例。

[0207] 图14示出从在预设选择单元1301的下拉菜单1401上显示的wav格式文件的文件名的列表中依次指定用作第一输入声音和第二输入声音的文件名的状态。在下拉菜单1401中指定的文件被分别显示在第一输入声音信息显示单元1302和第二输入声音信息显示单元1303上。在图14中所示的实例中,在下拉菜单1401上指定音频文件“Input_audio#001.wav”和“Input_audio#002.wav”。

[0208] 图15示出了响应于下拉菜单1401上的音频文件“input_audio#001.wav”和“input_audio#002.wav”的指定而在第一输入声音信息显示单元1302和第二输入声音信息显示单元1303上显示相应文件名的状态。用户可从显示在第一输入声音信息显示单元1302和第二输入声音信息显示单元1303上的文件名视觉上确认用于生成具有音高的乐器声音的输入声音的组合。此外,用户可通过使用第一输入声音信息显示单元1302的播放按钮1302-1和第二输入声音信息显示单元1303的播放按钮1303-1来单独地再现被指定为输入声音的每个wav格式文件“input_audio#001.wav”和“input_audio#002.wav”,并通过收听确认用于生成具有音高的乐器声音的输入声音的组合。

[0209] 图16示出混合比例指定单元1304的单选按钮用于指定分别被指定为第一输入声音和第二输入声音的“input_audio#001.wav”和“input_audio#002.wav”的音频波形的混合比例的状态。混合比例指定单元1304包括可选择地指定第二输入声音相对于第一输入声音的混合比例0.0、0.1、0.2、...、0.8、0.9以及1.0的单选按钮。较接近于0.0的混合比例可请求生成俘获第一输入声音的特征的乐器声音,且较接近于1.0的混合比例可请求生成俘获第二输入声音的特征的乐器声音(上文所描述)。在图16所示的实例中,指定了混合比例0.3。

[0210] 图16还示出使用音高信息指定单元1305指定要生成的乐器声音的音高的状态。音高信息指定单元1305具有使用钢琴的键盘的布局的设计,并且在键盘1305-1的每个键上显示表示相应音高的字符。此外,可使用设置在键盘1305-1的左下端附近的正/负按钮1305-3

来指示倍频程的向上和向下。因此,用户能够通过组合键盘1305-1和劈起信息指定单元1305(如上所述)的正/负按钮1305-3的操作来指定从A-1至C7的88个劈起。应注意,用于接通/断开“低声音质量”的拨动开关1305-4被切换至“断开”。

[0211] 当用户在音高信息指定单元1305上选择键盘1305-1的任何键(在图16中所示的实施例中,设置键“A”)时,用于生成具有音高的乐器声音的请求被输出到服务器1101(可替代地,在信息处理设备中激活生成具有音高的乐器声音的过程)。然后,具有反映输入声音的特征的音高的乐器声音再现在服务器1101侧生成的大约1秒或几秒的相对短时间(或者在信息处理设备内部生成)。图17说明其中根据乐器声音的再现在所生成的乐器声音信息呈现单元1306的呈现域1306-1中显示乐器声音的梅尔频谱图的状态。呈现域1306-1反映对应于由混合比例指定单元1304指定的混合比例生成的梅尔频谱图。当在混合比例指定单元1304中选择另一单选按钮时,在混合比例指定单元1304(未示出)中切换单选按钮的开/关显示,在与新选择的混合比例对应的乐器声音中反映呈现域1306-1中的梅尔频谱图,并且用户可以收听并确认乐器声音。

[0212] 在用户想要搜索另一乐器声音的样本的情况下,用户可通过预设选择单元1301再次选择预设,或通过第一输入声音信息显示单元1302和第二输入声音信息显示单元1303单独地改变第一输入声音和第二输入声音。图18示出了第一输入声音和第二输入声音分别从第一输入声音信息显示单元1302和第二输入声音信息显示单元1303的下拉菜单改变的状态。在图18中所示的实例中,通过在第一输入声音信息显示单元1302的下拉菜单1302-2上的选择操作,将第一输入声音改变为音频文件“input_audio#101.wav”。此外,第二输入声音通过在第二输入声音信息显示单元1303的下拉菜单1303-2上的选择操作而变成“input_audio#203.wav”。与混合比例的指定、音高信息的指定以及在单独改变每个输入声音之后以音高再现所生成的乐器声音相关的操作与上述描述类似,因此这里省略其描述。

[0213] 此外,图19示出在用户保持输入声音的组合的同时生成具有另一音高的乐器声音的情况下的操作实施例。用户通过操作音高信息指定单元1305的键盘1305-1和加/减按钮1305-3,再次指定希望以相同的输入声音的组合新生成的乐器声音的音高,然后选择在键盘1305-1上方大致中央的更新按钮1305-1,以指示以另一音高再生相同的乐器声音。

[0214] 应注意,图18示出了通过第一输入声音信息显示单元1302和第二输入声音信息显示单元1303的下拉菜单指定或改变输入声音的操作实例。下拉菜单上显示预设的wav格式文件的文件名列表。另一方面,尽管不是预设的,但是用户的手处的声源可被选择为输入声音。图20示出了在用户的手处的声源被指定为第一输入声音的情况下的操作实施例。这里提及的“用户手上的声源”是可以经由网络从客户端1102的本地盘(可替代地,信息处理设备)或者从外部累积装置获取的wav格式文件。由于要生成音高的乐器声音是约1秒或几秒的相对短的时间,所以声源还优选地是约1秒或几秒的相对短的时间的音频数据。在图20中所说明的实施例中,用于选择用户的手处的声音源的输入框1302-3出现在第一输入声音信息显示单元1302中。输入框1302-3在左半部显示用户手边的声源(wav格式文件)的列表,并在右半部显示当前选择的声源(高亮文件)的属性信息。用户可使用输入框1302-3的左半部搜索手边的声源,并基于属性信息和显示在右半部的再现声音来检查输入声音是否是期望的输入声音。然后,当在输入框1302-3上确认对任何wav格式文件的选择时,输入框1302-3消失,并且其选择被确认的wav格式文件的文件名显示在第一输入声音信息显示单元1302

(未示出)上。

[0215] 在用户喜欢具有在服务器1101侧生成的音高的乐器声音的情况下,用户可以将wav格式文件作为声源下载到客户端1102(例如,用户自己的信息终端)。图21示出下载具有在服务器1101侧生成的音高的乐器声音时的操作实例。在生成的乐器声音信息呈现单元1306的呈现域1306-1中,显示待下载的乐器声音的梅尔频谱图。用户可通过选择布置在呈现域1306-1的右下端附近的下载按钮1306-4来指示下载具有喜爱的音调的乐器声音。

[0216] 此外,“12音高的下载文件”按钮1305-5基本上布置在音高信息指定单元1305的键盘1305-1下方的中心处。“12音高的下载文件”按钮1305-5是用于指示不仅下载由键盘1305-1中的键指定的一个特定音高而且下载用于12音高的乐器声音的按钮。当选择“12音高的下载文件”按钮1305-5时,服务器1101侧生成具有12音高的乐器声音,并将该乐器声音下载到请求客户端1102。然而,生成12音高的乐器声音需要花费处理时间。

[0217] 图22示出了选择基本上布置在生成的乐器声音信息呈现单元1306上方的中心处的“内插全部”按钮1306-3的状态。“内插全部”按钮1306-3是指示组合第一输入声音和第二输入声音以生成具有从混合比例0.1到0.9(或从0.0到1.0,包括单个声音)的顺序的音调的乐器声音的按钮。当选择“全部插入”按钮1306-3时,在服务器1101侧按照每个混合比例依次自动生成具有音高的乐器声音,并且在客户端1102侧依次再现自动生成的乐器声音(可替代地,在信息处理设备中激活具有每个混合比例的音高的乐器声音的自动生成处理,并且再现并输出依次生成的乐器声音)。

[0218] 图23示出了在执行包络处理的情况下的操作实例。包络线是对所生成的乐器声音给出随着时间的典型改变的过程。当选择大体上安置于所生成的乐器声音信息呈现单元1306下方的中心处的包络按钮1306-5时,呈现域1306-1从梅尔频谱图的显示切换到用于调整包络的每一参数的滑块的显示。该包络参数包括ADSR(攻击、衰减、持续、释放)。攻击时间越短,响应越好,攻击时间越长,上升越柔软。衰减时间越长,衰减发生的时间越长,并且衰减时间越短,衰减发生的时间越短。持续水平是用于控制音量而不是时间的参数,并且通过继续开启笔记来控制最终达到的音量。释放时间越长,谐振发生的时间越长,并且释放时间越短,声音变得更清晰。注意,拨动开关1305-4如上所述。当拨动开关1305-4被用于拨动至“低声音质量”的断开状态(换言之,高声音质量的状态)时,反映输入声音的特征的乐器声音以高品质生成并且由包络线等定制。

[0219] G-2.第二实施例

[0220] 图24示出根据本公开的用于生成并再现具有音高的乐器声音的GUI屏幕的另一配置实施例。然而,虽然图13示出在组合两个输入声音以生成乐器声音的情况下使用的GUI屏幕,但是图24示出在组合三个输入声音以生成具有音高的乐器声音的情况下使用的GUI屏幕。

[0221] 图24所示的GUI屏幕2400包括声源指定域2410、乐器声音生成操作域2420和预处理操作域2430。

[0222] 声源指定域2410是用于选择第一至第三输入声音的每个声源的操作区域。可使用下拉菜单从预设置中选择用作每个输入声音的声源的wav格式文件,或者可使用输入框选择用户的手处的声源(例如,见图20)。使用下拉菜单和输入框的声源选择操作如上所述,这里将省略其详细描述。在图24所示的实施例中,通过用户的操作选择“第一输入声音.wav”、

“第二输入声音.wav”和“第三输入声音.wav”的三个wav格式文件,并且在声源指定域2410中显示这些文件中的每一个的文件名和创建日期和时间。

[0223] 乐器声音生成操作域2420是用于在合成在声源指定域2410中选择的第一至第三输入声音时进行设置的操作区域。乐器声音生成操作域2420具有三角形背景。第一输入声音至第三输入声音被分配至三角形的相应顶点2421至2423,并且相应输入声音的采样波形显示在相应顶点2421至2423附近。

[0224] 预处理操作域2430是用于对每个输入声音的采样波形执行预处理的操作区域。当在乐器声音生成操作域2420中选择第一输入声音至第三输入声音的样本波形中的任一个时,均衡器(EQ)的操作屏幕和指定用于所选择的输入声音的声源的包络处理(ADSR)显示在预处理操作域2430中,并且可对声源执行EQ和ADSR的预处理。

[0225] 将再次描述乐器声音生成操作域2420。当选择背景三角形中的任何位置2424时,基于选择的位置2424与每个顶点2421至2423之间的距离的比率来设置第一输入声音至第三输入声音的混合比例。例如,当选择更靠近三角形中分配第一输入声音的顶点2421的位置时,第一输入声音的混合比例变得更高,并可请求生成更多捕捉第一输入声音的特征的乐器声音。此外,当选择三角形的任何顶点位置时,这意味着仅指定分配给第一至第三顶点中的顶点的输入声音的单个声音,并且可从单个声音的特性请求乐器声音生成。

[0226] 半透明的圆首先显示在三角形中选择的位置处。此后,当选择乐器声音生成操作域2420上方的“生成”按钮2425时,将乐器声音生成请求输出到服务器1101(可选地,在信息处理设备中激活生成乐器声音的处理)。此外,通过选择“生成”按钮2425,确定在○位置中指定的第一输入声音至第三输入声音的混合比例,并且○的显示从半透明变成不透明(未示出)。

[0227] 在图24所示的GUI屏幕2400中,假设乐器声音生成请求请求生成88个键的乐器声音。当然,所述请求可以是用于生成具有89个或更多键或87个或更少键的乐器声音的请求。可选地,用于指定音高信息的GUI组件(例如,键盘)可被布置在乐器声音生成操作域2420中或另一域中,并且可做出用于生成具有通过GUI组件指定的仅一个键的音高的乐器声音的请求。

[0228] 在服务器1101侧(备选地,在信息处理设备中),将第一至第三输入声音的每个输入声音的音频波形转换为梅尔频谱图,从每个梅尔频谱图提取音色特征量并以指定的混合比例混合,然后使用音色特征量和88个键的每个音高信息作为实例条件,生成具有指定的声音质量(高声音质量或低声音质量)的音调的乐器声音。另一方面,在客户端1102侧,与在服务器1101侧生成的88个键对应的乐器声音被流传输和再现(可替换地,其被下载和再现和输出)(然而,在信息处理设备内部生成具有音高的乐器声音的情况下,信息处理设备再现并输出所生成的声音)。

[0229] 注意,当在预处理操作域2430中改变任何输入声音的参数时,突出显示“生成”按钮2425,并且在视觉上警告用户除非选择“生成”按钮2425以请求生成乐器声音,否则参数的改变将不会反映在乐器声音中。

[0230] 用户收听在服务器1101侧生成的88个键的乐器声音,如果用户喜欢乐器声音,则用户可选择喜爱按钮2426并将乐器声音登记在喜爱列表中。登记到喜爱列表可以包括将访问方法(例如,用于区分存储在服务器1101侧的wav格式文件的位置的统一资源标识符

(URI)等)记录到相应乐器声音的声源的处理和从服务器1101下载乐器声音的wav格式文件的处理。

[0231] 图25示出图24所示的GUI屏幕的变形例。图25中示出的GUI屏幕2500与图24中示出的GUI屏幕的不同之处在于乐器声音生成操作域2420在背景中具有沿水平方向布置的两个三角形2510和2520。

[0232] 左侧的三角形2510用于显示第一至第三输入声音的采样波形,并设置混合比例,类似于图24所示的GUI屏幕2400的乐器声音生成操作域2420中的三角形。这里,省略对左三角形2510的详细描述。

[0233] 另一方面,当使用由本公开模型生成的发生器以实例条件生成乐器声音时,使用直角三角形2520来微调输入到发生器的随机数z。同样在直角三角形2520中,第一输入声音至第三输入声音被分配至该三角形2520的顶点。当在三角形中选择任何位置2521时,基于所选择的位置2521与每个顶点之间的距离的比率来微调输入到发生器的随机数z。例如,当选择更接近三角形中分配第一输入声音的顶点的位置时,精细地调整随机数z,以便包括第一输入声音的更多特征。

[0234] H.信息处理设备的配置

[0235] 在该部分H中,将描述提供用于实现本公开的信息处理设备的具体配置。

[0236] 图26示出信息处理设备2000的具体硬件配置实施例。图26中所示的信息处理设备2000包括例如PC等。信息处理设备2000可以操作为图3中示出的乐器声音生成系统300或者可以操作为图11中示出的服务器1101或客户端1102。

[0237] 图26中所示的信息处理设备2000包括CPU 2001、只读存储器(ROM)2002、随机存取存储器(RAM)2003、主机总线2004、桥接器2005、扩展总线2006、接口单元2007、输入单元2008、输出单元2009、存储单元2010、驱动器2011以及通信单元2013。

[0238] CPU 2001用作运算处理装置和控制装置,并且根据各种程序控制信息处理设备2000的整体操作。ROM 2002以非易失性方式存储由CPU 2001使用的程序(基本输入/输出系统等)和计算参数。RAM 2003用于加载CPU 2001的执行中使用的程序,并且临时存储在程序执行中适当改变的诸如工作数据等参数。加载到RAM 2003中并由CPU 2001执行的程序是例如各种应用程序、操作系统(OS)等。

[0239] CPU 2001、ROM 2002和RAM 2003通过包括CPU总线等的主机总线2004相互连接。然后,CPU 2001可以通过ROM 2002和RAM 2003的协作操作,在OS提供的执行环境下通过执行各种应用程序来实现各种功能和服务。在信息处理设备2000是PC的情况下,OS例如是微软公司的Windows或Unix。另外,应用程序包含执行处理作为波形频谱图转换单元301、音色特征提取单元302、生成单元303和频谱图波形逆转换单元304中的每一者的应用程序、执行在音色特征提取单元302和生成单元303中的每一者中使用的机器学习模型(DNN等)的学习处理的应用程序,以及通过GUI屏幕处理用户操作的应用程序,如图13至25中所说明。

[0240] 主机总线2004经由桥接器2005连接至扩展总线2006。例如,扩展总线2006是外围组件互连(PCI)总线或PCI Express,并且桥接器2005基于PCI标准。然而,信息处理设备2000不必具有电路部件通过主机总线2004、桥接器2005和扩展总线2006分离的配置,并且可以采用其中几乎所有的电路部件通过单个总线(未示出)互连的实现方式。

[0241] 接口单元2007根据扩展总线2006的标准连接诸如输入单元2008、输出单元2009、

存储单元2010、驱动器2011和通信单元2013之类的外围装置。然而,并非图26中示出的所有外围装置都是必要的,并且信息处理设备2000可进一步包括外围装置(未示出)。此外,外围装置可内置在信息处理设备2000的主体中,或者一些外围装置可外部连接到信息处理设备2000的主体。

[0242] 输入单元2008包括输入控制电路等,该输入控制电路基于来自用户的输入来生成输入信号并且将该输入信号输出到CPU 2001。输入单元2008可以包括诸如键盘、鼠标、触摸面板或麦克风的输入装置。输出单元2009包括例如显示装置,诸如液晶显示器(LCD)装置、有机电致发光(EL)显示装置和发光二极管(LED)。使用输入单元2008和输出单元2009的装置中的至少一部分执行GUI操作,以指定输入声音和音高并指示生成具有音高的乐器声音,或指示所生成的具有音高的乐器声音的再现、下载等。

[0243] 存储单元2010包括例如诸如固态驱动器(SSD)或硬盘驱动器(HDD)的大容量存储装置,但是可包括外部存储装置。存储单元2010存储诸如CPU 2001执行的程序(应用、OS等)的文件和各种数据。另外,存储单元2010存储作为乐器音的声源的音频波形的wav格式文件,并存储所生成的乐器音的MIDI数据。

[0244] 可移动记录介质2012是诸如微SD卡的盒式存储介质。驱动器2011在加载的可移动记录介质2012上执行读取和写入操作。驱动器2011将从可移动记录介质2012读取的数据输出到RAM 2003和存储单元2010,并且将RAM 2003和存储单元2010上的数据写入到可移动记录介质2012。可移动记录介质2012用于读取作为乐器声音的声源的音频波形的wav格式文件,并用于存储所生成的乐器声音的MIDI数据。

[0245] 通信单元2013是执行诸如Wi-Fi(注册商标)、蓝牙(注册商标)的无线通信或诸如4G或5G的蜂窝通信网络的装置。在信息处理设备2000作为服务器1101操作的情况下,经由通信单元2013执行客户端1102之间的相互通信。此外,通信单元2013可以包括诸如通用串行总线(USB)或高清晰度多媒体接口(HDMI(注册商标))的端子,并且可以进一步包括执行与诸如扫描仪或打印机、显示器等的USB装置的数据通信的功能。

[0246] 本说明书中描述的一系列处理可以通过硬件、软件或硬件和软件相结合的配置执行。在由软件执行处理的情况下,利用与本公开的实现相关的处理序列记录的程序被安装并且执行于结合在计算机中的专用硬件中的存储器中。也可以在能够执行各种类型的处理的通用计算机中安装程序,并且使计算机执行与本公开的实施方式相关的处理。

[0247] 程序可以预先存储在设置在计算机中的记录介质中,诸如HDD、SSD或ROM作为记录介质。可替换地,程序可以临时或永久地存储在诸如软盘、致密盘只读存储器(CD-ROM)、磁光(MO)盘、数字通用盘(DVD)、蓝光盘(BD)(注册商标)、磁盘、通用串行总线(USB)存储器等的可移动记录介质中。通过使用这种可移动记录介质,可以提供与本公开的实现方式有关的程序作为所谓的包软件。

[0248] 此外,程序可以经由诸如广域网的网络从下载站点传输到计算机。局域网(WAN),由蜂窝网络、局域网(LAN)或互联网以无线或有线方式表示。在计算机中,这样传输的程序可以被接收并安装在诸如计算机中的HDD或SSD的大容量存储装置中。

[0249] I.与相关研究的比较

[0250] 本节I描述了本公开与对乐器声音的生成的其他研究的比较。

[0251] NSynth(参见非专利文献2)使用基于Wavenet(参见非专利文献3)的自动编码器生

成乐器声音的波形,但是由于自回归采样存在生成缓慢的问题,并且在生成的声音中可能出现伪影。另一方面,本发明可生成反映交互时间内的输入声音的乐器声音。

[0252] GANSynth(参见非专利文献4)可通过使用图像生成模型生成包含相位信息的频谱图来改进生成速度和声音质量,但是由于GANSynth是无条件的生成模型且不接受输入,因此难以在复杂潜在空间中搜索所要音色。另一方面,因为本发明是具有实例条件的生成模型,所以可接收输入声音且在复杂潜在空间中搜索反映输入声音的音色。

[0253] 工业实用性

[0254] 在此之前参考具体实施方式详细描述了本公开。然而,本公开不应被解释为限于上述实施方式,并且显而易见的是,在不背离本公开的要旨的情况下,本领域技术人员可对实施方式进行修改和替换。此外,在本说明书中描述的效果仅是实施例,并且本公开带来的效果不受限制,并且可以存在未在本说明书中描述的另外的效果。

[0255] 本公开可应用于例如个人计算机、电子乐器等,其执行与诸如作品或音乐编辑的音乐生成有关的处理,并可从任意声音启示生成唯一的乐器声音,以自由定制乐器或通过声音向音乐分配含义。

[0256] 简而言之,迄今为止,本公开是以实施例的形式描述的,并且本说明书中描述的内容不应以限制的方式解释。为了确定本公开的要点,应当考虑权利要求。

[0257] 应注意,本公开可具有以下配置。

[0258] (1)一种信息处理系统,包括:

[0259] 电路系统,被配置为

[0260] 接收输入声音和音高信息;

[0261] 从输入声音提取音色特征量;并且

[0262] 基于音色特征量和音高信息来生成具有音高的乐器声音的信息。

[0263] (2)根据(1)的信息处理系统,其中,

[0264] 电路系统被配置为使用学习模型来生成乐器声音的信息。

[0265] (3)根据(1)至(2)中任一项的信息处理系统,其中,

[0266] 电路系统被配置为以通过对输入声音进行预处理而生成的信息和音高信息作为实例条件使用学习模型来生成乐器声音的信息。

[0267] (4)根据(1)至(3)中任一项的信息处理系统,其中,

[0268] 电路系统被配置为提取音色特征量以使得不保留音高信息。

[0269] (5)根据(1)至(4)中任一项的信息处理系统,其中,

[0270] 电路系统被配置为使用音色特征提取器来提取音色特征,音色提取器已执行关于音高的对手学习。

[0271] (6)根据(1)至(5)中任一项的信息处理系统,其中,电路系统被配置为:

[0272] 将输入声音转换成梅尔频谱图;并且

[0273] 基于来自输入声音的梅尔频谱图提取输入声音的音色特征量。

[0274] (7)根据(6)的信息处理系统,其中,电路系统被配置为:

[0275] 使用音色特征量和音高信息生成具有音高的乐器声音的梅尔频谱图,并且

[0276] 基于梅尔频谱图构造音频波形。

[0277] (8)根据(7)的信息处理系统,其中,电路系统被配置为:

[0278] 将梅尔频谱图转换成线性频谱图中的频率刻度;

[0279] 恢复线性频谱图的相位;并且

[0280] 在恢复线性频谱图的相位之后,对线性频谱图执行傅立叶逆转换。

[0281] (9)根据(8)的信息处理设备,其中,电路系统被配置为:

[0282] 将校正为非负值的解设置为无非负值的最小二乘法的解作为迭代计算的初始值;并且

[0283] 根据依照梯度法重复更新并校正为非负值的迭代法执行频率刻度转换。

[0284] (10)根据(1)至(9)中任一项的信息处理系统,其中,电路系统被配置为:

[0285] 接收多个输入声音的输入;

[0286] 从每个输入声音提取音色特征量;并且

[0287] 基于通过混合多个输入声音的音色特征量而获得的音色特征量和音高信息来生成乐器声音信息。

[0288] (11)根据(10)的信息处理系统,其中,电路系统被配置为:

[0289] 接收关于多个输入声音的混合比例的信息;以及

[0290] 基于通过混合多个音色特征量获得的音色特征量、混合比例和音高信息来生成乐器声音信息。

[0291] (12)根据(1)至(11)中任一项的信息处理系统,其中,

[0292] 电路系统被配置为基于用户操作来接收输入声音和音高信息。

[0293] (13)根据(1)至(12)中任一项的信息处理系统,其中,

[0294] 电路系统被配置为输出乐器声音的信息。

[0295] (14)根据(1)至(13)中任一项的信息处理系统,其中,

[0296] 电路系统被配置为显示用户界面,用户界面被配置为接收与输入声音和音高信息相对应的用户输入。

[0297] (15)根据(14)的信息处理系统,其中,

[0298] 用户界面被配置为接收与第一输入声音相对应的第一输入和与第二输入声音相对应的第二输入,并且

[0299] 用户界面被配置为接收与第一输入声音和第二输入声音相对应的混合比例。

[0300] (16)根据(15)的信息处理系统,其中,

[0301] 图形用户界面包括至少第一图形和第二图形,其中,

[0302] 第一图形被配置为接收与第一输入声音相对应的第一输入和与第二输入声音相对应的第二输入,并且

[0303] 第二图形被配置为接收音色特征量。

[0304] (17)一种信息处理方法,包括:

[0305] 接收输入声音和音高信息;

[0306] 从输入声音提取音色特征量;并且

[0307] 基于音色特征量和音高信息生成具有音高的乐器声音的信息。

[0308] (18)一种或多种非暂时性计算机可读介质,当由电路系统执行时,使电路系统:

[0309] 接收输入声音和音高信息;

[0310] 从输入声音提取音色特征量;并且

[0311] 基于音色特征量和音高信息生成具有音高的乐器声音的信息。

[0312] (19)一种声音生成系统,包括:

[0313] 终端,被配置为请求乐器声音的生成;以及

[0314] 信息处理设备,生成乐器声音,其中,

[0315] 信息终端被配置为:

[0316] 接收输入声音和音高信息;

[0317] 从输入声音提取音色特征量;并且

[0318] 基于音色特征量和音高信息来生成具有音高的乐器声音的信息。

[0319] (20)一种信息终端,包括:

[0320] 通信接口,被配置为与信息处理系统通信;以及

[0321] 用户界面,被配置为接收与包括生成输入声音和音高信息的乐器声音相关的指示,其中,

[0322] 通信接口被配置为:

[0323] 将用于生成包括输入声音和音高信息的乐器声音的请求发送至信息处理系统,并且

[0324] 从信息处理系统接收乐器声音的信息。

[0325] 参考标号列表

[0326] 300 声音生成系统

[0327] 301 波形频谱图转换单元

[0328] 302 音色特征提取单元

[0329] 303 生成单元

[0330] 304 频谱图波形逆转换单元

[0331] 501 音色特征提取器

[0332] 502 乐器鉴别器

[0333] 601 音色特征提取器

[0334] 602 乐器鉴别器

[0335] 603 音高鉴别器

[0336] 801 频率刻度转换单元

[0337] 802 相位恢复单元

[0338] 803逆短时傅立叶转换单元(iSTFT)

[0339] 901 初始化单元

[0340] 902 更新单元

[0341] 903 校正单元

[0342] 1001 初始化单元

[0343] 1002 初始值校正单元

[0344] 1003 更新单元

[0345] 1004 校正单元

[0346] 1100声音生成系统(客户端服务器模型)

[0347] 1101 服务器

[0348] 1102 客户端

[0349] 2000 信息处理设备

[0350] 2001CPU

[0351] 2002ROM

[0352] 2003RAM

[0353] 2004 主机总线

[0354] 2005 桥接器

[0355] 2006 扩展总线

[0356] 2007 接口单元

[0357] 2008 输入单元

[0358] 2009 输出单元

[0359] 2010 存储单元

[0360] 2011 驱动器

[0361] 2012 可移动记录介质

[0362] 2013 通信单元。

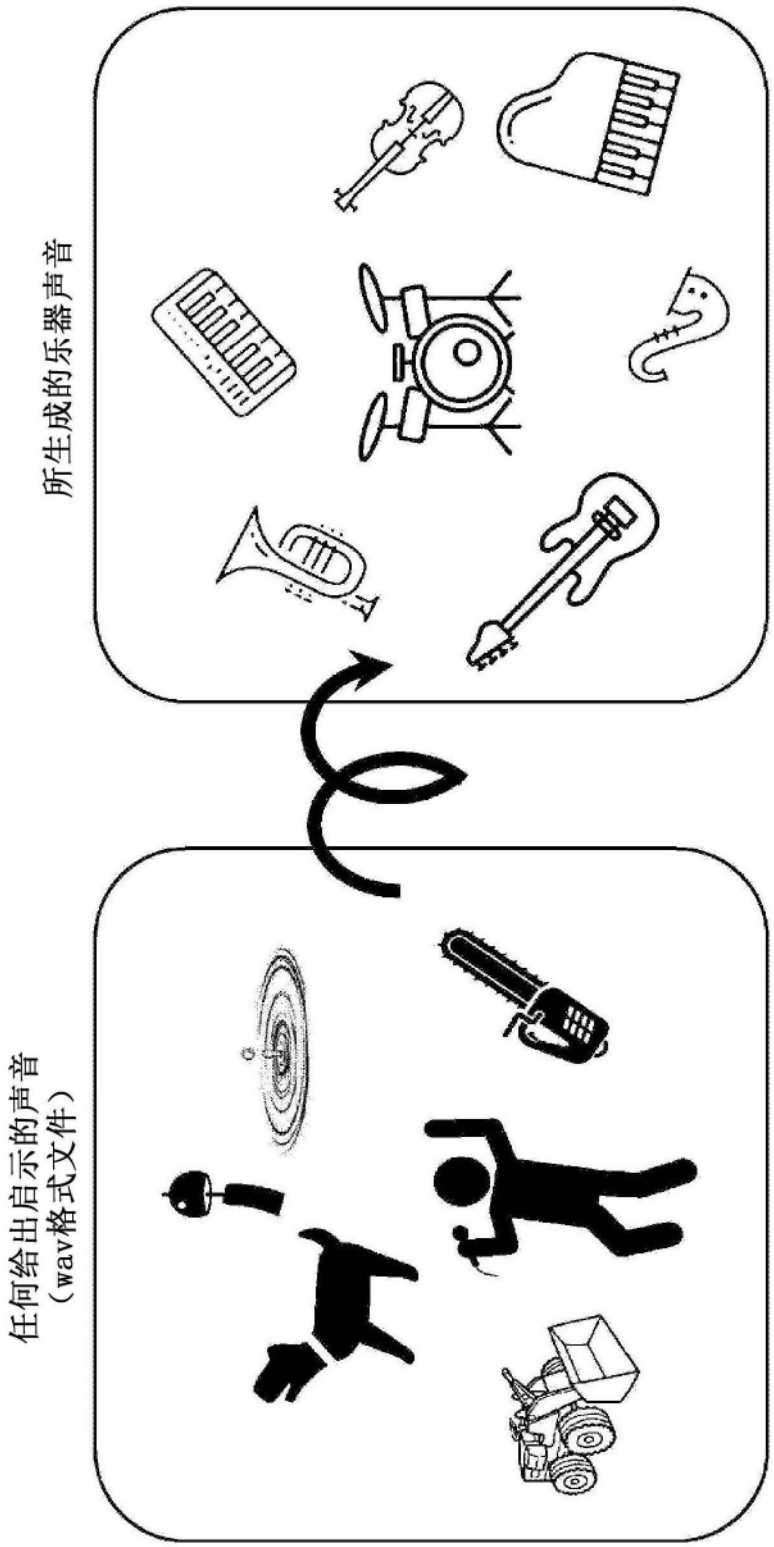


图1

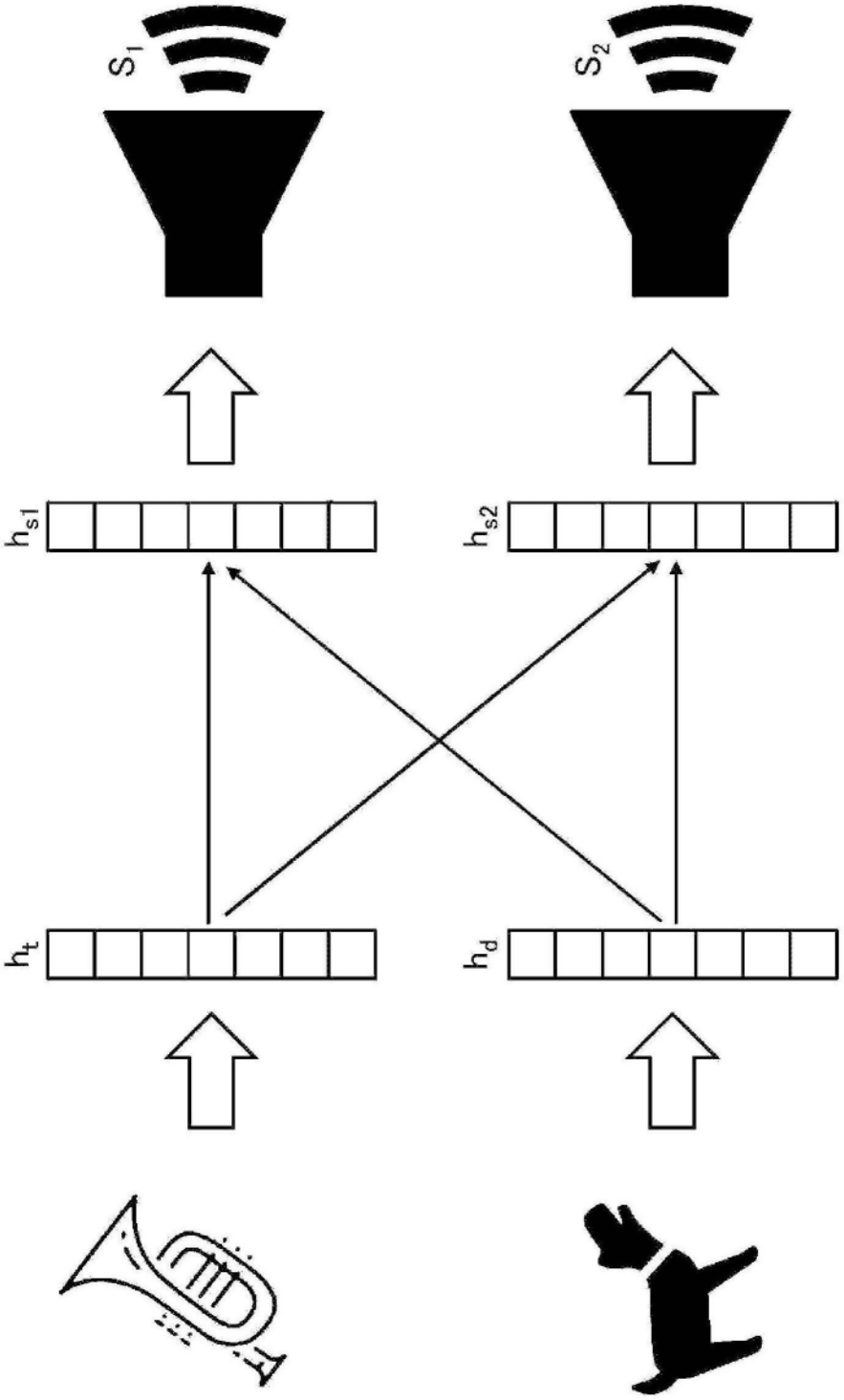


图2

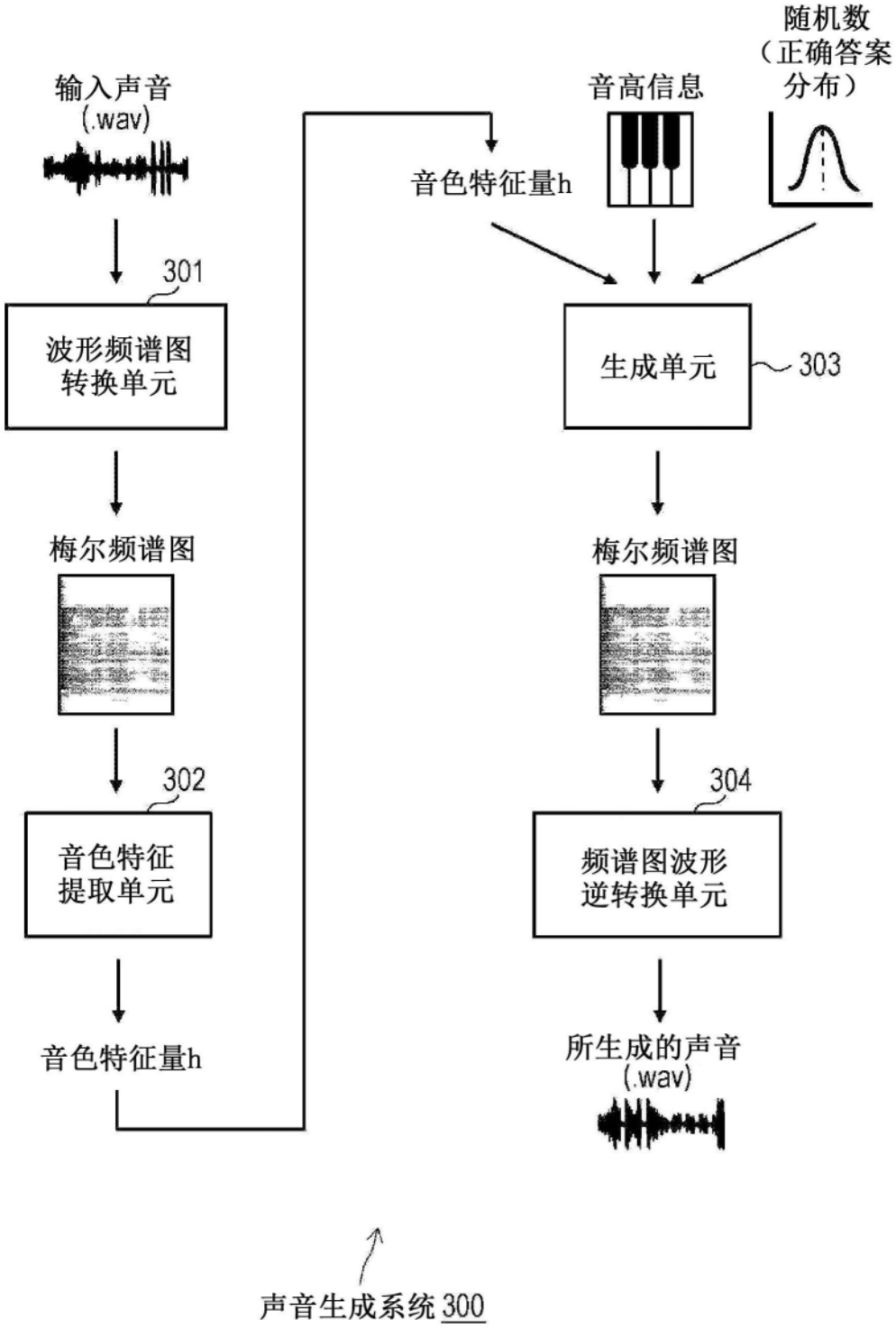

输入声音
(.wav)
301
波形频谱图
转换单元
梅尔频谱图
302
音色特征
提取单元
音色特征量h
音色特征量h
音高信息
随机数
（正确答案
分布）
生成单元
303
梅尔频谱图
304
频谱图波形
逆转换单元
所生成的声音
(.wav)
声音生成系统 300

图3

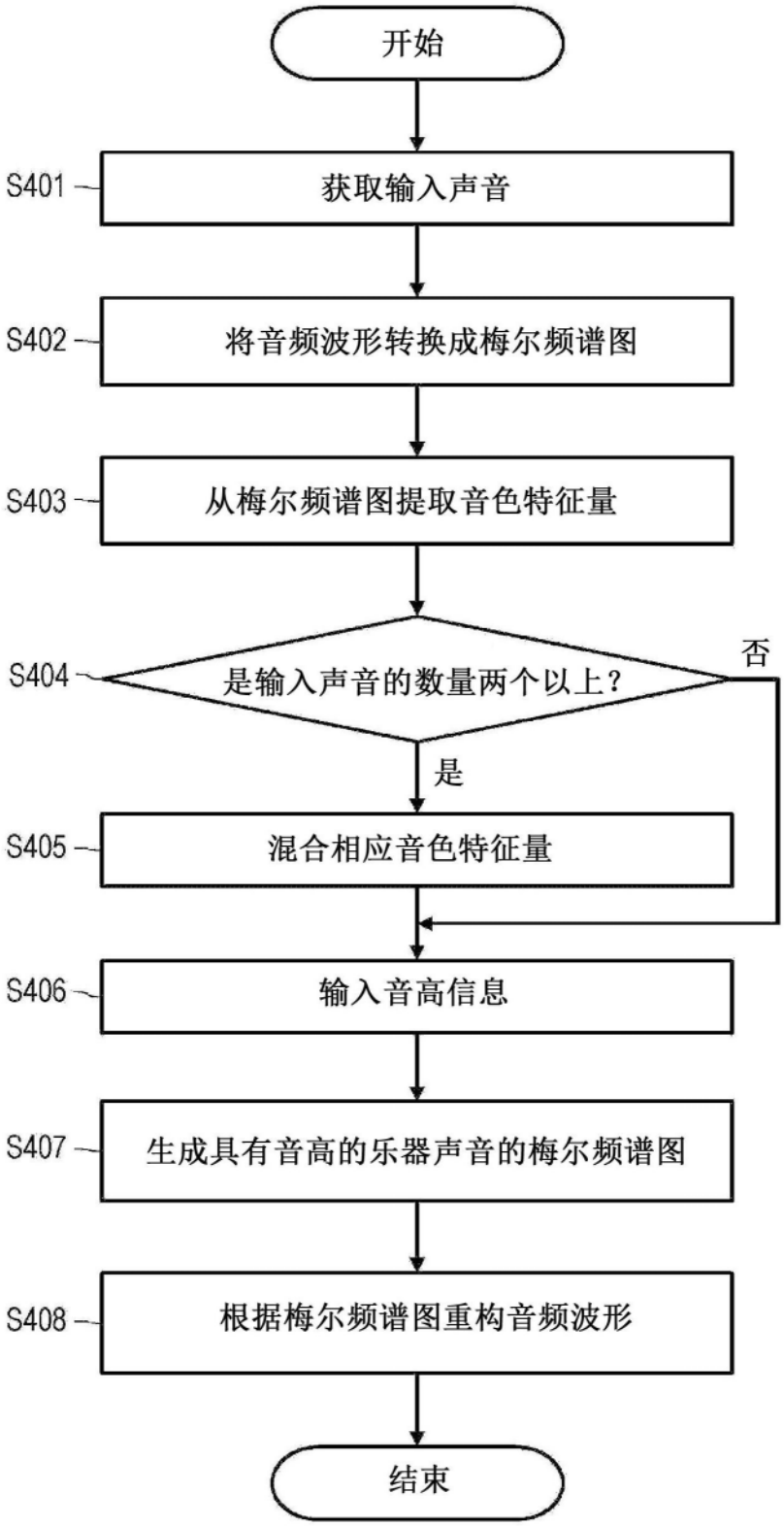

开始
S401
获取输入声音
S402
将音频波形转换成梅尔频谱图
S403
从梅尔频谱图提取音色特征量
S404
是输入声音的数量两个以上？
否
是
S405
混合相应音色特征量
S406
输入音高信息
S407
生成具有音高的乐器声音的梅尔频谱图
S408
根据梅尔频谱图重构音频波形
结束

图4

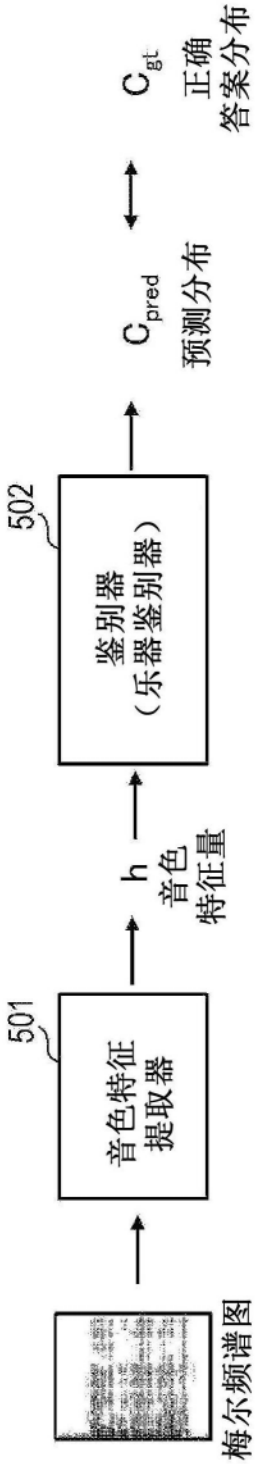

梅尔频谱图
501
音色特征提取器
h
音色特征量
502
鉴别器（乐器鉴别器）
C_{pred}
预测分布
C_{gt}
正确答案分布

图5

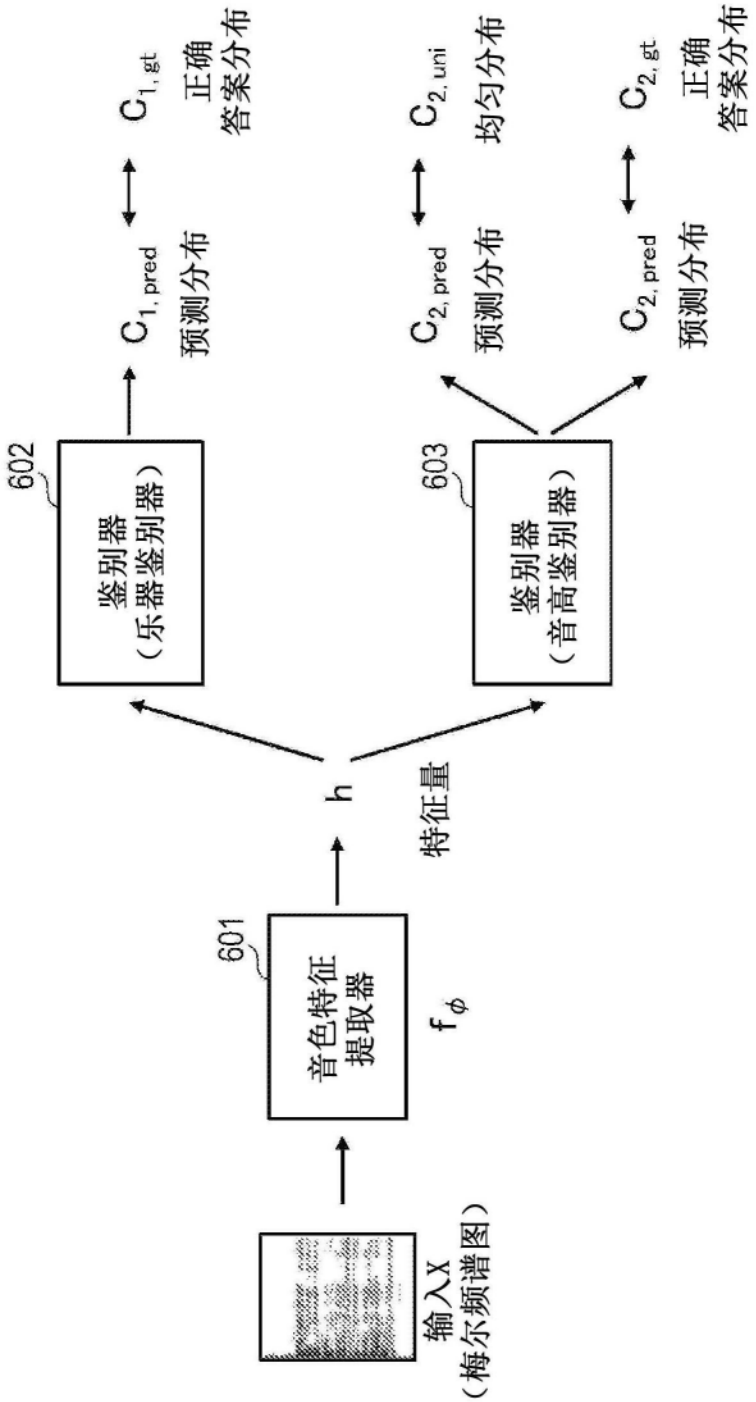

输入X
（梅尔频谱图）
601
音色特征
提取器
f_ϕ
h
特征量
602
鉴别器
（乐器鉴别器）
603
鉴别器
（音高鉴别器）
$C_{1,pred}$
预测分布
$C_{1,gt}$
正确
答案分布
$C_{2,pred}$
预测分布
$C_{2,uni}$
均匀分布
$C_{2,pred}$
预测分布
$C_{2,gt}$
正确
答案分布

图6

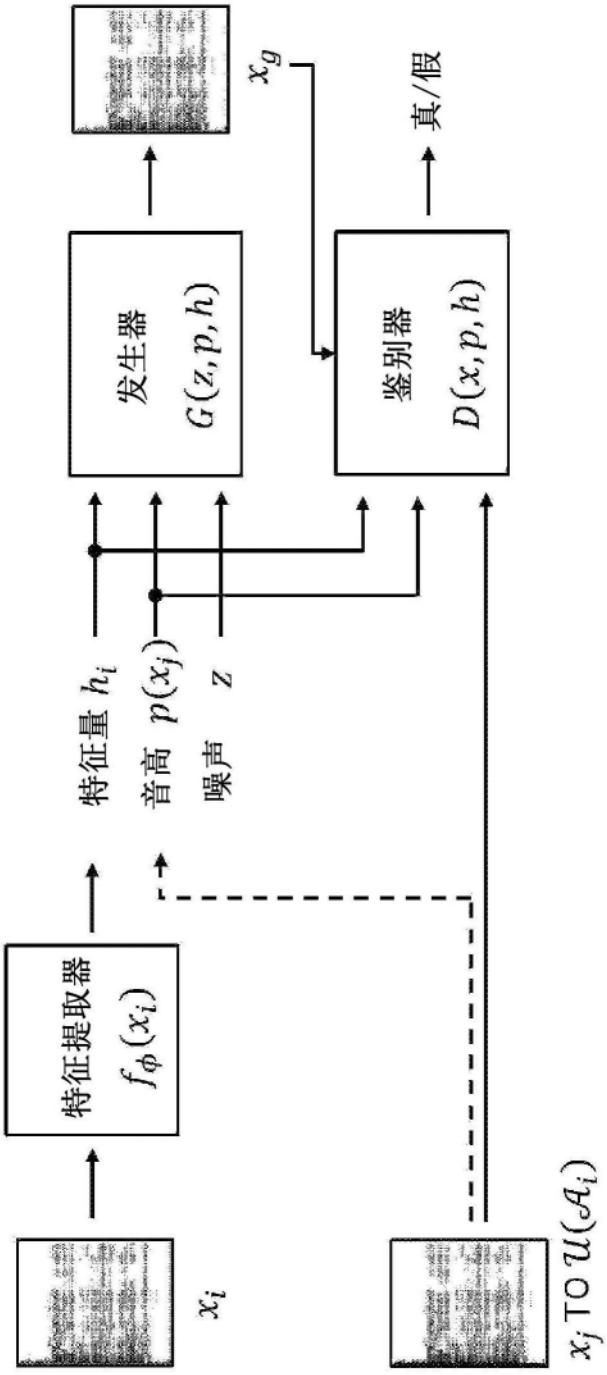

x_g
真/假
发生器
$G(z,p,h)$
鉴别器
$D(x,p,h)$
特征量 h_i
音高 $p(x_j)$
噪声 z
特征提取器
$f_\phi(x_i)$
x_i
x_j TO $\mathcal{U}(\mathcal{A}_i)$

图7

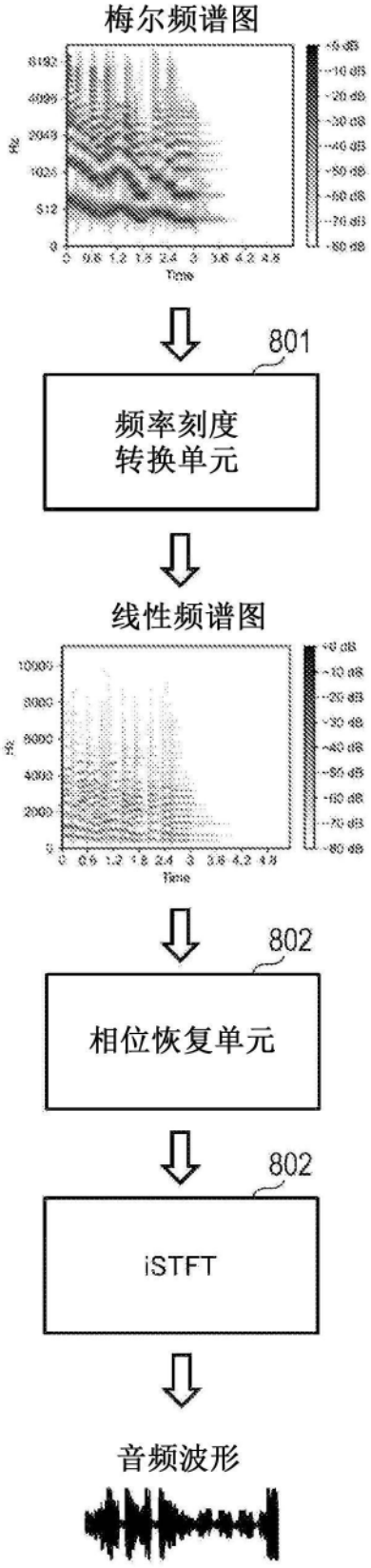


图8

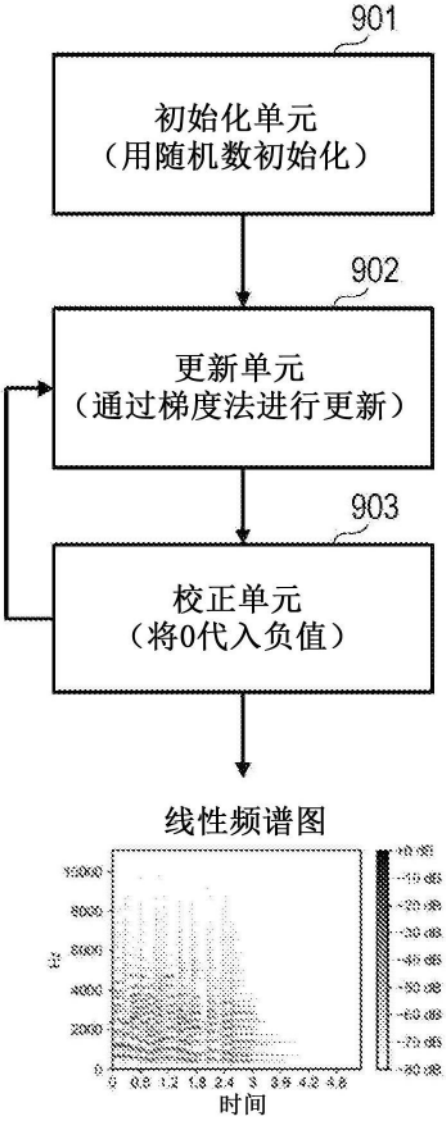

901
初始化单元
（用随机数初始化）
902
更新单元
（通过梯度法进行更新）
903
校正单元
（将0代入负值）
线性频谱图
时间

图9

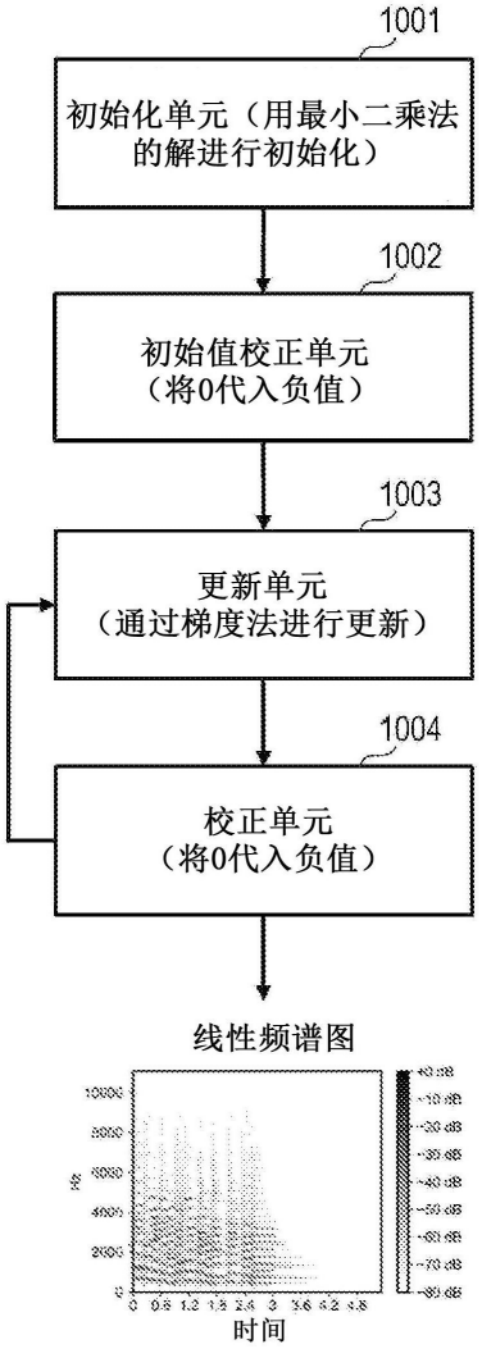

1001
初始化单元（用最小二乘法的解进行初始化）
1002
初始值校正单元（将0代入负值）
1003
更新单元（通过梯度法进行更新）
1004
校正单元（将0代入负值）
线性频谱图
时间

图10

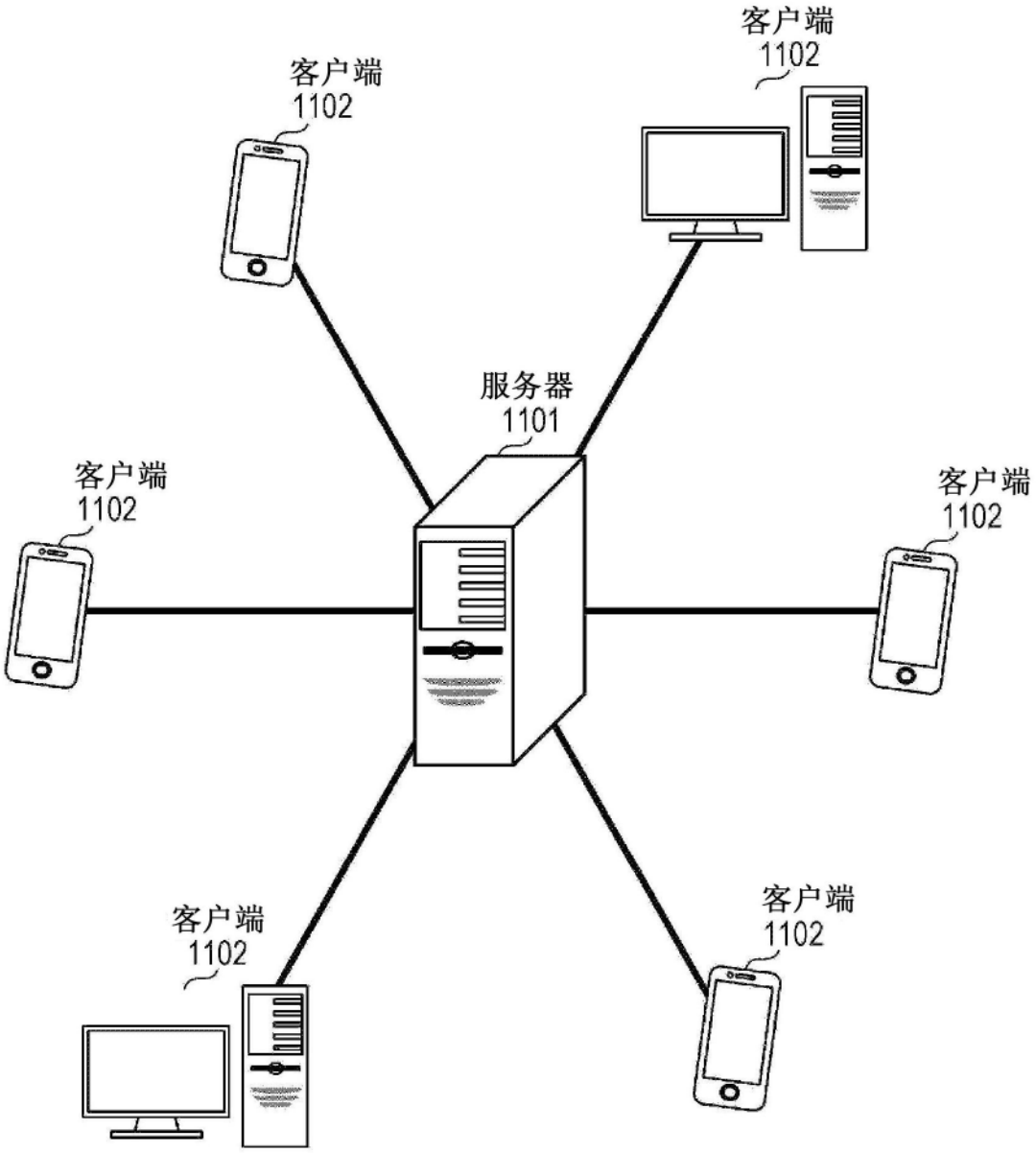

客户端
1102
客户端
1102
服务器
1101
客户端
1102
客户端
1102
客户端
1102
客户端
1102
客户端服务器模型 1100
(声音生成系统 300)

图11

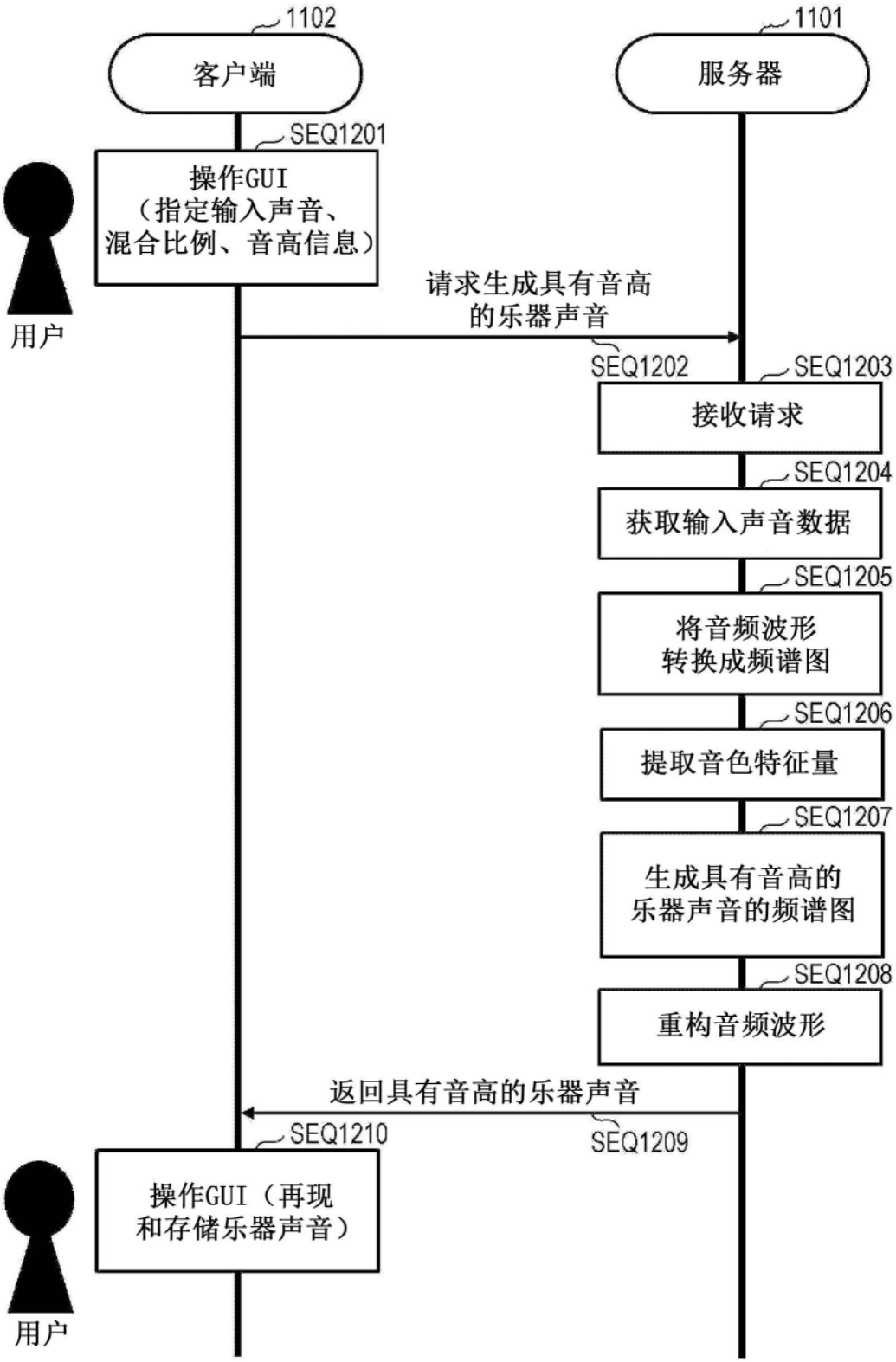

1102
客户端
1101
服务器
SEQ1201
操作GUI
（指定输入声音、
混合比例、音高信息）
用户
请求生成具有音高
的乐器声音
SEQ1202
SEQ1203
接收请求
SEQ1204
获取输入声音数据
SEQ1205
将音频波形
转换成频谱图
SEQ1206
提取音色特征量
SEQ1207
生成具有音高的
乐器声音的频谱图
SEQ1208
重构音频波形
返回具有音高的乐器声音
SEQ1209
SEQ1210
操作GUI（再现
和存储乐器声音）
用户

图12

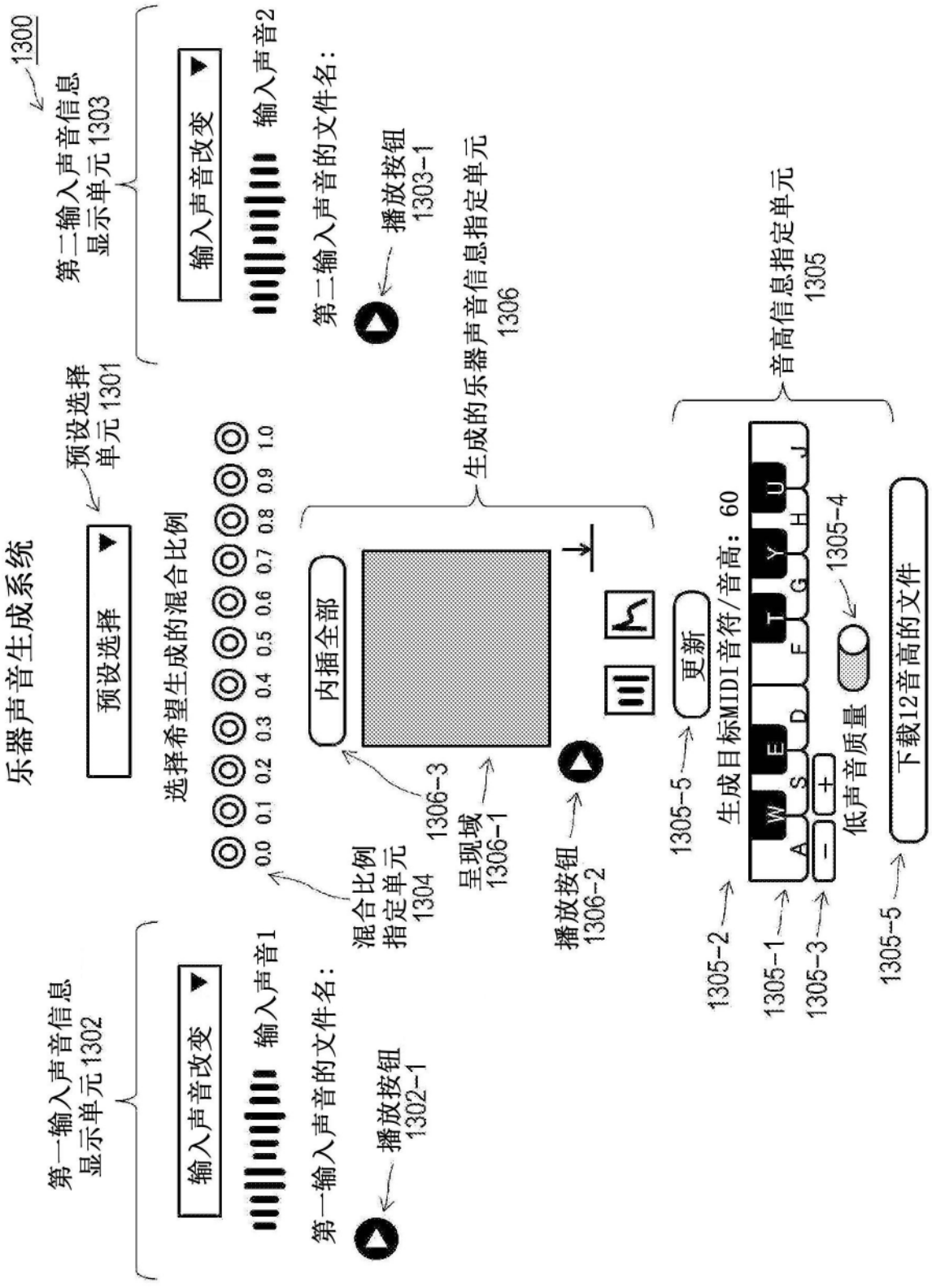

1300
乐器声音生成系统
第一输入声音信息显示单元 1302
预设选择单元 1301
第二输入声音信息显示单元 1303
预设选择
选择希望生成的混合比例
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
输入声音改变
输入声音1
输入声音2
第一输入声音的文件名：
第二输入声音的文件名：
播放按钮 1302-1
播放按钮 1303-1
混合比例指定单元 1304
内插全部
生成的乐器声音信息指定单元 1306
1306-3
呈现域 1306-1
播放按钮 1306-2
1305-5
更新
音高信息指定单元 1305
生成目标MIDI音符/音高：60
1305-2
1305-1
1305-3
低声音质量
1305-4
1305-5
下载12音高的文件

图13

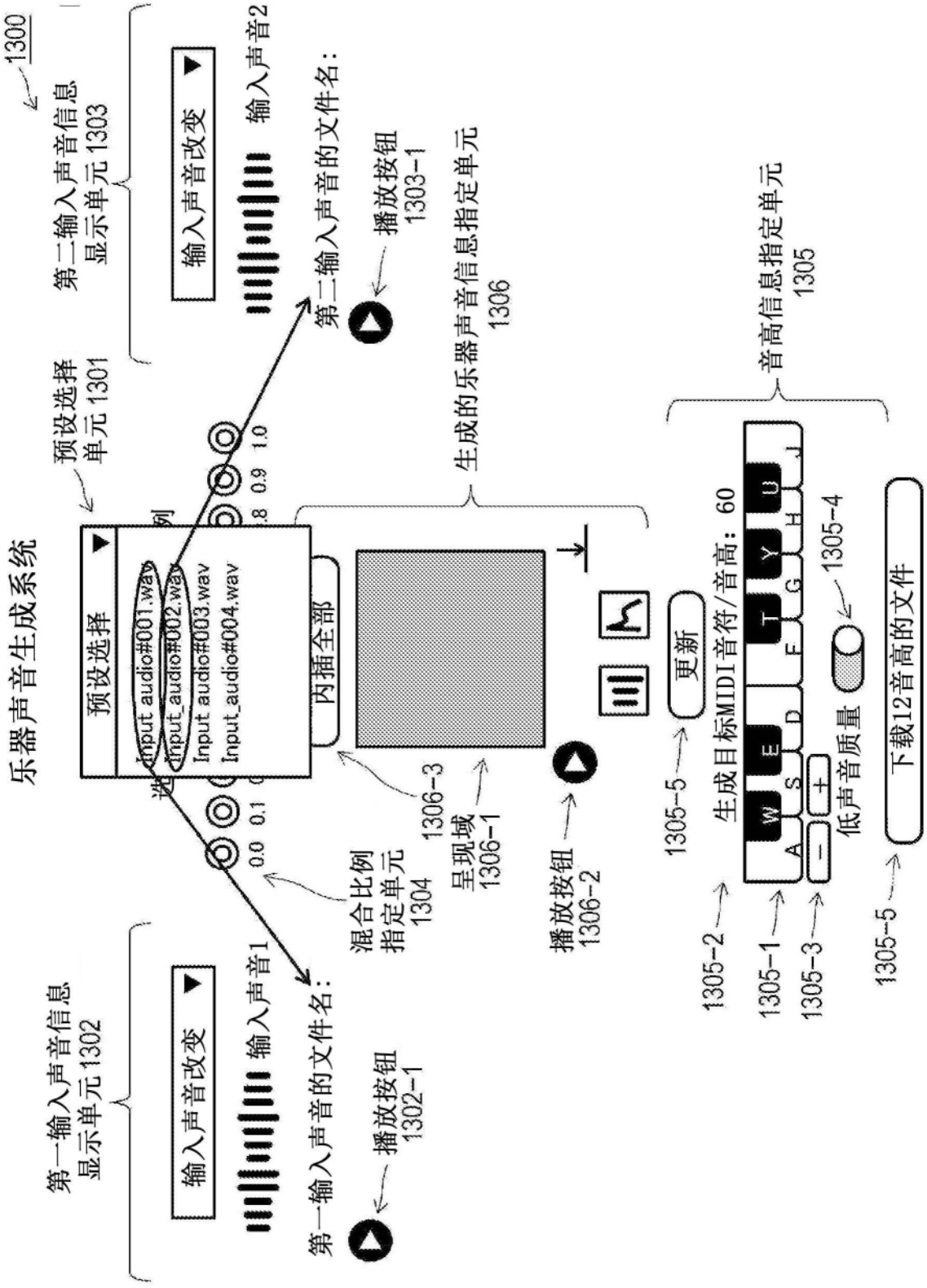

1300
乐器声音生成系统
第一输入声音信息显示单元 1302
输入声音改变
输入声音1
第一输入声音的文件名：
播放按钮 1302-1
预设选择单元 1301
预设选择
Input audio#001.wav
Input_audio#002.wav
Input audio#003.wav
Input_audio#004.wav
内插全部
混合比例指定单元 1304
0.0 0.1 0.8 0.9 1.0
第二输入声音信息显示单元 1303
输入声音改变
输入声音2
第二输入声音的文件名：
播放按钮 1303-1
生成的乐器声音信息指定单元 1306
1306-3
呈现域 1306-1
播放按钮 1306-2
音高信息指定单元 1305
1305-5
更新
生成目标MIDI音符/音高：60
1305-2
A S D F G H J
W E T Y U
1305-1
1305-3
−
+
低声音质量
1305-4
1305-5
下载12音高的文件

图14

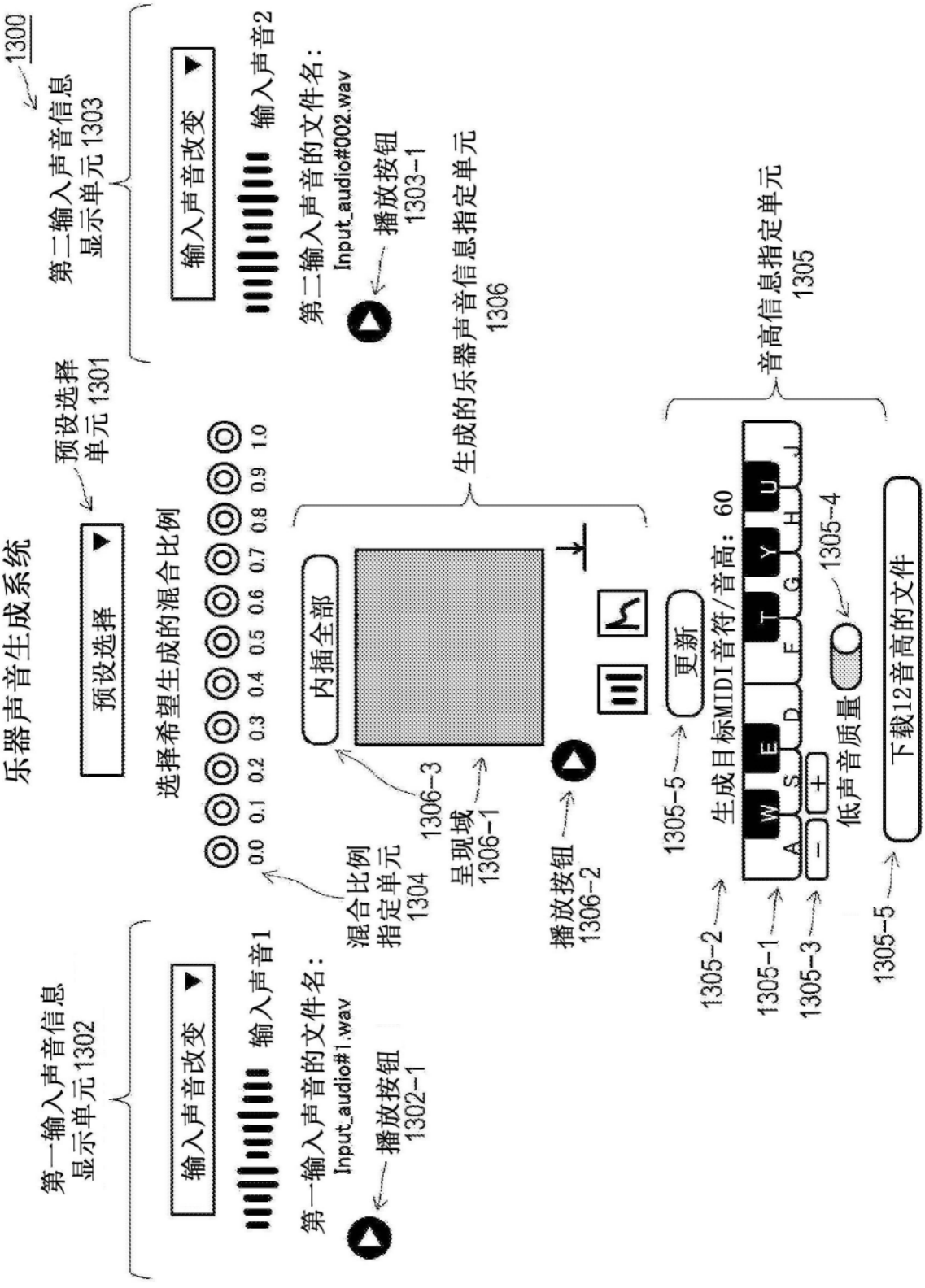

1300
乐器声音生成系统
第一输入声音信息显示单元 1302
输入声音改变
输入声音1
第一输入声音的文件名：
Input_audio#1.wav
播放按钮 1302-1
预设选择单元 1301
预设选择
选择希望生成的混合比例
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
混合比例指定单元 1304
内插全部
1306-3
呈现域 1306-1
播放按钮 1306-2
第二输入声音信息显示单元 1303
输入声音改变
输入声音2
第二输入声音的文件名：
Input_audio#002.wav
播放按钮 1303-1
生成的乐器声音信息指定单元 1306
1305-5
更新
音高信息指定单元 1305
1305-2
生成目标MIDI音符/音高：60
1305-1
1305-3
低声音质量
1305-4
1305-5
下载12音高的文件

图15

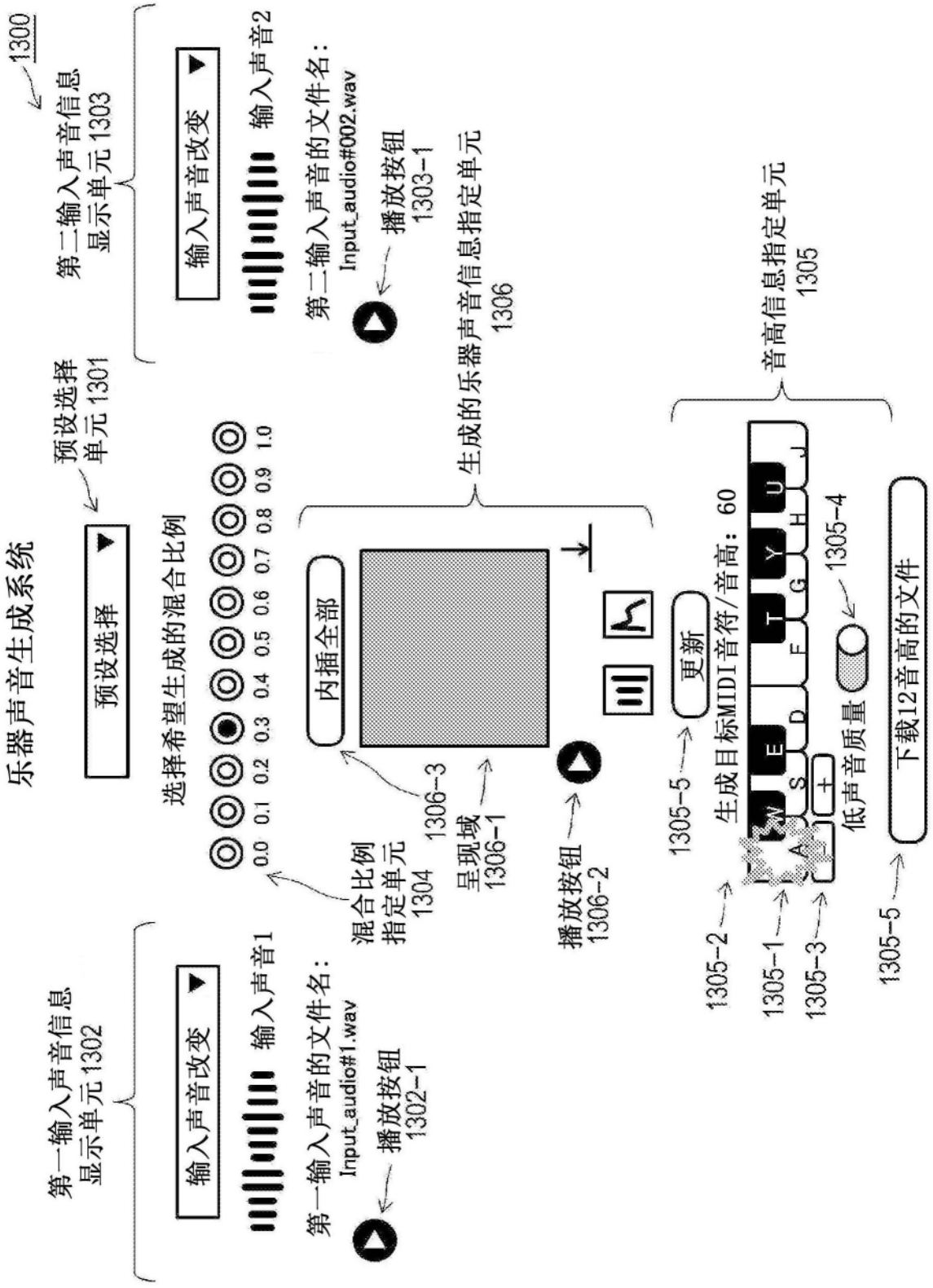

1300
乐器声音生成系统
第一输入声音信息显示单元 1302
预设选择单元 1301
第二输入声音信息显示单元 1303
预设选择
输入声音改变
输入声音1
输入声音2
选择希望生成的混合比例
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
第一输入声音的文件名：
Input_audio#1.wav
播放按钮 1302-1
混合比例指定单元 1304
内插全部
1306-3
呈现域 1306-1
播放按钮 1306-2
第二输入声音的文件名：
Input_audio#002.wav
播放按钮 1303-1
生成的乐器声音信息指定单元 1306
1305-5
更新
生成目标MIDI音符/音高：60
音高信息指定单元 1305
1305-2
1305-1
1305-3
低声音质量
1305-4
1305-5
下载12音高的文件

图16

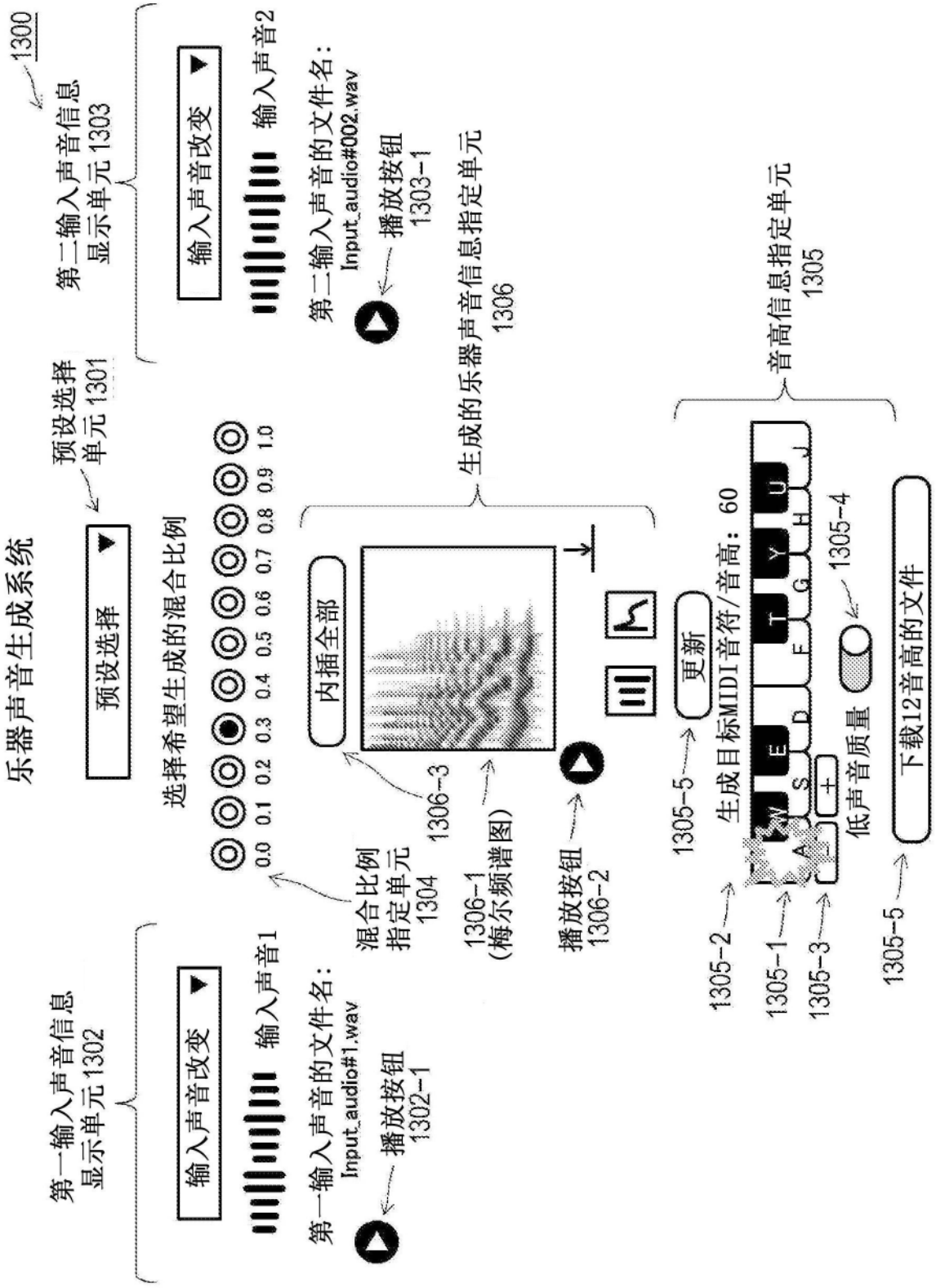

1300
乐器声音生成系统
第一输入声音信息
显示单元 1302
输入声音改变
输入声音1
第一输入声音的文件名：
Input_audio#1.wav
播放按钮
1302-1
预设选择
单元 1301
预设选择
选择希望生成的混合比例
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
混合比例
指定单元
1304
第二输入声音信息
显示单元 1303
输入声音改变
输入声音2
第二输入声音的文件名：
Input_audio#002.wav
播放按钮
1303-1
生成的乐器声音信息指定单元
1306
内插全部
1306-3
1306-1
(梅尔频谱图)
播放按钮
1306-2
1305-5
更新
生成目标MIDI音符/音高：60
音高信息指定单元
1305
1305-2
1305-1
1305-3
低声音质量
1305-4
1305-5
下载12音高的文件

图17

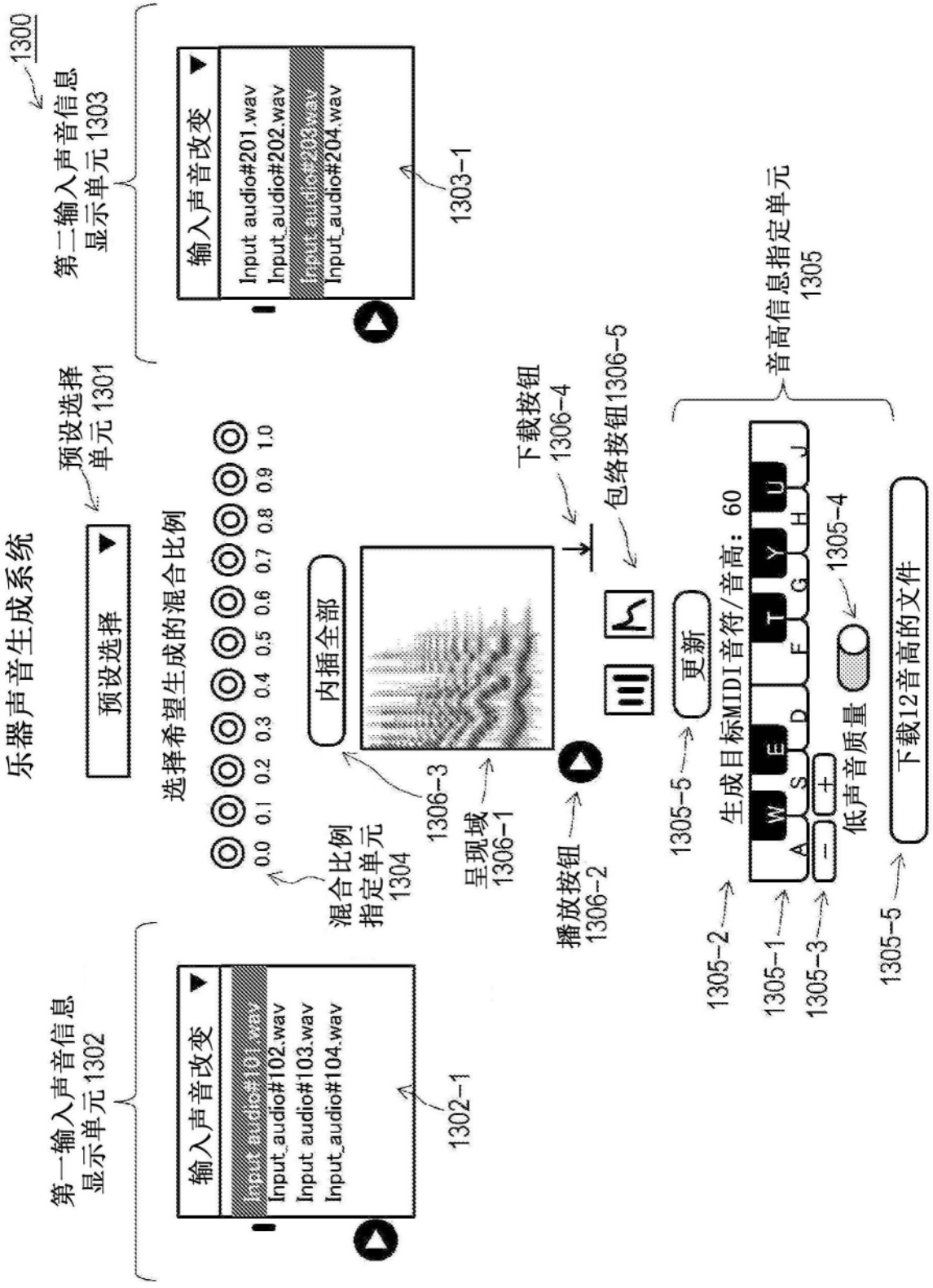

1300
乐器声音生成系统
第一输入声音信息显示单元1302
输入声音改变
Input audio#101.wav
Input_audio#102.wav
Input audio#103.wav
Input_audio#104.wav
1302-1
预设选择单元1301
预设选择
选择希望生成的混合比例
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
混合比例指定单元1304
内插全部
1306-3
呈现域1306-1
播放按钮1306-2
下载按钮1306-4
包络按钮1306-5
第二输入声音信息显示单元1303
输入声音改变
Input audio#201.wav
Input_audio#202.wav
Input audio#203.wav
Input_audio#204.wav
1303-1
音高信息指定单元1305
1305-5
更新
生成目标MIDI音符/音高：60
A S D F G H J
W E T Y U
1305-2
1305-1
1305-3
低声音质量
1305-4
1305-5
下载12音高的文件

图18

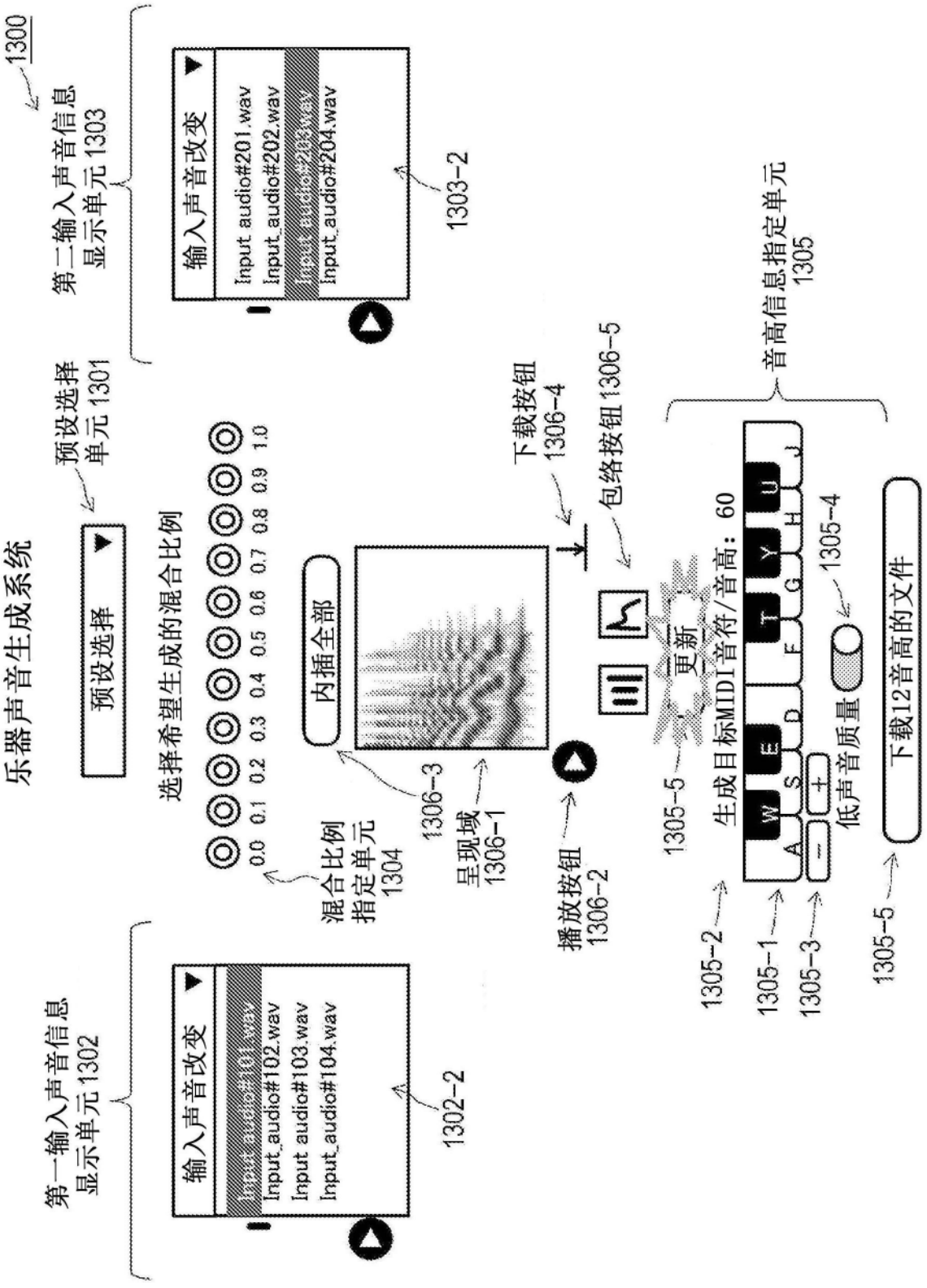

1300
乐器声音生成系统
第一输入声音信息显示单元 1302
输入声音改变
Input_audio#101.wav
Input_audio#102.wav
Input_audio#103.wav
Input_audio#104.wav
1302-2
预设选择单元 1301
预设选择
选择希望生成的混合比例
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
混合比例指定单元 1304
内插全部
1306-3
呈现域 1306-1
播放按钮 1306-2
下载按钮 1306-4
包络按钮 1306-5
更新
1305-5
第二输入声音信息显示单元 1303
输入声音改变
Input_audio#201.wav
Input_audio#202.wav
Input_audio#203.wav
Input_audio#204.wav
1303-2
音高信息指定单元 1305
生成目标MIDI音符/音高：60
A W S E D F T G Y H U J
1305-2
1305-1
1305-3
低声音质量
1305-4
1305-5
下载12音高的文件

图19

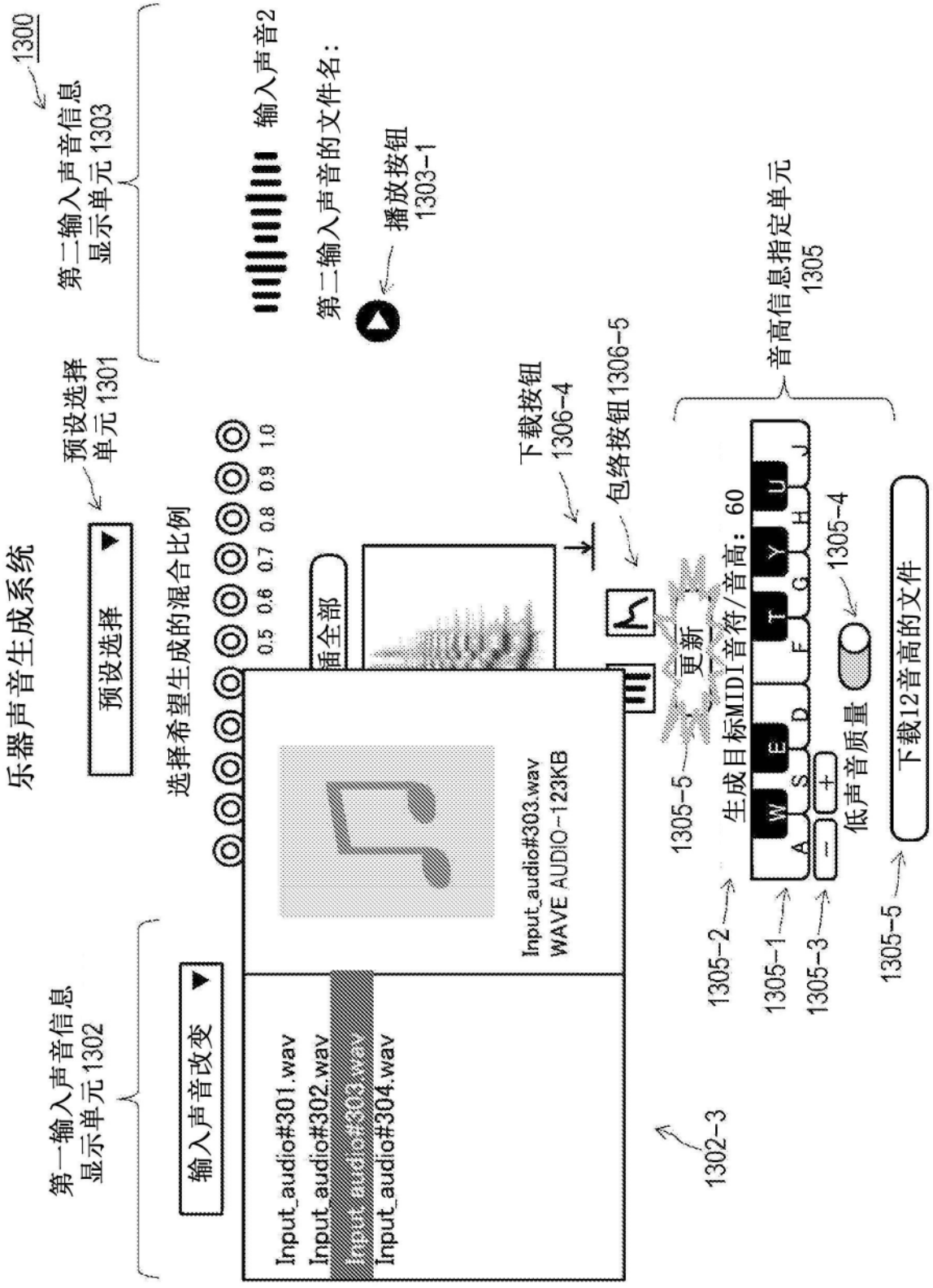

1300
乐器声音生成系统
第二输入声音信息显示单元 1303
预设选择单元 1301
预设选择
选择希望生成的混合比例
0.5 0.6 0.7 0.8 0.9 1.0
输入声音2
第二输入声音的文件名：
播放按钮 1303-1
下载按钮 1306-4
包络按钮 1306-5
插全部
音高信息指定单元 1305
第一输入声音信息显示单元 1302
输入声音改变
Input_audio#301.wav
Input_audio#302.wav
Input_audio#303.wav
Input_audio#304.wav
Input_audio#303.wav
WAVE AUDIO-123KB
更新
生成目标MIDI音符/音高：60
低声音质量
下载12音高的文件
1305-1
1305-2
1305-3
1305-4
1305-5
1305-5
1302-3

图20

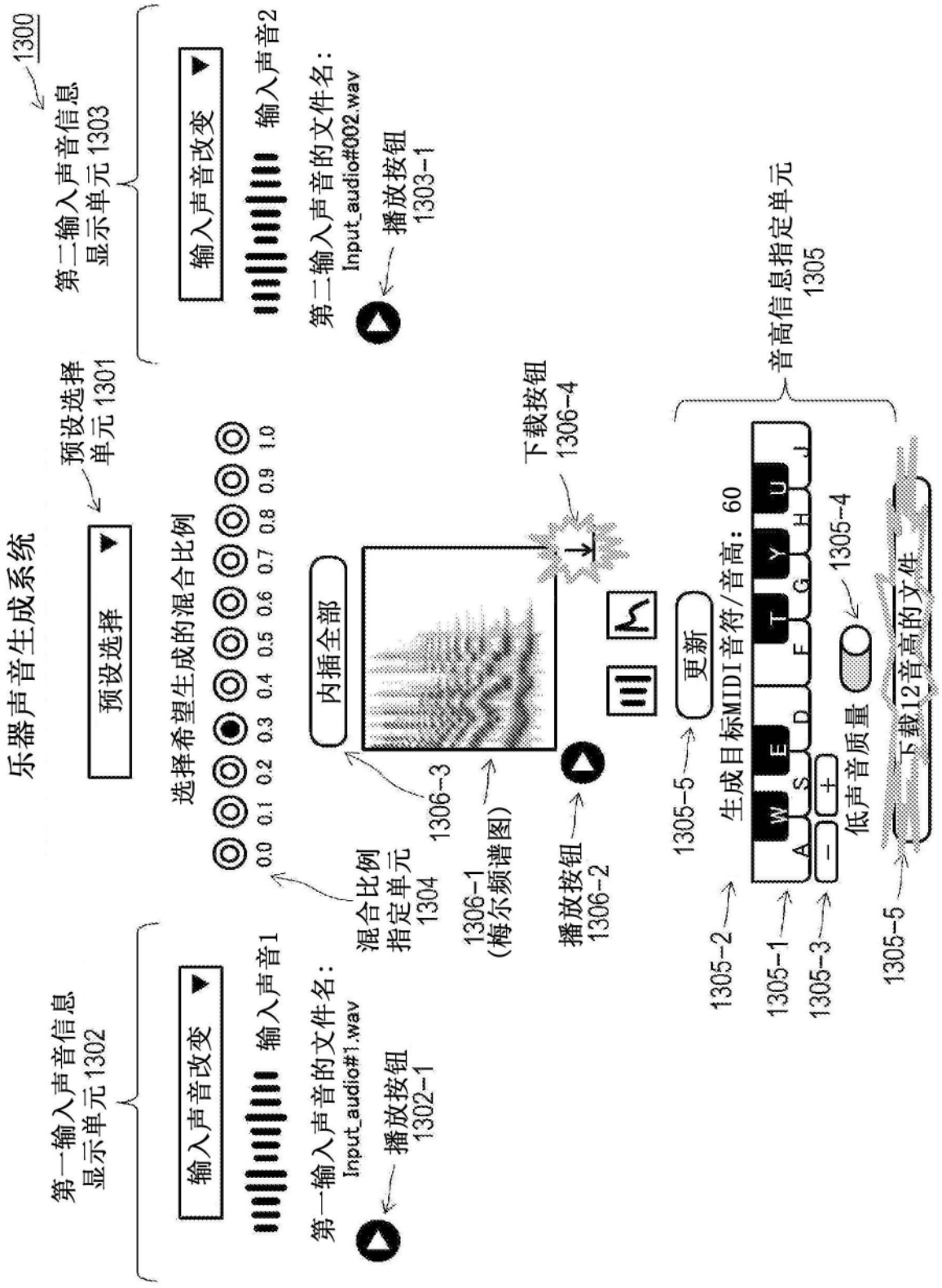

1300
乐器声音生成系统
第一输入声音信息显示单元1302
第二输入声音信息显示单元1303
预设选择单元1301
预设选择
输入声音改变
输入声音1
输入声音2
选择希望生成的混合比例
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
第一输入声音的文件名：
Input_audio#1.wav
播放按钮
1302-1
第二输入声音的文件名：
Input_audio#002.wav
播放按钮
1303-1
混合比例指定单元1304
内插全部
1306-3
1306-1
(梅尔频谱图)
播放按钮
1306-2
下载按钮
1306-4
1305-5
更新
音高信息指定单元
1305
生成目标MIDI音符/音高：60
1305-2
1305-1
1305-3
低声音质量
1305-4
1305-5
下载12音高的文件

图21

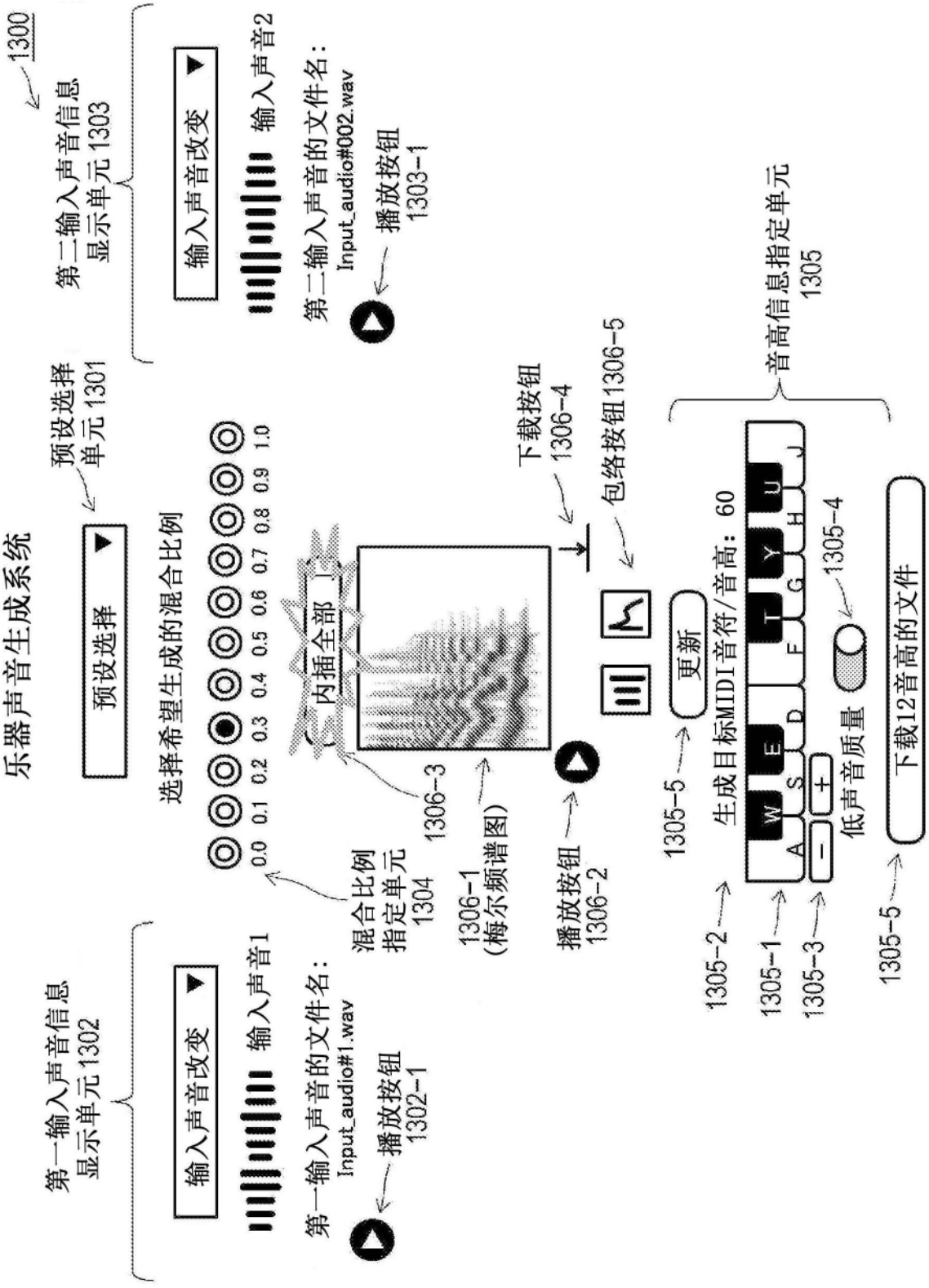

1300
乐器声音生成系统
预设选择单元 1301
预设选择
选择希望生成的混合比例
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
第一输入声音信息显示单元 1302
输入声音改变
输入声音1
第一输入声音的文件名：
Input_audio#1.wav
播放按钮 1302-1
混合比例指定单元 1304
内插全部
1306-3
1306-1
(梅尔频谱图)
播放按钮 1306-2
下载按钮 1306-4
包络按钮 1306-5
第二输入声音信息显示单元 1303
输入声音改变
输入声音2
第二输入声音的文件名：
Input_audio#002.wav
播放按钮 1303-1
音高信息指定单元 1305
更新
1305-5
生成目标MIDI音符/音高：60
A S D F G H J
W E T Y U
1305-2
1305-1
1305-3
1305-4
低声音质量
1305-5
下载12音高的文件

图22

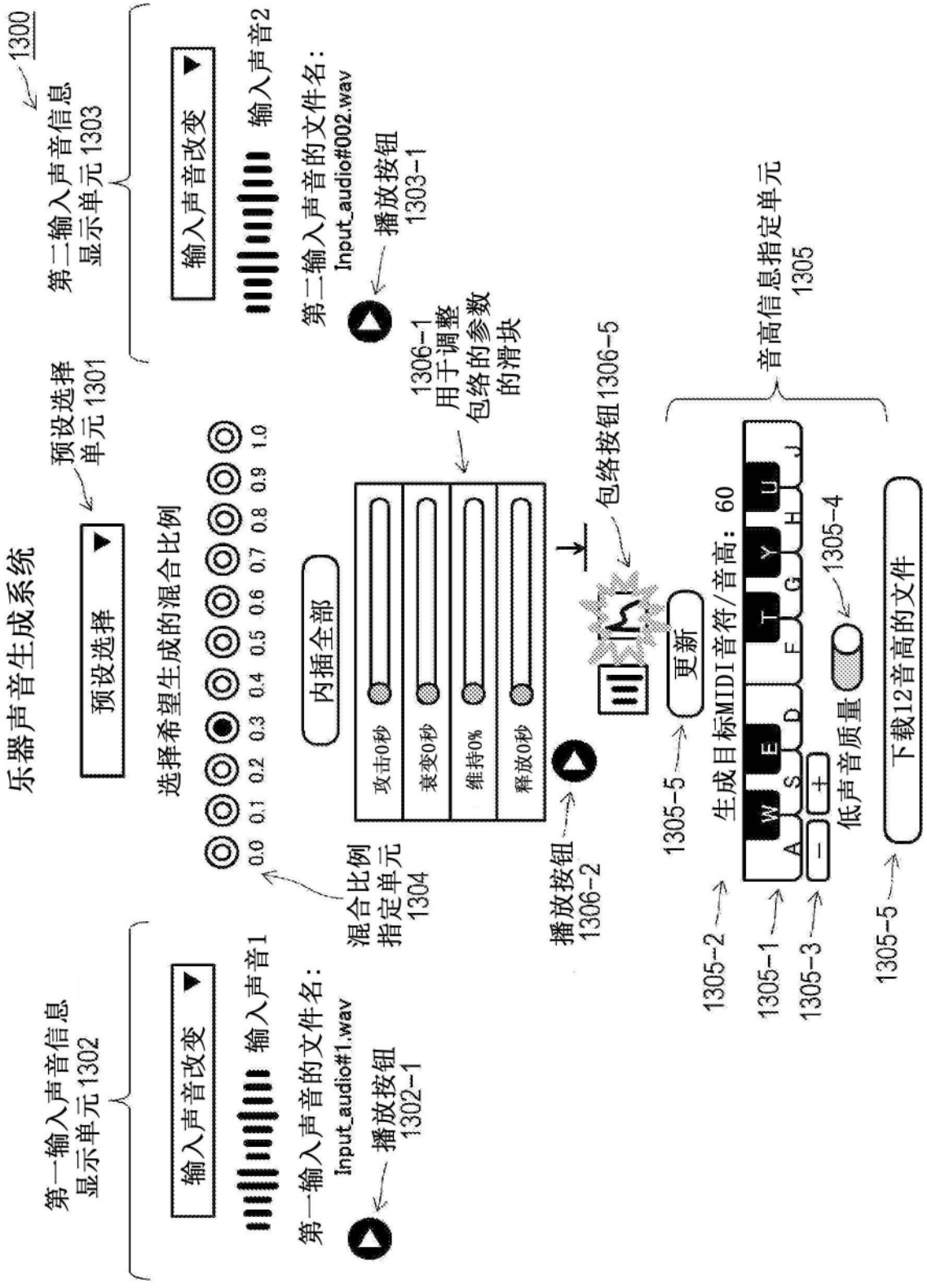

1300
乐器声音生成系统
第一输入声音信息显示单元 1302
预设选择单元 1301
第二输入声音信息显示单元 1303
预设选择
输入声音改变
输入声音改变
输入声音1
输入声音2
选择希望生成的混合比例
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
第一输入声音的文件名：
Input_audio#1.wav
播放按钮 1302-1
混合比例指定单元 1304
内插全部
攻击0秒
衰变0秒
维持0%
释放0秒
播放按钮 1306-2
第二输入声音的文件名：
Input_audio#002.wav
播放按钮 1303-1
1306-1 用于调整包络的参数的滑块
包络按钮 1306-5
更新
1305-5
音高信息指定单元 1305
1305-2 生成目标MIDI音符/音高：60
1305-1
1305-3
低声音质量
1305-4
下载12音高的文件
1305-5

图23

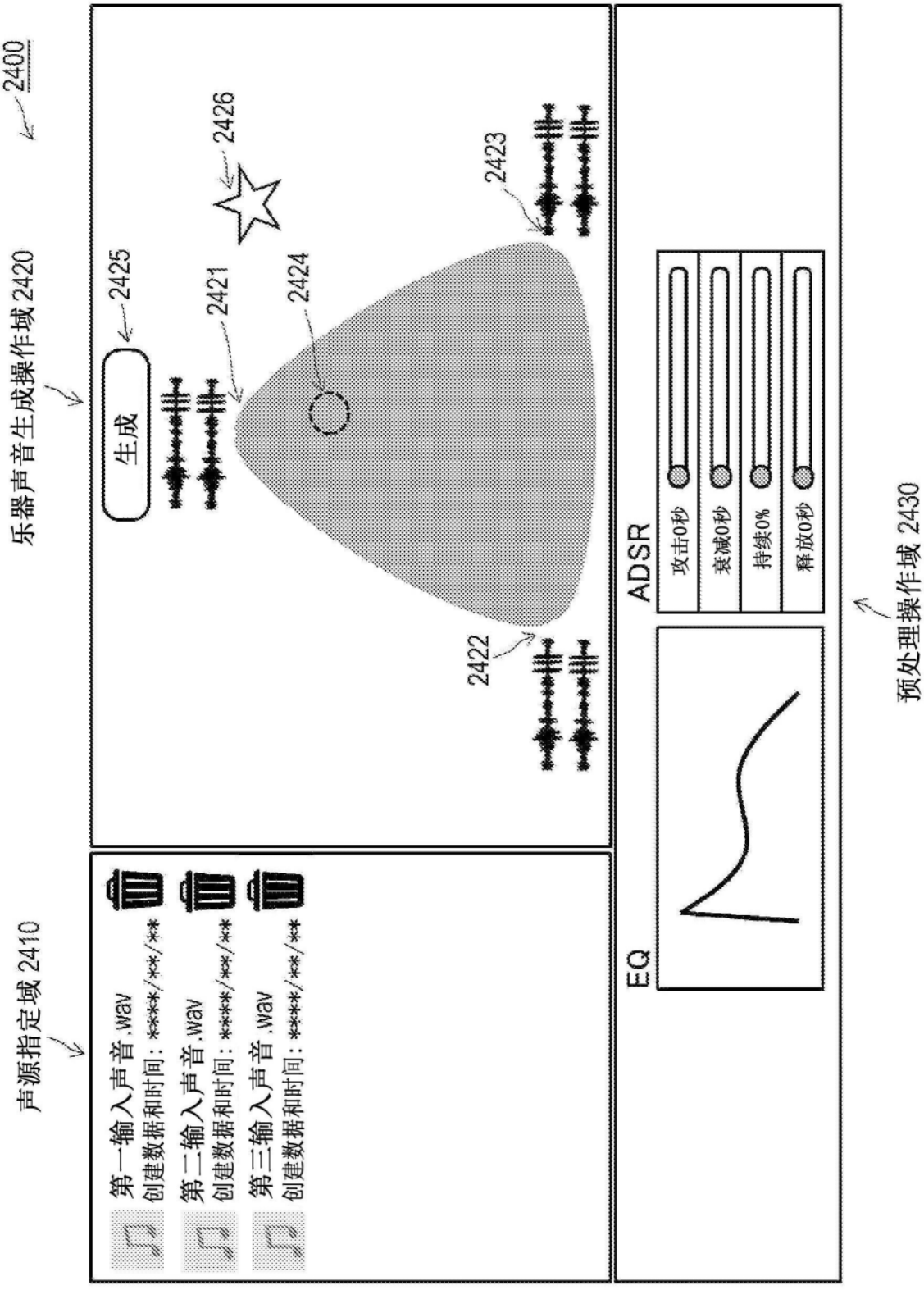

2400
乐器声音生成操作域 2420
声源指定域 2410
第一输入声音.wav
创建数据和时间：*****/**/**
第二输入声音.wav
创建数据和时间：*****/**/**
第三输入声音.wav
创建数据和时间：*****/**/**
生成
2425
2421
2424
2426
2423
2422
EQ
ADSR
攻击0秒
衰减0秒
持续0%
释放0秒
预处理操作域 2430

图24

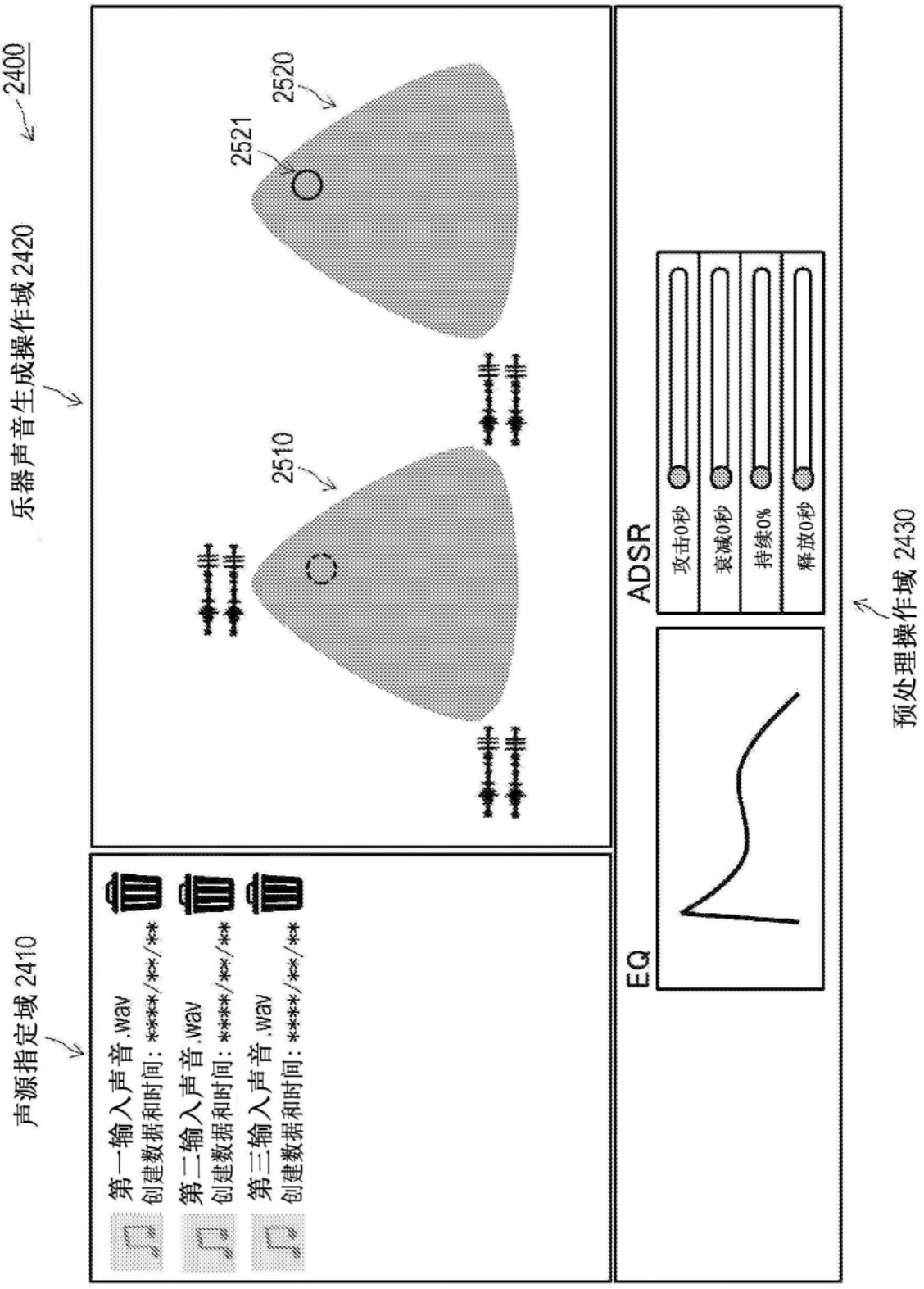

2400
乐器声音生成操作域 2420
2520
2521
2510
声源指定域 2410
第一输入声音.wav
创建数据和时间：*****/**/**
第二输入声音.wav
创建数据和时间：*****/**/**
第三输入声音.wav
创建数据和时间：*****/**/**
ADSR
攻击0秒
衰减0秒
持续0%
释放0秒
EQ
预处理操作域 2430

图25

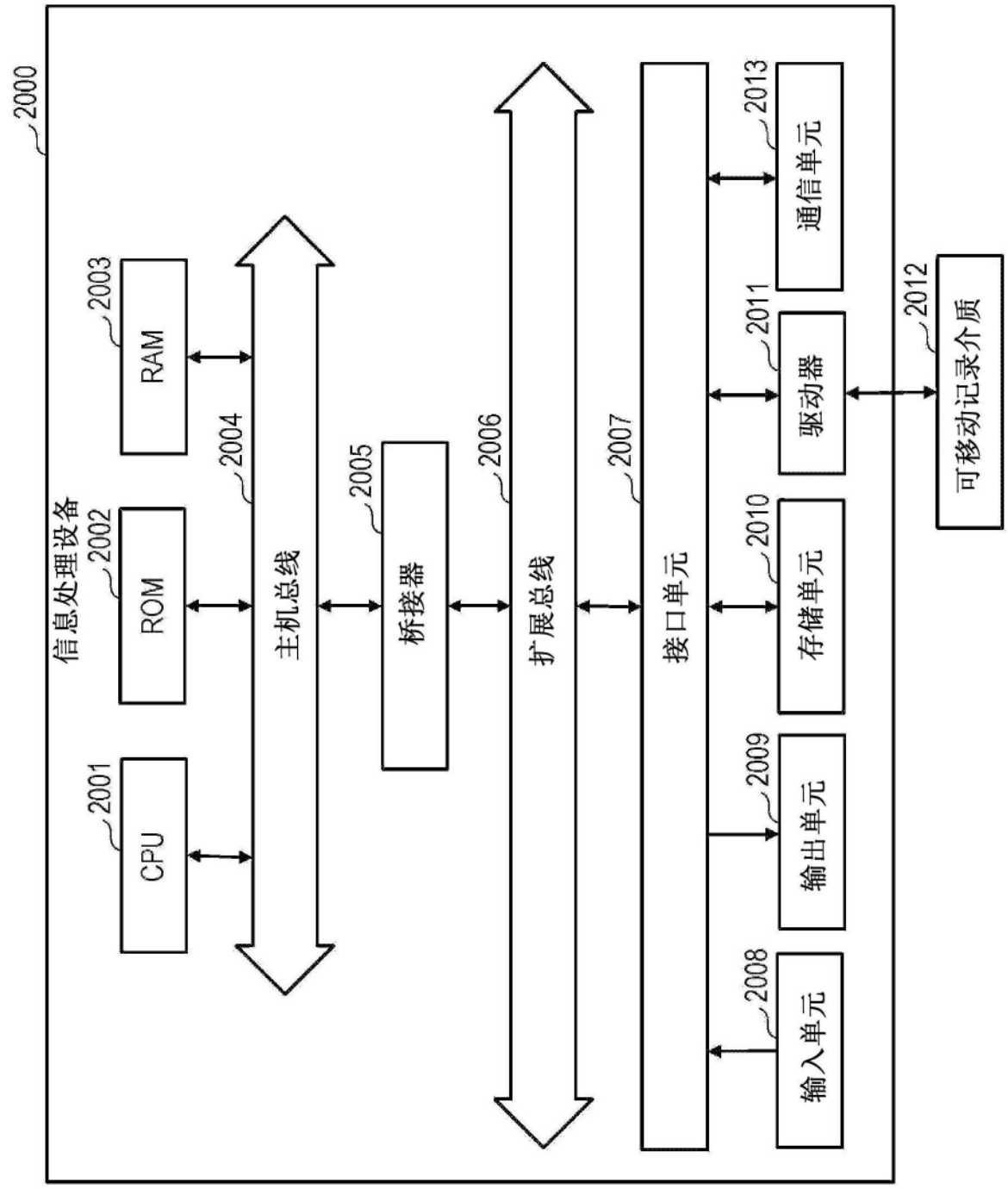

2000
信息处理设备
2001
CPU
2002
ROM
2003
RAM
2004
主机总线
2005
桥接器
2006
扩展总线
2007
接口单元
2008
输入单元
2009
输出单元
2010
存储单元
2011
驱动器
2012
可移动记录介质
2013
通信单元

图26

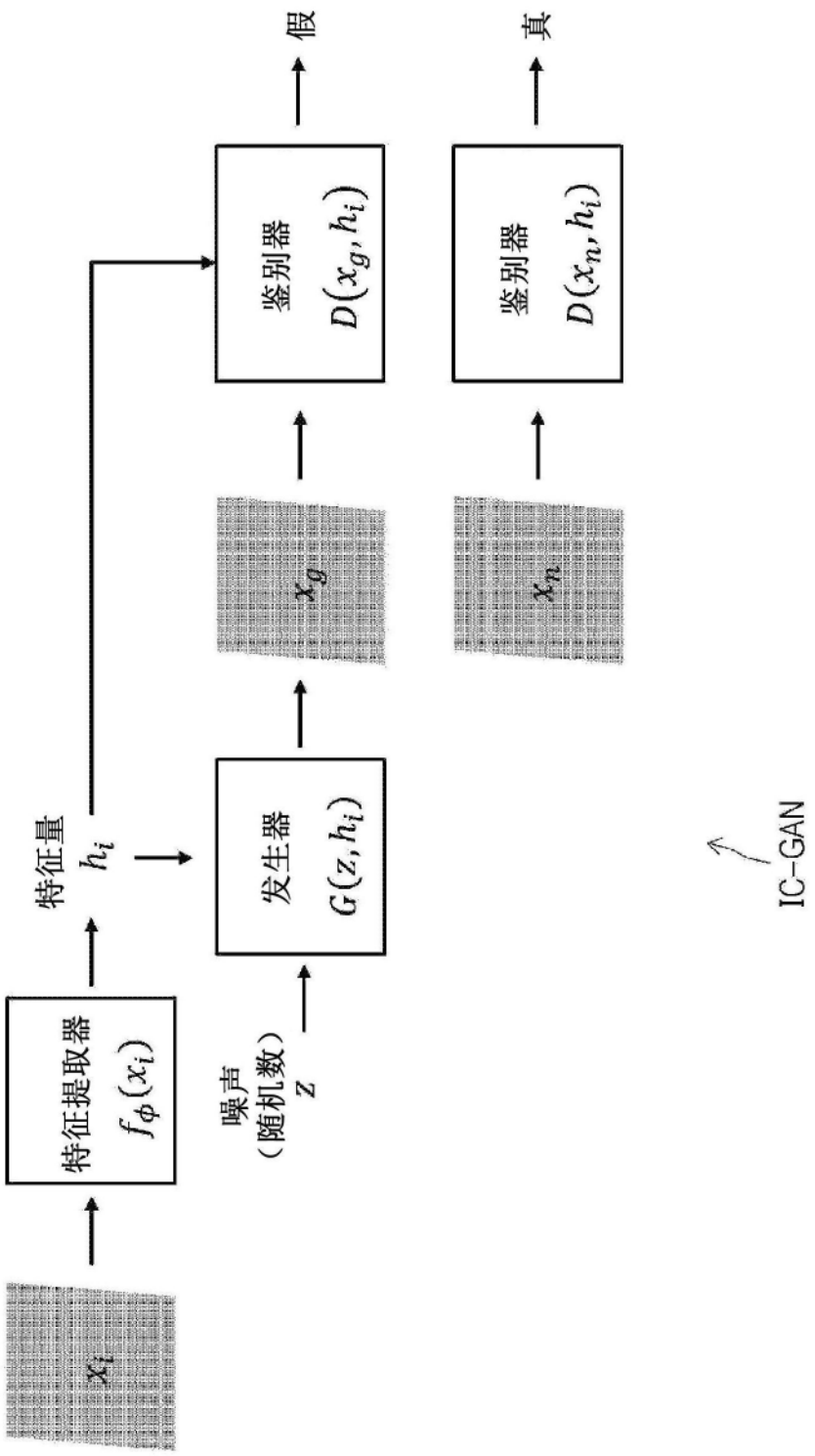

假
真
鉴别器
$D(x_g, h_i)$
鉴别器
$D(x_n, h_i)$
x_g
x_n
特征量
h_i
发生器
$G(z, h_i)$
特征提取器
$f_\phi(x_i)$
噪声
（随机数）
z
x_i
IC-GAN

图27