(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117561570 A
(43)申请公布日 2024.02.13

(21)申请号 202280045017.1
(22)申请日 2022.02.09
(30)优先权数据
2021-107651 2021.06.29 JP
(85)PCT国际申请进入国家阶段日
2023.12.22
(86)PCT国际申请的申请数据
PCT/JP2022/005001 2022.02.09
(87)PCT国际申请的公布数据
WO2023/276234 JA 2023.01.05
(71)申请人 索尼集团公司
地址 日本东京
(72)发明人 高桥直也
(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 沈丹阳
(51)Int.Cl.
G10L 21/007(2006.01)

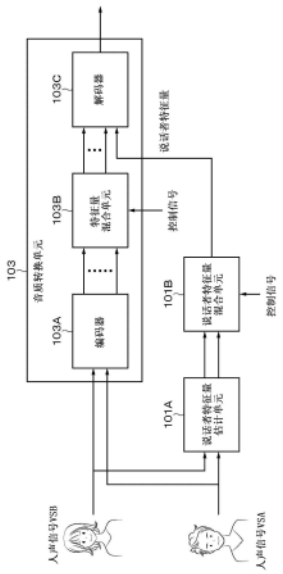
权利要求书2页 说明书14页 附图7页

(54)发明名称

信息处理装置、信息处理方法和程序

(57)摘要

为了执行有效的音质转换处理，例如，本发明提供了具有音质转换单元的信息处理装置，该音质转换单元用于从混合的声音信号执行人声信号和伴奏信号的声源分离，并使用声源分离的结果执行音质转换。



CN 117561570 A

1.一种信息处理装置,包括:

音质转换单元,从混合声音信号执行人声信号和伴奏信号的声源分离,并使用所述声源分离的结果执行音质转换。

2.根据权利要求1所述的信息处理装置,其中

第一人声信号通过所述声源分离从所述混合声音信号中分离,

收集的第二人声信号被输入至所述音质转换单元,并且

所述音质转换单元让所述第一人声信号和所述第二人声信号中的一个人声信号更接近另一人声信号。

3.根据权利要求2所述的信息处理装置,其中

能设置让所述一个人声信号更接近所述另一人声信号的改变量。

4.根据权利要求2所述的信息处理装置,进一步包括:

说话者特征量估计单元,估计与说话者相关的特征量,

其中,所述音质转换单元包括编码器和解码器。

5.根据权利要求4所述的信息处理装置,其中

与所述说话者相关的特征量是与不随时间改变的特征对应的特征量,

所述编码器从输入的人声信号中提取与随着时间改变的特征对应的特征量,以及

所述解码器基于由所述说话者特征量估计单元估计的特征量和由所述编码器提取的特征量,生成人声信号。

6.根据权利要求5所述的信息处理装置,其中

与不随时间改变的特征对应的特征量是说话者信息,并且与随着时间改变的特征对应的特征量包括声音音高信息、音量信息和对话信息中的至少一者。

7.根据权利要求6所述的信息处理装置,其中

所述特征量由嵌入矢量限定。

8.根据权利要求7所述的信息处理装置,其中

所述编码器通过使用学习模型,提取与随着时间改变的特征对应的特征量的嵌入矢量,所述学习模型通过执行针对从仅反映特定特征的特征量获得嵌入矢量的学习或者针对从人声信号中仅提取特定特征的学习而获得。

9.根据权利要求6所述的信息处理装置,其中

所述说话者特征量估计单元通过使用学习模型,估计所述说话者的特征量,所述学习模型通过基于预定说话者的人声信号针对估计所述说话者的所述说话者信息的学习而获得。

10.根据权利要求6所述的信息处理装置,其中

所述说话者特征量估计单元通过使用学习模型来估计所述说话者的特征量,所述学习模型通过基于预定人声信号针对估计所述说话者的说话者信息的学习而获得。

11.根据权利要求4所述的信息处理装置,其中

所述说话者特征量估计单元包括第一说话者特征量估计单元和第二说话者特征量估计单元,

所述信息处理装置进一步包括特征量结合单元,所述特征量结合单元对由所述第一说话者特征量估计单元估计的与所述说话者相关的特征量和由所述第二说话者特征量估计

单元估计的与所述说话者相关的特征量进行结合。

12.根据权利要求11所述的信息处理装置,其中

所述第一说话者特征量估计单元基于预定时间以上的人声信号估计与所述说话者相关的特征量,并且所述第二说话者特征量估计单元基于比所述预定时间短的时间的人声信号估计与所述说话者相关的特征量。

13.根据权利要求11所述的信息处理装置,其中

所述特征量结合单元中的结合系数根据所述第一人声信号和所述第二人声信号之间的相似度而改变。

14.根据权利要求13所述的信息处理装置,其中

所述结合系数是针对由所述第一说话者特征量估计单元估计的与所述说话者相关的特征量和由所述第二说话者特征量估计单元估计的与所述说话者相关的特征量中的每一者的权重。

15.一种信息处理方法,包括:

通过音质转换单元从混合声音信号执行人声信号和伴奏信号的声源分离,并使用所述声源分离的结果执行音质转换。

16.一种用于使计算机执行信息处理方法的程序,所述信息处理方法包括:

通过音质转换单元从混合声音信号执行人声信号和伴奏信号的声源分离,并使用所述声源分离的结果执行音质转换。

信息处理装置、信息处理方法和程序

技术领域

[0001] 本公开涉及信息处理装置、信息处理方法和程序。

背景技术

[0002] 已经提出了用于将自己对话(包括歌声)的音质转换成另一公司的音质的音质转换技术。音质是由说话者生成的人类语音,并且是指收听者在多个语音单元(例如,音素)上感知的语音属性,并且更具体地,是指即使对话具有相同的声音音高和音调,如果取决于收听者存在差异,则变得更近的元素。下面的专利文献1描述了一种在保持对话内容的同时将一般对话语音转换为另一说话者的音质的音质转换技术。

[0003] 引用列表

[0004] 专利文献

[0005] 专利文献1:日本专利申请公开号2018-005048。

发明内容

[0006] 本发明要解决的问题

[0007] 在该领域中,期望执行适当的音质转换处理。

[0008] 本公开的目的是提供用于执行适当的音质转换处理的信息处理装置、信息处理方法和程序。

[0009] 问题的解决方案

[0010] 例如,本公开提供,

[0011] 信息处理装置,包括:

[0012] 音质转换单元,从混合声音信号执行人声信号和伴奏信号的声源分离,并使用所述声源分离的结果执行音质转换。

[0013] 例如,本公开提供,

[0014] 信息处理方法,包括:

[0015] 通过音质转换单元从混合声音信号执行人声信号和伴奏信号的声源分离,并使用所述声源分离的结果执行音质转换。

[0016] 例如,本公开提供,

[0017] 用于使计算机执行信息处理方法的程序,所述信息处理方法包括:

[0018] 通过音质转换单元从混合声音信号执行人声信号和伴奏信号的声源分离,并使用所述声源分离的结果执行音质转换。

附图说明

[0019] 图1是示出用于描述一个实施方式的概况的示图。

[0020] 图2是示出根据实施方式的智能电话的配置实施例的框图。

[0021] 图3是示出根据实施方式的音质转换单元的配置实施例的框图。

[0022] 图4是示出根据实施方式的用于描述由音质转换单元执行的学习的实施例的示图。

[0023] 图5是示出根据实施方式的用于描述智能电话的操作时参考的示图。

[0024] 图6是示出用于描述与在实施方式中执行的音质转换处理相关联执行的处理的实施例的示图。

[0025] 图7是示出用于描述与在实施方式中执行的音质转换处理相关联地执行的处理的另一实施例的示图。

[0026] 图8是示出用于描述变形例的示图。

[0027] 图9是示出用于说明变形例的示图。

具体实施方式

[0028] 在下文中,将参考附图描述本公开的实施方式等。注意,将按照以下顺序给出描述。

[0029] <本公开的背景>

[0030] <一个实施方式>

[0031] <变形例>

[0032] 在下文中将描述的实施方式等是本公开的优选的具体实施例,并且本公开的内容不限于实施方式等。

[0033] <本公开的背景>

[0034] 首先,将描述本公开的背景,以便于理解本公开。近年来,在卡拉OK中,已经越来越多地对包含人声语音的原始声源进行声源分离,以获得人声信号和伴奏信号并使用分离的伴奏信号,而不是使用先前创建的乐器数字接口(MIDI)声源或记录的声源作为伴奏。

[0035] 随着这种声源分离技术的发展,能够获得伴奏声源制作成本降低、原始音乐享受卡拉OK等优点。同时,在卡拉OK中通常使用诸如混响、通过改变歌声语音的音高而添加的合唱、以及将音质改变为未指定的音质的语音改变器之类的效果,但是仍然难以对特定人的歌声语音进行改变。因此,例如,难以平滑地将音质转换成特定歌手的音质,诸如“让一人的语音稍微更接近原始歌曲的艺人的语音”。

[0036] 提出了一种音质转换技术,用于在保持对话内容的同时将一般对话语音转换为另一说话者的音质,如在上述专利文献1中描述的技术中。然而,一般来说,歌声的语音在声音音高和音质以及各种音乐表达方法(颤音等)上具有比普通对话更多的改变,并且歌声语音的转换是困难的。因此,目前只能转换为未指定的音质,例如转换为机器人风格或动画风格和性别转换,以及事先能够获得足够干净语音量的特定说话者的音质转换,难以转换为无法事先获得足够干净语音量的说话者。通常,获得足够量的干净语音需要大量时间和成本,并且例如,将音质转换成著名歌手的语音基本上非常困难。

[0037] 此外,因为必须实时执行音质转换,并且不能使用未来的信息,所以为了在卡拉OK中使用执行高质量转换更加困难。此外,通过声源分离而分离的声源可以包括在声源分离时产生的噪声,参考这种分离的语音转换的语音可能包括大量噪声,并且难以以更高质量转换。考虑到以上几点,将详细地描述本公开的一个实施方式。

[0038] <一个实施方式>

[0039] [一个实施方式的概述]

[0040] 首先,将参照图1描述一个实施方式的概要。在图1中所示的混合声源上执行声源分离处理PA。可以通过经由诸如光盘(CD)或网络的记录媒体的分配来提供混合声源。例如,混合声源包括艺术家的声音信号(这是第一人声信号的实施例,并且在下文中,视情况而定也称为人声信号VSA)。此外,混合声源包括除了人声信号VSA之外的信号(乐器声音等,并且在下文中适当地也称为伴奏信号)。

[0041] 同时,通过麦克风等收集卡拉OK用户的歌声。用户的歌声(第二人声信号的示例)也视情况被称为人声信号VSB。

[0042] 对人声信号VSA和人声信号VSB进行音质转换处理PB。在音质转换处理PB中,进行让人声信号VSA和人声信号VSB中的任一个人声信号更接近(类似)另一人声信号的处理。此时,可以根据预定控制信号,设定让任意一个人声信号更接近另一人声信号的改变量。例如,进行让卡拉OK用户的人声信号VSB更接近艺术家的人声信号VSA的音质转换处理。然后,进行将经过音质转换处理的人声信号VSB与伴奏信号添加的添加处理PC,并对添加处理PC获得的信号进行再生处理PD。因此,再生经历音质转换处理以近似艺术家的人声信号的用户的歌声语音。

[0043] [信息处理装置的配置实施例]

[0044] (总体配置实施例)

[0045] 图2是示出根据本实施方式的信息处理装置的配置实施例的框图。根据本实施方式的信息处理装置的实施例包括智能电话(智能电话100)。用户可以使用智能电话100容易地执行具有音质转换的卡拉OK。应注意,卡拉OK(即,歌声)在本实施方式中被描述为实施例,但是本公开不限于歌声,并且可应用于针对诸如对话的语音的音质转换处理。此外,根据本公开的信息处理装置不仅可应用于智能电话,而且可应用于便携式电子设备,诸如智能手表、个人计算机、固定卡拉OK设备等。

[0046] 例如,智能电话100包括控制单元101、声源分离单元102、音质转换单元103、麦克风104以及扬声器105。

[0047] 控制单元101整体控制整个智能电话100。控制单元101被配置为例如中央处理单元(CPU),并且包括存储程序的只读存储器(ROM)、用作工作存储器的随机存取存储器(RAM)等(注意,省略对这些存储器的说明)。

[0048] 控制单元101包括作为功能块的说话者特征量估计单元101A。说话者特征量估计单元101A估计与歌声的进行没有随时间改变的特征对应的特征量,具体而言,估计与说话者相关的特征量(以下,适当地称为说话者特征量)。

[0049] 此外,控制单元101包括作为功能块的特征量混合单元101B。特征量混合单元101B以适当的权重混合例如2个以上的说话者特征量。

[0050] 声源分离单元102将输入的混合声音信号分离成人声信号和伴奏信号(声源分离处理)。将通过声源分离获得的人声信号提供给音质转换单元103。此外,将通过声源分离获得的伴奏信号提供给扬声器105。

[0051] 音质转换单元103执行音质转换处理,使得由麦克风104收集的与用户的歌声语音相应的人声信号的音质近似于通过声源分离单元102进行声源分离而获得的人声信号。注意,稍后将描述由音质转换单元103执行的处理的细节。注意,本实施方式中的音质除了说

话者特征量之外进一步包括诸如声音音高和音量的特征量。

[0052] 例如,麦克风104收集智能电话100的用户的歌声或对话(在该实施例中,歌声)。与收集的歌声对应的人声信号被提供给音质转换单元103。

[0053] 添加单元(未示出)添加从声源分离单元102提供的伴奏信号和从音质转换单元103输出的人声信号。通过扬声器105再生所添加的信号。

[0054] 要注意的是,智能电话100可以具有除了在图2中示出的配置之外的配置(例如,配置为触摸面板的显示器或按钮)。

[0055] (音质转换单元的配置实施例)

[0056] 图3是示出了音质转换单元103的配置实施例的框图。音质转换单元103包括编码器103A、特征量混合单元103B和解码器103C。编码器103A使用通过预定学习获得的学习模型从人声信号提取特征量。编码器103A提取出的特征量例如是随着歌声的进行而随时间改变的特征量,具体包括声音音高信息、音量信息或对话(歌词)信息中的至少一者。

[0057] 特征量混合单元103B混合由编码器103A提取的特征量。由特征量混合单元103B混合的特征量被提供给解码器103C。

[0058] 解码器103C基于从特征量混合单元103B提供的特征量和说话者特征量生成人声信号。

[0059] (关于由音质转换单元执行的学习)

[0060] 接下来,将参考图4描述由音质转换单元103执行的学习方法的实施例。注意,在图4中,省略了音质转换单元103中的特征量混合单元103B和特征量混合单元101B的说明。

[0061] 在学习时,使用多个歌手的人声信号(可包括普通对话)来学习音质转换单元103。人声信号可以是多个歌手唱相同内容的并行数据片段,或者不一定是并行数据。在本示例中,它被视为更现实且难以学习的非平行数据。如图4所示,多个歌手的人声信号被存储在适当的数据库110中。

[0062] 对说话者特征量估计单元101A和编码器103A输入预定人声信号,作为歌声语音数据x的输入。说话者特征量估计单元101A根据所输入的歌声语音数据x估计说话者特征量。此外,编码器103A从所输入的歌声语音数据x中提取例如声音音高信息、音量信息和对话内容(歌词)作为特征量的实施例。这些特征量例如由以多维矢量表示的嵌入矢量来限定。由嵌入矢量限定的每个特征量被适当地称为如下:

[0063] 说话者嵌入;

[0064] e^{id}

[0065] 声音音高嵌入;

[0066] e^{pitch}

[0067] 音量嵌入;以及

[0068] e^{loud}

[0069] 内容嵌入;

[0070] e^{cont}。

[0071] 解码器103C执行以这些特征量作为输入来构造语音的处理。在学习时,解码器103C执行学习,使得解码器103C的输出重构歌声语音数据x的输入。例如,解码器103C执行学习以使在图4中示出的由损失函数计算器115计算的歌声语音数据x的输入与解码器103C

的输出之间的损失函数最小化。

[0072] 因为学习说话者特征量估计单元101A和编码器10AC,使得每个嵌入仅反映对应特征而不具有其他特征的信息,所以可以在推断时通过将一个嵌入替换为另一来仅转换对应特征。例如,当只有说话者嵌入时

[0073] e^{id}

[0074] 用另人的音质代替,可以在保持声音音高、音量和对话内容的同时转换音质(在不包括声音音高的狭义上的音质)。作为获得以这种方式分离特征的嵌入矢量的方法,存在从仅反映特定特征的特征量获得嵌入的方法和学习从数据(预定人声信号)仅提取特定特征的编码器的方法。

[0075] 作为前者,存在通过底座声音提取器提取底座声音f0并且获得的方法。

[0076] 声音音高嵌入;

[0077] $e^{pitch}=E^{pitch}(f_0)$,

[0078] 获得音量嵌入的方法

[0079] $e^{loud}=E^{loud}(p)$

[0080] 来自平均功率P,

[0081] 获得说话者嵌入的方法

[0082] $e^{id}=E^{id}(n)$

[0083] 来自说话者标签n,

[0084] 获得特征量的方法

[0085] V^{ASR}

[0086] 从语音识别中获得,

[0087] 获得内容嵌入的方法

[0088] $e^{cont}=E^{cont}(v^{ASR})$

[0089] 来自自动对话识别等。

[0090] 作为后一方法(学习从数据仅提取特定特征的编码器的方法),可考虑基于对手学习或量化导致的信息损失的技术。例如,使用对手学习来获得以下中的每一者:

[0091] 声音音高嵌入,

[0092] e^{pitch}

[0093] 音量嵌入,以及

[0094] e^{loud}

[0095] 说话者嵌入

[0096] e^{id}

[0097] 。此外,内容嵌入

[0098] e^{cont}

[0099] 其中,很难获取正确的标签可以通过学习使用数据来获得。

[0100] 作为具体实施例,由编码器103A执行的提取内容嵌入的学习的实施例

[0101] e^{cont}

[0102] 被描述。首先,将描述使用基于对手学习的技术的具体示例。

[0103] 编码器

[0104] $E^{cont}(x,\theta^{cont})$

[0105] 提取内容嵌入

[0106] e^{Cont}

[0107] 从输入的歌声语音数据x中可以通过添加学习到

[0108] 损失函数

[0109] L^{j}

[0110] 使用评论员的

[0111] C^{j}

[0112] 用于估计另一特征量

[0113] y^{j}

[0114] 从内容嵌入

[0115] e^{cont}

[0116] 到失函数

[0117] L^{rec}

[0118] 关于输入的重新配置。

[0119] 具体地,使用以下公式执行学习。

[0120] $$L_{ED}(\theta) = L_{rec}\left(x, D\left(E^{id}(n,\theta^{id}), E^{pitch}(f_0,\theta^{pitch}), E^{loud}(p,\theta^{loud}), E^{cont}(x,\theta^{cont}), \theta^{dec}\right)\right) - \sum_{j\neq i}\lambda_j L^j(C^j(E^{cont}(x,\theta^{cont}),\phi^j), y^j)$$

[0121] $$L_{cj}(\phi^j) = L^j(C^j(E^{cont}(x,\theta^{cont}),\phi^j), y^j)$$

[0122] 然而,在上述公式中,

[0123] L_{ED}

[0124] 表示用于学习编码器103A和解码器103C的损失函数。此外,

[0125] L_{cj}

[0126] 是评论员的损失函数

[0127] C^{j}

[0128] 以及

[0129] λ_j

[0130] 是权重参数。

[0131] θ^{id}

[0132] θ^{pitch}

[0133] θ^{loud}

[0134] θ^{cont}

[0135] θ^{dec}

[0136] 是编码器103A和解码器103C的参数,并且

[0137] φ^{j}

[0138] 是评论员的参数

[0139] C^{j}。

[0140] 接下来,将描述基于通过量化的信息损失的技术的特定实施例。

[0141] 当编码器的输出时

[0142] $E^{cont}(x,\theta^{cont})$

[0143] 提取内容嵌入

[0144] e^{Cont}

[0145] 从输入的歌声语音数据x被矢量量化并且信息被压缩,内容嵌入

[0146] e^{cont}

[0147] 可以引导仅保存未被包括在其他信息中的信息

[0148] $(e^{id},e^{pitch},e^{loud})$

[0149] 提供给解码器。

[0150] 可以通过最小化以下损失函数来执行学习。

[0151] $L(\theta)=L_{rec}(x,D(E^{id}(n,\theta^{id}),E^{pitch}(f_0,\theta^{pitch}),E^{loud}(p,\theta^{loud}),E^{cont}(x,\theta^{cont}),\theta^{dec}))+|sg(E(x)-V(E(x)))|^2+\beta|E(x)-sg(V(E(x))|^2$

[0152] 在此,sg()是不将神经网络的梯度信息发送到以下层的停止梯度算子,并且V()是矢量量化运算。

[0153] 关于用于重新配置的损失函数

[0154] L^{rec},

[0155] 根据解码器和编码器的类型,可想到各种形式。例如,下界(ELBO)的证据

[0156] $$L_{rec}=\mathbb{E}\left[\log\left(p(X|e^{id},e^{pitch},e^{loud},e^{cont})\right)\right]-D_{KL}\left[q\left(e^{id},e^{pitch},e^{loud},e^{cont}|X\right)||p(e^{id},e^{pitch},e^{loud},e^{cont})\right]$$

[0157] 可以在变分自动编码器(VAE)或向量量化VAE的情况下使用。在生成的对手网络的情况下,它可以表示为输入、输出情况错误和对手损失的加权和(以下公式)

[0158] L_{adv}。

[0159] $L_{rec}=||x-D(e^{id},e^{pitch},e^{loud},e^{cont})||^2+\lambda L_{adv}$

[0160] 上述学习不改变说话者特征量估计单元估计的说话者信息而进行。一旦学习,说话者信息就可改变。此外,未来信息可以在学习时使用。

[0161] 在上文中,已经给出了关于获得用于确定音质的说话者嵌入的方法的描述:

[0162] $e^{id}=E^{id}(n)$

[0163] 使用说话者标记n。然而,在该方法中,转换目的地歌手需要被预先包括在学习数据中,并且不能对任意歌手(未知说话者)执行音质转换。对此,将描述从语音信号获得嵌入说话者的方法。例如,可想到以下两种方法。

[0164] 第一方法是基于说话者的人声信号执行说话者嵌入估计的方法,该说话者嵌入估计用于估计预定说话者(例如,具有与作为转换目的地的歌手的歌声语音数据相似的特征的歌声语音数据的说话者)的说话者信息。说话者特征量估计单元F(),估计说话者嵌入;

[0165] $e_n^{id}=E^{id}(n)$

[0166] 使用说话者标签n从说话者n的歌声声音中学习

[0167] x_n

[0168] 被学习。F可以由神经网络等配置,并被学习以最小化到说话者嵌入的距离。作为距离,Lp范数

[0169] $\left\| e_n^{id} - F(x_n) \right\|_p$

[0170] 可以使用。

[0171] 第二方法是执行歌手识别模型学习以基于预定人声信号估计说话者的说话者信息的方法。

[0172] 提取说话者嵌入的说话者特征量估计单元G()

[0173] e_n^{id}

[0174] 来自歌声声音

[0175] x_n

[0176] 在音质转换单元103的学习之前被学习。通过使用具有歌手标签的多个歌手的歌声声音数据来最小化以下目标函数L,可以学习G。

[0177] L=-min(K(G(x_n),G(x_m))-K(G(x_n),G(x'_n))-1,0)

[0178] 在此,K(x,y)是x和y之间的余弦距离,

[0179] x_n, x'_n

[0180] 是歌手n的不同的歌声声音,以及

[0181] x_n

[0182] 是歌手歌声声音(m≠n)。

[0183] 说话者嵌入

[0184] e_n^{id}

[0185] 利用以此方式学习的G如下获得,并用于学习音质转换单元103。

[0186] $e_n^{id} = \frac{G(x_n)}{|G(x_n)|}$

[0187] 在上述任何方法中,优选的是,输入到说话者特征量估计单元G()的输入语音足够长,以便获得准确的说话者嵌入。这是因为不能从短声音充分地提取歌手的特征。另一方面,过长的输入具有必要的存储器变得巨大的缺点。在这点上,对于G(),可使用具有递归结构的递归神经网络,或者可使用利用多个短时间段获得的说话者嵌入的平均值等。

[0188] [操作实施例]

[0189] 音质转换由如上所述学习的音质转换单元103执行。将参考图5描述由智能电话100执行的音质转换处理。

[0190] 在图5中,人声信号VSB唱卡拉OK用户的歌声语音数据。此外,人声信号VSA是卡拉OK用户想要使音质更近的歌手的歌声语音数据,是通过声源分离获得的人声信号。

[0191] 人声信号VSA和人声信号VSB中的每一者被输入到音质转换单元103。编码器103A从人声信号VSA和人声信号VSB中提取诸如声音音高和音量的特征量。

[0192] 例如,指定要被替换的特征量的控制信号被输入到特征量混合单元103B。例如,在输入用于将从人声信号VSB提取的声音音高信息转换为从人声信号VSA提取的声音音高信息的控制信号的情况下,特征量混合单元101B使用从人声信号VSA提取的声音音高信息来替换从人声信号VSB提取的声音音高信息。由特征量混合单元101B混合的特征量被输入到

解码器103C。

[0193] 人声信号VSA和人声信号VSB被输入到说话者特征量估计单元101A。说话者特征量估计单元101A从各人声信号估计说话者信息。估计的说话者信息被提供给特征量混合单元101B。

[0194] 向特征量混合单元101B输入表示是否更换说话者特征量、更换时的说话者特征量的更换重量的控制信号。根据控制信号,特征量混合单元101B适当地替换说话者特征量。例如,在将从人声信号VSB获得的说话者特征量替换为从人声信号VSA获得的说话者特征量的情况下,将由说话者特征量限定的音质(狭义的音质)从卡拉OK用户的音质替换为与人声信号VSA对应的歌手的音质。将由特征量混合单元101B混合的说话者特征量提供给解码器103C。

[0195] 解码器103C基于从特征量混合单元101B提供的特征量和从特征量混合单元101B提供的说话者特征量来生成歌声语音数据。通过扬声器105再生产生的歌声语音数据。因此,再生其中卡拉OK用户的音质的一部分已经被歌手的音质的一部分(诸如专业人员)替换的歌声语音。

[0196] [与音质转换处理相关联地执行的处理]

[0197] 接下来,将描述与音质转换处理相关联地执行的处理。首先,将描述用于实现平滑音质转换的处理。在将自己的歌声声音改变为在卡拉OK等中使用的原始歌曲的歌手的歌声声音时,需要享受。例如,这可以通过替换嵌入歌手A的说话者为

[0198] e_A^{id}

[0199] 嵌入歌手B的说话者来实现

[0200] e_B^{id}

[0201] 为了在推断时(在执行音质转换处理时)将该歌手A(自身)的歌声语音改变成另一歌手(原始歌曲的歌手)的音质。

[0202] 然而,在卡拉OK等中,存在这样的需求:自己的歌声语音没有完全改变为歌手B的音质,而是歌手B稍微模仿。为了实现这一点,插值函数

[0203] $g(e_A^{id}, e_B^{id}, \alpha)$

[0204] 用于平稳地改变嵌入歌手A的说话者

[0205] e_A^{id}

[0206] 为嵌入歌手B的说话者

[0207] e_B^{id}

[0208] 被使用。在此,α是用于确定改变量的标量变量,并且还可以由用户确定。可以使用线性插值或球面线性插值作为插值函数。

[0209] 注意,除了

[0210] e_A^{id},

[0211] e^{pitch},

[0212] e^{loud},和

[0213] e^{cont}

[0214] 还可以类似地使用线性插值或球面线性插值来进行插入。例如,在卡拉OK用户的

音调的情况下

[0215] $f_0^{original}$

[0216] 期望更接近原始声源的歌手的音调

[0217] f_0^{target},

[0218] 可以如下执行线性插值。

[0219] $E^{pitch}\left(\beta f_0^{original}+(1-\beta)f_0^{target},\theta^{pitch}\right)$

[0220] 接下来,将描述实时处理。通过使用过去和未来信息的批处理来执行歌声语音转换的许多一般算法。另一方面,在卡拉OK等中使用的情况下,需要实时转换。此时,不能使用未来信息,因此,难以执行高质量转换。

[0221] 对此,本实施方式着眼于在卡拉OK中的多种情况下在原始声源中的歌声与音质转换中的用户歌声之间的对话(歌词)具有相同内容的并行数据的关系,并且即使在使用这种特征的实时处理中也能够实现高质量转换。在下文中,将描述用于实现这种转换的处理的具体实施例。

[0222] 首先,设置在音质转换单元103中的编码器103A和解码器103C都被设置为不使用未来信息的功能。在使用递归神经网络(RNN)或卷积神经网络(CNN)来配置编码器103A和解码器103C的情况下,这可以通过使用单向RNN或不使用未来信息的因果卷积来形成编码器103A和解码器103C来实现。

[0223] 因此,可以实时地执行该处理。但是,为了高精度地推定,需要基于足够长的输入来获得说话者嵌入,因此,在刚开始歌声后不久无法获得足够长度的输入,难以进行高品质变换。对此,在卡拉OK的音质转换中,可想到在推断时使用并行数据的关系,并且仅在短时间内使用输入来估计说话者嵌入。在此,短时间是包括一个或少量音素的歌声语音的持续时间,并且例如是大约几百毫秒到几秒。通常,不同说话者的相同音素之间的音质转换相对容易,并且能够以高质量执行转换。对此,当说话者嵌入依赖于音素时,即使使用短时间信息也可执行高质量转换。然而,假设学习时没有并行数据的情形,因此,需要在说话者嵌入是时间不变的约束下学习模型。也就是说,不可能简单地从短时间信息中获得说话者嵌入,换言之,不可能学习音素相关的说话者嵌入。

[0224] 对此,编码器103A和解码器103C是利用时不变的说话者嵌入来学习的,并且说话者特征量估计机

[0225] $F^{short}()$

[0226] 学习冻结这些模型的参数并使用这些模型估计异常说话者嵌入的方法。因此,将本处理时的说话者嵌入处理为异常特征量。

[0227] 学习目标函数

[0228] F^{short}

[0229] 可以表示为

[0230] $L(\psi)=L_{rec}(x,D(F^{short}(x,\psi),e^{pitch},e^{loud},e^{cont}))$。

[0231] 在此,应当注意,编码器103A和解码器103C的参数是固定的。

[0232] 感受野

[0233] F^{short}

[0234] 限于上述短时间,通过使上述目标函数最小化来获得。

[0235] 以这种方式学习的说话者特征量估计单元F是获得取决于由下式指定的对话内容(音素)的说话者嵌入的估计器:

[0236] e^{cont},

[0237] 并且仅仅基于短时间信息实现高质量转换。

[0238] 另一方面,若歌声持续一定时间,从充分长的输入声音能够获得说话者嵌入,则在使用执行参照图4等描述的学习的说话者特征量估计单元F的情况下,时间稳定性有时提高。

[0239] 对此,如图6所示,例如,说话者特征量估计单元101A包括使用预定时间或更长时间的长时间信息的说话者特征量估计单元(在下文中,适当地称为全局特征量估计单元121A)、使用短于预定时间的时间的短时间信息的说话者特征量估计单元(在下文中,适当地称为局部(音素)特征量估计单元121B)以及特征量结合单元121C。然后,使用全局特征量估计单元121A和局部特征量估计单元121B两者,可以获得说话者特征量。由特征量结合单元121C结合从两个估计单元获得的说话者特征量,并使用该说话者特征量来获得最终的说话者嵌入。可以使用加权线性结合、球上线性结合等用于结合,并且可以从持续时间、输入信号等获得结合权重参数。例如,说话者嵌入

[0240] e^{id}

[0241] 可以如下获得。

[0242] $e^{id}=\alpha(T,x)F^{short}(x^{short})+(1-\alpha(T,x))F(x)$

[0243] 在此,T是从转换开始起的输入长度。在此,α也可以仅取决于T如下获得。

[0244] $$\alpha(T)=(1-\alpha_{\infty})\frac{e^{-\frac{T}{T_0}}}{e}+\alpha_{\infty}$$

[0245] 可替代地,它可以使用神经网络(像α(x))从输入x获得,或可以使用T或x的任何信息获得。

[0246] 接下来,将描述处理歌声错误的处理。上述实时处理具有这样的前提,即,包括在推断时的原始歌曲中的歌声内容和用户的歌声内容彼此一致(假设并行数据)。另一方面,用户可能错误地唱了歌曲等,并且该前提不一定成立。在通过仅使用上述短时间输入的方法在大大不同的音素之间获得说话者嵌入的情况下,转换的质量可能大大劣化。

[0247] 在这方面,在进行本处理的情况下,如图7所示,相似性计算器103D设置在音质转换单元103中。相似性计算器103D计算内容嵌入的相似性

[0248] e^{cont}

[0249] 目标歌手与原歌手之间。相似度计算器103D的计算结果被提供给说话者特征量估计单元101A。

[0250] 说话者特征量估计单元101A根据相似度使说话者特征量估计时的全局特征量和局部特征量的结合系数(由每个说话者特征量估计单元估计的每个说话者特征量的权重)与其他特征量的混合的权重改变。即,在相似度低的情况下,由于对话内容不同,所以基于短时间信息的说话者特征量的结合的权重减小,相依性降低。换言之,主要使用全局特征量估计单元121A的处理结果。此外,在其他特征量的混合中,通过增加相对于原始说话者的特征量的权重来抑制过度转换,由此抑制音质的显著劣化。

[0251] 接下来,将描述用于使分离的声源鲁棒的机制。通常,用于学习歌声语音转换的数据优选地是干净的,而没有噪声。另一方面,在本发明中,目标说话者歌声的语音是通过声源分离获得的语音,且包括由该分离引起的噪声。因此,每个嵌入的估计精度由于噪声而劣化,并且转换后的语音的音质可能包括噪声。为了防止这种情况,将描述针对声源分离噪声构建鲁棒系统的方法。

[0252] 可以通过在编码器、解码器和说话者特征量估计单元的学习期间应用约束,使得从通过声源分离获得的语音提取的嵌入矢量和原始干净语音相同,来实现针对声源分离噪声的鲁棒性。具体地,当干净语音信号是x,伴奏信号是b,声源分离器是h()时,正则化项

[0253] $L_{reg}=||E(x)-E(h(x+b))||_p$

[0254] 添加到学习的目标函数中。

[0255] 在此,E是编码器或特征量提取器。关于损失函数的计算

[0256] L_{rec}

[0257] 与重构相关使得能够学习编码器103A,使得来自分离的语音的特征量提取结果与来自干净语音的特征量提取结果一致,同时通过仅使用干净语音来保持解码器103C的输出干净。

[0258] 优选的是,执行与上述音质转换处理相关联地执行的所有处理,但是一些处理可以被执行或者不必被执行。

[0259] <变形例>

[0260] 虽然上面已经描述了本公开的实施方式,但是本公开不限于上述实施方式,并且在不背离本公开的主旨的情况下可以做出各种修改。

[0261] 并非实施方式中描述的所有处理都需要由智能电话100执行。一些处理可以由与智能电话100不同的装置(例如,服务器)执行。例如,如图8所示,声源分离处理和说话者特征量估计处理可以由服务器执行,音质转换处理和再生处理可以由智能电话执行。此外,如图9所示,可以由服务器执行声源分离处理,并且可以由智能电话执行音质转换处理、再生处理和说话者特征量估计处理。经由网络在服务器与智能电话之间发送和接收处理结果。

[0262] 此外,本公开还可以通过诸如装置、方法、程序或者系统的任何模式来实现。例如,通过使得能够下载执行在上述实施方式中描述的功能的程序和通过不具有在实施方式中描述的功能的装置下载并安装程序,可以在装置中执行在实施方式中描述的控制。本公开也可以通过分配这样的程序的服务器来实现。此外,在各个实施方式和变形例中描述的项可以适当地结合。此外,本公开的内容并不被解释为受本说明书中示例的效果的限制。

[0263] 本公开可以具有以下配置。

[0264] (1)

[0265] 一种信息处理装置,包括:

[0266] 音质转换单元,从混合声音信号执行人声信号和伴奏信号的声源分离,并使用所述声源分离的结果执行音质转换。

[0267] (2)

[0268] 根据(1)所述的信息处理装置,其中

[0269] 第一人声信号通过所述声源分离从所述混合声音信号中分离,

[0270] 收集的第二人声信号被输入至所述音质转换单元,并且

[0271] 所述音质转换单元让所述第一人声信号和所述第二人声信号中的一个人声信号更接近另一人声信号。

[0272] (3)

[0273] 根据(2)所述的信息处理装置,其中

[0274] 能够设置让所述一个人声信号更接近所述另一人声信号的改变量。

[0275] (4)

[0276] 根据(2)所述的信息处理装置,进一步包括:

[0277] 说话者特征量估计单元,估计与说话者相关的特征量,

[0278] 其中,所述音质转换单元包括编码器和解码器。

[0279] (5)

[0280] 根据(4)所述的信息处理装置,其中

[0281] 与所述说话者相关的特征量是与不随时间改变的特征对应的特征量,

[0282] 所述编码器从输入的人声信号中提取与随着时间改变的特征对应的特征量,以及

[0283] 所述解码器基于由所述说话者特征量估计单元估计的特征量和由所述编码器提取的特征量,生成人声信号。

[0284] (6)

[0285] 根据(5)所述的信息处理装置,其中

[0286] 所述与不随时间改变的特征对应的特征量是说话者信息,并且

[0287] 所述与随着时间改变的特征对应的特征量包括声音音高信息、音量信息和对话信息中的至少一者。

[0288] (7)

[0289] 根据(6)所述的信息处理装置,其中

[0290] 所述特征量由嵌入矢量限定。

[0291] (8)

[0292] 根据(7)所述的信息处理装置,其中

[0293] 所述编码器通过使用学习模型,提取对应于随时间改变的特征的所述特征量的嵌入矢量,所述学习模型通过执行针对从仅反映特定特征的特征量获得嵌入矢量的学习或者针对从声音信号中仅提取特定特征的学习而获得。

[0294] (9)

[0295] 根据(6)至(8)中任一项所述的信息处理装置,其中

[0296] 所述说话者特征量估计单元通过使用学习模型,估计所述说话者的特征量,所述学习模型通过基于预定说话者的人声信号针对估计所述说话者的所述说话者信息的学习而获得。

[0297] (10)

[0298] 根据(6)至(8)中任一项所述的信息处理装置,其中

[0299] 所述说话者特征量估计单元通过使用学习模型来估计所述说话者的特征量,所述学习模型通过基于预定人声信号针对估计所述说话者的说话者信息的学习而获得。

[0300] (11)

[0301] 根据(4)至(10)中任一项所述的信息处理装置,其中

[0302] 所述说话者特征量估计单元包括第一说话者特征量估计单元和第二说话者特征量估计单元，

[0303] 所述信息处理装置进一步包括特征量结合单元，所述特征量结合单元对由所述第一说话者特征量估计单元估计的与所述说话者相关的特征量和由所述第二说话者特征量估计单元估计的与所述说话者相关的特征量进行结合。

[0304] (12)

[0305] 根据(11)所述的信息处理装置，其中

[0306] 所述第一说话者特征量估计单元基于预定时间以上的人声信号估计与所述说话者相关的特征量，并且所述第二说话者特征量估计单元基于比所述预定时间短的时间的人声信号估计与所述说话者相关的特征量。

[0307] (13)

[0308] 根据(11)所述的信息处理装置，其中

[0309] 所述特征量结合单元中的结合系数根据所述第一人声信号和所述第二人声信号之间的相似度而改变。

[0310] (14)

[0311] 根据(13)所述的信息处理装置，其中

[0312] 所述结合系数是针对由所述第一说话者特征量估计单元估计的与所述说话者相关的特征量和由所述第二说话者特征量估计单元估计的与所述说话者相关的特征量中的每一者的权重。

[0313] (15)

[0314] 一种信息处理方法，包括：

[0315] 通过音质转换单元从混合声音信号执行人声信号和伴奏信号的声源分离，并使用所述声源分离的结果执行音质转换。

[0316] (16)

[0317] 一种用于使计算机执行信息处理方法的程序，所述信息处理方法包括：

[0318] 通过音质转换单元从混合声音信号执行人声信号和伴奏信号的声源分离，并使用所述声源分离的结果执行音质转换。

[0319] 参考符号列表

[0320] 100 智能电话

[0321] 102 声源分离单元

[0322] 101A 说话者特征量估计单元

[0323] 101B 说话者特征量混合单元

[0324] 103 音质转换单元

[0325] 103A 编码器

[0326] 103C 解码器

[0327] 103D 相似性计算器

[0328] 121A 全局特征量估计单元

[0329] 121B 局部特征量估计单元。

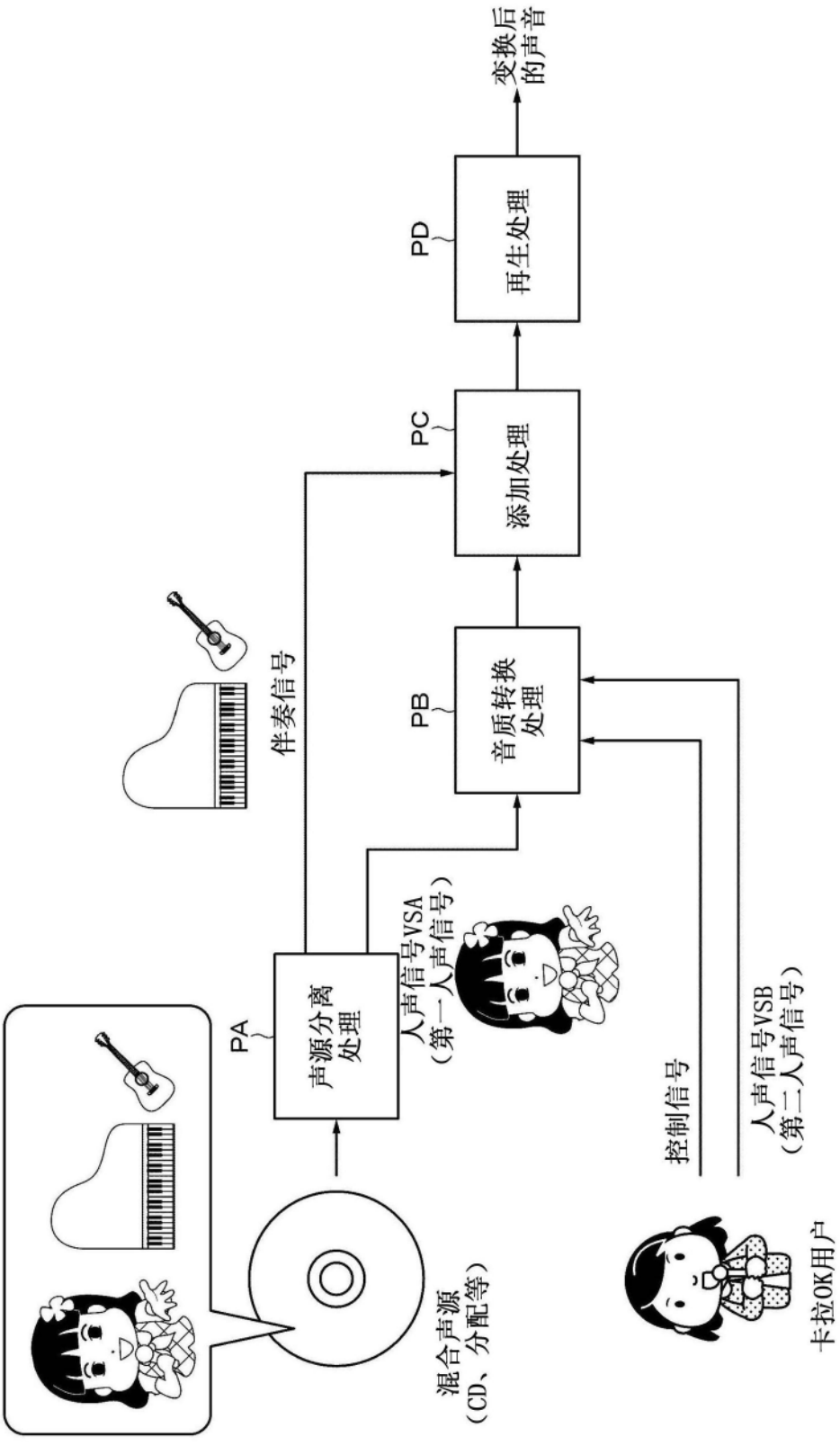

变换后的声音
PD
再生处理
PC
添加处理
PB
音质转换处理
伴奏信号
PA
声源分离处理
人声信号VSA（第一人声信号）
控制信号
人声信号VSB（第二人声信号）
混合声源（CD、分配等）
卡拉OK用户

图1

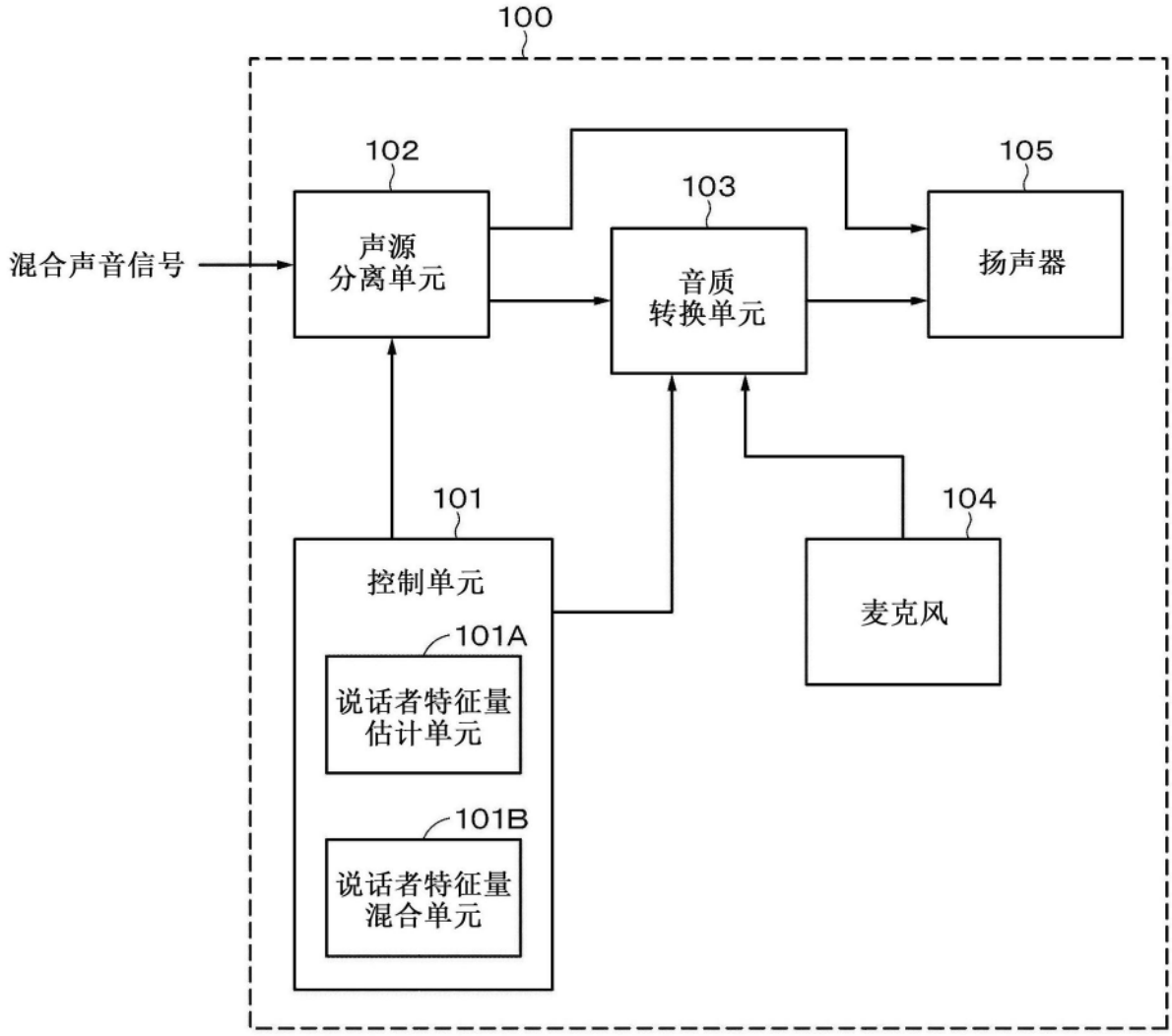

100
102
声源
分离单元
混合声音信号
103
音质
转换单元
105
扬声器
101
控制单元
101A
说话者特征量
估计单元
101B
说话者特征量
混合单元
104
麦克风

图2

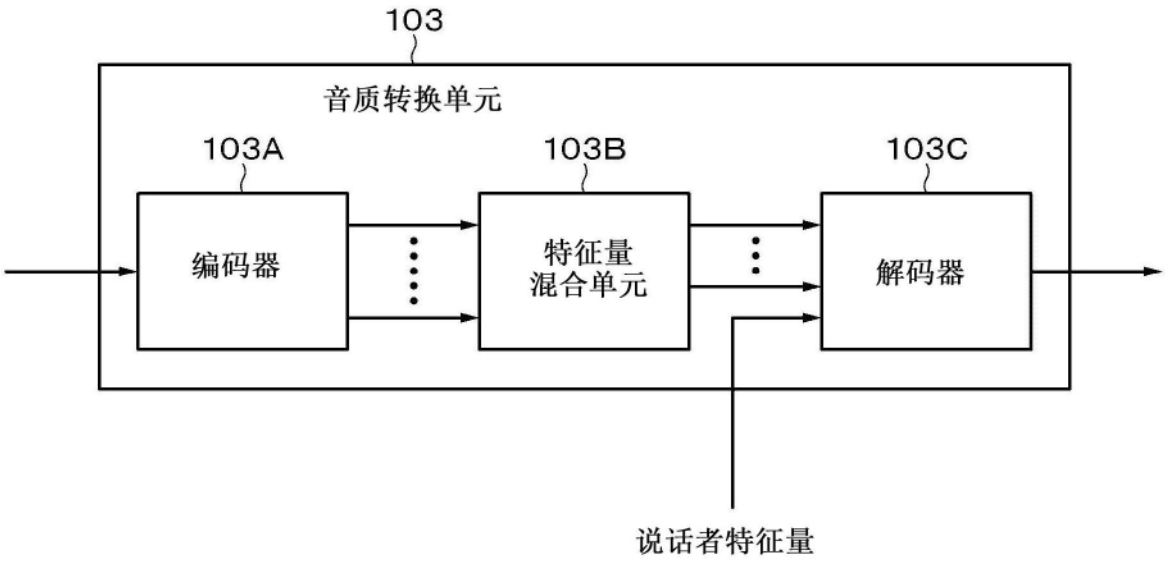

103
音质转换单元
103A
103B
103C
编码器
特征量
混合单元
解码器
说话者特征量

图3

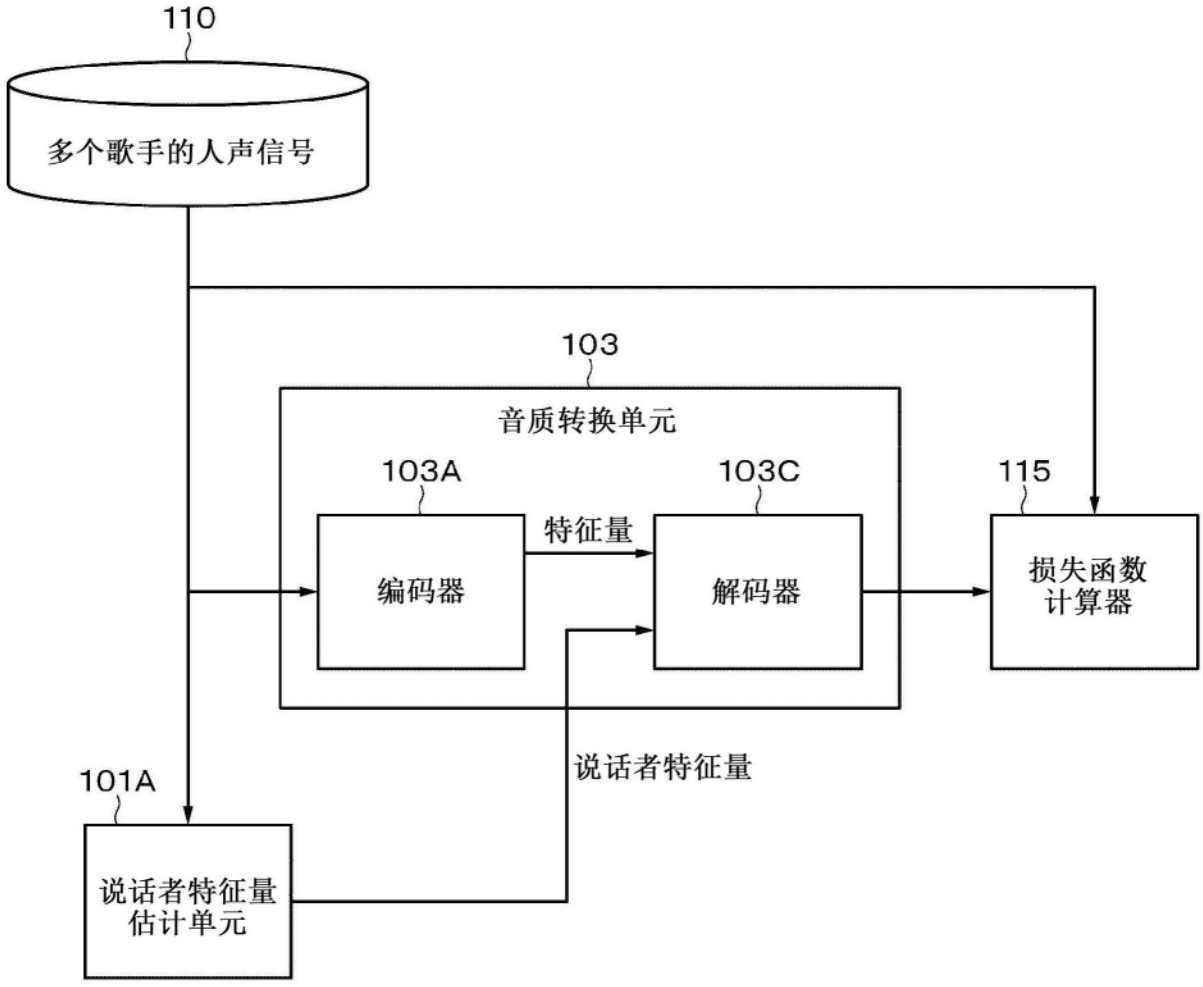

110
多个歌手的人声信号
103
音质转换单元
103A
103C
115
特征量
编码器
解码器
损失函数
计算器
说话者特征量
101A
说话者特征量
估计单元

图4

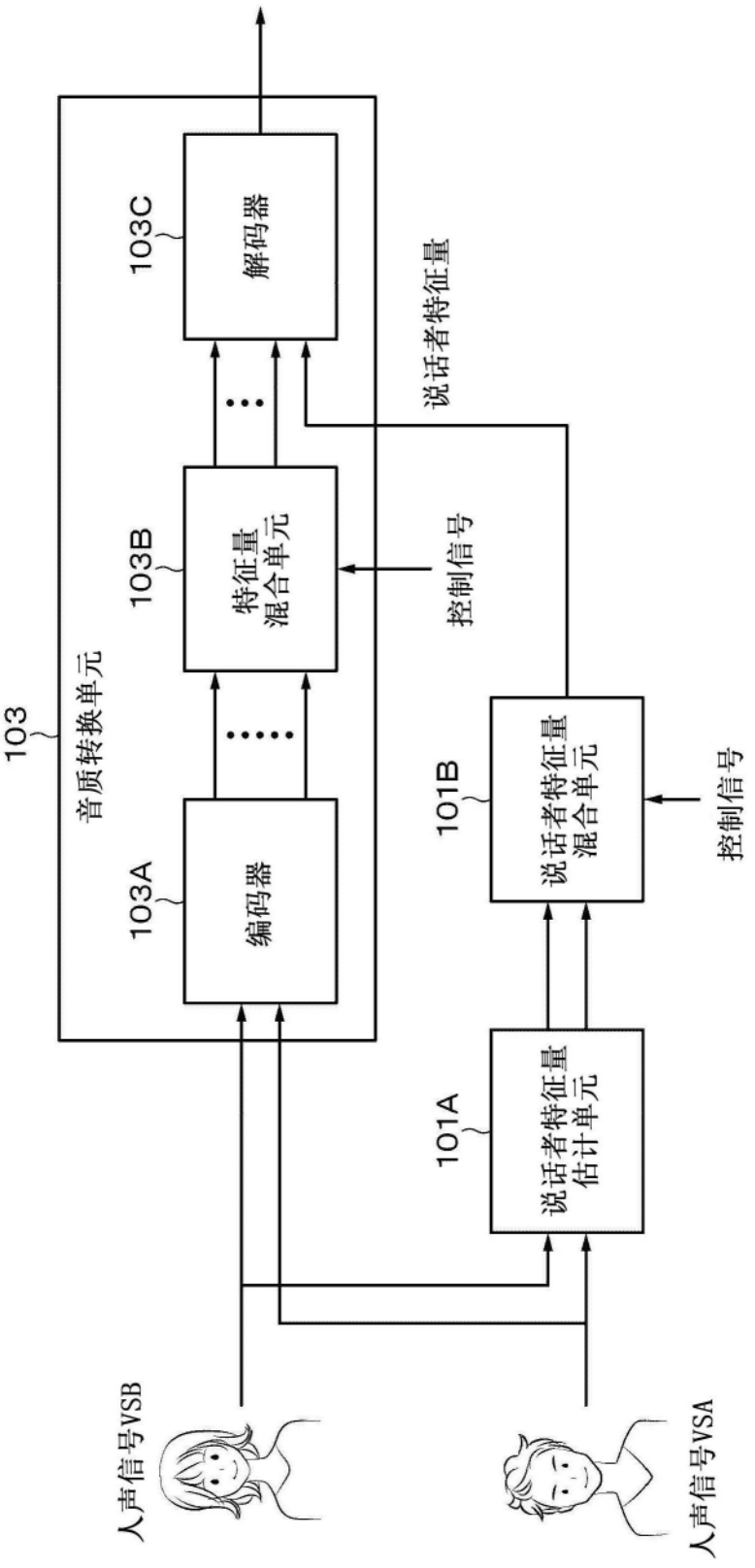

103
音质转换单元
103A
编码器
103B
特征量
混合单元
103C
解码器
101A
说话者特征量
估计单元
101B
说话者特征量
混合单元
说话者特征量
控制信号
控制信号
人声信号VSB
人声信号VSA

图5

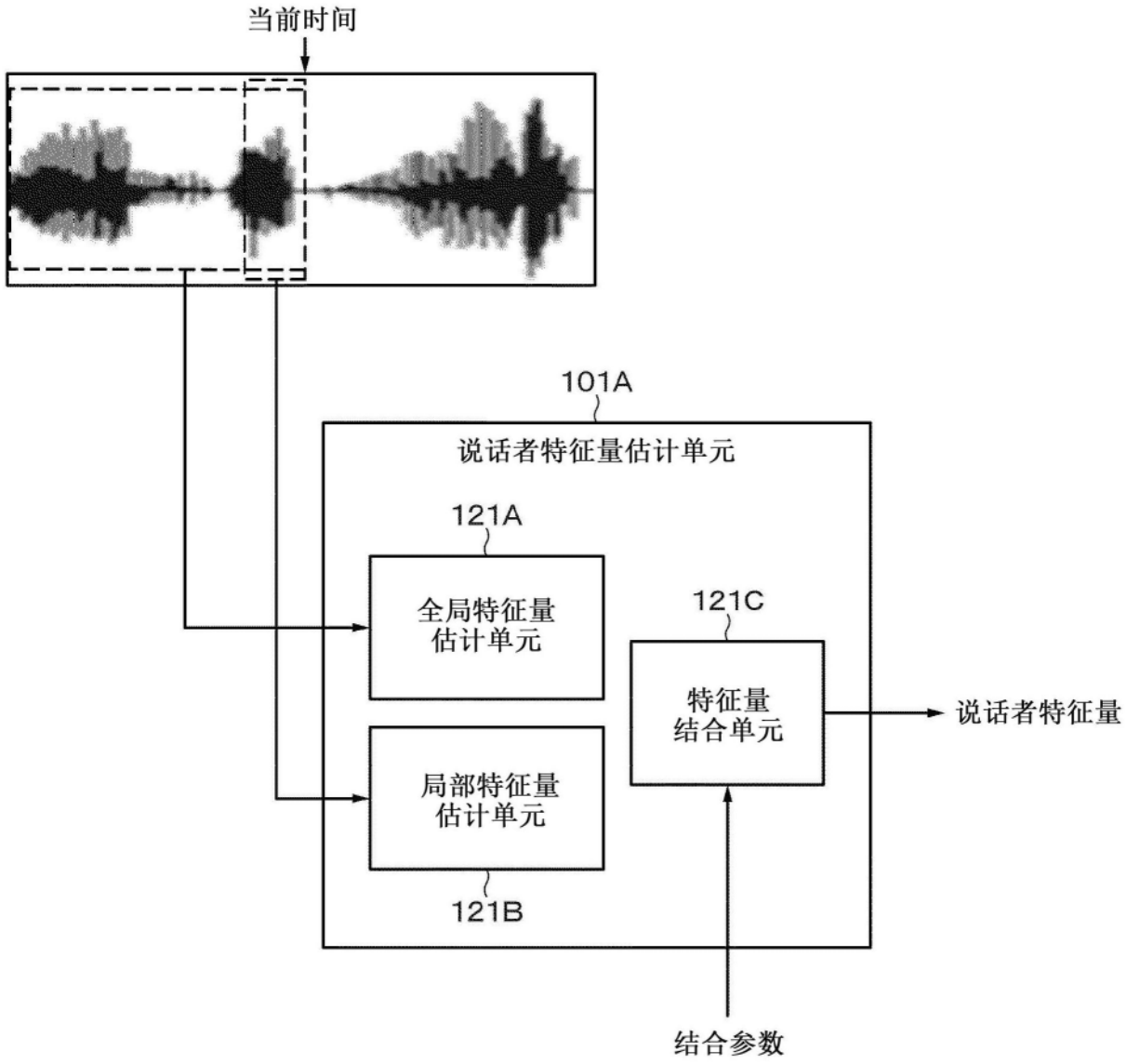

当前时间
101A
说话者特征量估计单元
121A
全局特征量估计单元
121C
特征量结合单元
说话者特征量
局部特征量估计单元
121B
结合参数

图6

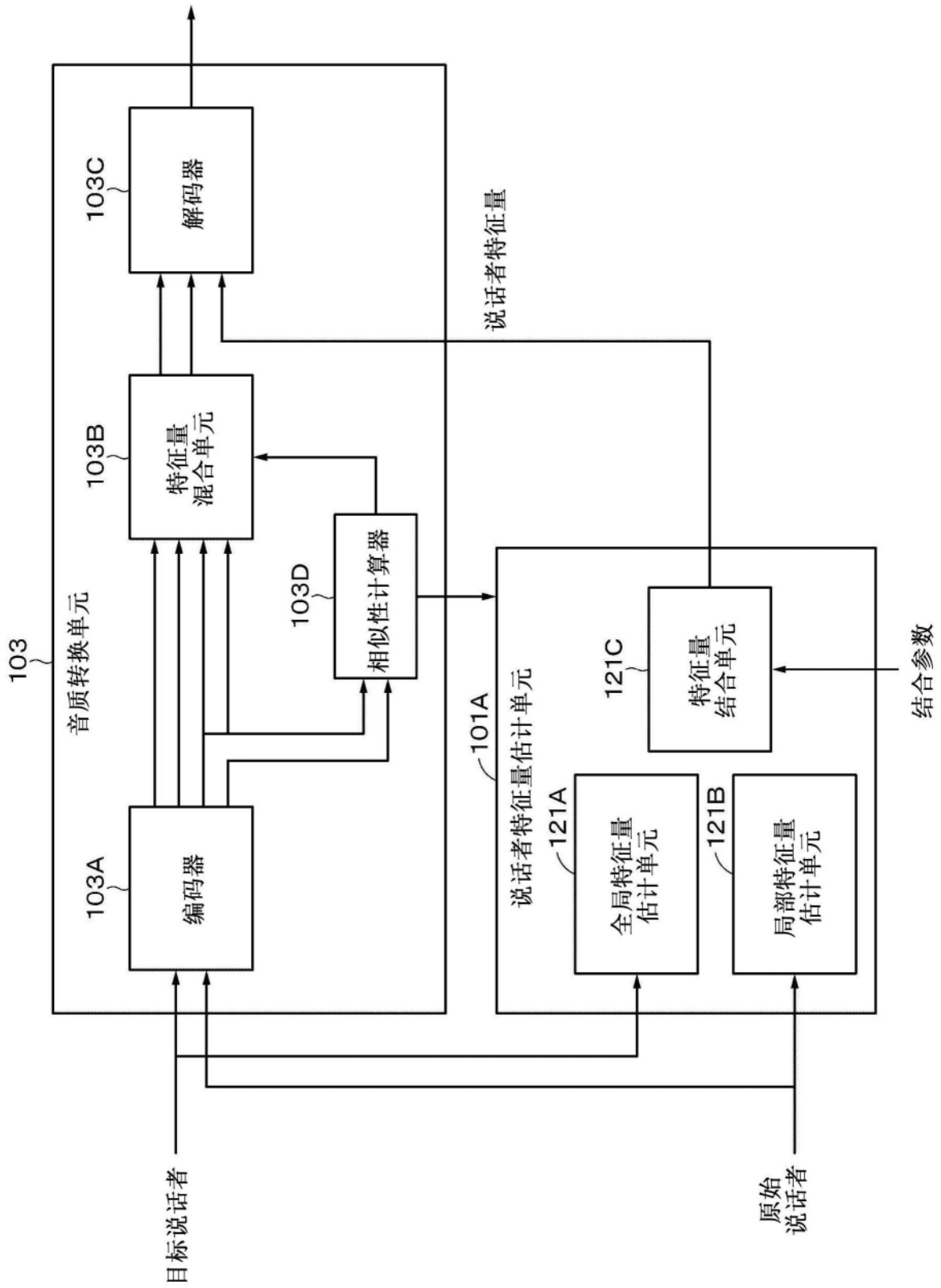

103
音质转换单元
103A
编码器
103B
特征量
混合单元
103C
解码器
103D
相似性计算器
说话者特征量
101A
说话者特征量估计单元
121A
全局特征量
估计单元
121B
局部特征量
估计单元
121C
特征量
结合单元
结合参数
目标说话者
原始
说话者

图7

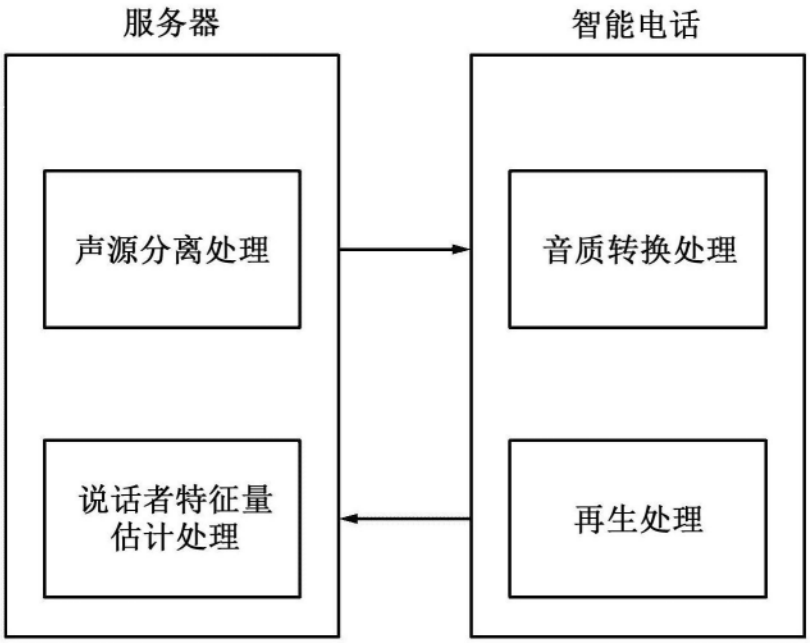

服务器
智能电话
声源分离处理
音质转换处理
说话者特征量
估计处理
再生处理

图8

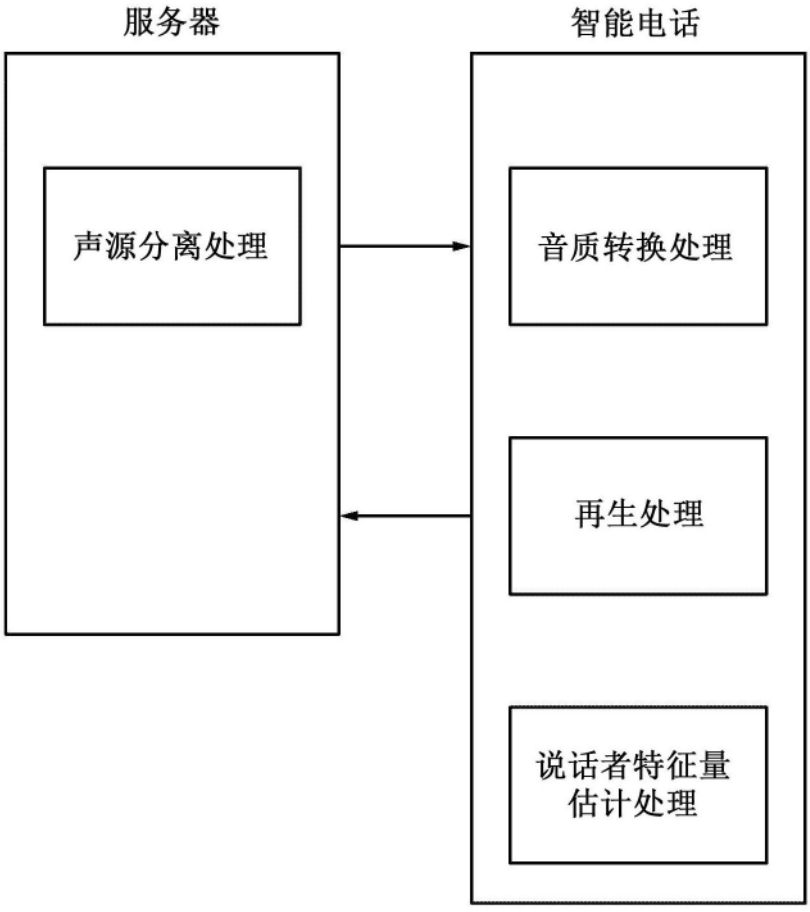

服务器
智能电话
声源分离处理
音质转换处理
再生处理
说话者特征量
估计处理

图9