(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2023年1月5日(05.01.2023)

WIPO | PCT

(10) 国際公開番号
WO 2023/276234 A1

(51) 国際特許分類:
G10L 13/00 (2006.01) *G10L 21/0272* (2013.01)
G10L 21/007 (2013.01)

(21) 国際出願番号： PCT/JP2022/005001

(22) 国際出願日： 2022年2月9日(09.02.2022)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(30) 優先権データ：
特願 2021-107651 2021年6月29日(29.06.2021) JP

(71) 出願人：ソニーグループ株式会社(**SONY GROUP CORPORATION**) [JP/JP]；〒1080075 東京都港区港南１丁目７番１号 Tokyo (JP).

(72) 発明者：高 橋 直 也 (**TAKAHASHI, Naoya**)；〒1080075 東京都港区港南１丁目７番１号 ソニーグループ株式会社内 Tokyo (JP).

(74) 代理人：弁理士法人杉浦特許事務所，外 (**SUGIURA PATENT OFFICE** et al.)；〒1710022 東京都豊島区南池袋１－１－１１ カドラービル４０２ Tokyo (JP).

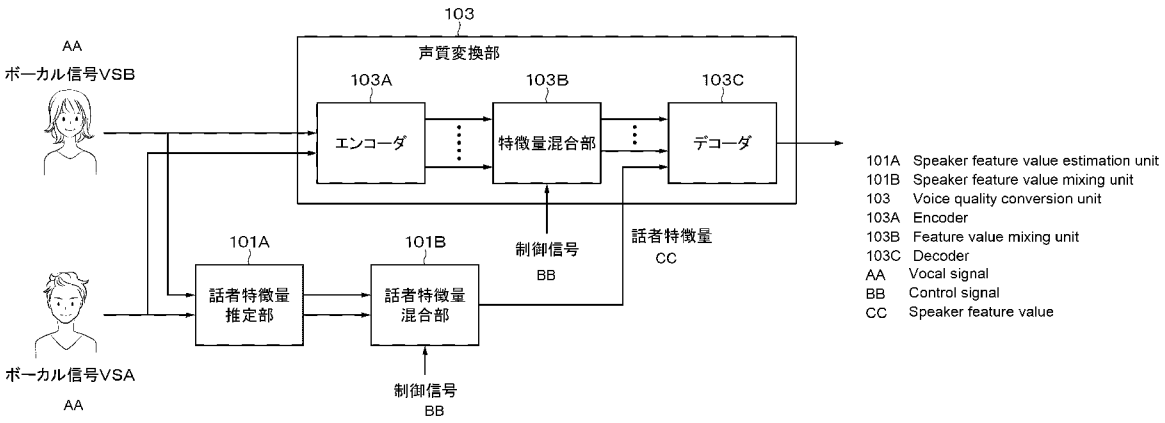
(81) 指定国(表示のない限り、全ての種類の国内保護が可能)：AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, 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(表示のない限り、全ての種類の広域保護が可能)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, 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).

(54) **Title:** INFORMATION PROCESSING DEVICE, INFORMATION PROCESSING METHOD, AND PROGRAM

(54) 発明の名称： 情報処理装置、情報処理方法およびプログラム

図5



(57) **Abstract:** In order to perform effective voice quality conversion processing, for example, the present invention provides an information processing device having a voice quality conversion unit for performing sound source separation of a vocal signal and an accompaniment signal from a mixed sound signal and performing voice quality conversion using the result of the sound source separation.

(57) 要約：例えば、効果的な声質変換処理を行う。 混合音信号からボーカル信号と伴奏信号とを音源分離し、当該音源分離の結果を用いて声質変換を行う声質変換部を有する情報処理装置である。

［続葉有］

WO 2023/276234 A1

添付公開書類：

一　国際調査報告（条約第21条(3)）

明　細　書

発明の名称 ： 情報処理装置、情報処理方法およびプログラム

技術分野

[0001]　本開示は、情報処理装置、情報処理方法およびプログラムに関する。

背景技術

[0002]　自身の発話（歌唱を含む）の声質を他社の声質に変換する声質変換技術に関する提案がなされている。声質とは、話者により生成された人間の音声であって、かつ複数個の音声単位（例えば音素）にわたって聴者により知覚された音声の属性のことをいい、より具体的には、音高と音色が同じの発話であっても聴者により違うと近くされる要素をいう。下記特許文献１には、一般的な発話音声を、発話内容を保ったまま別の話者の声質に変換する声質変換技術が記載されている。

先行技術文献

特許文献

[0003] 特許文献1：特開２０１８－００５０４８号公報

発明の概要

発明が解決しようとする課題

[0004]　この分野では、適切な声質変換処理が行われることが望まれる。

[0005]　本開示は、適切な声質変換処理が行われる情報処理装置、情報処理方法およびプログラムを提供することを目的の一つとする。

課題を解決するための手段

[0006]　本開示は、例えば、

混合音信号からボーカル信号と伴奏信号とを音源分離し、当該音源分離の結果を用いて声質変換を行う声質変換部を有する

情報処理装置である。

[0007]　本開示は、例えば、

声質変換部が、混合音信号からボーカル信号と伴奏信号とを音源分離し、

当該音源分離の結果を用いて声質変換を行う

情報処理方法である。

[0008] 本開示は、例えば、

声質変換部が、混合音信号からボーカル信号と伴奏信号とを音源分離し、当該音源分離の結果を用いて声質変換を行う

情報処理方法をコンピュータに実行させるプログラムである。

図面の簡単な説明

[0009] [図1]図１は、一実施形態の概要を説明するための図である。

[図2]図２は、一実施形態に係るスマートホンの構成例を示すブロック図である。

[図3]図３は、一実施形態に係る声質変換部の構成例を示すブロック図である。

[図4]図４は、一実施形態に係る声質変換部で行われる学習の例を説明するための図である。

[図5]図５は、一実施形態に係るスマートホンの動作を説明する際に参照される図である。

[図6]図６は、一実施形態で行われる声質変換処理に付随して行われる処理の一例を説明するための図である。

[図7]図７は、一実施形態で行われる声質変換処理に付随して行われる処理の他の例を説明するための図である。

[図8]図８は、変形例を説明するための図である。

[図9]図９は、変形例を説明するための図である。

発明を実施するための形態

[0010] 以下、本開示の実施形態等について図面を参照しながら説明する。なお、説明は以下の順序で行う。

＜本開示の背景＞

＜一実施形態＞

＜変形例＞

以下に説明する実施形態等は本開示の好適な具体例であり、本開示の内容がこれらの実施形態等に限定されるものではない。

[0011] ＜本開示の背景＞

始めに、本開示の理解を容易とするために、本開示の背景について説明する。近年カラオケにおいて、あらかじめ作成されたＭＩＤＩ(Musical Instrument Digital Interface)音源や録音音源を伴奏として用いるのではなく、ボーカル音声入りの原音源をボーカル信号と伴奏信号とに音源分離し、分離された伴奏信号を用いることが増えている。

[0012] このような音源分離技術の進化により、伴奏音源作成のコスト削減や、原曲そのままのサウンドでカラオケを楽しめるといったメリットが得られる。一方で、カラオケにおいては残響、歌声のピッチを変化させて加えるコーラス、声質を不特定な声質に変えるボイスチェンジャーなどのエフェクトが一般的に使われているが、特定の人物の歌声に変化させることは未だに困難である。そのため、例えば、「自分の声を少しだけ原曲のアーティストの声に近づける」といった、声質を特定の歌手の声質に滑らかに変換することが困難である。

[0013] 上述した特許文献１に記載の技術のように、一般的な発話音声を、発話内容を保ったまま別の話者の声質に変換する声質変換技術は提案されているが、一般に歌声は普通の発話に比べ音高や声質、様々な音楽的表現方法（ビブラートなど）のバリエーションが多く、歌声の変換は難しい。そのため、現状ではロボット風・アニメ風に変換、性別変換などの不特定な声質への変換や、あらかじめ十分な量のクリーンな音声が得られる特定話者の声質変換しか行なえず、あらかじめ十分な量のクリーン音声が得られない話者への変換が困難である。十分な量のクリーン音声が得るのには一般的に多くの時間やコストがかかり、例えば有名歌手の声に声質変換を行うことは実質的に非常に困難である。

[0014] また、カラオケ用途においてはリアルタイムで声質変換を行うことが必要で、未来情報が用いることが出来ないため、高品質な変換はさらに困難であ

る。加えて、音源分離により分離された音源は音源分離時に発生するノイズを含みうるため、そのような分離音声を参照して変換された音声はノイズを多く含みやすく、さらに高品質な変換は困難である。以上の点を踏まえつつ、本開示の一実施形態について以下、詳細に説明する。

[0015] ＜一実施形態＞

［一実施形態の概要］

始めに図１を参照しつつ、一実施形態の概要について説明する。図１に示す混合音源に対して、音源分離処理ＰＡが行われる。混合音源は、ＣＤ(Compact Disc)等の記録媒体やネットワークを介した配信によって提供され得る。混合音源には、例えば、アーティストのボーカル信号（第１のボーカル信号の一例であり、以下、ボーカル信号ＶＳＡとも適宜、称する）が含まれる。また、混合音源には、ボーカル信号ＶＳＡ以外の信号（楽器音等であり、以下、伴奏信号とも適宜、称する）が含まれる。

[0016] 一方、カラオケのユーザーの歌唱音声がマイクロホン等によって収音される。ユーザーの歌唱音声（第２のボーカル信号の一例）をボーカル信号ＶＳＢとも適宜、称する。

[0017] ボーカル信号ＶＳＡおよびボーカル信号ＶＳＢに対して、声質変換処理ＰＢが行われる。声質変換処理ＰＢでは、ボーカル信号ＶＳＡおよびボーカル信号ＶＳＢの何れか一方のボーカル信号を他方のボーカル信号に近づける（似せる）処理が行われる。この際、所定の制御信号に応じて、何れか一方のボーカル信号を他方のボーカル信号に近づける変化量を設定することができる。例えば、カラオケのユーザーのボーカル信号ＶＳＢを、アーティストのボーカル信号ＶＳＡに近づける声質変換処理が行われる。そして、声質変換処理が行われたボーカル信号ＶＳＢと伴奏信号とを加算する加算処理ＰＣが行われ、加算処理ＰＣが行われた信号に対して再生処理ＰＤが行われる。これにより、アーティストのボーカル信号に近づける声質変換処理がなされたユーザーの歌声が再生される。

[0018] ［情報処理装置の構成例］

（全体の構成例）

図２は、一実施形態に係る情報処理装置の構成例を示すブロック図である。本実施形態に係る情報処理装置としては、例えば、スマートホン（スマートホン１００）が挙げられる。スマートホン１００を用いて、ユーザーは、声質変換が可能なカラオケを手軽に行うことができる。なお、本実施形態では、カラオケ、即ち、歌唱を例にして説明するが、本開示は歌唱に限らず、会話等の発話に対する声質変換処理に対しても適用可能である。また、本開示に係る情報処理装置は、スマートホンに限らず、スマートウォッチ等の携帯型の電子機器や、パーソナルコンピュータや据え置き型のカラオケ装置等に対しても適用可能である。

[0019] スマートホン１００は、例えば、制御部１０１、音源分離部１０２、声質変換部１０３、マイクロホン１０４、および、スピーカー１０５を有している。

[0020] 制御部１０１は、スマートホン１００全体を統括的に制御する。制御部１０１は、例えば、ＣＰＵ（Central Processing Unit）に構成されており、プログラムが格納されるＲＯＭ（Read Only Memory）やワークメモリとして使用されるＲＡＭ（Random Access Memory）等を有している（なお、これらのメモリの図示は省略している。）。

[0021] 制御部１０１は、機能ブロックとして話者特徴量推定部１０１Ａを有している。話者特徴量推定部１０１Ａは、歌唱の進行に伴い時間的に変化しない特徴に対応する特徴量、具体的には、話者に関する特徴量（以下、話者特徴量と適宜、称する）を推定する。

[0022] また、制御部１０１は、機能ブロックとして特徴量混合部１０１Ｂを有している。特徴量混合部１０１Ｂは、例えば、２以上の話者特徴量を適宜な重みで混合する。

[0023] 音源分離部１０２は、入力される混合音信号をボーカル信号と伴奏信号とに分離する（音源分離処理）。音源分離されたボーカル信号が声質変換部１０３に供給される。また、音源分離された伴奏信号がスピーカー１０５に供

給される。

[0024] 声質変換部１０３は、マイクロホン１０４により収音されたユーザーの歌声に対応するボーカル信号の声質を、音源分離部１０２により音源分離されたボーカル信号に近づけるように声質変換処理を行う。なお、声質変換部１０３で行われる処理の詳細については後述する。なお、本実施形態における声質とは、話者特徴量の他に、音高、音量等の特徴量を含む。

[0025] マイクロホン１０４は、例えば、スマートホン１００のユーザーの歌唱や発話（本例では歌唱）を収音する。収音された歌唱に対応するボーカル信号が、声質変換部１０３に供給される。

[0026] 音源分離部１０２から供給される伴奏信号と、声質変換部１０３から出力されるボーカル信号とが、不図示の加算部により加算される。加算された信号がスピーカー１０５から再生される。

[0027] なお、スマートホン１００が、図２に図示した構成以外の構成（例えば、タッチパネルとして構成されるディスプレイやボタン）を有していてもよい。

[0028] （声質変換部の構成例）

図３は、声質変換部１０３の構成例を示すブロック図である。声質変換部１０３は、エンコーダ１０３Ａ、特徴量混合部１０３Ｂ、および、デコーダ１０３Ｃを有している。エンコーダ１０３Ａは、所定の学習により得られる学習モデルを用いて、ボーカル信号から特徴量を抽出する。エンコーダ１０３Ａにより抽出される特徴量は、例えば、歌唱の進行に伴って時間的に変化する特徴量であり、具体的には、音高情報、音量情報、発話（歌詞）情報の少なくとも一つを含む。

[0029] 特徴量混合部１０３Ｂは、エンコーダ１０３Ａにより抽出された特徴量を混合する。特徴量混合部１０３Ｂにより混合された特徴量がデコーダ１０３Ｃに供給される。

[0030] デコーダ１０３Ｃは、特徴量混合部１０３Ｂから供給される特徴量および話者特徴量に基づいて、ボーカル信号を生成する。

[0031]（声質変換部で行われる学習について）

次に、図４を参照しつつ、声質変換部１０３で行われる学習方法の一例について説明する。なお、図４では、声質変換部１０３における特徴量混合部１０３Ｂ、および、特徴量混合部１０１Ｂに関する図示は省略している。

[0032] 学習時、声質変換部１０３は、複数歌手のボーカル信号（通常発話を含んでもよい）を用いて学習される。ボーカル信号は、複数歌手が同内容を歌うパラレルデータであってもよいし、パラレルデータでなくてもよい。本例では、より現実的かつ学習が困難な非パラレルデータとして扱う。図４に示すように、複数歌手のボーカル信号は、適宜なデータベース１１０に記憶されている。

[0033] 所定のボーカル信号は、入力歌声データｘとして、上述した話者特徴量推定部１０１Ａ及びエンコーダ１０３Ａに入力される。話者特徴量推定部１０１Ａは、入力歌声データｘから話者特徴量を推定する。また、エンコーダ１０３Ａは、入力歌声データｘから、特徴量の一例として、例えば、音高情報、音量情報、発話内容（歌詞）を抽出する。これらの特徴量は、例えば、多次元のベクトルで表されるエンベディングベクトル（埋め込みベクトル）により規定される。エンベディングベクトルで規定された各特徴量を、それぞれ、

話者エンベディング

$$e^{id}$$

音高エンベディング

$$e^{pitch}$$

音量エンベディング

$$e^{loud}$$

コンテンツエンベディング

$$e^{cont}$$

と適宜、称する。

[0034] デコーダ１０３Ｃは、これらの特徴量を入力とし、音声を構築する処理を行う。学習時には、デコーダ１０３Ｃの出力が入力歌声データｘを再構築するように、デコーダ１０３Ｃは学習を行う。例えば、図４に示す損失関数算出部１１５により算出される入力歌声データｘとデコーダ１０３Ｃの出力との間の損失関数を最小化するように、デコーダ１０３Ｃは学習を行う。

[0035] 各エンベディングが対応する特徴のみを反映し、他の特徴の情報を持たないように話者特徴量推定部１０１Ａやエンコーダ１０ＡＣを学習することで、推論時に一部のエンベディングを他のものに置き換えることで、対応する特徴のみを変換することができる。例えば、

話者エンベディング

$$e^{id}$$

のみを他者のものに置き換えることで、音高、音量、発話内容を保ったまま声質（音高等を含まない狭義の声質）を変換することができる。このように、特徴を分離するようなエンベディングベクトルを得る方法として特定の特徴のみを反映した特徴量からエンベディングを得る方法や、データ（所定のボーカル信号）から特定の特徴のみを抽出するエンコーダを学習する方法がある。

[0036] 前者として基音抽出機により基音ｆ０を抽出し、

音高エンベディング

$$e^{pitch} = E^{pitch}(f_0)$$

を得る、

平均パワーｐから音量エンベディング

$$e^{loud} = E^{loud}(p)$$

を得る、

話者ラベルｎから話者エンベディング

$$e^{id} = E^{id}(n)$$

を得る、

音声認識から得られる特徴量

v^{ASR}

(Automatic Speech Recognition) からコンテンツエンベディング

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

を得るなどの方法がある。

[0037] 後者の方法（データから特定の特徴のみを抽出するエンコーダを学習する方法）として敵対的学習や量子化による情報損失による手法が考えられる。例えば、

音高エンベディング

e^{pitch}

音量エンベディング

e^{loud}

話者エンベディング

e^{id}

のそれぞれについては、敵対的学習により得られる。また、正確なラベルの取得が困難なコンテンツエンベディング

e^{cont}

についてデータから学習することで得られる。

[0038] 具体例として、コンテンツエンベディング

e^{cont}

を抽出するエンコーダ１０３Ａで行われる学習の例について説明する。始めに敵対的学習による手法を用いた具体例について説明する。

[0039] 入力歌声データｘから、

コンテンツエンベディング

e^{cont}

を抽出するエンコーダ

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

は、

コンテンツエンベディング

e^{cont}

から他の特徴量

y^{j}

を推定するクリティック

C^{j}

を用いた損失関数

L^{j}

を入力の再構成についての損失関数

L^{rec}

に加えることで学習できる。

[0040] 具体的には、下記の式を用いて学習が行われる。

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

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

但し、上記式における

L_{ED}

はエンコーダ１０３Ａ及びデコーダ１０３Ｃの学習のための損失関数を示す。

また、

L_{cj}

はクリティック

C^{j}

のための損失関数であり、

λ_j

は重みパラメータである。

θ^{id}

θ^{pitch}

θ^{loud}

θ^{cont}

θ^{dec}

のそれぞれは、エンコーダ１０３Ａ及びデコーダ１０３Ｃのパラメータであり、

ϕ^j

はクリティック

C^j

のパラメータである。

[0041] 次に、量子化による情報損失による手法の具体例について説明する。

入力歌声データｘからコンテンツエンベディング

e^{cont}

を抽出するエンコーダ

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

の出力をベクトル量子化し、情報を圧縮することで、デコーダに与えられている他の情報

$(e^{id}, e^{pitch}, e^{loud})$

に含まれない情報のみをコンテンツエンベディング

e^{cont}

に保持するように誘導することができる。

[0042] 学習は以下の損失関数の最小化によって行うことができる。

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

但し、ｓｇ（）はニューラルネットワークの勾配情報を以下の層に伝えないようにする勾配停止演算子、Ｖ（）はベクトル量子化演算である。

再構成についての損失関数

L^{rec}

についてはデコーダやエンコーダの種類により色々な形が考えられる。例えば、variational autoencoder (VAE)やベクトル量子化VAEの場合は変分下界（ELBO）

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

を用いることができ、Generative adversarial networkの場合は入力と出力の事情誤差と敵対的損失

L_{adv}

の重み付き和（下記の式）として表すことができる。

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

[0043] 以上説明した学習は、話者特徴量推定部で推定された話者情報を変えずに行われる。一度、学習された後は、話者情報が変化しても構わない。また、学習時には、未来情報を使用してもよい。

[0044] 上記では、声質を決定する話者エンベディングは話者ラベルｎを用いて

$$e^{id} = E^{id}(n)$$

と求める方法について説明した。しかしながら、この方法では変換先の歌手があらかじめ学習データになくてはならず、任意の歌手（未知の話者）に対して声質変換を行うことができない。そこで、音声信号から話者エンベディングを求める方法を説明する。例えば、以下の２つの方法が考えらえる。

[0045] 第１の方法は、所定の話者（例えば、変換先の歌手の歌声データと似た特徴の歌声データの話者）のボーカル信号に基づいて当該話者の話者情報を推定する話者エンベディング推定を行う方法である。話者ラベルｎを用いて学習した話者エンベディング

$$e_n^{id} = E^{id}(n)$$

を話者ｎの歌唱音

$$x_n$$

から推定する話者特徴量推定部Ｆ（）を学習する。Ｆはニューラルネットワークなどで構成することができ、話者エンベディングとの距離を最小化するように学習される。距離としてはＬｐノルム

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

を利用することができる。

[0046] 第２の方法は、所定のボーカル信号に基づいて当該話者の話者情報を推定する歌手識別モデル学習を行う方法である。

歌唱音

$$x_n$$

から話者エンベディング

$$e_n^{id}$$

を抽出する話者特徴量推定部Ｇ（）を声質変換部１０３の学習に先立って学習する。Ｇは歌手ラベルの付いた複数歌手の歌唱音データを用いて以下の目的関数Ｌを最小化することで学習できる。

$$\mathrm{L} = -\min\left(K\left(G(x_n), G(x_m)\right) - K\left(G(x_n), G(x_n')\right) - 1, 0\right)$$

但し、Ｋ（ｘ，ｙ）はｘとｙのコサイン距離、

$$x_n, x_n'$$

は歌手ｎによる異なる歌唱音声、

$$x_n$$

は歌手（ｍ≠ｎ）による歌唱音声である。

この様にして学習されたＧを用いて話者エンベディング

$$e_n^{id}$$

を以下のように求め、声質変換部１０３の学習に用いる。

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

[0047] 上記何れの方法においても、正確な話者エンベディングを得るためには話者特徴量推定部Ｇ（）に入力される入力音声は十分に長いことが好ましい。これは短い音声からでは十分に歌手の特徴を抽出できないためである。一方で、あまりにも長い入力は必要メモリが膨大になるというデメリットがある。そこでＧ（）に再帰構造を持つニューラルネットワーク（recurrent neural network）を用いたり、複数の短時間セグメントを用いて求めた話者エンベディングの平均などを用いたりすることができる。

[0048] ［動作例］

以上のようにして学習された声質変換部１０３により声質変換が行われる。図５を参照しつつ、スマートホン１００で行われる声質変換の処理について説明する。

[0049] 図５において、ボーカル信号ＶＳＢは、カラオケユーザーの歌声データである。また、ボーカル信号ＶＳＡは、カラオケユーザーが声質を近づけたい歌手の歌声データであり、音源分離されたボーカル信号である。

[0050] ボーカル信号ＶＳＡ及びボーカル信号ＶＳＢのそれぞれが声質変換部１０３に入力される。エンコーダ１０３Ａは、ボーカル信号ＶＳＡ及びボーカル信号ＶＳＢから音高、音量等の特徴量を抽出する。

[0051] 特徴量混合部１０３Ｂには、例えば、置き換える特徴量を指定する制御信号が入力される。例えば、ボーカル信号ＶＳＢから抽出された音高情報を、ボーカル信号ＶＳＡから抽出された音高情報にする制御信号が入力されている場合には、特徴量混合部１０１Ｂは、ボーカル信号ＶＳＢから抽出された音高情報をボーカル信号ＶＳＡから抽出された音高情報に置き換える。特徴

量混合部１０１Ｂにより混合された特徴量がデコーダ１０３Ｃに入力される。

[0052] ボーカル信号ＶＳＡ及びボーカル信号ＶＳＢは、話者特徴量推定部１０１Ａに入力される。話者特徴量推定部１０１Ａは、各ボーカル信号から話者情報を推定する。推定された話者情報が特徴量混合部１０１Ｂに供給される。

[0053] 特徴量混合部１０１Ｂには、話者特徴量を置き換えるか否か、置き換える場合にはどの程度の重みで置き換えるかを示す制御信号が入力される。制御信号に応じて、特徴量混合部１０１Ｂは、話者特徴量を適宜、置き換える。例えば、ボーカル信号ＶＳＢから得られた話者特徴量をボーカル信号ＶＳＡから得られた話者特徴量に置き換えた場合には、話者特徴量で規定される声質（狭義の声質）が、カラオケユーザーの声質からボーカル信号ＶＳＡに対応する歌手の声質に置き換わる。特徴量混合部１０１Ｂにより混合された話者特徴量がデコーダ１０３Ｃに供給される。

[0054] デコーダ１０３Ｃは、特徴量混合部１０１Ｂから供給される特徴量および特徴量混合部１０１Ｂから供給される話者特徴量に基づいて、歌声データを生成する。生成された歌声データがスピーカー１０５から再生される。これにより、カラオケユーザーの声質の一部がプロ等の歌手の声質の一部に置き換わった歌声が再生される。

[0055] ［声質変換処理に付随して行われる処理］

次に、声質変換処理に付随して行われる処理について説明する。始めに、滑らかな声質変換を実現する処理について説明する。カラオケなどの用途で自分の歌声を、原曲の歌手の歌声に変えて楽しみたいという要求がある。これは、推論時（声質変換処理の実行時）に歌手Ａ（自分）の歌声を他の歌手（原曲歌手）の声質に変更するため、例えば、歌手Ａの話者エンベディング

e_A^{id}

を歌手Ｂの話者エンベディング

e_B^{id}

に置き換えることで実現できる。

[0056] しかしながら、カラオケなどの用途では自分の歌声を完全に歌手Ｂの声質に変えるのではなく、少しだけ歌手Ｂに似せたいといった要求がある。これを実現するために、歌手Ａの話者エンベディング

$$e_A^{id}$$

を歌手Ｂの話者エンベディング

$$e_B^{id}$$

に滑らかに変化させる内挿関数

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

を用いる。αは変化量を決定するスカラー変数であり、ユーザーが決定することもできる。内挿関数は線形補間や、球面線形補間を用いることができる。

[0057] なお、

$$e_A^{id}$$

だけでなく、

$$e^{pitch}$$

$$e^{loud}$$

$$e^{cont}$$

も同様に線形補間や球面線形補間を用いて内挿できる。例えば、カラオケユーザーの音程

$$f_0^{original}$$

を元音源歌手の音程

$$f_0^{target}$$

に近づけたい場合、

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

の様に線形補間することができる。

[0058] 次に、リアルタイム化処理について説明する。一般的な多くの歌声変換のアルゴリズムは過去と未来の情報を用いるバッチ処理で行われている。一方、カラオケなどで利用する場合、リアルタイムでの変換が必要となる。この際に未来情報が使えないため、高品質な変換を行うことが困難であった。

[0059] そこで本実施形態では、カラオケでの声質変換では多くの場合、原音源中の歌唱とユーザーの歌唱とは同内容の発話（歌詞）であるパラレルデータの関係に着目し、その特徴を利用してリアルタイム処理でも高品質な変換を可能にする。以下、係る変換を実現する処理の具体例について説明する。

[0060] まず、声質変換部１０３が有するエンコーダ１０３Ａ及びデコーダ１０３Ｃを、全て未来情報を利用しない関数にする。これは、エンコーダ１０３Ａやデコーダ１０３Ｃがリカーレントニューラルネットワーク（ＲＮＮ）や畳み込みニューラルネットワーク（ＣＮＮ）で構成されている場合にこれらを、未来情報を利用しない単方向のＲＮＮやCausal convolutionを利用して構成することで実現できる。

[0061] これにより、リアルタイムでの処理が可能となるものの、特に話者エンベディングの正確な推定には十分に長い入力に基づいて求める必要があるため、歌い始めてしばらくの時間は十分な長さの入力が得られず、高品質な変換は難しい。そこで、カラオケでの声質変換では推論時にパラレルデータの関係を利用し、話者エンベディングの推定に短時間の入力のみを利用することを考える。ここで短時間とは１つまたは少数の音素が含まれるような歌唱音声の持続時間で、例えば、数１００ミリ秒から数秒程度である。一般に異なる話者の同音素間での声質変換は比較的容易であり、高品質に変換が行える。そこで話者エンベディングを音素依存にすることで短時間の情報でも高品質な変換を可能にする。しかしながら、学習時にはパラレルデータがない状況を仮定しているため、話者エンベディングは時不変であるとの制約のもとでモデルを学習する必要がある。すなわち、単純に話者エンベディングを短時間情報から求めるようにすること、換言すれば、音素依存の話者エンベデ

ィングの学習はできない。

[0062] そこで、一旦、時不変の話者エンベディングでエンコーダ１０３Ａ及びデコーダ１０３Ｃを学習し、それらのモデルのパラメータを凍結した上でそれらのモデルを使い、事変の話者エンベディングを推定する話者特徴量推定機

$$F^{short}()$$

を学習する。従って、本処理を行う際の話者エンベディングは事変の特徴量として取り扱われる。

$$F^{short}$$

の学習のための目的関数は

$$L(\psi) = L_{rec}\left(x, D\left(F^{short}(x,\psi), e^{pitch}, e^{loud}, e^{cont}\right)\right)$$

と表すことができる。

ここで、エンコーダ１０３Ａやデコーダ１０３Ｃのパラメータは固定されていることに注意されたい。

$$F^{short}$$

の受容野は上記短時間に限られており、上記目的関数を最小化することで求められる。

[0063] このように学習された話者特徴量推定部Ｆは

$$e^{cont}$$

で指定された発話内容（音素）依存で話者エンベディングを求める推定機となっており、短時間情報のみ基づいてリアルタイムに高品質な変換を可能とする。

[0064] 一方で、ある程度長時間歌唱が持続し、十分に長い入力音声から話者エンベディングを求められる時は、図４等を参照して説明した学習を行った話者特徴量推定部Ｆを用いた方時間的安定性が高いことがある。

[0065] そこで、図６に示すように、例えば、話者特徴量推定部１０１Ａが、所定時間以上の長時間情報を用いる話者特徴量推定部（以下、大域的特徴量推定部１２１Ａと適宜、称する）、所定時間より短い短時間情報を用いる話者特

徴量推定部（以下、局所的（音素）特徴量推定部１２１Ｂと適宜、称する）、及び、特徴量結合部１２１Ｃを有する構成とする。そして、大域的特徴量推定部１２１Ａ及び局所的特徴量推定部１２１Ｂを両方用いて話者特徴量を求めるようにすることができる。両推定部から求められた話者特徴量は特徴量結合部１２１Ｃによって結合され、最終的な話者エンベディングを求めることに利用される。結合は重み付き線形結合や球面上線形結合などが利用でき、結合重みパラメータは持続時間や入力信号などから求めることができる。例えば話者エンベディング

$$e^{id}$$

は以下のように求めることができる。

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

但し、Ｔは変換を始めてからの入力長である。αはＴのみに依存して以下のように求めることもできる。

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

またはα（ｘ）の様に入力ｘからニューラルネットワークを用いて求めることもできるし、Ｔ、ｘどちらの情報を用いて求めることも可能である。

[0066] 次に、歌い間違えに対応する処理について説明する。上述したリアルタイム化の処理は推論時に原曲に含まれる歌唱内容と、ユーザーの歌唱内容が一致するという前提（パラレルデータの仮定）がある。一方で、ユーザーは歌い間違いなどする場合があり、必ずしもこの前提は成り立たない。大きく異なる音素間で上記の短時間入力のみを用いる方法で話者エンベディングを求める場合、大きく変換の品質が劣化する場合がある。

[0067] そこで、本処理を行う場合は、図７に示すように、声質変換部１０３に類似度計算部１０３Ｄを設ける。類似度計算部１０３Ｄは、目的歌手と元歌手のコンテンツエンベディング

$$e^{cont}$$

の類似度を計算する。類似度計算部１０３Ｄによる計算結果が話者特徴量推定部１０１Ａに供給される。

[0068] 話者特徴量推定部１０１Ａは、類似度に応じて話者特徴量推定の際の大域的特徴量と局所的特徴量の結合係数（各話者特徴量推定部により推定された話者に関する特徴量のそれぞれに対する重み付け）や他の特徴量混合の重みを変更する。具体的には類似度が低い場合、発話内容が異なるため、短時間情報に基づく話者特徴量に対する結合の重みを小さくしてその依存度を下げる。換言すれば、主として、大域的特徴量推定部１２１Ａの処理結果を用いる。また、その他の特徴量混合において、元話者の特徴量に対する重みを大きくすることで過度に変換を行うことを抑制し、大きな音質の劣化を抑制する。

[0069] 次に、分離音源に対するロバスト化について説明する。歌声変換の学習のためのデータは一般にノイズのないクリーンであるものが好ましい。一方で、本開示では目的話者の歌声音声は音源分離された音声であり、この分離に伴うノイズが含まれている。そのためノイズにより各エンベディングの推定精度が悪化し、変換音声の音質もノイズを含んだものになりやすい。これを防ぐために音源分離ノイズに対して頑健なシステムを構築する方法について説明する。

[0070] 音源分離ノイズに対しての頑健性は音源分離された音声と、元のクリーンな音声とで抽出されるエンベディングベクトルが同一になるようにエンコーダ、デコーダ、話者特徴量推定部の学習中に拘束をかけることで実現できる。具体的にはクリーンな音声信号をｘ、伴奏信号をｂ、音源分離機をｈ（）とすると、正則化項

$$L_{reg} = \left\| E(x) - E\left(h(x+b)\right) \right\|_p$$

を学習の目的関数に加える。

ここでＥはエンコーダ、又は特徴量抽出器である。再構築に関する損失関数

$$L_{rec}$$

に関する計算はクリーン音声のみを用いることでデコーダ１０３Ｃの出力をクリーンに保ったまま、分離音声からの特徴量抽出結果がクリーン音声に対するそれと一致するようにエンコーダ１０３Ａを学習することが可能となる。

[0071] 以上説明した声質変換処理に付随して行われる処理は、全て行われることが好ましいが、一部の処理が行われてもよいし、必ずしも行われなくてもよい。

[0072] ＜変形例＞

以上、本開示の一実施形態について説明したが、本開示は、上述した実施形態に限定されることはなく、本開示の趣旨を逸脱しない範囲で種々の変形が可能である。

[0073] 一実施形態で説明した処理の全てがスマートホン１００で行われる必要はない。一部の処理がスマートホン１００とは別の装置、例えば、サーバによって行われてもよい。例えば、図８に示すように、音源分離処理及び話者特徴量推定処理がサーバによって行われ、声質変換処理及び再生処理がスマートホンで行われるようにしてもよい。また、図９に示すように、音源分離処理がサーバによって行われ、声質変換処理、再生処理及び話者特徴量推定処理がスマートホンで行われてもよい。サーバ及びスマートホンの間では、処理結果がネットワークを介して送受信される。

[0074] また、本開示は、装置、方法、プログラム、システム等、任意の形態により実現することもできる。例えば、上述した実施形態で説明した機能を行うプログラムをダウンロード可能とし、実施形態で説明した機能を有しない装置が当該プログラムをダウンロードしてインストールすることにより、当該装置において実施形態で説明した制御を行うことが可能となる。本開示は、このようなプログラムを配布するサーバにより実現することも可能である。また、各実施形態、変形例で説明した事項は、適宜組み合わせることが可能である。また、本明細書で例示された効果により本開示の内容が限定して解釈されるものではない。

[0075] 本開示は、以下の構成も採ることができる。

（１）

混合音信号からボーカル信号と伴奏信号とを音源分離し、当該音源分離の結果を用いて声質変換を行う声質変換部を有する

情報処理装置。

（２）

前記音源分離により前記混合音信号から第１のボーカル信号が分離され、

前記声質変換部に対して、収音された第２のボーカル信号が入力され、

前記声質変換部は、前記第１のボーカル信号および前記第２のボーカル信号の何れか一方を他方のボーカル信号に近づける

（１）に記載の情報処理装置。

（３）

何れか一方を他方のボーカル信号に近づける変化量が設定可能とされる

（２）に記載の情報処理装置。

（４）

さらに、話者に関する特徴量を推定する話者特徴量推定部を有し、

前記声質変換部は、エンコーダおよびデコーダを有する

（２）に記載の情報処理装置。

（５）

前記話者に関する特徴量は、時間的に変化しない特徴に対応する特徴量であり、

前記エンコーダは、入力されたボーカル信号から、時間的に変化する特徴に対応する特徴量を抽出し、

前記デコーダは、前記話者特徴量推定部により推定された特徴量および前記エンコーダにより抽出された特徴量に基づいてボーカル信号を生成する

（４）に記載の情報処理装置。

（６）

前記時間的に変化しない特徴に対応する特徴量は話者情報であり、

前記時間的に変化する特徴に対応する特徴量は、音高情報、音量情報、発話情報の少なくとも一つを含む

（５）に記載の情報処理装置。

（７）

前記特徴量は、エンベディングベクトルにより規定される

（６）に記載の情報処理装置。

（８）

前記エンコーダは、特定の特徴のみを反映した特徴量からエンベディングベクトルを得る学習またはボーカル信号から特定の特徴のみを抽出するような学習を行うことで得られる学習モデルを用いて、前記時間的に変化する特徴に対応する特徴量のエンベディングベクトルを抽出する

（７）に記載の情報処理装置。

（９）

前記話者特徴量推定部は、所定の話者のボーカル信号に基づいて当該話者の話者情報を推定する学習により得られる学習モデルを用いて話者の特徴量を推定する

（６）から（８）までの何れかに記載の情報処理装置。

（１０）

前記話者特徴量推定部は、所定のボーカル信号に基づいて当該話者の話者情報を推定する学習により得られる学習モデルを用いて話者の特徴量を推定する

（６）から（８）までの何れかに記載の情報処理装置。

（１１）

話者特徴量推定部は、第１の話者特徴量推定部および第２の話者特徴推定部を含み、

前記第１の話者特徴量推定部により推定された話者に関する特徴量と前記第２の話者特徴推定部により推定された話者に関する特徴量とを結合する特徴量結合部を有する

（４）から（１０）までの何れかに記載の情報処理装置。

（１２）

前記第１の話者特徴量推定部は所定時間以上のボーカル信号に基づいて話者に関する特徴量を推定し、前記第２の話者特徴量推定部は前記所定時間より短いボーカル信号に基づいて話者に関する特徴量を推定する

（１１）に記載の情報処理装置。

（１３）

前記第１のボーカル信号と前記第２のボーカル信号との類似度に応じて、前記特徴量結合部における結合係数を変化させる

（１１）に記載の情報処理装置。

（１４）

前記結合係数は、前記第１の話者特徴量推定部により推定された話者に関する特徴量および前記第２の話者特徴量推定部により推定された話者に関する特徴量のそれぞれに対する重み付けである

（１３）に記載の情報処理装置。

（１５）

声質変換部が、混合音信号からボーカル信号と伴奏信号とを音源分離し、当該音源分離の結果を用いて声質変換を行う

情報処理方法。

（１６）

声質変換部が、混合音信号からボーカル信号と伴奏信号とを音源分離し、当該音源分離の結果を用いて声質変換を行う

情報処理方法をコンピュータに実行させるプログラム。

符号の説明

[0076] １００・・・スマートホン

１０２・・・音源分離部

１０１Ａ・・・話者特徴量推定部

１０１Ｂ・・・話者特徴量混合部

１０３・・・声質変換部

１０３Ａ・・・エンコーダ

１０３Ｃ・・・デコーダ

１０３Ｄ・・・類似度計算部

１２１Ａ・・・大域的特徴量推定部

１２１Ｂ・・・局所的特徴量推定部

請求の範囲

[請求項1] 混合音信号からボーカル信号と伴奏信号とを音源分離し、当該音源分離の結果を用いて声質変換を行う声質変換部を有する

情報処理装置。

[請求項2] 前記音源分離により前記混合音信号から第１のボーカル信号が分離され、

前記声質変換部に対して、収音された第２のボーカル信号が入力され、

前記声質変換部は、前記第１のボーカル信号および前記第２のボーカル信号の何れか一方を他方のボーカル信号に近づける

請求項１に記載の情報処理装置。

[請求項3] 何れか一方を他方のボーカル信号に近づける変化量が設定可能とされる

請求項２に記載の情報処理装置。

[請求項4] さらに、話者に関する特徴量を推定する話者特徴量推定部を有し、

前記声質変換部は、エンコーダおよびデコーダを有する

請求項２に記載の情報処理装置。

[請求項5] 前記話者に関する特徴量は、時間的に変化しない特徴に対応する特徴量であり、

前記エンコーダは、入力されたボーカル信号から、時間的に変化する特徴に対応する特徴量を抽出し、

前記デコーダは、前記話者特徴量推定部により推定された特徴量および前記エンコーダにより抽出された特徴量に基づいてボーカル信号を生成する

請求項４に記載の情報処理装置。

[請求項6] 前記時間的に変化しない特徴に対応する特徴量は話者情報であり、

前記時間的に変化する特徴に対応する特徴量は、音高情報、音量情報、発話情報の少なくとも一つを含む

請求項５に記載の情報処理装置。

[請求項7] 前記特徴量は、エンベディングベクトルにより規定される

請求項６に記載の情報処理装置。

[請求項8] 前記エンコーダは、特定の特徴のみを反映した特徴量からエンベディングベクトルを得る学習またはボーカル信号から特定の特徴のみを抽出するような学習を行うことで得られる学習モデルを用いて、前記時間的に変化する特徴に対応する特徴量のエンベディングベクトルを抽出する

請求項７に記載の情報処理装置。

[請求項9] 前記話者特徴量推定部は、所定の話者のボーカル信号に基づいて当該話者の話者情報を推定する学習により得られる学習モデルを用いて話者の特徴量を推定する

請求項６に記載の情報処理装置。

[請求項10] 前記話者特徴量推定部は、所定のボーカル信号に基づいて当該話者の話者情報を推定する学習により得られる学習モデルを用いて話者の特徴量を推定する

請求項６に記載の情報処理装置。

[請求項11] 話者特徴量推定部は、第１の話者特徴量推定部および第２の話者特徴推定部を含み、

前記第１の話者特徴量推定部により推定された話者に関する特徴量と前記第２の話者特徴推定部により推定された話者に関する特徴量とを結合する特徴量結合部を有する

請求項４に記載の情報処理装置。

[請求項12] 前記第１の話者特徴量推定部は所定時間以上のボーカル信号に基づいて話者に関する特徴量を推定し、前記第２の話者特徴量推定部は前記所定時間より短いボーカル信号に基づいて話者に関する特徴量を推定する

請求項１１に記載の情報処理装置。

[請求項13] 前記第１のボーカル信号と前記第２のボーカル信号との類似度に応じて、前記特徴量結合部における結合係数を変化させる

請求項１１に記載の情報処理装置。

[請求項14] 前記結合係数は、前記第１の話者特徴量推定部により推定された話者に関する特徴量および前記第２の話者特徴量推定部により推定された話者に関する特徴量のそれぞれに対する重み付けである

請求項１３に記載の情報処理装置。

[請求項15] 声質変換部が、混合音信号からボーカル信号と伴奏信号とを音源分離し、当該音源分離の結果を用いて声質変換を行う

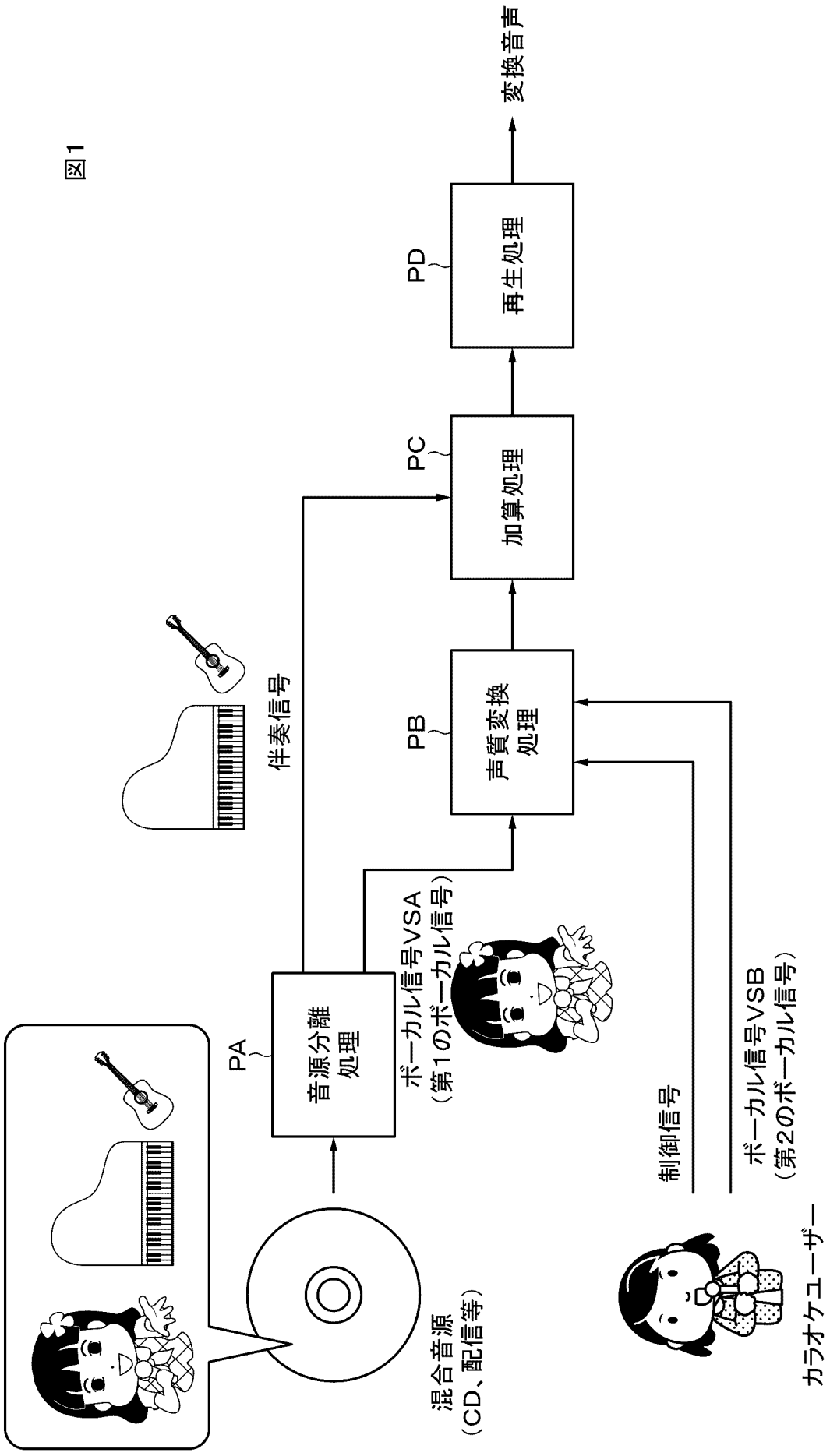
情報処理方法。

[請求項16] 声質変換部が、混合音信号からボーカル信号と伴奏信号とを音源分離し、当該音源分離の結果を用いて声質変換を行う

情報処理方法をコンピュータに実行させるプログラム。

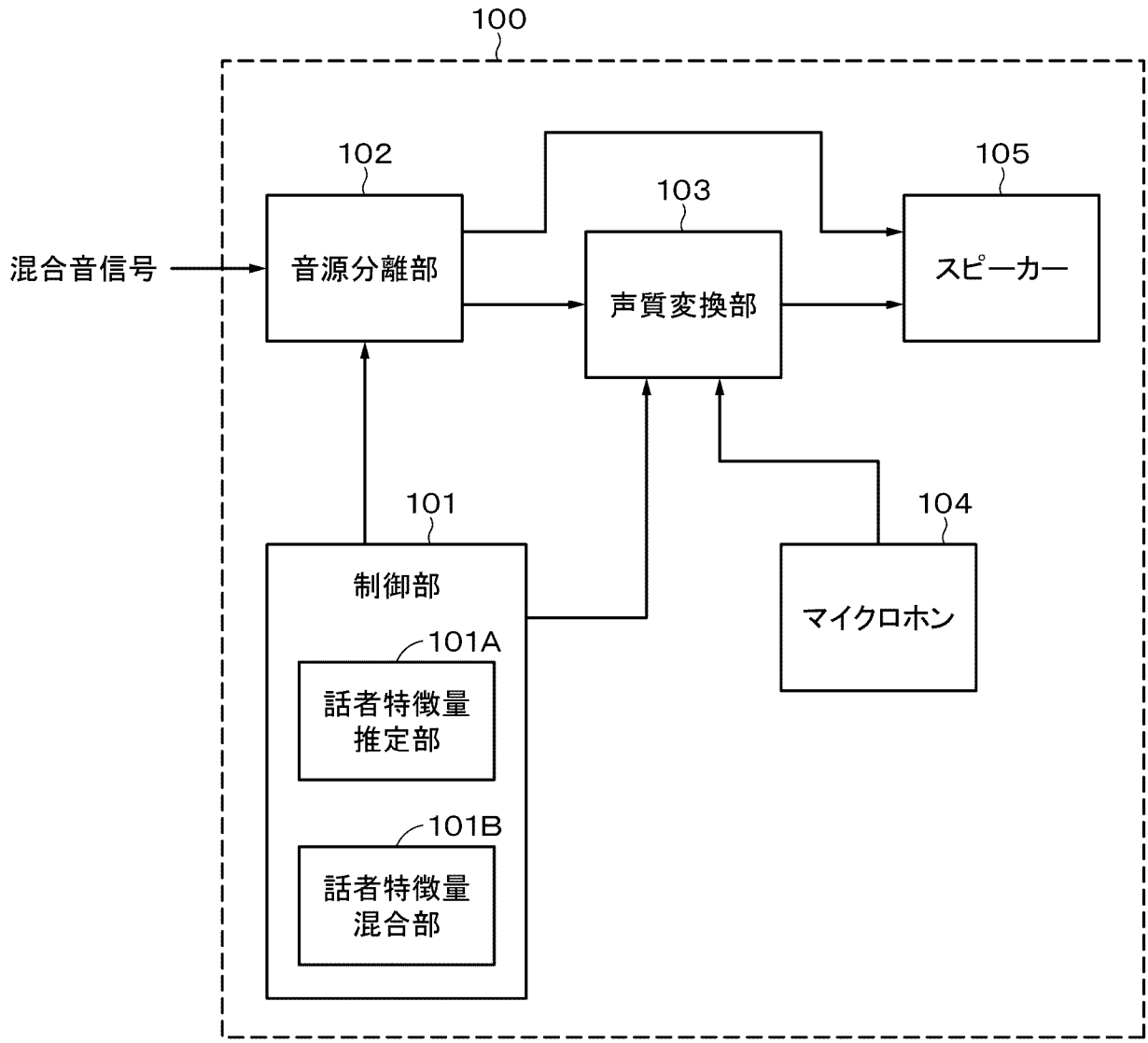
[図1]

図1


変換音声
PD
再生処理
PC
加算処理
PB
声質変換
処理
伴奏信号
PA
音源分離
処理
ボーカル信号VSA
（第1のボーカル信号）
制御信号
ボーカル信号VSB
（第2のボーカル信号）
混合音源
（CD、配信等）
カラオケユーザー

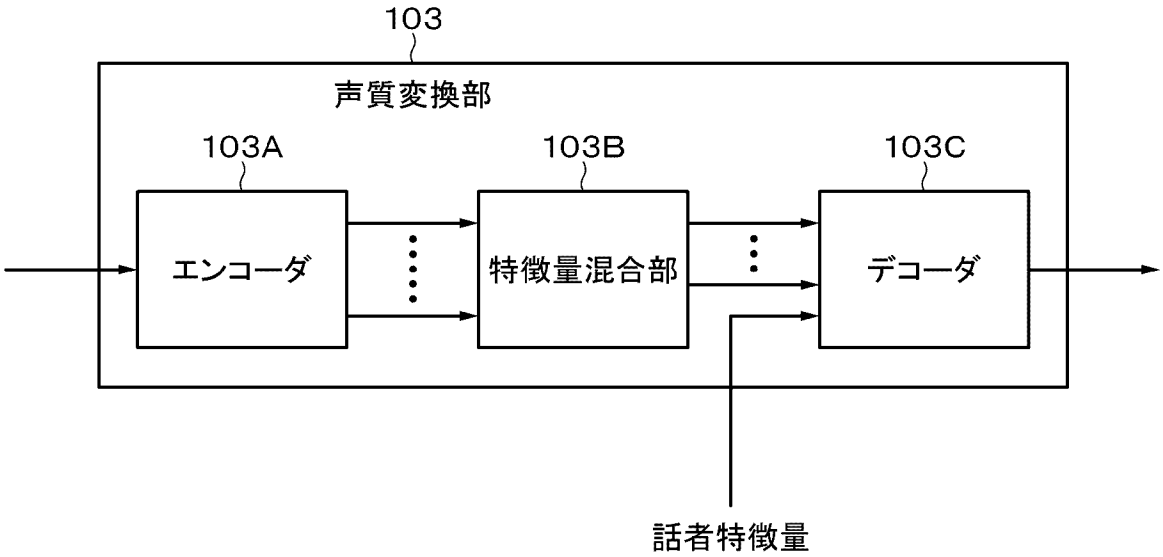
[図2]

図2


100
102
103
105
混合音信号
音源分離部
声質変換部
スピーカー
101
104
制御部
マイクロホン
101A
話者特徴量
推定部
101B
話者特徴量
混合部

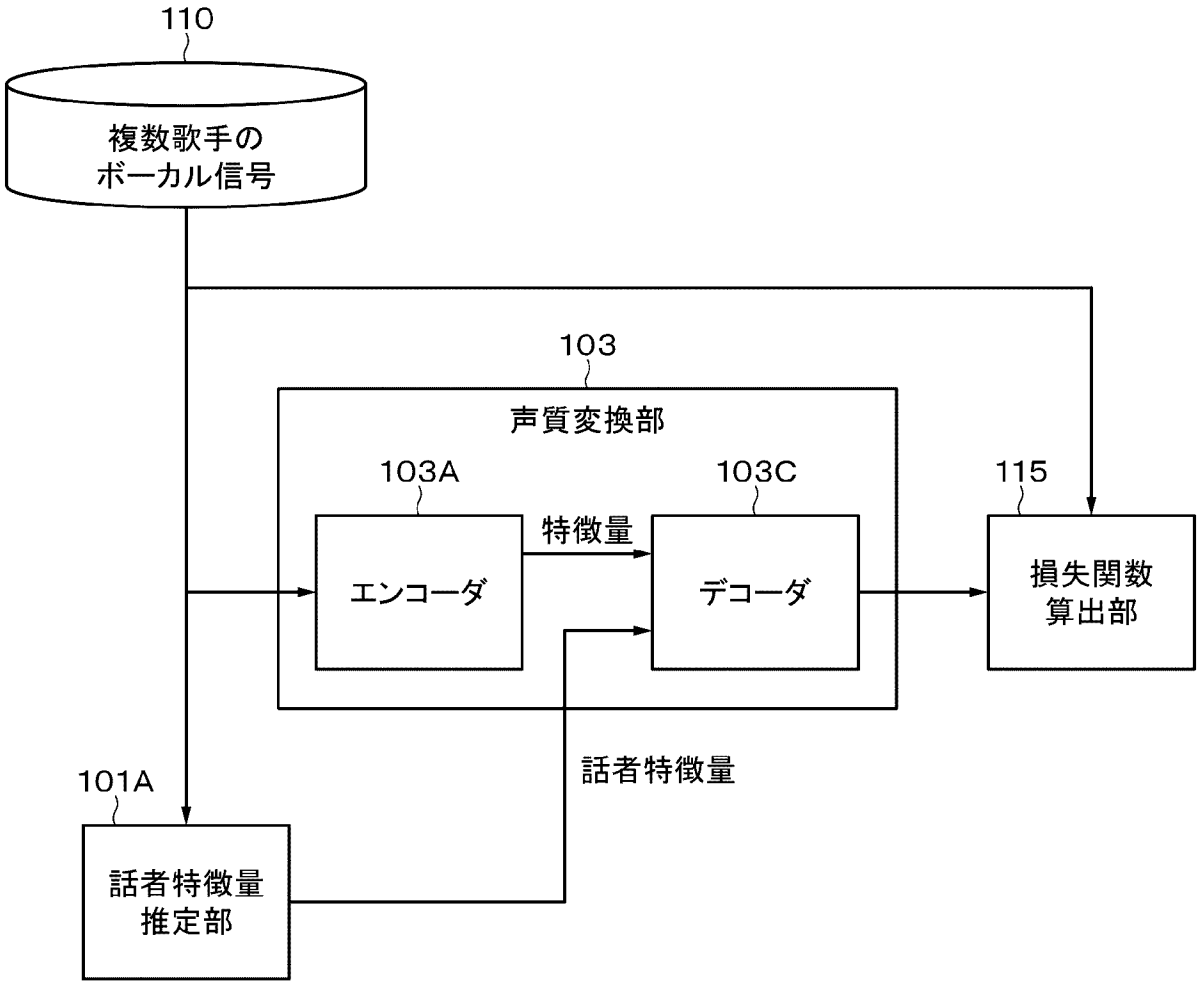
[図3]

図3



[図4]

図4


110
複数歌手の
ボーカル信号
103
声質変換部
103A
103C
115
特徴量
エンコーダ
デコーダ
損失関数
算出部
話者特徴量
101A
話者特徴量
推定部

[図5]

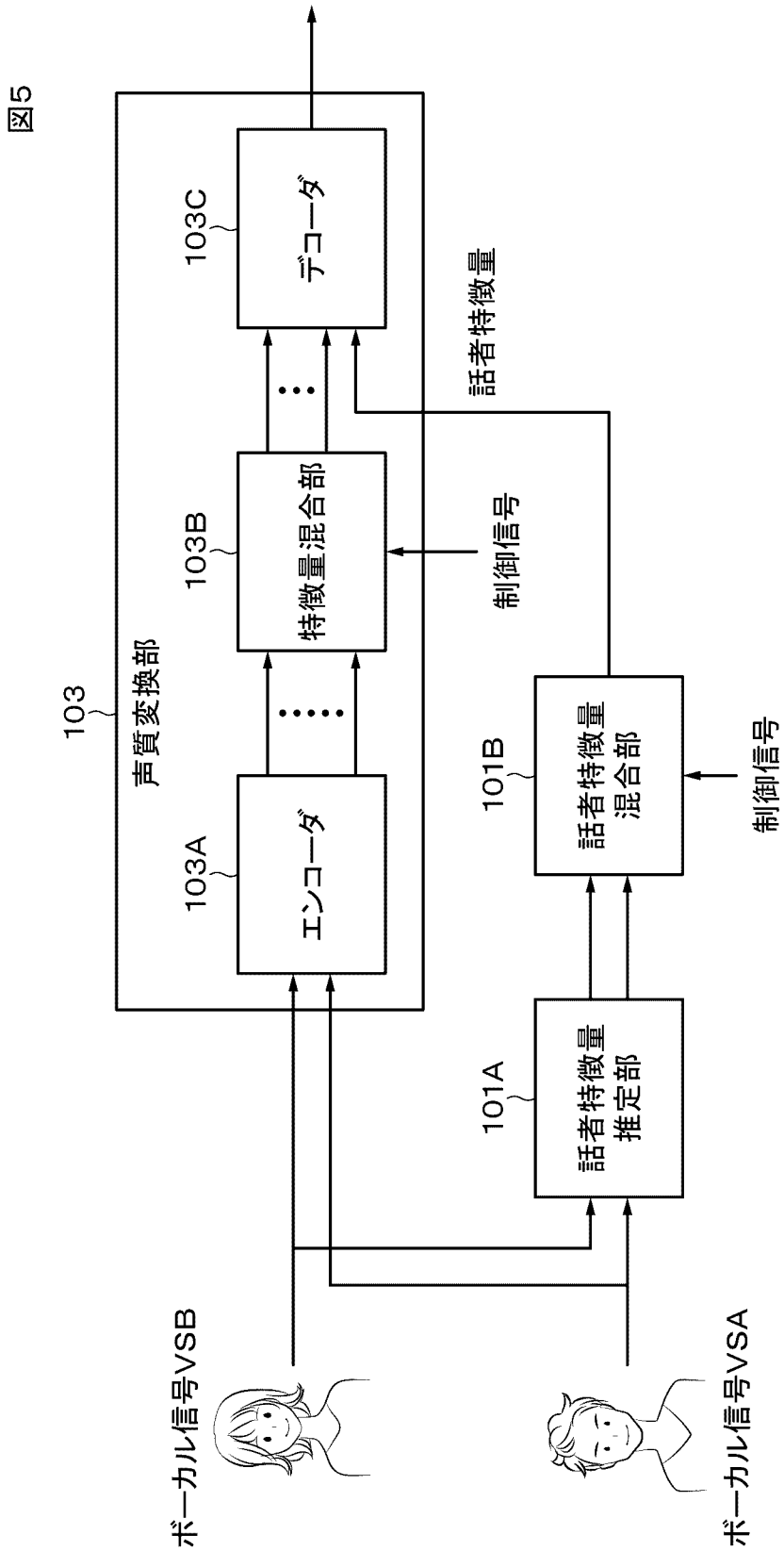
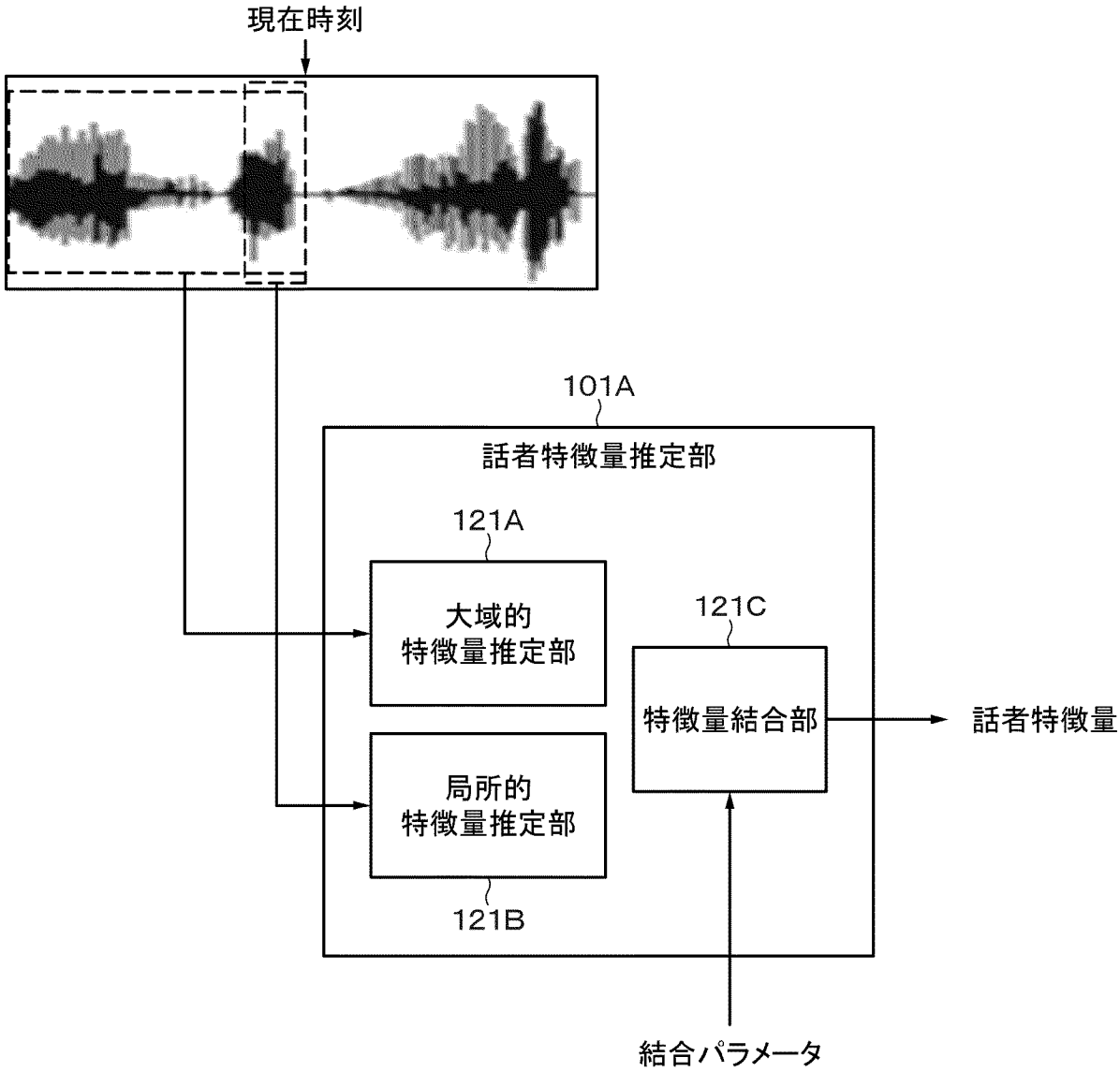


図5
103
声質変換部
103A
エンコーダ
103B
特徴量混合部
103C
デコーダ
101A
話者特徴量
推定部
101B
話者特徴量
混合部
制御信号
話者特徴量
制御信号
ボーカル信号VSB
ボーカル信号VSA

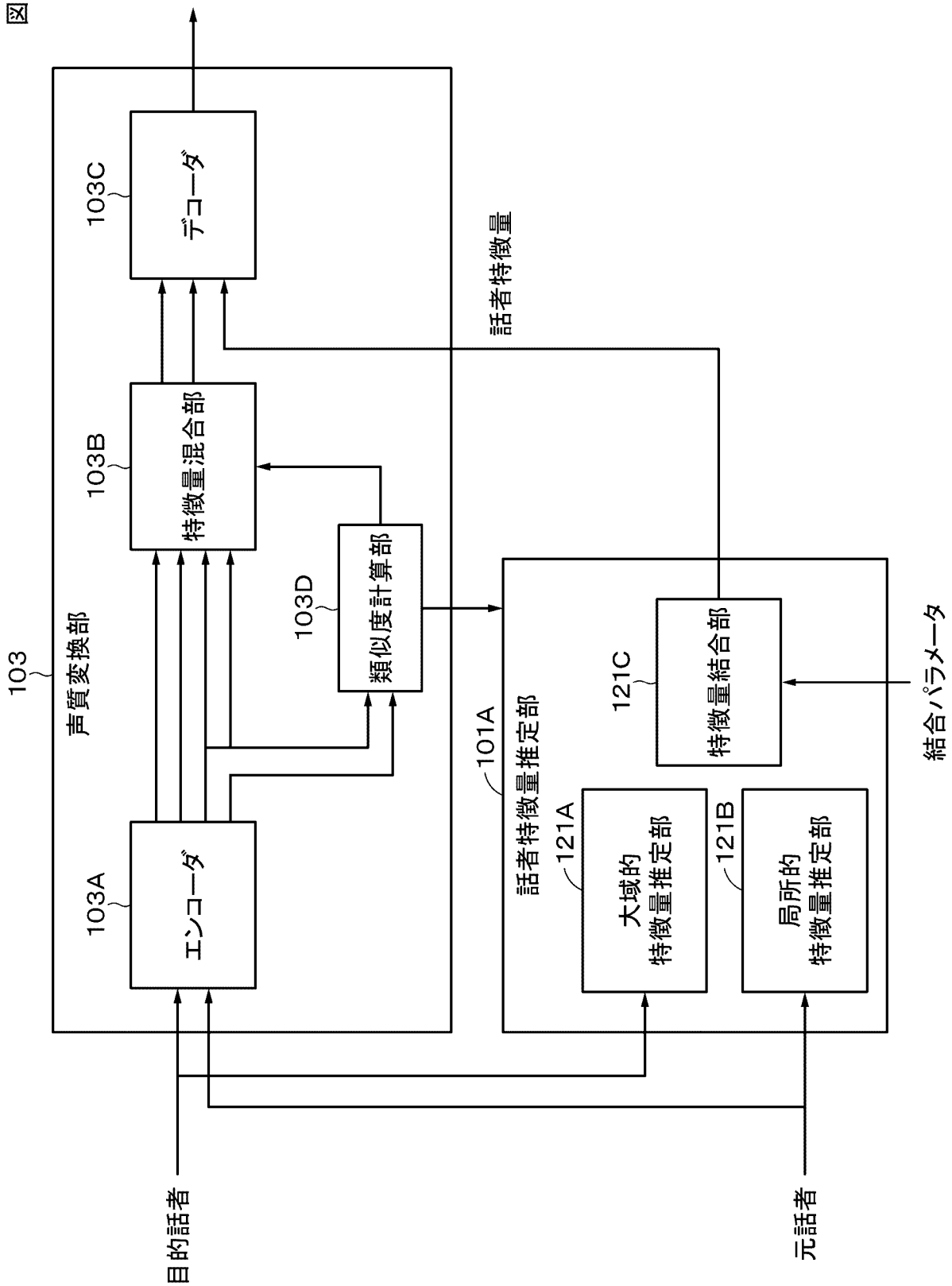
[図6]

図6


現在時刻
101A
話者特徴量推定部
121A
大域的
特徴量推定部
121C
特徴量結合部
話者特徴量
局所的
特徴量推定部
121B
結合パラメータ

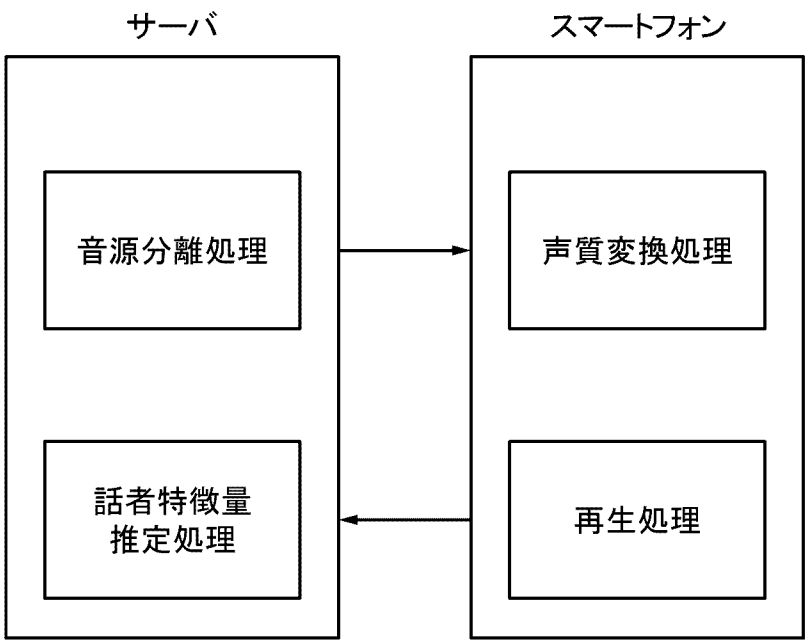
[図7]

図7


103
声質変換部
103A
エンコーダ
103B
特徴量混合部
103C
デコーダ
103D
類似度計算部
101A
話者特徴量推定部
121A
大域的
特徴量推定部
121B
局所的
特徴量推定部
121C
特徴量結合部
話者特徴量
結合パラメータ
目的話者
元話者

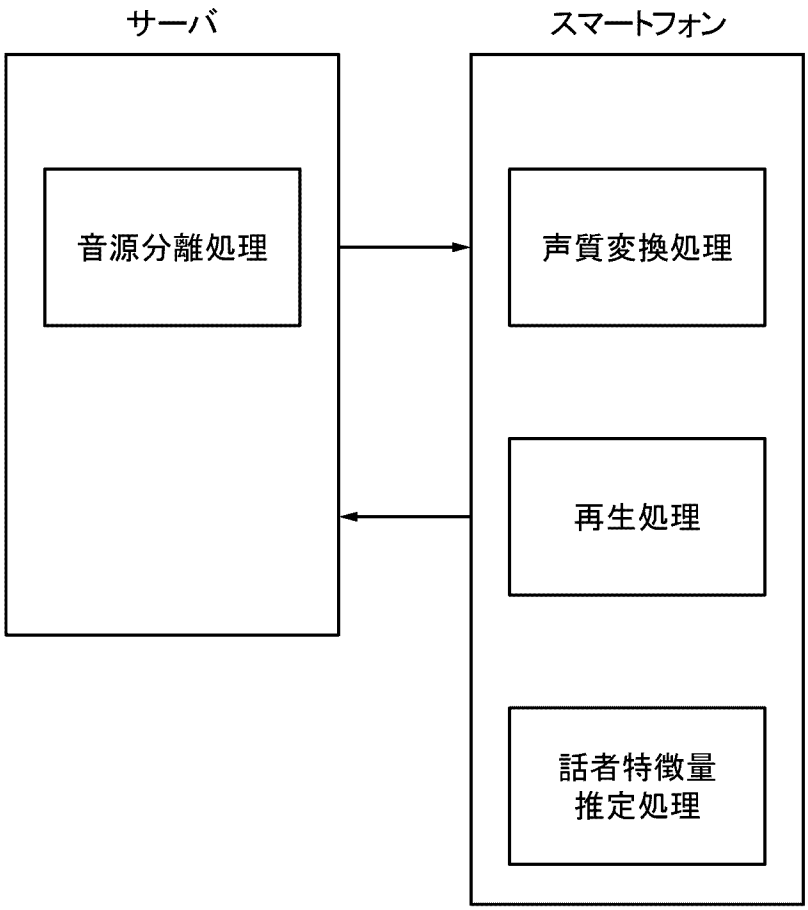
[図8]

図8



[図9]

図9



INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2022/005001

A. CLASSIFICATION OF SUBJECT MATTER

G10L 13/00(2006.01)i; ***G10L 21/007***(2013.01)i; ***G10L 21/0272***(2013.01)i
FI: G10L21/007; G10L21/0272 100Z; G10L13/00 100Y

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)

G10L13/00-25/93

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

Published examined utility model applications of Japan 1922-1996
Published unexamined utility model applications of Japan 1971-2022
Registered utility model specifications of Japan 1996-2022
Published registered utility model applications of Japan 1994-2022

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2019/116889 A1 (SONY CORP) 20 June 2019 (2019-06-20) paragraphs [0147]-[0157]	1, 15-16
Y		1-11, 15-16
A		12-14
Y	DENG, Chengqi et al. PitchNet: Unsupervised Singing Voice Conversion with Pitch Adversarial Network. [online]. 18 February 2020, [retrieval date 14 April 2022], Internet<URL: https://arxiv.org/pdf/1912.01852.pdf> chapter 2, fig. 1	1-2, 4-11, 15-16
A		3, 12-14
Y	QIAN, Kaizhi et al. AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. [online]. 06 June 2019, [retrieval date 14 April 2022], Internet:<URL: https://arxiv.org/pdf/1905.05879.pdf> section 3.2, fig. 1	2, 4-11

☑ 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
"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
18 April 2022	**26 April 2022**
Name and mailing address of the ISA/JP **Japan Patent Office (ISA/JP)** **3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915** **Japan**	Authorized officer Telephone No.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	JP 2001-117598 A (YAMAHA CORP) 27 April 2001 (2001-04-27) paragraphs [0013]-[0023]	1-4, 15-16
A		5-14
A	KR 10-2020-0065248 A (KOREA ADVANCED INST SCI & TECH) 09 June 2020 (2020-06-09) entire text, all drawings	1-16
P, A	CN 113781993 A (BEIJING WODONG TIANJUN INFORMATION TECHNOLOGY CO LTD) 10 December 2021 (2021-12-10) entire text, all drawings	1-16
A	山田 智也ほか, 歌声分離ならびに統計的歌声声質変換に基づく楽曲中の歌声加工, 情報処理学会研究報告, June 2017, vol. 2017-MUS-115, no. 30, pp. 1-6, ISSN 2188-8752 entire text, all drawings, (YAMADA, Tomoya et al. IPSJ SIG Technical Report.), non-official translation (Singing voice processing in music based on singing voice separation and statistical singing voice quality conversion)	1-16

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/JP2022/005001

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
WO 2019/116889 A1	20 June 2019	US 2021/0225383 A1 paragraphs [0151]-[0162] CN 111465982 A	
JP 2001-117598 A	27 April 2001	(Family: none)	
KR 10-2020-0065248 A	09 June 2020	(Family: none)	
CN 113781993 A	10 December 2021	(Family: none)	

国際調査報告

国際出願番号
PCT/JP2022/005001

A. 発明の属する分野の分類（国際特許分類（ＩＰＣ））

G10L 13/00(2006.01)i; G10L 21/007(2013.01)i; G10L 21/0272(2013.01)i
FI: G10L21/007; G10L21/0272 100Z; G10L13/00 100Y

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

G10L13/00-25/93

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０２２年
日本国実用新案登録公報	１９９６－２０２２年
日本国登録実用新案公報	１９９４－２０２２年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献のカテゴリー*	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
X	WO 2019/116889 A1 (ソニー株式会社) 20.06.2019 (2019-06-20) 段落[0147]-[0157]	1,15-16
Y		1-11,15-16
A		12-14
Y	DENG, Chengqi et al., PitchNet: Unsupervised Singing Voice Conversion with Pitch Adversarial Network, [online], 2020.02.18, [検索日 2022.04.14], インターネット<URL: https://arxiv.org/pdf/1912.01852.pdf> 2章, 図1	1-2,4-11,15-16
A		3,12-14
Y	QIAN, Kaizhi et al., AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss, [online], 2019.06.06, [検索日 2022.04.14], インターネット:<URL: https://arxiv.org/pdf/1905.05879.pdf> 3.2節, 図1	2,4-11
Y	JP 2001-117598 A (ヤマハ株式会社) 27.04.2001 (2001-04-27) 段落[0013]-[0023]	1-4,15-16
A		5-14

☑ Ｃ欄の続きにも文献が列挙されている。

☑ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー
"A" 特に関連のある文献ではなく、一般的技術水準を示すもの
"E" 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
"L" 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
"O" 口頭による開示、使用、展示等に言及する文献
"P" 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
"T" 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの
"X" 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
"Y" 特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
"&" 同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
18.04.2022	26.04.2022

名称及びあて先

日本国特許庁(ISA/JP)
〒100-8915
日本国
東京都千代田区霞が関三丁目４番３号

権限のある職員（特許庁審査官）

中村　天真 5Z 1786

電話番号 03-3581-1101 内線 3591

C.　関連すると認められる文献

引用文献のカテゴリー*	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	KR 10-2020-0065248 A (KOREA ADVANCED INST SCI & TECH) 09.06.2020 (2020-06-09) 全文，全図	1-16
P, A	CN 113781993 A (BEIJING WODONG TIANJUN INFORMATION TECHNOLOGY CO LTD) 10.12.2021 (2021-12-10) 全文，全図	1-16
A	山田 智也ほか，歌声分離ならびに統計的歌声声質変換に基づく楽曲中の歌声加工，情報処理学会研究報告，2017.06，Vol.2017-MUS-115，No.30，p.1-6，ISSN 2188-8752 全文，全図	1-16

国際調査報告
パテントファミリーに関する情報

国際出願番号
PCT/JP2022/005001

引用文献	公表日	パテントファミリー文献	公表日
WO 2019/116889 A1	20.06.2019	US 2021/0225383 A1 段落[0151]-[0162] CN 111465982 A	
JP 2001-117598 A	27.04.2001	（ファミリーなし）	
KR 10-2020-0065248 A	09.06.2020	（ファミリーなし）	
CN 113781993 A	10.12.2021	（ファミリーなし）	