(19)日本国特許庁(JP)

(12)特許公報(B1)

(11)特許番号
特許第7164793号
(P7164793)

(45)発行日　令和4年11月2日(2022.11.2)

(24)登録日　令和4年10月25日(2022.10.25)

(51)国際特許分類

		FI		
G10L 13/08 (2013.01)		G10L	13/08	124
G10L 13/033 (2013.01)		G10L	13/033	102A
G10L 13/02 (2013.01)		G10L	13/02	130Z
G10L 15/10 (2006.01)		G10L	15/10	500N

請求項の数　12　（全25頁）

(21)出願番号　特願2021-190678(P2021-190678)
(22)出願日　令和3年11月25日(2021.11.25)
審査請求日　令和3年11月25日(2021.11.25)
早期審査対象出願

(73)特許権者　501440684
ソフトバンク株式会社
東京都港区海岸一丁目７番１号
(74)代理人　100079108
弁理士　稲葉 良幸
(74)代理人　100109346
弁理士　大貫 敏史
(74)代理人　100117189
弁理士　江口 昭彦
(74)代理人　100134120
弁理士　内藤 和彦
(72)発明者　中谷 敏之
東京都港区海岸一丁目７番１号
(72)発明者　末永 君慧
東京都港区海岸一丁目７番１号

最終頁に続く

(54)【発明の名称】　音声処理システム、音声処理装置及び音声処理方法

(57)【特許請求の範囲】

【請求項１】

第１のユーザの発話音声の信号である発話音声信号を取得する取得部と、

前記発話音声信号に基づいて抽出される特徴量を音声認識モデルに入力して、一以上の単語からなる単語列を含むテキストデータを生成する音声認識部と、

前記テキストデータに基づいて抽出される特徴量を音声合成モデルに入力して、合成音声の信号である合成音声信号を生成する音声合成部と、

第２のユーザに対して前記合成音声を出力する音声出力部と、

前記発話音声信号に対応する第１のユーザの感情情報を生成する感情認識部と、

前記感情情報に基づいて、前記音声出力部から前記合成音声又は前記発話音声のどちらを出力するかを切り替える制御部と、を備え、

前記制御部は、前記感情情報が第２のユーザにストレスを与えうるものである場合に前記合成音声を出力する、

音声処理システム。

【請求項２】

第１のユーザの発話音声の信号である発話音声信号を取得する取得部と、

前記発話音声信号に基づいて抽出される特徴量を音声認識モデルに入力して、一以上の単語からなる単語列を含むテキストデータを生成する音声認識部と、

前記テキストデータに基づいて抽出される特徴量を音声合成モデルに入力して、合成音声の信号である合成音声信号を生成する音声合成部と、

第２のユーザに対して前記合成音声を出力する音声出力部と、
前記第２のユーザのストレス状況に関するストレス情報を生成するストレス認識部と、
前記ストレス情報に基づいて、前記音声出力部から前記合成音声又は前記発話音声のどちらを出力するかを切り替える制御部と、を備え、
前記制御部は、前記ストレス情報がストレスの高い状態を示している場合に前記合成音声を出力する、
音声処理システム。
【請求項３】
第１のユーザの発話音声の信号である発話音声信号を取得する取得部と、
前記発話音声信号に基づいて抽出される特徴量を音声認識モデルに入力して、一以上の単語からなる単語列を含むテキストデータを生成する音声認識部と、
前記テキストデータに基づいて抽出される特徴量を音声合成モデルに入力して、合成音声の信号である合成音声信号を生成する音声合成部と、
第２のユーザに対して前記合成音声を出力する音声出力部と、
前記第２のユーザによって入力される切り替え情報に基づいて、前記音声出力部から前記合成音声又は前記発話音声のどちらを出力するかを切り替える制御部と、
前記第２のユーザによって入力された切り替え情報を、前記切り替え情報が入力された際の発話音声信号と時間軸上で関連付けた情報を生成し、当該情報に基づいて、発話音声信号、当該発話音声信号から抽出した特徴量、当該発話音声信号から生成したテキストデータ、当該テキストデータから抽出された特徴量、又はこれらの少なくとも二つの組み合わせを入力とし、前記合成音声と前記発話音声とを切り替えるタイミングを出力とする感情抑制切替モデルを機械学習する学習部と、を備え、
前記制御部は、前記感情抑制切替モデルに、前記取得部が取得した発話音声信号、当該発話音声信号から抽出した特徴量、当該発話音声信号から生成したテキストデータ、当該テキストデータから抽出された特徴量、又はこれらの少なくとも二つの組み合わせを入力することにより、前記合成音声と前記発話音声とを切り替えるタイミングを生成する、
音声処理システム。
【請求項４】
前記感情認識部は、発話音声信号、当該発話音声信号から抽出した特徴量、当該発話音声信号から生成したテキストデータ、当該テキストデータから抽出された特徴量、又はこれらの少なくとも二つの組み合わせを入力とし、当該発話音声信号の発話者の感情情報を出力するよう機械学習された感情認識モデルに、前記取得部が取得した発話音声信号、当該発話音声信号から抽出した音声特徴量、当該発話音声信号から生成したテキストデータ、当該テキストデータに対応するテキスト特徴量、又はこれらの少なくとも二つの組み合わせを入力することにより、前記取得部が取得した発話音声信号に対応する第１のユーザの感情情報を生成する、請求項１に記載の音声処理システム。
【請求項５】
前記発話音声信号に対応する第１のユーザの感情情報を生成する感情認識部と、
前記第２のユーザに対して前記感情情報を表示する表示部と、を備え、
前記表示部は、前記音声出力部による前記合成音声の出力タイミングに合わせて、前記合成音声に対応する前記感情情報を表示する、
請求項１又は４に記載の音声処理システム。
【請求項６】
前記音声合成部は、前記感情認識部が生成した感情情報に基づいて、前記感情情報が示す感情が前記合成音声に反映されるように、前記合成音声信号を生成する、
請求項１又は４に記載の音声処理システム。
【請求項７】
第１のユーザの発話音声の信号である発話音声信号を取得する取得部と、
前記発話音声信号に基づいて抽出される特徴量を音声認識モデルに入力して、一以上の単語からなる単語列を含むテキストデータを生成する音声認識部と、

前記テキストデータに基づいて抽出される特徴量を音声合成モデルに入力して、第２のユーザに対して出力される合成音声の信号である合成音声信号を生成する音声合成部と、
前記発話音声信号に対応する第１のユーザの感情情報を生成する感情認識部と、
前記感情情報に基づいて、前記音声出力部から前記合成音声又は前記発話音声のどちらを出力するかを切り替える制御部と、を備え、
前記制御部は、前記感情情報が第２のユーザにストレスを与えうるものである場合に前記合成音声を出力する、
音声処理装置。
【請求項８】
第１のユーザの発話音声の信号である発話音声信号を取得する取得部と、
前記発話音声信号に基づいて抽出される特徴量を音声認識モデルに入力して、一以上の単語からなる単語列を含むテキストデータを生成する音声認識部と、
前記テキストデータに基づいて抽出される特徴量を音声合成モデルに入力して、第２のユーザに対して出力される合成音声の信号である合成音声信号を生成する音声合成部と、
前記第２のユーザのストレス状況に関するストレス情報を生成するストレス認識部と、
前記ストレス情報に基づいて、前記音声出力部から前記合成音声又は前記発話音声のどちらを出力するかを切り替える制御部と、を備え、
前記制御部は、前記ストレス情報がストレスの高い状態を示している場合に前記合成音声を出力する、
音声処理装置。
【請求項９】
第１のユーザの発話音声の信号である発話音声信号を取得する取得部と、
前記発話音声信号に基づいて抽出される特徴量を音声認識モデルに入力して、一以上の単語からなる単語列を含むテキストデータを生成する音声認識部と、
前記テキストデータに基づいて抽出される特徴量を音声合成モデルに入力して、第２のユーザに対して出力される合成音声の信号である合成音声信号を生成する音声合成部と、
前記第２のユーザによって入力される切り替え情報に基づいて、前記音声出力部から前記合成音声又は前記発話音声のどちらを出力するかを切り替える制御部と、
前記第２のユーザによって入力された切り替え情報を、前記切り替え情報が入力された際の発話音声信号と時間軸上で関連付けた情報を生成し、当該情報に基づいて、発話音声信号、当該発話音声信号から抽出した特徴量、当該発話音声信号から生成したテキストデータ、当該テキストデータから抽出された特徴量、又はこれらの少なくとも二つの組み合わせを入力とし、前記合成音声と前記発話音声とを切り替えるタイミングを出力とする感情抑制切替モデルを機械学習する学習部と、
を備え、
前記制御部は、前記感情抑制切替モデルに、前記取得部が取得した発話音声信号、当該発話音声信号から抽出した特徴量、当該発話音声信号から生成したテキストデータ、当該テキストデータから抽出された特徴量、又はこれらの少なくとも二つの組み合わせを入力することにより、前記合成音声と前記発話音声とを切り替えるタイミングを生成する、
音声処理装置。
【請求項１０】
第１のユーザの発話音声の信号である発話音声信号を取得する工程と、
前記発話音声信号に基づいて抽出される特徴量を音声認識モデルに入力して、一以上の単語からなる単語列を含むテキストデータを生成する工程と、
前記テキストデータに基づいて抽出される特徴量を音声合成モデルに入力して、合成音声の信号である合成音声信号を生成する工程と、
第２のユーザに対して前記合成音声を出力する工程と、
前記発話音声信号に対応する第１のユーザの感情情報を生成する工程と、
前記感情情報に基づいて、前記合成音声又は前記発話音声のどちらを出力するかを切り替える工程と、を含み、

前記切り替える工程は、前記感情情報が第２のユーザにストレスを与えうるものである場合に前記合成音声を出力する、
音声処理方法。
【請求項１１】
第１のユーザの発話音声の信号である発話音声信号を取得する工程と、
前記発話音声信号に基づいて抽出される特徴量を音声認識モデルに入力して、一以上の単語からなる単語列を含むテキストデータを生成する工程と、
前記テキストデータに基づいて抽出される特徴量を音声合成モデルに入力して、合成音声の信号である合成音声信号を生成する工程と、
第２のユーザに対して前記合成音声を出力する工程と、
前記第２のユーザのストレス状況に関するストレス情報を生成する工程と、
前記ストレス情報に基づいて、前記合成音声又は前記発話音声のどちらを出力するかを切り替える工程と、を含み、
前記切り替える工程は、前記ストレス情報がストレスの高い状態を示している場合に前記合成音声を出力する、
音声処理方法。
【請求項１２】
第１のユーザの発話音声の信号である発話音声信号を取得する工程と、
前記発話音声信号に基づいて抽出される特徴量を音声認識モデルに入力して、一以上の単語からなる単語列を含むテキストデータを生成する工程と、
前記テキストデータに基づいて抽出される特徴量を音声合成モデルに入力して、合成音声の信号である合成音声信号を生成する工程と、
第２のユーザに対して前記合成音声を出力する工程と、
前記第２のユーザによって入力される切り替え情報に基づいて、前記音声出力部から前記合成音声又は前記発話音声のどちらを出力するかを切り替える工程と、
前記第２のユーザによって入力された切り替え情報を、前記切り替え情報が入力された際の発話音声信号と時間軸上で関連付けた情報を生成し、当該情報に基づいて、発話音声信号、当該発話音声信号から抽出した特徴量、当該発話音声信号から生成したテキストデータ、当該テキストデータから抽出された特徴量、又はこれらの少なくとも二つの組み合わせを入力とし、前記合成音声と前記発話音声とを切り替えるタイミングを出力とする感情抑制切替モデルを機械学習する工程と、を備え、
前記切り替える工程は、前記感情抑制切替モデルに、前記発話音声信号、当該発話音声信号から抽出した特徴量、当該発話音声信号から生成したテキストデータ、当該テキストデータから抽出された特徴量、又はこれらの少なくとも二つの組み合わせを入力することにより、前記合成音声と前記発話音声とを切り替えるタイミングを生成する、
音声処理方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、音声処理システム、音声処理装置及び音声処理方法に関する。
【背景技術】
【０００２】
従来、顧客満足度（Customer Satisfaction：ＣＳ）向上のために、顧客の苦情等に対してオペレータが電話で応対する各種のコールセンターが運用されている。このような顧客応対業務では、顧客がオペレータに対して威圧的な言動や理不尽な要求を行う「カスタマーハラスメント」により、オペレータの精神不調を招いたり、オペレータの離職率が高くなったりすることが問題視されている。
【０００３】
近年、このようなカスタマーハラスメントから、企業側が従業員であるオペレータを守るための音声変換システムも検討されている。例えば、特許文献１では、入力音声信号か

ら音量及びピッチ変動量を算出し、音量及びピッチ変動量が所定値を超える場合に、音量及びピッチ変動量が所定内に収まるように音量及びピッチを変換して出力するように制御することが記載されている。

【先行技術文献】

【特許文献】

【０００４】

【文献】特開２００４－２５２０８５号公報

【発明の概要】

【発明が解決しようとする課題】

【０００５】

しかしながら、例えば、特許文献１に記載の方法で話し手の発話音声を変換するだけでは、話し手（第１のユーザ）の感情が十分に抑制されず、聞き手（第２のユーザ）のストレスを十分に軽減できない恐れがある。一方、聞き手のストレスを軽減するために、聞き手に出力される話し手の発話音声を変換すると、聞き手が話し手の感情を十分に認識できず、聞き手が適切な応対を行うことができない恐れもある。

【０００６】

そこで、本発明は、聞き手のストレスの十分な軽減、及び／又は、聞き手の適切な応対を可能とする音声処理システム、音声処理装置及び音声処理方法を提供する。

【課題を解決するための手段】

【０００７】

本発明の一つの態様に係る音声処理システムは、第１のユーザの発話音声の信号である発話音声信号を取得する取得部と、前記発話音声信号に基づいて抽出される特徴量を音声認識モデルに入力して、一以上の単語からなる単語列を含むテキストデータを生成する音声認識部と、前記テキストデータに基づいて抽出される特徴量を音声合成モデルに入力して、合成音声の信号である合成音声信号を生成する音声合成部と、第２のユーザに対して前記合成音声を出力する音声出力部と、前記発話音声信号に対応する第１のユーザの感情情報を生成する感情認識部と、前記感情情報に基づいて、前記音声出力部から前記合成音声又は前記発話音声のどちらを出力するかを切り替える制御部と、を備え、前記制御部は、前記感情情報が第２のユーザにストレスを与えうるものである場合に前記合成音声を出力する、を備える。

【０００８】

本発明の一つの態様に係る音声処理システムは、第１のユーザの発話音声の信号である発話音声信号を取得する取得部と、前記発話音声信号に基づいて抽出される特徴量を音声認識モデルに入力して、一以上の単語からなる単語列を含むテキストデータを生成する音声認識部と、前記テキストデータに基づいて抽出される特徴量を音声合成モデルに入力して、合成音声の信号である合成音声信号を生成する音声合成部と、第２のユーザに対して前記合成音声を出力する音声出力部と、前記第２のユーザのストレス状況に関するストレス情報を生成するストレス認識部と、前記ストレス情報に基づいて、前記音声出力部から前記合成音声又は前記発話音声のどちらを出力するかを切り替える制御部と、を備え、前記制御部は、前記ストレス情報がストレスの高い状態を示している場合に前記合成音声を出力する。

【０００９】

本発明の一つの態様に係る音声処理システムは、第１のユーザの発話音声の信号である発話音声信号を取得する取得部と、前記発話音声信号に基づいて抽出される特徴量を音声認識モデルに入力して、一以上の単語からなる単語列を含むテキストデータを生成する音声認識部と、前記テキストデータに基づいて抽出される特徴量を音声合成モデルに入力して、合成音声の信号である合成音声信号を生成する音声合成部と、第２のユーザに対して前記合成音声を出力する音声出力部と、前記第２のユーザによって入力される切り替え情報に基づいて、前記音声出力部から前記合成音声又は前記発話音声のどちらを出力するかを切り替える制御部と、前記第２のユーザによって入力された切り替え情報を、前記切り替

え情報が入力された際の発話音声信号と時間軸上で関連付けた情報を生成し、当該情報に基づいて、発話音声信号、当該発話音声信号から抽出した特徴量、当該発話音声信号から生成したテキストデータ、当該テキストデータから抽出された特徴量、又はこれらの少なくとも二つの組み合わせを入力とし、前記合成音声と前記発話音声とを切り替えるタイミングを出力とする感情抑制切替モデルを機械学習する学習部と、を備え、前記制御部は、前記感情抑制切替モデルに、前記取得部が取得した発話音声信号、当該発話音声信号から抽出した特徴量、当該発話音声信号から生成したテキストデータ、当該テキストデータから抽出された特徴量、又はこれらの少なくとも二つの組み合わせを入力することにより、前記合成音声と前記発話音声とを切り替えるタイミングを生成する。

【図面の簡単な説明】

【００１０】

【図１】本実施形態に係る音声処理システム１の概略の一例を示す図である。

【図２】本実施形態に係る音声処理システム１を構成する各装置の物理構成の一例を示す図である。

【図３】本実施形態に係る音声処理装置１０の機能構成の一例を示す図である。

【図４】本実施形態に係る合成音声信号の生成の一例を示す図である。

【図５Ａ】本実施形態に係る顧客の感情情報の生成の一例を示す図である。

【図５Ｂ】本実施形態に係る顧客の感情情報の生成の一例を示す図である。

【図６】本実施形態に係るオペレータ端末２０の機能構成の一例を示す図である。

【図７】本実施形態に係る画面Ｄ１の一例を示す図である。

【図８】本実施形態に係る画面Ｄ２の一例を示す図である。

【図９】本実施形態に係る感情抑制動作の一例を示すフローチャートである。

【図１０】本実施形態に係る感情抑制機能の自動切り替え動作を示すフローチャートである。

【図１１】本実施形態の変更例に係る合成音声信号の生成の一例を示す図である。

【図１２】本実施形態に係る画面Ｄ３の一例を示す図である。

【発明を実施するための形態】

【００１１】

添付図面を参照して、本発明の実施形態について説明する。なお、各図において、同一の符号を付したものは、同一又は同様の構成を有する。

【００１２】

以下、本実施形態に係る音声処理システムをコールセンター等の顧客応対業務において使用することを想定して説明を行うが、本発明の適用形態はこれに限られない。本実施形態は、第１のユーザの発話音声の信号（以下、「発話音声信号」という）に所定の処理を施して生成される音声を第２のユーザに対して出力するどのような場面にも適用可能である。以下では、第１のユーザが顧客であり、第２のユーザがオペレータであるものとするが、これに限られない。

【００１３】

（音声処理システムの構成）

＜全体構成＞

図１は、本実施形態に係る音声処理システム１の概略の一例を示す図である。図１に示すように、音声処理システム１は、音声処理装置１０と、第２のユーザ（以下、「オペレータ」という）によって使用される端末（以下、「オペレータ端末」という）２０と、第１のユーザ（以下、「顧客」という）によって使用される端末（以下、「顧客端末」という）３０と、を備える。

【００１４】

音声処理装置１０は、顧客端末３０で取得される発話音声信号を、ネットワーク４０を介して受信する。ネットワーク４０は、インターネット等の外部ネットワークであってもよいし、外部ネットワーク、及び、Local Access Network（ＬＡＮ）等の内部ネットワークを含んでもよい。音声処理装置１０は、顧客の発話音声信号に対して所定の処理を施

した音声をオペレータ端末２０に送信する。なお、音声処理装置１０は、一つ又は複数のサーバで構成されてもよい。

【００１５】

オペレータ端末２０は、例えば、電話、スマートフォン、パーソナルコンピュータ、タブレット等である。オペレータ端末２０は、音声処理装置１０で所定の処理で生成される音声信号又は顧客端末３０からの発話音声信号に基づいて、音声をオペレータに出力する。

【００１６】

顧客端末３０は、例えば、電話、スマートフォン、パーソナルコンピュータ、タブレット等である。顧客端末３０は、顧客の発話音声をマイクにより収音して、当該発話音声の信号である発話音声信号を音声処理装置１０に送信する。

【００１７】

＜物理構成＞

図２は、本実施形態に係る音声処理システム１を構成する各装置の物理構成の一例を示す図である。各装置（例えば、音声処理装置１０、オペレータ端末２０及び顧客端末３０）は、演算部に相当するプロセッサ１０ａと、記憶部に相当するＲＡＭ（Random Access Memory）１０ｂと、記憶部に相当するＲＯＭ（Read Only Memory）１０ｃと、通信部１０ｄと、入力部１０ｅと、表示部１０ｆと、カメラ１０ｇ、音声入力部１０ｈと、音声出力部１０ｉと、を有する。これらの各構成は、バスを介して相互にデータ送受信可能に接続される。なお、図２で示す構成は一例であり、各装置はこれら以外の構成を有してもよいし、これらの構成のうち一部を有さなくてもよい。

【００１８】

プロセッサ１０ａは、例えば、ＣＰＵ（Central Processing Unit）である。プロセッサ１０ａは、ＲＡＭ１０ｂ又はＲＯＭ１０ｃに記憶されているプログラムを実行することにより、各装置における各種処理を制御する制御部である。プロセッサ１０ａは、各装置が備える他の構成と、プログラムとの協働により、各装置の機能を実現し、処理の実行を制御する。プロセッサ１０ａは、入力部１０ｅや通信部１０ｄから種々のデータを受け取り、データの演算結果を表示部１０ｆに表示したり、ＲＡＭ１０ｂに格納したりする。

【００１９】

ＲＡＭ１０ｂ及びＲＯＭ１０ｃは、各種処理に必要なデータ及び処理結果のデータを記憶する記憶部である。各装置は、ＲＡＭ１０ｂ及びＲＯＭ１０ｃ以外に、ハードディスクドライブ等の大容量の記憶部を備えてもよい。ＲＡＭ１０ｂ及びＲＯＭ１０ｃは、例えば、半導体記憶素子で構成されてもよい。

【００２０】

通信部１０ｄは、各装置を他の機器に接続するインターフェースである。通信部１０ｄは、他の機器と通信する。入力部１０ｅは、ユーザからデータの入力を受け付けるためのデバイスや、各装置の外部からデータを入力するためのデバイスである。入力部１０ｅは、例えば、キーボード、マウス及びタッチパネル等を含んでよい。表示部１０ｆは、プロセッサ１０ａによる制御に従って、情報を表示するデバイスである。表示部１０ｆは、例えば、ＬＣＤ（Liquid Crystal Display）により構成されてよい。

【００２１】

カメラ１０ｇは、静止画像又は動画像を撮像する撮像素子を含み、所定の領域の撮像により撮像画像（例えば、静止画像又は動画像）を生成する。音声入力部１０ｈは、音声を収音するデバイスであり、例えば、マイクである。音声出力部１０ｉは、音声を出力するデバイスであり、例えば、スピーカーである。

【００２２】

各装置を実行させるためのプログラムは、ＲＡＭ１０ｂやＲＯＭ１０ｃ等のコンピュータによって読み取り可能な記憶媒体に記憶されて提供されてもよいし、通信部１０ｄにより接続されるネットワーク４０を介して提供されてもよい。各装置では、プロセッサ１０ａが当該プログラムを実行することにより、各装置を制御するための様々な動作が実現される。なお、これらの物理的な構成は例示であって、必ずしも独立した構成でなくてもよ

い。例えば、各装置は、プロセッサ１０ａとＲＡＭ１０ｂやＲＯＭ１０ｃが一体化したＬＳＩ（Large-Scale Integration）を備えていてもよい。

【００２３】

＜機能的構成＞

≪音声処理装置≫

図３は、本実施形態に係る音声処理装置１０の機能構成の一例を示す図である。音声処理装置１０は、記憶部１０１、送受信部１０２、音声認識部１０３、除去部１０４、音声合成部１０５、感情認識部１０６、ストレス認識部１０７、制御部１０８、学習部１０９を含む。

【００２４】

記憶部１０１は、各種情報、プログラム、アルゴリズム、モデル、操作ログ等を記憶する。具体的には、記憶部１０１は、後述する音声認識モデル１０１ａ、音声合成モデル１０１ｂ、感情認識モデル１０１ｃ、ストレス認識モデル１０１ｄ、感情抑制切替モデル１０１ｅ等を記憶する。

【００２５】

送受信部１０２は、オペレータ端末２０及び／又は顧客端末３０との間で、種々の情報及び／又は信号を送信及び／又は受信する。例えば、送受信部１０２（取得部）は、顧客端末３０で収音された顧客の発話音声の信号である発話音声信号を取得する。送受信部１０２は、オペレータ端末２０に対して、合成音声信号及び／又は発話音声信号を送信する。また、送受信部１０２は、オペレータ端末２０からオペレータによる操作ログを取得してもよい。操作ログにはオペレータによる顧客の感情の主観的評価に関する情報（以下、「主観的評価情報」という）、後述する「ストレスの度合い」、後述する「手動切替履歴データ」が含まれてよい。また、送受信部１０２は、オペレータ端末２０に対して、顧客の感情に関する情報（以下、「感情情報」という）等を送信してもよい。

【００２６】

音声認識部１０３は、送受信部１０２で取得された発話音声信号に基づいて抽出される特徴量（以下、「音声特徴量」という）を音声認識モデル１０１ａに入力して、一以上の単語からなる単語列を含むテキストデータを生成する。具体的には、音声認識部１０３は、音声認識モデル１０１ａの音響モデルを用いて上記音声特徴量から単語列を生成し、言語モデルを用いた単語列の分析結果に従って上記テキストデータを生成してもよい。音声認識部１０３は、発話音声信号に対して前処理（例えば、アナログ信号のディジタル化、ノイズの除去、フーリエ変換等）を実施して、音声特徴量を抽出してもよい。

【００２７】

音声認識モデル１０１ａは、音声信号に基づいて音声の内容を推定するアルゴリズムである。音声認識モデル１０１ａは、ある単語がどのような音となって現れやすいかということをモデル化した音響モデル、及び／又は、特定の言語においてある単語列がどのくらいの確率で現れるかをモデル化した言語モデルを含んでもよい。音響モデルとしては、例えば、隠れマルコフモデル（Hidden Markov Model：ＨＭＭ）及び／又はディープニューラルネットワーク（Deep Neural Network：ＤＮＮ）が用いられてもよい。言語モデルとしては、例えば、ｎグラム言語モデル等の確率的言語モデルが用いられてもよい。

【００２８】

除去部１０４は、音声認識部１０３で生成されたテキストデータに含まれる特定の単語列を検出し、当該特定の単語列を除去又は前記特定の単語列を他の単語列に置換したテキストデータを生成し、音声合成部１０５に出力する。除去部１０４は、音声認識部１０３で生成されたテキストデータ内で特定の単語列が検出されない場合、当該テキストデータを音声合成部１０５に出力してもよい。

【００２９】

当該特定の単語列は、例えば、聞き手を侮辱したり、聞き手の人格を否定したりする、聞き手を不快にする等、聞き手に心理的悪影響を与える一以上の単語であってもよい。ここで、各単語は、名詞、動詞、副詞、助詞、形容詞、助動詞等の少なくとも一つの品詞、

当該品詞が音変化したもの等を含んでもよい。例えば、特定の単語列は、「お前、ぶっ殺すぞ」というような「文」であってもよいし、「困るっつってんの」の「っつってん」等、乱暴な言葉遣いであることを示す「文の一部」であってもよい。除去部１０４は、テキストデータ内で検出された特定の単語列のみを他の単語列に置き換えたテキストデータを音声合成部１０５に出力してもよいし、又は、当該特定の単語列を含む文全体を他の単語列に置き換えたテキストデータを音声合成部１０５に出力してもよい。当該他の単語列は、空白等であってもよい。

【００３０】

除去部１０４は、記憶部１０１に予め記憶された特定の単語列に基づいて、テキストデータ内の特定の単語列の検出及び／又は他の単語列への置き換えを実施してもよい。

【００３１】

或いは、除去部１０４は、機械学習により学習されたモデルに基づいて、テキストデータ内の特定の単語列の検出、及び／又は、意味的感情を緩和した他の単語列への置き換えを実施してもよい。例えば、テキストデータ内の特定の単語列「お前」は、「あなた」に置換されてもよい。機械学習に基づくモデルに基づいて、テキストデータ内の特定の単語列の検出及び／又は他の単語列への置き換えを実施してもよい。

【００３２】

なお、除去部１０４は、テキストデータ内で特定の単語列が検出される場合、当該特定の単語列の検出に関する情報（以下、「検出情報」という）を生成してもよい。当該検出情報は、例えば、当該特定の単語列が検出されたことを示す情報（例えば、「ＮＧワード」又は「ＮＧワード検出」という文字列）、当該特定の単語列を示す情報、及び、顧客に対する警告に関する情報（以下、「警告情報」という）の少なくとも一つを含んでもよい。当該警告情報は、例えば、オペレータに対する顧客の発話内容が侮辱罪、名誉棄損罪等の刑事告訴対象となり得ることを通知するための情報であってもよい。検出情報は、送受信部１０２によってオペレータ端末２０に送信されてもよい。検出情報が生成された場合、音声処理装置１０は、顧客端末３０に対して警告情報（例えば、「当社オペレータに対して侮辱罪等の恐れがあります。当社の不手際もあるとは思いますが、当社オペレータに過度な負担になる場合がありますのでご協力を頂けますと幸いです。」）を出力させてもよい。このような警告情報は、カスタマーハラスメントに対する事前告知として利用することができる。

【００３３】

音声合成部１０５は、除去部１０４から入力されるテキストデータに基づいて抽出される特徴量（以下、「テキスト特徴量」という）を音声合成モデル１０１ｂに入力して、合成音声の信号（以下、「合成音声信号」という）を生成する。具体的には、除去部１０４は、テキスト特徴量に基づいて音声合成パラメータを予測し、予測された音声合成パラメータを用いて合成音声信号を生成してもよい。音声合成部１０５は、合成音声信号を送受信部１０２に出力する。合成音声信号は、テキストデータの内容を読み上げた音声の信号ともいえる。

【００３４】

音声合成モデル１０１ｂは、テキストデータを入力として当該テキストデータの内容に対応する合成音声信号を出力するアルゴリズムである。音声合成モデル１０１ｂとしては、例えば、上記ＨＭＭ及び／又はＤＮＮが用いられてもよい。

【００３５】

当該音声合成モデル１０１ｂは、複数の音声種別に対応してもよい。音声合成部１０５は、複数の音声種別の中から合成音声信号に用いる音声種別を選択し、選択した音声種別とテキストデータとを音声合成モデル１０１ｂに入力して、選択した音声種別の合成音声信号を合成してもよい。当該複数の音声種別は、例えば、抑揚が少ない音声、機械音、キャラクターの音声、芸能人の音声及び声優の音声の少なくとも一つ等であってもよい。音声合成部１０５は、オペレータからオペレータ端末２０を介して音声種別の選択を受け付けてもよい。

【００３６】
図４は、本実施形態に係る合成音声信号の生成の一例を示す図である。図４では、送受信部１０２で取得された発話音声信号Ｓ１～Ｓ３に基づいて、音声認識部１０３においてテキストデータＴ１～Ｔ３が生成されるものとする。例えば、図４では、除去部１０４は、テキストデータＴ１内で特定の単語列を検出しないので、テキストデータＴ１をそのまま音声合成部１０５に出力する。一方、除去部１０４は、テキストデータＴ２及びＴ３内で特定の単語列（Ｔ２では「お前、ぶっ殺すぞ」、Ｔ３では「っつってん」）を検出するので、当該特定の単語列を除去又は置換したテキストデータＴ２’及びＴ３’を音声合成部１０５に出力する。例えば、テキストデータＴ２’では、テキストデータＴ２内の特定の単語列が空白（□）に置換される。また、テキストデータＴ３’では、テキストデータＴ３内の特定の単語列「っつってん」が「という」に置換される。音声合成部１０５は、テキストデータＴ１、Ｔ２及びＴ３からそれぞれ合成音声信号Ｓ１、Ｓ２’及びＳ３’を生成する。
【００３７】
感情認識部１０６は、送受信部１０２で取得された発話音声信号、音声認識部１０３で生成されたテキストデータ、及び、送受信部１０２で受信される主観的評価情報の少なくとも一つに基づいて、顧客の感情情報を生成する。感情認識部１０６は、発話音声信号に基づいて抽出された音声特徴量（例えば抑揚や音量など）に基づいて顧客の感情情報を生成してよい。感情認識部１０６は、発話音声信号に基づいて生成されたテキストデータに特定の単語列が検出されたこと、又は、特定の単語列が所定時間以上検出されなかったことに基づいて顧客の感情情報を生成してよい。感情認識部１０６は、カメラ１０ｇで取得される顧客の撮像画像に基づいて、顧客の感情情報を生成してもよい。感情認識部１０６は感情認識モデル１０１ｃを用いて顧客の感情情報を生成してもよい。
【００３８】
感情認識モデル１０１ｃは、発話音声信号、当該発話音声信号から抽出した音声特徴量、当該発話音声信号から生成したテキストデータ、テキスト特徴量又はこれらの少なくとも二つの組み合わせを入力とし、当該発話音声信号に対応する顧客の感情である感情情報を出力するモデルである。
【００３９】
図５Ａは感情認識モデル１０１ｃの学習処理の説明図である。例えば、感情認識モデル１０１ｃの学習には、発話音声信号から抽出される音声特徴量、テキストデータから抽出されるテキスト特徴量、及び、オペレータによる「主観的評価情報」（又は主観的評価情報から抽出される特徴量）の少なくとも一つをそれぞれ含む複数のデータのセット（以下、「データセット」という）を用いてよい。主観的評価情報は、オペレータが顧客の発話音声信号を聞いて顧客の感情を主観で評価した情報である。例えば、怒りレベル１～１０のように、オペレータが複数のレベルで顧客の怒りを評価するものであってもよい。感情認識モデル１０１ｃを学習するためのデータセットは例えば以下のように生成されてもよい。オペレータは、顧客の生の発話音声信号を聞いて、当該発話音声信号から推定される顧客の感情をアノテーションする（すなわち発話音声信号に対して「主観的評価情報」を付与する）。これにより、発話音声信号と当該発話音声信号から推定される顧客の感情とが時間軸上で関連付けされた情報が得られる。複数のオペレータが複数の発話音声信号に対して主観的評価情報の付与を行うことにより、このような情報の束であるデータセットが得られる。感情認識モデル１０１ｃは、このようなデータセットを用いて教師有り機械学習されてもよい。なお、感情認識モデル１０１ｃの学習に用いられるデータセットは、音声特徴量に加えて又は代えて発話音声信号を含んでもよいし、テキスト特徴量に加えて又は代えてテキストデータを含んでもよい。
【００４０】
図５Ｂは感情認識モデル１０１ｃを用いた推定処理の説明図である。例えば、図５Ｂに示すように、発話音声信号Ｓ１から抽出した音声特徴量、及び／又は、当該発話音声信号Ｓ１から生成したテキストデータＴ１から抽出したテキスト特徴量を感情認識モデル１０１ｃに入力することにより、入力に対応する出力、すなわち発話音声信号に対応する感情

情報が得られる。なお、感情認識モデル１０１ｃには、音声特徴量に加えて又は代えて発話音声信号Ｓ１が入力されてもよいし、テキスト特徴量に加えて又は代えてテキストデータＴ１が入力されてもよい。

【００４１】

主観的評価情報は、一以上の感情（例えば、「幸福」、「驚き」、「恐怖」、「怒り」、「嫌悪」及び「悲しみ」の少なくとも一つ等）の度合を数値で示すものであってもよい。又は、感情情報は、顧客が感じている可能性が高い特定の感情（例えば、「怒り」）を示すものであってもよい。

【００４２】

ストレス認識部１０７は、オペレータのストレス状況に関する情報（以下、「ストレス情報」という）を生成する。例えば、ストレス認識部１０７は、オペレータの心拍数、発汗量、呼吸量などのバイタルデータあるいは、カメラを用いて収集したオペレータの視線、表情などの画像情報に基づいて、従来周知の方法によってオペレータのストレス状況を推定してよい。例えば、ストレス認識部１０７は、オペレータによる発話音声に基づいてオペレータのストレス状況を推定してよい。具体的には、ストレス認識部１０７は、オペレータの発話のトーンやスピードの変化、謝罪に関する単語の出現、顧客の発言に被せて発言すること等に基づいて、オペレータのストレス状況を推定してよい。例えば、ストレス認識部１０７は、オペレータ端末２０の操作ログに基づいてオペレータのストレス状況を推定してよい。具体的には、ストレス認識部１０７は、マウス等の動きや、操作すべき場面で操作入力が無いことなどに応じて、オペレータのストレス状況を推定してよい。ストレス認識部１０７は、ストレス認識モデル１０１ｄに基づいてストレス情報を生成してよい。ストレス認識モデル１０１ｄは、発話音声信号、当該発話音声信号から抽出した音声特徴量、当該発話音声信号から生成したテキストデータ、テキスト特徴量又はこれらの少なくとも二つの組み合わせを入力とし、当該発話音声を聞いているオペレータが感じるストレスの推定値を出力するモデルである。ストレス認識モデル１０１ｄの学習には、顧客の発話音声を聞いてオペレータが実際に感じたストレスの実測値を用いてよい。ストレス認識モデル１０１ｄを学習するためのデータセットは例えば以下のように生成されてもよい。オペレータは、顧客の発話音声を聞いて感じたストレスの度合い（例えば１～１０のようなレベル）をアノテーションする（すなわち発話音声信号に対して自身が感じた「ストレスの度合い」を付与する）。これにより、発話音声信号と当該発話音声信号を聞いた際のオペレータのストレスとが時間軸上で関連付けされた情報が得られる。複数のオペレータが複数の発話音声信号に対してストレスの度合いの付与を行うことにより、このような情報の束であるデータセットが得られる。ストレス認識モデル１０１ｄは、このようなデータセットを用いて教師有り機械学習されてもよい。

【００４３】

制御部１０８は、音声処理装置１０に関する種々の制御を行う。具体的には、制御部１０８は、ストレス認識部１０７において生成されるストレス情報に基づいて、オペレータ端末２０において音声合成部１０５で生成された合成音声又は顧客の発話音声のどちらを出力するかを切り替えてもよい。制御部１０８は、発話音声信号に基づいて合成音声信号を生成するか否かをストレス情報に基づいて切り替えてもよい。例えば、制御部１０８は、ストレス情報が示すストレス度数が所定の閾値以上又はより大きい場合、顧客の発話音声ではなく合成音声をオペレータに出力するように制御してもよい。一方、制御部１０８は、ストレス情報が示すストレス度数が所定の閾値より小さい又は以下である場合、発話音声をオペレータに出力するように制御してもよい。制御部１０８は、オペレータから感情抑制機能の自動切り替えについての指示情報が入力された場合、ストレス情報に基づいて上記切り替えを行ってもよい。感情抑制機能とは、顧客の発話音声に代えて合成音声をオペレータに出力する機能である。

【００４４】

制御部１０８は、感情情報に基づいて上記切り替えを行ってもよい。制御部１０８は、当該切り替えを感情抑制切替モデル１０１ｅの出力に基づいて行ってもよい。感情抑制切

替モデル１０１ｅは、発話音声信号、音声特徴量、テキストデータ、テキスト特徴量又はこれらの少なくとも二つの組み合わせを入力として、感情抑制機能のオン・オフを切り替えるタイミングを出力とするモデルである。感情抑制切替モデル１０１ｅは更にストレス情報又は感情情報を入力としてもよい。感情抑制切替モデル１０１ｅの詳細については後述する。

【００４５】

また、制御部１０８は、オペレータによって入力される切り替え情報に基づいて上記切り替えを行ってもよい。ここで、切り替え情報は、顧客の感情抑制機能の適用（オン）又は非適用（オフ）の切り替えに関する情報である。例えば、制御部１０８は、切り替え情報が顧客の感情抑制機能の適用を示す場合、合成音声をオペレータに出力するように制御してもよい。一方、制御部１０８は、切り替え情報が顧客の感情抑制機能の非適用を示す場合、発話音声をオペレータに出力するように制御してもよい。制御部１０８は、オペレータから感情抑制機能の手動切り替えについての指示情報が入力された場合、上記切り替え情報に基づいて上記切り替えを行ってもよい。

【００４６】

学習部１０９は、感情認識モデル１０１ｃ、ストレス認識モデル１０１ｄ及び感情抑制切替モデル１０１ｅの学習処理を行ってよい。

【００４７】

音声処理装置１０は、以下１）乃至７）に示すいずれかの情報、又は、少なくとも二つの情報の組み合わせを時間軸上で関連付け、送受信部１０２を介して、オペレータ端末２０に対して送信してよい。１）顧客の発話音声信号、２）発話音声信号から生成されたテキストデータ、３）除去部１０４の処理を経たあとのテキストデータ、４）検出情報、５）合成音声信号、６）顧客の発話音声信号から推定される顧客の感情情報、７）感情抑制機能のオン・オフを切り替えるタイミング。感情抑制機能がオンである場合、音声処理装置１０は顧客の発話音声信号をオペレータ端末２０に送らなくてもよい。感情抑制機能がオフである場合、音声処理装置１０は合成音声信号をオペレータ端末２０に送らなくてもよい。感情抑制機能のオン・オフに関わらず、音声処理装置１０は顧客の発話音声信号と合成音声信号との両方をオペレータ端末２０に送ってもよい。

【００４８】

≪オペレータ端末≫

【００４９】

図６は、本実施形態に係るオペレータ端末の機能構成の一例を示す図である。オペレータ端末２０は、送受信部２０１、入力受付部２０２、制御部２０３を備える。なお、図６に示す機能構成は一例にすぎず、図示しない他の構成を備えてもよい。

【００５０】

送受信部２０１は、音声処理装置１０及び／又は顧客端末３０との間で、種々の情報及び／又は信号を送信及び／又は受信する。例えば、送受信部２０１は、顧客端末３０で収音された顧客の発話音声の信号である発話音声信号を受信してもよい。送受信部１０２は、音声処理装置１０から、合成音声信号を受信してもよい。また、送受信部２０１は、音声処理装置１０に対して、主観的評価情報を送信してもよい。また、送受信部２０１は、音声処理装置１０から、顧客の感情情報を受信してもよい。

【００５１】

入力受付部２０２は、オペレータによる入力部１０ｅの操作に基づいて、種々の情報の入力を受け付ける。例えば、入力受付部２０２は、感情認識モデル１０１ｃやストレス認識モデル１０１ｄを学習するためのデータセットを生成するための作業の一環として、顧客の生の発話音声信号に対して主観的評価情報やストレスの度合いの入力を受け付けてもよい。以降、オペレータが、オペレータ端末２０において主観的評価情報やストレスの度合いを入力する作業を「アノテーション作業」と呼ぶ。アノテーション作業は、通常のコールセンター業務とは別の業務として位置付けられていてもよい。また、入力受付部２０２は、顧客の感情抑制機能の切り替え情報の入力を受け付けてもよい。また、入力受付部

２０２は、感情抑制機能の手動切り替え又は自動切り替えのどちらかを指示する指示情報の入力を受け付けてもよい。

【００５２】

制御部２０３は、オペレータ端末２０に関する種々の制御を行う。例えば、制御部２０３は、表示部１０ｆにおける情報及び／又は画像の表示を制御する。また、制御部２０３は、音声出力部１０ｉにおける音声の出力を制御する。制御部２０３は、音声処理装置１０から送信される情報に基づいて音声の出力を制御してもよいし、入力受付部２０２が受け付けた情報に基づいて音声の出力を制御してもよい。

【００５３】

制御部２０３は、音声処理装置１０から受信した合成音声信号に基づいて合成音声を音声出力部１０ｉから出力させる。制御部２０３は、顧客端末３０からの発話音声信号に基づいて発話音声を音声出力部１０ｉから出力させてもよい。

【００５４】

また、制御部２０３は、音声処理装置１０から受信した感情情報に基づいて、合成音声信号に対応する感情情報を表示部１０ｆに表示させてもよい。また、制御部２０３は、音声処理装置１０から受信した合成音声信号に対応するテキストデータを表示部１０ｆに表示させてもよい。例えば、制御部２０３は、感情情報、テキストデータ及び検出情報の少なくとも一つを含む画面Ｄ１を表示部１０ｆに表示させてもよい。また、制御部２０３は、ストレス情報を表示部１０ｆに表示させてもよい。例えば、制御部２０３は、ストレス情報を含む画面Ｄ２を表示部１０ｆに表示させてもよい。

【００５５】

図７は、本実施形態に係る画面Ｄ１の一例を示す図である。図７に示すように、画面Ｄ１において、制御部２０３は、音声出力部１０ｉからの合成音声の出力タイミングＴに合わせて、感情情報Ｉ１を表示部１０ｆに表示させてもよい。合成音声の出力タイミングＴ毎に感情情報Ｉ１を表示させることにより、オペレータは、感情抑制機能により顧客の感情が抑制された合成音声を聞く場合でも、顧客の感情をリアルタイムで認識することができる。

【００５６】

また、画面Ｄ１において、制御部２０３は、当該合成音声の出力タイミングＴに合わせて、当該合成音声に対応するテキストデータＩ２の内容を表示部１０ｆに表示させてもよい。テキストデータＩ２の内容を表示させることにより、オペレータは、合成音声だけでなく、視覚的にも顧客の発話内容を把握可能となる。

【００５７】

また、画面Ｄ１では、制御部２０３は、音声処理装置１０から受信した検出情報に基づいて、特定の単語列そのものの表示に代えて、特定の単語列の検出を示す情報Ｉ３（例えば、「ＮＧワード検出」）を表示部１０ｆに表示させてもよい。この機能を「ＮＧワード非表示機能」と呼ぶ。これにより、心理的悪影響を与える顧客の発話の内容をそのままオペレータに認識させるのを回避できるのでオペレータのストレスを抑制できる。また、当該発話があったことはオペレータに通知できるので、オペレータが顧客に対する応対を適切に行うことができる。

【００５８】

また、画面Ｄ１において、制御部２０３は、音声処理装置１０からの感情情報に基づいて、合成音声の出力タイミングＴ毎に、顧客の特定の感情のレベルＩ４を時系列に表示部１０ｆに表示させてもよい。例えば、図７では、合成音声の出力タイミングＴ毎の顧客の「怒り」のレベルＩ４が折れ線グラフで示される。これにより、オペレータが顧客の特定の感情（例えば、「怒り」）の遷移を容易に把握できるので、顧客に対するオペレータの応対の満足度を向上できる。

【００５９】

画面Ｄ１において、制御部２０３は選択ボタンＩ５を表示部１０ｆに表示させてもよい。選択ボタンＩ５は、感情抑制機能の適用（オン）又は非適用（オフ）を自動又は手動のど

ちらで切り替えるかをオペレータが選択可能とするインターフェースである。オペレータは選択ボタンＩ５に対してクリック、タップ又はスライド等の操作を行うことにより「自動切換モード」と「手動切替モード」を切り替えることができる。自動切換モードにおいては、例えば感情情報、ストレス情報、又は感情抑制切替モデル１０１ｅからの出力等に基づいて感情抑制機能のオン・オフが自動で切り替わる。

【００６０】

「手動切替モード」が選択された場合、制御部２０３は、感情抑制機能の適用又は非適用をオペレータが選択可能とするインターフェースである切替ボタンＩ６を表示部１０ｆに表示させてよい。オペレータが感情抑制機能のオンとオフを切り替えたタイミングは、顧客の発話音声（及び/又は発話音声に基づいて抽出される各種特徴量）と時間軸上で関連付けされて「手動切替履歴データ」として不図示の記憶部に蓄積される。「手動切替履歴データ」には更にオペレータの識別情報が関連付けられてもよい。

【００６１】

切り替えボタンＩ７は、「ＮＧワード非表示機能」のオン・オフを切り替えるためのボタンである。「ＮＧワード非表示機能」がオフの場合には、テキストデータＩ２の内に特定の単語列が検出された場合でも、除去部１０４による処理が行われる前のテキストデータＩ２がそのまま表示部１０ｆに表示される。感情抑制機能をオンにしつつＮＧワード非表示機能をオフにした場合、オペレータは顧客による特定の単語列を直接聞くことは無いのでストレスが軽減される一方で、顧客の発話内容を正確に把握することにより顧客の感情をより正確に把握することができる。

【００６２】

感情抑制切替モデル１０１ｅを学習するためのデータセットは、ストレス情報、感情情報、発話音声信号Ｓ１、音声特徴量、テキストデータ、テキスト特徴量又はこれらの少なくとも二つの組み合わせと、オペレータが感情抑制機能のオン・オフを切り替えたタイミングとが、時間軸上で関連付けされたデータの束であってよい。感情抑制切替モデル１０１eを学習する方法は、例えば下記１）から３）に述べるような様々な方法がある。１）感情抑制切替モデル１０１ｅはオペレータ毎に学習されてもよい。すなわち、或るオペレータに対して適用される感情抑制切替モデル１０１ｅは、そのオペレータによる感情抑制機能の「手動切替履歴データ」のみに基づいて学習されてもよい。この方法によれば、感情抑制切替モデル１０１ｅはそのオペレータの好みに合わせたタイミングで感情抑制機能を切り替えることができるようになる。あるいは、２）或るオペレータに対して適用される感情抑制切替モデル１０１ｅは、不特定多数のオペレータによる「手動切替履歴データ」に基づいて学習されてもよい。この方法によれば、学習に用いることができるデータが多くなるため、感情抑制切替モデル１０１ｅを早く学習することができるようになる。あるいは、３）或るオペレータに対して適用される感情抑制切替モデル１０１ｅは、そのオペレータと年齢・性別・その他の特性が類似したオペレータによる「手動切替履歴データ」に基づいて学習されてもよい。この方法によれば、１）の方法と比較して学習に用いることができるデータが多くなるため感情抑制切替モデル１０１ｅを早く学習することができ、２）の方法と比較して自分の好みに合った切替タイミングを学習することができるようになる。

【００６３】

図８は、本実施形態に係る画面Ｄ２の一例を示す図である。画面Ｄ２において、制御部２０３は、音声処理装置１０からのストレス情報を表示させてもよい。例えば、図８では、ストレス情報として、オペレータが感じるストレスの推定値を示す情報（例えば、「５６％」）と、当該オペレータの平常時の状態からの相対的な評価値を示す情報（例えば、「平常時より８．１％減」）とが表示される。

【００６４】

図１２は、本実施形態に係る画面Ｄ３の一例を示す図である。画面Ｄ３において、制御部２０３は、オペレータがアノテーション作業を行うためのインターフェースＩ８を表示させてもよい。オペレータは、例えば、顧客の生の音声（サンプル音声）を聞きながら、

サンプル音声から感じられる顧客の感情をインターフェースＩ８から都度選択する。図１２において、顧客感情Ｉ１はオペレータによる顧客感情の主観的評価情報である。例えば、オペレータが、サンプル音声「今日の夕方までにどうにかして届けてよ」に対して「怒り」という感情をアノテーションしたならば、図１２に示すように、「今日の夕方までにどうにかして届けてよ」というサンプル音声と「怒り」という情報が時間軸上で関連付けられる。アノテーションは文単位で行われてもよいし所定の時間間隔ごとに行われてもよい。

【００６５】

（音声処理システムの動作）

図９は、本実施形態に係る感情抑制動作の一例を示すフローチャートである。なお、図９は、例示にすぎず、少なくとも一部のステップ（例えば、ステップＳ１０６）の順番は入れ替えられてもよいし、不図示のステップが実施されてもよいし、一部のステップが省略されてもよい。

【００６６】

音声処理装置１０は、顧客端末３０の音声入力部１０ｈで収音される顧客の発話音声の信号である発話音声信号を取得する（Ｓ１０１）。

【００６７】

音声処理装置１０は、Ｓ１０１で取得された発話音声信号に基づいて抽出される特徴量を音声認識モデル１０１ａに入力して、一以上の単語からなる単語列を含むテキストデータを生成する（Ｓ１０２）。

【００６８】

音声処理装置１０は、Ｓ１０２で生成されたテキストデータ内に特定の単語列が含まれるか否かを判定する（Ｓ１０３）。当該テキストデータ内に特定の単語列が含まれる場合、音声処理装置１０は、当該特定の単語列を除去又は前記特定の単語列を他の単語列に変換したテキストデータを生成する（Ｓ１０４）。

【００６９】

音声処理装置１０は、テキストデータに基づいて抽出される特徴量を音声合成モデル１０１ｂに入力して、合成音声の信号である合成音声信号を生成する（Ｓ１０５）。

【００７０】

音声処理装置１０は、Ｓ１０１で取得された発話音声信号、Ｓ１０２で生成されたテキストデータ、及び、オペレータによって入力される顧客の感情の主観的評価情報の少なくとも一つに基づいて抽出される特徴量を感情認識モデル１０１ｃに入力して、顧客の感情情報を生成する（Ｓ１０６）。

【００７１】

オペレータ端末２０は、Ｓ１０５で生成された合成音声信号に基づいて合成音声を音声出力部１０ｉから出力させるとともに、当該合成音声の出力タイミングＴに合わせて当該合成音声に対応する感情情報を表示部１０ｆに表示させる（Ｓ１０７、例えば、図７）。

【００７２】

音声処理装置１０は、処理を終了するか否かを判定する（Ｓ１０８）。処理を終了しない場合（Ｓ１０８：ＮＯ）、音声処理装置１０は、処理Ｓ１０１～Ｓ１０７を再び実行する。一方、音声変換処理を終了する場合（Ｓ１０８：ＹＥＳ）、音声処理装置１０は、処理を終了する。

【００７３】

図１０は、本実施形態に係る感情抑制機能の自動切り替え動作を示すフローチャートである。なお、図１０は、例示にすぎず、少なくとも一部のステップの順番は入れ替えられてもよいし、不図示のステップが実施されてもよいし、一部のステップが省略されてもよい。

【００７４】

音声処理装置１０は、オペレータのストレス情報を生成する（Ｓ２０１）。

【００７５】

音声処理装置１０は、ストレス情報が所定の条件を満たすか否かを判定する（Ｓ２０２）。例えば、所定の条件は、ストレス情報が示すストレス度数が所定の閾値以上又はより大きいことであってもよい。

【００７６】

音声処理装置１０は、ストレス情報が所定の条件を満たす場合（Ｓ２０２：ＹＥＳ）、感情抑制機能を適用（すなわち、オペレータ端末２０から合成音声を出力）してもよい（Ｓ２０３）。一方、音声処理装置１０は、ストレス情報が所定の条件を満たさない場合（Ｓ２０２：ＮＯ）、感情抑制機能を非適用（すなわち、オペレータ端末２０から顧客の発話音声を出力）してもよい（Ｓ２０４）。

【００７７】

音声処理装置１０は、処理を終了するか否かを判定する（Ｓ２０５）。処理を終了しない場合（Ｓ２０５：ＮＯ）、音声処理装置１０は、処理Ｓ２０１～Ｓ２０４を再び実行する。一方、音声変換処理を終了する場合（Ｓ２０５：ＹＥＳ）、音声処理装置１０は、処理を終了する。なお、Ｓ２０１及びＳ２０２において、音声処理装置１０は、感情情報や感情抑制切替モデル１０１ｅの出力に基づいて、感情抑制機能を適用するか否を決定してもよい。

【００７８】

以上のように、本実施形態に係る音声処理システム１によれば、顧客の発話音声信号に基づいてテキストデータを生成し、当該テキストデータに基づいて生成される合成音声をオペレータに出力する。このため、顧客の発話音声に含まれる顧客の感情を十分に抑制した合成音声をオペレータに聞かせることができ、顧客の感情的発話に起因するオペレータのストレスを軽減できる。本発明の発明者は、約５０名の被験者に対して、１）顧客の発話音声そのもの、２）顧客の発話音声の音量を調整した音声、３）顧客の発話音声の声質を変換した音声、４）顧客の発話音声をテキスト化してから生成した合成音声、の４種類の音声を聞き比べてもらい、音声から感じられる怒りの度合いを７段階の尺度で評価してもらう実験を行った。その結果、２）や３）と比較して４）が、被験者に伝わった怒りの軽減度合いが顕著であった。

【００７９】

また、本実施形態に係る音声処理システム１によれば、オペレータに対して、合成音声を出力するだけでなく顧客の感情情報を合成音声出力のタイミングに合わせて通知することができるので、合成音声を聞いたオペレータが顧客の感情をリアルタイムに認識でき、顧客に対して適切な応対を行うことができる。

【００８０】

また、本実施形態に係る音声処理システム１によれば、オペレータのストレス情報又は顧客の感情情報等に基づいて、感情抑制機能を適用するか否か（すなわち、オペレータに対して合成音声又は発話音声のどちらを出力するか）が切り替えられるので、オペレータのストレスと顧客の満足度とのバランスを適切に図ることができる。

【００８１】

（変更例）

上記音声処理システム１では、音声認識部１０３は、発話音声信号から、一つ又は複数の文として確定された単語列を含むテキストデータを生成したが、これに限られない。音声認識部１０３は、発話音声信号から認識された単語列が一つ又は複数の文として確定される前に、一つ又は複数の単語（品詞又は形態素）からなる単語列を含むテキストデータを生成してもよい。除去部１０４は、当該文として確定されていないテキストデータ内の特定の単語列を除去し、音声合成部１０５は、当該文として確定されていないテキストデータから合成音声信号を生成してもよい。

【００８２】

図１１は、本実施形態の変更例に係る合成音声信号の生成の一例を示す図である。図１１では、送受信部１０２で取得された発話音声信号Ｓ４に基づいて、音声認識部１０３においてテキストデータＴ４１～Ｔ４３が生成されるものとする。図１１に示すように、テ

キストデータＴ４１～Ｔ４３は、「はやく送ってください」という一文の確定前に、意味を持つ形態素単位（「はやく」、「送って」、「ください」）でテキストデータが生成される点で、図４と異なる。除去部１０４は、テキストデータＴ４１～Ｔ４３それぞれに対して特定の単語列が含まれるか否かを判定して、当該特定の単語列を除去して音声合成部１０５に出力する。音声合成部１０５は、テキストデータＴ４１～Ｔ４３からそれぞれ合成音声信号Ｓ４１～Ｓ４３を生成する。

【００８３】

図１１に示すように、文の確定前に一つ又は複数の形態素単位でテキストデータを生成して合成音声を出力することにより、テキストデータの生成によりオペレータの応答遅延を軽減できる。なお、形態素単位での複数のテキストデータ（又は合成音声）が意味的に不自然でないかを判定するモデルなどが用いられてもよい。

【００８４】

また、応答遅延を軽減するために、図４に示す合成音声信号Ｓ１～Ｓ３、図１１に示す合成音声信号Ｓ４１～Ｓ４３それぞれの前及び／又は後に、例えば、「あ～」、「え～」、「まあ」等のフィラー音が追加されてもよい。これにより、オペレータも応答遅延による顧客の満足度の低下を防止できる。

【００８５】

また、音声合成部１０５は、感情認識部１０６が推定した顧客の感情に基づいて、複数の音声合成モデル１０１ｂのうちから顧客の感情に合った音声合成モデル１０１ｂを選択してもよい。例えば、感情認識部１０６が推定した顧客の感情が「激昂」である場合、音声合成部１０５は、ピッチが速く抑揚が激しい音声合成モデル１０１ｂを用いてよい。例えば、感情認識部１０６が推定した顧客の感情が「号泣」である場合、音声合成部１０５は、泣き声のような音声を出力する音声合成モデル１０１ｂを用いてよい。或いは、音声合成部１０５は、感情認識部１０６が推定した顧客の感情に基づいて音声合成モデル１０１ｂのパラメータを変更し、顧客の感情に合った音声が出力されるように調整してよい。顧客が激昂している際の生の音声を直接聞いたオペレータは極めて強いストレスを感じてしまう。他方、オペレータは顧客対応業務を適切に遂行するために、顧客の感情をリアルタイムで正確に把握する必要がある。オペレータに発話音声を直接聞かせないことによりオペレータは過剰なストレスを感じることがなく、合成音声に顧客の感情を乗せることにより、オペレータは聴覚を通じて顧客の感情をリアルタイムに把握することができる。

【００８６】

（その他の実施形態）

上記実施形態では、顧客の発話音声信号をテキスト化して、合成音声信号をオペレータに出力するものとしたがこれに限られない。音声処理装置１０は、顧客の発話音声信号に基づいて抽出される音声特徴量を音声変換モデルに入力して、変換音声の信号を生成し、オペレータ端末２０から変換音声を出力してもよい。

【００８７】

特許請求の範囲に記載の「音声変換モデル」は、発話音声信号を一旦テキスト化して合成音声として出力するモデルと、発話音声信号をテキスト化せずに声質を変換させて出力するモデルとの両方を包含する概念である。顧客の発話音声に代えて合成音声または変換音声をオペレータに対して出力することにより、効果の程度の差こそあれ、オペレータが感じるストレスを軽減できる。他方で、顧客対応業務の遂行のためには、オペレータが顧客の感情をリアルタイムに把握することも欠かせない。

【００８８】

本変形例における音声処理装置１０は、顧客の発話音声信号に基づいて抽出される音声特徴量を音声変換モデルに入力して、変換音声信号を生成する。音声処理装置１０は、１）変換音声信号と、２）顧客の発話音声から推定される顧客の感情情報とを時間軸上で関連付けた情報を生成し、オペレータ端末２０に対して送信する。音声処理装置１０が送信する情報には、発話音声信号、発話音声信号から生成されたテキストデータ、除去部１０４の処理を経たあとのテキストデータ、検出情報、感情抑制機能のオン・オフを切り替え

るタイミングが関連付けされていてもよい。
【００８９】
オペレータ端末２０は、音声処理装置１０から受信した変換音声の信号を音声出力部１０ｉから出力し、且つ、音声出力部１０ｉからの変換音声の出力タイミングＴに合わせて、感情情報を示す情報を表示部１０fに表示してよい。オペレータ端末２０は更に、音声出力部１０ｉからの変換音声の出力タイミングＴに合わせて、テキストデータを表示部１０fに表示してよい。かかる表示の態様は図７に図示するようであってよい。
【００９０】
本変形例における音声処理装置１０は、感情情報に基づいて、感情情報が示す感情が変換音声に反映されるように、変換音声の信号を生成してもよい。例えば感情情報が示す感情が「激昂」である場合、ピッチが速く抑揚が激しい音声変換モデルを用いてよい。例えば感情情報が示す感情が「号泣」である場合、泣き声のような音声を出力する音声変換モデルを用いてよい。音声処理装置１０は、感情情報が示す感情が変換音声に反映されるように、変換音声の信号を生成してよい。オペレータに発話音声を直接聞かせないことによりオペレータは過剰なストレスを感じるがことなく、変換音声に顧客の感情を乗せることにより、オペレータは聴覚を通じて顧客の感情をリアルタイムに把握することができる。
【００９１】
本変形例における音声処理システム１においては、オペレータによるアノテーション作業は、オペレータによる通常のコールセンター業務中において、変換音声に対して行われても良い。オペレータが変換音声に対して「怒りの感情」をアノテーションした場合、当該アノテーションの結果に基づいて、音声変換モデルがより柔らかい音声を出力するようにリアルタイムに調整されてもよい。
【００９２】
以上説明した実施形態では、第１のユーザが顧客であり、第２のユーザがオペレータであるコールセンターを想定したが、本実施形態の適用場面はコールセンターに限られない。例えば、Ｗｅｂミーティング等、第１のユーザの感情を抑制した音声を第２のユーザに出力するどのような場面にも適用可能である。すなわち、本実施形態は、カスタマーハラスメント対策だけでなく、社内のパワーハラスメント等、様々なハラスメントに対する企業側の対策として利用可能である。
【００９３】
以上説明した実施形態における、感情情報と合成音声とを「時間軸上で関連付け」する処理は、図７に示すように、合成音声または変換音声の出力タイミングに合わせて、それらの元となった発話音声から推定される感情情報を表示することが実現可能な態様であれば、その具体的な態様を問わない。以上説明した実施形態における「時間軸上で関連付け」する処理は、何時何分何秒といった時刻情報に基づいて関連付けする処理でも良いし、発話音声情報の開始から何分何秒経過時といった情報に基づいて関連付けする処理でも良いし、文単位、単語単位又は形態素単位で関連付けする処理でもよい。
【００９４】
以上説明した実施形態における音声処理システム１において、顧客からは、自身の音声が感情抑制されてオペレータに届いていることが分からないようにしてもよい。すなわち、感情抑制機能がオンになっているかオフになっているかは、顧客からは把握できないようにしてもよい。
【００９５】
アノテーション作業は、オペレータがオペレータ端末２０上で行っても良いし、別途、アノテーション作業用の専用のアプリケーションや端末が用意されていてもよい。
【００９６】
また、以上説明した実施形態は、本発明の理解を容易にするためのものであり、本発明を限定して解釈するためのものではない。実施形態が備える各要素並びにその配置、材料、条件、形状及びサイズ等は、例示したものに限定されるわけではなく適宜変更することができる。また、異なる実施形態で示した構成同士を部分的に置換し又は組み合わせるこ

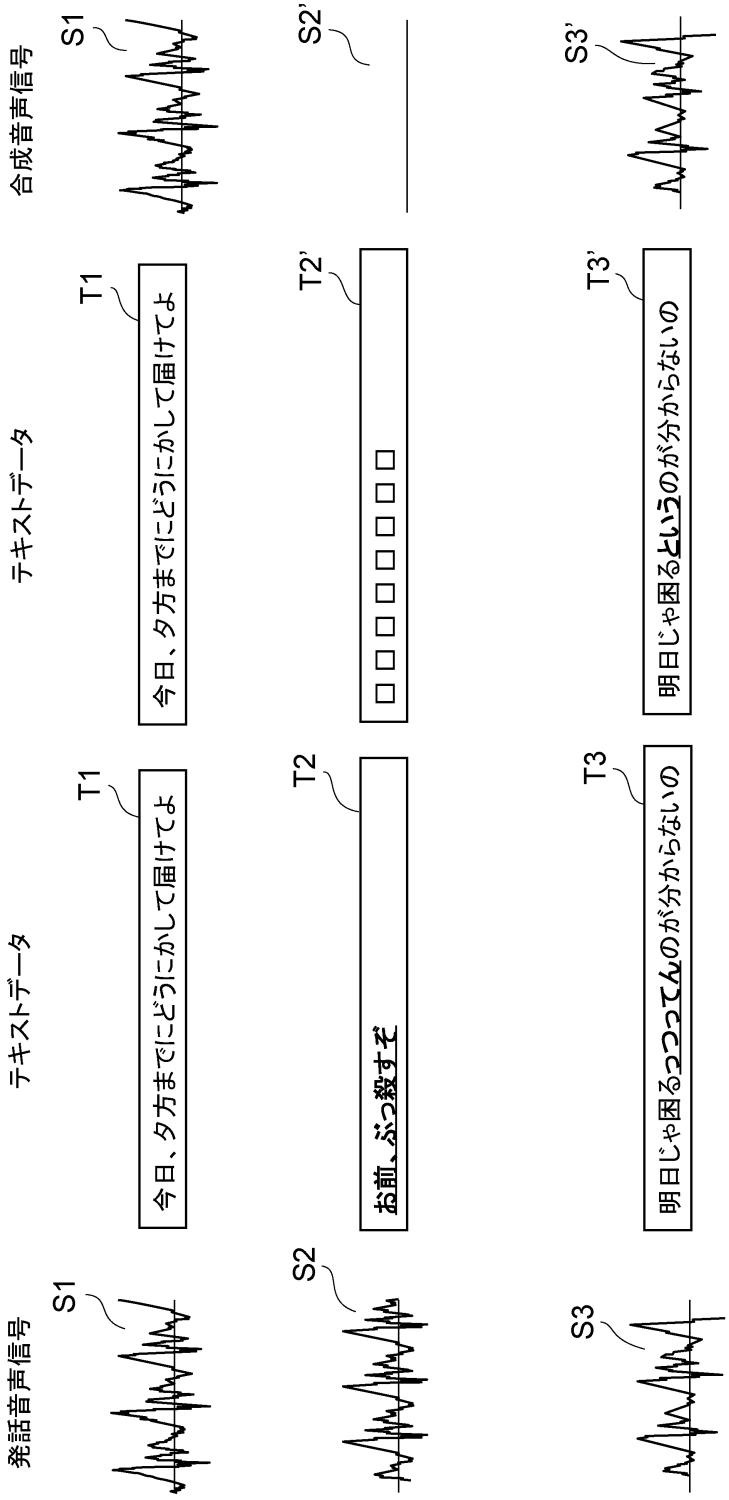
とが可能である。また、音声処理装置１０の機能として記載した機能をオペレータ端末２０が備えていてもよい。また、オペレータ端末２０の機能として記載した機能を音声処理装置１０が備えていてもよい。

【符号の説明】

【００９７】

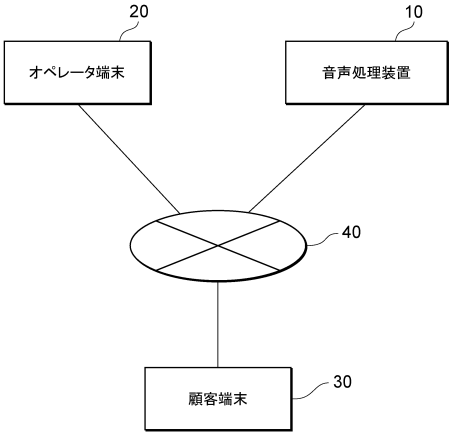
１…音声処理システム、１０…音声処理装置、２０…オペレータ端末、３０…顧客端末、１０ａ…プロセッサ、１０ｂ…ＲＡＭ、１０ｃ…ＲＯＭ、１０ｄ…通信部、１０ｅ…入力部、１０ｆ…表示部、１０ｇ…カメラ、１０ｈ…音声入力部、１０ｉ…音声出力部、１０１…記憶部、１０２…送受信部、１０３…音声認識部、１０４…除去部、１０５…音声合成部、１０６…感情認識部、１０７…ストレス認識部、１０８…制御部、１０９…学習部、２０１…送受信部、２０２…入力受付部、２０３…制御部

【要約】
【課題】聞き手のストレスの軽減を可能とすること。
【解決手段】音声処理システム１は、第１のユーザの発話音声の信号である発話音声信号を取得する取得部と、前記発話音声信号に基づいて抽出される特徴量を音声認識モデルに入力して、一以上の単語からなる単語列を含むテキストデータを生成する音声認識部と、前記テキストデータに基づいて抽出される特徴量を音声合成モデルに入力して、合成音声の信号である合成音声信号を生成する音声合成部と、第２のユーザに対して前記合成音声を出力する音声出力部と、を備える。
【選択図】図４

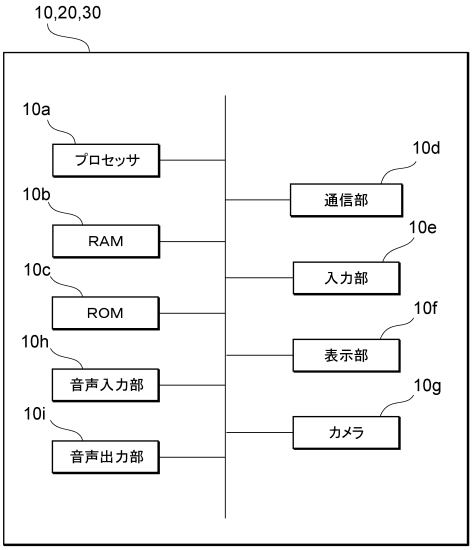


【図面】

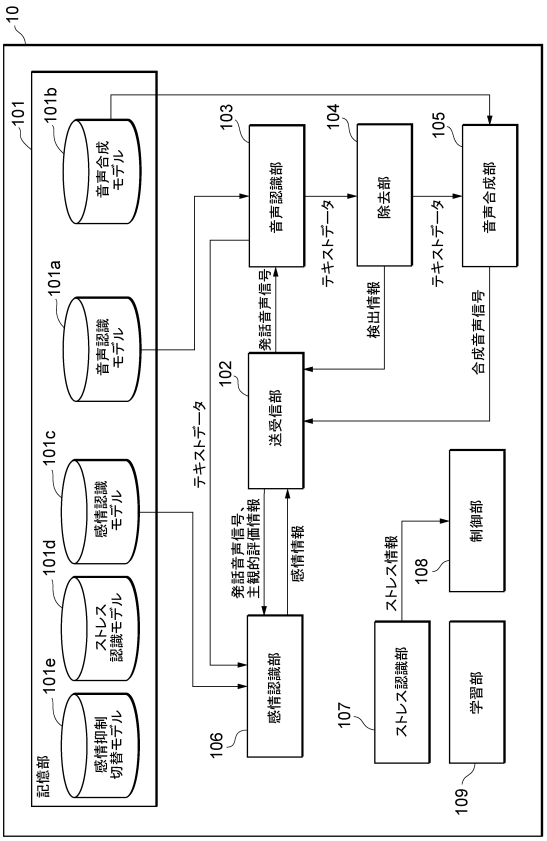
【図１】


20
10
オペレータ端末
音声処理装置
40
顧客端末
30

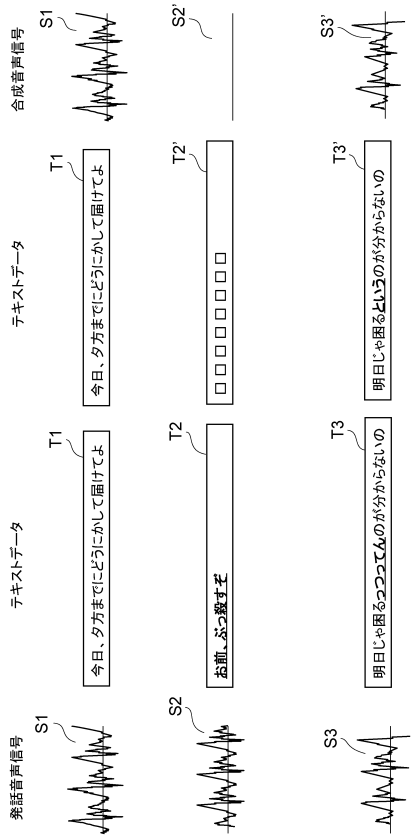
【図２】


10,20,30
10a
プロセッサ
10b
RAM
10c
ROM
10h
音声入力部
10i
音声出力部
10d
通信部
10e
入力部
10f
表示部
10g
カメラ

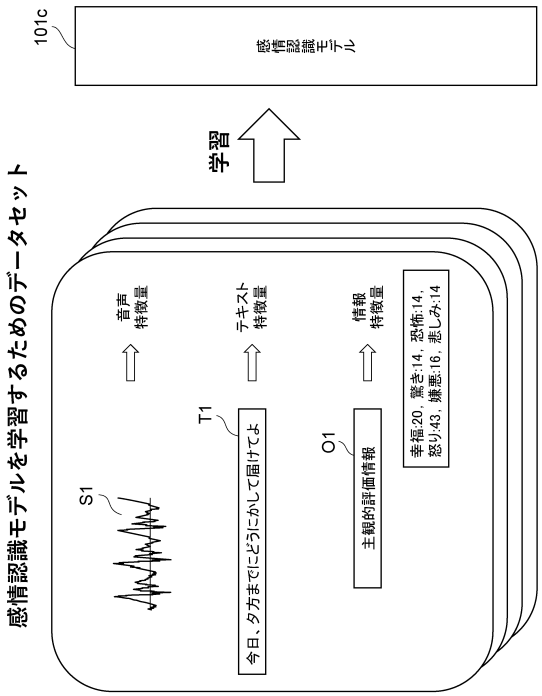
【図３】


10
101
記憶部
101b
音声合成モデル
101a
音声認識モデル
101c
感情認識モデル
101d
ストレス認識モデル
101e
感情抑制切替モデル
103
音声認識部
104
除去部
105
音声合成部
102
送受信部
106
感情認識部
107
ストレス認識部
108
制御部
109
学習部
テキストデータ
発話音声信号
検出情報
合成音声信号
発話音声信号、主観的評価情報
感情情報
ストレス情報

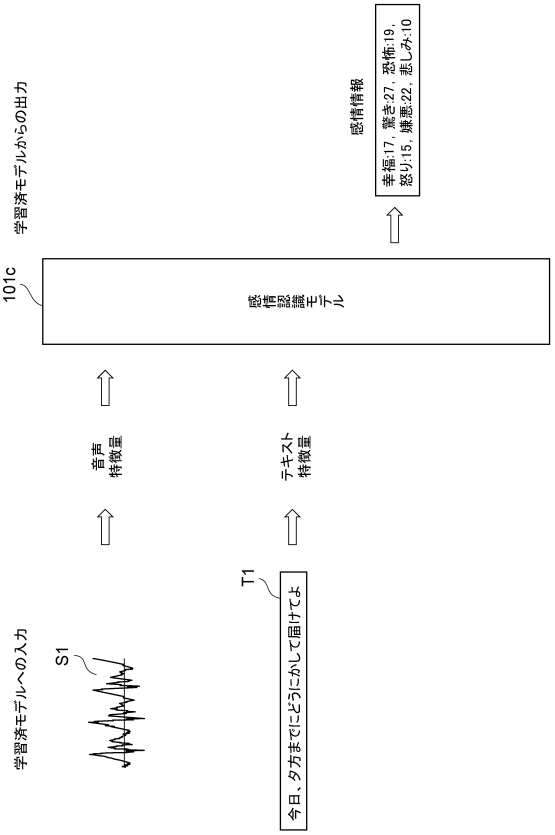
【図４】


合成音声信号
S1
S2'
S3'
テキストデータ
T1
今日、夕方までにどうにかして届けてよ
T2'
□□□□□□□□
T3'
明日じゃ困るというのが分からないの
テキストデータ
T1
今日、夕方までにどうにかして届けてよ
T2
お前、ぶっ殺すぞ
T3
明日じゃ困るっつってんのが分からないの
発話音声信号
S1
S2
S3

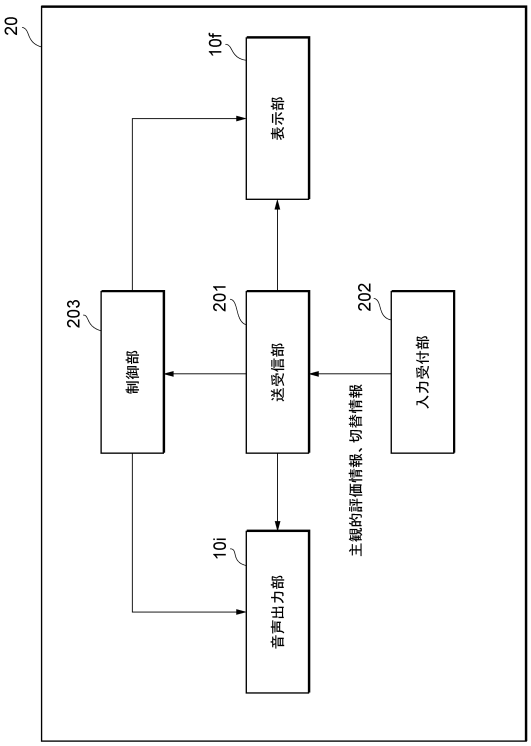
【図５Ａ】


感情認識モデルを学習するためのデータセット
S1
T1
今日、夕方までにどうにかして届けてよ
O1
主観的評価情報
幸福:20，驚き:14，恐怖:14，
怒り:43，嫌悪:16，悲しみ:14
音声
特徴量
テキスト
特徴量
情報
特徴量
学習
101c
感情認識モデル

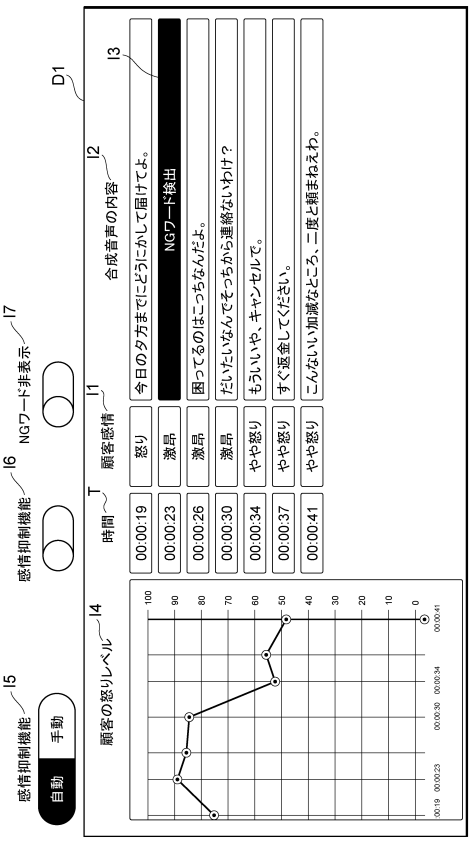
【図５Ｂ】


学習済モデルへの入力
S1
T1
今日、夕方までにどうにかして届けてよ
音声
特徴量
テキスト
特徴量
101c
感情認識モデル
学習済モデルからの出力
感情情報
幸福:17，驚き:27，恐怖:19，
怒り:15，嫌悪:22，悲しみ:10

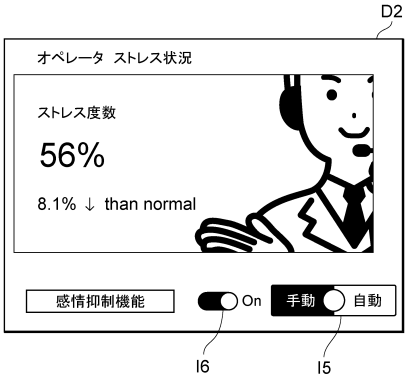
【図６】


20
10f
表示部
203
制御部
201
送受信部
202
入力受付部
10i
音声出力部
主観的評価情報、切替情報

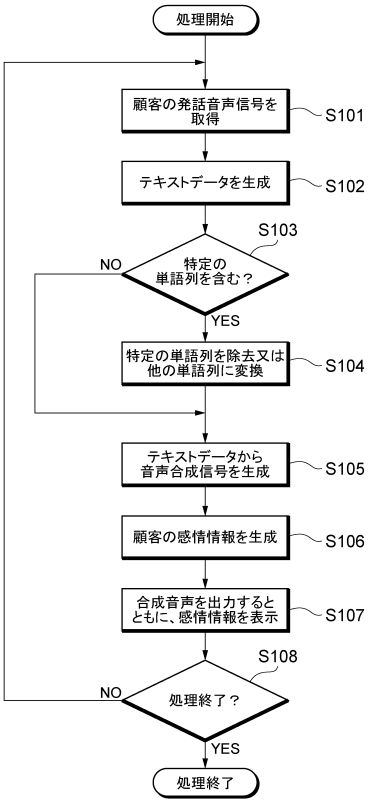
【図７】


D1
I3
I2
合成音声の内容
I1
顧客感情
T
時間
I7
NGワード非表示
I6
感情抑制機能
I5
感情抑制機能
自動
手動
I4
顧客の怒りレベル
00:00:19 怒り 今日の夕方までにどうにかして届けてよ。
00:00:23 激昂 NGワード検出
00:00:26 激昂 困ってるのはこっちなんだよ。
00:00:30 激昂 だいたいなんでそっちから連絡ないわけ？
00:00:34 やや怒り もういいや、キャンセルで。
00:00:37 やや怒り すぐ返金してください。
00:00:41 やや怒り こんないい加減なところ、二度と頼まねえわ。

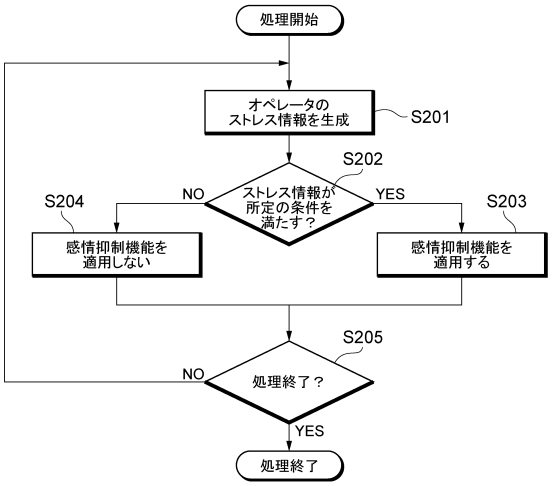
【図８】


D2
オペレータ ストレス状況
ストレス度数
56%
8.1% ↓ than normal
感情抑制機能
On
手動
自動
I6
I5

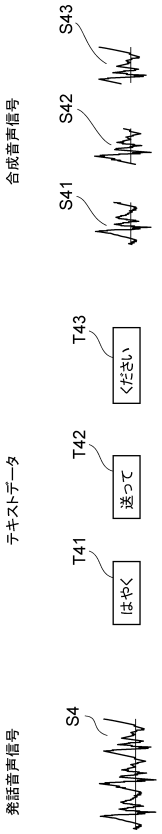
【図９】


処理開始
顧客の発話音声信号を取得
S101
テキストデータを生成
S102
S103
特定の単語列を含む？
NO
YES
特定の単語列を除去又は他の単語列に変換
S104
テキストデータから音声合成信号を生成
S105
顧客の感情情報を生成
S106
合成音声を出力するとともに、感情情報を表示
S107
S108
処理終了？
NO
YES
処理終了

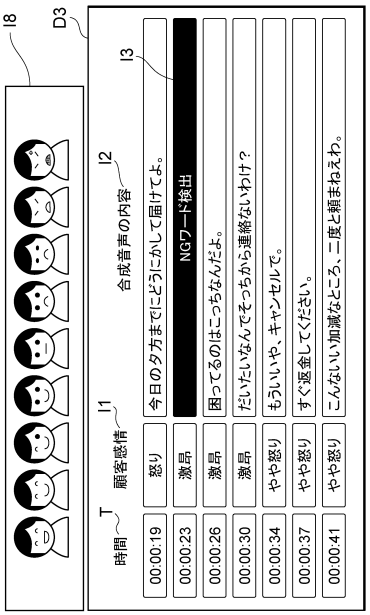
【図１０】


処理開始
オペレータのストレス情報を生成
S201
S202
ストレス情報が所定の条件を満たす？
NO
YES
S204
S203
感情抑制機能を適用しない
感情抑制機能を適用する
S205
処理終了？
NO
YES
処理終了

【図１１】


合成音声信号
S43
S42
S41
テキストデータ
T43
ください
T42
送って
T41
はやく
発話音声信号
S4

【図１２】


I8
D3
T
時間
I1
顧客感情
I2
合成音声の内容
I3
00:00:19
怒り
今日の夕方までにどうにかして届けてよ。
00:00:23
激昂
NGワード検出
00:00:26
激昂
困ってるのはこっちなんだよ。
00:00:30
激昂
だいたいなんでそっちから連絡ないわけ？
00:00:34
やや怒り
もういいや、キャンセルで。
00:00:37
やや怒り
すぐ返金してください。
00:00:41
やや怒り
こんないい加減なところ、二度と頼まねえわ。

フロントページの続き

(72)発明者　今村 俊雄
東京都港区海岸一丁目７番１号
(72)発明者　阪下 啓祐
東京都港区海岸一丁目７番１号
(72)発明者　▲高▼原 周平
東京都港区海岸一丁目７番１号
審査官　堀 洋介
(56)参考文献　特開２００１－１１７７５２（ＪＰ，Ａ）
特開２０２０－０２１０２５（ＪＰ，Ａ）
特開２０１０－１６６３２４（ＪＰ，Ａ）
国際公開第２０１９／１１１３４６（WO，Ａ１）
特開２０１９－１１０４５１（ＪＰ，Ａ）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１３／００－１５／３４
Ｇ０６Ｆ　３／１６