

(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11月2日(02.11.2023)



(10) 国際公開番号

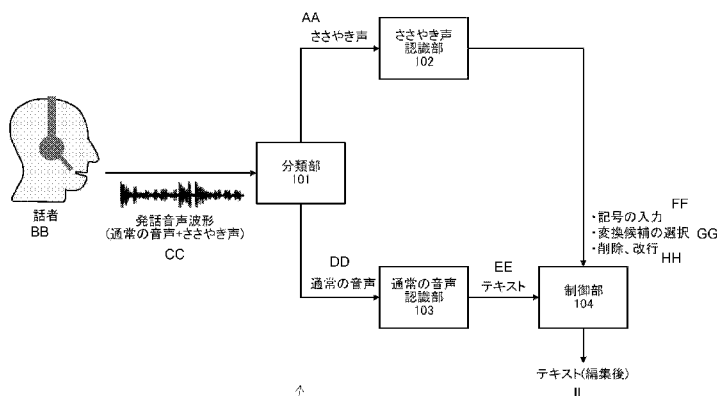
WO 2023/210149 A1

- (51) 国際特許分類:
G10L 15/32 (2013.01) G10L 25/51 (2013.01)
- (21) 国際出願番号: PCT/JP2023/007479
- (22) 国際出願日: 2023年3月1日(01.03.2023)
- (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 2022-072123 2022年4月26日(26.04.2022) 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: 暦本 純一 (REKIMOTO, Junichi); 〒1410022 東京都品川区東五反田 3 丁目 1 4 番 1 3 号 株式会社ソニーコンピュータサイエンス研究所株式会社内 Tokyo (JP).
- (74) 代理人: 宮田 正昭, 外(MIYATA, Masaaki et al.); 〒1040032 東京都中央区八丁堀三丁目 2 5 番 9 号 D a i w a 八丁堀駅前ビル西館 8 階 特許業務法人 大同特許事務所 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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE,

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

(54) 発明の名称: 情報処理装置及び情報処理方法、並びにコンピュータプログラム

【図1】



音声入力システム100

- 100 Speech input system
101 Classification unit
102 Whispered speech recognition unit
103 Normal speech recognition unit
104 Control unit
AA Whispered speech
BB Speaker
CC Uttered speech waveform (normal speech + whispered speech)
DD Normal speech
EE Text
FF Symbol input
GG Selection of conversion candidate
HH Delete, line break
II Text (edited)

(57) Abstract: Provided is an information processing device that carries out a process related to speech input. This information processing device comprises: a classification unit for classifying uttered speech as normal speech or whispered speech on the basis of a speech feature amount; a recognition unit for recognizing the whispered speech classified by the classification unit; and a control unit for controlling a process on the basis of a recognition result from the recognition unit. This information processing device also comprises a normal speech recognition unit for recognizing the normal speech

KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, 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, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類：

一 国際調査報告 (条約第21条(3))

classified by the classification unit, and the control unit executes a process corresponding to the recognition result of the whispered speech by the recognition unit on the recognition result from the normal speech recognition unit.

(57) 要約：音声入力に関する処理を行う情報処理装置を提供する。情報処理装置は、発話音声をも音声特徴量に基づいて通常の音声とささやき声に分類する分類部と、前記分類部によって分類されたささやき声を認識する認識部と、前記認識部の認識結果に基づく処理を制御する制御部を具備する。情報処理装置は、前記分類部によって分類された通常の音声を認識する通常の音声認識部をさらに備え、前記制御部は、前記通常の音声認識部の認識結果に対して前記認識部によるささやき声の認識結果に対応する処理を実施する。

明 細 書

発明の名称：

情報処理装置及び情報処理方法、並びにコンピュータプログラム

技術分野

[0001] 本明細書で開示する技術（以下、「本開示」とする）は、音声の処理を行う情報処理装置及び情報処理方法、並びにコンピュータプログラムに関する。

背景技術

[0002] 音声認識は多く利用されている。音声認識は発話内容をテキストに変換する文字起こしなどにも利用されているが、元の発話内容が文字のみからなるとは限らず、例えば句読点や引用符などの記号や、入力テキストに対する編集コマンドを含む場合（又は、話者が記号の入力や編集コマンドを指示した場合）がある。しかしながら、通常の音声認識技術では、テキストの入力と、記号の入力や編集コマンドを区別することが難しい。例えば、「山」を「かぎっこ やま かぎっこことじる」と発声して入力すると、それがそのまま入力されてしまうかもしれない。「1行削除」のようなコマンドも、そのままテキストとして入力されてしまう。

[0003] かぎっこ と発声された場合には「」を入力する、というように、頻繁に使用する記号を入力するためのルールあらかじめ決めておくことも考えられるが、今度は 鉤括弧 と入力したいときに手段がないことになる。また、どの発話がテキストと入力され、どの発話は記号やコマンドとして入力されるのかをあらかじめ把握しなければならないので、利用者の負担になる。

[0004] 例えば、対話状況特徴に基づいて複数の文末記号付与規範の中からその対話が行われた状況に対応する文末記号付与規範を選択し、選択された文末記号付与規範、その対話の音響特徴及び言語特徴を用いてその対話の音声認識結果テキストに対する文末記号を推定する装置が提案されているが（特許文

献 1 を参照のこと）、この装置は、音声認識されたテキストに句読点を付与することくらいしかできず、鉤括弧などテキスト中に記号を入力することはできない。

発明の概要

発明が解決しようとする課題

[0005] 本開示の目的は、音声入力に関する処理を行う情報処理装置及び情報処理方法、並びにコンピュータプログラムを提供することにある。

課題を解決するための手段

[0006] 本開示は、上記課題を参酌してなされたものであり、その第 1 の側面は、発話音声をも音声特徴量に基づいて通常の音声とささやき声に分類する分類部と、

前記分類部によって分類されたささやき声を認識する認識部と、

前記認識部の認識結果に基づく処理を制御する制御部と、

を具備する情報処理装置である。

[0007] 前記分類部は、第 1 の学習済みニューラルネットワークを用いて通常の音声とささやき声の分類を行い、前記認識部は、第 2 の学習済みニューラルネットワークを用いてささやき声の認識を行うように構成されている。例えば、前記第 2 の学習済みニューラルネットワークは特徴抽出層及びトランスフォーマ層を含み、前記第 1 の学習済みニューラルネットワークは前記第 2 の学習済みニューラルネットワークと前記特徴抽出層を共有して構成される。

[0008] 第 1 の側面に係る情報処理装置は、前記分類部によって分類された通常の音声を認識する通常の音声認識部をさらに備え、前記制御部は、前記通常の音声認識部の認識結果に対して前記認識部によるささやき声の認識結果に対応する処理を実施するように構成されている。具体的には、前記制御部は、前記通常の音声認識部が通常の音声を変換したテキストに対して、前記認識部が認識したささやき声コマンドの処理を実行する。

[0009] また、本開示の第 2 の側面は、

発話音声をも音声特徴量に基づいて通常の音声とささやき声に分類する分類

ステップと、

前記分類ステップにおいて分類されたささやき声を認識する認識ステップと、

前記認識ステップにおける認識結果に基づく処理を制御する制御ステップと、

を有する情報処理方法である。

[0010] また、本開示の第3の側面は、

発話音声を音声特徴量に基づいて通常の音声とささやき声に分類する分類部、

前記分類部によって分類されたささやき声を認識する認識部、

前記認識部の認識結果に基づく処理を制御する制御部、

としてコンピュータを機能させるようにコンピュータ可読形式で記述されたコンピュータプログラムである。

[0011] 本開示の第3の側面に係るコンピュータプログラムは、コンピュータ上で所定の処理を実現するようにコンピュータ可読形式で記述されたコンピュータプログラムを定義したものである。換言すれば、本開示の第3の側面に係るコンピュータプログラムをコンピュータにインストールすることによって、コンピュータ上では協働的作用が発揮され、本開示の第1の側面に係る情報処理装置と同様の作用効果を得ることができる。

発明の効果

[0012] 本開示によれば、ささやき声を含む音声を処理する情報処理装置及び情報処理方法、並びにコンピュータプログラムを提供することができる。

[0013] なお、本明細書に記載された効果は、あくまでも例示であり、本開示によりもたらされる効果はこれに限定されるものではない。また、本開示が、上記の効果以外に、さらに付加的な効果を奏する場合もある。

[0014] 本開示のさらに他の目的、特徴や利点は、後述する実施形態や添付する図面に基づくより詳細な説明によって明らかになるであろう。

図面の簡単な説明

[0015] [図1]図1は、本開示を適用した音声入力システム100の基本構成を示した図である。

[図2]図2は、通常の音声とささやき声の周波数帯域毎のスペクトログラムを示した図である。

[図3]図3は、音声入力システム100の動作を示した図である。

[図4]図4は、ニューラルネットワークを用いてささやき声認識を行うシステムの全体構造を示した図である。

[図5]図5は、ささやき声と通常の音声の分類サンプルの結果を示した図である。

[図6]図6は、分類部101内で特徴ベクトルが処理される状況をt-SNEアルゴリズムに基づき次元圧縮して2次元表現した結果を示した図である。

[図7]図7は、分類部101の他のニューラルネットワーク構成例を示した図である。

[図8]図8は、本開示に係る音声入力システム100を利用した音声入力プロセスの具体的な動作例を示した図である。

[図9]図9は、本開示に係る音声入力システム100を利用した音声入力プロセスの具体的な動作例を示した図である。

[図10]図10は、本開示に係る音声入力システム100を利用した音声入力プロセスの具体的な動作例を示した図である。

[図11]図11は、本開示に係る音声入力システム100を利用した音声入力プロセスの具体的な動作例を示した図である。

[図12]図12は、本開示に係る音声入力システム100を利用した音声入力プロセスの具体的な動作例を示した図である。

[図13]図13は、本開示に係る音声入力システム100を利用した音声入力プロセスの具体的な動作例を示した図である。

[図14]図14は、本開示に係る音声入力システム100を利用した音声入力プロセスの具体的な動作例を示した図である。

[図15]図15は、本開示に係る音声入力システム100を利用した音声入力

プロセスの具体的な動作例を示した図である。

[図16]図 1 6 は、本開示を遠隔会議に適用した場合の動作例を示した図である。

[図17]図 1 7 は、会議端末 1 7 0 0 の機能的構成例を示した図である。

[図18]図 1 8 は、アバター制御システム 1 8 0 0 の機能的構成例を示した図である。

[図19]図 1 9 は、マスクタイプのウェアラブルインターフェース 1 9 0 0 の機能的構成例を示した図である。

[図20]図 2 0 は、情報処理装置 2 0 0 0 の構成例を示した図である。

発明を実施するための形態

[0016] 以下、図面を参照しながら本開示について、以下の順に従って説明する。

[0017] A. 概要

B. 基本構成

C. ささやき声の利用

D. 認識アーキテクチャ

D-1. ささやき声の認識

D-2. ささやき声の分類

D-3. トレーニング用データセット

E. 対話システムへの適用

F. 他のアプリケーション

F-1. 遠隔会議でのささやきコマンドの使用

F-2. 複数のアバターの制御

F-3. サイレントスピーチとの組み合わせ

G. 先行技術との対比

H. 情報処理装置の構成

I. 結論

[0018] A. 概要

自動音声認識（ASR）は、スマートスピーカーやカーナビゲーションシ

ステムなどのデバイスの操作からテキスト入力の手法に至るまで、さまざまなアプリケーションで既に使用されている。音声入力はマイク以外の特別な入力機器を必要としないので、ハンズフリーで操作可能である。さらに、音声に基づくテキスト入力はタイプ入力よりもはるかに高速であることから、音声入力はアイデアを書き留めたり、原稿の下書きをすばやく入力したりするためにも使用される。

[0019] しかしながら、音声認識では認識エラーの問題がある。音声による誤認識を発話コマンドによって修正することが難しく、キーボード又はマウスを使用して認識エラーを修正する必要があるため、ハンズフリー操作の利点が失われる。また、音声を使用してテキストを入力する場合、通常のテキスト以外の記号などの特殊文字やコマンドを音声で入力するときに問題が発生する。例えば、「？」を入力する場合は、「疑問符」と言うことができるが、「？」という記号ではなく「疑問符」というテキスト自体として入力されることもある。新しい段落を開始したくて「新しい段落」と言うと、文字通り認識され、「新しい段落」というテキストとして入力されてしまう場合がある。テキスト編集でよく使われる「単語の削除」や「新しい段落」などのコマンドを入力するときにも、同様の問題が発生し得る。

[0020] 例えば、「改行」という単語が発話されると、テキスト入力ではなく、常に改行コマンドを意味することを事前に取り決めておくこともできるが、その場合はテキストとして「改行」を入力できなくなる。さらに、発話コマンドとして定義されるすべての単語と覚えておくことはユーザにとって負担である。

[0021] ユーザが発話する音声はテキスト入力と発話コマンドが混在するが、従来の自動音声認識では音声からコマンドとテキスト入力を区別する簡単な方法がない。コンピュータのキーボード入力では、テキスト入力が可能であるとともに、Shiftキー、Controlキー、Altキーなどの修飾キーを英数字キーと同時に押すことでコマンドを入力することもできる。例えば修飾キーの入力で音声入力時におけるモダリティを切り替えるようにしても

よいが、ハンズフリー操作の利点が失われる。

[0022] そこで、本開示では、ささやき声と通常の音声で、発話に複数の意味を与えるようにした。具体的には、本開示では、発話音声を音声特徴量に基づいて通常の音声とささやき声に分類して、分類されたささやき声を認識し、その認識結果に基づいて、元の発話音声のうち通常の音声の部分に関連する処理を制御する情報処理装置について提案する。本開示に係る情報処理装置は、通常のマイク以外の特別なハードウェアを必要とせずに、完全なハンズフリー操作による音声入力を実現することができる。

[0023] 例えば、本開示を音声認識に適用する場合、入力された音声のうち通常の音声の部分を使ってテキスト入力を行う一方、入力音声からささやき声を分類して、ささやき声の認識結果をピリオド、コンマ、引用符などの記号又は特殊文字の入力や、テキストの削除、改行といったテキスト編集上のコマンドなどとして利用することができる。すなわち、本開示を音声認識に適用することにより、テキストを通常の音声で入力できる一方、コマンドなどの非テキスト情報はささやき声で入力することができる。

[0024] B. 基本構成

図1には、本開示を適用した音声入力システム100の基本構成を模式的に示している。音声入力システム100は、ささやき声と通常の音声で、発話に複数の意味を与えるシステムである。音声入力システム100は、例えばパーソナルコンピュータ（PC）のような一般的な情報処理装置を用いて構成することができ、音声入力にはマイク以外の特別な入力機器を必要としない。一般的な情報処理装置はキーボードやマウスといったユーザが入力操作を行う入力装置を備えているが、本開示により音声入力を行う上では、マイク以外の入力装置を使用する必要はない。

[0025] 図1に示す音声入力システム100は、分類部101と、ささやき声認識部102と、通常の音声認識部103と、制御部104を備えている。

[0026] ユーザが発話した音声は、マイク（又はヘッドセット）を使って、音声入力システムに入力される。分類部101は、入力音声を、音声特徴量に基づ

いて通常の音声とささやき声に分類する。入力音声は、通常の音声の区間と、ささやき声の区間と、無音の区間が含まれる。したがって、分類部101は、入力された音声信号に対して、通常の音声の区間とささやき声の区間をそれぞれ識別するマークを付けて、後段のささやき声認識部102及び通常の音声認識部103の各々に出力する。

[0027] 通常の音声認識部103は、分類部101から入力される時系列の音声信号のうち、通常の音声の区間のみを使って通常の音声認識処理を実行して、テキストに変換する。

[0028] ささやき声認識部102は、分類部101から入力される音声信号のうち、ささやき声の区間のみを使ってささやき声の認識処理を実行して、通常の音声によって入力されたテキストに対する記号や特殊文字、編集コマンド（テキストの削除や改行など）といった非テキスト情報に変換する。

[0029] そして、制御部104は、通常の音声認識部102から出力されるテキストに対して、ささやき声認識部103から出力される非テキスト情報に基づく処理を実施する。例えば、音声入力システム100がテキスト入力システムとして利用される場合、制御部104は、通常の音声として入力されたテキストに対して、ささやき声で入力された記号や特殊文字のテキスト中の該当箇所への入力や、ささやき声で指定されたテキスト変換候補の選択、ささやき声で入力された削除や改行などのコマンドの実行を行う。

[0030] もちろん、音声入力システム100を、テキスト入力以外の潜在的なアプリケーションに適用することもできるが、詳細は後述に譲る。

[0031] 本開示を適用した音声入力システム100を実現するために、2つの新たなニューラルネットワークを用いている。1つは分類部101で使用される、通常の音声とささやき声を区別するためのニューラルネットワークである。もう1つはささやき声認識部102で使用される、ささやき声を認識するためのニューラルネットワークである。各ニューラルネットワークの詳細については後続のD項に譲る。

[0032] 音声認識が既に利用可能なさまざまな状況では、本開示を広範に利用する

ことができる。本開示の1つの実装例では、既存の音声認識システム（例えば、米Google社のCloud Speech-to-Text）を使用して通常の音声認識を行うとともに、カスタマイズしたニューラルネットワークを使用してささやき声の認識を行う。前者の通常の音声認識は大規模なコーパスで訓練されており、話者に依存せず、特別な訓練なしで 사용할 ことができる。対照的に、ささやき声認識の部分は、既存の大規模なコーパスがないため、より小さなコーパスでトレーニングするしかなく、話者適応（すなわちカスタマイズ）のための個々のユーザが例文をささやき声で読み上げることが求められるトレーニング段階を必要とする。

[0033] 将来、ささやき声に基づくインタラクションが普及するにつれて、使用履歴からささやき声のデータセットを取得できるようになると予想される。その結果、話者に依存しない方法でささやき音声認識を使用できるようになることが期待される。

[0034] C. ささやき声の利用

ささやき声は、多くの人が特別なトレーニングなしで発音できる発話のモードであり、発話中に通常の音声とささやき声に自在に使い分けることができる。これが、本開示において、通常の音声の別の音声入力モードとしてささやき声を使用する理由の1つである。

[0035] 人が発音できる発話モードとして、通常の音声とささやき声の他に、音声のピッチの切り替えを挙げることができる。しかしながら、人が高いピッチで発話することは負担が大きく、またピッチの高低の閾値は個人差がある。すなわち、音声のピッチの切り替えは人にとって不自然であるとともに、音声を入力したシステム側でも通常の音声との分類が難しい。一方、ささやき声の音量を無理やりに上げたとしても、通常の音声と同じものにもならない。要するに、人はささやき声と通常の音声を容易に使い分けることができるとともに、音声を入力したシステム側でも通常の音声との分類は容易である。

[0036] 一般には、通常の音声はユーザが正常に話す時の音声として使用されるの

に対し、ささやき声はユーザが内緒話をするに使用される音声である。通常の音声とささやき声とは、発音方式が異なる。具体的に、通常の音声が出されると、人の声帯は規則的且つ周期的な振動を呈しており、このような振動周波数は基本周波数と呼ばれる。一方、ささやき声で話した場合、声帯の振動は目立たず、不規則的且つランダムな振動を呈しており、すなわち基本周波数がないということもできる。したがって、ささやき声の音量を無理やりに上げたとしても、通常の音声と同じものにもならない。

[0037] 図2（A）及び図2（B）には、通常の音声とささやき声の各音声の周波数帯域毎のスペクトログラムをそれぞれ示している。但し、話者は、「A quick brown fox jumps over the lazy black dog.」という同じフレーズを通常の音声とささやき声でそれぞれ発話した場合である。通常の音声で発話するときには声帯は規則的且つ周期的な振動を呈するのに対し、ささやき声を発話するときには声帯の振動はほとんどない、という生体现象の相違に起因して、両者のスペクトログラムのパターンの違いは一目瞭然である。パターン認識はニューラルネットワークの主要なアプリケーションの1つであり、ニューラルネットワークを活用すれば、通常の音声とささやき声の周波数スペクトログラムをリアルタイムで正確に識別することができる。

[0038] そこで、本開示を適用した音声入力システム100を実現するために、2つの新たなニューラルネットワークを用いた。1つは分類部101で使われる、通常の音声とささやき声を区別するためのニューラルネットワークである。もう1つはささやき声認識部102で使われる、ささやき声を認識するためのニューラルネットワークである。

[0039] 図3には、ニューラルネットワークを用いて周波数スペクトログラムを認識するという観点から、図1に示した音声入力システム100の動作を図解している。

[0040] 音声入力システム100には、参照番号301で示す、通常の音声とささやき声が混在した話者の音声が入力される。第1のニューラルネットワーク

を用いた分類部 101 は、所定の時間単位（例えば、100 ミリ秒毎）で、入力音声が通常の音声とささやき声のいずれであるかを分類する。

[0041] ささやき声認識部 102 は、入力音声のうち、参照番号 302 で示す、ささやき声に分類された区間に対して、第 2 のニューラルネットワークを用いて認識処理を行う。また、通常の音声認識部 103 は、入力音声のうち、参照番号 303 で示す、通常の音声に分類された区間に対して、第 3 のニューラルネットワークを用いて通常の音声認識処理を行う。

[0042] 例えば、音声入力システム 100 がテキスト編集に適用される場合、通常の音声認識部 103 は、通常の音声区間の音声をテキストに変換する。また、ささやき声認識部 102 は、ささやき声の区間を認識処理して、通常の音声によって入力されたテキストに対する記号や特殊文字、編集コマンド（テキストの削除や改行など）といった非テキスト情報に変換する。そして、制御部 104 は、通常の音声認識部 102 から出力されるテキストに対して、ささやき声認識部 103 から出力される非テキスト情報に基づく処理を実施する。

[0043] 音声入力システム 100 は、ささやき声と通常の音声で、発話に複数の意味を与えることができる。例えば、通常の音声で「改行」と言うと、そのテキストを入力することを意味するが、ささやき声で「改行」と言うと、改行コマンドを呼び出すことを意味する。同様に、認識エラーを修正するために、ユーザは「候補」をささやいて代替の認識候補を表示し、その候補に付けられた番号（「1」、「2」など）をささやくことによってそれらの 1 つを選択することができる。

[0044] 従来の音声認識方式は、コマンドキーや記号キーを含まないキーボードに喩えることができる。コンピュータの操作に使用されるキーボードには、文字入力と同様に、ファンクションキー、コマンドキー、記号のキー、及びコマンドを入力するためのその他の手段が用意されている。インタラクションモードを明示的に変更したり、キーボードから手を離したりしなくても、これらの関数を呼び出すことができる。すなわち、コンピュータキーボードで

は、特別なモード変換操作なしでテキスト入力とコマンド入力を共存させている。これに対し、音声入力システム 100 を用いることで、通常の音声とささやき声の関係を、テキスト入力とコマンド入力の関係に喩えることができる。音声入力システム 100 では、ユーザが通常の音声とささやき声を使い分けることで、ハンズフリー操作でテキスト入力とコマンド入力を共存させている。

[0045] D. 認識アーキテクチャ

この D 項では、本開示を適用した音声入力システム 100 において、ささやき声の認識と分類を可能にするニューラルネットワークの構成について説明する。

[0046] D-1. ささやき声の認識

本開示の最良の実施形態では、ささやき声の認識（すなわち、ささやき声認識部 102）には、wav2vec2.0（非特許文献 1 及び非特許文献 2 を参照のこと）又は HuBERT（非特許文献 3 を参照のこと）を使用する。wav3vec2.0 及び HuBERT はいずれも音声処理システム用に設計された自己教師ありニューラルネットワークである。

[0047] wav2vec2.0 と HuBERT のどちらも、事前トレーニングと、ラベルなし音声データを用いた自己教師あり表現学習及びラベルあり音声データを用いたファインチューニングとの組み合わせを想定している。これらのシステムは、主に音声認識アプリケーションを対象としているが、話者認識、言語認識、及び感情認識にも適用されている。

[0048] 図 4 中央には、ニューラルネットワークを用いて構成されるささやき声認識を行うシステム（ささやき声認識ニューラルネットワーク 400）の全体構造を示している。図示のシステムにおいて、wav2vec2.0 又は HuBERT を用いて構成されるニューラルネットワーク 400 は、特徴抽出層 410 と変換層 420 に大別できる。

[0049] このささやき声認識ニューラルネットワーク 400 の事前トレーニングの方法は、自然言語処理における BERT のマスクあり言語モデルに似ている

。これは、入力の一部をマスクし、残りの入力からそれに対応する表現特徴（特徴ベクトル）を推定するように設計されている。この事前トレーニングにより、入力データの音響特性と音声の特徴を学習できるはずである。

[0050] 事前トレーニングとは対照的に、音響データにテキストラベルを必要とするファインチューニングでは、少量のデータしか必要としない。図4に示す、WAV2vec 2.0又はHuBERTからなるささやき声認識用のニューラルネットワーク400の出力に、Projection層430とConnectionist Temporal Classification (CTC)層440をさらに追加して、音声波形からテキストの文字起こしを生成できるように構成されている。

[0051] 図4に示すささやき声認識ニューラルネットワーク400は、少量のラベル付き音声データセットのみを用いたファインチューニングによって、従来の最先端のASRに匹敵する音声認識精度を実現できる（非特許文献1及び非特許文献2を参照のこと）。したがって、このアーキテクチャは、限られたささやき声コーパスの下でささやき声を認識するのに適していると考えられる。

[0052] 通常の音声のコーパスは数多あるが、ささやき声のコーパスはLIMIT（非特許文献4を参照のこと）などごくわずかしかない。そこで、本実施形態では、通常の音声コーパスで事前トレーニングされたささやき声認識ニューラルネットワーク400に対してさらにささやき声によるファインチューニングを行うという学習方法を採用している。詳細については後続のD-3項を参照されたい。

[0053] 特徴抽出層410は、生の音声波形を潜在的な特徴ベクトルに変換する。図4に示す例では、特徴抽出層410は、複数（例えば5個）の畳み込みニューラルネットワーク（CNN）からなる。各CNNには、ユーザが発音した音声波形信号を所定長の時間区間毎に分割した区間データが入力される。但し、隣接する時間区間で重複領域を含むように各区間データに分割される。

[0054] 図4左には、特徴抽出層410内の各CNNの構成を示している。各CNNは、オリジナルのwav2vec2及びHuBERTと同様に、7ブロックの1次元畳み込み層（Conv1D）と時間的畳み込み層（GroupNorm）で構成される。各ブロックは、ストライド（5, 2, 2, 2, 2, 2, 2）及びカーネル幅（10, 3, 3, 3, 3, 2, 2）からなる512チャンネルを持つ。特徴抽出層410は、20ミリ秒毎に512個の潜在次元の特徴ベクトルを出力するように設計されている。

[0055] D-2. ささやき声の分類

分類部101は、固定長（100ミリ秒など）の音声信号により、ささやき声と通常の音声入力を区別する。図4右には、wav2vec2.0又はHuBERTからなるささやき声認識ニューラルネットワーク400のうち、特徴抽出層410の構成を詳細に示している。特徴抽出層410は、音響信号を20ミリ秒毎に512次元の特徴ベクトルに変換する。分類部101は、ささやき声認識ニューラルネットワーク400とニューラルネットワーク（すなわち、特徴抽出層410）を部分的に共有することによって、音声入力システム100全体としてネットワークサイズが縮小される。

[0056] 分類部101は、図4中央に示すささやき声認識ニューラルネットワーク400内の特徴抽出層410のうち、破線で囲んだ部分をささやき声認識ニューラルネットワーク400と共有している。図4右には、分類部101内を拡大して示している。分類部101は、音声波形信号から抽出された特徴ベクトルに対して正規化層（Layer Norm）411と、平均プーリング層（Avg Pool）412と、それに続く2つのFC（全結合）層413及び414と、多クラス分類の出力層としてのLogSoftmax層415を順に適用して、音声波形信号からささやき声と通常の音声分類を取得することができる。

[0057] 図5には、図4右に示した分類部101によって取得される、ささやき声と通常の音声の分類サンプルの結果を、周波数スペクトラムを用いて示している。

[0058] 音声入力システム100に入力される音声波形信号には、ささやき声の区間と、通常の音声の区間と、無音の区間が含まれる。まず、分類部101は、図5(A)に示すように、入力された音声波形信号の周波数スペクトログラムからささやき声の区間を識別して「Classified as whisper」とマークするとともに、通常の音声の区間を識別して「Classified as normal」とマークする。

[0059] そして、分類部101は、「Classified as normal」のマークに基づいて、図5(B)に示すような通常の音声削除された音響ストリームを生成して、後段のささやき声認識部102に出力する。また、分類部101は、「Classified as whisper」のマークに基づいて、図5(C)に示すようなささやき声削除された音響ストリームを生成して、後段の通常の音声認識部103に出力する。

[0060] 図6(A)には、2番目のFC(全結合)層414への入力として分類される512次元の特徴ベクトルを、最初の正規化層への入力時の特徴ベクトルとともに、t-SNE(t-Distributed Stochastic Neighbor Embedding)と呼ばれる次元削減アルゴリズムによる可視化方法に基づいて2次元表現した結果を示している。図6(A)を図6(B)に示す正規化層411への入力時と2番目のFC(全結合)層414への入力時と比較すると、分類部101内での処理によって通常の発話の特徴ベクトルとささやき声の特徴ベクトルがよく区別されていることが分かる。

[0061] 図7には、ささやき声認識ニューラルネットワーク400内の特徴抽出層410と共有される、分類部101の他の構成例を示している。図示の分類部101は、音声波形信号から抽出された特徴ベクトルに対して正規化層gMLP701と、平均プーリング層(Avg Pool)702と、FC(全結合)層703と、多クラス分類の出力層としてのLogSoftmax層704を順に適用して、図4右に示した構成例と同様に、音声波形信号からささやき声と通常の音声分類を取得することができる。なお、gMLP(

multi-layer perceptron with gating) は、Google Brain から発表された深層学習モデルであり、多層パーセプトロンにゲート機構を組み合わせるだけで、Attention 機構を使わずにトランスフォーマに匹敵する性能を有するとされる。

[0062] D-3. トレーニング用データセット

ささやき声認識ニューラルネットワーク400のトレーニングに関しては、まず、通常の音声データを用いて事前トレーニング及びファインチューニングされたニューラルネットワークから開始する。具体的には、通常の音声データとしてLibriSpeechデータセット（非特許文献5を参照のこと）を使用し、通常の音声を用いた960時間のトレーニングを実施する。事前トレーニングには教師データ（音声に対応するテキスト）は必要でなく、音声データのみが必要であるという点には留意されたい。

[0063] 続いて、wTIMIT及びPer-user databaseという2種類のささやき声データセットを使用して、さらにささやき声認識ニューラルネットワーク400のファインチューニングを行う。

[0064] ここで、wTIMIT (whisper TIMIT) は、ささやき声用のコーパスである（非特許文献4を参照のこと）。各話者は、TIMITのプロンプトに従って、450の音声的にバランスの取れた文を通常の発話とささやき声の両方と言う。話者は29名で、性別のバランスがとれている。使用した発話の総数は11,324 (1,011分) である。また、wTIMITのコーパスには通常の音声が付属しており、分類部101のささやき声と通常の音声の分類トレーニングにも使用することができる（後述）。データは、トレーニングとテストの2つの部分で提供されている。本実施形態ではトレーニングとテストの区分をそのまま使用する。

[0065] また、Per-user databaseは、各ユーザがささやき声で吹き替えた音声データである。選択した音声コマンドのシーケンスがスクリプトとして使用される。これらのフレーズは、主にテキスト入力中に使用されるコマンドであると想定されている。各ユーザは各フレーズを5回繰り返

し、合計 110 のフレーズが録音される。これらのフレーズはさらにランダムに連結され、データセットとして使用される。この連結後の発話の総数は 936（約 82 分）である。

[0066] ファインチューニングは、トレーニング済みのモデルを各タスクに合わせて調整するようにトレーニングするという当業界で周知の学習手法である。具体的には、ファインチューニングでは、ラベルなしデータを使ってモデルを学習させた後、さらに解きたい特定のタスクの教師ありデータを使ってモデルのパラメータをチューニングする。

[0067] 本実施形態では、ささやき声認識ニューラルネットワーク 400 のファインチューニングを 2 段階で行う。最初の段階では、w T I M I T（ささやき声）でトレーニングし、2 番目の段階では、ユーザが吹くささやき声コマンドセット（ユーザ毎のささやき声のデータセット）を使用する。

[0068] また、分類部 101（又は、特徴抽出層 410）のささやき声と通常の音声の分類トレーニング用のトレーニングデータセットとして、w T I M I T に含まれる通常の音声及びささやき声の音声データを使用する（上述したように、w T I M I T のコーパスには、ささやき声とともに通常の音声が付属している）。分類部 101 に供給される音声の長さは、1,600 サンプルに設定されている（16 K p s のオーディオサンプリングで 100 ミリ秒のサイズ）。これは、後の段階において音声認識クラウドサービスで 사용되는音声チャンクの長さとは一致する。

[0069] E. 対話システムへの適用

本開示に係る音声入力システム 100 では、通常の音声入力は従来の音声入力テキスト作成と同じようにテキストに変換される。対照的に、「C O M M A」、「P E R I O D」、「Q U O T E」などの記号をささやき声で話すと、音声入力システム 100 は、これらを記号として扱う。

[0070] また、新しい行を作成することを意味する「N E W L I N E S」をささやき声で話すと、音声入力システム 100 はコマンドとして扱い、音声入力中のテキストに対して新しい行を作成する。

- [0071] (通常の音声で入力されたテキストに対して) 誤認識がある場合、「戻る」又は「文を削除」とささやくことで、音声入力システム100はコマンドとして扱い、最後の単語を削除したり、関心のある単語を変更したりすることができる。
- [0072] また、「MENU」とささやくことで、音声入力システム100はコマンドとして扱い、別の認識候補を提示することも可能である。メニュー上の候補には1、2、3のラベルが付いており、「ONE」、「TWO」、「THREE」などをささやくことでメニューの中から好きな候補を選択できる。
- [0073] 本開示に係る音声入力システム100によれば、通常の音声入力とささやき声のコマンドを組み合わせることも可能である。例えば、「w a v 2 v e c」という綴りを通常の音声で入力した直後に「SPELL」とささやくと、通常のASRでは入力がかかり難しい「wav2vec2」という単語が生成される。同様に、通常の音声で「smile」と言った直後に「EMOTION」とささやくことで、絵文字入力も可能である。
- [0074] もちろん、「COMMA」、「PERIOD」、「QUOTE」、「NEW LINES」、「戻る」又は「文を削除」と通常の音声で話すと、音声入力システム100は、記号やコマンドとしては扱わずに、従来の音声入力テキスト作成と同じようにテキストに変換する。
- [0075] 本開示に係る音声入力システム100を利用した音声入力プロセスの具体的な動作例について、図8～図15を参照しながら説明する。図8～図15は、音声入力ベースのテキスト編集画面である。
- [0076] 図8～図10には、通常の音声でテキストを入力中にささやき声によってコマンドを使用する場合の動作例を示している。図8に示すテキスト編集画面では、通常の音声により入力されたテキストのうち確定した部分を黒い文字で表示し、音声入力直後の未確定の部分をグレーのテキストで表示している。ここで、ユーザが「MENU」とささやく。音声入力システム100は、「MENU」というささやき声をコマンドとして扱う。具体的には、分類部101が「MENU」に対応する音声波形をささやき声に分類し、ささや

き声認識部102がその音声波形が「MENU」というコマンドであると認識する。

[0077] 音声入力システム100は、図9に示すように、画面内にグレーで表示されているテキストの音声認識候補をリストアップしたメニューウィンドウをポップアップ表示する。図示のメニューウィンドウには、未確定の入力音声に対する5つのテキスト変換候補が表示されている。音声入力システム100内では、制御部104が、ささやき声認識部102による「MENU」コマンドの認識結果に応答して、対象とする（又は、直近で入力された）通常の音声に対する通常の音声認識部103によるテキスト変換候補を取得して、これらの変換候補をリストアップしたメニューウィンドウを生成して、テキスト編集画面上にポップアップ表示させる。

[0078] 図9に示すように、メニュー上の候補には1、2、3、…のラベルが付いており、ユーザは「ONE」、「TWO」、「THREE」、…などをささやくことでメニューの中から好きな候補を選択することができる。ここで、ユーザが「FOUR」とささやいて、4番目の候補「of sitting by her sister」を選択する。

[0079] 音声入力システム100内では、分類部101が「FOUR」に対応する音声波形をささやき声に分類し、ささやき声認識部102がその音声波形が「FOUR」という4番目の候補を選択するコマンドであると認識する。そして、制御部104が、ささやき声認識部102による「FOUR」すなわち4番目の候補を選択という認識結果に応答して、4番目のテキスト変換候補の選択を確定させると、図10に示すように、メニューウィンドウを閉じて、図8に示したテキスト編集画面では未確定のグレー表示されていたテキストを選択された「of sitting by her sister」というテキスト候補に置き換えて表示させる。

[0080] 続いて、図11を参照しながら、通常の音声でテキストを入力中にささやき声で記号を入力する場合の動作について説明する。図10に示したように、ささやき声を使ってテキスト変換候補を選択した後、ユーザは通常の音声

による入力をしばらく続けたとする。音声入力システム１００内では、分類部１０１が通常の音声に分類した音声波形を、通常の音声認識部１０３で認識してテキストに変換する。そして、制御部１０４は、テキスト編集画面上で、入力が確定したテキストの最後尾に、新たに変換されたテキストを逐次連結させていく。

[0081] ここで、ユーザが、通常の音声で「…c o n v e r s a t i o n s i n i t」と発話した後に、続けてささやき声で「COMMA DOUBLE QUOTE」と発話して、さらに通常の音声で「a n d w h a t i s …」と発話したとする。

[0082] 音声入力システム１００内では、分類部１０１は「…c o n v e r s a t i o n s i n i t」までの音声波形を通常の音声に分類し、通常の音声認識部１０３はテキストに変換し、制御部１０４はテキスト編集画面上で「…c o n v e r s a t i o n s i n i t」までのテキストを表示させる。次いで、分類部１０１が「COMMA DOUBLE QUOTE」に対応する音声波形をささやき声に分類し、ささやき声認識部１０２がその音声波形が「COMMA」及び「DOUBLE QUOTE」という記号入力であると認識する。そして、制御部１０４は、ささやき声認識部１０２による各記号「COMMA」及び「DOUBLE QUOTE」の入力という認識結果に応答して、図１１に示すように、テキスト編集画面上で、入力が確定したテキストの最後尾に、各記号「，」及び「”」を逐次連結させていく。なお、図示を省略するが、特殊文字についても記号入力と同様の方法によりささやき声で入力することができる。

[0083] 図１２及び図１３には、略称のように通常のＡＳＲでは入力がかかり難しい単語（又は、まだ辞書登録されていない単語）を、通常の音声入力とささやき声のコマンドを組み合わせることによって入力する、音声入力システム１００の動作例を示している。

[0084] ユーザは、「w a v 2 v e c 2」という綴りを通常の音声で入力し、その直後にささやき声で「S P E L L」と発話する。音声入力シ

ステム１００内では、まず、分類部１０１は「w a v 2 v e c 2」に該当する音声波形を通常の音声に分類し、通常の音声認識部１０３はその音声波形をアルファベットの各文字「w」、「a」、「v」、「2」、「v」、「e」、「c」、及び「2」に変換する。そして、制御部１０４は、通常の音声認識部１０３による単語の綴りの認識結果に応答して、図１２に示すように、テキスト編集画面上で、入力が確定したテキストの最後尾に、アルファベットの各文字「w」、「a」、「v」、「2」、「v」、「e」、「c」、及び「2」を逐次連結させていく。

[0085] 続いて、分類部１０１は「S P E L L」に対応する音声波形をささやき声に分類し、ささやき声認識部１０２がその音声波形が「S P E L L」という単語の綴りの入力を指示するコマンドであると認識すると、制御部１０４は、直前に入力したアルファベットの文字をそのまま発話された順に結合して、「w a v 2 v e c 2」という単語を生成する。そして、制御部１０４は、ささやき声認識部１０２による単語の綴りの認識結果に応答して、図１３に示すように、テキスト編集画面上で、入力が確定したテキストの最後尾に、「w a v 2 v e c 2」という単語を連結させる。

[0086] 図１４及び図１５には、略称のように通常のＡＳＲでは入力がかかり難しい絵文字を、通常の音声入力とささやき声のコマンドを組み合わせることによって入力する、音声入力システム１００の動作例を示している。

[0087] ユーザが、通常の音声で「w h a t a g o o d d a y t o d a y」と発話した後に、さらに続けて通常の音声で単語「s m i l e」と言った直後に「E M O T I O N」とささやいたとする。

[0088] 音声入力システム１００内では、分類部１０１は「w h a t a g o o d d a y t o d a y」及び「s m i l e」までの音声波形を通常の音声に分類し、通常の音声認識部１０３はこの音声波形を「w h a t a g o o d d a y t o d a y s m i l e」というテキストに変換する。そして、制御部１０４は、図１４に示すように、テキスト編集画面上で「w h a t a g o o d d a y t o d a y」及び「s m i l e」までのテキス

トを表示させる。次いで、分類部101が単語「EMOTION」に対応する音声波形をささやき声に分類し、ささやき声認識部102がその音声波形が絵文字入力を指示するコマンドであると認識する。そして、制御部104は、ささやき声認識部102による絵文字の入力という認識結果に応答して、図15に示すように、テキスト編集画面上で、「smile」という最後のテキストを絵文字に変換して表示する。

[0089] F. 他のアプリケーション

このF項では、上記E項で説明した音声対話入力以外の、本開示を利用できるいくつかのアプリケーションについて説明する。

[0090] F-1. 遠隔会議でのささやきコマンドの使用

本開示の他のアプリケーションとして、遠隔会議中の音声コマンドの使用が挙げられる。遠隔会議中に音声コマンドを使用することを想定する。コマンドが会議中の発話の一部になるのを防ぐために、会議の参加者はささやき声を使ってコマンドを与える。

[0091] 一般に、遠隔会議では、各参加者はTCP/IPなどのネットワーク経由で接続され、互いの顔画像の映像及び発話音声をリアルタイムで通信するとともに、各参加者は、それぞれ自分の会議端末上で、他の参加者のリアルタイムの顔画像と発話音声を視聴し、必要に応じて会議資料を参加者間で共有する。

[0092] さらに本開示を遠隔会議に適用すれば、会議の参加者はささやき声を使ってコマンドを与えることができ、音声コマンドが会議中の発話の一部になるのを防ぐことができる。すなわち、会議の参加者が発話した音声が入力システム100に入力されると、分類部101はささやき声と通常の音声を区別する。そして、ささやき声の部分はささやき認識部102に送られ、その認識結果に基づいてコマンドの処理が実施される。また、ささやき声が会議中の発話の一部になるのを防ぐために、会議参加者の音声ストリームからささやき声の部分を削除した後、参加者の顔画像の映像と多重化して、他の参加者の会議端末にリアルタイム送信する。

[0093] ところで、通常の音声及びささやき声のいずれを発話する場合にも口を動かす。このため、他の参加者からは、（ささやき声の部分が削除されているのに）ささやき声を発しながら口を動かす映像が見えて不自然であるという問題がある。図16（A）には、通常の音声とささやき声を交互に発話している参加者の元の映像を例示している。ささやき声を発話している区間は、他の参加者にとっては発話していない無音区間になっているのに、通常の音声聴こえている区間と同様に参加者の口が動いているので、このような映像を視聴している他の参加者にとっては不自然に感じられる。

[0094] そこで、本開示を遠隔会議に適用する際には、図16（B）に示すように、参加者がささやき声を発話している区間において、参加者が発話していないかのように参加者の映像の唇の部分を入れ替えるという映像加工を実施する。その結果、他の参加者は、ささやき声の部分が削除されて無音になっている区間は参加者の口が動いていない映像を視聴することになり、聴覚及び視覚の各々から得られる情報が整合して、自然な映像を観察できるようになる。

[0095] 参加者の映像の唇の部分を入れ替える目的のために、例えばW a v 2 L i p（非特許文献6を参照のこと）を使用して、深層学習を使用した発声から唇の画像を生成することができる。ユーザがささやき声で話すと、図16（C）に示すように、顔画像のうちが口の部分を、口が開いて発話している状態から口を閉じて黙っている状態の画像に置き換えられる。

[0096] このようなメカニズムを使用すると、顔画像ストリームと音声ストリームの両方を調整して、ささやき声でコマンドが発声された部分が他の参加者に見えなくなり且つ聞こえなくなるようにすることができる。

[0097] 図17には、本開示に係る音声入力システム100を適用した会議端末1700の機能的構成例を示している。但し、図17では、会議端末1700のうち、参加者の映像及び音声を取り込んで他の会議端末へ送信する入力側の機能ブロックのみを図示し、説明の便宜及び図面の簡素化の観点から、他の会議端末から受信した映像及び音声を出力する出力側の機能ブロックにつ

いては図示を省略している。また、会議端末1700に含まれる機能ブロックのうち、図1に示した音声入力システム100内に含まれるものと同じ機能ブロックについては同じ名称及び符号を付けて示している。

[0098] 会議端末1700は、音声入力システム100の機能ブロックの他に、マイク1701と、カメラ1702と、データ処理部1703と、データ送信部1704をさらに備えている。

[0099] マイク1701は会議参加者が発話する音声を入力し、カメラ1702は会議参加者を撮像する。

[0100] 分類部101は、マイク1701から入力された会議参加者の音声を通常の音声とささやき声に分類する。ささやき声認識部102は、ささやき声に分類された音声信号を認識処理してコマンドに変換する。制御部104は、ささやき声認識部102で認識されたコマンドの処理を実行する。コマンドがどのような内容であるかは特に限定されない。また、会議端末1700においてテキスト入力を行わない場合には、通常の音声認識部103は不要である。

[0101] データ処理部1703は、分類部101によって通常の音声に分類された音声ストリームと、カメラ1702によって参加者を撮像した映像ストリームとを、所定のフォーマット（例えばMPEG（Moving Picture Experts Group）など）に多重化及び符号化する。

[0102] 上述したように、分類部101からデータ処理部1703には、ささやき声の部分が削除された音声ストリームが入力される。一方、カメラ170からデータ処理部1703に入力される元の映像では会議参加者は通常の音声及びささやき声のいずれを発話する場合にも口を動かしており、ささやき声の部分が削除された区間では映像と音声が整合していない。そこで、データ処理部1703は、図16（B）を参照しながら説明したように、参加者がささやき声を発話している区間において、例えばWav2Lipを使用して、参加者が話していないかのように参加者の映像の唇の部分を入れ替えるという映像加工を施してから、音声ストリームとの多重化及び符号化処理を行

う。

[0103] そして、データ通信部 1704 は、データ処理部 1703 による処理後のデータ（映像音声ストリーム）を、TCP/IP などの所定の通信プロトコルに則ってネットワーク経由で他の会議端末に送信する。

[0104] 会議端末 1700 において顔画像ストリームと音声ストリームの両方を調整することにより、受信側である他の会議端末では、ささやき声が聞こえなくなったクリアな通常の音声を聞くことができるとともに、ささやき声でコマンドが発声された部分の口の動きが見えなくなった自然な映像を視聴することができる。

[0105] F-2. 複数のアバターの制御

本開示のさらに他のアプリケーションとして、1人のユーザが仮想空間における複数のアバターの制御を同時に行うことができる。

[0106] 複数のアバターの同時制御の一例として、ユーザの通常の音声で1番目のアバターを喋らせ、ユーザのささやき声を使って2番目のアバターを操作することができる。ここで言うアバターの「操作」は、アバターの身体運動や発話を始めアバターのさまざまな動作を含むものとする。

[0107] また、複数のアバターの同時制御の他の例として、ユーザの通常の音声を使って1番目のアバターの発話を行い、ユーザのささやき声を使って2番目のアバターの発話を行うようにすることができる。但し、ささやき声自体は聞き取り難いので、例えば非特許文献7及び非特許文献8に開示されている音声変換技術を用いて、ささやき声を通常の音声に音声変換してアバターの声として利用するようにしてもよい。

[0108] 図18には、本開示に係る音声入力システム100の機能を組み込んで構成される、複数のアバターを同時制御するアバター制御システム1800の機能的構成例を模式的に示している。但し、アバター制御システム1800に含まれる機能ブロックのうち、図1に示した音声入力システム100内に含まれるものと同じ機能ブロックについては同じ名称及び符号を付けて示している。

[0109] ユーザの通常の音声を使って1番目のアバターの発話を行い、ユーザのささやき声を使って2番目のアバターの発話を行う場合のアバター制御システム1800の動作について説明する。

[0110] 分類部101は、マイクやヘッドセットから入力されたユーザの音声を通常の音声とささやき声に分類する。第1のアバター音声生成部1801は、分類部101によって通常の音声に分類された音声信号を音声変換して、第1のアバターの音声を生成する。通常の音声を他の音声に変換するアルゴリズムは任意であり、現在利用可能なボイスチェンジャーを活用してもよい。

[0111] また、第2のアバター音声生成部1802は、分類部101によってささやき声に分類された音声信号を通常の音声に変換して、第2のアバターの音声を生成する。例えば非特許文献7及び非特許文献8に開示されている音声変換技術を用いて、ささやき声を他の通常の音声に変換して、アバターの発話として利用するようにしてもよい。

[0112] F-3. サイレントスピーチとの組み合わせ

サイレントスピーチは、無声音又は小声で周囲に気づかれずに音声入力を可能にするインターフェースであり、音声コマンドの音声周囲にノイズにならないようにすることと、機密情報が開示されないことによってプライバシーを保護することを主な目的とする。一般的な会話の音圧レベルは約60 dBであるが、ささやき声の音圧レベルは30~40 dBである。このように、ささやき声を発話コマンドとして使用することで、サイレントスピーチの目的を大幅に達成することができる。

[0113] 例えば、大気汚染や感染症から保護するために、呼吸用の動力付き換気を可能にするマスクが提案されている（特許文献2を参照のこと）。このようなマスクに本開示に係る音声入力システム100の機能を組み込むことで、マスクタイプのウェアラブルインターフェースを実現することができる。マスクタイプのウェアラブルインターフェースの場合、口元にマイクを配置できるので、音圧レベルの低いささやき声も收音することができる。

[0114] 図19には、本開示に係る音声入力システム100の機能を組み込んでマ

マスクタイプのウェアラブルインターフェース１９００の機能的構成例を示している。但し、ウェアラブルインターフェース１９００に含まれる機能ブロックのうち、図１に示した音声入力システム１００内に含まれるものと同じ機能ブロックについては同じ名称及び符号を付けて示している。

[0115] ウェアラブルインターフェース１９００は、マスクタイプの本体に、少なくともマイク１９０１とスピーカー１９０２が搭載されている。分類部１０１やささやき声認識部１０２など、音声入力システム１００の少なくとも一部の構成要素も、ウェアラブルインターフェース１９００内に搭載されていてもよいし、音声入力システム１００の機能はウェアラブルインターフェース１９００外に配置されていてもよい。

[0116] 分類部１０１は、マイク１９０１から入力されたユーザの音声を通常の音声とささやき声に分類する。ささやき声認識部１０２は、分類部１０１によってささやき声に分類された音声信号から音声コマンドを認識する。そして、制御部１０４は、ささやき声コマンドの処理を実施する。ここではささやき声コマンドの内容は特に限定されない。ささやき声コマンドは、マイク１９０１から入力された通常の音声に対する処理であってもよいし、ウェアラブルインターフェース１９００を接続するパーソナルコンピュータ（ＰＣ）やその他の情報端末に対するコマンドであってもよい。マスクタイプのウェアラブルインターフェース１９００の場合、口元に配置したマイク１９０１を用いて音圧レベルの低いささやき声も收音できるので、音声コマンドの聞き逃しを防ぐことができる。

[0117] 一方、分類部１０１によって通常の音声に分類された音声信号は、増幅器１９０３で増幅された後、マスクすなわちウェアラブルインターフェース１９００に取り付けられたスピーカー１９０２から音声出力される。マスクタイプのウェアラブルインターフェース１９００を装着することによって、ユーザの口が覆われるため、通常の音声を聞き取り難くなるが、スピーカー１９０２で拡声することによって、マスクによる音声の減衰を補うことができる。

[0118] マスクタイプのウェアラブルインターフェース1900によれば、マスクの中のマイクでささやき声を拾うことによって、無音の発話とほぼ同等の効果を達成することが可能である。すなわち、常時着用のマスクに本開示に係る音声入力システム100を導入することで、通常の会話を妨げることなく常時利用可能な音声インターフェースを構築することができる。

[0119] また、マスクタイプのウェアラブルインターフェース1900によれば、サイレントスピーチ、通常の音声、及びささやき声の3つのモダリティを使用する可能性がある。例えば、読唇術やささやき声を認識できれば、通常の発話と合わせて3種類の発話モダリティを得ることができる。

[0120] G. 先行技術との対比

このG項では、本出願人が調査した先行技術と本開示との比較結果及び本開示の優位な点について説明する。

[0121] 非特許文献9には、意図的に音声のピッチを制御することによって音声入力のモードを指定する「音声シフト」が開示されている。音声シフトでは、発話の基本周波数（F0）が指定された閾値を超えると別のモードと判断される。音声シフトでは、2つの音声入力モードを安定して認識するために、ユーザは不自然に高いピッチで話す必要がある。そもそも、ピッチを使い分ける周波数がユーザにとって分かり難い。これに対し、本開示によれば、ささやき声と通常の発話の切り替えはより自然であり、（ユーザ不明の）閾値を設定しなくてもより明確に行うことができる。

[0122] 非特許文献10には、音声コマンドの発話の躊躇現象の1つである、埋められた（発声された）一時停止を自動的に検出し、コマンドを埋めることができる候補を提案する方法が開示されている。この方法によれば、例えば、ユーザが「p l a y、B e e e e…」と言って停止すると、システムは「e e e e」を躊躇の声で埋められた一時停止として検出し、「B e a t l e s」や「B e a c h」などの埋められた候補を提案することができる。しかしながら、この方法は、音声における非言語的意図を示す可能性を示すが、躊躇を利用しているだけであり、本開示のように恣意的なコマンドを利用する

ことはできない。さらに、この方法において発声された一時停止を行うことは、子音の後ではなく、発話の母音の後でのみ可能である。

[0123] 非特許文献 11 に開示されている「PrivateTalk」という技術では、音声コマンドをアクティブにするために、片側から口を部分的に覆っている手を使用する。PrivateTalkの主な目的はプライバシーを保護することであるが、通常の発話（手で覆うことなく）とコマンド（手で覆うこと）を区別するためにも使用できる。しかしながら、明示的な手のジェスチャーが必要であるため、もはや「ハンズフリー」のインタラクションではないという点で本開示とは相違する。さらに、PrivateTalkではハンドカバーの効果を認識するために2つのマイク（左右のイヤホンに接続）が必要である。これに対し、本開示では、1つの標準マイクのみで十分である。また、本開示によれば、口を覆ってささやき声で話すという自然で且つ効果的な方法でプライバシーを保護することができる。

[0124] 非特許文献 12 に開示されている「DualBreath」という技術では、呼吸をコマンドとして使用し、鼻と口から同時に空気を吸い込むか吐き出すかを区別することで、通常の呼気と区別する。DualBreathは、ボタンを押すだけのトリガーを表現できるが、本開示のようにささやき声ほど豊富なコマンドを表現することはできない。

[0125] 非特許文献 13 に開示されている「ProxiMic」は、口の近くに配置されたマイクデバイスによってユーザの発話を検出するセンシング技術であり、スマートスピーカーや同様の音声制御デバイスにおける「Raise-to-Speak」のような「wake-up-free」の発話として使用することを目的としている。但し、ProxiMicは、マイクを口の近くに移動するといった物理的な動きが必要であるため、通常の発話と口の近くの発話を混合することは容易でない。

[0126] 非特許文献 14 には、吸い込みしながら発話する「吸気音」を使って音声を入力できる「SilentVoice」という技術が開示されている。SilentVoiceは、主に無音の発話用に設計されているが、通常の発

話と侵入音を区別することもできる。しかしながら、口のすぐ近くに特別なマイクを配置する必要がある、ユーザが吸気モードで正しく話すにはトレーニングが必要である。さらに、人が通常の発話と吸気発話を頻繁に切り替えることは困難である。

[0127] 米グーグル社がクラウドサービスとして提供しているCloud Speech-to-Textなどの一部の音声入力システムでは、「ピリオド」ということによって「。（句点）」を入力することができる。ただし、この発話を文中の「ピリオド」と区別するためには、前後に一定の休止を残す必要があり、全体の入力速度が低下してしまう。

[0128] ささやき声の認識に関する研究もいくつかあるが、いずれも通常の音声とささやき声の混合の使用に関しては考慮されていない。

[0129] Alexaスマートスピーカーはウィisperモードをサポートしている。このモードが設定されている場合、ささやき声でAlexaに話しかけるとAlexaもささやき声で応答するが、本開示のようにささやき声で音声コマンドを入力するというものではない。

[0130] H. 情報処理装置の構成

このH項では、本開示に係る音声入力システム100を実現するために、さらには上記F項で紹介した本開示の各種アプリケーションを実現するために利用される情報処理装置について説明する。

[0131] 図20には、通常の音声とささやき声の分類とささやき声の認識を行い、又は、ささやき声の認識結果を利用する各種アプリケーションを実現する情報処理装置2000の構成例を示している。また、情報処理装置2000は、上記F-1項で説明した会議端末1700として活用することもできる。

[0132] 図20に示す情報処理装置2000は、CPU (Central Processing Unit) 2001と、ROM (Read Only Memory) 2002と、RAM (Random Access Memory) 2003と、ホストバス2004と、ブリッジ2005と、拡張バス2006と、インターフェース部2007と、入力部2008と、出力部2

009と、ストレージ部2010と、ドライブ2011と、通信部2013を含んでいる。

[0133] CPU2001は、演算処理装置及び制御装置として機能し、各種プログラムに従って情報処理装置2000の動作全般を制御する。ROM2002は、CPU2001が使用するプログラム（基本入出力システムなど）や演算パラメータを不揮発的に格納している。RAM2003は、CPU2001の実行において使用するプログラムをロードしたり、プログラム実行において適宜変化する作業データなどのパラメータを一時的に格納したりするのに使用される。RAM2003にロードしてCPU2001において実行するプログラムは、例えば各種アプリケーションプログラムやオペレーティングシステム（OS）などである。

[0134] CPU2001とROM2002とRAM2003は、CPUバスなどから構成されるホストバス2004により相互に接続されている。そして、CPU2001は、ROM2002及びRAM2003の協働的な動作により、OSが提供する実行環境下で各種アプリケーションプログラムを実行して、さまざまな機能やサービスを実現することができる。情報処理装置100がパーソナルコンピュータの場合、OSは例えば米マイクロソフト社のWindowsやUnixである。情報処理装置2000がスマートフォンやタブレットなどの情報端末の場合、OSは例えば米アップル社のiOS又は米グーグル社のAndroidである。また、アプリケーションプログラムには、通常の音声とささやき声の分類とささやき声の認識を行うアプリケーションや、ささやき声の認識結果を利用する各種アプリケーションが含まれるものとする。

[0135] ホストバス2004は、ブリッジ2005を介して拡張バス2006に接続されている。拡張バス2006は、例えばPCI（Peripheral Component Interconnect）バス又はPCI Expressであり、ブリッジ2005はPCI規格に基づく。但し、情報処理装置2000がホストバス2004、ブリッジ2005及び拡張バス20

０６によって回路コンポーネントを分離される構成する必要はなく、単一のバス（図示しない）によってほぼすべての回路コンポーネントが相互接続される実装であってもよい。

[0136] インターフェース部２００７は、拡張バス２００６の規格に則って、入力部２００８、出力部２００９、ストレージ部２０１０、ドライブ２０１１、及び通信部２０１３といった周辺装置を接続する。但し、図２０に示す周辺装置がすべて必須であるとは限らず、また図示しない周辺装置を情報処理装置２０００がさらに含んでもよい。また、周辺装置は情報処理装置２０００の本体に内蔵されていてもよいし、一部の周辺装置は情報処理装置２０００本体に外付け接続されていてもよい。

[0137] 入力部２００８は、ユーザからの入力に基づいて入力信号を生成し、ＣＰＵ２００１に出力する入力制御回路などから構成される。情報処理装置２０００がパーソナルコンピュータの場合、入力部２００８は、キーボードやマウス、タッチパネルを含んでもよく、さらにカメラやマイクを含んでもよい。また、情報処理装置２０００がスマートフォンやタブレットなどの情報端末の場合、入力部２００８は、例えばタッチパネルやカメラ、マイクロホンであるが、ボタンなどのその他の機械式的操作子をさらに含んでもよい。但し、通常の音声とささやき声を利用したハンズフリーの音声対話を実施する際には、マイク以外の入力装置はほぼ不要である。

[0138] 出力部２００９は、例えば、液晶ディスプレイ（ＬＣＤ）装置、有機ＥＬ（Electro-Luminescence）ディスプレイ装置、及びＬＥＤ（Light Emitting Diode）などの表示装置を含む。本実施形態のように、情報処理装置２０００上で音声対話を行う場合には、通常の音声によって入力されたテキストや、ささやき声を使って入力した記号などの特殊文字並びに削除や改行などの編集コマンドの実行結果を、表示装置を使って提示する。また、出力部２００９は、スピーカー及びヘッドホンなどの音声出力装置を含み、ＵＩ画面上で表示するユーザへのメッセージの少なくとも一部を音声メッセージとして出力するようにしてもよい。

[0139] ストレージ部2010は、CPU2001で実行されるプログラム（アプリケーション、OSなど）や各種データなどのファイルを格納する。ストレージ部2010が格納するデータとして、ニューラルネットワークのトレーニングのための通常の音声及びささやき声のコーパス（前述）を含んでもよい。ストレージ部2010は、例えば、SSD（Solid State Drive）やHDD（Hard Disk Drive）などの大容量記憶装置で構成されるが、外付けの記憶装置を含んでもよい。

[0140] リムーバブル記憶媒体2012は、例えばmicroSDカードのようなカートリッジ式で構成される記憶媒体である。ドライブ2011は、装填したリムーバブル記憶媒体113に対して読み出し及び書き込み動作を行う。ドライブ2011は、リムーバブル記録媒体2012から読み出したデータをRAM2003やストレージ部2010に出力したり、RAM2003やストレージ部2010上のデータをリムーバブル記録媒体2012に書き込んだりする。

[0141] 通信部2013は、Wi-Fi（登録商標）、Bluetooth（登録商標）や4Gや5Gなどのセルラー通信網などの無線通信を行うデバイスである。また、通信部2013は、USB（Universal Serial Bus）やHDMI（登録商標）（High-Definition Multimedia Interface）などの端子を備え、スキャナやプリンタなどのUSBデバイスやディスプレイなどとのHDMI（登録商標）通信を行う機能をさらに備えていてもよい。

[0142] 情報処理装置2000としてパーソナルコンピュータ（PC）を想定しているが、1台とは限らず、2台以上の装置に分散して図1に示した音声入力システム100を実現したり、あるいは、上記F項で紹介した各種アプリケーションの処理を実行したりするようにすることもできる。

[0143] 1. 結論

最後に、本開示の利点及び本開示によってもたらされる効果についてまとめておく。

[0144] (1) 本開示によれば、ささやき声で非テキストコマンドを入力し、通常の音声でテキストを入力することができる音声入力システムを実現することができる。本開示に係る音声入力システムによれば、通常の音声入力をテキスト入力に使用するとともに、ささやくだけでさまざまなコマンドを入力することができる。本開示を実現するために、通常のマイク以外の特別なハードウェアを必要とせず、音声認識が既に利用可能である幅広い状況で 사용할ことができる。

[0145] (2) さらに、本開示によれば、2つの有益なニューラルネットワークを提供することができる。1つはささやき声と通常の音声を区別することに使用できるニューラルネットワークであり、もう1つはささやき声を認識することによって使用できるニューラルネットワークである。後者に関しては、通常の音声で事前にトレーニングされたモデルをささやき声でファインチューニングすることによって、優れたささやき声認識精度を実現することができる。

先行技術文献

特許文献

[0146] 特許文献1：特開2015-219480号公報

特許文献2：特表2019-524331号公報

非特許文献

[0147] 非特許文献1：Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv [cs.CL] (June 2020).

非特許文献2：Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu. 2020. Applying Wav2vec2.0 to Speech Recognition in Various Low-resource Languages. (Dec. 2020). arXiv:2012.12121 [cs.CL]

非特許文献3：Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. (June 2021). arXiv:2106.07447 [cs.CL]

非特許文献4 : Boon Pang Lim. 2010. Computational differences between whispered and non-whispered speech. Ph.D. Dissertation. University of Illinois Urbana-Champaign.

非特許文献5 : Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 5206-5210. <https://doi.org/10.1109/ICASSP.2015.7178964>

非特許文献6 : K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In Proceedings of the 28th ACM International Conference on Multimedia (Seattle, WA, USA) (MM '20). Association for Computing Machinery, New York, NY, USA, 484-492. <https://doi.org/10.1145/3394171.3413532>

非特許文献7 : Abhishek Niranjan, Mukesh Sharma, Sai Bharath Chandra Gupta, and M Ali Basha Shaik. 2020. End-to-End Whisper to Natural Speech Conversion using Modified Transformer Network. <https://doi.org/10.48550/ARXIV.2004.09347>

非特許文献8 : Santiago Pascual, Antonio Bonafonte, Joan Serra, and Jose A. Gonzalez. 2018. Whispered-to-voiced Alaryngeal Speech Conversion with Generative Adversarial Networks. <https://doi.org/10.48550/ARXIV.1808.10687>

非特許文献9 : Masataka Goto, Yukihiro Omoto, Katunobu Itou, and Tetsunori Kobayashi. 2003. Speech shift: direct speech-input-mode switching through intentional control of voice pitch. In 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 – INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003.

非特許文献10 : Masataka Goto. 1999. A Real-time Filled Pause Detection

System for Spontaneous Speech Recognition. Proc. of Eurospeech '99 (1999).

非特許文献11: Yukang Yan, Chun Yu, Yingtian Shi, and Minxing Xie. 2019. PrivateTalk: Activating Voice Input with Hand-On-Mouth Gesture Detected by Bluetooth Earphones. In Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (New Orleans, LA, USA) (UIST '19). Association for Computing Machinery, New York, NY, USA, 1013-1020. <https://doi.org/10.1145/3332165.3347950>

非特許文献12: Ryoya Onishi, Tao Morisaki, Shun Suzuki, Saya Mizutani, Takaaki Kamigaki, Masahiro Fujiwara, Yasutoshi Makino, and Hiroyuki Shinoda. 2021. DualBreath: Input Method Using Nasal and Mouth Breathing. In Augmented Humans Conference 2021 (Rovaniemi, Finland) (AHs'21). Association for Computing Machinery, New York, NY, USA, 283-285

非特許文献13: Yue Qin, Chun Yu, Zhaoheng Li, Mingyuan Zhong, Yukang Yan, and Yuanchun Shi. 2021. ProxiMic: Convenient Voice Activation via Close-to-Mic Speech Detected by a Single Microphone. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1-12.

非特許文献14: Masaaki Fukumoto. 2018. SilentVoice: Unnoticeable Voice Input by Ingressive Speech. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (Berlin, Germany) (UIST '18). Association for Computing Machinery, New York, NY, USA, 237-246.

産業上の利用可能性

[0148] 以上、特定の実施形態を参照しながら、本開示について詳細に説明してきた。しかしながら、本開示の要旨を逸脱しない範囲で当業者が該実施形態の修正や代用を成し得ることは自明である。

[0149] 本開示は、自動音声認識（ASR）を始め、遠隔会議におけるささやき声

を用いたコマンド、複数のアバターの音声切り替え、サイレントスピーチとの組み合わせなど、さまざまな用途に適用することができる。

[0150] 要するに、例示という形態により本開示について説明してきたのであり、本明細書の記載内容を限定的に解釈するべきではない。本開示の要旨を判断するためには、特許請求の範囲を参酌すべきである。

[0151] なお、本開示は、以下のような構成をとることも可能である。

[0152] (1) 発話音声を音声特徴量に基づいて通常の音声とささやき声に分類する分類部と、

前記分類部によって分類されたささやき声を認識する認識部と、

前記認識部の認識結果に基づく処理を制御する制御部と、

を具備する情報処理装置。

[0153] (2) 前記分類部は、第1の学習済みニューラルネットワークを用いて通常の音声とささやき声の分類を行い、

前記認識部は、第2の学習済みニューラルネットワークを用いてささやき声の認識を行う、

上記(1)に記載の情報処理装置。

[0154] (3) 前記第2の学習済みニューラルネットワークは、wave2vec2、0又はHuBERTからなる、

上記(2)に記載の情報処理装置。

[0155] (4) 前記第2の学習済みニューラルネットワークは、通常の音声コーパスで事前トレーニングされた後にささやき声によるファインチューニングが実施されている、

上記(2)又は(3)のいずれか1つに記載の情報処理装置。

[0156] (5) 前記ファインチューニングは、汎用のささやき声コーパスを用いた第1段階のファインチューニングと、ユーザ毎のささやき声のデータベースを用いた第2段階のファインチューニングを含む、

上記(4)に記載の情報処理装置。

[0157] (6) 前記第2の学習済みニューラルネットワークは、特徴抽出層及びトラ

ンスフォーマ層を含み、

前記第1の学習済みニューラルネットワークは、前記第2の学習済みニューラルネットワークと前記特徴抽出層を共有して構成される、
上記(2)乃至(5)のいずれか1つに記載の情報処理装置。

[0158] (7) 前記分類部によって分類された通常の音声を認識する通常の音声認識部をさらに備え、

前記制御部は、前記通常の音声認識部の認識結果に対して前記認識部によるささやき声の認識結果に対応する処理を実施する、
上記(1)乃至(6)のいずれか1つに記載の情報処理装置。

[0159] (8) 前記制御部は、前記通常の音声認識部が通常の音声を変換したテキストに対して、前記認識部が認識したささやき声コマンドの処理を実行する、
上記(7)に記載の情報処理装置。

[0160] (9) 前記制御部は、前記ささやき声コマンドに基づいて、前記テキストに対する記号又は特殊文字の入力、テキスト変換候補の選択、テキストの削除、テキストの改行のうち少なくとも1つを実行する、
上記(8)に記載の情報処理装置。

[0161] (10) 通常の音声で複数の文字が発話された後にささやき声で「SPELL」と発話されたときに、前記認識部は単語の綴りの入力を指示するコマンドであることを認識し。前記制御部は前記複数の文字を発話された順に連結した単語を生成する、
上記(8)又は(9)のいずれか1つに記載の情報処理装置。

[0162] (11) 前記認識部が絵文字入力を指示するささやき声コマンドを認識したことに応答して、前記制御部はそのささやき声の直前の通常の音声を絵文字に変換する、
上記(8)乃至(10)のいずれか1つに記載の情報処理装置。

[0163] (12) 前記制御部は、ささやき声に分類された音声を元の発話音声から除去して、外部の装置に送信する、
上記(1)乃至(11)のいずれか1つに記載の情報処理装置。

- [0164] (13) 前記制御部は、前記分類部がささやき声に分類した区間の話者の映像の唇の部分が発話していないときの映像に置き換える処理を行う、
上記(12)に記載の情報処理装置。
- [0165] (14) 前記認識部が通常の音声に分類した音声に基づいて第1のアバターの音声を生成する第1の音声生成部と、
前記認識部がささやき声に分類した音声に基づいて第2のアバターの音声を生成する第2の音声生成部と、
をさらに備える上記(1)乃至(13)のいずれか1つに記載の情報処理装置。
- [0166] (15) 話者が装着するマスクに搭載されたマイク及びスピーカーと、
前記分類部によって通常の音声に分類された音声信号を増幅する増幅部と、
をさらに備え、
前記増幅部で増幅した音声信号を前記スピーカーから音声出力する、
上記(1)乃至(14)のいずれか1つに記載の情報処理装置。
- [0167] (16) 発話音声を音声特徴量に基づいて通常の音声とささやき声に分類する分類ステップと、
前記分類ステップにおいて分類されたささやき声を認識する認識ステップと、
前記認識ステップにおける認識結果に基づく処理を制御する制御ステップと、
を有する情報処理方法。
- [0168] (17) 発話音声を音声特徴量に基づいて通常の音声とささやき声に分類する分類部、
前記分類部によって分類されたささやき声を認識する認識部、
前記認識部の認識結果に基づく処理を制御する制御部、
としてコンピュータを機能させるようにコンピュータ可読形式で記述されたコンピュータプログラム。

符号の説明

- [0169] 100…音声入力システム、101…分類部
102…ささやき声認識部、103…通常の音声認識部
104…制御部
400…ささやき声認識ニューラルネットワーク
410…特徴抽出層、411…正規化層 (Layer Norm)
412…平均プーリング層 (Avg Pool)
413、414…FC (全結合) 層
415…出力層 (Log softmax)、420…トランスフォーマ層
430…Projection層、440…CTC層
701…正規化層 (gMLP)
702平均プーリング層 (Avg Pool)
703…FC (全結合) 層、704…出力層 (Log softmax)
1700…会議端末、1701…マイク、1702…カメラ
1703…データ処理部、1704…データ送信部
1800…アバター制御システム
1801…第1のアバター音声生成部
1802…第2のアバター音声生成部
1900…ウェアラブルインターフェース、1901…マイク
1902…スピーカー、1903…増幅部
2000…情報処理装置、2001…CPU、2002…ROM
2003…RAM、2004…ホストバス、2005…ブリッジ
2006…拡張バス、2007…インターフェース部
2008…入力部、2009…出力部、2010…ストレージ部
2011…ドライブ、2012…リムーバブル記録媒体
2013…通信部

請求の範囲

- [請求項1] 発話音声を音声特徴量に基づいて通常の音声とささやき声に分類する分類部と、
前記分類部によって分類されたささやき声を認識する認識部と、
前記認識部の認識結果に基づく処理を制御する制御部と、
を具備する情報処理装置。
- [請求項2] 前記分類部は、第1の学習済みニューラルネットワークを用いて通常の音声とささやき声の分類を行い、
前記認識部は、第2の学習済みニューラルネットワークを用いてささやき声の認識を行う、
請求項1に記載の情報処理装置。
- [請求項3] 前記第2の学習済みニューラルネットワークは、wave2vec2.0又はHuBERTからなる、
請求項2に記載の情報処理装置。
- [請求項4] 前記第2の学習済みニューラルネットワークは、通常の音声コーパスで事前トレーニングされた後にささやき声によるファインチューニングが実施されている、
請求項2に記載の情報処理装置。
- [請求項5] 前記ファインチューニングは、汎用のささやき声コーパスを用いた第1段階のファインチューニングと、ユーザ毎のささやき声のデータベースを用いた第2段階のファインチューニングを含む、
請求項4に記載の情報処理装置。
- [請求項6] 前記第2の学習済みニューラルネットワークは、特徴抽出層及びトランスフォーマ層を含み、
前記第1の学習済みニューラルネットワークは、前記第2の学習済みニューラルネットワークと前記特徴抽出層を共有して構成される、
請求項2に記載の情報処理装置。
- [請求項7] 前記分類部によって分類された通常の音声を認識する通常音声認識

部をさらに備え、

前記制御部は、前記通常音声認識部の認識結果に対して前記認識部によるささやき声の認識結果に対応する処理を実施する、

請求項 1 に記載の情報処理装置。

[請求項8] 前記制御部は、前記通常音声認識部が通常の音声を変換したテキストに対して、前記認識部が認識したささやき声コマンドの処理を実行する、

請求項 7 に記載の情報処理装置。

[請求項9] 前記制御部は、前記ささやき声コマンドに基づいて、前記テキストに対する記号又は特殊文字の入力、テキスト変換候補の選択、テキストの削除、テキストの改行のうち少なくとも 1 つを実行する、

請求項 8 に記載の情報処理装置。

[請求項10] 通常の音声で複数の文字が発話された後にささやき声で「S P E L L」と発話されたときに、前記認識部は単語の綴りの入力を指示するコマンドであることを認識し。前記制御部は前記複数の文字を発話された順に連結した単語を生成する、

請求項 8 に記載の情報処理装置。

[請求項11] 前記認識部が絵文字入力を指示するささやき声コマンドを認識したことに応答して、前記制御部はそのささやき声の直前の通常の音声を絵文字に変換する、

請求項 8 に記載の情報処理装置。

[請求項12] 前記制御部は、ささやき声に分類された音声を元の発話音声から除去して、外部の装置に送信する、

請求項 1 に記載の情報処理装置。

[請求項13] 前記制御部は、前記分類部がささやき声に分類した区間の話者の映像の唇の部分を発話していないときの映像に置き換える処理を行う、

請求項 1 2 に記載の情報処理装置。

[請求項14] 前記認識部が通常の音声に分類した音声に基づいて第 1 のアバター

の音声を生成する第1の音声生成部と、

前記認識部がささやき声に分類した音声に基づいて第2のアバターの音声を生成する第2の音声生成部と、

をさらに備える請求項1に記載の情報処理装置。

[請求項15]

話者が装着するマスクに搭載されたマイク及びスピーカーと、

前記分類部によって通常の音声に分類された音声信号を増幅する増幅部と、

をさらに備え、

前記増幅部で増幅した音声信号を前記スピーカーから音声出力する

、

請求項1に記載の情報処理装置。

[請求項16]

発話音声を音声特徴量に基づいて通常の音声とささやき声に分類する分類ステップと、

前記分類ステップにおいて分類されたささやき声を認識する認識ステップと、

前記認識ステップにおける認識結果に基づく処理を制御する制御ステップと、

を有する情報処理方法。

[請求項17]

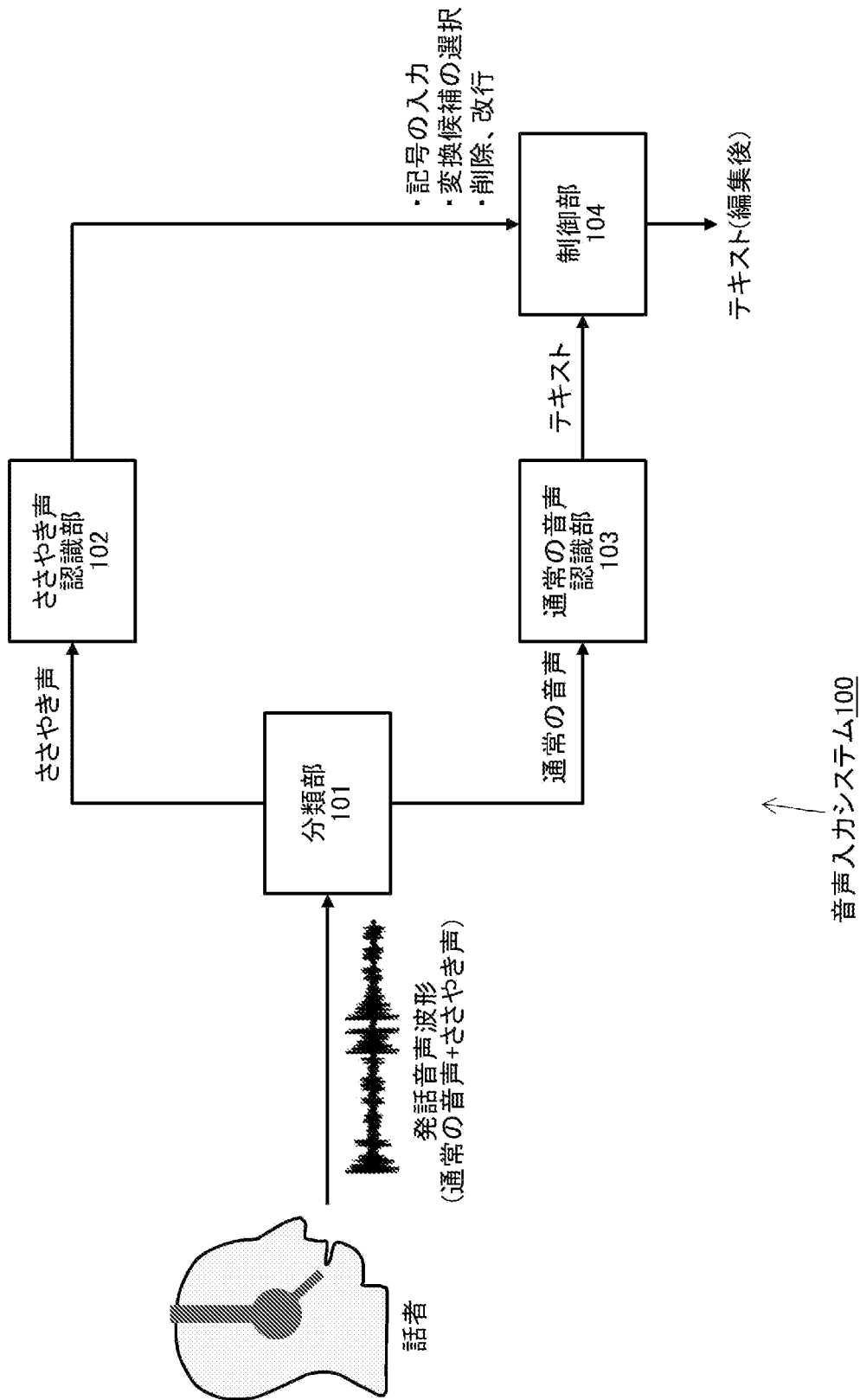
発話音声を音声特徴量に基づいて通常の音声とささやき声に分類する分類部、

前記分類部によって分類されたささやき声を認識する認識部、

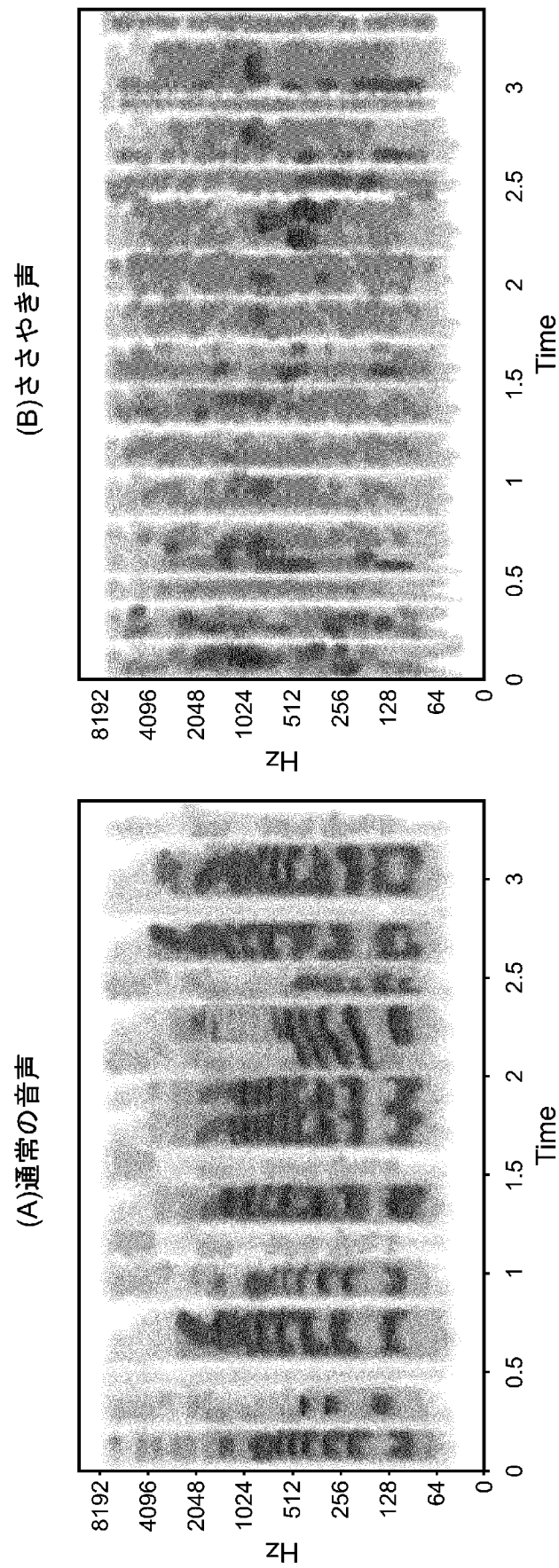
前記認識部の認識結果に基づく処理を制御する制御部、

としてコンピュータを機能させるようにコンピュータ可読形式で記述されたコンピュータプログラム。

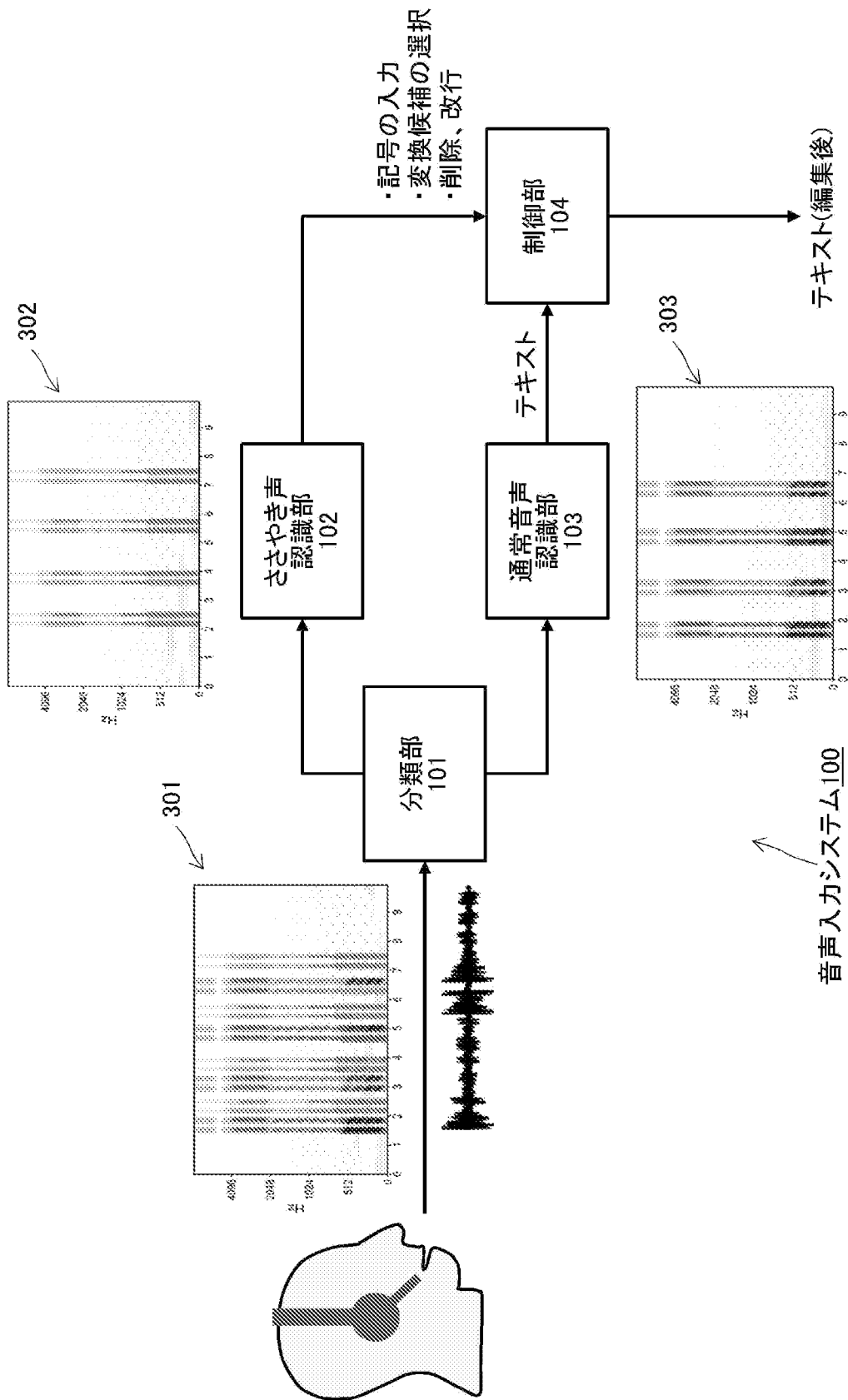
[図1]



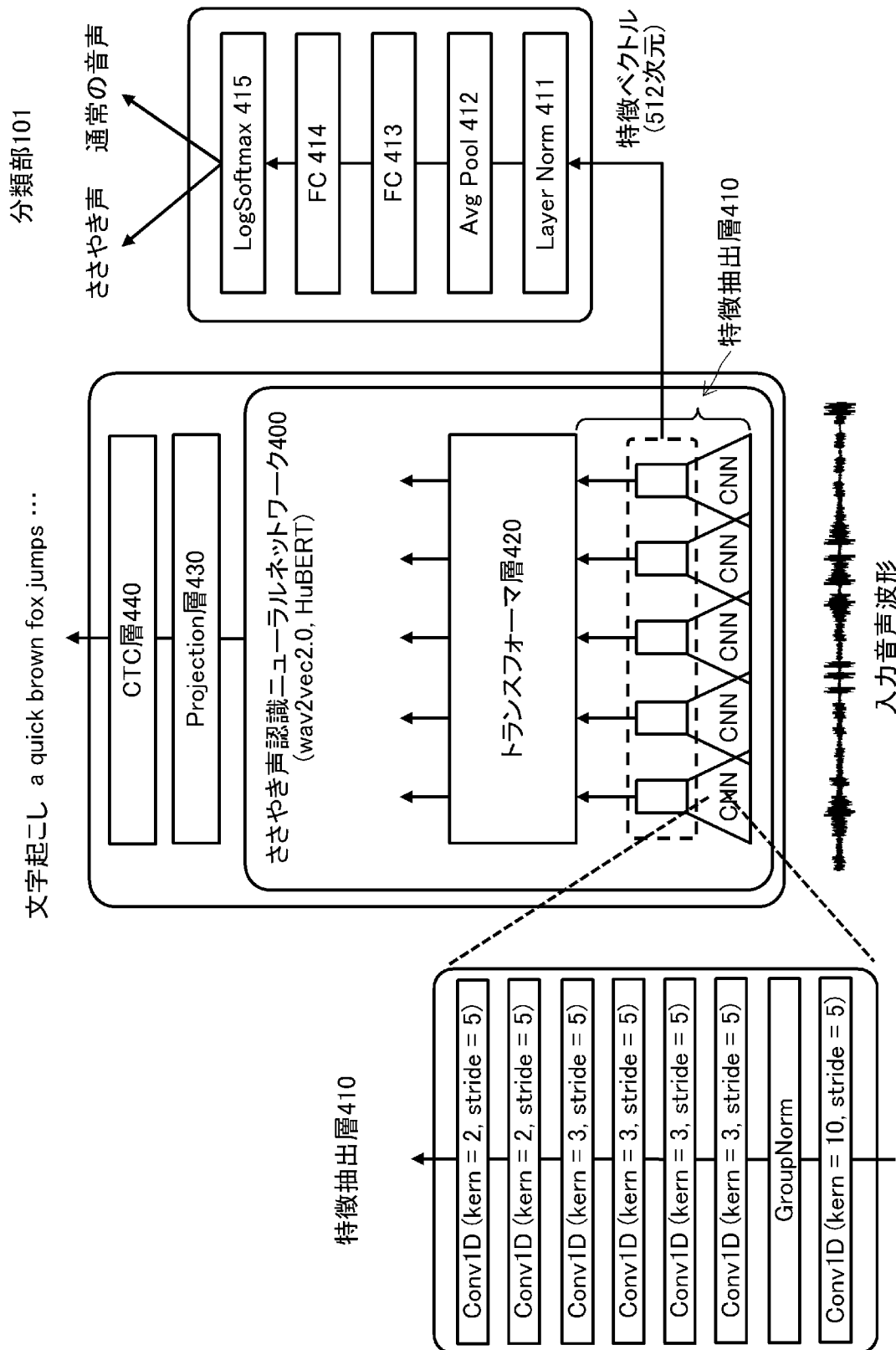
[図2]



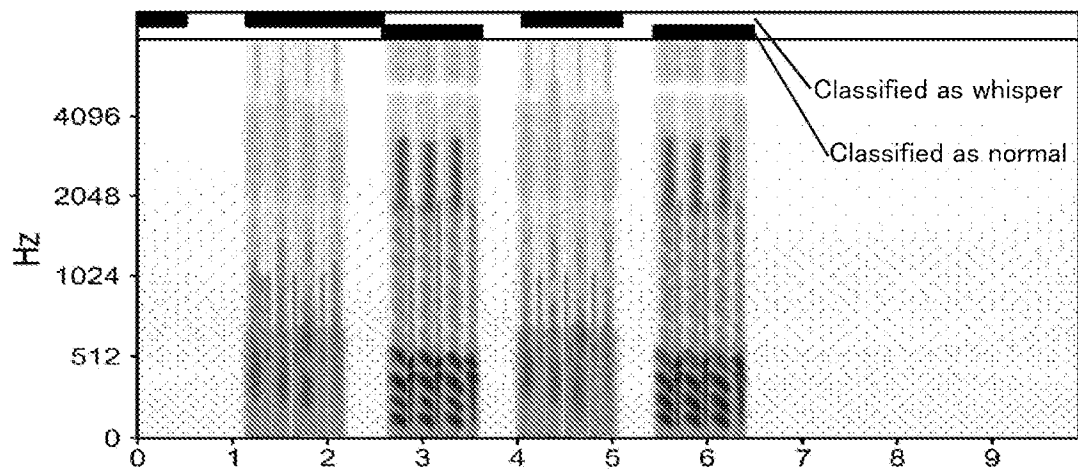
[図3]



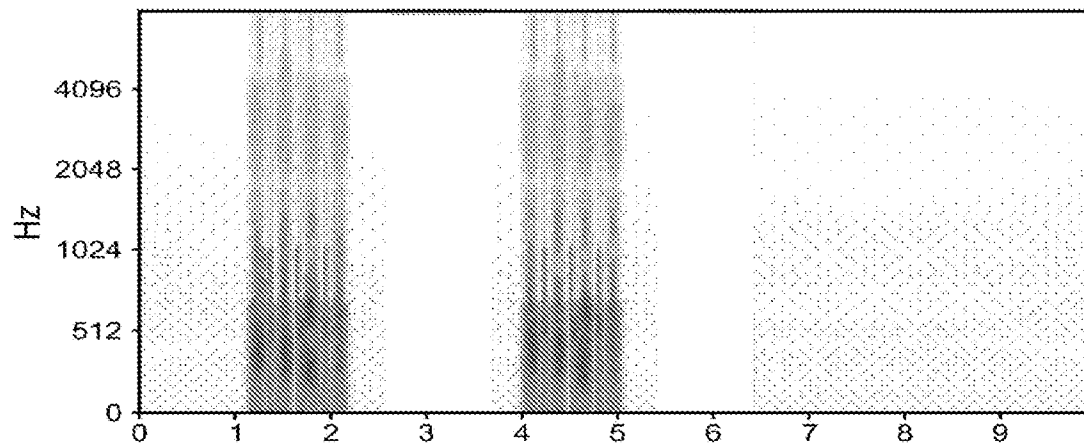
[図4]



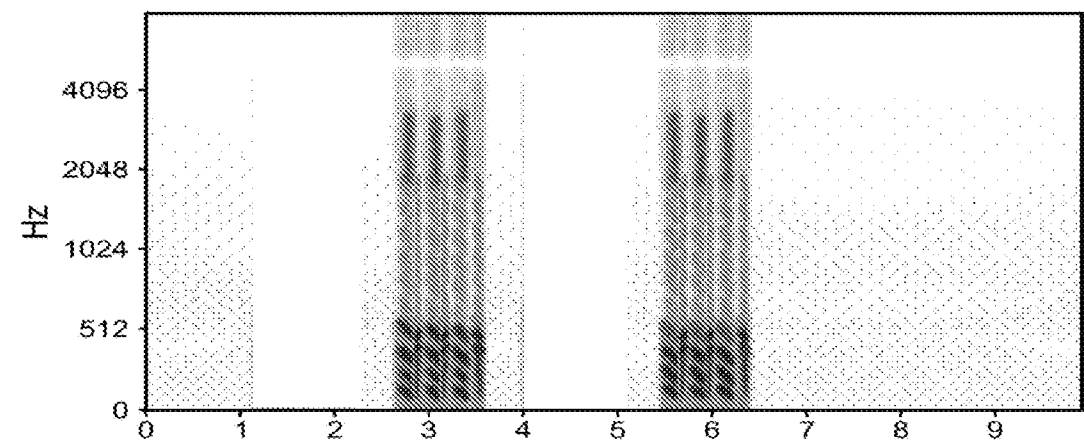
[図5]



(A) 入力音声の周波数スペクトログラム

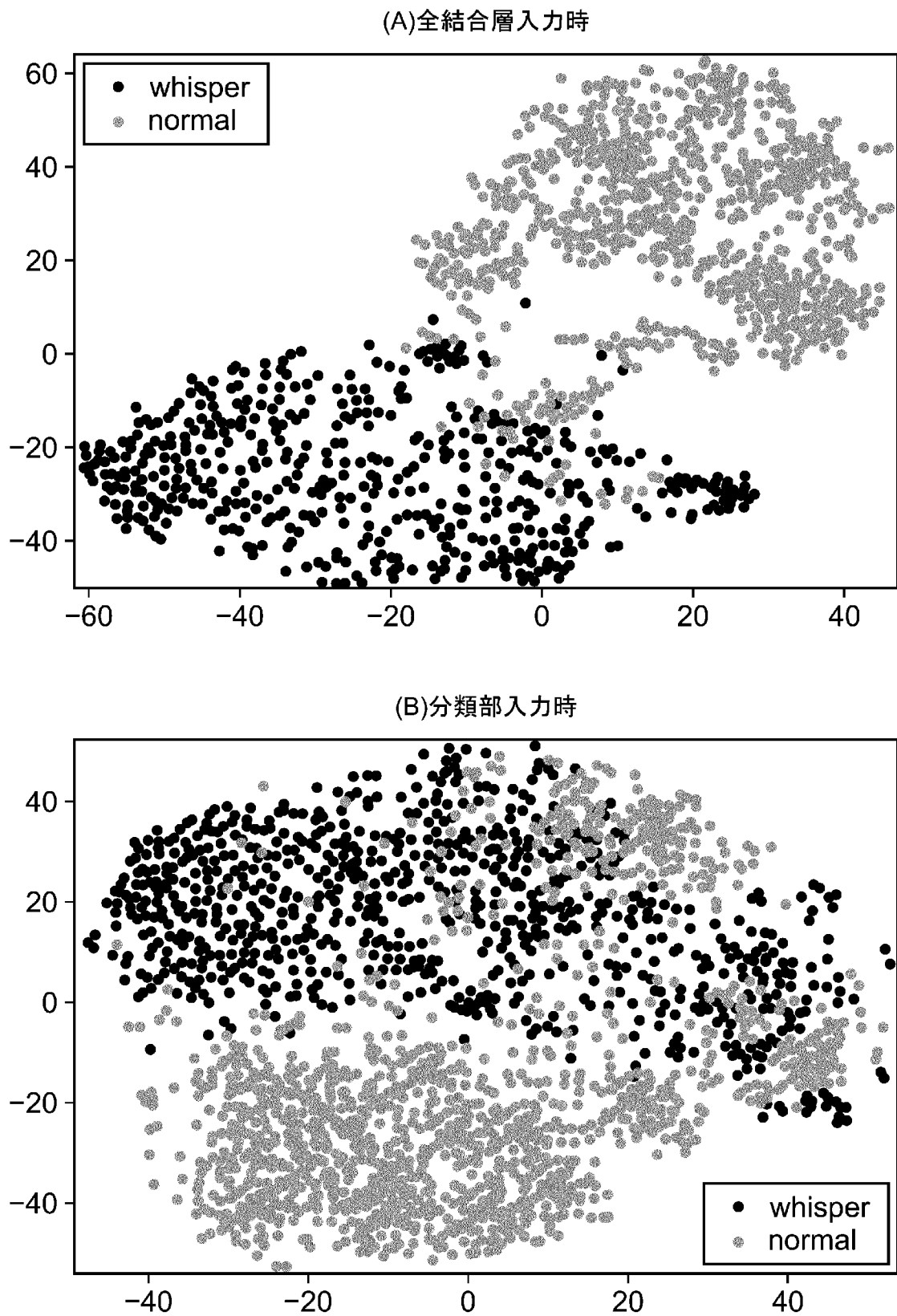


(B) 通常の音声除去後の周波数スペクトログラム

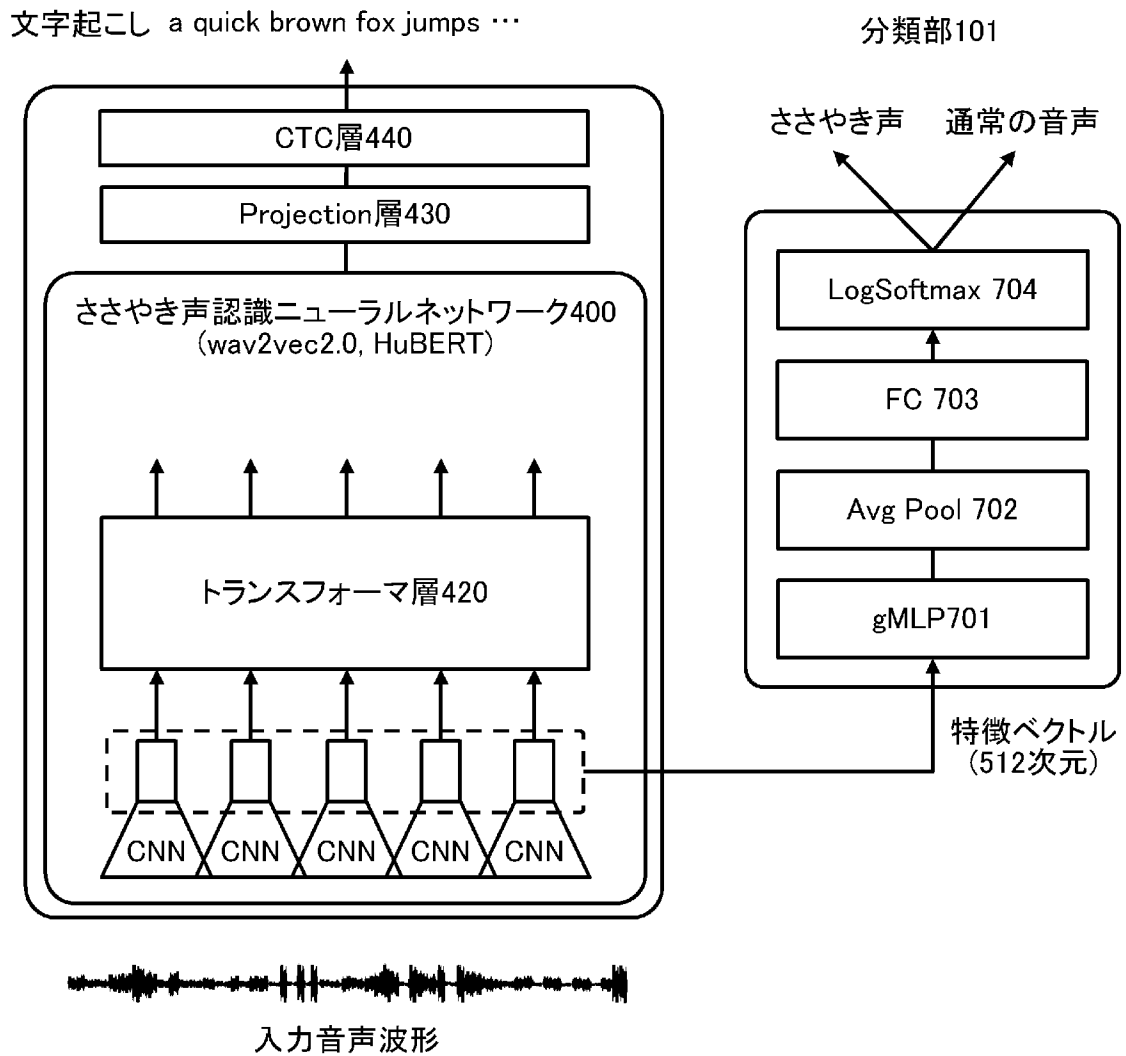


(C) ささやき声除去後の周波数スペクトログラム

[図6]



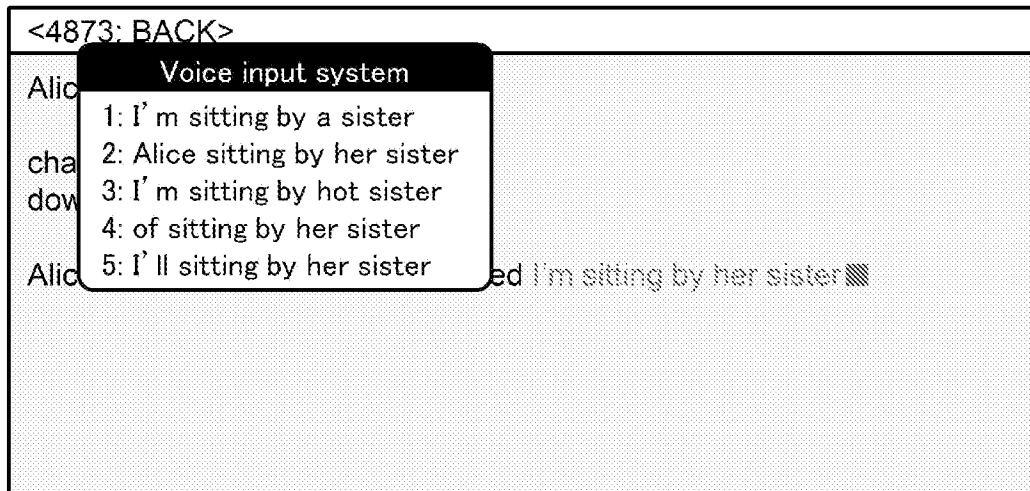
[図7]



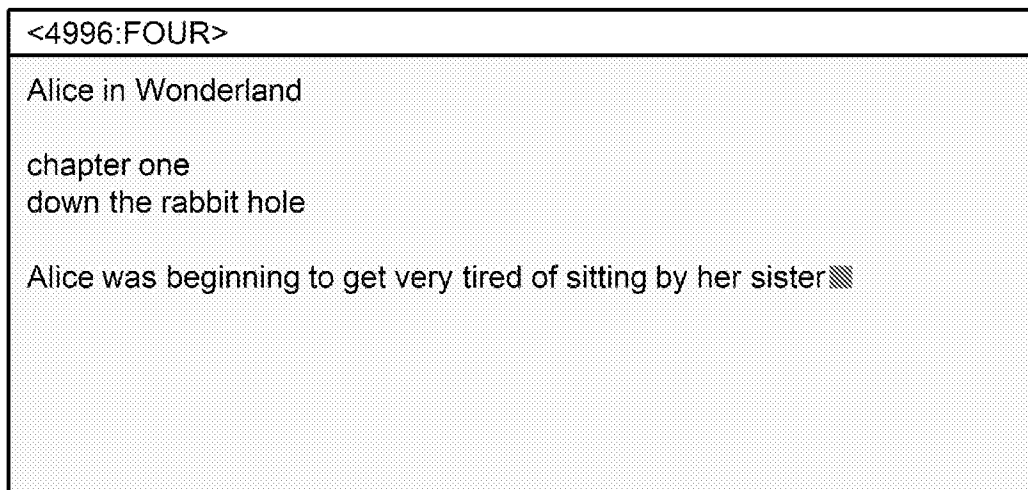
[図8]

<4873: BACK>
Alice in Wonderland
chapter one
down the rabbit hole
Alice was beginning to get very tired I'm sitting by her sister:■

[図9]



[図10]



[図11]

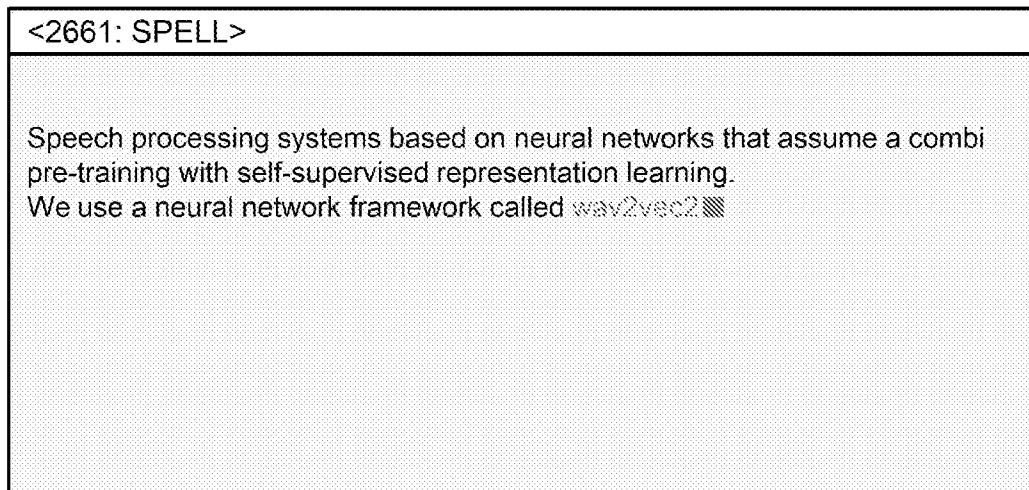
Alice was beginning to get very tired of sitting by her sister on the ...
Nothing to do: once or twice she had peeped into the book her ...
had no picture or conversations in it, " and what is

[図12]

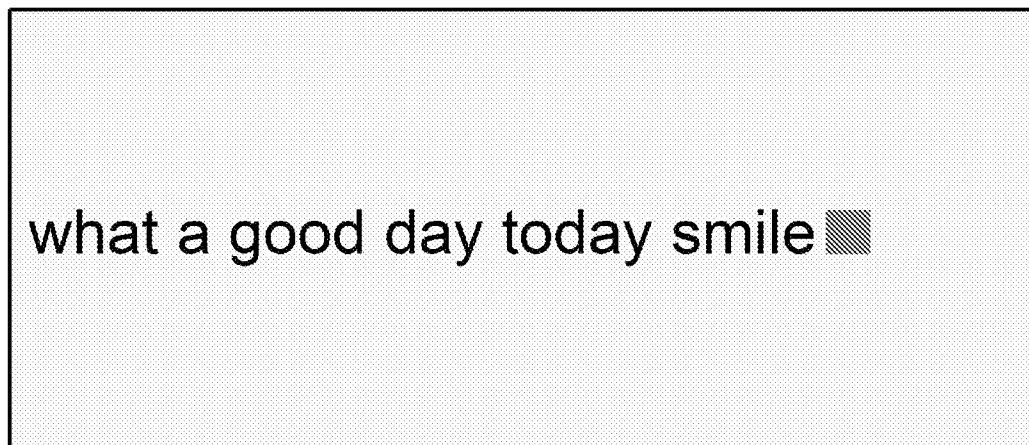
<3696: DELETE>

Speech processing systems based on neural networks that assume a combi
pre-training with self-supervised representation learning.
We use a neural network framework called `wav2vec2`

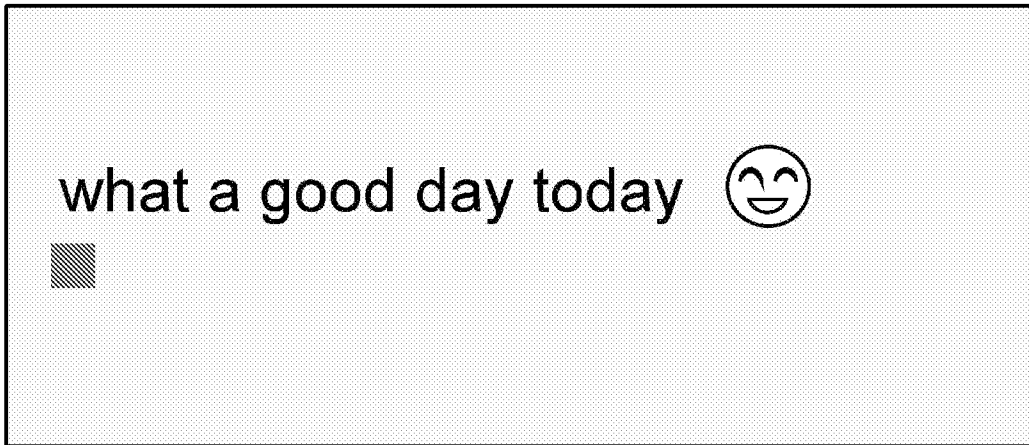
[図13]



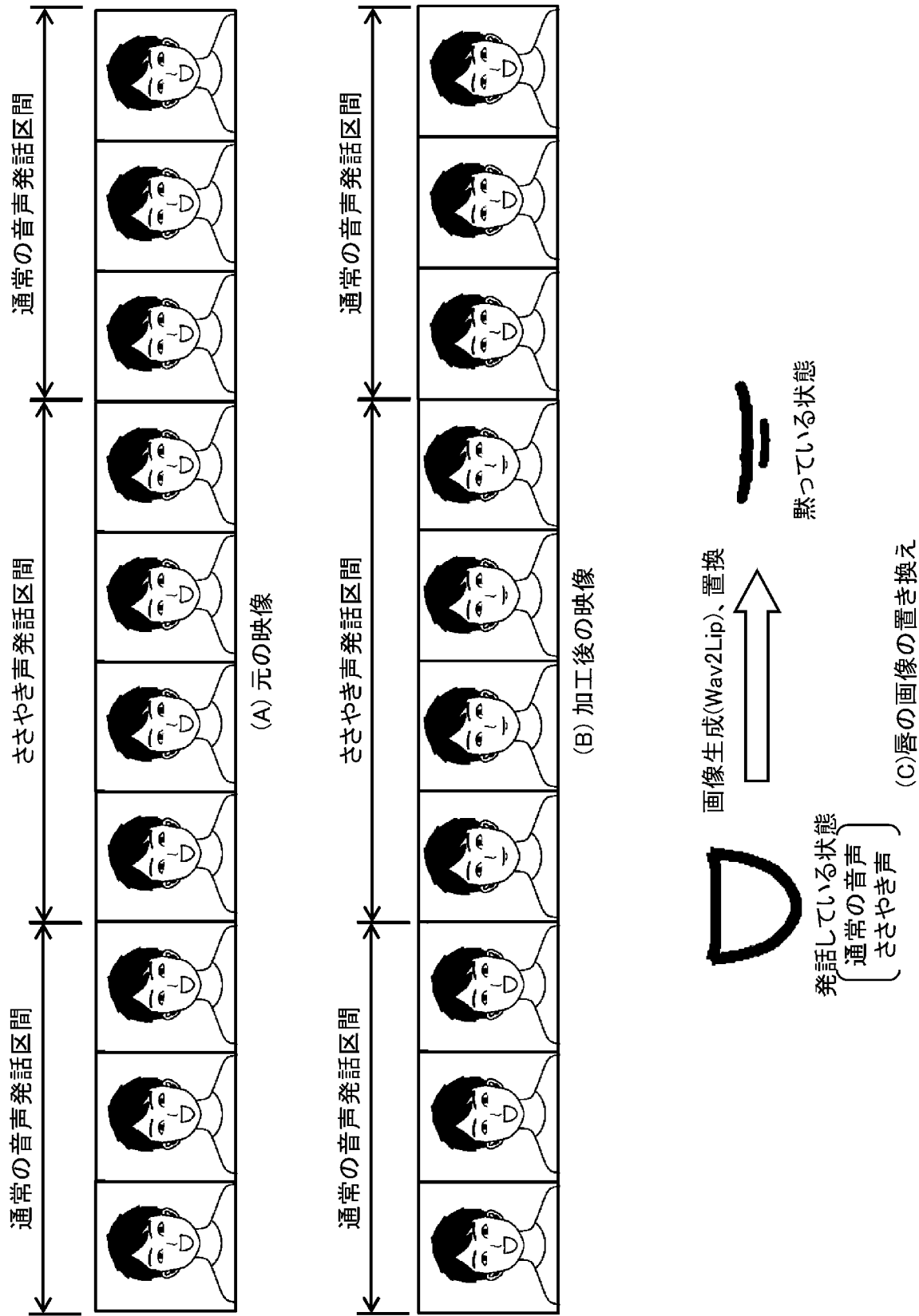
[図14]



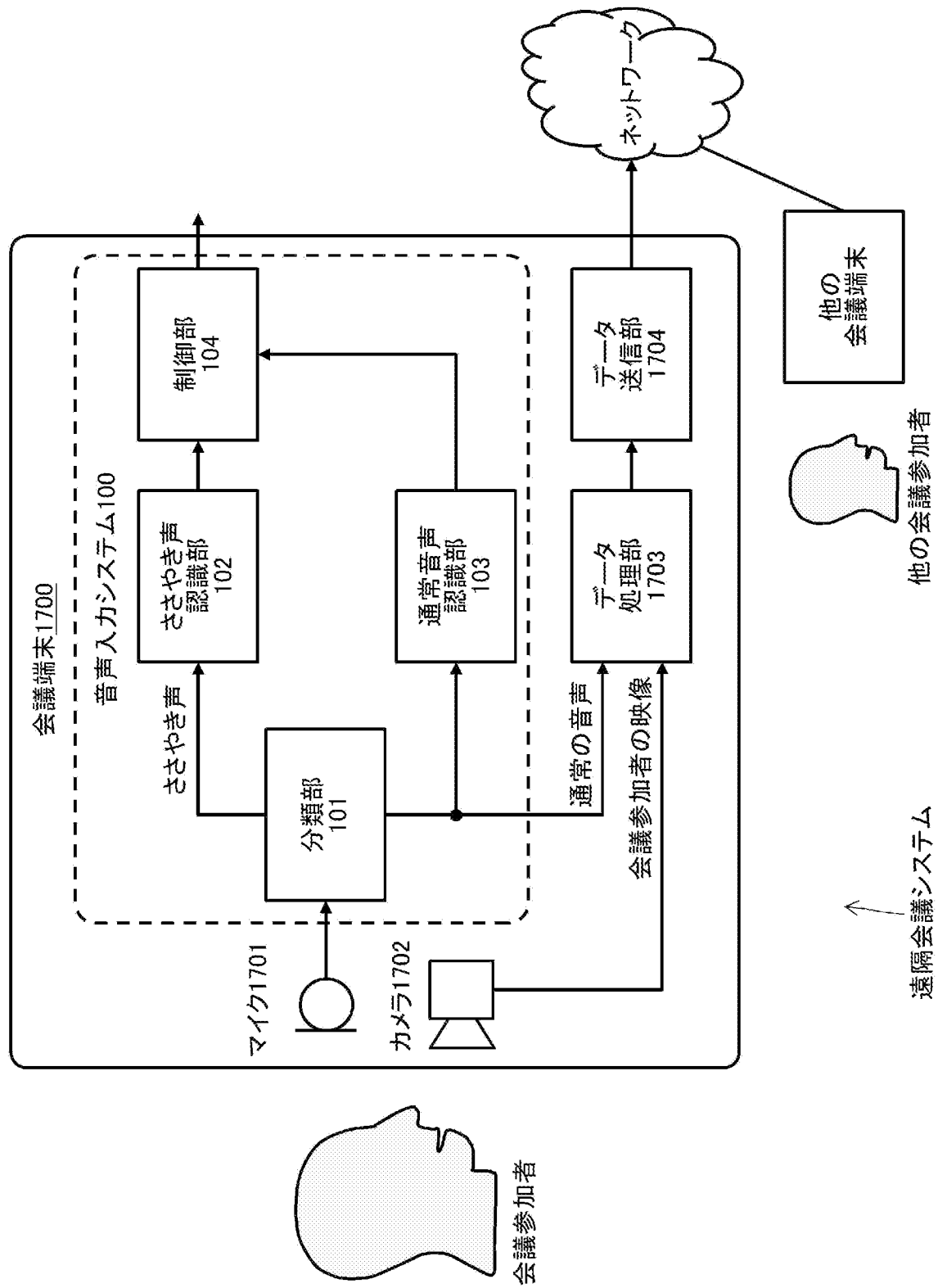
[図15]



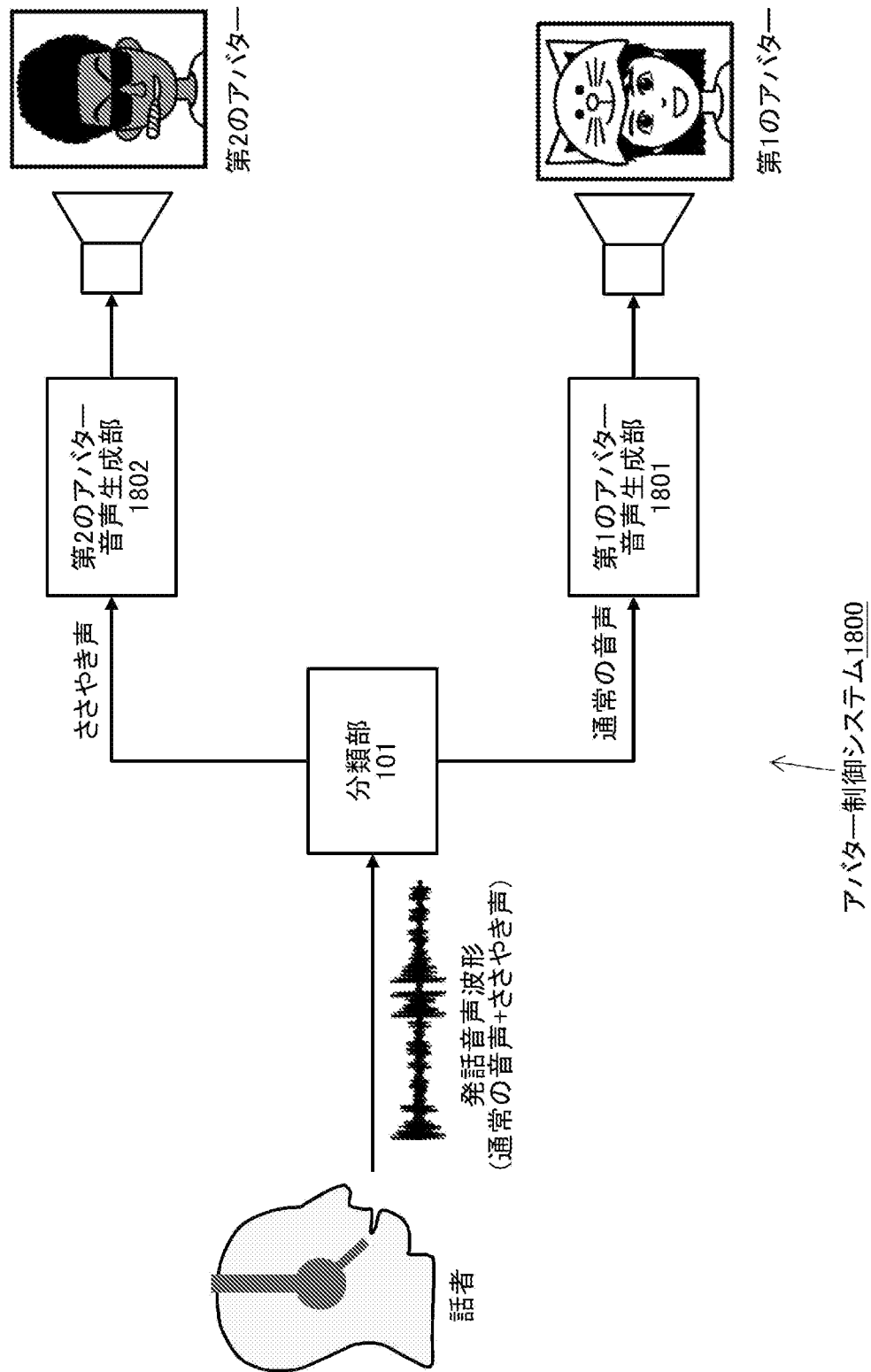
[図16]



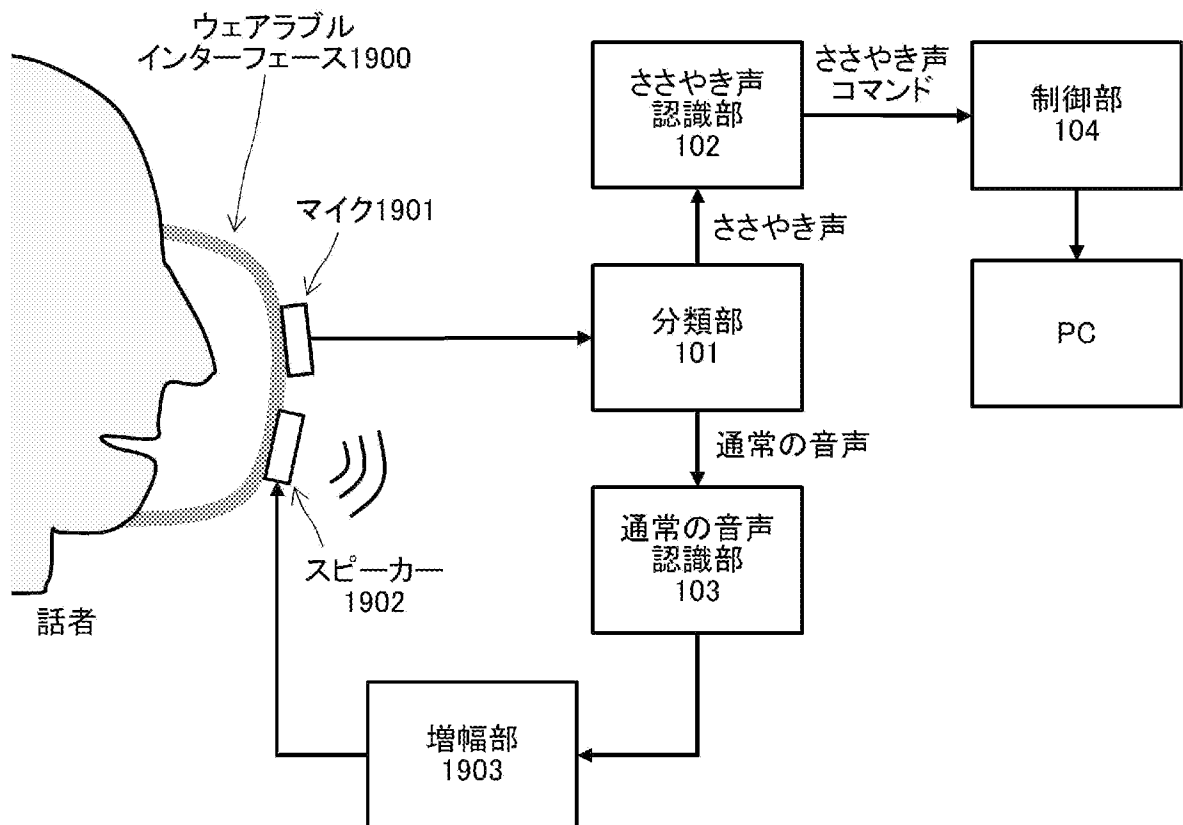
[図17]



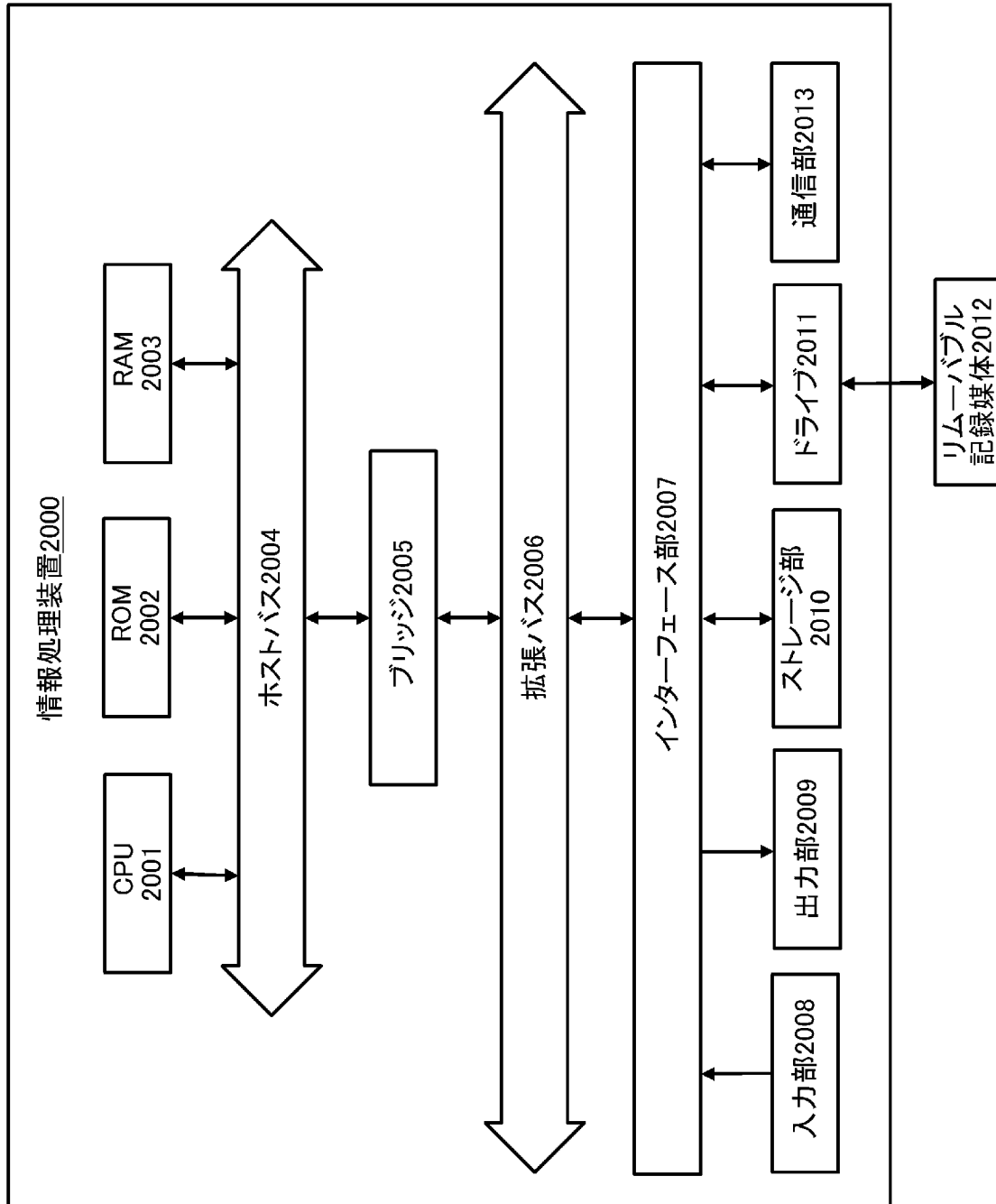
[図18]



[図19]



[図20]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2023/007479

A. CLASSIFICATION OF SUBJECT MATTER**G10L 15/32**(2013.01)i; **G10L 25/51**(2013.01)i

FI: G10L15/32 220Z; G10L25/51

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; A63F13/00-13/98

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 1971-2023
 Registered utility model specifications of Japan 1996-2023
 Published registered utility model applications of Japan 1994-2023

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	KR 10-2019-0133325 A (CAFE24 CORP.) 03 December 2019 (2019-12-03) paragraphs [0046]-[0057], [0074]-[0079]	1, 7, 12, 16-17
Y		2-4
A		5-6, 8-11, 13-15
Y	JP 2016-186515 A (NIPPON TELEGRAPH AND TELEPHONE CORP.) 27 October 2016 (2016-10-27) paragraph [0031]	2-4
Y	YOON, Ji Won et al. HuBERT-EE: Early Exiting HuBERT for Efficient Speech Recognition, [online], 13 April 2022, [retrieved on 24 April 2023], Internet:<URL: https://arxiv.org/pdf/2204.06328.pdf > in particular, abstract	3
Y	JP 2020-515877 A (IFLYTEK CO., LTD.) 28 May 2020 (2020-05-28) paragraphs [0047]-[0052]	4
A	JP 2005-140859 A (CANON KABUSHIKI KAISHA) 02 June 2005 (2005-06-02) entire text, all drawings	1-17

☒ 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

01 May 2023

Date of mailing of the international search report

16 May 2023

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.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2023/007479**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2001-079265 A (SEGA CORP.) 27 March 2001 (2001-03-27) entire text, all drawings	1-17
A	JP 2000-338986 A (CANON KABUSHIKI KAISHA) 08 December 2000 (2000-12-08) entire text, all drawings	1-17
A	US 2022/0054870 A1 (LACOMBE, Joseph) 24 February 2022 (2022-02-24) entire text, all drawings	1-17
P, X	REKIMOTO, Jun. DualVoice: Speech Interaction that Discriminates between Normal and Whispered Voice Input. Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, 28 October 2022, pp. 1-10 entire text, all drawings	1-17

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/JP2023/007479

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
KR	10-2019-0133325	A	03 December 2019	(Family: none)	
JP	2016-186515	A	27 October 2016	(Family: none)	
JP	2020-515877	A	28 May 2020	US 2020/0211550 A1 paragraphs [0104]-[0113] WO 2019/196196 A1 CN 108520741 A	
JP	2005-140859	A	02 June 2005	(Family: none)	
JP	2001-079265	A	27 March 2001	(Family: none)	
JP	2000-338986	A	08 December 2000	(Family: none)	
US	2022/0054870	A1	24 February 2022	(Family: none)	

A. 発明の属する分野の分類（国際特許分類（IPC）） G10L 15/32(2013.01)i; G10L 25/51(2013.01)i FI: G10L15/32 220Z; G10L25/51										
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; A63F13/00-13/98										
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										
<table border="0"> <tr> <td>日本国実用新案公報</td> <td>1922-1996年</td> </tr> <tr> <td>日本国公開実用新案公報</td> <td>1971-2023年</td> </tr> <tr> <td>日本国実用新案登録公報</td> <td>1996-2023年</td> </tr> <tr> <td>日本国登録実用新案公報</td> <td>1994-2023年</td> </tr> </table>			日本国実用新案公報	1922-1996年	日本国公開実用新案公報	1971-2023年	日本国実用新案登録公報	1996-2023年	日本国登録実用新案公報	1994-2023年
日本国実用新案公報	1922-1996年									
日本国公開実用新案公報	1971-2023年									
日本国実用新案登録公報	1996-2023年									
日本国登録実用新案公報	1994-2023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	KR 10-2019-0133325 A (CAFE24 CORP.) 03.12.2019 (2019 - 12 - 03) [0046]-[0057], [0074]-[0079]	1, 7, 12, 16-17								
Y		2-4								
A		5-6, 8-11, 13-15								
Y	JP 2016-186515 A (日本電信電話株式会社) 27.10.2016 (2016 - 10 - 27) [0031]	2-4								
Y	YOON, Ji Won et al., HuBERT-EE: Early Exiting HuBERT for Efficient Speech Recognition, [online], 2022.04.13, [検索日2023.04.24], インターネット: <URL: https://arxiv.org/pdf/2204.06328.pdf > 特にAbstract	3								
Y	JP 2020-515877 A (アイフライテック カンパニー, リミテッド) 28.05.2020 (2020 - 05 - 28) [0047]-[0052]	4								
A	JP 2005-140859 A (キヤノン株式会社) 02.06.2005 (2005 - 06 - 02) 全文, 全図	1-17								
☒ 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。										
<table border="0"> <tr> <td> * 引用文献のカテゴリー “A” 特に関連のある文献ではなく、一般的な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 </td> <td> “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 </td> </tr> </table>			* 引用文献のカテゴリー “A” 特に関連のある文献ではなく、一般的な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						
* 引用文献のカテゴリー “A” 特に関連のある文献ではなく、一般的な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 01.05.2023		国際調査報告の発送日 16.05.2023								
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	JP 2001-079265 A (株式会社セガ) 27.03.2001 (2001 - 03 - 27) 全文, 全図	1-17
A	JP 2000-338986 A (キャノン株式会社) 08.12.2000 (2000 - 12 - 08) 全文, 全図	1-17
A	US 2022/0054870 A1 (LACOMBE, Joseph) 24.02.2022 (2022 - 02 - 24) 全文, 全図	1-17
P, X	REKIMOTO, Jun, DualVoice: Speech Interaction that Discriminates between Normal and Whispered Voice Input, Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, 2022.10.28, p.1-10 全文, 全図	1-17

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/JP2023/007479

引用文献			公表日	パテントファミリー文献	公表日
KR	10-2019-0133325	A	03.12.2019	(ファミリーなし)	
JP	2016-186515	A	27.10.2016	(ファミリーなし)	
JP	2020-515877	A	28.05.2020	US 2020/0211550 A1	
				[0104]-[0113]	
				WO 2019/196196 A1	
				CN 108520741 A	
JP	2005-140859	A	02.06.2005	(ファミリーなし)	
JP	2001-079265	A	27.03.2001	(ファミリーなし)	
JP	2000-338986	A	08.12.2000	(ファミリーなし)	
US	2022/0054870	A1	24.02.2022	(ファミリーなし)	