(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7548482号
(P7548482)

(45)発行日　令和6年9月10日(2024.9.10)

(24)登録日　令和6年9月2日(2024.9.2)

(51)国際特許分類

Ｇ１０Ｌ　25/51　(2013.01)	Ｇ１０Ｌ　25/51
Ｇ１０Ｌ　21/0272(2013.01)	Ｇ１０Ｌ　21/0272　１００Ｚ

請求項の数　14　（全23頁）

(21)出願番号　特願2023-528949(P2023-528949)
(86)(22)出願日　令和4年1月18日(2022.1.18)
(65)公表番号　特表2023-549411(P2023-549411A)
(43)公表日　令和5年11月24日(2023.11.24)
(86)国際出願番号　PCT/CN2022/072460
(87)国際公開番号　WO2022/156655
(87)国際公開日　令和4年7月28日(2022.7.28)
審査請求日　令和5年5月16日(2023.5.16)
(31)優先権主張番号　202110083388.6
(32)優先日　令和3年1月21日(2021.1.21)
(33)優先権主張国・地域又は機関
中国(CN)

(73)特許権者　514187420
テンセント・テクノロジー・（シェンジェン）・カンパニー・リミテッド
中華人民共和国 ５１８０５７ グアンドン，シェンジェン，ナンシャン・ディストリクト，ミッドウェスト・ディストリクト・オブ・ハイテックパーク ケジジョンギ・ロード テンセント・ビルディング ３５エフ
(74)代理人　100107766
弁理士　伊東 忠重
(74)代理人　100070150
弁理士　伊東 忠彦
(74)代理人　100135079
弁理士　宮崎 修

最終頁に続く

(54)【発明の名称】　音声通話の制御方法、装置、コンピュータプログラム及び電子機器

(57)【特許請求の範囲】

【請求項１】

サーバが実行する、音声通話の制御方法であって、

混合された通話音声を取得するステップであって、前記混合された通話音声は、少なくとも１つの分岐音声を含む、ステップと、

前記通話音声に対して周波数領域変換を行い、前記通話音声の周波数領域における各周波数点に対応するエネルギー情報を決定するステップと、

ニューラルネットワークに基づいて各周波数点における前記エネルギー情報に対して分離処理を行い、各周波数点における対応する各分岐音声の該周波数点の全てのエネルギーにおけるエネルギー占有比率を決定するステップと、

各周波数点における各分岐音声の前記エネルギー占有比率に基づいて前記通話音声に含まれる分岐音声の数を決定するステップと、

前記分岐音声の数に基づいて、通話音声の制御方式を設定して前記音声通話を制御するステップと、を含む、方法。

【請求項２】

前記通話音声に対して周波数領域変換を行い、前記通話音声の周波数領域における各周波数点に対応するエネルギー情報を決定するステップは、

前記通話音声に対してフレーム分割処理を行い、少なくとも１つのフレームの音声情報を取得するステップと、

各フレームの前記音声情報に対して周波数領域変換を行い、周波数領域の音声エネルギ

ースペクトルを取得するステップと、
前記音声エネルギースペクトルに基づいて、前記通話音声の周波数領域における各周波数点に対応するエネルギー情報を決定するステップと、を含む、請求項１に記載の方法。
【請求項３】
各フレームの前記音声情報に対して周波数領域変換を行い、周波数領域の音声エネルギースペクトルを取得するステップは、
時間領域の各フレームの前記音声情報に対してフーリエ変換を行い、各フレームの前記音声情報に対応する周波数領域の音声エネルギースペクトルを取得するステップ、を含む、請求項２に記載の方法。
【請求項４】
前記音声エネルギースペクトルに基づいて、前記通話音声の周波数領域における各周波数点に対応するエネルギー情報を決定するステップは、
前記音声エネルギースペクトルにおける各周波数点に対応する振幅に対してモジュラス求め処理を行い、前記音声エネルギースペクトルに対応する振幅スペクトルを取得するステップと、
前記振幅スペクトルの二乗値を求め、前記二乗値に対して対数演算を行い、前記通話音声の周波数領域における各周波数点に対応するエネルギー情報を生成するステップと、を含む、請求項２又は３に記載の方法。
【請求項５】
前記ニューラルネットワークは、長短期記憶ニューラルネットワークを含み、
ニューラルネットワークに基づいて各周波数点における前記エネルギー情報に対して分離処理を行い、各周波数点における対応する各分岐音声の該周波数点の全てのエネルギーにおけるエネルギー占有比率を決定するステップは、
前記エネルギー情報を予め設定された音声分離モデルに入力し、長短期記憶ニューラルネットワークに基づく畳み込み処理を行い、各周波数点における対応する分岐音声を決定するステップと、
各周波数点における対応する各分岐音声の該周波数点におけるエネルギー情報に基づいて、各周波数点における前記各分岐音声の該周波数点におけるエネルギー占有比率を決定するステップと、を含む、請求項１乃至４の何れかに記載の方法。
【請求項６】
単一音声に対応する第１の音声サンプル、及び前記単一音声を含む混合音声に対応する第２の音声サンプルを取得するステップと、
前記第１の音声サンプルから第１の音声特徴を抽出し、前記第２の音声サンプルから第２の音声特徴を抽出するステップと、
前記第２の音声特徴を、長短期記憶人工ニューラルネットワークに基づいて構築された音声分離モデルに入力し、前記第２の音声特徴から分離された予測音声、及び前記予測音声の前記第２の音声サンプルにおける対応する予測エネルギー占有比率を決定するステップと、
前記第１の音声サンプルの前記第２の音声サンプルにおける実際エネルギー占有比率と前記予測エネルギー占有比率との比較結果に基づいて、前記音声分離モデルのパラメータを更新するステップと、をさらに含む、請求項５に記載の方法。
【請求項７】
各周波数点における各分岐音声の前記エネルギー占有比率に基づいて前記通話音声に含まれる分岐音声の数を決定するステップは、
各分岐音声について、該分岐音声の各周波数点における対応するエネルギー占有比率に基づいて、該分岐音声の前記エネルギー占有比率の平均値を求めるステップと、
各分岐音声の前記平均値及び所定閾値に基づいて、前記通話音声に含まれる分岐音声の数を決定するステップと、を含む、請求項５又は６に記載の方法。
【請求項８】
各分岐音声の前記平均値及び所定閾値に基づいて、前記通話音声に含まれる分岐音声の

数を決定するステップは、
各分岐音声の前記平均値と前記所定閾値との差の絶対値が差閾値よりも小さい場合、前記分岐音声の数が複数であると判定するステップと、
各分岐音声の前記平均値と前記所定閾値との差の絶対値が前記差閾値以上である場合、前記分岐音声の数が１つであると判定するステップと、を含む、請求項７に記載の方法。
【請求項９】
前記分岐音声の数が複数であると判定された場合、前記分岐音声の数に基づいて、通話音声の制御方式を設定して前記音声通話を制御するステップは、
設定された音声抽出方式に基づいて、主要発言者の音声を抽出するステップ、を含む、請求項１乃至８の何れかに記載の方法。
【請求項１０】
前記分岐音声の数が複数であると判定された場合、設定された音声抽出方式に基づいて、主要発言者の音声を抽出するステップは、
各周波数点における複数の分岐音声のそれぞれに対応するエネルギー占有比率に基づいて、前記エネルギー占有比率のうちの最大値に対応する分岐音声を前記主要発言者の音声として認識するステップと、
前記エネルギー情報から前記主要発言者の音声に対応する周波数情報を決定するステップと、
前記周波数情報に基づいて前記通話音声から前記主要発言者の音声を抽出するステップと、を含む、請求項９に記載の方法。
【請求項１１】
前記分岐音声の数は、１つ又は少なくとも２つを含み、
前記分岐音声の数に基づいて、通話音声の制御方式を設定して前記音声通話を制御するステップは、
前記分岐音声の数が１つである場合、設定されたシングルトークのエコー処理方式に基づいて、前記分岐音声のエコー音声を認識し、前記エコー音声に対してシングルトークエコー除去を行うステップと、
前記分岐音声の数が少なくとも２つである場合、設定されたダブルトークのエコー処理方式に基づいて、前記分岐音声に対応するエコー音声をそれぞれ認識し、前記エコー音声に対してダブルトークエコー除去を行うステップと、を含む、請求項１乃至１０の何れかに記載の方法。
【請求項１２】
音声通話の制御装置であって、
混合された通話音声を取得する取得部であって、前記混合された通話音声は、少なくとも１つの分岐音声を含む、取得部と、
前記通話音声に対して周波数領域変換を行い、前記通話音声の周波数領域における各周波数点に対応するエネルギー情報を決定する変換部と、
ニューラルネットワークに基づいて各周波数点における前記エネルギー情報に対して分離処理を行い、各周波数点における対応する各分岐音声の該周波数点の全てのエネルギーにおけるエネルギー占有比率を決定する分離部と、
各周波数点における各分岐音声の前記エネルギー占有比率に基づいて前記通話音声に含まれる分岐音声の数を決定する数決定部と、
前記分岐音声の数に基づいて、通話音声の制御方式を設定して前記音声通話を制御する制御部と、を含む、装置。
【請求項１３】
プロセッサにより実行される際に、請求項１乃至１１の何れかに記載の音声通話の制御方法を実現する、コンピュータプログラム。
【請求項１４】
１つ又は複数のプロセッサと、
１つ又は複数のプログラムを記憶する記憶装置と、を含む電子機器であって、

前記１つ又は複数のプログラムは、前記１つ又は複数のプロセッサにより実行される際に、前記１つ又は複数のプロセッサに請求項１乃至１１の何れかに記載の音声通話の制御方法を実現させる、電子機器。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、２０２１年１月２１日に出願した出願番号が２０２１１００８３３８８．６であり、発明の名称が「音声通話の制御方法、装置、コンピュータ読み取り可能な記憶媒体及び電子機器」である中国特許出願に基づく優先権を主張し、その全ての内容を参照により本発明に援用する。

【０００２】

本発明の実施例は、コンピュータ技術の分野に関し、特に音声通話の制御方法、装置、コンピュータ読み取り可能な記憶媒体及び電子機器に関する。

【背景技術】

【０００３】

多くの音声通話のシナリオでは、その後の音声制御のために、発言者の数や音色などを判別する必要がある。関連技術では、多数のラベル付きの音声セグメントに基づいて、発言者シナリオの検出システムを訓練する。ここで、各セグメントのラベルは発言者の数であり、テストの際に１つの音声セグメントを与え、システムは現在の発言者の数を予測する。このような処理方式は、音声検出に長い遅延をもたらし、特にリアルタイムの通信シナリオにおいて、音声認識の効率を大幅に低下し、リアルタイムの音声制御効果に影響を与えてしまう。

【発明の概要】

【０００４】

本発明の実施例は、少なくともある程度で音声人数の検出精度を保証することができると共に、音声人数の識別効率及び音声通話の制御効率を向上させることができる、音声通話の制御方法、装置、コンピュータ読み取り可能な記憶媒体及び電子機器を提供する。

【０００５】

本発明の他の特徴及び利点は、以下の詳細な説明によって明らかになり、又は本発明の実施により部分的に明らかになるであろう。

【０００６】

本発明の実施例の１つの態様では、音声通話の制御方法であって、混合された通話音声を取得するステップであって、前記混合された通話音声は、少なくとも１つの分岐音声を含む、ステップと、前記通話音声に対して周波数領域変換を行い、前記通話音声の周波数領域における各周波数点に対応するエネルギー情報を決定するステップと、ニューラルネットワークに基づいて各周波数点における前記エネルギー情報に対して分離処理を行い、各周波数点における対応する各分岐音声の該周波数点の全てのエネルギーにおけるエネルギー占有比率を決定するステップと、各周波数点における各分岐音声の前記エネルギー占有比率に基づいて前記通話音声に含まれる分岐音声の数を決定するステップと、前記分岐音声の数に基づいて、通話音声の制御方式を設定して前記音声通話を制御するステップと、を含む、方法を提供する。

【０００７】

本発明の実施例の１つの態様では、音声通話の制御装置であって、混合された通話音声を取得する取得部であって、前記混合された通話音声は、少なくとも１つの分岐音声を含む、取得部と、前記通話音声に対して周波数領域変換を行い、前記通話音声の周波数領域における各周波数点に対応するエネルギー情報を決定する変換部と、ニューラルネットワークに基づいて各周波数点における前記エネルギー情報に対して分離処理を行い、各周波数点における対応する各分岐音声の該周波数点の全てのエネルギーにおけるエネルギー占有比率を決定する分離部と、各周波数点における各分岐音声の前記エネルギー占有比率に基づいて前記通話音声に含まれる分岐音声の数を決定する数決定部と、前記分岐音声の数

に基づいて、通話音声の制御方式を設定して前記音声通話を制御する制御部と、を含む、装置を提供する。

【０００８】

本発明の実施例の１つの態様では、コンピュータプログラムが記憶されたコンピュータ読み取り可能な記憶媒体であって、前記コンピュータプログラムは、プロセッサにより実行される際に、本発明の実施例に記載の音声通話の制御方法を実現する、記憶媒体を提供する。

【０００９】

本発明の実施例の１つの態様では、１つ又は複数のプロセッサと、１つ又は複数のプログラムを記憶する記憶装置と、を含む電子機器であって、前記１つ又は複数のプログラムは、前記１つ又は複数のプロセッサにより実行される際に、前記１つ又は複数のプロセッサに本発明の実施例に記載の音声通話の制御方法を実現させる、電子機器を提供する。

【００１０】

本発明の実施例の１つの態様では、コンピュータ読み取り可能な記憶媒体に記憶されたコンピュータ命令を含むコンピュータプログラム製品又はコンピュータプログラムであって、コンピュータ装置のプロセッサは、コンピュータ読み取り可能な記憶媒体からコンピュータ命令を読み取り、該コンピュータ命令を実行することによって、前記コンピュータ装置に本発明の実施例に記載の音声通話の制御方法を実行させる、コンピュータプログラム製品又はコンピュータプログラムを提供する。

【００１１】

なお、上記の一般的な説明及び後述の詳細な説明は、単なる例示的なもの及び解釈的なものであり、本発明を限定するものではない。

【図面の簡単な説明】

【００１２】

【図１】本発明の実施例の技術を適用可能な例示的なシステムアーキテクチャを示す概略図である。

【図２】本発明の幾つかの実施例に係る会議システムを概略的に示す概略図である。

【図３】本発明の幾つかの実施例に係る音声通話の制御方法を概略的に示すフローチャートである。

【図４】本発明の幾つかの実施例に係る音声分離の流れを概略的に示す概略図である。

【図５】本発明の幾つかの実施例に係る音声抽出を概略的に示す概略図である。

【図６】本発明の幾つかの実施例に係る会議音声抽出を概略的に示す概略図である。

【図７】本発明の幾つかの実施例に係るエコー除去の適用シナリオを概略的に示す図である。

【図８】本発明の幾つかの実施例に係るエコー除去を概略的に示す概略図である。

【図９】本発明の幾つかの実施例に係るエネルギー情報の抽出を概略的に示すフローチャートである。

【図１０】本発明の幾つかの実施例に係るエネルギー情報の抽出を概略的に示す概略図である。

【図１１】本発明の幾つかの実施例に係る分離モデルの訓練を概略的に示すフローチャートである。

【図１２】本発明の幾つかの実施例に係る主要発言者の設定のインターフェースを概略的に示す図である。

【図１３】本発明の幾つかの実施例に係る通信リソースの割り当てのインターフェース図を概略的に示す図である。

【図１４】本発明の幾つかの実施例に係る音声通話の制御装置を概略的に示すブロック図である。

【図１５】本発明の実施例の電子機器を実現可能なコンピュータシステムの構成を示す概略図である。

【発明を実施するための形態】

【００１３】
クラウドコンピューティング（ｃｌｏｕｄ　ｃｏｍｐｕｔｉｎｇ）は計算モードの一種であり、計算タスクを大量のコンピュータにより構成されたリソースプールに分布することによって、各種の応用システムが需要に応じて計算力、記憶空間及び情報サービスを取得できる。リソースを提供するネットワークは「クラウド」と呼ばれる。「クラウド」内のリソースは、ユーザの目には無限に拡張可能であり、いつでも利用可能であり、オンデマンドで、いつでも拡張可能であり、従量課金される。クラウドコンピューティングの基礎能力の提供者として、クラウドコンピューティングリソースプール（クラウドプラットフォームと略称し、一般的にＩａａＳ（Ｉｎｆｒａｓｔｒｕｃｔｕｒｅ　ａｓ　ａ　Ｓｅｒｖｉｃｅ：サービスとしてのインフラストラクチャ）プラットフォーム）と呼ばれ、リソースプールの中に多種類の仮想リソースを構成し、外部の顧客が選択して使用することができる。クラウドコンピューティングリソースプールは、主に計算機器（仮想化機器であり、オペレーティングシステムを含む）、ストレージ機器、ネットワーク機器を含む。論理的な機能別では、ＩａａＳ（Ｉｎｆｒａｓｔｒｕｃｔｕｒｅ　ａｓ　ａ　Ｓｅｒｖｉｃｅ：サービスとしてのインフラストラクチャ）層にＰａａＳ（Ｐｌａｔｆｏｒｍ　ａｓ　ａ　Ｓｅｒｖｉｃｅ：サービスとしてのプラットフォーム）層を配備し、ＰａａＳ層にＳａａＳ（Ｓｏｆｔｗａｒｅ　ａｓ　ａ　Ｓｅｒｖｉｃｅ：サービスとしてのソフトウェア）層を配備してもよいし、ＩａａＳにＳａａＳを直接配備してもよい。ＰａａＳは、例えばデータベース、ウェブコンテナなどのソフトウェアを実行するプラットフォームである。ＳａａＳは、例えばＷｅｂポータルやＳＭＳグループプロバイダなどの様々なサービスソフトウェアである。一般的には、ＳａａＳやＰａａＳはＩａａＳと比べて上位層である。

【００１４】
クラウドコールセンター（Ｃｌｏｕｄ　Ｃａｌｌ　Ｃｅｎｔｅｒ）は、クラウドコンピューティング技術に基づいて構築されたコールセンターシステムであり、企業はいかなるソフトウェア、ハードウェアシステムを購入する必要がなく、人員、場所などの基本条件を備えるだけで、迅速に自分のコールセンターを所有することができ、ソフトウェアとハードウェアのプラットフォーム、通信リソース、日常のメンテナンスとサービスはサーバ業者から提供される。建設周期が短く、投入が少なく、リスクが低く、配備が柔軟であり、システム容量の拡張性が強く、運営メンテナンスコストが低いなどの特徴がある。電話マーケティングセンター、顧客サービスセンターにかかわらず、企業は必要に応じてサービスをレンタルするだけで、機能が全面的で、安定的で、信頼性があり、座席が全国各地に分布でき、全国コールが接続できるコールセンターシステムを確立することができる。

【００１５】
本発明の実施例では、クラウドコールセンターの方式でセンターシステムをコールすることができ、同時に前記システムに音声通話の制御方法を埋め込み、コール中の音声制御を実現し、さらにクラウドコールセンターをよりスマート化させ、クラウドコールセンターの信頼性と安全性を向上させる。

【００１６】
クラウド会議は、クラウドコンピューティング技術に基づく効率的で、便利で、低コストの会議形式である。ユーザは、インターネットインターフェースを通じて、簡単で使いやすい操作を行うだけで、迅速かつ効率的に世界各地のチームと顧客と同時に音声、データファイルとビデオを共有することができ、会議中のデータの転送、処理などの複雑な技術はクラウド会議サービス業者がユーザの操作を助ける。現在、国内クラウド会議は、主にＳａａＳ（Ｓｏｆｔｗａｒｅ　ａｓ　ａ　Ｓｅｒｖｉｃｅ：サービスとしてのソフトウェア）モードを主体とするサービス内容に集中し、電話、ネットワーク、ビデオなどのサービス形式を含み、クラウドコンピューティングに基づくビデオ会議はクラウド会議と呼ばれる。クラウド会議時代に、データの転送、処理、保存は全てビデオ会議メーカーのコンピュータリソースによって処理され、ユーザは、高価なハードウェアとインストールの面倒なソフトウェアを購入する必要がなく、ブラウザを開き、対応するインターフェースを登録するだけで、効率的な遠隔会議を行うことができる。クラウド会議システムは、マル

チサーバの動的クラスタ配備をサポートし、かつ複数の高性能サーバを提供し、会議の安定性、安全性、可用性を大幅に向上させる。近年、ビデオ会議は、コミュニケーション効率を大幅に高め、コミュニケーションコストを持続的に低下させ、内部管理レベルのアップグレードをもたらすため、多くのユーザに歓迎され、すでに政府、軍隊、交通、運輸、金融、運営者、教育、企業などの各分野に広く応用されている。勿論、ビデオ会議がクラウドコンピューティングを利用した後、利便性、快速性、使いやすさの面で更に強い魅力があり、きっとビデオ会議応用の新しいクライマックスの到来につながる。

【００１７】

クラウド会議の応用シナリオにおいて、本発明の実施例は、音声通話に基づく制御方法をクラウド会議に応用することができ、クラウド会議の過程における音声通話をよりはっきりさせ、音声通信過程をよりインテリジェント化させ、さらに会議の効率を向上させる。

【００１８】

クラウドソーシャル（Ｃｌｏｕｄ　Ｓｏｃｉａｌ）は、Ｉｎｔｅｒｎｅｔ　ｏｆ　Ｔｈｉｎｇｓ（ＩｏＴ）、クラウドコンピューティングとモバイルインターネットの相互作用応用の仮想ソーシャルアプリケーションモデルの１つであり、有名な「リソース共有関係マップ」を創立することを目的とし、さらにネットソーシャルを展開する。クラウドソーシャルの主要な特徴は、大量の社会リソースを統一的に整合と評価し、１つのリソース有効プールを構成してユーザにオンデマンドでサービスを提供することである。共有に参加するユーザが増えれば増えるほど、生み出せる利用価値は大きくなる。

【００１９】

人工知能（Ａｒｔｉｆｉｃｉａｌ　Ｉｎｔｅｌｌｉｇｅｎｃｅ：ＡＩ）は、人間の知能をシミュレーションし、拡張し、環境を感知し、知識を取得し、知識を使用して最適な結果を得るために、デジタルコンピュータ又はデジタルコンピュータによって制御される機械を利用する理論、方法、技術及び応用システムである。言い換えれば、人工知能は、計算機科学の総合技術であり、知能の本質を理解し、人間の知能と同様の方法で反応する新しい知能機械を生産しようとする。人工知能は、各種の知能機械の設計原理と実現方法を研究し、機械に感知、推理と決定の機能を持たせる。人工知能技術は、１つの総合的な学科であり、領域が広く、ハードウェアレベルの技術もソフトウェアレベルの技術もある。人工知能基礎技術は一般的に例えばセンサー、専用人工知能チップ、クラウドコンピューティング、分散型ストレージ、ビッグデータ処理技術、操作／インタラクティブシステム、メカトロニクスなどの技術を含む。人工知能ソフトウェア技術は、主にコンピュータビジョン技術、音声処理技術、自然言語処理技術、及び機械学習／深層学習などの幾つかの方面を含む。

【００２０】

音声技術（Ｓｐｅｅｃｈ　Ｔｅｃｈｎｏｌｏｇｙ）のキーテクノロジーには、自動音声認識（ＡＳＲ）と音声合成（ＴＴＳ）、声紋認識がある。コンピュータが聞くことができ、見ることができ、話すことができ、感じることができるようにすることは、未来のヒューマンコンピュータインタラクションの発展方向であり、その中で音声は、未来の最も有望なヒューマンコンピュータインタラクション方式の１つとなっている。機械学習（Ｍａｃｈｉｎｅ　Ｌｅａｒｎｉｎｇ：ＭＬ）は１つの多分野の交差学科であり、確率論、統計学、近似論、凸分析、算法複雑度理論などの多学科に関わる。コンピュータがどのように人類の学習行為を模擬或いは実現するかを専念に研究し、新しい知識或いは技能を取得し、既存の知識構造を再組織し、絶えず自身の性能を改善させる。機械学習は、人工知能の核心であり、コンピュータに知能を持たせる根本的なルートであり、その応用は人工知能の各領域に及んでいる。機械学習とディープラーニングは、通常、人工ニューラルネットワーク、信頼ネットワーク、強化学習、遷移学習、帰納学習式教育学習などの技術を含む。

【００２１】

人工知能技術の研究と進歩に伴い、人工知能技術は多くの領域で研究と応用を展開し、例えば一般的なスマートホーム、スマートウエアデバイス、バーチャルアシスタント、スマートスピーカー、スマートマーケティング、無人運転、自動運転、ドローン、ロボット

、知能医療、知能カスタマーサービスなど、技術の発展に伴い、人工知能技術は更に多くの領域で応用され、ますます重要な価値を発揮すると信じている。

【００２２】

関連技術では、多数のラベル付きの音声セグメントに基づいて、発言者シナリオの検出システムを訓練する。ここで、各セグメントのラベルは発言者の数であり、テストの際に１つの音声セグメントを与え、システムは現在の発言者の数を予測する。しかし、このスキームは、検出過程において、現在の発言者数を判定するために多くのコンテキスト情報を必要とし、例えば、発言者の数を決定するために、比較的に長い時間の音声セグメントをデータ基礎として取り込む必要がある。このような処理方式は、音声検出に長い遅延をもたらし、特にリアルタイムの通信シナリオにおいて、音声認識の効率を大幅に低下し、リアルタイムの音声制御効果に影響を与えてしまう。

【００２３】

本発明の実施例は、人工知能の音声技術及び機械学習等の技術に関するものであり、これらの技術により、本発明の実施例に係る音声通話の制御方法をより正確にすることができる。具体的には、以下の実施例を参照しながら説明する。

【００２４】

図１は、本発明の実施例の技術を適用可能な例示的なシステムアーキテクチャを示す概略図である。

【００２５】

図１に示すように、システムアーキテクチャは、端末装置（図１に示すように、スマートフォン１０１、タブレットコンピュータ１０２及びポータブルコンピュータ１０３のうちの１つ以上であってもよいし、デスクトップコンピュータなどであってもよい）、ネットワーク１０４、及びサーバ１０５を含むことができる。ネットワーク１０４は、端末装置とサーバ１０５との間の通信リンクを提供するための媒体である。ネットワーク１０４は、有線通信リンク、無線通信リンクなどの様々な接続タイプを含むことができる。

【００２６】

なお、図１における端末装置、ネットワーク及びサーバの数は、単に例示的なものである。実装要件に応じて、任意の数の端末装置、ネットワーク、及びサーバを有することができる。例えば、サーバ１０５は、複数のサーバからなるサーバクラスタ等であってもよい。

【００２７】

なお、本実施例の各端末装置は、異なる通話用クラスタを対象とすることができ、通話用クラスタ内の参加者数は、１人、２人、又はそれ以上などであってもよい。例えば、ポータブルコンピュータ１０３を対象とする通話クラスタには複数の参加者が含まれてもよく、タブレットコンピュータ１０２を対象とする通話クラスタには他の参加者が含まれてもよく、ユーザはスマートフォン１０１を介して会議に参加してもよい。

【００２８】

一例として、会議の進行中に、複数のユーザ又は１人のユーザが端末装置を使用して会議通話を行うことができる。同時に、サーバ１０５は、ネットワーク１０４を介して端末装置との通話音声を取得し、通話音声に対して周波数領域変換を行い、通話音声の周波数領域に対応するエネルギー情報を決定することができる。ニューラルネットワークに基づいてエネルギー情報に対して分離処理を行い、通話音声に含まれる各分岐音声の通話音声におけるエネルギー占有比率を決定する。エネルギー占有比率に基づいて通話音声に含まれる分岐音声の数を決定する。分岐音声の数に基づいて、通話音声制御方式を設定して音声通話を制御する。

【００２９】

上記のスキームは、通話進行過程においてリアルタイムに通話音声を取得し、通話音声に対して周波数領域変換を行って通話音声の周波数領域に対応するエネルギー情報を決定し、その後、ニューラルネットワークに基づいてエネルギー情報に対して分離処理を行い、通話音声に含まれる各分岐音声の通話音声におけるエネルギー占有比率を決定し、エネ

ルギー占有比率に基づいて通話音声に含まれる分岐音声の数を決定し、最後に、分岐音声の数に基づいて、通話音声制御方式を設定することによって音声通話を制御する。これによって、音声通話過程における音声人数の即時検出、及び音声通話のリアルタイム制御を実現し、音声人数の検出精度を保証すると共に、音声人数の認識効率及び音声通話の制御効率を向上させることができる。

【００３０】

これに加えて、図２に示すように、本実施例では、１つの通話クラスタのみを対象として処理することも可能であり、該通話クラスタは、１つ、２つ、又は複数の参加者を含む。上述した音声通話の制御方法によれば、通話クラスタ内のリアルタイムの発言者の数を検出し、通話中の音声品質を保証すると共に、対応する通話制御を行うことで、通話効率を向上させることができる。

【００３１】

なお、本発明の実施例に係る音声通話の制御方法は、一般にサーバ１０５によって実行されるため、音声通話の制御装置は、一般にサーバ１０５に設けられる。しかしながら、本発明の他の実施例では、端末装置は、本発明の実施例に係る音声通話の制御スキームを実行するために、サーバと同様の機能を有することもできる。

【００３２】

なお、本実施例におけるサーバは、独立した物理サーバであってもよいし、複数の物理サーバからなるサーバクラスタや分散型システムであってもよいし、クラウドコンピューティングサービスを提供するクラウドサーバであってもよい。端末としては、スマートフォン、タブレットコンピュータ、ノートパソコン、デスクトップパソコン、スマートスピーカー、スマートウォッチ等が挙げられるが、これらに限定されない。端末及びサーバは、有線又は無線通信により直接又は間接に接続することができるが、本発明はここに限定されない。

【００３３】

以下は、本発明の実施例の技術的解決策の実施の詳細を説明する。

【００３４】

図３は、本発明の幾つかの実施例に係る音声通話の制御方法を概略的に示すフローチャートである。該音声通話の制御方法はサーバにより実行されてもよく、該サーバは図１に示すサーバであってもよい。図３に示すように、該音声通話の制御方法は、少なくともステップＳ１１０～ステップＳ１５０を含み、以下に詳細に説明する。

【００３５】

ステップＳ１１０において、混合された通話音声を取得し、該混合された通話音声は、少なくとも１つの分岐音声を含む。

【００３６】

本発明の幾つかの実施例では、通話クラスタが通話を行う間に、混合された通話音声を取得してもよい。本実施例の通話音声の長さは、制限されず、リアルタイムで取得された１フレームの通話音声であってもよいし、時間長が１秒間又は１分間の通話音声であってもよい。

【００３７】

例えば、該通話は、リアルタイム通信会議のシナリオであってもよい。該リアルタイム通信会議の間に、リアルタイムで通話音声を収集し、収集された通話音声に基づいて対応する認識処理を行い、生成された識別結果に基づいてその後の制御を行い、通話音声のリアルタイム制御の効果を達成する。

【００３８】

ステップＳ１２０において、該通話音声に対して周波数領域変換を行い、該通話音声の各周波数点に対応するエネルギー情報を決定する。

【００３９】

本発明の１つの実施例では、本実施例で取得される通話音声は、時間領域の通話音声であり、時間を独立変数とし、音量を従属変数とする音声信号である。本実施例では、通話

音声を取得した後、通話音声に対して周波数領域変換を行い、時間領域の音声信号を周波数領域の音声信号に変換して、通話音声の周波数領域でのエネルギー情報を表す。

【００４０】

図４に示すように、音声に基づいて分離された発言者シーンの分類フレームワークは、「信号前処理」段階において、本発明の実施例では、オーディオ信号を取得し、オーディオ信号を前処理することによって音響特徴を抽出し、オーディオ信号に対応する対数エネルギースペクトルを、通話音声の周波数領域における各周波数点に対応するエネルギー情報として生成する。

【００４１】

具体的には、本実施例におけるエネルギー情報は、通話音声の各周波数点に対応するエネルギー値、エネルギースペクトル等の情報を含んでもよい。本実施例では、エネルギー情報によって各周波数点のエネルギーなどの属性を評価し、エネルギー情報に基づいて各周波数点における対応する分岐音声を区別することができる。

【００４２】

ステップＳ１３０において、ニューラルネットワークに基づいて各周波数点におけるエネルギー情報に対して分離処理を行い、各周波数点における対応する各分岐音声の該周波数点の全てのエネルギーにおけるエネルギー占有比率を決定する。

【００４３】

本発明の１つの実施例では、各周波数点におけるエネルギー情報を取得した後、ニューラルネットワークに基づいてエネルギー情報に対して分離処理を行う。即ち、既に訓練された分離モデルにエネルギー情報を入力して、各分岐音声の該周波数点におけるエネルギー占有比率を取得し、該エネルギー占有比率は、音声分離に基づく周波数点係数とも呼ばれる。例えば、図４において、２つの分岐音声を一例にすると、１つの周波数点における２つの分岐音声の周波数点係数、即ち、周波数点係数P_A及び周波数点係数P_Bを取得することができる。なお、図４では、単なる２つの分岐音声を一例として、１つの周波数点における２つの周波数点係数を示しており、それぞれ発言者Ａと発言者Ｂに対応しているが、実際には、１つの周波数点で得られる周波数点係数は、同時に話している人の数と関連し、２つに限らない。また、エネルギー値が０でない各周波数点では、何れも音声分離に基づく周波数点係数が得られる。

【００４４】

具体的には、本実施例における周波数点係数は、１つの周波数点における発言者に対応するエネルギーの、該周波数点の全てのエネルギー情報に占める割合を表すものである。本実施例における音声分離の考え方は、周波数領域の各周波数点係数の方式に基づいており、ある発言者がある周波数点におけるエネルギー占有比率は、混合信号において予測された周波数点係数の大きさに正比例する。周波数点係数（Ｐ）の計算方式は、１つの周波数点におけるある発言者の音声エネルギー値（Ｅ）と該周波数点における混合発言者の音声エネルギー値との比である。二人（ＡとＢ）を仮定し、以下の式に従って１つの周波数点における発言者Ａの周波数点係数を計算する。

【００４５】

【数１】

$$P_A = \frac{E_A}{E_A + E_B}$$

以上の式の計算により１つの周波数点における周波数点係数P_A及び周波数点係数P_Bが得られた後、P_AがP_Bよりも大きい場合、該周波数点は発言者Ａが主に行うものであり、P_AがP_Bよりも小さい場合、該周波数点は発言者Ｂが主に行うものである。

【００４６】

上記の方法は、特に複数人が同時に発話するシーンにおいて、エネルギー情報を分離することによって、各周波数点における各分岐音声に対応するエネルギー占有比率を決定し

、エネルギー占有比率に基づいて各分岐音声の分布状況を決定し、音声数の認識の正確性とリアルタイム性を向上させることができる。

【００４７】

ステップＳ１４０において、各周波数点における各分岐音声のエネルギー占有比率に基づいて通話音声に含まれる分岐音声の数を決定する。

【００４８】

本発明の１つの実施例では、通話音声中の各周波数点における各分岐音声のエネルギー占有比率を決定した後、本実施例では、エネルギー占有比率に基づいて、通話音声に含まれる分岐音声の数を平均化することにより決定する。

【００４９】

本発明の１つの実施例では、発言者Ａの音声の各フレーム内の各周波数点に対応するエネルギー占有比率について、各フレーム内の各周波数点のエネルギー占有比率を平均化し、１フレーム時間内の安定したエネルギーの平均値を取得し、そして、所定の閾値に基づいて現在のエネルギーの平均値が一人の発言に対応するか、それとも複数の人の発言に対応するかを判定し、最後に、現在フレームの発言者数情報を出力する。例えば、各フレームの現在の発言者数に対応する離散的な０（一人の発言）又は１（複数の人の発言）を出力することができる。

【００５０】

本発明の１つの実施例では、収集された通話音声は多数のフレームからなり、１フレーム中に複数の周波数点が存在する。例えば、周波数点の個数は、フーリエ変換されたポイント数であってもよく、１フレーム中の周波数点の個数をｆとし、F_iはある発言者のその中のｉ番目の周波数点に対応するエネルギー占有比率、即ち周波数点係数であり、平均値を求めることによって、該発言者の該フレームに対応するエネルギー占有比率の平均値は、以下のようになる。

【００５１】

【数２】

$$\frac{1}{f}\sum_{i=0}^{i=f-1} Fi \text{ 。}$$

そして、各分岐音声に対する平均値と閾値とを比較することによって、同時に発話する人の数を決定することができる。例えば、２人の場合、$P_A+P_B=1$であるため、P_A及びP_Bのうちの何れかと閾値とを比較すればよい。例えば、P_Aを一例にすると、実際の音声人数判定では、その値が０又は１である場合、現在発話しているエネルギー（P_A又はP_B）が音声エネルギー全体を占めていることを示しているため、ある１人だけが発話している、即ちＢ又はＡが発話していることを意味する。この場合、発言者数は１である。その値が０．５である場合、２人とも同時に発話しており、且つその時の発話エネルギーの大きさが同一であることを意味する。この場合、発言者数は２である。計算されたエネルギー占有比率の平均値と所定の閾値とを比較することで、現在の分岐音声の数を決定することができる。実際の応用では、応用シナリオに応じて閾値の具体的な値を設定することができる。

【００５２】

本実施例では、上記の閾値検出方式により現在の分岐音声の数を決定し、リアルタイムにフレームレベルの非常に短い時間内で多発言者シナリオを判断し、リアルタイムに音声ストリームを処理することができる。また、多対多のラベルを使用し、音声情報を十分に活用し、シナリオ検出の正確率を向上させることができる。

【００５３】

ステップＳ１５０において、決定された分岐音声の数に基づいて、通話音声の制御方式を設定して前記音声通話を制御する。

【００５４】
本発明の１つの実施例では、分岐音声の数を決定した後、現在の分岐音声の数に基づいて現在の通話状況を決定し、さらに設定された通話音声制御方式により音声通話を制御してもよい。これによって、音声通話のリアルタイム制御を実現し、音声制御の精確性とリアルタイム性を向上させることができる。
【００５５】
例えば、図５に示すように、分岐音声の数が複数である場合、分岐音声の数に基づいて、背景発言者をフィルタリングにより除去し、主要な発言者のオーディオストリームのみを抽出する。この際、フロントエンドで現在発言者の数を検出する必要があり、現在の発言者の数が１よりも大きい場合、主要発言者抽出をオンにし、現在の発言者数が１であることを検出した場合、音声への損傷を回避するために、音声抽出をオフにする。
【００５６】
本実施例では、音声を抽出する過程では、各周波数点における複数の分岐音声のそれぞれに対応するエネルギー占有比率に基づいて、エネルギー占有比率のうちの最大値に対応する分岐音声を主要発言者の音声として認識し、エネルギー情報から主要発言者の音声に対応する周波数情報を決定し、周波数情報に基づいて通話音声から主要発言者の音声を抽出する。
【００５７】
図６に示すように、複数のユーザが発話しているシーンでは、上述した周波数検出方式により、そのうちの主要発言者、例えば図６におけるユーザ４を特定してもよい。また、音声通話を明瞭にするように、主要発言者の音声を抽出し、或いは他のユーザの音声をフィルタリングにより除去する。
【００５８】
以上のように、複数の人が話すシーンでその中の一人の主要発言者の発話音声を抽出することができ、通話中の音声をよりはっきりさせ、通話品質と効果を向上させることができる。
【００５９】
図７に示す音声を外部に出力する場合、通話者の一方は、他方から戻ってきた音声から自分のエコーを聞こえ、通話品質が低下するという問題がある。
【００６０】
この問題を避けるために、図８に示すように、通信相手側と自分側とが交互に話しているシングルトークのシナリオの場合、即ち、分岐音声の数が１つである場合、設定されたシングルトークのエコー処理方式に基づいて、分岐音声のエコー音声を認識し、エコー音声に対してシングルトークエコー除去を行う。
【００６１】
通信相手側と自分側とが同時に発話するダブルトークのシナリオの場合、即ち、分岐音声の数が少なくとも２つである場合、設定されたダブルトークのエコー処理方式に基づいて、分岐音声に対応するエコー音声をそれぞれ認識し、エコー音声に対してダブルトークエコー除去を行う。通信システムでは、できるだけ自分側の信号がエコー除去過程において最大限に保留できることを保証する。
【００６２】
上記のスキームは、通話進行過程においてリアルタイムに通話音声を取得し、通話音声に対して周波数領域変換を行って通話音声の周波数領域に対応するエネルギー情報を決定し、その後、ニューラルネットワークに基づいてエネルギー情報に対して分離処理を行い、通話音声に含まれる各分岐音声の通話音声におけるエネルギー占有比率を決定し、エネルギー占有比率に基づいて通話音声に含まれる分岐音声の数を決定し、最後に、分岐音声の数に基づいて、通話音声制御方式を設定することによって音声通話を制御する。これによって、音声通話過程における音声人数の即時検出、及び音声通話のリアルタイム制御を実現し、音声人数の検出精度を保証すると共に、音声人数の認識効率及び音声通話の制御効率を向上させることができる。

【００６３】
本発明の１つの実施例では、図９に示すように、ステップＳ１２０における通話音声に対して周波数領域変換を行い、通話音声の周波数領域における各周波数点に対応するエネルギー情報を決定するプロセスは、ステップＳ１２１０～ステップＳ１２３０を含み、以下のように詳細に説明する。
【００６４】
ステップＳ１２１０において、通話音声に対してフレーム分割処理を行い、少なくとも１つのフレームの音声情報を取得する。
【００６５】
ステップＳ１２２０において、各フレームの音声情報に対して周波数領域変換を行い、周波数領域の音声エネルギースペクトルを取得する。
【００６６】
ステップＳ１２３０において、音声エネルギースペクトルに基づいて、通話音声の周波数領域における各周波数点に対応するエネルギー情報を決定する。
【００６７】
本発明の１つの実施例では、各フレームの前記音声情報に対して周波数領域変換を行い、周波数領域の音声エネルギースペクトルを取得するステップは、時間領域の各フレームの音声情報に対してフーリエ変換（他の時間領域を周波数領域に変換する方式を含む）を行い、各フレームの音声情報に対応する周波数領域の音声エネルギースペクトルを取得する。
【００６８】
本実施例のステップＳ１２３０における音声エネルギースペクトルに基づいて、通話音声の周波数領域における各周波数点に対応するエネルギー情報を決定するステップは、音声エネルギースペクトルにおける各周波数点に対応する振幅に対してモジュラス求め処理を行い、音声エネルギースペクトルに対応する振幅スペクトルを取得するステップと、振幅スペクトルの二乗値を求め、二乗値に対して対数演算を行い、通話音声の周波数領域における各周波数点に対応するエネルギー情報を生成するステップと、を含む。
【００６９】
図１０に示すように、本発明の１つの実施例では、収集された時間領域音声に対してフレーム分割処理、ウィンドウ追加処理を行い、その後、フレーム毎にそれぞれＮ点フーリエ変換を行い、フーリエ変換して得られたＮ個の周波数点のフーリエ変換係数を求め、それに対してモジュラスを求めて周波数領域の振幅スペクトルを取得し、得られた振幅スペクトルに対して二乗を求めて対数エネルギースペクトルを取得し、音声のＮ個の周波数点におけるエネルギー情報を取得する。
【００７０】
本発明の１つの実施例では、ニューラルネットワークは、長短期記憶ニューラルネットワークを含む。ステップＳ１３０におけるニューラルネットワークに基づいて各周波数点におけるエネルギー情報に対して分離処理を行い、各周波数点における対応する各分岐音声の該周波数点の全てのエネルギーにおけるエネルギー占有比率を決定するステップは、エネルギー情報を予め設定された音声分離モデルに入力し、長短期記憶ニューラルネットワークに基づく畳み込み処理を行い、各周波数点における、通話音声に含まれる各分岐音声の該周波数点に対応するエネルギー占有比率を決定するステップを含む。
【００７１】
本発明の１つの実施例では、該方法は、図１１に示すように、音声分離モデルを訓練するプロセスにおいて、以下のステップをさらに含む。
【００７２】
ステップＳ１１１０において、単一音声に対応する第１の音声サンプル、及び単一音声を含む混合音声に対応する第２の音声サンプルを取得する。
【００７３】
ステップＳ１１２０において、第１の音声サンプルから第１の音声特徴を抽出し、第２

の音声サンプルから第２の音声特徴を抽出する。

【００７４】

ステップＳ１１３０において、第２の音声特徴を、長短期記憶人工ニューラルネットワークに基づいて構築された音声分離モデルに入力し、第２の音声特徴から分離された予測音声、及び予測音声の第２の音声サンプルにおける対応する予測エネルギー占有比率を決定する。

【００７５】

ステップＳ１１４０において、第１の音声サンプルの第２の音声サンプルにおける実際エネルギー占有比率と予測エネルギー占有比率との比較結果に基づいて、音声分離モデルのパラメータを更新する。

【００７６】

本発明の１つの実施例では、まず、訓練用のデータセットを構築して２つの音声ライブラリ、即ち、一人発話用のコーパスと複数人発話用のコーパスを取得する。ここで、一人発話用のコーパスには、単一の音声に対応する第１の音声サンプルが含まれ、複数人発話用のコーパスは、それぞれランダムに複数の一人発話用のセグメントを抽出して重ね合わせ、その後、それぞれこの２つのデータベースに対して信号前処理により音声中の対数エネルギースペクトル特徴を抽出し、さらに分離モデルを経てそれぞれ音声セグメントの周波数係数を取得し、さらに後処理により分離後の音声を取得する。本実施例では、抽出された対数エネルギースペクトルの特徴を入力とし、この特徴を２層の長短期記憶ネットワーク（Ｌｏｎｇ　Ｓｈｏｒｔ－Ｔｅｒｍ　Ｍｅｍｏｒｙ：ＬＳＴＭ）及び１層の出力層からなる分離モデルに入力して、周波数点係数を取得する。

【００７７】

なお、本実施例でＬＳＴＭネットワークを用いる理由は、現時点の入力を考慮することだけではなく、ネットワークに前の内容の記憶機能を付与するためである。同時に、本実施例のネットワーク構造における追加した入力ゲート、出力ゲート、忘却ゲート、細胞状態ユニットは、ＬＳＴＭのタイミングモデリング能力を著しく向上させ、より多くの情報を記憶することができ、データの長時間依存性を効果的に把握することができる。

【００７８】

モデル全体の訓練では、訓練セットにマルチ発言者とシングル発言者が含まれており、複数のシングル発言者音声を用いて加算してマルチ発言者音声を取得し、シングル発言者はミュートとの混合と見なすことができる。ここで、混合音声から音声特徴ａを抽出し、対応するきれいな音声から音声特徴ｂを抽出し、特徴ベクトルａを訓練入力とし、特徴ベクトルｂを訓練目標とし、モデル出力は周波数点係数ｍ、ｎとする。後処理により分離された音声を取得し、分離された音声ときれいな音声との誤差でＬＳＴＭモデルを訓練し、ＬＳＴＭモデルにおけるパラメータを調整することによって、最終の分離モデルを取得し、得られた分離モデルを更に精確で完全にすることができる。

【００７９】

上記の訓練プロセスで得られた音声分離モデルは、リアルタイムに多発言者シナリオをフレームレベルで短い時間で判断し、音声ストリームをリアルタイムに処理することができる。また、フレームにおける各周波数点に対応するラベルに基づいて、複数のラベルに複数の周波数点を対応付けることで、音声情報を十分に活用し、シナリオ検出の正確率を向上させることができる。

【００８０】

さらに、本実施例では、ステップＳ１１３０における第２の音声特徴を、長短期記憶人工ニューラルネットワークに基づいて構築された音声分離モデルに入力し、第２の音声特徴から分離された予測音声を決定するプロセスにおいて、得られた周波数係数に混合音声の周波数スペクトルを乗算し、さらに逆フーリエ変換を経て、混合信号の位相を結合し、分離された音声信号を取得することができる。

【００８１】

ステップＳ１４０では、各周波数点における各分岐音声のエネルギー占有比率に基づい

て通話音声に含まれる分岐音声の数を決定するステップは、各分岐音声について、該分岐音声の各周波数点における対応するエネルギー占有比率に基づいて、該分岐音声のエネルギー占有比率の平均値を求めるステップと、各分岐音声の平均値及び所定閾値に基づいて、通話音声に含まれる分岐音声の数を決定するステップと、を含む。

【００８２】

本発明の１つの実施例では、収集された通話音声は多数のフレームからなり、１フレーム中に複数の周波数点が存在する。１フレーム中の周波数点の個数をｆとし、Ｆｉはある発言者のその中のｉ番目の周波数点に対応するエネルギー占有比率、即ち周波数点係数であり、平均値を求めることによって、該発言者の該フレームに対応するエネルギー占有比率の平均値は、以下のようになる。

【００８３】

【数３】

$$\frac{1}{f}\sum_{i=0}^{i=f-1} Fi$$ 。

本発明の１つの実施例では、各分岐音声の平均値及び所定閾値に基づいて、通話音声に含まれる分岐音声の数を決定するステップは、各分岐音声の平均値と所定閾値との差の絶対値が差閾値よりも小さい場合、分岐音声の数が複数であると判定するステップと、各分岐音声の平均値と所定閾値との差の絶対値が差閾値以上である場合、分岐音声の数が１つであると判定するステップと、を含む。

【００８４】

具体的には、本実施例では、２人が同時に発話することを一例にすると、平均値が０に近いほど、又は１に近いほど、１人発話の確率が大きくなり、０．５に近いほど、２人同時発話の確率が大きくなる。閾値の決定は、具体的なタスクに応じて決定される。例えば、主要発言者抽出のアルゴリズムのように、１人発話の場合にアルゴリズムが音声への損傷を避けるために、１人発言者の誤検出率が低いと判断する必要があり、その場合には閾値は０又は１に近い値に設定してもよい。

【００８５】

図１２に示すように、実際の会議適用シナリオにおいて、現在の発言者数が複数の人であると検出された場合、検出された発言者又は参加者をインターフェースに表示し、ユーザがトリガした主要発言者設定の指示に応じて、そのうちの何れか１人又は複数の人を主要発言者として設定し、残りの人の音声をフィルタリングにより除去して、会議の通話品質を保証することができる。

【００８６】

図１３に示すように、会話制御サーバは、複数の音声会話が同時に行われている場合、発言者数の多い会議により多くの通信リソースを割り当て、通話の品質を保証することができる。

【００８７】

以下は、本発明の上記の実施例における音声通話の制御方法を実行可能な本発明の装置の実施例を説明する。なお、該装置は、コンピュータ装置内で実行される１つのコンピュータプログラム（プログラムコードを含む）であってもよく、例えば、該装置は、アプリケーションソフトウェアである。該装置は、本発明の実施例に係る方法における対応するステップを実行するために使用されてもよい。本発明の装置の実施例で開示されていない詳細について、本発明の上記の音声通話の制御方法の実施例を参照してもよい。

【００８８】

図１４は、本発明の幾つかの実施例に係る音声通話の制御装置を概略的に示すブロック図である。

【００８９】

図１４に示すように、本発明の１つの実施例に係る音声通話の制御装置１４００は、以下の各部を含む。取得部１４１０は、混合された通話音声を取得し、混合された通話音声は、少なくとも１つの分岐音声を含む。変換部１４２０は、通話音声に対して周波数領域変換を行い、通話音声の周波数領域における各周波数点に対応するエネルギー情報を決定する。分離部１４３０は、ニューラルネットワークに基づいて各周波数点におけるエネルギー情報に対して分離処理を行い、各周波数点における対応する各分岐音声の該周波数点の全てのエネルギーにおけるエネルギー占有比率を決定する。数決定部１４４０は、各周波数点における各分岐音声のエネルギー占有比率に基づいて通話音声に含まれる分岐音声の数を決定する。制御部１４５０は、分岐音声の数に基づいて、通話音声の制御方式を設定して音声通話を制御する。

【００９０】

本発明の幾つかの実施例では、上記の実施例をベースとして、変換部１４２０は、通話音声に対してフレーム分割処理を行い、少なくとも１つのフレームの音声情報を取得するフレーム分割部と、各フレームの音声情報に対して周波数領域変換を行い、周波数領域の音声エネルギースペクトルを取得する周波数領域変換部と、音声エネルギースペクトルに基づいて、通話音声の周波数領域における各周波数点に対応するエネルギー情報を決定するエネルギー決定部と、を含む。

【００９１】

本発明の幾つかの実施例では、上記の実施例をベースとして、周波数領域変換部は、時間領域の各フレームの音声情報に対してフーリエ変換を行い、各フレームの前記音声情報に対応する周波数領域の音声エネルギースペクトルを取得する。

【００９２】

本発明の幾つかの実施例では、上記の実施例をベースとして、エネルギー決定部は、音声エネルギースペクトルにおける各周波数点に対応する振幅に対してモジュラス求め処理を行い、音声エネルギースペクトルに対応する振幅スペクトルを取得し、振幅スペクトルの二乗値を求め、二乗値に対して対数演算を行い、通話音声の周波数領域における各周波数点に対応するエネルギー情報を生成する。

【００９３】

本発明の幾つかの実施例では、上記の実施例をベースとして、ニューラルネットワークは、長短期記憶ニューラルネットワークを含む。分離部１４３０は、エネルギー情報を予め設定された音声分離モデルに入力し、長短期記憶ニューラルネットワークに基づく畳み込み処理を行い、各周波数点における対応する分岐音声を決定し、各周波数点における対応する各分岐音声の該周波数点におけるエネルギー情報に基づいて、各周波数点における各分岐音声の該周波数点におけるエネルギー占有比率を決定する。

【００９４】

本発明の幾つかの実施例では、上記の実施例をベースとして、音声通話の制御装置１４００は、更新部をさらに含む。更新部は、単一音声に対応する第１の音声サンプル、及び単一音声を含む混合音声に対応する第２の音声サンプルを取得し、第１の音声サンプルから第１の音声特徴を抽出し、第２の音声サンプルから第２の音声特徴を抽出し、第２の音声特徴を、長短期記憶人工ニューラルネットワークに基づいて構築された音声分離モデルに入力し、第２の音声特徴から分離された予測音声、及び予測音声の第２の音声サンプルにおける対応する予測エネルギー占有比率を決定し、第１の音声サンプルの第２の音声サンプルにおける実際エネルギー占有比率と予測エネルギー占有比率との比較結果に基づいて、音声分離モデルのパラメータを更新する。

【００９５】

本発明の幾つかの実施例では、上記の実施例をベースとして、数決定部１４４０は、各分岐音声について、該分岐音声の各周波数点における対応するエネルギー占有比率に基づいて、該分岐音声のエネルギー占有比率の平均値を求める平均部と、各分岐音声の平均値及び所定閾値に基づいて、通話音声に含まれる分岐音声の数を音声数決定部と、を含む。

【００９６】

本発明の幾つかの実施例では、上記の実施例をベースとして、音声数決定部は、各分岐音声の平均値と所定閾値との差の絶対値が差閾値よりも小さい場合、分岐音声の数が複数であると判定する第１の数判定部と、各分岐音声の平均値と所定閾値との差の絶対値が差閾値以上である場合、分岐音声の数が１つであると判定する第２の数判定部と、を含む。

【００９７】

本発明の幾つかの実施例では、制御部１４５０は、設定された音声抽出方式に基づいて、主要発言者の音声を抽出する抽出部、を含む。

【００９８】

本発明の幾つかの実施例では、上記の実施例をベースとして、抽出部は、各周波数点における複数の分岐音声のそれぞれに対応するエネルギー占有比率に基づいて、エネルギー占有比率のうちの最大値に対応する分岐音声を主要発言者の音声として認識し、エネルギー情報から主要発言者の音声に対応する周波数情報を決定し、周波数情報に基づいて通話音声から主要発言者の音声を抽出する。

【００９９】

本発明の幾つかの実施例では、上記の実施例をベースとして、分岐音声の数は、１つ又は少なくとも２つを含む。制御部１４５０は、分岐音声の数が１つである場合、設定されたシングルトークのエコー処理方式に基づいて、分岐音声のエコー音声を認識し、エコー音声に対してシングルトークエコー除去を行い、分岐音声の数が少なくとも２つである場合、設定されたダブルトークのエコー処理方式に基づいて、分岐音声に対応するエコー音声をそれぞれ認識し、エコー音声に対してダブルトークエコー除去を行う。

【０１００】

図１５は、本発明の実施例の電子機器を実現可能なコンピュータシステムの構成を示す概略図である。

【０１０１】

なお、図１５に示す電子機器のコンピュータシステム１５００は一例に過ぎず、本発明の実施例の機能や使用範囲に何ら制限を加えるべきものではない。

【０１０２】

図１５に示すように、コンピュータシステム１５００は、読み出し専用メモリ１５０２（Ｒｅａｄ－Ｏｎｌｙ　Ｍｅｍｏｒｙ：ＲＯＭ）に記憶されたプログラムや、記憶部１５０８からランダムアクセスメモリ１５０３（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ：ＲＡＭ）にロードされたプログラムに応じて、各種の適宜の動作や処理を実行可能な中央処理部１５０１（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ：ＣＰＵ）を有する。ランダムアクセスメモリ１５０３には、システム動作に必要な各種プログラムやデータも記憶されている。中央処理部１５０１、読み出し専用メモリ１５０２及びランダムアクセスメモリ１５０３は、バス１５０４を介して互いに接続されている。入力／出力インターフェース１５０５（Ｉｎｐｕｔ／Ｏｕｔｐｕｔインターフェース、即ちＩ／Ｏインターフェース）もバス１５０４に接続される。

【０１０３】

入力部１５０６（キーボード、マウスなどを含む）、出力部１５０７（ディスプレイ、例えば陰極線管（Ｃａｔｈｏｄｅ　Ｒａｙ　Ｔｕｂｅ：ＣＲＴ）、液晶ディスプレイ（Ｌｉｑｕｉｄ　Ｃｒｙｓｔａｌ　Ｄｉｓｐｌａｙ：ＬＣＤ）など、及びスピーカなどを含む）、記憶部１５０８（例えばハードディスクなどを含む）、通信部１５０９（例えばネットワークのインタフェースカード、例えばＬＡＮカード、モデムなどを含む）は、入力／出力インターフェース１５０５に接続されている。通信部１５０９は、ネットワーク、例えばインターネットを介して通信処理を実行する。必要に応じて、ドライバ１５１０は、入力／出力インターフェース１５０５に接続されてもよい。取り外し可能な媒体１５１１は、例えば磁気ディスク、光ディスク、光磁気ディスク、半導体メモリなどであり、必要に応じてドライバ１５１０にセットアップされて、その中から読みだされたコンピュータプログラムは必要に応じて記憶部１５０８にインストールされている。

【０１０４】

特に、本発明の実施例によれば、様々な方法のフローチャートに記載されたプロセスは、コンピュータソフトウェアプログラムとして実装することができる。例えば、本発明の実施例は、コンピュータ読み取り可能な媒体に記憶されたコンピュータプログラムを含むコンピュータプログラム製品を含み、コンピュータプログラムは、フローチャートに示された方法を実行するためのプログラムコードを含む。このような実施例では、コンピュータプログラムは、通信部１５０９を介してネットワークにダウンロードされてインストールされ、且つ／或いは、取り外し可能な媒体１５１１からインストールされる。このコンピュータプログラムが中央処理部（ＣＰＵ）１５０１により実行されると、本発明のシステムに限定される各種機能が実行される。

【０１０５】

なお、本発明の実施例に示すコンピュータ読み取り可能な媒体は、コンピュータ読み取り可能な信号媒体、コンピュータ読み取り可能な記憶媒体、又はこれらの何れかの組み合わせであってもよい。コンピュータ読み取り可能な記憶媒体は、例えば、電気、磁気、光、電磁、赤外線、又は半導体のシステム、機器又はデバイス、又は任意の以上の組み合わせであってもよいが、これらに限定されない。コンピュータ読み取り可能な記憶媒体のより具体的な例としては、１つ以上のリード線を有する電気接続、ポータブルコンピュータディスク、ハードディスク、ランダムアクセスメモリ（ＲＡＭ）、読取り専用メモリ（ＲＯＭ）、消去可能プログラマブル読取り専用メモリ（Ｅｒａｓａｂｌｅ Ｐｒｏｇｒａｍｍａｂｌｅ Ｒｅａｄ Ｏｎｌｙ Ｍｅｍｏｒｙ：ＥＰＲＯＭ）、フラッシュメモリ、光ファイバ、ポータブルコンパクトディスク読取り専用メモリ（Ｃｏｍｐａｃｔ Ｄｉｓｃ Ｒｅａｄ－Ｏｎｌｙ Ｍｅｍｏｒｙ：ＣＤ－ＲＯＭ）、光記憶装置、磁気記憶装置、又はこれらの任意の適切な組み合わせが挙げられるが、これらに限定されない。本発明では、コンピュータ読み取り可能な記憶媒体は、命令実行システム、機器又はデバイスによって使用されるか、又はそれらと組み合わせて使用されることができるプログラムを含むか、又は格納する任意の有形媒体であってもよい。本発明では、コンピュータ読み取り可能な信号媒体は、コンピュータ読み取り可能なプログラムコードが搬送されるベースバンド又は搬送波の一部として伝搬されるデータ信号を含んでもよい。そのような伝搬データ信号は、限定されるものではないが、電磁信号、光信号、又は上述の任意の適切な組み合わせを含む様々な形態であってもよい。コンピュータ読み取り可能な信号媒体は、コンピュータ読み取り可能な記憶媒体以外の任意のコンピュータ読み取り可能な媒体であってもよく、コンピュータ読み取り可能な媒体は、命令実行システム、機器又はデバイスによって使用するために、又はそれらと組み合わせて使用するために、プログラムを送信、伝播、又は送信してもよい。コンピュータ読み取り可能な媒体上に含まれるプログラムコードは、無線、有線等、又は上記の任意の適切な組み合わせを含むが、これらに限定されない任意の適切な媒体で送信してもよい。

【０１０６】

本発明の実施例の１つの態様では、コンピュータ読み取り可能な記憶媒体に記憶されたコンピュータ命令を含むコンピュータプログラム製品又はコンピュータプログラムであって、コンピュータ装置のプロセッサは、コンピュータ読み取り可能な記憶媒体からコンピュータ命令を読み取り、該コンピュータ命令を実行することによって、前記コンピュータ装置に上記の各態様に記載の方法を実行させる、コンピュータプログラム製品又はコンピュータプログラムを提供する。

【０１０７】

他の態様では、本発明は、コンピュータ読み取り可能な媒体をさらに提供する。該コンピュータ読み取り可能な媒体は、上記の実施例で説明した電子機器に含まれてもよいし、単独に存在してもよいし、該電子機器に組み込まれなくてもよい。上記のコンピュータ読み取り可能な媒体は、１つ又は複数のプログラムを保持し、上記の１つ又は複数のプログラムは、１つの電子機器により実行される際に、上記の実施例に記載の方法を電子機器に実現させる。

【０１０８】

なお、本出願は、上述して図面に示した正確な構成に限定されるものではなく、その範囲から逸脱することなく種々の修正及び変更が可能である。本出願の範囲は、添付の特許請求の範囲によってのみ限定される。

【図面】

【図１】

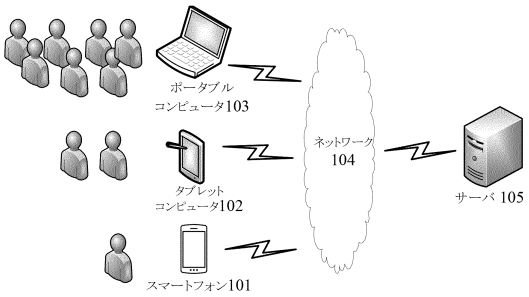

図１

【図２】

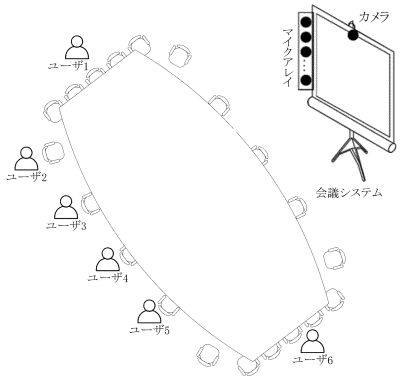

図２

【図３】

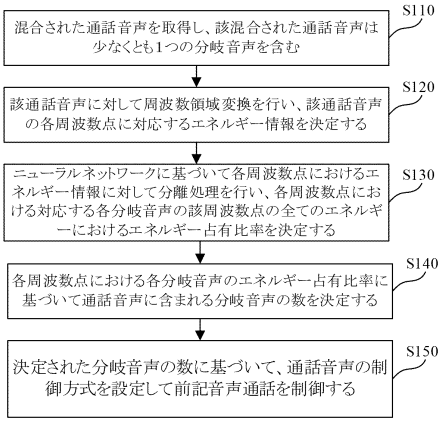

図３

【図４】

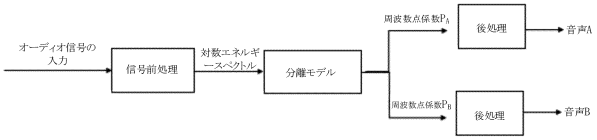

図４

【図５】

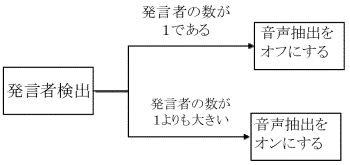

図 5

【図６】

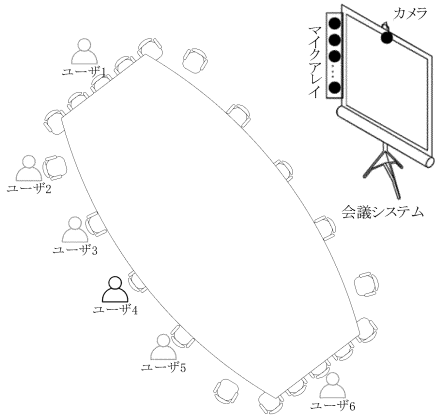

図 6

【図７】

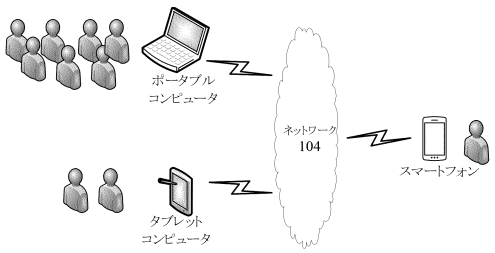

図 7

【図８】

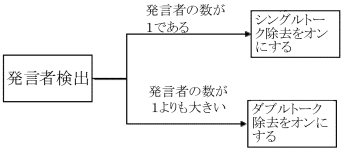

図 8

【図９】

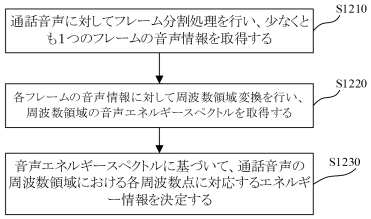

図 9

【図１０】

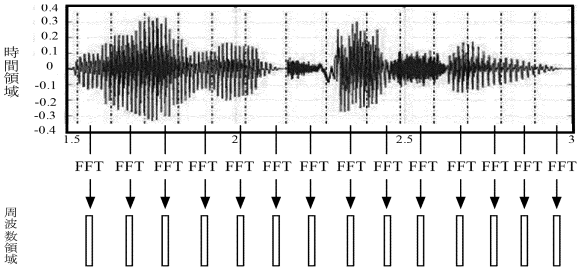

図 10

【図１１】

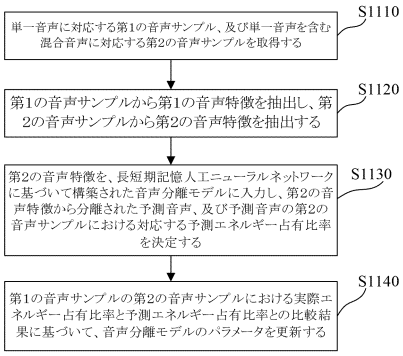

図 11

【図１２】

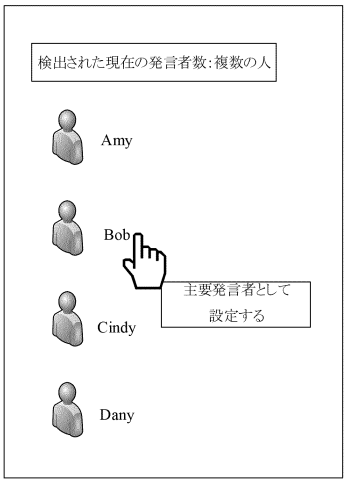

図 12

【図１３】

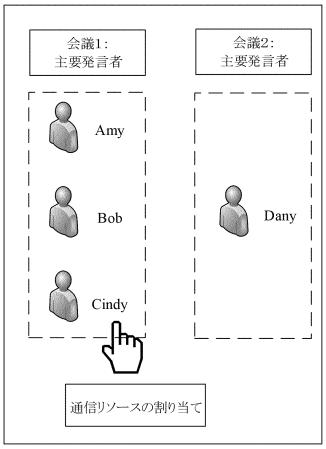

図 13

【図１４】

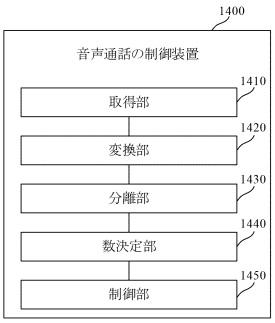

図 14

【図１５】

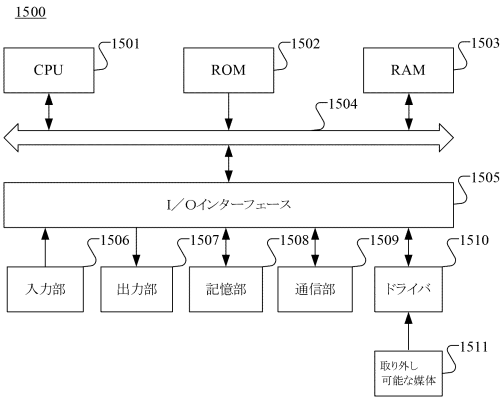

図 15

フロントページの続き

(72)発明者　リー，ジュアンジュアン
中華人民共和国 ５１８０５７ グアンドン シェンジェン ナンシャン・ディストリクト ミッドウェスト・ディストリクト・オブ・ハイテックパーク ケジジョンギ・ロード テンセント・ビルディング ３５エフ
(72)発明者　シア，シャンジュン
中華人民共和国 ５１８０５７ グアンドン シェンジェン ナンシャン・ディストリクト ミッドウェスト・ディストリクト・オブ・ハイテックパーク ケジジョンギ・ロード テンセント・ビルディング ３５エフ

審査官　中村 天真
(56)参考文献
特開２０１８－２０５４４９（ＪＰ，Ａ）
特開２０１２－１７３５８４（ＪＰ，Ａ）
国際公開第２０２０／１１０２２８（WO，Ａ１）
特開２０２０－１３４６５７（ＪＰ，Ａ）
特開２０１０－０６６５０６（ＪＰ，Ａ）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　２１／００－２５／９３
ＩＥＥＥ Ｘｐｌｏｒｅ