(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6305955号
(P6305955)

(45)発行日　平成30年4月4日(2018.4.4)　(24)登録日　平成30年3月16日(2018.3.16)

(51)Int.Cl.		F I		
G10L 15/065	(2013.01)	G10L	15/065	
G10L 15/06	(2013.01)	G10L	15/06	300C
G10L 21/003	(2013.01)	G10L	15/06	500Z
		G10L	21/003	

請求項の数 5　(全 13 頁)

(21)出願番号　特願2015-65787(P2015-65787)
(22)出願日　平成27年3月27日(2015.3.27)
(65)公開番号　特開2016-186515(P2016-186515A)
(43)公開日　平成28年10月27日(2016.10.27)
審査請求日　平成29年2月10日(2017.2.10)

(73)特許権者　000004226
日本電信電話株式会社
東京都千代田区大手町一丁目５番１号
(74)代理人　100121706
弁理士　中尾　直樹
(74)代理人　100128705
弁理士　中村　幸雄
(74)代理人　100147773
弁理士　義村　宗洋
(72)発明者　芦原　孝典
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内
(72)発明者　浅見　太一
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内

最終頁に続く

(54)【発明の名称】音響特徴量変換装置、音響モデル適応装置、音響特徴量変換方法、およびプログラム

(57)【特許請求の範囲】
【請求項１】
所定の音声現象である対象音声現象を含む音声信号から音響特徴量系列を抽出する音響特徴量抽出部と、
上記音響特徴量系列に音素ラベルを付与する音素ラベル付与部と、
上記音響特徴量系列に上記対象音声現象であるか否かを示す対象ラベルを付与する対象ラベル付与部と、
上記音響特徴量系列のうち上記音素ラベルが等しく上記対象ラベルが異なる音響特徴量同士の対応関係に基づいて、上記対象音声現象以外の音響特徴量を上記対象音声現象の音響特徴量へ変換する変換モデルを学習する変換モデル生成部と、
上記変換モデルを用いて上記音響特徴量系列のうち上記対象音声現象以外の音響特徴量を上記対象音声現象の音響特徴量へ変換した疑似音響特徴量系列を生成する疑似音響特徴量生成部と、
を含み、
上記対象音声現象は、声帯振動を伴わない発声により生成される音声であるささやき声、もしくは声帯声門がわずかに開き緩やかな声帯震動により生成される低周波数の音声であるボーカルフライである
音響特徴量変換装置。
【請求項２】
請求項１に記載の音響特徴量変換装置であって、

上記対象ラベル付与部は、上記対象音声現象の音響特徴量と上記対象音声現象以外の音響特徴量とを識別するニューラルネットワークを用いて、発話単位、単語単位、フレーム単位のいずれかの単位で上記音響特徴量系列に上記対象ラベルを付与するものである
音響特徴量変換装置。
【請求項３】
請求項１または２に記載の音響特徴量変換装置により生成した疑似音響特徴量系列を記憶する疑似音響特徴量記憶部と、
所定の音声現象である対象音声現象を含む音声信号から抽出した音響特徴量系列を記憶する音響特徴量記憶部と、
上記音響特徴量系列と上記疑似音響特徴量系列とを用いて音響モデルを学習する音響モデル学習部と、
を含み、
上記対象音声現象は、声帯振動を伴わない発声により生成される音声であるささやき声、もしくは声帯声門がわずかに開き緩やかな声帯震動により生成される低周波数の音声であるボーカルフライである
音響モデル適応装置。
【請求項４】
音響特徴量抽出部が、所定の音声現象である対象音声現象を含む音声信号から音響特徴量系列を抽出する特徴量抽出ステップと、
音素ラベル付与部が、上記音響特徴量系列に音素ラベルを付与する音素ラベル付与ステップと、
対象ラベル付与部が、上記音響特徴量系列に上記対象音声現象であるか否かを示す対象ラベルを付与する対象ラベル付与ステップと、
変換モデル生成部が、上記音響特徴量系列のうち上記音素ラベルが等しく上記対象ラベルが異なる音響特徴量同士の対応関係に基づいて、上記対象音声現象以外の音響特徴量を上記対象音声現象の音響特徴量へ変換する変換モデルを学習する変換モデル生成ステップと、
疑似音響特徴量生成部が、上記変換モデルを用いて上記音響特徴量系列のうち上記対象音声現象以外の音響特徴量を上記対象音声現象の音響特徴量へ変換した疑似音響特徴量系列を生成する疑似音響特徴量生成ステップと、
を含み、
上記対象音声現象は、声帯振動を伴わない発声により生成される音声であるささやき声、もしくは声帯声門がわずかに開き緩やかな声帯震動により生成される低周波数の音声であるボーカルフライである
音響特徴量変換方法。
【請求項５】
請求項１または２に記載の音響特徴量変換装置もしくは請求項３に記載の音響モデル適応装置としてコンピュータを機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
この発明は、音声認識技術に関し、特に、音声認識のタスクに音響モデルを適応させるために用いる学習データを生成する技術に関する。
【背景技術】
【０００２】
特許文献１には、音声認識において実用レベルの性能を担保するために、音声認識の対象とするタスク（以下、認識対象タスクと呼ぶ）に対して音響モデルを適応させる技術が記載されている。ここで、認識対象タスクとは、元々の音響モデルに対して、話者や雑音タイプ、喋り方などの音響的特徴が異なるタスクである。
【０００３】

一般的に、音声認識の性能は認識対象タスクの学習データ量に依存して上下する。つまり、認識対象タスクの学習データが満足に存在しない状況で、従来の技術により音響モデルを適応させたとしても満足のいく認識率は得られない場合が多い。そこで通常は、認識対象タスクの音声を十分に集め、その音声を書き起こしすることで所望の量の学習データを収集するのであるが、そのためには莫大な金銭的・時間的コストを要する。また、認識対象タスクの音声が十分に入手可能であるならば、書き起こしによる学習データの収集を実施することが可能だが、そもそもあらゆるタスクにおいて十分な量の音声が入手可能というわけではない。例えば、方言や日本人が英語を話す音声など、十分な量の音声を入手することが難しいタスクも存在する。

【０００４】

認識対象タスクのデータベースを所有していたとしても、認識対象タスク内で出現頻度の少ない音声現象が存在する場合、その音声現象に対しても頑健な音響モデルを構築するためには、その音声現象の学習データも十分な量を収集する必要がある。例えば、人間同士の自然な会話では、様々な種類の発声を発話の一部でしており、「ささやくような発声（以降、「ささやき声」と呼ぶ）」や「低周波数でのブツブツした音がなる発声（以降、「ボーカルフライ」と呼ぶ）」等の現象が存在している。「ささやき声」とは声帯振動を伴わない発声により生成される音声を指す。例えば、周囲に声を漏らさないためにコソコソ話す際（例えば、公の場でモバイル端末に話しかける場合等）に、しばしば現れる発声である。「ボーカルフライ」とは「きしみ声」や「エッジボイス」とも呼ばれ、声帯声門がわずかに開き緩やかな声帯震動により生成される低周波数の音声を指す。例えば、議論の場において頭の中で考えながら発話する場合や少し自信が無くなった場合等に、発話の全体もしくはその一部に出現する発声である。このようなささやき声やボーカルフライは、通常の発声に比べると圧倒的に頻度が少ないため、さまざまな話者で音響モデルの学習をするために十分な量を収集することは困難である。したがって、ささやき声やボーカルフライ等で発声された発話は誤認識となる可能性が高くなってしまう。

【０００５】

ささやき声やボーカルフライが通常の発話とどのように音響特性が異なるのかについては、非特許文献１や非特許文献２が詳しい。なお、非特許文献１でもささやき声を学習することで認識精度の改善を実現しているが、ここでは既に学習データを十分所有している場合を想定している。

【０００６】

非特許文献３には、声道長正規化（VTLN: Vocal Tract Length Normalization）のWarping Factorを複数の値で実行することで、学習データにおける話者バリエーションを疑似的に作成する方法が記載されている。なお、VTLNについては非特許文献４に記されている。

【先行技術文献】

【特許文献】

【０００７】

【特許文献１】特開２００７－２４９０５１号公報

【非特許文献】

【０００８】

【非特許文献１】伊藤太介，武田一哉，板倉文忠，“ささやき声の音響分析と音声認識への応用”，信学技報，DSP2001-98，SP2001-71，pp. 59-64，2001

【非特許文献２】M. Blomgren，Y. Chen ，M. L. Ng，H. R. Gilbert，“Acoustic，aerodynamic，physiologic，and perceptual properties of modal and vocal fry registers”，Journal of the Acoustical Society of America，vol. 103，pp. 2649-2658，1998

【非特許文献３】N. Jaitly，G. E. Hinton，“Vocal Tract Length Perturbation (VTLP) improves speech recognition”，ICML Workshop on Deep Learning for Audio，Speech，and Language Processing，2013

【非特許文献４】E. Eide，H. Gish，“A parametric approach to vocal tract length

normalization”, ICASSP, pp. 346-348, 1996

【発明の概要】

【発明が解決しようとする課題】

【０００９】

しかしながら、非特許文献３に記載の従来技術では、話者の声質を変換し、話者のバリエーションを拡充することのみを目的としており、ささやき声やボーカルフライのような出現頻度の少ない音声現象に関する学習データを疑似生成する場合には利用することができない。

【００１０】

この発明の目的は、ささやき声やボーカルフライのような出現頻度が少ない音声現象を認識対象タスクとする音響モデル適応において、十分な量の学習データを収集できない状況であっても、認識率を向上させることである。

【課題を解決するための手段】

【００１１】

上記の課題を解決するために、この発明の音響特徴量変換装置は、出現頻度が低い音声現象である対象音声現象を含む音声信号から音響特徴量系列を抽出する音響特徴量抽出部と、音響特徴量系列に音素ラベルを付与する音素ラベル付与部と、音響特徴量系列に対象音声現象であるか否かを示す対象ラベルを付与する対象ラベル付与部と、音響特徴量系列のうち音素ラベルが等しく対象ラベルが異なる音響特徴量同士の対応関係に基づいて、対象音声現象以外の音響特徴量を対象音声現象の音響特徴量へ変換する変換モデルを学習する変換モデル生成部と、変換モデルを用いて音響特徴量系列のうち対象音声現象以外の音響特徴量を対象音声現象の音響特徴量へ変換した疑似音響特徴量系列を生成する疑似音響特徴量生成部と、を含む。

【発明の効果】

【００１２】

この発明の音響特徴量変換技術は、ささやき声やボーカルフライのような出現頻度が少ない音声現象を認識対象タスクとする場合に、学習データが十分に入手できない状況下であっても、統計モデルに基づいて認識対象タスクの音響特徴量を疑似生成し、その疑似音響特徴量を用いて音響モデルを適応させる。これにより、ささやき声とボーカルフライに頑健な音響モデルを生成でき、認識率を向上することができる。

【図面の簡単な説明】

【００１３】

【図１】図１は、音響特徴量変換装置および音響モデル適応装置の機能構成を例示する図である。

【図２】図２は、音響特徴量変換方法および音響モデル適応方法の処理フローを例示する図である。

【発明を実施するための形態】

【００１４】

この発明では、ささやき声やボーカルフライのような出現頻度が少ない音声現象を認識対象タスクとする音響モデル適応において、音響モデルを適応させるために十分な量の学習データを収集できない状況を想定する。まず、ささやき声またはボーカルフライと通常の発声との音響特性の違いを統計的に学習して変換モデルのパラメータを生成する。次に、その変換モデルを用いて通常の発声による学習データからささやき声またはボーカルフライによる学習データを疑似的に生成する。変換モデルのパラメータを学習するためには、音響特性の違いを統計的に学習するために必要な量のささやき声またはボーカルフライの学習データはあらかじめ用意しておく必要がある。

【００１５】

この発明では、大きく以下の流れで音響モデルの適応を行う。

【００１６】

１．ささやき声またはボーカルフライの元々入手できた少量の学習データ（Ｂ）と、さ

さやき声およびボーカルフライではないが十分な量の学習データ（Ａ）とを用意し、学習データ（Ａ）から学習データ（Ｂ）へ変換する変換器を生成する。

【００１７】

２．上記の変換器を利用して十分な量の学習データ（Ａ）から十分な量の疑似学習データ（Ｃ）へ変換する。

【００１８】

３．元々の学習データ（Ｂ）と疑似学習データ（Ｃ）とを用いて、音響モデルを認識対象タスクへ適応する学習処理を行う。

【００１９】

以下、この発明の実施の形態について詳細に説明する。なお、図面中において同じ機能を有する構成部には同じ番号を付し、重複説明を省略する。

【００２０】

［第一実施形態］

第一実施形態では、統計的な変換パラメータによりささやき声の音響特徴量を疑似生成し、その疑似音響特徴量を用いて音響モデルを適応する方法について説明する。

【００２１】

第一実施形態の音響特徴量変換装置１は、図１に例示するように、入力端子１０、音声信号取得部１１、音響特徴量抽出部１２、音素ラベル付与部１３、対象ラベル付与部１４、変換モデル生成部１５、疑似音響特徴量生成部１６、音声信号記憶部２１、音響特徴量記憶部２２、変換モデル記憶部２３、および疑似音響特徴量記憶部２４を含む。

【００２２】

第一実施形態の音響モデル適応装置２は、図１に例示するように、音響特徴量変換装置１の各構成部に加えて、音響モデル学習部１７および音響モデル記憶部２５を含む。図１では、音響モデル適応装置２に音響特徴量変換装置１のすべての構成部が含まれる構成を例示したが、音響特徴量変換装置１の出力を記憶させた音響特徴量記憶部２２と疑似音響特徴量記憶部２４のみを含む構成とすることも可能である。

【００２３】

音響特徴量変換装置１および音響モデル適応装置２の各装置は、例えば、中央演算処理装置（CPU: Central Processing Unit）、主記憶装置（RAM: Random Access Memory）などを有する公知又は専用のコンピュータに特別なプログラムが読み込まれて構成された特別な装置である。各装置は、例えば、中央演算処理装置の制御のもとで各処理を実行する。各装置に入力されたデータや各処理で得られたデータは、例えば、主記憶装置に格納され、主記憶装置に格納されたデータは必要に応じて読み出されて他の処理に利用される。また、各装置が備える各処理部の少なくとも一部が集積回路等のハードウェアによって構成されていてもよい。

【００２４】

音響特徴量変換装置１および音響モデル適応装置２が備える各記憶部は、例えば、RAM（Random Access Memory）などの主記憶装置、ハードディスクや光ディスクもしくはフラッシュメモリ（Flash Memory）のような半導体メモリ素子により構成される補助記憶装置、またはリレーショナルデータベースやキーバリューストアなどのミドルウェアにより構成することができる。各装置が備える各記憶部は、それぞれ論理的に分割されていればよく、一つの物理的な記憶装置に記憶されていてもよい。

【００２５】

図２を参照して、第一実施形態の音響特徴量変換方法の処理手続きを説明する。

【００２６】

ステップＳ１０において、入力端子１０へ、学習データとする音声信号が入力される。学習データの音声信号には、認識対象タスクであるささやき声（以下、対象音声現象とも呼ぶ）による音声信号と、通常の発声による音声信号とが含まれる。入力される音声信号は、マイクロホン等の収音手段を入力端子１０へ接続してリアルタイムに人間の発話を収音したものであってもよいし、あらかじめ人間の発話をICレコーダーやスマートフォンの

録音機能のような録音手段で不揮発性メモリやハードディスクドライブのような記録媒体へ録音し、入力端子１０へ接続した再生手段により再生することで入力してもよい。

【００２７】

ステップＳ１１において、音声信号取得部１１は、アナログの入力音声信号をディジタル信号に変換する。入力端子１０からディジタルの音声信号が入力される場合には、音声信号取得部１１は備えなくともよい。ディジタルの入力音声信号は、音声信号記憶部２１へ記憶される。

【００２８】

ステップＳ１２において、音響特徴量抽出部１２は、音声信号記憶部２１に記憶されたディジタルの入力音声信号を読み込み、入力音声信号の各フレームから音響特徴量を抽出し、音響特徴量系列を生成する。入力音声信号の音響特徴量系列は、音響特徴量記憶部２２へ記憶される。抽出する音響特徴量としては、例えば、音声信号の短時間フレーム分析に基づくメル周波数ケプストラム係数（MFCC: Mel-Frequency Cepstrum Coefficient）の1～12次元と、その動的特徴量であるΔMFCC、ΔΔMFCCなどの動的パラメータや、パワー、Δパワー、ΔΔパワー等を用いる。また、MFCCに対してはケプストラム平均正規化（CMN: Cepstral Mean Normalization）処理を行ってもよい。抽出する音響特徴量は、MFCCやパワーに限定したものではなく、音声認識に用いられるパラメータを用いてもよい。

【００２９】

ステップＳ１３において、音素ラベル付与部１３は、音響特徴量記憶部２２へ記憶された入力音声信号の音響特徴量系列を読み込み、フレーム単位で音素ラベルを付与する。音響特徴量系列に付与された音素ラベルは、対象ラベル付与部１４および変換モデル生成部１５へ送られる。音素ラベルの付与方法としては、手動獲得による方法と、自動獲得による方法が挙げられる。手動獲得による方法は、発話内容を鑑みながら音声波形に対して該当する時間領域の音素を手動でラベル付けする方法である。自動獲得による方法は、音響特徴量系列から強制アラインメントを実行することで、音素ラベル付き音響特徴量系列を生成する方法である。強制アラインメントとは、音響特徴量系列の発話内容が既知である前提で、その発話内容に一致する正解テキストに対する音声認識を実行し、認識処理過程における状態遷移を観測することで、入力した分析フレーム毎の音響特徴量に対応する隠れマルコフモデル（HMM: Hidden Markov Model）の状態番号を割り当てる処理である。なお、音声認識ではしばしば音素認識のために隠れマルコフモデルを用い、状態番号はトライフォン（triphone）までを考える。トライフォンは分類すべき音素の前後の音素関係も含めた音素の３つ組みである。トライフォンでは、例えば「a-k-a」のように３音素を１つの状態番号として考える。なお、モノフォン（monophone）は音素１つ、バイフォン（biphone）は音素２つの組を１つの状態番号として考える。強制アラインメントは正解テキストを用いてビタビアルゴリズム等を利用して実行される。なお、音声認識における隠れマルコフモデルやビタビアルゴリズムについては下記参考文献１に記載されている。

〔参考文献１〕鹿野清宏他、“IT Text 音声認識システム”、オーム社、2001年

【００３０】

ステップＳ１４において、対象ラベル付与部１４は、音響特徴量記憶部２２へ記憶された入力音声信号の音響特徴量系列を読み込み、対象音声現象（すなわち、ささやき声）の音声であるか否かを表す対象ラベル（以下、ささやき声ラベルと呼ぶ）を付与する。音響特徴量系列に付与されたささやき声ラベルは、変換モデル生成部１５および疑似音響特徴量生成部１６へ送られる。ささやき声ラベルの付与方法としては、例えば、（１）音声の収録時に発話者がささやき声か否かを予め指定する方法、（２）人間が実際に音声を聴取しささやき声か否かを判断する方法、（３）自動でささやき声か否かを判別する方法が挙げられる。（１）発話者が指定する方法は、音声を収録する際に発話者がこれから発話する音声がささやき声か否かを予め指定する。（２）人間が判断する方法は、収録済みの音声を発話者本人もしくはそれ以外の人間が音声を聴取しささやき声か否かを判断する。（３）自動で判別する方法は、例えば、音声信号をケプストラム分析した上で、その高次成分の大きさを予め定めた閾値と比較することでささやき声か否かを判別する。ささやき声

は通常の発声とは異なり、ホワイトノイズのような非周期的な駆動音源信号となるため、駆動音源信号成分と考えられるケプストラムの高次成分の値が大きい場合は周期的と捉えて通常の発声であると判別し、小さい場合は非周期的と捉えてささやき声であると判別する。この際、音素ラベルから[p][t][k][f][s]のような無声音はささやき声と判別が難しいため、予め除去しておいてもよい。

【００３１】

ささやき声ラベルの付与方法は上記に限定されない。他には、ささやき声か否かを判別するモデルを予め構築しておき、そのモデルに基づいてささやき声ラベルを付与する方法でもよい。例えば、ささやき声と通常の発声をそれぞれ混合ガウス分布（GMM: Gaussian Mixture Model）により予めモデル化しておき、そのモデルに基づいた尤度比較によりささやき声か通常の発声かを識別する方法や、ささやき声と通常の発声の二つのクラスによるディープニューラルネットワーク（DNN: Deep Neural Networks）により識別する方法等も考えられる。

【００３２】

上記（１）～（３）の方法は、それぞれ単独で利用することも可能であるが、組み合わせて利用することも可能である。また、ささやき声ラベルを付与する単位は、発話単位、単語単位、フレーム単位など、どのような単位でもよい。例えば、（１）の方法により発話単位もしくは単語単位で大まかにささやき声ラベルを付与した後に、さらに（３）の方法によりフレーム単位でささやき声ラベルを付与してもよい。

【００３３】

ステップＳ１５において、変換モデル生成部１５は、音響特徴量記憶部２２へ記憶された入力音声信号の音響特徴量系列を読み込み、音素ラベル付与部１３から受け取った音素ラベルと対象ラベル付与部１４から受け取ったささやき声ラベルとを用いて、音素ラベルが等しく対象ラベルが異なる音響特徴量同士の対応関係に基づいて、通常の発話による音響特徴量をささやき声による音響特徴量に変換する変換モデルのパラメータ（以下、音響特徴量変換パラメータと呼ぶ）を学習する。学習済みの音響特徴量変換パラメータは、変換モデル記憶部２３へ記憶される。変換モデルとしては、例えば、声質変換に利用されるようなモデルが考えられる。なお、声質変換に用いられるモデルとしては、混合ガウス分布やディープニューラルネットワークが挙げられる。混合ガウス分布やディープニューラルネットワークによる声質変換手法は、下記参考文献２や下記参考文献３が詳しい。

〔参考文献２〕S. Desai, A.W. Black, B. Yegnanarayana, K. Prahallad, “Spectral Mapping Using Artificial Neural Networks for Voice Conversion”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 5, pp. 954-964, 2010

〔参考文献３〕T. Toda, A.W. Black, K. Tokuda, “Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222-2235, 2007

【００３４】

ステップＳ１６において、疑似音響特徴量変換部１６は、音響特徴量記憶部２２へ記憶された入力音声信号の音響特徴量系列を読み込み、変換モデル記憶部２３に記憶された音響特徴量変換パラメータと対象ラベル付与部１４から受け取ったささやき声ラベルとを用いて、入力音声信号の音響特徴量系列のうち通常の発声による（ささやき声ラベルが付与されていない）音響特徴量をささやき声による音響特徴量に変換してささやき声の疑似音響特徴量系列を生成する。すなわち、変換モデルの入力はささやき声ではない通常の発声による音響特徴量であり、出力は疑似的に生成したささやき声による音響特徴量となる。変換後の疑似音響特徴量系列は、疑似音響特徴量記憶部２４へ記憶される。

【００３５】

引き続き、図２を参照して、第一実施形態の音響モデル適応方法の処理手続きを説明する。

【００３６】

音響特徴量記憶部２２には、対象音声現象であるささやき声を含む学習データの音声信

号から抽出した音響特徴量系列が記憶されている。

【００３７】

疑似音響特徴量記憶部２４には、学習データの音声信号から抽出した音響特徴量系列を上述の音響特徴量変換装置１により生成したささやき声の疑似音響特徴量系列が記憶されている。

【００３８】

ステップＳ１７において、音響モデル学習部１７は、音響特徴量記憶部２２に記憶された入力音声信号の音響特徴量系列と疑似音響特徴量記憶部２４に記憶された疑似音響特徴量系列とを利用して音響モデルを学習する。学習済みの音響モデルは、音響モデル記憶部２５へ記憶される。音声認識における音響モデルとしては、GMM-HMMなどが用いられており、音響モデルを認識対象タスクに適応させる手法は、例えば、下記参考文献４などに記載されている。

〔参考文献４〕篠田浩一、“確率モデルによる音声認識のための話者適応化技術”、電子情報通信学会論文誌、J87-D-II(2)、pp．371-386、2004年

【００３９】

音響モデルの適応に用いる音響特徴量は、音響モデルを用いる音声認識装置に求められる機能によって選択するとよい。具体的には、以下の２パターンが考えられる。

【００４０】

（１）ささやき声に関する音響特徴量のみを用いて適応した音響モデルにより構築される音声認識装置の場合は、ささやき声の少量の学習データ（Ｂ）と疑似学習データ（Ｃ）のみを用いる。つまり、ささやき声に関する音響特徴量系列と疑似音響特徴量系列で適応した音響モデルを生成する。認識時には、ささやき声で発声した発話にのみ、この音響モデルを利用することができる。したがって、予めささやき声しか入力されないことがわかっている場合は、この音響モデルだけで音声認識装置を構築する。

【００４１】

通常の発声による発話も含まれる場合は、通常の発声による十分な量の学習データ（Ａ）のみで適応した音響モデルも併用すればよい。この場合、上述の対象ラベル付与部で説明したささやき声であるか否かを判別する方法を用いて、二つの音響モデルのうちどちらを利用するかを判別するとよい。すなわち、認識対象の入力音声がささやき声である場合は、ささやき声に関するデータだけで適応された音響モデルを利用し、ささやき声でない場合は、通常の発声による音響特徴量だけで適応した音響モデルを利用する。なお、後述の第二実施形態で説明するボーカルフライに関する音響特徴量だけで適応した音響モデルも所有している場合には、さらにこの音響モデルも併用してよい。

【００４２】

（２）ささやき声だけでなくすべての発声を含めた音響特徴量を用いて適応した音響モデルにより構築される音声認識装置の場合は、ささやき声ではない十分な量の学習データ（Ａ）とささやき声の少量の学習データ（Ｂ）と疑似学習データ（Ｃ）とをすべて用いる。つまり、ささやき声に関する音響特徴量系列と疑似音響特徴量系列だけでなく、それ以外の全発話の音響特徴量系列で適応した音響モデルを生成する。このとき、第二実施形態で生成されるボーカルフライの疑似音響特徴量系列も含めてもよい。この場合、認識時には、すべての発話を一様に音声認識装置に入力することになる。

【００４３】

上述のように構成することで、第一実施形態の音響特徴量変換装置および方法は、認識対象であるささやき声の学習データが十分に入手できない場合であっても、ささやき声と通常の発声の音響特性の違いを統計的に学習した特徴量変換パラメータに基づいて、ささやき声の音響特徴量系列を疑似的に生成することができる。したがって、第一実施形態の音響モデル適応装置および方法は、十分な量の疑似音響特徴量系列を用いて音響モデルの適応を行うことで、ささやき声に頑健な音響モデルを作成することができ、この音響モデルを用いて音声認識をすることで認識率が向上する。

【００４４】

［第二実施形態］
第二実施形態では、統計的な変換パラメータによるボーカルフライの音響特徴量を疑似生成し、その疑似音響特徴量を用いて音響モデルを適応する方法について説明する。
【００４５】
以下、第二実施形態の音響特徴量変換方法を説明する。以下では、上述の第一実施形態との相違点を中心に説明する。
【００４６】
第二実施形態の対象ラベル付与部１４は、音響特徴量記憶部２２へ記憶された入力音声信号の音響特徴量系列を読み込み、対象音声現象（すなわち、ボーカルフライ）の音声であるか否かを表す対象ラベル（以下、ボーカルフライラベルと呼ぶ）を付与する。ボーカルフライラベルの付与方法としては、例えば、（１）音声の収録時に発話者がボーカルフライか否かを予め指定する方法、（２）人間が実際に音声を聴取しボーカルフライか否かを判断する方法、（３）自動でボーカルフライか否かを判別する方法が挙げられる。
【００４７】
（１）発話者が指定する方法は、音声を収録する際に発話者がこれから発話する音声がボーカルフライか否かを予め指定することでボーカルフライラベルを付与する。
【００４８】
（２）人間が判断する方法は、収録済みの音声を発話者本人もしくはそれ以外の人間が音声を聴取しボーカルフライか否かを判断してボーカルフライラベルを付与する。
【００４９】
（３）自動で判別する方法は、例えば、下記参考文献５に記載されるように自己相関を利用してボーカルフライか否かを判別してボーカルフライラベルを付与する。
〔参考文献５〕C. T. Ishi, “Analysis of autocorrelation-based parameters for creaky voice detection”, Proceedings of The 2nd International Conference on Speech Prosody, pp. 643-646, 2004
【００５０】
ボーカルフライラベルの付与方法は上記に限定されない。他には、母音が継続している部分に対して、複数の窓幅を用いてケプストラム分析を実行し、その差の大きさからボーカルフライなのか否かを判別する方法でもよい。通常の発声では、20ミリ秒の窓幅によるケプストラム分析も30ミリ秒の窓幅によるケプストラム分析も結果は大きく変わらないが、ボーカルフライでは20～40ミリ秒毎に音声が消失しているような不規則な音声波形を有しているため、ケプストラム分析の値が大きく変わる。したがって、窓幅を変えたケプストラム分析の値の差を予め定めた閾値と比較することでボーカルフライか否かを判別しボーカルフライラベルを付与する。
【００５１】
さらに、ボーカルフライか否かを判別するモデルを予め構築しておき、そのモデルに基づいてボーカルフライラベルを付与する方法でもよい。例えば、ボーカルフライと通常の発声をそれぞれ混合ガウス分布により予めモデル化しておき、そのモデルに基づいた尤度比較によりボーカルフライか通常の発声かを識別する方法や、ボーカルフライと通常の発声の二つのクラスによるディープニューラルネットワークにより識別する方法等も考えられる。
【００５２】
上記（１）～（３）の方法は、それぞれ単独で利用することも可能であるが、組み合わせて利用することも可能である。また、ボーカルフライラベルを付与する単位は、発話単位、単語単位、フレーム単位など、どのような単位でもよい。例えば、（１）の方法により発話単位もしくは単語単位で大まかにボーカルフライラベルを付与した後に、さらに（３）の方法によりフレーム単位でボーカルフライラベルを付与してもよい。
【００５３】
第二実施形態の変換モデル生成部１５は、音響特徴量記憶部２２へ記憶された入力音声信号の音響特徴量系列を読み込み、音素ラベル付与部１３から受け取った音素ラベルと対

象ラベル付与部１４から受け取ったボーカルフライラベルとを用いて、音素ラベルが等しく対象ラベルが異なる音響特徴量同士の対応関係に基づいて、通常の発話による音響特徴量をボーカルフライによる音響特徴量に変換する変換モデルのパラメータ（以下、音響特徴量変換パラメータと呼ぶ）を学習する。学習済みの音響特徴量変換パラメータは、変換モデル記憶部２３へ記憶される。

【００５４】

第二実施形態の疑似音響特徴量変換部１６は、音響特徴量記憶部２２へ記憶された入力音声信号の音響特徴量系列を読み込み、変換モデル記憶部２３に記憶された音響特徴量変換パラメータと対象ラベル付与部１４から受け取ったボーカルフライラベルとを用いて、入力音声信号の音響特徴量系列のうち通常の発声による（ボーカルフライラベルが付与されていない）音響特徴量をボーカルフライによる音響特徴量に変換してボーカルフライの疑似音響特徴量系列を生成する。すなわち、変換モデルの入力はボーカルフライではない通常の発話による音響特徴量であり、出力は疑似的に生成したボーカルフライの音響特徴量となる。変換後の疑似音響特徴量系列は、疑似音響特徴量記憶部２４へ記憶される。

【００５５】

上述のように構成することで、第二実施形態の音響特徴量変換装置および方法は、認識対象であるボーカルフライの学習データが十分に入手できない場合であっても、ボーカルフライと通常の発声の音響特性の違いを統計的に学習した特徴量変換パラメータに基づいて、ボーカルフライの音響特徴量系列を疑似的に生成することができる。したがって、第二実施形態の音響モデル適応装置および方法は、十分な量の疑似音響特徴量系列を用いて音響モデルの適応を行うことで、ボーカルフライに頑健な音響モデルを作成することができ、この音響モデルを用いて音声認識をすることで認識率が向上する。

【００５６】

この発明は上述の実施形態に限定されるものではなく、この発明の趣旨を逸脱しない範囲で適宜変更が可能であることはいうまでもない。上記実施形態において説明した各種の処理は、記載の順に従って時系列に実行されるのみならず、処理を実行する装置の処理能力あるいは必要に応じて並列的にあるいは個別に実行されてもよい。

【００５７】

［プログラム、記録媒体］

上記実施形態で説明した各装置における各種の処理機能をコンピュータによって実現する場合、各装置が有すべき機能の処理内容はプログラムによって記述される。そして、このプログラムをコンピュータで実行することにより、上記各装置における各種の処理機能がコンピュータ上で実現される。

【００５８】

この処理内容を記述したプログラムは、コンピュータで読み取り可能な記録媒体に記録しておくことができる。コンピュータで読み取り可能な記録媒体としては、例えば、磁気記録装置、光ディスク、光磁気記録媒体、半導体メモリ等どのようなものでもよい。

【００５９】

また、このプログラムの流通は、例えば、そのプログラムを記録したDVD、CD-ROM等の可搬型記録媒体を販売、譲渡、貸与等することによって行う。さらに、このプログラムをサーバコンピュータの記憶装置に格納しておき、ネットワークを介して、サーバコンピュータから他のコンピュータにそのプログラムを転送することにより、このプログラムを流通させる構成としてもよい。

【００６０】

このようなプログラムを実行するコンピュータは、例えば、まず、可搬型記録媒体に記録されたプログラムもしくはサーバコンピュータから転送されたプログラムを、一旦、自己の記憶装置に格納する。そして、処理の実行時、このコンピュータは、自己の記録媒体に格納されたプログラムを読み取り、読み取ったプログラムに従った処理を実行する。また、このプログラムの別の実行形態として、コンピュータが可搬型記録媒体から直接プログラムを読み取り、そのプログラムに従った処理を実行することとしてもよく、さらに、

このコンピュータにサーバコンピュータからプログラムが転送されるたびに、逐次、受け取ったプログラムに従った処理を実行することとしてもよい。また、サーバコンピュータから、このコンピュータへのプログラムの転送は行わず、その実行指示と結果取得のみによって処理機能を実現する、いわゆるASP（Application Service Provider）型のサービスによって、上述の処理を実行する構成としてもよい。なお、本形態におけるプログラムには、電子計算機による処理の用に供する情報であってプログラムに準ずるもの（コンピュータに対する直接の指令ではないがコンピュータの処理を規定する性質を有するデータ等）を含むものとする。

【００６１】

また、この形態では、コンピュータ上で所定のプログラムを実行させることにより、本装置を構成することとしたが、これらの処理内容の少なくとも一部をハードウェア的に実現することとしてもよい。

【符号の説明】

【００６２】

１　音響特徴量変換装置
２　音響モデル適応装置
１１　音声信号取得部
１２　音響特徴量抽出部
１３　音素ラベル付与部
１４　対象ラベル付与部
１５　変換モデル生成部
１６　疑似音響特徴量生成部
１７　音響モデル学習部
２１　音声信号記憶部
２２　音響特徴量記憶部
２３　変換モデル記憶部
２４　疑似音響特徴量記憶部
２５　音響モデル記憶部

【図１】

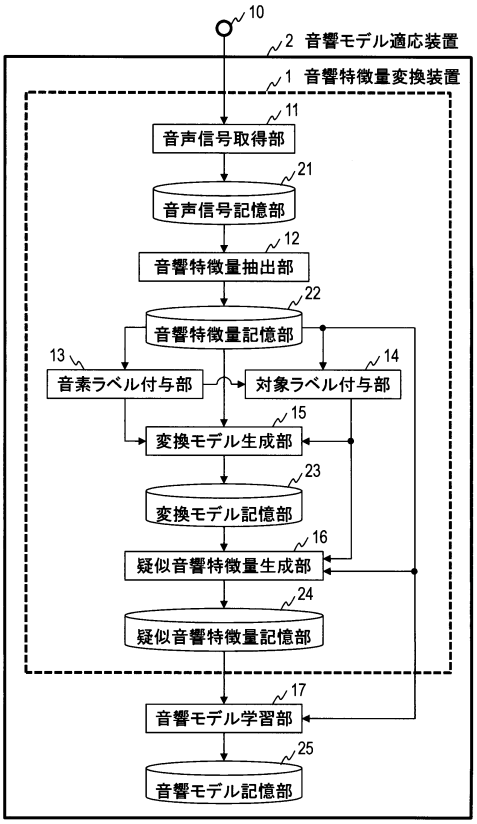
10
2 音響モデル適応装置
1 音響特徴量変換装置
11
音声信号取得部
21
音声信号記憶部
12
音響特徴量抽出部
22
音響特徴量記憶部
13
音素ラベル付与部
14
対象ラベル付与部
15
変換モデル生成部
23
変換モデル記憶部
16
疑似音響特徴量生成部
24
疑似音響特徴量記憶部
17
音響モデル学習部
25
音響モデル記憶部

図1

【図２】

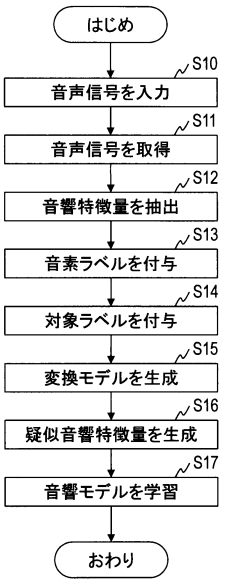
はじめ
S10
音声信号を入力
S11
音声信号を取得
S12
音響特徴量を抽出
S13
音素ラベルを付与
S14
対象ラベルを付与
S15
変換モデルを生成
S16
疑似音響特徴量を生成
S17
音響モデルを学習
おわり

図2

フロントページの続き

(72)発明者　井島　勇祐
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内

審査官　鈴木　圭一郎

(56)参考文献　特開２００７－０７９３６３（ＪＰ，Ａ）
特開２００８－１３９５７３（ＪＰ，Ａ）
石井カルロス寿憲，Vocal Fry発声区間の自動検出法，電子情報通信学会論文誌Ｄ（Ｊ８９－Ｄ），日本，一般社団法人電子情報通信学会，２００６年１２月　１日，No.12，P2679-2687
芦原孝典，声質変換を用いた音声特徴量疑似生成による話者適応，電子情報通信学会技術研究報告　Ｖｏｌ．１１４　Ｎｏ．４１１，日本，一般社団法人電子情報通信学会，２０１５年　１月２２日，第114巻，p13-18

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ１５／００－１５／３４
Ｇ１０Ｌ２１／００－２１／１８