(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6073952号
(P6073952)
(45)発行日　平成29年2月1日(2017.2.1)　(24)登録日　平成29年1月13日(2017.1.13)

(51)Int.Cl.　F I
G10L 15/06 (2013.01)　G10L 15/06 3OOY
G10L 15/06 5OOL

請求項の数 5　(全 15 頁)

(21)出願番号　特願2015-59213(P2015-59213)
(22)出願日　平成27年3月23日(2015.3.23)
(65)公開番号　特開2016-177233(P2016-177233A)
(43)公開日　平成28年10月6日(2016.10.6)
審査請求日　平成27年10月28日(2015.10.28)

(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　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
(72)発明者　増村　亮
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】学習データ生成装置、方法及びプログラム

(57)【特許請求の範囲】
【請求項１】
ターゲットとなる発話セットをターゲットセットとして、上記ターゲットセットの各発話の音声信号から音響特徴量系列を抽出する第一音響特徴量抽出部と、
上記第一音響特徴量抽出部で抽出された音響特徴量系列を１個以上のコンポーネントに分解し、上記ターゲットセットにおける各コンポーネントの構成比である目標構成比を求めるコンポーネント分解部と、
母体セットに含まれる各発話の音声信号から音響特徴量系列を抽出する第二音響特徴量抽出部と、
上記第二音響特徴量抽出部で抽出された音響特徴量系列を用いて、上記母体セットに含まれる各発話における各コンポーネントの含有量を計算するコンポーネント含有量計算部と、
上記母体セットに含まれる各発話における各コンポーネントの含有量に基づいて、上記母体セットに含まれる発話の中から、学習セットにおけるコンポーネントの構成比が上記目標構成比に近くなるように発話を選択することにより学習データを生成する発話選択部と、
を含む学習データ生成装置。
【請求項２】
請求項１の学習データ生成装置において、
上記発話選択部は、上記母体セットに含まれる各発話における各コンポーネントのそれ

ぞれの含有量に基づいて、上記母体セットに含まれる発話の中から、発話を追加した後の学習セットにおけるコンポーネントの構成比が上記目標構成比に近くなるように１個の発話を選択し学習セットに追加する処理を繰り返すことにより学習データを生成する、
学習データ生成装置。
【請求項３】
請求項１又は２の学習データ生成装置において、
上記ターゲットセットの各発話を非発声区間と発生区間とに分割することにより非発声ターゲットセット及び発声ターゲットセットを生成する第一発声／非発声区間抽出部と、
上記母体セットの各発話を非発声区間と発声区間とに分割することにより非発声母体セット及び発声母体セットを生成する第二発声／非発声区間抽出部と、
上記第一音響特徴量抽出部は、上記非発声ターゲットセットの各発話の音声信号から音響特徴量系列を抽出し、上記発声ターゲットセットの各発話の音声信号から音響特徴量系列を抽出し、
上記コンポーネント分解部は、上記第一音響特徴量抽出部で抽出された上記非発声ターゲットセットの各発話の音響特徴量系列を１個以上の非発声コンポーネントに分解し、上記非発声ターゲットセットにおける各非発声コンポーネントの構成比である非発声目標構成比を求め、上記第一音響特徴量抽出部で抽出された上記発声ターゲットセットの各発話の音響特徴量系列を１個以上の発声コンポーネントに分解し、上記発声ターゲットセットにおける各発声コンポーネントの構成比である発声目標構成比を求め、
上記第二音響特徴量抽出部は、上記非発声母体セットの各発話の音声信号から音響特徴量系列を抽出し、上記発声母体セットの各発話の音声信号から音響特徴量系列を抽出し、
上記コンポーネント含有量計算部は、上記第二音響特徴量抽出部で抽出された上記非発声母体セットの各発話の音響特徴量系列を用いて、上記非発声母体セットに含まれる各発話における各非発声コンポーネントの含有量を計算し、上記第二音響特徴量抽出部で抽出された上記発声母体セットの各発話の音響特徴量系列を用いて、上記発声母体セットに含まれる各発話における各発声コンポーネントの含有量を計算し、
上記発話選択部は、上記非発声母体セットに含まれる各発話における各非発声コンポーネントの含有量及び上記発声母体セットに含まれる各発話における各発声コンポーネントの含有量に基づいて、上記母体セットに含まれる発話の中から、学習セットにおける非発声コンポーネントの構成比が上記非発声目標構成比に近くなるように、かつ、学習セットにおける発声コンポーネントの構成比が上記発声目標構成比に近くなるように、発話を選択することにより学習データを生成する、
学習データ生成装置。
【請求項４】
第一音響特徴量抽出部が、ターゲットとなる発話セットをターゲットセットとして、上記ターゲットセットの各発話の音声信号から音響特徴量系列を抽出する第一音響特徴量抽出方法と、
コンポーネント分解部が、上記第一音響特徴量抽出部で抽出された音響特徴量系列を１個以上のコンポーネントに分解し、上記ターゲットセットにおける各コンポーネントの構成比である目標構成比を求めるコンポーネント分解ステップと、
第二音響特徴量抽出部が、母体セットに含まれる各発話の音声信号から音響特徴量系列を抽出する第二音響特徴量抽出ステップと、
コンポーネント含有量計算部が、上記第二音響特徴量抽出部で抽出された音響特徴量系列を用いて、上記母体セットに含まれる各発話における各コンポーネントの含有量を計算するコンポーネント含有量計算ステップと、
学習データを生成する発話選択部が、上記母体セットに含まれる各発話における各コンポーネントの含有量に基づいて、上記母体セットに含まれる発話の中から、学習セットにおけるコンポーネントの構成比が上記目標構成比に近くなるように発話を選択することにより学習データを生成する発話選択ステップと、
を含む学習データ生成方法。

【請求項５】
請求項１から３の何れかの学習データ生成装置の各部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、高い精度で音声認識を行える音響モデルの学習データ生成技術に関する。
【背景技術】
【０００２】
様々な話者、発声スタイル、収録機器、周辺雑音環境（これら４つをまとめて「ドメイン」と書く）で収録された音声を大規模に集積した（１０００時間超の）音声データセット（以下、「母体セット」と書く）から、実際に音声認識システムが認識対象とするドメイン（以下、「ターゲットドメイン」と書く）で高精度に認識を行える音響モデルを構築するためには、母体セットから適切に音声データを選別し、ターゲットドメインに適合した学習セットを構成する必要がある。余分な音声が学習セットに含まれる場合、また必要な音声を学習セットに含めなかった場合は、いずれも音響モデルに悪影響を及ぼし認識精度が低下する可能性がある。母体セット中の適切な（ターゲットドメインに適合する）部分集合を音響モデルの学習セットとして用いることは認識精度を高めるために重要である。
【０００３】
ターゲットドメインで収録された数時間程度の少量の音声データをもとに、母体セットから学習セットを選別する方法が、例えば特許文献１に記載されている。特許文献１の方法では、ターゲットドメインで収録された音声を用いて、事前に用意されたベース音響モデルをターゲットドメインの音声に適応させた適応音響モデルを生成し、母体セットの音声をベース音響モデルおよび適応音響モデルで音声認識して、ベース音響モデルおよび適応音響モデルの音声認識スコアであるベース認識スコアと適応認識スコアを求め、適応認識スコアからベース認識スコアを減じた値が大きい音声を母体セットから選択することで、学習セットを選別している。
【先行技術文献】
【特許文献】
【０００４】
【特許文献１】特開２００９－１２８４９０号公報
【発明の概要】
【発明が解決しようとする課題】
【０００５】
特許文献１に記載されている従来技術は、適応認識スコアとベース認識スコアの差が大きくなる音声を母体セットから選択する。母体セットの音声の中で、話者、発声スタイル、収録機器、周辺雑音環境の全てがターゲットドメインに適合する音声では適応認識スコアとベース認識スコアの差は大きくなるが、いずれかが適合しない音声では適応認識スコアとベース認識スコアの差は小さくなる傾向がある。そのため、特許文献１に記載されている従来技術では、例えば周辺雑音環境は適合していないが話者と発声スタイルと収録機器はターゲットドメインに適合している（本来は学習セットに含めるべき）音声を母体セットから選択できず、学習セットが不十分なものとなるため、音響モデルの精度が高くならない可能性があった。
【０００６】
この発明の目的は、ターゲットドメインに従来よりも適合した学習データを生成する学習データ生成装置、方法及びプログラムを提供することである。
【課題を解決するための手段】
【０００７】
この発明の一態様による学習データ生成装置は、ターゲットとなる発話セットをターゲ

ットセットとして、ターゲットセットの各発話の音声信号から音響特徴量系列を抽出する第一音響特徴量抽出部と、第一音響特徴量抽出部で抽出された音響特徴量系列を１個以上のコンポーネントに分解し、ターゲットセットにおける各コンポーネントの構成比である目標構成比を求めるコンポーネント分解部と、母体セットに含まれる各発話の音声信号から音響特徴量系列を抽出する第二音響特徴量抽出部と、第二音響特徴量抽出部で抽出された音響特徴量系列を用いて、母体セットに含まれる各発話における各コンポーネントの含有量を計算するコンポーネント含有量計算部と、母体セットに含まれる各発話における各コンポーネントの含有量に基づいて、母体セットに含まれる発話の中から、学習セットにおけるコンポーネントの構成比が目標構成比に近くなるように発話を選択することにより学習データを生成する発話選択部と、を備えている。

【発明の効果】

【０００８】

ターゲットドメインに従来よりも適合した学習データを生成することができる。

【図面の簡単な説明】

【０００９】

【図１】第一実施形態の学習データ生成装置の例を説明するためのブロック図。

【図２】第一実施形態の学習データ生成方法の例を説明するための流れ図。

【図３】第二実施形態の学習データ生成装置の例を説明するためのブロック図。

【図４】第二実施形態の学習データ生成方法の例を説明するための流れ図。

【発明を実施するための形態】

【００１０】

［技術的背景］

本発明のポイントの中で主要なものは、以下の２点である。

【００１１】

１．ターゲットドメインの音声を潜在的な構成要素（コンポーネント）に分解し、「コンポーネント」と「その構成比」によってターゲットドメインの音声を捉えること。

【００１２】

２．学習セットのコンポーネント構成比が、ターゲットドメインの音声のコンポーネント構成比と近くなるように学習セットを母体セットから選別すること。

【００１３】

ポイント１のコンポーネントへの分解は（混合正規分布の当てはめなどの方法で）自動的に行われる。各コンポーネントは話者、発声スタイル、収録機器、周辺雑音環境のいずれか（あるいは複数）の特徴を表し、コンポーネントの構成比は話者、発声スタイル、収録機器、周辺雑音環境を全て考慮したターゲットドメイン全体の特徴を表すことになる。

【００１４】

ポイント２により、ターゲットドメイン全体の特徴を表すコンポーネント構成比を観点として学習セットを選別するため、話者、発声スタイル、収録機器、周辺雑音環境が全てターゲットドメインに適合する学習セットを選別することができる。そして、従来技術とは違い、例えば話者のみがターゲットドメインに適合する音声も学習セットとして選択されることになる。話者のみがターゲットドメインに適合する音声を選択したとしても、発声スタイルや収録機器や周辺雑音環境が適合した別の音声を選択することによって、全体のコンポーネント構成比をターゲットドメインに近づけることが可能なためである。コンポーネント構成比を観点として学習セットを選別することにより、従来技術の問題を解消し、ターゲットドメインに十分に適合した学習セットを作ることができる。

【００１５】

なお、母体セットから部分集合を選別する選び方（組み合わせ）は無数に存在するため、全ての部分集合についてコンポーネント構成比を算出し、ターゲットドメインのコンポーネント構成比と最も構成比が近くなる部分集合を学習セットとする方法では、現実的な時間で処理できない。そこで、音声データセットのコンポーネント構成比がターゲットドメインに近いほど値が高くなる構成比類似スコアを定式化し、構成比類似スコアを最も大

きく向上させる音声データを１つずつ選んでいく貪欲法により学習セットを選別する。構成比類似スコアを劣モジュラ関数として定式化することにより、貪欲法でも準最適な学習セットを選別可能となっている。

【００１６】

［第一実施形態］

以下、ターゲットドメインで収録した数時間の音声を「ターゲットセット」と書く。母体セット及びターゲットセットの音声は、既存のＶＡＤ技術により、一呼吸で発声された音声区間（以下「発話」と書く）ごとに分割されているものとする。したがって、母体セットとターゲットセットは発話の集合である。通常、母体セットは数百万～数千万発話、ターゲットセットは数千発話程度のサイズである。

【００１７】

第一実施形態の学習データ生成装置は、図１に示すように、第一音響特徴量抽出部１０１１、第二音響特徴量抽出部１０１２、コンポーネント分解部１０２、コンポーネント含有量計算部１０３及び発話選択部１０４を例えば備えている。第一実施形態の学習データ生成装置の各部が、図２に例示する各ステップの処理を行うことにより、第一実施形態の学習データ生成方法が実現される。以下、学習データ生成装置の各部について説明する。

【００１８】

＜第一音響特徴量抽出部１０１１＞

入力：発話セット（ターゲットセット）

出力：各発話の音響特徴量系列（コンポーネント分解部１０２へ）

処理：第一音響特徴量抽出部１０１１は、入力されたターゲットセットである発話セットの各発話から音響特徴量系列を抽出し、得られた各発話の音響特徴量系列をコンポーネント分解部１０２に出力する（ステップＳ１）。

【００１９】

音響特徴量系列の抽出では、各発話の音声信号を数十ｍｓｅｃの音響分析フレームに分割し、各音響分析フレームから音響特徴量を抽出することで、音響特徴量系列を得る。各フレームの音響特徴量は実数値ベクトルであり、ＭＦＣＣやＬＰＣケプストラムなど既存のいずれの手法で抽出しても構わない。

【００２０】

＜コンポーネント分解部１０２＞

入力：ターゲットセットの各発話の音響特徴量系列（第一音響特徴量抽出部１０１１から）、コンポーネント数Ｍ

出力：コンポーネント群（コンポーネント含有量計算部１０３へ）、目標構成比（発話選択部１０４へ）

処理：コンポーネント分解部１０２は、入力されたターゲットセットの各発話の音響特徴量系列をＭ個のコンポーネントに分解し、ターゲットセットにおける各コンポーネントの構成比（目標構成比）を算出し、Ｍ個のコンポーネント（コンポーネント群）をコンポーネント含有量計算部１０３と、目標構成比を発話選択部１０４とに出力する（ステップＳ２）。入力されるコンポーネント数Ｍは所定の１以上の整数であり、ターゲットセットのサイズに応じて適切な値が異なるパラメータである。数千発話のターゲットセットに対しては通常は５１２程度の値に設定する。

【００２１】

分解方法しだいでコンポーネントの形式と構成比の算出方法は異なる。ここでは混合正規分布の当てはめによって分解する場合について説明する。この場合、コンポーネント数Ｍは混合正規分布の混合数を表す。入力された全てのターゲットセットの発話の音響特徴量系列に対して、例えば参考文献１などに記載されている一般的なＥＭアルゴリズムを用いて混合数Ｍの混合正規分布を当てはめる（各正規分布の混合重みと平均ベクトルと共分散行列を求める）。各正規分布（平均ベクトルおよび共分散行列）がコンポーネントであり、混合重みが当該コンポーネントの構成比を表すため、各正規分布の平均ベクトルと共分散行列をコンポーネント群としてコンポーネント含有量計算部１０３へ出力し、各正規

分布の混合重みを目標構成比として発話選択部１０４へ出力する。

【００２２】

〔参考文献１〕C.M. ビショップ, “パターン認識と機械学習　下”, pp.154-155, シュプリンガー・ジャパン株式会社, 2008-07-01.

なお、混合正規分布の当てはめ以外の分解方法として、例えばＫ－ｍｅａｎｓ法のようなクラスタリングを用いてＭ個のコンポーネントと構成比を得ることも可能である。クラスタリングを用いる場合、各クラスタの重心ベクトルをコンポーネントとし、各コンポーネントの構成比は当該クラスタに属する音響特徴量の個数の全体に対する割合として計算される。

【００２３】

＜第二音響特徴量抽出部１０１２＞

入力：発話セット（母体セット）

出力：各発話の音響特徴量系列（コンポーネント含有量計算部１０３へ）

処理：第二音響特徴量抽出部１０１２は、入力された母体セットである発話セットの各発話から音響特徴量系列を抽出し、得られた各発話の音響特徴量系列をコンポーネント含有量計算部１０３に出力する（ステップＳ３）。

【００２４】

音響特徴量系列の抽出では、各発話の音声信号を数十ｍｓｅｃの音響分析フレームに分割し、各音響分析フレームから音響特徴量を抽出することで、音響特徴量系列を得る。各フレームの音響特徴量は実数値ベクトルであり、ＭＦＣＣやＬＰＣケプストラムなど既存のいずれの手法で抽出しても構わない。なお、第二音響特徴量抽出部１０１２における音響特徴量系列の抽出方法は、第一音響特徴量抽出部１０１１の音響特徴量系列の抽出方法と同じであるとする。

【００２５】

＜コンポーネント含有量計算部１０３＞

入力：母体セットの各発話の音響特徴量系列（第二音響特徴量抽出部１０１２から）、コンポーネント群（コンポーネント分解部１０２から）

出力：母体セットの各発話のコンポーネント含有量（発話選択部１０４へ）

処理：コンポーネント含有量計算部１０３は、入力された母体セットの各発話の音響特徴量系列とコンポーネント群から、各発話が各コンポーネントをどの程度含有しているか（コンポーネント含有量）を算出し、コンポーネント含有量を発話選択部１０４へ出力する（ステップＳ４）。

【００２６】

ある発話のコンポーネント含有量は、当該発話の各フレームの音響特徴量のコンポーネント含有量の、全フレーム分の総和として計算する。コンポーネント分解部１０２で混合正規分布の当てはめにより分解した場合、Ｍ個の各コンポーネントは、各正規分布の平均ベクトルと共分散行列である。この場合、ある音響特徴量ｘのコンポーネント含有量は以下のように計算する。

【００２７】

（１）まず、コンポーネント含有量計算部１０３は、音響特徴量ｘに対して、１番目からＭ番目までの全ての正規分布における尤度を計算する。ｍ番目の正規分布（平均ベクトルμ_m、共分散行列S_m）の尤度L_mは以下の式で計算される。ｄは音響特徴量ベクトルの次元数である。

【００２８】

【数１】

$$L_m = \frac{1}{(\sqrt{2\pi})^d \sqrt{|S_m|}} \exp\left(-\frac{1}{2}(x-\mu_m)' S_m^{-1}(x-\mu_m)\right) \quad \cdots(1)$$

【００２９】

（２）コンポーネント含有量計算部１０３は、得られたL_1～L_MまでのＭ個の尤度を

和が１となるように正規化して、１番目からＭ番目までの各コンポーネント含有量を計算する。ｍ番目のコンポーネント含有量$Ｐ_{m}$の計算式は以下の通りである。

【００３０】

【数２】

$$P_m = \frac{L_m}{\sum_1^M L_m} \quad \cdots(2)$$

【００３１】

（２）で得られる$Ｐ_{1}$～$Ｐ_{M}$が、音響特徴量ｘの各コンポーネントの含有量である。当該発話中の各音響特徴量のコンポーネント含有量を計算し、コンポーネントごとに発話内で総和を取ることで、当該発話のコンポーネント含有量を計算する。

【００３２】

コンポーネント含有量計算部１０３は、母体セットの各発話に対して以上の手順でコンポーネント含有量を計算し、母体セットの各発話のコンポーネント含有量（各発話が各コンポーネントをどれだけ含有しているか）を発話選択部１０４へ出力する。

【００３３】

なお、コンポーネント分解部１０２でクラスタリングにより分解した場合、入力されるＭ個のコンポーネントは各クラスタの重心ベクトルである。この場合、音響特徴量ｘのコンポーネント含有量の計算方法が異なる。

【００３４】

この場合、１番目からＭ番目の各重心ベクトルと音響特徴量ｘとのユークリッド距離を計算し、最も距離の小さい重心ベクトルの含有量を１、その他の重心ベクトルの含有量を０とする。

【００３５】

音響特徴量ｘのコンポーネント含有量を計算した後の処理（発話内で総和を取り発話選択部１０４へ出力する）は混合正規分布を使った場合と同様である。

【００３６】

＜発話選択部１０４＞

入力：母体セット、母体セットの各発話のコンポーネント含有量（コンポーネント含有量計算部１０３から）、目標構成比（コンポーネント分解部１０２から）、選択停止条件Ｃ

出力：学習データ

処理：発話選択部１０４は、入力された母体セットの各発話のコンポーネント含有量と目標構成比と選択停止条件Ｃを用いて、母体セットから発話を選択して学習セットとして出力する（ステップＳ５）。

【００３７】

発話選択は以下の手順で行われる。

【００３８】

（０）発話選択部１０４は、学習セットＵを空集合に初期化する。

【００３９】

（１）発話選択部１０４は、学習セットＵに母体セット中の各発話を追加したときの構成比類似スコアの上昇値を計算する。母体セット中のｎ番目の発話$ｕ_{n}$を学習セットに追加したときの構成比類似スコアの上昇値は以下の式で計算する。

【００４０】

【数３】

$$\mathrm{Improve}_n = D(U \cup \{u_n\}) - D(U) \quad \cdots(3)$$

$$D(U) = \sum_{i=1}^{M} w_i \log f_{iU} \quad \cdots(4)$$

【００４１】

w_iはｉ番目のコンポーネントの目標構成比の値、f_{iU}はＵに含まれる全発話のｉ番目のコンポーネント含有量の総和である。

【００４２】

（２）発話選択部１０４は、最も大きく構成比類似スコアを上昇させる発話を母体セットから学習セットＵに移動する（Ｕに追加し、母体セットから削除する）。

【００４３】

（３）発話選択部１０４は、学習セットＵの発話数がＣ未満であれば（１）に戻って処理を繰り返す。発話選択部１０４は、学習セットＵの発話数がＣになれば終了し学習セットＵを学習データとして出力する。

【００４４】

選択停止条件Ｃとしてはいくつかの条件設定方法が考えられる。例えばＣとして学習セットＵの発話数の上限値を設定する方法がある。この場合は最終的に出力される学習セットの発話数をＣによって調整することができる、（３）において学習セットＵの発話数をチェックし、Ｃになっていれば終了する。Ｃは１以上母体セットの発話数以下の整数となり、例えば母体セットが１００万発話から為るときに通常はＣ＝１０万程度に設定する。また、Ｃとして構成比類似スコアの上昇値の下限値を設定する方法がある。この場合は構成比類似スコアがほとんど上昇しなくなった（＝必要十分な発話を選択できた）タイミングで発話選択を終了できる。Ｃは０以上の実数値となり、通常は$C=10^{-6}$程度に設定する。

【００４５】

式（４）の構成比類似スコアは、学習セットＵの中の各コンポーネントの構成比が目標構成比に近い場合に高くなる。（１）と（２）の繰り返しにより、構成比類似スコアをできるだけ高くするように発話を順次選択していくため、学習セットＵの中の各コンポーネントの構成比は目標構成比に近づいていく。そのため、上記の手順で最終的に出力された学習セットは、各コンポーネントを目標構成比に近い構成比で含んだ（ターゲットドメインに適合した）学習セットとなる。

【００４６】

なお、構成比類似スコアＤ(Ｕ)は劣モジュラ関数であるため、例えば参考文献２に記載されている、最大化したい関数が劣モジュラ関数である場合に上記の貪欲法と同一の学習セットをより少ない処理量で得られる高速化法を用いても構わない。

【００４７】

〔参考文献２〕Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen and Natalie Glance, “Cost-effective outbreak detection in networks,” in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.420-429, 2007.

ターゲットドメインに十分に適合した学習セットを選別できるようになると、ターゲットドメインで高精度に音声認識を行うシステムが構築可能になり、音声認識システム利用者にとっての利便性が向上する。

【００４８】

また、［発明が解決しようとする課題］の欄で述べた通り従来技術には問題があったため、あるドメイン向けに特化した音声認識システムを構築する場合、これまでは音声認識について深い知識を持つ人間がターゲットドメインの音声を聴取し、ターゲットドメインに十分に適合した学習セットを人手で選別していた。この人手による選別には大きなコストがかかるため、多くの導入先それぞれに特化した音声認識システムを構築することは困難なことが多かった。ターゲットドメインで数時間の音声を収録すれば、自動的にターゲットドメインに十分に適合する学習セットを選別できるようになるため、低コストにドメイン特化音声認識システムを構築することが可能となり、より多くの導入先で高精度な音声認識システムを運用できるようになる。

【００４９】

［第二実施形態］

第二実施形態の学習データ生成装置は、図３に示すように、第一発声／非発声区間抽出部２０１１、第二発声／非発声区間抽出部２０１２、第一音響特徴量抽出部１０１１、第二音響特徴量抽出部１０１２、コンポーネント分解部１０２、コンポーネント含有量計算部１０３及び発話選択部１０４を例えば備えている。第二実施形態の学習データ生成装置の各部が、図４に例示する各ステップの処理を行うことにより、第二実施形態の学習データ生成方法が実現される。以下、学習データ生成装置の各部について説明する。

【００５０】

母体セット、ターゲットセットの各発話を既存のＶＡＤ技術で切り出すときは、切り出し誤り（発声区間冒頭／末尾の音声が切れてしまう問題）を防ぐため、前後に１秒程度の非発声区間（ポーズ区間）を含むように切り出すことが多い。この各発話に付随する非発声区間を利用して周辺雑音環境をより精緻にコンポーネント分解することで、よりターゲットドメインに適合した学習セットを選別するのが第二実施形態である。

【００５１】

第二実施形態では、はじめに母体セットとターゲットセットのそれぞれの各発話を非発声区間と発声区間に分割し、以降、非発声区間と発声区間のそれぞれで第一音響特徴量抽出部１０１１、第二音響特徴量抽出部１０１２、コンポーネント分解部１０２及びコンポーネント含有量計算部１０３の処理を行い、非発声区間のコンポーネント・目標構成比と発声区間のコンポーネント・目標構成比とを別々に求める。母体セットからの発話選択は、非発声区間のコンポーネント含有量と目標構成比、発声区間のコンポーネント含有量と目標構成比をともに用いて行われる。

【００５２】

各発話の非発声区間は当該発話の周辺雑音のみを音として含むため、非発声区間をコンポーネントに分解することで、周辺雑音の特徴を精緻に捉えたコンポーネントが得られる。この非発声区間のコンポーネントを用いて発話を選択することで、第一実施形態以上に周辺雑音環境がターゲットセットに適合した学習セットを得ることができる。

【００５３】

＜第一発声／非発声区間抽出部２０１１＞

入力：発話セット（ターゲットセット）

出力：非発声発話セットと発声発話セット（第一音響特徴量抽出部１０１１へ）

処理：第一発声／非発声区間抽出部２０１１は、入力されたターゲットセットである発話セットの各発話を既存のＶＡＤ技術を用いて非発声区間と発声区間に分割し、各発話の非発声区間を非発声ターゲットセット、各発話の発声区間を発声ターゲットセットとして出力する。

【００５４】

このようにして、第一発声／非発声区間抽出部２０１１は、ターゲットセットの各発話を非発声区間と発生区間とに分割することにより非発声ターゲットセット及び発声ターゲットセットを生成する（ステップＳ０１）。

【００５５】

＜第二発声／非発声区間抽出部２０１２＞

入力：発話セット（母体セット）

出力：非発声母体セットと発声母体セット（第二音響特徴量抽出部１０１２へ）

処理：第二発声／非発声区間抽出部２０１２は、入力された母体セットである発話セットの各発話を既存のＶＡＤ技術を用いて非発声区間と発声区間に分割し、各発話の非発声区間を非発声母体セット、各発話の発声区間を発声母体セットとして出力する。

【００５６】

このようにして、第二発声／非発声区間抽出部２０１２は、上記母体セットの各発話を非発声区間と発声区間とに分割することにより非発声母体セット及び発声母体セットを生成する（ステップＳ０２）。

【００５７】

＜第一音響特徴量抽出部１０１１＞

入力：非発声ターゲットセット及び発声ターゲットセット（第一発声／非発声区間抽出部２０１１から）
出力：非発声ターゲットセットの各発話の音響特徴量系列及び発声ターゲットセットの各発話の音響特徴量系列（コンポーネント分解部１０２へ）
処理：第一音響特徴量抽出部１０１１は、非発声ターゲットセット及び発声ターゲットセットのそれぞれに対して、第一実施形態の第一音響特徴量抽出部１０１１と同様の処理を行う。

【００５８】
すなわち、第一音響特徴量抽出部１０１１は、非発声ターゲットセットの各発話の音声信号から音響特徴量系列を抽出し、発声ターゲットセットの各発話の音声信号から音響特徴量系列を抽出し、抽出されたそれぞれの音響特徴量系列をコンポーネント分解部１０２に出力する（ステップＳ１）。

【００５９】
音響特徴量系列の抽出の詳細については、第一実施形態と同様であるため、ここでは重複説明を省略する。

【００６０】
＜コンポーネント分解部１０２＞
入力：非発声ターゲットセットの各発話の音響特徴量系列及び発声ターゲットセットの各発話の音響特徴量系列（第一音響特徴量抽出部１０１１から）、非発声コンポーネント数M_1、発声コンポーネント数M_2
出力：非発声コンポーネント群及び発声コンポーネント群（コンポーネント含有量計算部１０３へ）、非発声目標構成比及び発声目標構成比（発話選択部１０４へ）
処理：コンポーネント分解部１０２は、非発声ターゲットセットの各発話の音響特徴量系列と、発声ターゲットセットの各発話の音響特徴量系列とのそれぞれに対して、第一実施形態のコンポーネント分解部１０２と同様の処理を行う。

【００６１】
すなわち、コンポーネント分解部１０２は、第一音響特徴量抽出部１０１１で抽出された非発声ターゲットセットの各発話の音響特徴量系列をM_1個の非発声コンポーネントに分解し、非発声ターゲットセットにおける各非発声コンポーネントの構成比である非発声目標構成比を求め、第一音響特徴量抽出部１０１１で抽出された発声ターゲットセットの各発話の音響特徴量系列をM_2個の発声コンポーネントに分解し、発声ターゲットセットにおける各発声コンポーネントの構成比である発声目標構成比を求める（ステップＳ２）。

【００６２】
入力されるコンポーネント数M_1，M_2は所定の１以上の整数であり、ターゲットセットのサイズに応じて適切な値が異なるパラメータである。数千発話のターゲットセットに対しては通常は５１２程度の値に設定する。M_1，M_2の値は、異なっていても同じでもよい。

【００６３】
M_1個の非発声コンポーネントである非発声コンポーネント群及びM_2個の発声コンポーネントである発声コンポーネント群は、コンポーネント含有量計算部１０３に出力される。

【００６４】
コンポーネント分解部１０２の処理の詳細については、第一実施形態と同様であるため、ここでは重複説明を省略する。

【００６５】
＜第二音響特徴量抽出部１０１２＞
入力：非発声母体セット及び発声母体セット（第二発声／非発声区間抽出部２０１２から）
出力：非発声母体セットの各発話の音響特徴量系列及び発声母体セットの各発話の音響

特徴量系列（コンポーネント含有量計算部１０３へ）
処理：第二音響特徴量抽出部１０１２は、非発声母体セット及び発声母体セットのそれぞれに対して、第一実施形態の第二音響特徴量抽出部１０１２と同様の処理を行う。
【００６６】
すなわち、第二音響特徴量抽出部１０１２は、非発声母体セットの各発話の音声信号から音響特徴量系列を抽出し、発声母体セットの各発話の音声信号から音響特徴量系列を抽出する（ステップＳ３）
音響特徴量系列の抽出の具体例については、第一実施形態と同様であるため、ここでは重複説明を省略する。
【００６７】
＜コンポーネント含有量計算部１０３＞
入力：非発声母体セットの各発話の音響特徴量系列及び発声母体セットの各発話の音響特徴量系列（第二音響特徴量抽出部１０１２から）、非発声コンポーネント群及び発声コンポーネント群（コンポーネント分解部１０２から）
出力：非発声母体セットの各発話のコンポーネント含有量及び発声母体セットの各発話のコンポーネント含有量（発話選択部１０４へ）
処理：コンポーネント含有量計算部１０３は、非発声母体セットの各発話の音響特徴量系列及び発声母体セットの各発話の音響特徴量系列のそれぞれに対して、第一実施形態のコンポーネント含有量計算部１０３と同様の処理を行う。
【００６８】
すなわち、コンポーネント含有量計算部１０３は、第二音響特徴量抽出部１０１２で抽出された非発声母体セットの各発話の音響特徴量系列を用いて、非発声母体セットに含まれる各発話における各非発声コンポーネントの含有量を計算し、第二音響特徴量抽出部１０１２で抽出された発声母体セットの各発話の音響特徴量系列を用いて、発声母体セットに含まれる各発話における各発声コンポーネントの含有量を計算する（ステップＳ４）。
【００６９】
コンポーネント含有量計算部１０３の処理の詳細については、第一実施形態と同様であるため、ここでは重複説明を省略する。
【００７０】
＜発話選択部１０４＞
入力：非発声母体セットの各発話のコンポーネント含有量（コンポーネント含有量計算部１０３から）、非発声目標構成比（コンポーネント分解部１０２から）、発声母体セットの各発話のコンポーネント含有量（コンポーネント含有量計算部１０３から）、発声目標構成比（コンポーネント分解部１０２から）、選択停止条件Ｃ
出力：学習データ
処理：発話選択部１０４は、入力された母体セット各発話の非発声区間および発声区間のコンポーネント含有量と、非発声区間および発声区間の目標構成比と選択停止条件Ｃを用いて、母体セットから発話を選択して学習セットとして出力する。
【００７１】
発話選択部１０４の発話選択手順のうち、構成比類似スコアの計算式（４）を以下の式（５）に置き換える以外はまったく同じ手順で発話を選択する。
【００７２】
【数４】

$$D(U)=\sum_{i=1}^{M_1} w_i^{NOISE}\log f_{iU}^{NOISE}+\sum_{j=1}^{M_2} w_j^{SPEECH}\log f_{jU}^{SPEECH} \quad \cdots(5)$$

【００７３】
M_1は非発声コンポーネント数（非発声区間のコンポーネント分解部１０２に設定された値、通常は２５６程度）、M_2は発声コンポーネント数（発声区間のコンポーネント分解部１０２に設定された値、通常は２５６程度）、w_i^{NOISE}はｉ番目の非発声コンポーネ

ントの目標構成比、f_{iU}^{NOISE}はＵに含まれる全発話のｉ番目の非発声コンポーネント含有量の総和、w_j^{SPEECH}はｊ番目の発声コンポーネントの目標構成比、f_{jU}^{SPEECH}はＵに含まれる全発話のｊ番目の発声コンポーネント含有量の総和である。

【００７４】

式（５）の構成比類似スコアは、学習セットＵの中の各非発声コンポーネントの構成比が非発声目標構成比に近く、かつ、各発声コンポーネントの構成比が発声目標構成比に近い場合に高くなる。手順（１）と（２）の繰り返しにより、構成比類似スコアをできるだけ高くするように発話を順次選択していくため、学習セットＵの中の各非発声／発声コンポーネントの構成比はそれぞれの目標構成比に近づいていく。そのため、上記の手順で最終的に出力された学習セットは、各非発声／発声コンポーネントを目標構成比に近い構成比で含んだ（ターゲットドメインに適合した）学習セットとなる。非発声区間だけに絞って分解した非発声コンポーネントの構成比を考慮するため、第一実施形態よりも周辺雑音環境がよりターゲットドメインに適合した学習セットとなる。

【００７５】

このようにして、発話選択部１０４は、非発声母体セットに含まれる各発話における各非発声コンポーネントの含有量及び発声母体セットに含まれる各発話における各発声コンポーネントの含有量に基づいて、母体セットに含まれる発話の中から、学習セットにおける非発声コンポーネントの構成比が非発声目標構成比に近くなるように、かつ、学習セットにおける発声コンポーネントの構成比が発声目標構成比に近くなるように、発話を選択することにより学習データを生成する（ステップＳ５）。

【００７６】

なお、式（５）の構成比類似スコアＤ(Ｕ)は劣モジュラ関数の和を取る関数となっている。劣モジュラ関数の和を取る関数もまた劣モジュラ関数であるため、式（５）の構成比類似スコアＤ(Ｕ)も劣モジュラ関数である。そのため、発話選択部１０４で利用可能な文献３に記載されている高速化法は、発声／非発声発話選択部２０２でも利用可能である。

【００７７】

[プログラム及び記録媒体]

上記学習データ生成装置及び方法において説明した処理は、記載の順にしたが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。

【００７８】

また、学習データ選択置における各処理をコンピュータによって実現する場合、その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、その各処理がコンピュータ上で実現される。

【００７９】

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。

【００８０】

また、各処理手段は、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構成することにしてもよいし、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。

【００８１】

その他、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。

【符号の説明】

【００８２】

２０１１第一発声／非発声区間抽出部
２０１２第二発声／非発声区間抽出部
１０１１第一音響特徴量抽出部

１０１２第二音響特徴量抽出部
１０２　コンポーネント分解部
１０３　コンポーネント含有量計算部
１０４　発話選択部

【図１】

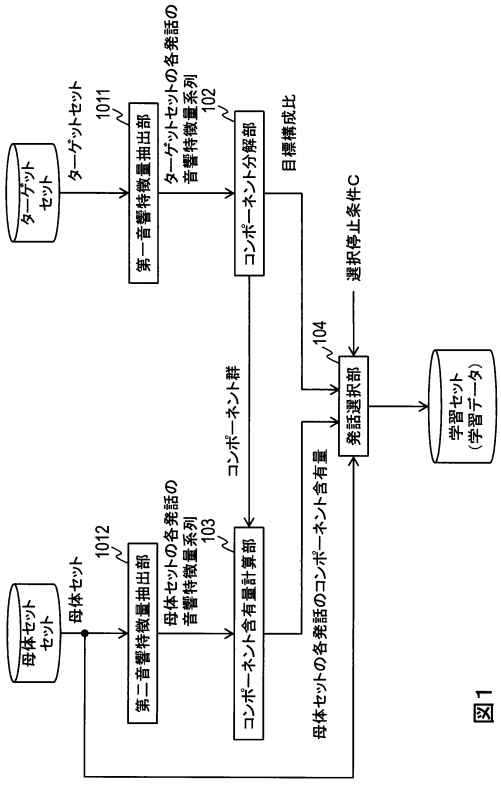

図1

【図２】

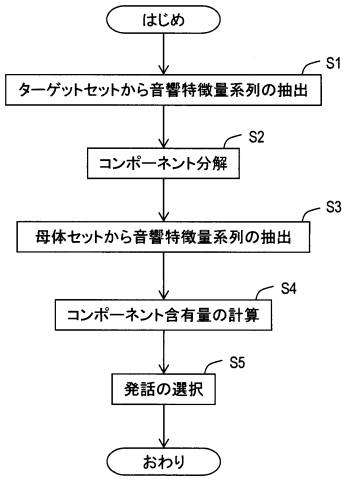

図2

【図３】

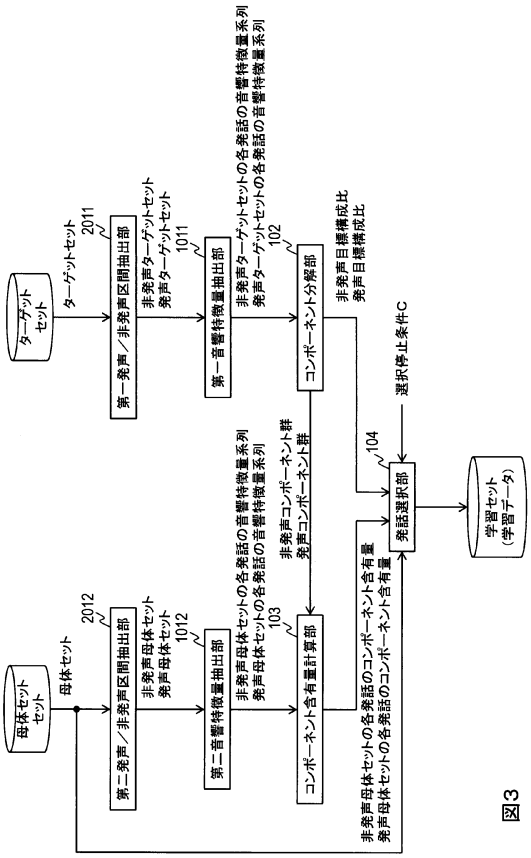
ターゲットセット
ターゲットセット
2011
第一発声／非発声区間抽出部
非発声ターゲットセット
発声ターゲットセット
1011
第一音響特徴量抽出部
非発声ターゲットセットの各発話の音響特徴量系列
発声ターゲットセットの各発話の音響特徴量系列
102
コンポーネント分解部
非発声目標構成比
発声目標構成比
選択停止条件C
104
母体セットセット
母体セット
2012
第二発声／非発声区間抽出部
非発声母体セット
発声母体セット
1012
第二音響特徴量抽出部
非発声母体セットの各発話の音響特徴量系列
発声母体セットの各発話の音響特徴量系列
103
コンポーネント含有量計算部
非発声コンポーネント群
発声コンポーネント群
非発声母体セットの各発話のコンポーネント含有量
発声母体セットの各発話のコンポーネント含有量
発話選択部
学習セット
（学習データ）

図3

【図４】

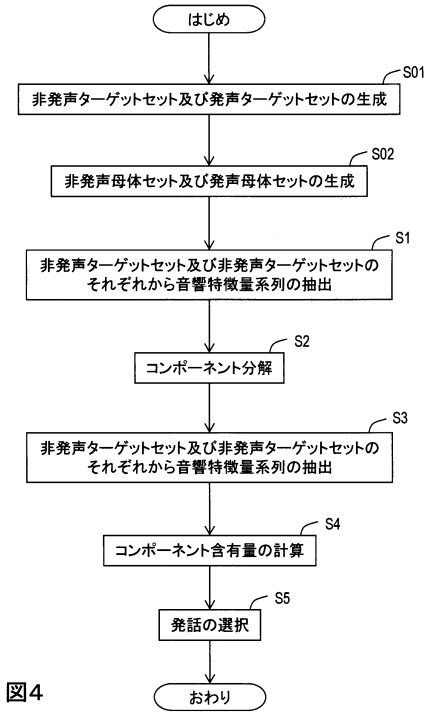
はじめ
S01
非発声ターゲットセット及び発声ターゲットセットの生成
S02
非発声母体セット及び発声母体セットの生成
S1
非発声ターゲットセット及び非発声ターゲットセットの
それぞれから音響特徴量系列の抽出
S2
コンポーネント分解
S3
非発声ターゲットセット及び非発声ターゲットセットの
それぞれから音響特徴量系列の抽出
S4
コンポーネント含有量の計算
S5
発話の選択
おわり

図4

フロントページの続き

審査官　上田　雄

(56)参考文献　特開２００９－１２８４９０（ＪＰ，Ａ）
国際公開第２０１０／０４７０１９（WO，Ａ1）
特開２００４－４５０９（ＪＰ，Ａ）
ツィンツァレク・トビアス 他，タスク依存音響モデルのための発話レベルでの選択学習法，情報処理学会研究報告，社団法人情報処理学会，２００５年１２月２２日，Vol.2005，No.127，pp.235-240
篠原　雄介，劣モジュラ最適化を用いた文部分集合選択によるコーパス構築法，電子情報通信学会技術研究報告，日本，一般社団法人電子情報通信学会，２０１４年１２月１５日，Vol.114，No.365，pp. 35-39
Olivier Siohan，Michiel Bacchiani，iVector-based Acoustic Data Selection，Proc. INTERSPEECH 2013，フランス，２０１３年　8月２５日，pp. 657-661

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１５／００－１５／３４