

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7472575号
(P7472575)

(45)発行日 令和6年4月23日(2024.4.23)

(24)登録日 令和6年4月15日(2024.4.15)

(51)国際特許分類

FI

G 1 0 L 21/0272(2013.01)

G 1 0 L 21/0272 1 0 0 B

G 1 0 L 17/00 (2013.01)

G 1 0 L 17/00 2 0 0 C

請求項の数 12 (全17頁)

(21)出願番号	特願2020-51019(P2020-51019)	(73)特許権者	000004075
(22)出願日	令和2年3月23日(2020.3.23)		ヤマハ株式会社
(65)公開番号	特開2021-149784(P2021-149784 A)	(74)代理人	静岡県浜松市中央区中沢町10番1号 110000154
(43)公開日	令和3年9月27日(2021.9.27)		弁理士法人はるか国際特許事務所
審査請求日	令和5年1月25日(2023.1.25)	(72)発明者	高橋 祐
			静岡県浜松市中区中沢町10番1号 ヤマハ株式会社内
		(72)発明者	大竹 徹郎
			静岡県浜松市中区中沢町10番1号 ヤマハ株式会社内
		審査官	堀 洋介

最終頁に続く

(54)【発明の名称】 処理方法、処理装置、及びプログラム

(57)【特許請求の範囲】

【請求項1】

音信号のスペクトログラムを取得し、
前記スペクトログラムに対し、周波数軸又は時間軸における所定幅ごとに第1の畳み込みを行い、
前記所定幅ごとに行われた第1の畳み込みの結果を合わせて、1次元の第1特徴データを得て、
前記第1特徴データに対し、少なくとも1回の第2の畳み込みを行って、前記スペクトログラムの特徴を示す1次元の第2特徴データを得る、
処理方法。

【請求項2】

前記第1の畳み込みの結果を合わせて、前記第1特徴データを得て、
前記第1特徴データに対し、少なくとも1回の前記第2の畳み込みとプーリングを行って、前記第2特徴データを得る、
請求項1に記載の処理方法。

【請求項3】

前記スペクトログラムに対し、前記所定幅ごとに、前記所定幅で所定長のフィルタで前記第1の畳み込みを行い、
前記第1特徴データに対し、1次元のフィルタで少なくとも1回の前記第2の畳み込みを行って、前記第2特徴データを得る、

請求項 1 又は 2 に記載の処理方法。

【請求項 4】

前記所定幅は、周波数軸における幅である、
請求項 1 ～ 3 の何れかに記載の処理方法。

【請求項 5】

前記所定幅は、1 周波数ビンの幅である、
請求項 4 に記載の処理方法。

【請求項 6】

前記第 1 の畳み込みの結果の和を計算して、前記第 1 特徴データを得る、
請求項 1 ～ 5 の何れかに記載の処理方法。

10

【請求項 7】

前記所定長の幅ごとに、独立にフィルタが用意されており、
前記スペクトログラムに対し、前記所定長の幅ごとに、対応するフィルタで畳み込みを行う、
請求項 1 ～ 6 の何れかに記載の処理方法。

【請求項 8】

前記スペクトログラムは、所定の音を含む複数の音が混合された音信号を示し、
前記第 2 特徴データに対し、少なくとも 1 回の逆畳み込みを行って、前記所定の音を分離するマスクを得て、
前記スペクトログラムに前記マスクを適用し、前記複数の音の中から前記所定の音を分離する、
請求項 1 ～ 7 の何れかに記載の処理方法。

20

【請求項 9】

前記少なくとも 1 回の逆畳み込みでは、各層の入力データに対し、対応する畳み込み層で得られたデータを付加して、逆畳み込みが行われる、
請求項 8 に記載の処理方法。

【請求項 10】

前記第 1 の畳み込み、前記第 2 の畳み込み、及び前記逆畳み込みに用いられる変数は、
複数の音が混合された音信号のスペクトログラムと、当該複数の音に含まれる前記所定の音と、を含む訓練データのスペクトログラムから、前記処理方法により前記訓練データの特定の音が分離されるように、繰り返し調整して決定された変数である、
請求項 8 又は 9 に記載の処理方法。

30

【請求項 11】

音信号のスペクトログラムを取得し、
前記スペクトログラムに対し、周波数軸又は時間軸における所定幅ごとに第 1 の畳み込みを行い、
前記所定幅ごとに行われた第 1 の畳み込みの結果を合わせて、1 次元の第 1 特徴データを得て、
前記第 1 特徴データに対し、少なくとも 1 回の第 2 の畳み込みを行って、前記スペクトログラムの特徴を示す 1 次元の第 2 特徴データを得る、
処理装置。

40

【請求項 12】

コンピュータに、
音信号のスペクトログラムを取得させ、
前記スペクトログラムに対し、周波数軸又は時間軸における所定幅ごとに第 1 の畳み込みを行わせ、
前記所定幅ごとに行われた第 1 の畳み込みの結果を合わせて、1 次元の第 1 特徴データを得させ、
前記第 1 特徴データに対し、少なくとも 1 回の第 2 の畳み込みを行って、前記スペクトログラムの特徴を示す 1 次元の第 2 特徴データを得させる、

50

ためのプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、処理方法、処理装置、及びプログラムに関する。

【背景技術】

【0002】

近年、学習モデルを利用して、音信号のスペクトログラムを解析する技術が検討されている。例えば、非特許文献1には、複数の音が混合された音信号のスペクトログラムに対し、2次元の畳み込みを繰り返し行って、2次元の特徴データを得る技術が記載されている。この技術では、2次元の特徴データに基づいて、複数の音の中から所定の音を分離するためのマスクが生成される。

10

【先行技術文献】

【非特許文献】

【0003】

【文献】ISMIR 2017, 「SINGING VOICE SEPARATION WITH DEEP U-NET CONVOLUTIONAL NETWORKS」, Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, Tillman Weyde

20

【発明の概要】

【発明が解決しようとする課題】

【0004】

しかしながら、非特許文献1のように2次元の特徴データを得る技術では、畳み込みの際にスペクトログラムの局所的な情報しか考慮されない。例えば、高域まで調波構造を有する音声は、周波数方向に広範囲に特徴的な情報を有するので、局所的な情報だけを考慮しても、音声の特徴データを精度良く得ることができない。スペクトログラム全体に分散された特徴量を考慮して精度の良い特徴データを得るには、学習モデルの層を深くする必要又は大きなフィルタを利用する必要があるので、スペクトログラムの特徴を効率良く表現する特徴データを得られない。

30

【0005】

本発明は上記課題を鑑みてなされたものであって、その目的は、音信号のスペクトログラムの特徴を効率良く表現する特徴データを得ることである。

【課題を解決するための手段】

【0006】

上記課題を解決するために、本発明に係る処理方法は、音信号のスペクトログラムを取得し、前記スペクトログラムに対し、周波数軸又は時間軸における所定幅ごとに第1の畳み込みを行い、前記所定幅ごとに行われた第1の畳み込みの結果を合わせて、1次元の第1特徴データを得て、前記第1特徴データに対し、少なくとも1回の第2の畳み込みを行って、前記スペクトログラムの特徴を示す1次元の第2特徴データを得る。

40

【0007】

本発明に係る処理装置は、音信号のスペクトログラムを取得し、前記スペクトログラムに対し、周波数軸又は時間軸における所定幅ごとに第1の畳み込みを行い、前記所定幅ごとに行われた第1の畳み込みの結果を合わせて、1次元の第1特徴データを得て、前記第1特徴データに対し、少なくとも1回の第2の畳み込みを行って、前記スペクトログラムの特徴を示す1次元の第2特徴データを得る。

【0008】

本発明に係るプログラムは、コンピュータに、音信号のスペクトログラムを取得させ、前記スペクトログラムに対し、周波数軸又は時間軸における所定幅ごとに第1の畳み込みを行わせ、前記所定幅ごとに行われた第1の畳み込みの結果を合わせて、1次元の第1特

50

徴データを得させ、前記第1特徴データに対し、少なくとも1回の第2の畳み込みを行って、前記スペクトログラムの特徴を示す1次元の第2特徴データを得させる。

【発明の効果】

【0009】

本発明によれば、音信号のスペクトログラムの特徴を効率良く表現する特徴データを得ることができる。

【図面の簡単な説明】

【0010】

【図1】実施形態に係る処理装置の一例を示す図である。

【図2】処理装置で実現される機能の一例を示すブロック図である。

10

【図3】音信号のスペクトログラムの一例を示す図である。

【図4】学習モデルにより実行される処理の全体的な流れを示す図である。

【図5】2次元のスペクトログラムが1次元の信号にみなされる様子を示す図である。

【図6】1次元の信号が畳み込まれる処理を示す図である。

【図7】調整処理の一例を示すフロー図である。

【図8】分離処理の一例を示すフロー図である。

【発明を実施するための形態】

【0011】

〔1. 処理装置のハードウェア構成〕

以下、本発明に係る実施形態の一例を図面に基づいて説明する。図1は、実施形態に係る処理装置の一例を示す図である。例えば、処理装置10は、デジタルミキサ、信号処理エンジン、オーディオ装置、電子楽器、エフェクタ、パーソナルコンピュータ、スマートフォン、又はタブレット端末である。図1に示すように、処理装置10は、CPU11、不揮発メモリ12、RAM13、操作部14、表示部15、入力部16、及びスピーカ17に接続される。

20

【0012】

CPU11は、少なくとも1つのプロセッサを含む。1チップの中の複数プロセッサに限られず、ネットワーク等で接続された複数の装置に分散された複数のプロセッサであってもよい。CPU11は、不揮発メモリ12に記憶されたプログラム及びデータに基づいて、所定の処理を実行する。不揮発メモリ12は、ROM、EEPROM、フラッシュメモリ、又はハードディスク等のメモリである。RAM13は、揮発メモリの一例である。操作部14は、タッチパネル、キーボード、マウス、ボタン、又はレバー等の入力デバイスである。表示部15は、液晶ディスプレイ又は有機ELディスプレイ等のディスプレイである。

30

【0013】

入力部16は、音信号を取得する。音信号は、音を示す信号である。音響信号又は音声信号は、音信号の一種である。音は、人間が発する音声に限られない。音信号は、任意の音を示せばよい。例えば、音信号は、人間以外の動物の音声、音楽、動画に含まれる音、機械の音、乗り物の音、自然現象の音、又はこれらの少なくとも2つが混合された音を示してもよい。本実施形態では、音信号がデジタルの信号である場合を説明する。音信号は、アナログの信号であってもよい。入力部16は、デジタルの音信号をアナログの音信号に変換し、スピーカ17に入力する。スピーカ17は、入力されたアナログの音信号に応じた音を出力する。

40

【0014】

本実施形態では、「得る」は、処理の結果として得ることを意味する。例えば、後述する特徴データは、後述する学習モデルによる処理の結果として得られるので、処理装置10は、特徴データを「得る」。「得る」は、作成する、定義する、又は生成すると言い換えることもできる。一方、「取得する」は、受け取ることを意味する。例えば、本実施形態では、音信号のスペクトログラムは、不揮発メモリ12から受け取るものなので、処理装置10は、スペクトログラムを取得する。「取得する」は、受信すると言い換えること

50

もできる。本実施形態では、このようにして「得る」と「取得する」を使い分ける。

【0015】

なお、処理装置10のハードウェア構成は、上記の例に限られない。例えば、処理装置10は、有線通信又は無線通信用の通信インタフェースを含んでもよい。また例えば、処理装置10は、コンピュータ読み取り可能な情報記憶媒体を読み取る読取装置（例えば、光ディスクドライブ又はメモリカードスロット）を含んでもよい。また例えば、処理装置10は、データの入出力をするための入出力端子（例えば、USBポート）を含んでもよい。本実施形態で不揮発メモリ12に記憶されるものとして説明するプログラム及びデータは、通信インタフェース、読取装置、又は入出力端子を介して処理装置10に供給されてもよい。

10

【0016】

〔2. 処理装置で実現される機能〕

図2は、処理装置10で実現される機能の一例を示すブロック図である。本実施形態では、音を分離する処理を例に挙げて、処理装置10で実現される機能を説明する。後述する変形例のように、処理装置10は、音を分離する処理以外の他の処理を実行してもよい。図2に示すように、処理装置10では、データ記憶部100、第1取得部101、第1畳み込み部102、合成部103、第2畳み込み部104、逆畳み込み部105、分離部106、及び調整部107が実現される。データ記憶部100は、不揮発メモリ12を主として実現され、他の各機能は、CPU11を主として実現される。

【0017】

〔2-1. データ記憶部〕

データ記憶部100は、本実施形態で説明する処理を実行するために必要なデータを記憶する。本実施形態では、このデータの一例として、音信号のスペクトログラム、訓練データ、及び学習モデルを説明する。

20

【0018】

図3は、音信号のスペクトログラムの一例を示す図である。スペクトログラムSGは、短時間フーリエ変換やバンドパスフィルタ等を用いて、時間領域の音信号を周波数領域に変換して得られる。本実施形態では、音分離の処理対象となるスペクトログラムに「SG」の符号を付す。訓練データに含まれるスペクトログラム等については、「SG」の符号を付さない。

30

【0019】

例えば、スペクトログラムSGは、2次元のデータである。横軸は、時間軸である。縦軸は、周波数軸である。例えば、スペクトログラムSGは、2次元形式で表現される。この2次元形式のデータは、画像データであってもよい。

【0020】

スペクトログラムSGの各値は、対応するフレームにおける各周波数成分の強さ（振幅）を示す。図3の例では、各画素の色を、模式的に網点の濃さで表現する。例えば、画素の色が明るさは、その画素に対応する時間における周波数の音信号が強さを示す。色と周波数の強さは、この関係に限られず、任意の関係にあってよい。本実施形態では、スペクトログラムSGのうち1回の処理に用いるデータのサイズを100×2000とするが、このサイズ（ビン数およびフレーム数）は、任意であってもよい。なお、本実施形態で「X×Y」（XとYは自然数）と記載した場合、この記載はデータのサイズを表す。例えば、Xは周波数軸におけるデータ数であり、Yは時間軸におけるデータ数である。

40

【0021】

なお、スペクトログラムSGは、図3の例に限られない。スペクトログラムSGは、任意の形式であってもよい。スペクトログラムSGは、リニアスケールではなく、対数スケールであってもよい。

【0022】

本実施形態のスペクトログラムSGは、所定の音を含む複数の音が混合された音信号から算出される。所定の音とは、分離の対象となる音である。所定の音は、単一の音（ソロ

50

信号)でもよいし、複数の音(混合信号)でもよい。

【0023】

例えば、所定の音が人間の音声であり、他の音が楽器の音であってもよい。この場合、スペクトログラムSGは、人間の音声と楽器の音が混合された音信号を示す。本実施形態の処理により、この音信号から人間の音声が分離される。

【0024】

データ記憶部100は、機械学習又は深層学習における訓練データを記憶する。機械学習又は深層学習自体は、画像や音声の処理における種々の手法を利用可能である。本実施形態では、畳み込みニューラルネットワークを例に挙げる。畳み込みニューラルネットワークの具体例としては、画像から特定の領域を抽出するU-Netと呼ばれる手法、又は、U-Netを利用した非特許文献1の手法であってもよい。本実施形態の手法は、従来の手法と比較して、大まかな枠組みはやや似ているが、具体的な処理が根本的に異なる。

【0025】

訓練データは、学習モデルを訓練するため(変数の調整)に用いられる。訓練データは、入力と出力(正解)のペアである。別の言い方をすれば、訓練データは、学習モデルに入力されるデータと同じ形式のデータと、学習モデルが出力すべき正解となるデータとのペアである。本実施形態では、訓練データは、1つのペアを意味する。例えば、データ記憶部100は、互いに異なる内容の複数の訓練データを記憶する。

【0026】

本実施形態では、訓練データは、入力としての、複数の音が混合された音信号のスペクトログラムと、出力としての、当該複数の音に含まれる所定の音信号のスペクトログラムと、を含む。このスペクトログラムは、学習モデルに入力されるスペクトログラムSG(分離の対象となるスペクトログラムSG)と同じ形式である。この所定の音は、学習モデルが出力するデータの形式と同じ形式で表現される。

【0027】

例えば、訓練データに含まれる音信号のスペクトログラムは、2次元形式のデータである。このスペクトログラムは、1つの軸が周波数軸であり、もう1つの軸が時間軸である。

【0028】

例えば、訓練データは、処理装置10のユーザによって用意される。ユーザは、分離の対象となる所定の音と他の音とを別々に録音する。ユーザは、録音した所定の音と他の音を混合し、混合音を得、その混合音を周波数領域のデータに変換してスペクトログラムを得る。ユーザは、このスペクトログラムを入力とし、最初に録音した所定の音を出力(正解)とするペアを訓練データとして作成する。ユーザは、種々の音について同様の作業を行い、複数の訓練データ(データセット)を作成する。

【0029】

データ記憶部100は、学習モデルを記憶する。本実施形態では、学習モデルは、教師あり学習により訓練される。例えば、学習モデルは、複数層からなるエンコーダと、複数層からなるデコーダと、を含む。本実施形態では、同じ階層のエンコーダとデコーダがスキップ接続される場合を説明するが、スキップ接続は省略してよい。

【0030】

エンコーダは、複数の畳み込み層と1以上のプーリング層を含む。デコーダは、エンコーダの各層に対応する、複数の逆畳み込み層と1以上のアップサンプリング層を含む。これらの層は、畳み込みニューラルネットワークである。例えば、学習モデルは、畳み込み係数などの変数を含む。フィルタの係数やバイアスが変数の一例である。

【0031】

例えば、データ記憶部100は、学習前の学習モデルを記憶する。学習前の学習モデルは、後述する調整部107により変数が調整される前の学習モデルである。変数が調整された学習モデルは、学習済みのモデルとしてデータ記憶部100に記憶される。追加学習が実行される場合には、学習済みのモデルの変数が追加学習によって更新される。

【0032】

10

20

30

40

50

図4は、学習モデルにより実行される処理の全体的な流れを示す図である。図5は、スライスされた2次元のスペクトログラムを処理して1次元のデータを得る処理を示す図である。図6は、1次元のデータを処理して2次元のデータを得る処理を示す図である。第1畳み込み部102から第2畳み込み部104がエンコーダであり、逆畳み込み部105がデコーダである。以降、図4～図6を参照し、これら各機能の詳細を説明する。

【0033】

[2-2. 第1取得部]

第1取得部101は、音信号のスペクトログラムSGを取得する。音信号が2000フレームより長い場合には、2000フレームごとのスペクトログラムに分割されて処理が実行される。この場合、同じ音信号の分離について学習モデルを訓練するために、複数のスペクトログラムが用いられてもよい。

10

【0034】

例えば、処理装置10は、公知のアルゴリズムに基づいて、音信号の周波数スペクトルを計算して、スペクトログラムSGを生成する。音信号は、データ記憶部100、外部の装置、又は外部の情報記憶媒体に記憶される。処理装置10は、入力部16から入力された音信号をデジタルデータに変換し、スペクトログラムSGを生成してもよい。

【0035】

[2-3. 第1畳み込み部]

第1畳み込み部102は、スペクトログラムSGに対し、周波数軸又は時間軸における所定幅ごとに同幅のフィルタによる第1の畳み込みを行う。所定幅とは、周波数軸又は時間軸における一定の長さの幅である。所定幅は、周波数軸又は時間軸の分解能と一致してもよいし、分解能の整数倍の幅であってもよい。

20

【0036】

本実施形態では、スペクトログラムSGが2次元形式で表現され、所定幅は、少なくとも1分解能の幅である。所定幅と、後述する第1特徴データ（畳み込みの結果）の次元数と、は相互に独立な値である。本実施形態では、第1畳み込み部102は、スペクトログラムSGに対し、周波数軸における所定幅ごとに第1の畳み込みを行う。

【0037】

本実施形態では、所定幅は、1周波数ピンの幅である。1周波数ピンとは、スペクトログラムSGにおける周波数の分解能である。なお、第1畳み込み部102は、2周波数ピンごと又は3周波数ピンごとに第1の畳み込みを行ってもよい。

30

【0038】

第1の畳み込みは、エンコーダにおける最初の畳み込み層（1段階目の畳み込み層）で行われる畳み込みである。第1の畳み込みとその直後の合成は、例えば48チャンネル分行われる。後述する第2の畳み込みは、第1の畳み込みの畳み込み層の後の複数の畳み込み層で行われる畳み込みである。これらの畳み込みは、学習モデルにより実行される処理の一部である。

【0039】

第1の畳み込みにおけるフィルタは、周波数軸方向の幅よりも、時間軸方向の長さが長いフィルタが利用される。例えば、 1×100 のサイズのフィルタが用いられる。フィルタは、他のサイズであってもよく、例えば、時間軸における幅が周波数軸における長さの数十倍～数百倍又はそれ以上であってもよい。フィルタの数も、任意の数であってもよい。例えば、スペクトログラムSGの成分数（例えばピン数）と同じ数のフィルタが用意される。

40

【0040】

2次元のスペクトログラムSGは、所定幅（例えば1ピン）の信号がデータ数をその所定幅で割った数（例えば、全周波数ピン数／1）だけ存在する所定幅の信号の集団とみなされる。例えば、スペクトログラムSGが 100×2000 の2次元データの場合、幅が1で長さが1000の1次元の信号が100個あるものとみなされる。別の言い方をすれば、スペクトログラムSGは、周波数方向に所定幅ずつスライスされる。図5では、個々

50

の1次元の信号を $s_{g1} \sim s_{g100}$ の符号で示す。

【0041】

第1畳み込み部102は、スペクトログラムSGに対し、所定幅（例えば1ビン）ごとに、所定幅で所定長（例えば100フレーム）のフィルタで第1の畳み込みを複数チャンネル分行う。即ち、スペクトログラムSGがスライスされる幅と、フィルタの幅と、は同じである。本実施形態では、所定長の幅ごとに、独立にフィルタが用意されている。第1畳み込み部102は、スペクトログラムSGに対し、所定長の幅ごとに、対応するフィルタで畳み込みを行う。

【0042】

図5に示すように、第1畳み込み部102は、1次元の信号 $s_{g1} \sim s_{g100}$ の各々に対し、1次元のフィルタを畳み込む。例えば、1列目の1次元信号は、1列目用の 1×100 のフィルタで第1の畳み込みが行われる。2列目の1次元信号は、2列目用の 1×100 のフィルタで第1の畳み込みが行われる。3列目以降も同様である。各列のフィルタは、独自の係数を有する。第1の畳み込みでは、時間軸方向の前後にそれぞれ50のパディングがあり、データサイズが維持される。特にパディングがなく、多少のデータサイズの縮小が許容されてもよい。後述する合成部103により、畳み込みの結果が合わせられて、 1×2000 のデータ $d1$ が得られる。

【0043】

なお、フィルタのストライド幅は、1である。フィルタは、1次元の信号（1周波数ビン）ごとに用意されるのではなく、複数の1次元の信号で共通であってもよい。例えば、全ての1次元の信号に共通の1つのフィルタが用意されていてもよい。

【0044】

[2-4. 合成部]

合成部103は、各チャンネルについて、所定幅ごとに行われた第1の畳み込みにより得られた全部の幅を所定幅で除算した数のデータを合わせて、1次元の第1特徴データ $D1$ を得る。図5の例であれば、1次元の信号 $s_{g1} \sim s_{g100}$ の各々が 1×100 のフィルタによって畳み込まれた個々の 1×2000 のデータは、第1の畳み込みの結果である。

【0045】

第1の畳み込みの結果を合わせるとは、個々の結果を1つのデータとしてまとめることである。別の言い方をすれば、第1の畳み込みの結果を合わせるとは、個々の 1×2000 のデータを、結合、合成、又は累積して同じサイズの1つのデータを得ることである。図5の例であれば、上記100個のデータ（ 1×2000 のサイズのデータ）を加算合成し、 1×2000 の第1特徴データ $D1$ を得ることは、第1の畳み込みの結果を合わせることに相当する。

【0046】

1次元の第1特徴データ $D1$ は、周波数軸又は時間軸におけるデータ数が1の特徴データである。例えば、周波数ビンごとに第1の畳み込みが行われ、時間軸のデータ数分の1次元データが得られる。

【0047】

特徴データとは、スペクトログラムSGが示す音信号の特徴を示すデータである。別の言い方をすれば、特徴データは、少なくとも1回の畳み込みによって得られたデータである。第1特徴データ $D1$ が 1×1000 のサイズである場合、第1特徴データ $D1$ は、 1000 個の特徴量を含む。なお、特徴データは、主に2次元データの場合には特徴マップと呼ばれることもある。第1特徴データ $D1$ は、周波数ビン間の特徴が1つにまとめられている。

【0048】

図4に示すように、第1の畳み込みと合成の結果として、サイズが 1×2000 の第1特徴データ $D1$ が48チャンネル分得られる。後述する第2畳み込み部104は、第1特徴データ $D1$ に1次元フィルタを畳み込み、48チャンネル分の第2特徴データ $D2-1$

10

20

30

40

50

(サイズは 1×2000)を得、プーリングを行って、48チャンネル分の 1×1000 の第2特徴データD2-2を得る。

【0049】

例えば、合成部103は、第1の畳み込みの結果の和を計算して、第1特徴データD1を得る。第1特徴データD1は、第1の畳み込みの結果の単純な和ではなく、所定の重み付けがなされた和であってもよい。第1特徴データD1は、第1の畳み込みの結果を、和以外の数式を含む計算式に代入して得てもよい。

【0050】

[2-5. 第2畳み込み部]

第2畳み込み部104は、第1特徴データD1に対し、少なくとも1回の第2の畳み込みを行って第1特徴データD1をエンコードし、スペクトログラムSGの特徴を示す1次元の第2特徴データD2を得る。第2特徴データD2として、第2の畳み込みの各層で得られたデータD2-1からデータD2-6までの何れを用いてもよい。何れか2以上の層で得られたデータから、第2特徴データD2を合成してもよい。第2の畳み込みは、第1の畳み込みよりも後に行われる畳み込みである。本実施形態では、第2の畳み込みにパディングがあり、データサイズが畳み込みの前後で維持されるものとする。特にパディングがなく、多少サイズが縮小してもよい。

【0051】

第1特徴データD1は1次元なので、第2の畳み込みは、1次元データに対する1次元の畳み込みとなる。例えば、第2畳み込み部104は、第1特徴データD1に対し、少なくとも1回の第2の畳み込みとプーリングを行って、第2特徴データD2（データD2-1からD2-6の何れか）を得る。プーリングは、第2の畳み込みのうちの所定の畳み込み層の直後に配置されたプーリング層によって行われるプーリングである。

【0052】

図4の例であれば、第2畳み込み部104は、48チャンネル分の 1×1000 の第1特徴データD1に対し、第1層目において、48チャンネルの第2の畳み込みを行って、48チャンネル分の 1×2000 のデータD2-1を得、プーリングによりデータD2-1のサイズを縮小し、48チャンネル分の 1×1000 のデータD2-2を得る。

【0053】

第2畳み込み部104は、データD2-2に対し、第2層における第2の畳み込みを行って、96チャンネル分の 1×1000 のデータD2-3を得る。第2畳み込み部104は、データD2-3に対し、第3層における第2の畳み込みを行って、96チャンネル分の 1×1000 のデータD2-4を得る。第2畳み込み部104は、プーリングによりデータD2-4のサイズを縮小し、96チャンネル分の 1×500 のデータD2-5を得る。第2畳み込み部104は、データD2-5に対し、第4層における第2の畳み込みを行って、192チャンネル分の 1×500 のデータD2-6を得る。

【0054】

本実施形態では、第2の畳み込みは1次元のフィルタで行われるので、第2畳み込み部104は、第1特徴データD1に対し、1次元のフィルタで少なくとも1回の第2の畳み込みとプーリングを行って、第2特徴データD2を得る。第2の畳み込みのフィルタは、任意のサイズのフィルタを利用可能である。本実施形態では、時間軸方向に長いフィルタ（周波数軸の幅よりも時間軸の幅の方が長いフィルタ）が利用される。例えば、 1×1000 のサイズのフィルタが用いられる。チャンネル数は、任意の数であってもよい。

【0055】

[2-6. 逆畳み込み部]

逆畳み込み部105は、第2特徴データD2に対し、少なくとも1回の逆畳み込みを行って、所定の音を分離するマスクMを得る。逆畳み込みは、畳み込みニューラルネットワークにおける逆畳み込み層で行われる処理である。逆畳み込み層は、エンコーダの畳み込み層と1対1に対応して存在するものとする。例えば、タD2-6が第2特徴データとして用いられる。図4における第1層の第2畳み込みからのスキップ接続や、第3層の第2

10

20

30

40

50

畳み込みからのスキップ接続を、第2特徴データと見做してもよい。

【0056】

図4に示すように、逆畳み込み部105は、192チャンネル分のデータD2-6に対し、第4層の第2畳み込みに対応する逆畳み込みを行って、192チャンネル分の 1×500 のデータD3-6を得る。逆畳み込み部105は、192チャンネル分のデータD3-6の算出過程の中で、同時に、アップサンプリングを行って、192チャンネル分の 1×1000 のデータD3-5を得る。アップサンプリングは、直前段の逆畳み込み時のストライドにより実現され、アンプーリングとも呼ばれる。

【0057】

逆畳み込み部105は、192チャンネル分のデータD3-5に対し、第3層の第2畳み込みに対応する逆畳み込みを行って、96チャンネル分の 1×1000 のデータD3-4を得る。逆畳み込み部105は、96チャンネル分のデータD3-4に対し、第2層の第2畳み込みに対応する逆畳み込みを行って、データD3-3を得る。逆畳み込み部105は、データD3-3の算出過程の中で、同時に、アップサンプリングを行って、96チャンネル分の 1×2000 のデータD3-2を得る。逆畳み込み部105は、96チャンネル分のデータD3-2に対し、第1層の第2畳み込みに対応する逆畳み込みを行って、48チャンネル分の 1×2000 のデータD3-1を得る。

【0058】

図6に示すように、逆畳み込み部105は、48チャンネル分のデータD3-1の各々に対し、1周波数ビンごとのフィルタ（サイズは、例えば 100×100 ）で $1D/2D$ 変換を兼ねた逆畳み込みを行い、データD4を得、さらに変換演算を行ってマスクMを得る。この変換演算は、全結合でもよいし、畳み込みでもよい。或いは、個々のデータごとの重み付けでもよい。マスクMは、分離すべき音を特定可能なデータである。マスクMは、音響信号処理用の時間変化するフィルタとも見做せる。

【0059】

例えば、データD4及びマスクMは、スペクトログラムSGと同じサイズのデータである。図6の例では、マスクMにおける各データの色によって、分離すべき音（透過すべき音）が表現される。

【0060】

例えば、マスクMのある時刻のあるビンが白なら、その時刻にそのビンの周波数の音は透過し、黒なら、そのビンの周波数の音は阻止（除去）される。分離すべき音は、先述した所定の音の成分である。分離すべきではない音は、先述した他の音である。なお、黒が分離すべき音を意味し、白が分離すべきではない音を意味してもよい。分離の度合いが色によって表現されてもよい。分離の度合いとは、分離すべき音である確率又は蓋然性である。例えば、マスクMが256段階である場合、ある時刻のあるビンが所定の音の成分である確率が50%であれば、その値は128といったような中間値で表現される。

【0061】

なお、少なくとも1回の逆畳み込みでは、各層の入力データに対し、対応する畳み込み層で得られたデータを付加して、逆畳み込みが行われてもよい。このデータの付加は、例えば、U-NetやRESNETなどで使われているスキップ接続を用いる。このスキップ接続には、concatenationとsummationの何れを用いてもよい。スキップ接続は、ある層の第2畳み込みの結果を、同じ層の逆畳み込みの入力に供給する。スキップ接続によれば、エンコーダのある層よりより下層の処理で失われる情報を、デコーダのその層で回復して用いることができる。図4の例であれば、第1層の第2畳み込みの出力D2-1が、第1層の逆畳み込みの入力にスキップ接続される。第3層の第2畳み込みの出力D2-4が、第3層の逆畳み込みの入力にスキップ接続される。第1の畳み込み及び合成（ $2D/1D$ 変換）の出力D1が、 $1D/2D$ 変換を兼ねた逆畳み込みの入力にスキップ接続される。

【0062】

[2-7. 分離部]

10

20

30

40

50

所定の音の分離が訓練された後であれば、分離部 106 は、スペクトログラム SG にマスク M を適用し、複数の音の中から所定の音を分離する。マスク M を適用するとは、マスク M を利用して音を分離することである。分離部 106 は、マスク M を利用して、スペクトログラム SG に示された複数の音の成分のうちの一部を、所定の音として分離する。例えば、分離部 106 は、スペクトログラム SG に対し、マスク M を乗算することによって、複数の音の中から所定の音を分離する。例えば、分離された音は、スペクトログラム PS として表現される。

【0063】

分離部 106 によって得られたスペクトログラム PS は、音信号に変換され、データ記憶部 100 に記録される。

【0064】

[2-8. 調整部]

調整部 107 は、機械学習の手法により第 1 の量み込み、第 2 の量み込み、及び逆量み込みに用いられる変数を調整する。これらの変数は、訓練データのスペクトログラム SG から、本実施形態で説明する処理方法により訓練データの特定の音が分離されるように、繰り返し調整して決定された変数である。調整部 107 は、訓練データに含まれる入力と出力の関係が得られるように、学習前の学習モデルの変数を調整する。例えば、調整部 107 の処理の詳細は、後述する図 7 の処理である。

【0065】

[3. 処理装置が実行する処理]

本実施形態では、処理装置 10 が実行する処理の一例として、学習モデルの変数を調整するための調整処理と、混合信号から所定の音信号を分離するための分離処理と、を説明する。調整処理と分離処理の各々は、CPU 11 が不揮発メモリ 12 に記憶されたプログラムに従って動作することによって実行される。調整処理と分離処理の各々は、図 2 に示す機能ブロックにより実行される処理の一例である。

【0066】

[3-1. 調整処理]

図 7 は、調整処理の一例を示すフロー図である。1 ないし複数のペアを用いた、この調整処理（訓練）が、学習モデルの損失が所定の基準をクリアするまで繰り返し行われる。図 7 に示すように、CPU 11 は、不揮発メモリ 12 に記憶された訓練データのデータセットから、混合信号のスペクトログラムと、ソロ信号のスペクトログラムと、のペアを取得する（S100）。不揮発メモリ 12 に複数のペアが記憶されている場合には、CPU 11 は、これら複数のペアを順次取得する。

【0067】

CPU 11 は、S100 で取得したペアに含まれる混合信号のスペクトログラムを、現状の学習モデル（変数を調整する前の学習モデル）に入力して、マスク M を推定する（S101）。混合信号のスペクトログラムが学習モデルに入力されると、図 4 を参照して説明した一連の処理（後述する分離処理と同様の処理）が実行される。学習モデルは、第 1 の量み込みを行って、混合信号のスペクトログラムの第 1 特徴データ D1 を得る。学習モデルは、第 1 特徴データ D1 に対し、少なくとも 1 回の第 2 の量み込みを行って、混合信号のスペクトログラムの第 2 特徴データ D2 を得る。学習モデルは、第 2 特徴データ D2 に対し、少なくとも 1 回の逆量み込みを行って、マスク M を推定する。

【0068】

CPU 11 は、マスク M を混合信号のスペクトログラムに適用して、分離信号のスペクトログラムを得る（S102）。S102 において得られる分離信号のスペクトログラムは、現状の学習モデルによって得られるスペクトログラムである。このスペクトログラムは、続く S103 の処理において、現状の学習モデルの性能を評価するために用いられる。

【0069】

CPU 11 は、分離信号のスペクトログラムと、ソロ信号のスペクトログラムと、を比較して、学習モデルの損失を得る（S103）。損失としては、非特許文献 1 と同じよう

10

20

30

40

50

にL1ノルムを用いてもよいし、その他のL2ノルムなどを用いてもよい。損失は、学習モデルの性能の指標となる情報である。別の言い方をすれば、損失は、分離信号のスペクトログラムと、ソロ信号のスペクトログラムと、の差異に相当する情報である。損失が大きいほど、現状の学習モデルの性能が低く変数を大幅に変更する必要がある。

【0070】

CPU11は、S103で得られた損失に基づいて、学習モデルの変数を調整する(S104)。変数の調整自体は、一般的な誤差逆伝搬で行えばよい。以降、損失が十分小さくなるまで、S100~S104の処理が繰り返され、学習モデルの訓練が完了する。

【0071】

【3-2. 分離処理】

図8は、分離処理の一例を示すフロー図である。図8に示すように、CPU11は、不揮発メモリ12に記憶された混合信号のスペクトログラムSGを取得する(S200)。S200において取得されるスペクトログラムSGは、音分離の対象となるスペクトログラムSGである。

【0072】

CPU11は、混合信号のスペクトログラムSGに対し、1周波数ビンの幅ごとに第1の畳み込みを行う(S201)。S201においては、CPU11は、混合信号のスペクトログラムSG(例えば100×2000)を、1周波数ビンの幅ごとの1次元の信号(例えば1×2000×100)とみなし、各周波数ビンに対応するフィルタ(例えば1×100×100×48)で第1の畳み込みを行う。

【0073】

CPU11は、S201で行われた第1の畳み込みの結果100個の和を計算して、1次元の第1特徴データD1(例えば1×2000×48)を得る(S202)。図4の例であれば、S202の処理により、第1特徴データD1が得られる。

【0074】

CPU11は、第1特徴データD1に対し、1次元のフィルタで少なくとも1回の第2の畳み込みと必要に応じてプーリングを行って、第2特徴データD2(サイズは様々)を得る(S203)。図4の例であれば、S203の処理により、データD2-1からD2-6が得られ、ここでは、データD2-6が第2特徴データD2として用いられる。S201からS203までの処理が、エンコード処理である。

【0075】

CPU11は、第2特徴データD2に対し、少なくとも1回の逆畳み込みを含むデコード処理を行って、マスクMを得る(S204)。図4の例であれば、S204の処理により、データD3-6からD3-1と、データD4と、マスクMと、が得られる。

【0076】

CPU11は、混合信号のスペクトログラムSGにマスクMを適用し、複数の音の中から所定の音を分離する(S205)。S205においては、CPU11は、混合信号のスペクトログラムSGに対し、マスクMを乗算することによって、複数の音の中から所定の音を分離する。CPU11は、分離された音のスペクトログラムPSを、逆短時間フーリエ変換等を用いて、周波数領域から時間領域へ変換し、分離された所定の音信号のデジタルデータを得る。このデジタルデータは、不揮発メモリ12に記録される。

【0077】

CPU11は、スピーカ17から、分離された所定の音を出力し(S206)、本処理は終了する。S206においては、CPU11は、S205において記録されたデジタルデータを再生し、分離された所定の音を出力する。

【0078】

本実施形態の処理装置10は、所定幅ごとに行われた第1の畳み込みの結果を合わせて、1次元の第1特徴データD1を得ることによって、音信号のスペクトログラムSGの特徴を効率良く表現する特徴データを得ることができる。例えば、周波数方向に広範囲に特徴的な情報を有する音(時間軸方向の特徴が局所的な音)の場合には、時間軸における所

10

20

30

40

50

定幅ごとに第1の畳み込みを行うことで、周波数方向に広範囲な情報を表す、周波数方向の1次元データ（例えば 100×1 ）が得られる。例えば、時間方向に広範囲に特徴的な情報を有する音（周波数方向の特徴が局所的な音）の場合には、周波数軸における所定幅ごとに第1の畳み込みを行うことで、時間方向に広範囲な情報を表す、時間軸方向の1次元データ（例えば 1×2000 ）が得られる。処理装置10によれば、エンコード処理のうち、第1特徴データD1を得た以降の処理は、全て1次元データが対象の処理なので、効率良く特徴データを得ることができる。その結果、特徴データを得る処理を高速化できる。処理装置10の処理負荷も軽減できる。時間軸方向の1次元データを用いる場合、同じデータ量及び演算量であれば、時間方向により長いフィルタを実現でき、その点でも効率的に時間方向の情報を加味できる。波形のスペクトル時系列をある軸方向の1次元データに変換して推論を行い、他方の軸方向の成分間で変数が融通されるので、同じ規模の学習モデルにより効率的に推論を行うことができる。

10

【0079】

処理装置10は、第1の畳み込みの結果を合わせて、第1特徴データD1を得る。処理装置10は、第1特徴データD1に対し、少なくとも1回の第2の畳み込みとプーリングを行って、第2特徴データD2を得る。プーリングにより特徴データのサイズが縮小され、より効率良く特徴データを得ることができる。

【0080】

処理装置10では、少なくとも1回の逆畳み込みでは、各層の入力データに対し、対応する畳み込み層で得られたデータを付加して、逆畳み込みが行われるので、逆畳み込みの精度が向上する。マスクMの精度が高まり、音分離の精度も高めることができる。

20

【0081】

〔4．変形例〕

なお、本発明は、以上に説明した実施形態に限定されるものではない。本発明の趣旨を逸脱しない範囲で、適宜変更可能である。

【0082】

例えば、畳み込みの後にプーリングが実行される場合を説明したが、特にプーリングを実行せずにデータサイズを縮小しなくてもよい。1次元のフィルタを利用した第1の畳み込みが実行される場合を説明したが、第1特徴データD1が1次元になればよく、第1の畳み込みは2次元のフィルタが利用されてもよい。

30

【0083】

実施形態では、処理装置10を音声分離に利用する場合を説明したが、処理装置10は、他の任意の場面に利用可能である。例えば、処理装置10を声紋鑑定に利用してもよい。ある特定の人間の声であるか否かを鑑定する声紋鑑定であれば、人間の声を示す音信号のスペクトログラムSGと、この人間であるか否かを示す情報（正例であるか負例であるかを示す情報）と、を含む訓練データに基づいて、学習モデルの変数が調整される。処理装置10は、声紋鑑定の対象となるスペクトログラムSGを学習モデルに入力する。学習モデルは、実施形態で説明したような第1の畳み込みと第2の畳み込みを行って、1次元の第2特徴データD2を得る。学習モデルは、第2特徴データD2に応じた情報を出力する。この情報は、学習済みの人間の声であるか否かを示す。声紋鑑定の場合、逆畳み込みは行われない。

40

【0084】

複数の人間の中から発声者を特定する声紋鑑定であれば、人間の声を示す音信号のスペクトログラムSGと、この人間を識別する識別情報（例えば、人間を一意に識別するラベルID）と、を含む訓練データに基づいて、学習モデルの変数が調整される。処理装置10は、声紋鑑定の対象となるスペクトログラムSGを学習モデルに入力する。学習モデルは、実施形態で説明したような第1の畳み込みと第2の畳み込みを行って、1次元の第2特徴データD2を得る。学習モデルは、第2特徴データD2に応じたラベルIDを出力する。音声分離及び声紋鑑定以外にも、楽曲のジャンル推定又は音信号におけるノイズ除去といった任意の場面に処理装置10を利用可能である。

50

【符号の説明】

【0085】

10 処理装置、11 CPU、12 不揮発メモリ、13 RAM、14 操作部、15 表示部、16 入力部、17 スピーカ、100 データ記憶部、101 第1取得部、102 第1畳み込み部、103 合成部、104 第2畳み込み部、105 逆畳み込み部、106 分離部、107 調整部。

10

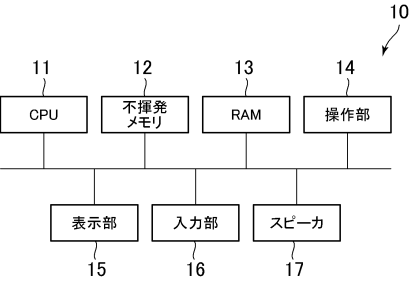
20

30

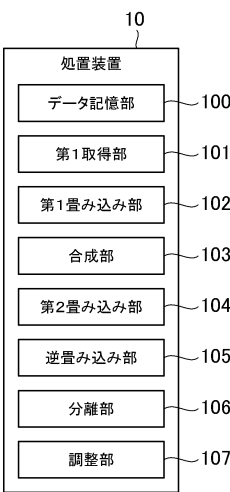
40

50

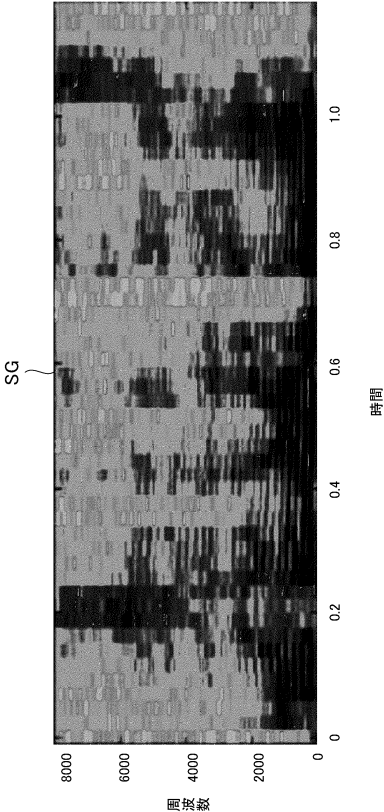
【図面】
【図 1】



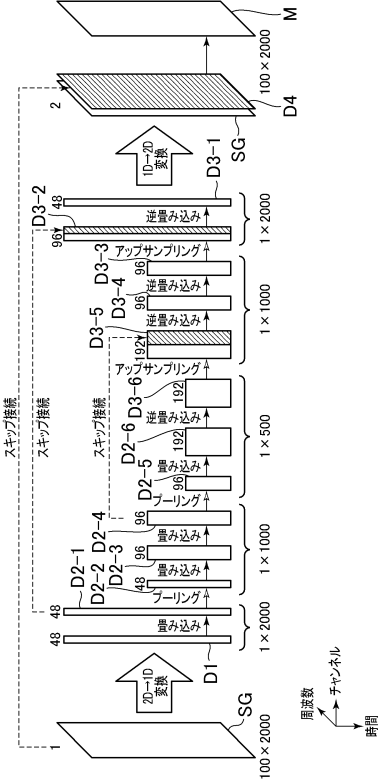
【図 2】



【図 3】



【図 4】



10

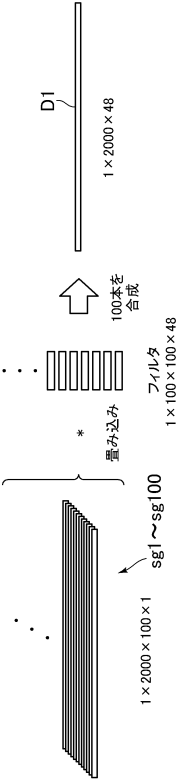
20

30

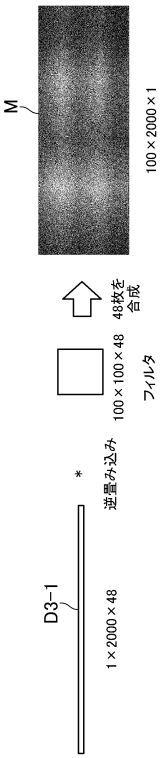
40

50

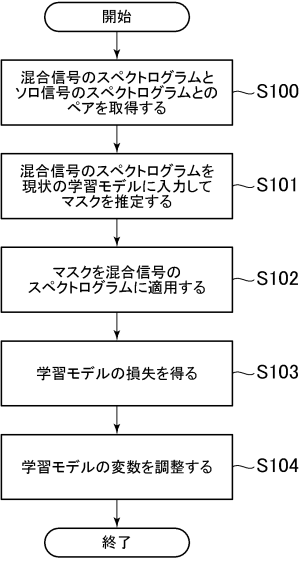
【図 5】



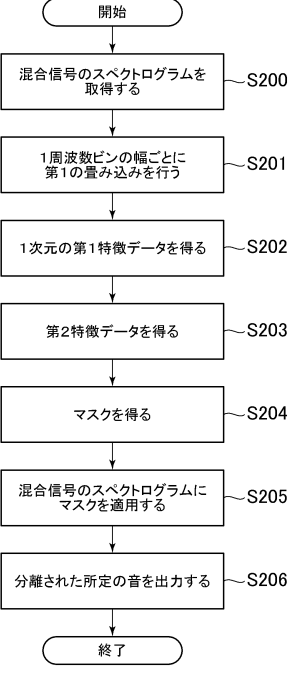
【図 6】



【図 7】



【図 8】



10

20

30

40

50

フロントページの続き

- (56)参考文献 米国特許出願公開第2019/0355347 (US, A1)
特開2019-191558 (JP, A)
特開2019-144511 (JP, A)
米国特許出願公開第2017/0140260 (US, A1)
- (58)調査した分野 (Int.Cl., DB名)
G10L 13/00-25/93