(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2020-3536
(P2020-3536A)
(43)公開日　令和2年1月9日(2020.1.9)

(51)Int.Cl.		F I			テーマコード（参考）
G10G 3/04	(2006.01)	G10G	3/04		5D182
G10H 1/00	(2006.01)	G10H	1/00	B	5D478
G10L 25/51	(2013.01)	G10H	1/00	Z	
G10L 25/30	(2013.01)	G10L	25/51	300	
G06N 20/00	(2019.01)	G10L	25/30		

審査請求 未請求　請求項の数 13　ＯＬ　（全 17 頁）　最終頁に続く

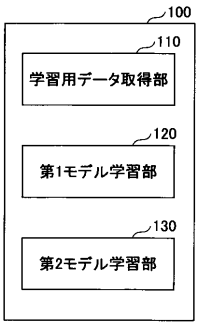
(21)出願番号　特願2018-120235 (P2018-120235)
(22)出願日　平成30年6月25日 (2018.6.25)

(71)出願人　000001443
カシオ計算機株式会社
東京都渋谷区本町１丁目６番２号
(74)代理人　100107766
弁理士　伊東　忠重
(74)代理人　100070150
弁理士　伊東　忠彦
(72)発明者　日暮　大輝
東京都羽村市栄町３丁目２番１号　カシオ計算機株式会社　羽村技術センター内

Ｆターム(参考)　5D182 AC01　AD01
5D478 DF02

(54)【発明の名称】学習装置、自動採譜装置、学習方法、自動採譜方法及びプログラム

(57)【要約】　（修正有）
【課題】各音の音高や区間が明確でないオーディオデータから楽譜を自動生成するための音響処理技術を提供する。
【解決手段】学習装置１００は、単音音源と音高情報とを第１の機械学習モデルの学習用データとして取得し、採譜対象の音源と楽譜情報とを第２の機械学習モデルの学習用データとして取得し、単音音源と前記採譜対象の音源とに対して前処理を実行し、それぞれのスペクトログラムを取得する学習用データ取得部と、単音音源のスペクトログラムを学習用入力データとして入力し、単音音源の音高の予測確率を出力するよう音高情報によって学習する第１モデル学習部と、採譜対象の音源のスペクトログラムを学習済みの第１の機械学習モデルに入力することによって生成される特徴マップを学習用入力データとして入力し、特徴マップの固定長の区間に音符が存在する予測確率を出力するよう楽譜情報によって学習する第２モデル学習部と、を有する。
【選択図】図２



【特許請求の範囲】
【請求項１】
単音音源と音高情報とを第１の機械学習モデルの学習用データとして取得し、採譜対象の音源と楽譜情報とを第２の機械学習モデルの学習用データとして取得し、前記単音音源と前記採譜対象の音源とに対して前処理を実行し、それぞれのスペクトログラムを取得する学習用データ取得部と、
前記単音音源のスペクトログラムを学習用入力データとして入力し、前記単音音源の音高の予測確率を出力するよう前記音高情報によって第１の機械学習モデルを学習する第１モデル学習部と、
前記採譜対象の音源のスペクトログラムを学習済みの前記第１の機械学習モデルに入力することによって生成される特徴マップを学習用入力データとして入力し、前記特徴マップの固定長の区間に音符が存在する予測確率を出力するよう前記楽譜情報によって第２の機械学習モデルを学習する第２モデル学習部と、
を有する学習装置。
【請求項２】
前記第１の機械学習モデルと前記第２の機械学習モデルとは、畳み込みニューラルネットワークにより構成される、請求項１記載の学習装置。
【請求項３】
前記第２モデル学習部は、前記第１の機械学習モデルにより生成される異なる時間解像度を有する複数の特徴マップを前記第２の機械学習モデルに入力する、請求項２記載の学習装置。
【請求項４】
前記第２モデル学習部は、前記第１の機械学習モデルと前記第２の機械学習モデルとをＳＳＤ（Ｓｉｎｇｌｅ　Ｓｈｏｔ　Ｄｅｔｅｃｔｉｏｎ）として実現する、請求項１乃至３何れか一項記載の学習装置。
【請求項５】
前記第１モデル学習部は、複数種別のオーディオ成分のそれぞれに対して前記第１の機械学習モデルを学習し、
前記第２モデル学習部は、複数種別のオーディオ成分を含む採譜対象の音源に対して各オーディオ成分種別毎に音符が存在する予測確率を出力するよう前記第２の機械学習モデルを学習する、請求項１乃至４何れか一項記載の学習装置。
【請求項６】
単音音源から音高の予測確率を出力する第１の学習済み機械学習モデルと、特徴マップから前記特徴マップの固定長の区間に音符が存在する予測確率を出力する第２の学習済み機械学習モデルとを利用し、採譜対象の音源を前記第１の学習済み機械学習モデルに入力し、前記第１の学習済み機械学習モデルによって生成された特徴マップを前記第２の学習済み機械学習モデルに入力し、前記特徴マップの固定長の区間に音符が存在する予測確率を出力するモデル処理部と、
前記音符が存在する予測確率に基づき楽譜情報を生成する楽譜生成部と、
を有する自動採譜装置。
【請求項７】
前記モデル処理部は、前記採譜対象の音源に対して前処理を実行することによってスペクトログラムを取得し、前記スペクトログラムを前記第１の学習済み機械学習モデルに入力する、請求項６記載の自動採譜装置。
【請求項８】
前記モデル処理部は、前記特徴マップ上の各点について前記第２の学習済み機械学習モデルから出力された最大の予測確率を有する音符を予測音符として決定する、請求項６又は７記載の自動採譜装置。
【請求項９】
前記楽譜生成部は、ＮＭＳ（Ｎｏｎ－Ｍａｘｉｍｕｍ　Ｓｕｐｐｒｅｓｓｉｏｎ）に従

って抽出された予測音符に基づき楽譜情報を生成する、請求項８記載の自動採譜装置。
【請求項１０】
プロセッサが、単音音源と音高情報とを第１の機械学習モデルの学習用データとして取得し、採譜対象の音源と楽譜情報とを第２の機械学習モデルの学習用データとして取得し、前記単音音源と前記採譜対象の音源とに対して前処理を実行し、それぞれのスペクトログラムを取得するステップと、
前記プロセッサが、前記単音音源のスペクトログラムを学習用入力データとして入力し、前記単音音源の音高の予測確率を出力するよう前記音高情報によって第１の機械学習モデルを学習するステップと、
前記プロセッサが、前記採譜対象の音源のスペクトログラムを学習済みの前記第１の機械学習モデルに入力することによって生成される特徴マップを学習用入力データとして入力し、前記特徴マップの固定長の区間に音符が存在する予測確率を出力するよう前記楽譜情報によって第２の機械学習モデルを学習するステップと、
を有する学習方法。
【請求項１１】
プロセッサが、単音音源から音高の予測確率を出力する第１の学習済み機械学習モデルに採譜対象の音源を入力するステップと、
前記プロセッサが、特徴マップから前記特徴マップの固定長の区間に音符が存在する予測確率を出力する第２の学習済み機械学習モデルに前記第１の学習済み機械学習モデルによって生成された特徴マップを入力するステップと、
前記プロセッサが、前記第２の学習済み機械学習モデルから出力された前記音符が存在する予測確率に基づき楽譜情報を生成するステップと、
を有する自動採譜方法。
【請求項１２】
単音音源と音高情報とを第１の機械学習モデルの学習用データとして取得し、採譜対象の音源と楽譜情報とを第２の機械学習モデルの学習用データとして取得し、前記単音音源と前記採譜対象の音源とに対して前処理を実行し、それぞれのスペクトログラムを取得するステップと、
前記単音音源のスペクトログラムを学習用入力データとして入力し、前記単音音源の音高の予測確率を出力するよう前記音高情報によって第１の機械学習モデルを学習するステップと、
前記採譜対象の音源のスペクトログラムを学習済みの前記第１の機械学習モデルに入力することによって生成される特徴マップを学習用入力データとして入力し、前記特徴マップの固定長の区間に音符が存在する予測確率を出力するよう前記楽譜情報によって第２の機械学習モデルを学習するステップと、
をプロセッサに実行させるプログラム。
【請求項１３】
単音音源から音高の予測確率を出力する第１の学習済み機械学習モデルに採譜対象の音源を入力するステップと、
特徴マップから前記特徴マップの固定長の区間に音符が存在する予測確率を出力する第２の学習済み機械学習モデルに前記第１の学習済み機械学習モデルによって生成された特徴マップを入力するステップと、
前記第２の学習済み機械学習モデルから出力された前記音符が存在する予測確率に基づき楽譜情報を生成するステップと、
をプロセッサに実行させるプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、音響処理技術に関する。
【背景技術】

【０００２】
オーディオデータから楽譜を自動生成する自動採譜技術が従来から知られている。例えば、特開２００７－０３３４７９には、同時に複数の音が演奏される場合でも単一楽器により演奏された音響信号から楽譜を自動採譜する技術が記載されている。
【先行技術文献】
【特許文献】
【０００３】
【特許文献１】特開２００７－０３３４７９
【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、従来の自動採譜では、楽譜に対して正確に演奏又は歌唱され、各音の音高や区間が明確なオーディオデータの場合には比較的高精度な採譜が可能であるが、例えば、各音の音高や区間が明確でないオーディオデータの場合には期待するような自動採譜が困難であった。
【０００５】
上記問題点を鑑み、本開示の課題は、様々なオーディオデータからより効果的に楽譜を自動生成するための音響処理技術を提供することである。
【課題を解決するための手段】
【０００６】
上記課題を解決するため、本開示の一態様は、単音音源と音高情報とを第１の機械学習モデルの学習用データとして取得し、採譜対象の音源と楽譜情報とを第２の機械学習モデルの学習用データとして取得し、前記単音音源と前記採譜対象の音源とに対して前処理を実行し、それぞれのスペクトログラムを取得する学習用データ取得部と、前記単音音源のスペクトログラムを学習用入力データとして入力し、前記単音音源の音高の予測確率を出力するよう前記音高情報によって第１の機械学習モデルを学習する第１モデル学習部と、前記採譜対象の音源のスペクトログラムを学習済みの前記第１の機械学習モデルに入力することによって生成される特徴マップを学習用入力データとして入力し、前記特徴マップの固定長の区間に音符が存在する予測確率を出力するよう前記楽譜情報によって第２の機械学習モデルを学習する第２モデル学習部と、を有する学習装置に関する。
【発明の効果】
【０００７】
本開示によると、各音の音高や区間が明確でないオーディオデータから楽譜を自動生成するための音響処理技術を提供することができる。
【図面の簡単な説明】
【０００８】
【図１】本開示の一実施例による学習済み機械学習モデルを有する自動採譜装置を示す概略図である。
【図２】本開示の一実施例による学習装置の機能構成を示すブロック図である。
【図３】本開示の一実施例による特徴マップ生成モデルの構成を示す概略図である。
【図４】本開示の一実施例による音符存在確率予測モデルの構成を示す概略図である。
【図５】本開示の一実施例による特徴マップとデフォルトボックスとの関係を示す概念図である。
【図６】本開示の一実施例による特徴マップ生成モデルの学習処理を示すフローチャートである。
【図７】本開示の一実施例による音符存在確率予測モデルの学習処理を示すフローチャートである。
【図８】本開示の一実施例による自動採譜装置の機能構成を示すブロック図である。
【図９】本開示の一実施例による自動採譜処理を示すフローチャートである。
【図１０】本開示の一実施例による学習装置及び自動採譜装置のハードウェア構成を示す

ブロック図である。
【発明を実施するための形態】
【０００９】
以下の実施例では、機械学習モデルによって音源（音の波形データであるオーディオデータ）から楽譜情報を生成する自動採譜装置が開示される。
【００１０】
従来の自動採譜技術では、音高の予測に注力され、音符の切れ目を示すオンセットとオフセットとの予測は自動採譜における課題の１つであった。本開示による自動採譜装置は、音源におけるオンセットとオフセットとを機械学習モデルの１つであるＳＳＤ（Ｓｉｎｇｌｅ　Ｓｈｏｔ　Ｄｅｔｅｃｔｉｏｎ）によって予測する。
【００１１】
ＳＳＤは、１つのニューラルネットワークを用いて入力画像における物体を検出する手法である。すなわち、当該ニューラルネットワークへの入力は画像であり、その出力は複数の矩形領域（ＳＳＤでは、デフォルトボックスと呼ばれる）の中心座標、高さ、幅及び物体の種類の予測確率である。デフォルトボックスは入力画像のサイズによって予め設定された個数の候補として用意され、後処理（ＮＭＳ：Ｎｏｎ－Ｍａｘｉｍｕｍ　Ｓｕｐｐｒｅｓｓｉｏｎなど）によって大部分のデフォルトボックスを候補から外し、残ったデフォルトボックスを検出結果とするというものである。
【００１２】
本開示による自動採譜装置におけるニューラルネットワークへの入力は、採譜対象の楽音の波形データ又はスペクトログラムであり、その出力は楽音のオンセット、オフセット及び音高であり、自動採譜装置は、ＳＳＤにおける中心座標及び幅に対応してオンセット及びオフセット（すなわち、楽音の形状又は長さ）を特定し、ＳＳＤにおける物体の種類に対応して音高を特定する。
【００１３】
後述される実施例を概略すると、自動採譜装置は２つの学習済み機械学習モデル（畳み込みニューラルネットワークなど）を利用し、一方のモデルは単音音源から音高の予測確率を出力するものであり、他方のモデルは特徴マップから当該特徴マップの固定長の区間に音符が存在する予測確率を出力するものである。自動採譜装置は、採譜対象の音源を前者の学習済み機械学習モデル（特徴マップ生成モデル）に入力し、当該学習済み特徴マップ生成モデルの畳み込み層から生成された各特徴マップを後者の学習済み機械学習モデル（音符存在確率予測モデル）に入力し、各特徴マップの各点に対して当該学習済み音符存在確率予測モデルから出力された固定長の区間又はデフォルトボックスにおける各音高の音符の予測存在確率に基づき楽譜情報を生成する。
【００１４】
学習済み特徴マップ生成モデルによって生成される特徴マップは、畳み込みの結果として異なる時間解像度を有し、固定長の区間又はデフォルトボックスは異なる時間的長さとなる。このため、音符存在確率予測モデルにより各特徴マップに対して固定長の区間と同じ長さの音符を検出することによって、異なる長さの音符のオンセット及びオフセットを特定することが可能になる。
【００１５】
まず、図１を参照して、本開示の一実施例による自動採譜装置を説明する。図１は、本開示の一実施例による学習済み機械学習モデルを有する自動採譜装置を示す概略図である。
【００１６】
図１に示されるように、本開示の一実施例による自動採譜装置２００は、限定することなく、畳み込みニューラルネットワークなどの何れかのタイプのニューラルネットワークとして実現される２種類の学習済みモデルを有し、学習用データストレージ５０を用いて学習装置１００によって学習された機械学習モデルを利用して、採譜対象の音源から楽譜情報を生成する。

【００１７】
次に、図２～７を参照して、本開示の一実施例による学習装置を説明する。学習装置１００は、学習用データストレージ５０における学習用データを利用して、自動採譜装置２００に利用される特徴マップ生成モデルと音符存在確率予測モデルとを学習する。図２は、本開示の一実施例による学習装置の機能構成を示すブロック図である。
【００１８】
図２に示されるように、学習装置１００は、学習用データ取得部１１０、第１モデル学習部１２０及び第２モデル学習部１３０を有する。
【００１９】
学習用データ取得部１１０は、単音音源と音高情報とを特徴マップ生成モデルの学習用データとして取得し、採譜対象の音源と楽譜情報とを音符存在確率予測モデルの学習用データとして取得し、単音音源と採譜対象の音源とに対して前処理を実行し、それぞれのスペクトログラムを取得する。
【００２０】
具体的には、学習用データ取得部１１０は、学習用データストレージ５０から、特徴マップ生成モデルを学習するための単音又はシングルノート音源（例えば、「ド」から「シ」までの１２種類の音源など）の波形データと音高情報（「ド」から「シ」までの音高など）とのペアを取得し、取得した単音音源の波形データに対して前処理（例えば、短時間フーリエ変換など）を実行することによって、各単音音源のスペクトログラムと音高情報との学習用データセットを生成する。
【００２１】
また、学習用データ取得部１１０は、学習用データストレージ５０から、音符存在確率予測モデルを学習するための単旋律音源（歌唱音源など）の波形データと楽譜情報（音高の時系列変化など）とのペアを取得し、取得したモノフォニック音源の波形データに対して前処理（例えば、短時間フーリエ変換など）を実行することによって、モノフォニック音源のスペクトログラムと楽譜情報との学習用データセットを生成する。ここで、楽譜情報は、例えば、ＭＩＤＩ（Ｍｕｓｉｃａｌ Ｉｎｓｔｒｕｍｅｎｔ Ｄｉｇｉｔａｌ Ｉｎｔｅｒｆａｃｅ）規格に従うものであってもよい。
【００２２】
典型的には、スペクトログラムは、時間軸及び周波数軸における信号成分の強度を表し、波形データを短時間フーリエ変換することによって生成される。短時間フーリエ変換には各種パラメータが設定される必要があるが、例えば、ＦＦＴ窓幅：１０２４、サンプリング周波数：１６ｋＨｚ、オーバラップ幅：７６８、窓関数：ハニング窓、及びフィルタバンク：メルフィルタバンク（１２８バンド）などに従って、短時間フーリエ変換が実行されてもよい。スペクトログラムに変換した後、時間軸方向に一定のサンプル数（例えば、１０２４サンプル）だけ抽出されてもよい。また、本実施例によるスペクトログラムは、低周波数成分を精細にするよう周波数軸が対数変換されたものであってもよい。
【００２３】
第１モデル学習部１２０は、単音音源のスペクトログラムを学習用入力データとして入力し、単音音源の音高の予測確率を出力するよう音高情報によって特徴マップ生成モデルを学習する。
【００２４】
例えば、特徴マップ生成モデルは、図３に示されるように、複数の畳み込み層を含む畳み込みニューラルネットワークにより構成され、入力された単音音源のスペクトログラムを音高の予測確率に変換するＳＳＤとして実現される。ここで、音高は連続値でなく離散値として表現され、ｏｎｅ－ｈｏｔベクトルとして表現されてもよい。なお、打楽器などの噪音音源も学習対象とする場合、噪音音源の単音又はシングルノートの音声をデータセットに含めてもよい。その場合、音高クラスとして噪音を表現するクラスを設定し、それを教師ラベルとしてもよい。
【００２５】

第１モデル学習部１２０は、学習用入力データの単音音源のスペクトログラムを特徴マップ生成モデルに入力し、特徴マップ生成モデルからの出力と学習用出力データの音高情報との誤差が小さくなるように、バックプロパゲーションによって特徴マップ生成モデルのパラメータを更新する。ここで、誤差を示す損失関数として、限定することなく、特徴マップ生成モデルの出力と学習用出力データの音高との交差エントロピーが利用されてもよい。

【００２６】

例えば、所定数の学習用データに対して更新処理が終了した、誤差が所定の閾値以下に収束した、誤差の改善が所定の閾値以下に収束したなどの所定の学習終了条件が充足されると、第１モデル学習部１２０は、更新された特徴マップ生成モデルを学習済み機械学習モデルとして設定する。

【００２７】

第２モデル学習部１３０は、採譜対象の音源のスペクトログラムを学習済みの特徴マップ生成モデルに入力することによって生成される特徴マップを学習用入力データとして入力し、特徴マップの固定長の区間に音符が存在する予測確率を出力するよう楽譜情報によって音符存在確率予測モデルを学習する。

【００２８】

例えば、音符存在確率予測モデルは、図４に示されるように、複数の畳み込み層を含む畳み込みニューラルネットワークにより構成され、モノフォニック音源のスペクトルグラムを学習済み特徴マップ生成モデルに入力することによって生成された特徴マップを当該特徴マップの各点を始点とする固定長の区間と同じ長さの音符が存在する予測確率に変換するＳＳＤとして実現される。例えば、ドからシの１２音で採譜する場合、特徴マップ上の各点は、ドからシの各音高及び休符（無音）の１３通りの音符又は音高クラスが存在する予測確率を有する。

【００２９】

上述したように、学習済み特徴マップ生成モデルは複数の畳み込み層を含み、各畳み込み層からモノフォニック音源のスペクトログラムの特徴マップが生成される。生成される特徴マップは、図３に示されるような畳み込み層のレベルに応じて時間解像度が異なる特徴マップとなる。典型的には、図５に示されるように、入力層に相対的に近い畳み込み層では、時間解像度が相対的に高い（図示された例では、３２Ｈｚ）特徴マップが生成され、出力層に相対的に近い畳み込み層では、時間解像度が相対的に低い（図示された例では、１６Ｈｚ）特徴マップが生成される。図示されるような固定長の区間又はデフォルトボックスが設定されると、時間解像度が相対的に高い特徴マップにおける区間は、時間解像度が相対的に低い特徴マップにおける区間より短い時間を占有する。このため、異なる時間的長さを有する音符の存在予測確率を導出することができ、音符の時間的長さを特定することが可能になる。

【００３０】

第２モデル学習部１３０は、学習用入力データの音源のスペクトログラムを学習済み特徴マップ生成モデルに入力し、学習済み特徴マップ生成モデルによって生成された各特徴マップを音符存在確率予測モデルに入力し、音符存在確率予測モデルからの出力と学習用出力データの楽譜情報との誤差が小さくなるように、バックプロパゲーションによって音符存在確率予測モデルのパラメータを更新する。

【００３１】

ここで、誤差を示す損失関数として、限定することなく、音符存在確率予測モデルの出力と音高の時系列変化とから算出されるタイミング誤差と信頼誤差との加重和が利用されてもよい。音高の時系列変化は、楽曲のスタートタイミング、エンドタイミング及び音高のセットが複数集まることによって表現され、楽譜情報から導出される。当該セットは発音と呼ばれてもよく、例えば、音高の時系列変化は、発音＃１："０：００～０：０２，Ａ（ラ）３"、発音＃２："０：０３～０：０５，Ｂ（シ）３"、発音＃３："０：０５～０：０８，Ｃ（ド）４"・・・などにより表現されてもよい。図５に示されるデフォルトボ

ックスは、１つの発音を表現しており、複数のチャネルを有する。デフォルトボックスの各チャネルの最初のサンプルはそれぞれ、当該デフォルトボックスの発音のオンセットの予測値、オフセットの予測値及び音高クラスの予測確率を有する。すなわち、トータルで２＋（音高のクラス数）のチャネルがある。

【００３２】

第２モデル学習部１３０は、各発音について、オンセットとオフセットとの和が最小となるデフォルトボックスを探索し、検出されたデフォルトボックスと発音とに対してタイミング誤差と信頼誤差を求める。ここで、タイミング誤差とは、予測したオンセットを考慮したスタートタイミングのずれと、予測したオフセットを考慮したエンドタイミングのずれとの和としてもよい。ただし、差分の表現として、デフォルトボックスの長さを基準にした相対値が利用されてもよい。また、信頼誤差は、発音の音高と予測した音高とから算出される交差エントロピーであってもよい。なお、無音を表すクラスも教師ラベルとして用意されてもよく、この場合、発音のない区間を予測することができる。

【００３３】

第２モデル学習部１３０は、ＮＭＳ（Ｎｏｎ－Ｍａｘｉｍｕｍ　Ｓｕｐｐｒｅｓｓｉｏｎ）に従って各特徴マップの各点について設定されたデフォルトボックスを減らしていき、残ったデフォルトボックスを予測発音としてもよい。具体的には、第２モデル学習部１３０はまず、各デフォルトボックスについて音高クラス毎の音符存在予測確率を求める。その後、第２モデル学習部１３０は、予測確率が所定の閾値（例えば、０．９など）以下であるデフォルトボックスを削除してもよい。第２モデル学習部は、残ったデフォルトボックスのうち積集合／和集合に閾値を設けて、閾値以上のデフォルトボックスの一方を削除し、重複したデフォルトボックスを排除する。第２モデル学習部１３０は、最終的に残ったデフォルトボックスを予測発音とする。

【００３４】

例えば、所定数の学習用データに対して更新処理が終了した、誤差が所定の閾値以下に収束した、誤差の改善が所定の閾値以下に収束したなどの所定の学習終了条件が充足されると、第２モデル学習部１３０は、更新された音符存在確率予測モデルを学習済みモデルとして設定する。

【００３５】

一実施例では、第１モデル学習部１２０は、複数種別のオーディオ成分のそれぞれに対して特徴マップ生成モデルを学習し、第２モデル学習部１３０は、複数種別のオーディオ成分を含む採譜対象の音源に対して各オーディオ成分種別毎に音符が存在する予測確率を出力するよう音符存在確率予測モデルを学習してもよい。

【００３６】

例えば、特徴マップ生成モデルと音符存在確率予測モデルとは、モノフォニックボーカルと伴奏とを含む楽曲に対して適用されてもよい。この場合、ボーカル用特徴マップ生成モデルと伴奏用特徴マップ生成モデルとが、ボーカルの単音音源と音高情報とのペアから構成されるボーカル用学習データと、伴奏の単音音源と音高情報とのペアから構成される伴奏用学習データとを利用して、上述した学習処理と同様に学習される。一方、ボーカル用音符存在確率予測モデルと伴奏用音符存在確率予測モデルとが、学習用の音源と楽譜情報と利用して、音源を学習済みボーカル用特徴マップ生成モデルと学習済み伴奏用特徴マップ生成モデルとに入力することによって生成された特徴マップを入力とし、上述した学習処理と同様に学習される。

【００３７】

あるいは、特徴マップ生成モデルと音符存在確率予測モデルとは、楽器毎などの複数のパートを含む楽曲に対して適用されてもよい。上述したボーカルと伴奏とを含む楽曲に対する学習処理と同様であるが、この場合、音符存在確率予測モデルの出力は、特徴マップの固定長の区間に特定パートの特定音符が存在する予測確率であってもよい。例えば、"男声のＡ３の音高"、"女声のＡ３の音高"などの特定パートの特定音符の存在の予測確率を出力するようにしてもよい。

【００３８】
あるいは、本開示は拍子を有する楽曲に対して適用されてもよい。この場合、音符存在確率予測モデルの出力は、拍子のオンセット及びオフセットに関するものであってもよく、例えば、デフォルトボックスが一拍である予測確率が出力されてもよい。
【００３９】
図６は、本開示の一実施例による特徴マップ生成モデルの学習処理を示すフローチャートである。当該学習処理は、上述した学習装置１００又は学習装置１００のプロセッサによって実現される。
【００４０】
図６に示されるように、ステップＳ１０１において、学習用データ取得部１１０は、学習用データストレージ５０から単音音源と音高情報とのペアを取得する。例えば、音高は、「ド」から「シ」の１２音と無音との１３通りであり、当該１３通りの音高に対応する単音音源が取得されてもよい。
【００４１】
ステップＳ１０２において、学習用データ取得部１１０は、取得した単音音源を前処理する。具体的には、学習用データ取得部１１０は、単音音源の波形データに対して前処理（例えば、短時間フーリエ変換など）を実行し、単音音源のスペクトログラムを取得する。
【００４２】
ステップＳ１０３において、第１モデル学習部１２０は、前処理された単音音源と音高情報とのペアによって特徴マップ生成モデルを学習する。例えば、特徴マップ生成モデルは、畳み込みニューラルネットワークにより構成され、入力音源を音高の予測確率に変換する。具体的には、第１モデル学習部１２０は、単音音源のスペクトログラムを特徴マップ生成モデルに入力し、特徴マップ生成モデルの出力と音高情報との誤差が小さくなるように、バックプロパゲーションによって特徴マップ生成モデルのパラメータを更新する。
【００４３】
ステップＳ１０４において、第１モデル学習部１２０は、学習終了条件が充足されたか判断する。所定の学習終了条件は、例えば、所定数の学習用データに対して更新処理が終了した、誤差が所定の閾値以下に収束した、誤差の改善が所定の閾値以下に収束したなどであってもよい。所定の学習終了条件が充足されている場合（Ｓ１０４：ＹＥＳ）、第１モデル学習部１２０は、更新された特徴マップ生成モデルを学習済みモデルとして設定してもよい。他方、所定の学習終了条件が充足されていない場合（Ｓ１０４：ＮＯ）、当該処理はステップＳ１０１に移行し、上述した各ステップを繰り返す。
【００４４】
図７は、本開示の一実施例による音符存在確率予測モデルの学習処理を示すフローチャートである。当該学習処理は、上述した学習装置１００又は学習装置１００のプロセッサによって実現される。
【００４５】
図７に示されるように、ステップＳ２０１において、学習用データ取得部１１０は、学習用データストレージ５０からモノフォニック音源と楽譜情報とのペアを取得する。例えば、モノフォニック音源は歌唱音源の波形データであってもよく、楽譜情報は当該モノフォニック音源の楽譜を示す。
【００４６】
ステップＳ２０２において、学習用データ取得部１１０は、取得したモノフォニック音源を前処理する。具体的には、学習用データ取得部１１０は、モノフォニック音源の波形データに対して前処理（例えば、短時間フーリエ変換など）を実行し、モノフォニック音源のスペクトログラムを取得する。
【００４７】
ステップＳ２０３において、第２モデル学習部１３０は、前処理されたモノフォニック音源を学習済み特徴マップ生成モデルに入力し、学習済み特徴マップ生成モデルによって

生成された特徴マップを取得する。具体的には、第２モデル学習部１３０は、学習済み特徴マップ生成モデルの各畳み込み層から生成された特徴マップを取得する。生成された特徴マップは、各畳み込み層の畳み込みの程度に応じて異なる時間解像度の特徴マップとなる。

【００４８】

ステップＳ２０４において、第２モデル学習部１３０は、取得した特徴マップと楽譜情報とのペアによって音符存在確率予測モデルを学習する。例えば、音符存在確率予測モデルは、畳み込みニューラルネットワークにより構成により構成され、入力された特徴マップを当該特徴マップの固定長の区間に音符が存在する音符存在予測確率に変換する。具体的には、第２モデル学習部１３０は、各特徴マップを音符存在確率予測モデルに入力し、音符存在確率予測モデルの出力と楽譜情報との誤差が小さくなるように、バックプロパゲーションによって音符存在確率予測モデルのパラメータを更新する。

【００４９】

ステップＳ２０５において、第２モデル学習部１３０は、学習終了条件が充足されたか判断する。所定の学習終了条件は、例えば、所定数の学習用データに対して更新処理が終了した、誤差が所定の閾値以下に収束した、誤差の改善が所定の閾値以下に収束したなどであってもよい。所定の学習終了条件が充足されている場合（Ｓ２０５：ＹＥＳ）、第２モデル学習部１３０は、更新された音符存在確率予測モデルを学習済みモデルとして設定してもよい。他方、所定の学習終了条件が充足されていない場合（Ｓ２０５：ＮＯ）、当該処理はステップＳ２０１に移行し、上述した各ステップを繰り返す。

【００５０】

次に、図８及び９を参照して、本開示の一実施例による自動採譜装置を説明する。図８は、本開示の一実施例による自動採譜装置の機能構成を示すブロック図である。

【００５１】

図８に示されるように、自動採譜装置２００は、モデル処理部２１０及び楽譜生成部２２０を有する。

【００５２】

モデル処理部２１０は、単音音源から音高の予測確率を出力する学習済み特徴マップ生成モデルと、特徴マップから当該特徴マップの固定長の区間に音符が存在する予測確率を出力する学習済み音符存在確率予測モデルとを利用し、採譜対象の音源を学習済み特徴マップ生成モデルに入力し、当該学習済み特徴マップ生成モデルによって生成された特徴マップを学習済み音符存在確率予測モデルに入力し、特徴マップの固定長の区間に音符が存在する予測確率を出力する。

【００５３】

具体的には、モデル処理部２１０は、採譜対象の音源に対して短時間フーリエ変換などの前処理を実行して当該音源のスペクトログラムを取得し、取得したスペクトログラムを学習装置１００による学習済み特徴マップ生成モデルに入力して当該学習済み特徴マップ生成モデルの各畳み込み層からの特徴マップを取得する。そして、モデル処理部２１０は、取得した各特徴マップを学習装置１００による学習済み音符存在確率予測モデルに入力し、入力した特徴マップの各点を始点とする固定長の区間又はデフォルトボックスと同じ長さの音符が存在する予測確率を取得し、取得した各特徴マップの音符存在予測確率を楽譜生成部２２０にわたす。例えば、音符存在予測確率は、特徴マップのデフォルトボックスに存在する各音高（例えば、「ド」、「レ」、・・・「シ」、無音など）の確率の予測値であり、高い予測確率を有する音高が当該時間的長さに対応する音符に相当すると判断できる。

【００５４】

楽譜生成部２２０は、音符が存在する予測確率に基づき楽譜情報を生成する。具体的には、楽譜生成部２２０は、ＳＳＤに用いられるＮＭＳ（Ｎｏｎ－Ｍａｘｉｍｕｍ　Ｓｕｐｐｒｅｓｓｉｏｎ）に従って学習済み音符存在確率予測モデルの出力を後処理する。典型的には、学習済み音符存在確率予測モデルから多数の予測音符候補が出力される。これら

の予測音符候補から予測音符を特定する必要があり、ＳＳＤではＮＭＳを利用して予測音符候補をしばしば絞っている。

【００５５】

例えば、楽譜生成部２２０はまず、学習済み音符存在確率予測モデルに入力された特徴マップ上の各点に対して出力された音符存在予測確率のうち最大となる音符を当該時間における予測音符とする。そして、楽譜生成部２２０は、特徴マップ上の各点について予測音符を決定し、各点、予測音符及び対応する音符存在予測確率のデータセットをリスト化し、音符存在予測確率に関して降順にリスト内のデータセットをソートする。そして、楽譜生成部２２０は、所定の抽出条件を適用し、リストから予測音符候補を絞る。例えば、楽譜生成部２２０は、音符存在予測確率が所定の閾値（例えば、０．９など）以下であるデータセットをリストから削除してもよい。また、楽譜生成部２２０は、重複して検出された音符の重複を排除するため、予測音符が同じであって、かつ、予測音符の重複度が所定の閾値（例えば、８０％など）以上のデータセットがリストの上位にある場合、当該上位のリストのみを残すようにしてもよい。楽譜生成部２２０は、最終的なリストにおけるデータセットに基づき楽譜を生成する。

【００５６】

図９は、本開示の一実施例による自動採譜処理を示すフローチャートである。当該自動採譜処理は、上述した自動採譜装置２００又は自動採譜装置２００のプロセッサによって実現される。

【００５７】

図９に示されるように、ステップＳ３０１において、モデル処理部２１０は、採譜対象の音源を取得する。例えば、当該音源はモノフォニック音源であってもよいし、複数種別のオーディオ成分を含んでもよい。

【００５８】

ステップＳ３０２において、モデル処理部２１０は、取得した音源を前処理する。具体的には、モデル処理部２１０は、取得した音源に対して短時間フーリエ変換などの前処理を実行し、当該音源のスペクトログラムを取得する。

【００５９】

ステップＳ３０３において、モデル処理部２１０は、前処理した音源を学習済み特徴マップ生成モデルに入力して特徴マップを取得し、取得した特徴マップを学習済み音符存在確率予測モデルに入力して入力した特徴マップの各点を始点とする固定長の区間又はデフォルトボックスと同じ長さの音符が存在する予測確率を取得する。

【００６０】

ステップＳ３０４において、楽譜生成部２２０は、特徴マップ上の各点に対して取得した音符存在予測確率に基づき予測音符を決定する。具体的には、楽譜生成部２２０は、各点について取得した音符存在予測確率のうち最大となる音符存在予測確率に対応する音符を当該点に対する予測音符として決定する。

【００６１】

ステップＳ３０５において、楽譜生成部２２０は、決定された特徴マップの各点の予測音符に対して後処理を実行する。具体的には、楽譜生成部２２０は、ＳＳＤにおけるＮＭＳに従って特徴マップの各点の予測音符を絞る。例えば、楽譜生成部２２０は、特徴マップ上の各点について決定された予測音符に基づき、各点、予測音符及び対応する音符存在予測確率のデータセットをリスト化し、音符存在予測確率に関して降順にリスト内のデータセットをソートし、音符存在予測確率が所定の閾値（例えば、０．９など）以下であるデータセットをリストから削除すると共に、予測音符が同じであって、かつ、予測音符の重複度が所定の閾値（例えば、８０％など）以上のデータセットがリストの上位にある場合、当該上位のリストのみを残すようにしてもよい。

【００６２】

ステップＳ３０６において、楽譜生成部２２０は、最終的なリストにおけるデータセットに基づき楽譜を生成する。

【００６３】
上述した学習装置１００及び自動採譜装置２００はそれぞれ、例えば、図１０に示されるように、ＣＰＵ（Ｃｅｎｔｒａｌ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）１０１、ＧＰＵ（Ｇｒａｐｈｉｃｓ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）１０２、ＲＡＭ（Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）１０３、通信インタフェース（ＩＦ）１０４、ハードディスク１０５、入力装置１０６及び出力装置１０７によるハードウェア構成を有してもよい。ＣＰＵ１０１及びＧＰＵ１０２は、プロセッサ又は処理回路として参照されてもよく、学習装置１００及び自動採譜装置２００の各種処理を実行し、特に、ＣＰＵ１０１は学習装置１００及び自動採譜装置２００における各種処理の実行を制御し、ＧＰＵ１０２は機械学習モデルを学習及び実行するための各種処理を実行する。ＲＡＭ１０３及びハードディスク１０５は、学習装置１００及び自動採譜装置２００における各種データ及びプログラムを格納するメモリとして機能し、特に、ＲＡＭ１０３は、ＣＰＵ１０１及びＧＰＵ１０２における作業データを格納するワーキングメモリとして機能し、ハードディスク１０５は、ＣＰＵ１０１及びＧＰＵ１０２の制御プログラム及び／又は学習用データを格納する。通信ＩＦ１０４は、学習用データストレージ５０から学習用データを取得するための通信インタフェースである。入力装置１０６は、情報及びデータを入力するための各種デバイス（例えば、ディスプレイ、スピーカ、キーボード、タッチ画面など）であり、出力装置１０７は、処理の内容、経過、結果等の各種情報を表示する各種デバイス（例えば、ディスプレイ、プリンタ、スピーカなど）である。しかしながら、本開示による学習装置１００及び自動採譜装置２００は、上述したハードウェア構成に限定されず、他の何れか適切なハードウェア構成を有してもよい。

【００６４】
本開示の一態様では、
単音音源と音高情報とを第１の機械学習モデルの学習用データとして取得し、採譜対象の音源と楽譜情報とを第２の機械学習モデルの学習用データとして取得し、前記単音音源と前記採譜対象の音源とに対して前処理を実行し、それぞれのスペクトログラムを取得する学習用データ取得部と、
前記単音音源のスペクトログラムを学習用入力データとして入力し、前記単音音源の音高の予測確率を出力するよう前記音高情報によって第１の機械学習モデルを学習する第１モデル学習部と、
前記採譜対象の音源のスペクトログラムを学習済みの前記第１の機械学習モデルに入力することによって生成される特徴マップを学習用入力データとして入力し、前記特徴マップの固定長の区間に音符が存在する予測確率を出力するよう前記楽譜情報によって第２の機械学習モデルを学習する第２モデル学習部と、
を有する学習装置が提供される。

【００６５】
一実施例では、
前記第１の機械学習モデルと前記第２の機械学習モデルとは、畳み込みニューラルネットワークにより構成されてもよい。

【００６６】
一実施例では、
前記第２モデル学習部は、前記第１の機械学習モデルにより生成される異なる時間解像度を有する複数の特徴マップを前記第２の機械学習モデルに入力してもよい。

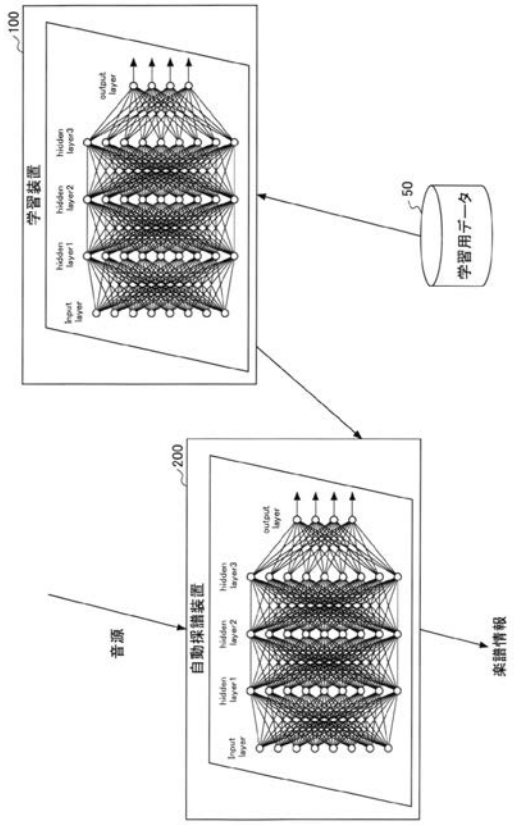
【００６７】
一実施例では、
前記第２モデル学習部は、前記第１の機械学習モデルと前記第２の機械学習モデルとをＳＳＤ（Ｓｉｎｇｌｅ Ｓｈｏｔ Ｄｅｔｅｃｔｉｏｎ）として実現してもよい。

【００６８】
一実施例では、
前記第１モデル学習部は、複数種別のオーディオ成分のそれぞれに対して前記第１の機

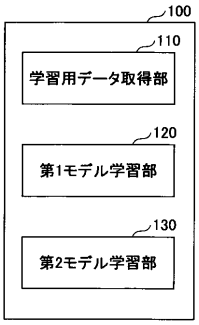
械学習モデルを学習し、
前記第２モデル学習部は、複数種別のオーディオ成分を含む採譜対象の音源に対して各オーディオ成分種別毎に音符が存在する予測確率を出力するよう前記第２の機械学習モデルを学習してもよい。
【００６９】
本開示の一態様では、
単音音源から音高の予測確率を出力する第１の学習済み機械学習モデルと、特徴マップから前記特徴マップの固定長の区間に音符が存在する予測確率を出力する第２の学習済み機械学習モデルとを利用し、採譜対象の音源を前記第１の学習済み機械学習モデルに入力し、前記第１の学習済み機械学習モデルによって生成された特徴マップを前記第２の学習済み機械学習モデルに入力し、前記特徴マップの固定長の区間に音符が存在する予測確率を出力するモデル処理部と、
前記音符が存在する予測確率に基づき楽譜情報を生成する楽譜生成部と、
を有する自動採譜装置が提供される。
【００７０】
一実施例では、
前記モデル処理部は、前記採譜対象の音源に対して前処理を実行することによってスペクトログラムを取得し、前記スペクトログラムを前記第１の学習済み機械学習モデルに入力してもよい。
【００７１】
一実施例では、
前記モデル処理部は、前記特徴マップ上の各点について前記第２の学習済み機械学習モデルから出力された最大の予測確率を有する音符を予測音符として決定してもよい。
【００７２】
一実施例では、
前記楽譜生成部は、ＮＭＳ（Ｎｏｎ－Ｍａｘｉｍｕｍ　Ｓｕｐｐｒｅｓｓｉｏｎ）に従って抽出された予測音符に基づき楽譜情報を生成してもよい。
【００７３】
本開示の一態様では、
プロセッサが、単音音源と音高情報とを第１の機械学習モデルの学習用データとして取得し、採譜対象の音源と楽譜情報とを第２の機械学習モデルの学習用データとして取得し、前記単音音源と前記採譜対象の音源とに対して前処理を実行し、それぞれのスペクトログラムを取得するステップと、
前記プロセッサが、前記単音音源のスペクトログラムを学習用入力データとして入力し、前記単音音源の音高の予測確率を出力するよう前記音高情報によって第１の機械学習モデルを学習するステップと、
前記プロセッサが、前記採譜対象の音源のスペクトログラムを学習済みの前記第１の機械学習モデルに入力することによって生成される特徴マップを学習用入力データとして入力し、前記特徴マップの固定長の区間に音符が存在する予測確率を出力するよう前記楽譜情報によって第２の機械学習モデルを学習するステップと、
を有する学習方法が提供される。
【００７４】
本開示の一態様では、
プロセッサが、単音音源から音高の予測確率を出力する第１の学習済み機械学習モデルに採譜対象の音源を入力するステップと、
前記プロセッサが、特徴マップから前記特徴マップの固定長の区間に音符が存在する予測確率を出力する第２の学習済み機械学習モデルに前記第１の学習済み機械学習モデルによって生成された特徴マップを入力するステップと、
前記プロセッサが、前記第２の学習済み機械学習モデルから出力された前記音符が存在する予測確率に基づき楽譜情報を生成するステップと、

を有する自動採譜方法が提供される。
【００７５】
本開示の一態様では、
単音音源と音高情報とを第１の機械学習モデルの学習用データとして取得し、採譜対象の音源と楽譜情報とを第２の機械学習モデルの学習用データとして取得し、前記単音音源と前記採譜対象の音源とに対して前処理を実行し、それぞれのスペクトログラムを取得するステップと、
前記単音音源のスペクトログラムを学習用入力データとして入力し、前記単音音源の音高の予測確率を出力するよう前記音高情報によって第１の機械学習モデルを学習するステップと、
前記採譜対象の音源のスペクトログラムを学習済みの前記第１の機械学習モデルに入力することによって生成される特徴マップを学習用入力データとして入力し、前記特徴マップの固定長の区間に音符が存在する予測確率を出力するよう前記楽譜情報によって第２の機械学習モデルを学習するステップと、
をプロセッサに実行させるプログラムが提供される。
【００７６】
本開示の一態様では、
単音音源から音高の予測確率を出力する第１の学習済み機械学習モデルに採譜対象の音源を入力するステップと、
特徴マップから前記特徴マップの固定長の区間に音符が存在する予測確率を出力する第２の学習済み機械学習モデルに前記第１の学習済み機械学習モデルによって生成された特徴マップを入力するステップと、
前記第２の学習済み機械学習モデルから出力された前記音符が存在する予測確率に基づき楽譜情報を生成するステップと、
をプロセッサに実行させるプログラムが提供される。
【００７７】
本開示の一態様では、
上述したプログラムを記憶するコンピュータ可読記憶媒体が提供される。
【００７８】
以上、本開示の実施例について詳述したが、本開示は上述した特定の実施形態に限定されるものではなく、特許請求の範囲に記載された本開示の要旨の範囲内において、種々の変形・変更が可能である。
【符号の説明】
【００７９】
５０　学習用データストレージ
１００　学習装置
２００　自動採譜装置

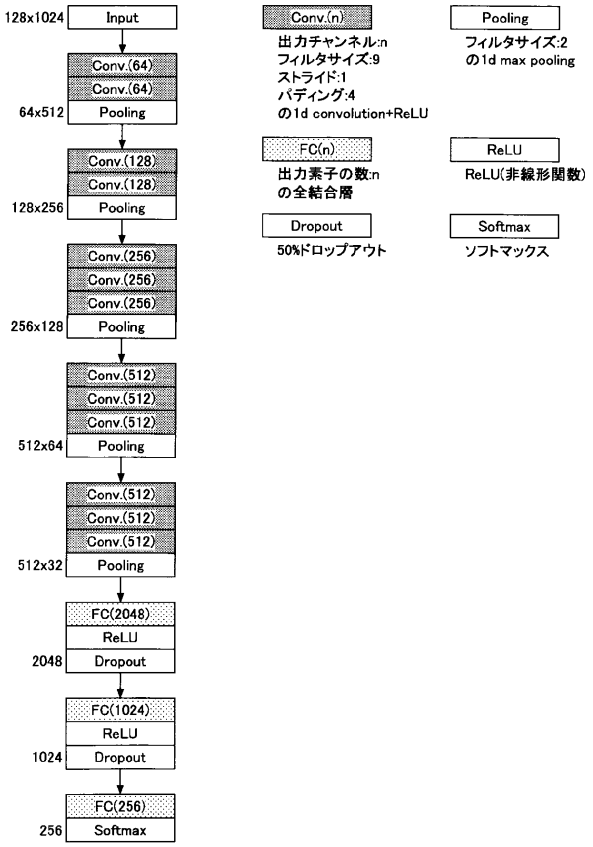
【図１】


100
学習装置
output layer
hidden layer3
hidden layer2
hidden layer1
input layer
50
学習用データ
200
自動採譜装置
音源
楽譜情報

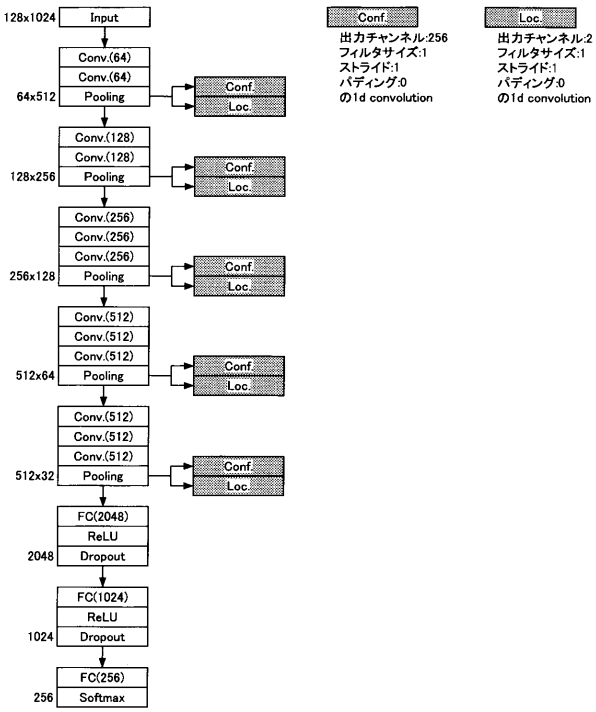
【図２】


100
110
学習用データ取得部
120
第1モデル学習部
130
第2モデル学習部

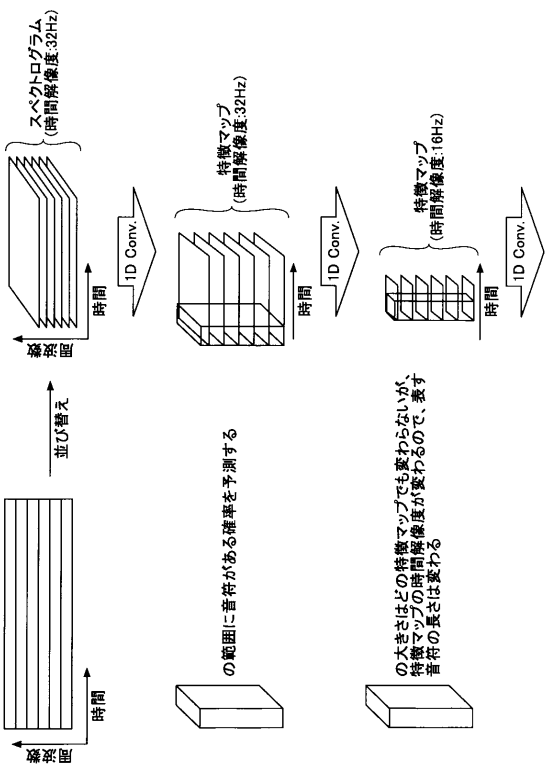
【図３】


128x1024
Input
Conv.(64)
Conv.(64)
64x512
Pooling
Conv.(128)
Conv.(128)
128x256
Pooling
Conv.(256)
Conv.(256)
Conv.(256)
256x128
Pooling
Conv.(512)
Conv.(512)
Conv.(512)
512x64
Pooling
Conv.(512)
Conv.(512)
Conv.(512)
512x32
Pooling
FC(2048)
ReLU
2048
Dropout
FC(1024)
ReLU
1024
Dropout
FC(256)
256
Softmax
Conv.(n)
出力チャンネル:n
フィルタサイズ:9
ストライド:1
パディング:4
の1d convolution+ReLU
Pooling
フィルタサイズ:2
の1d max pooling
FC(n)
出力素子の数:n
の全結合層
ReLU
ReLU(非線形関数)
Dropout
50%ドロップアウト
Softmax
ソフトマックス

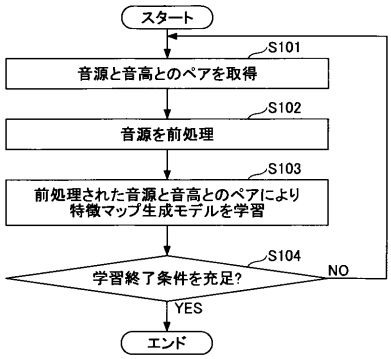
【図４】


128x1024
Input
Conv.(64)
Conv.(64)
64x512
Pooling
Conf.
Loc.
Conv.(128)
Conv.(128)
128x256
Pooling
Conf.
Loc.
Conv.(256)
Conv.(256)
Conv.(256)
256x128
Pooling
Conf.
Loc.
Conv.(512)
Conv.(512)
Conv.(512)
512x64
Pooling
Conf.
Loc.
Conv.(512)
Conv.(512)
Conv.(512)
512x32
Pooling
Conf.
Loc.
FC(2048)
ReLU
2048
Dropout
FC(1024)
ReLU
1024
Dropout
FC(256)
256
Softmax
Conf.
出力チャンネル:256
フィルタサイズ:1
ストライド:1
パディング:0
の1d convolution
Loc.
出力チャンネル:2
フィルタサイズ:1
ストライド:1
パディング:0
の1d convolution

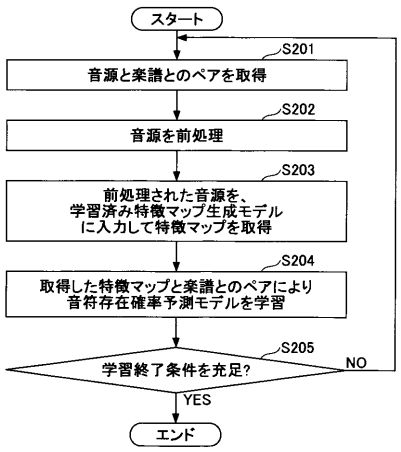
【図５】



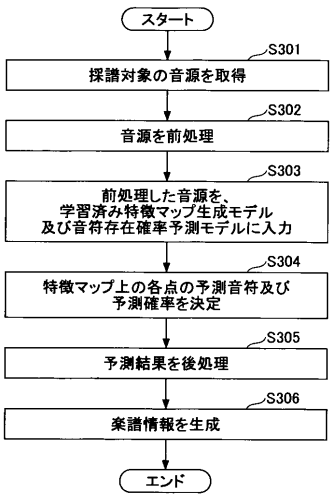
【図６】



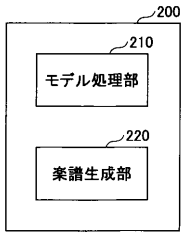
【図７】



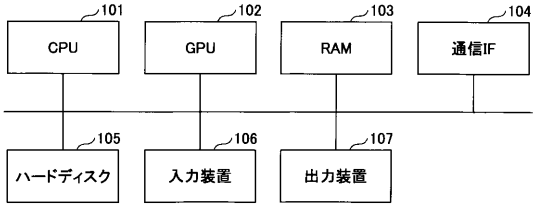
【図９】



【図８】



【図１０】



フロントページの続き

(51)Int.Cl.	FI		テーマコード（参考）
	Ｇ０６Ｎ	99/00 １５０	