(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2020年8月27日(27.08.2020)

WIPO | PCT

(10) 国際公開番号
WO 2020/171036 A1

(51) 国際特許分類:
G10L 13/00 (2006.01) *G10H 7/08* (2006.01)

(21) 国際出願番号： PCT/JP2020/006162

(22) 国際出願日： 2020年2月18日(18.02.2020)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(30) 優先権データ：
特願 2019-028684 2019年2月20日(20.02.2019) JP

(71) 出願人: ヤ マ ハ 株 式 会 社 (**YAMAHA CORPORATION**) [JP/JP]; 〒4308650 静岡県浜松市中区中沢町１０番１号 Shizuoka (JP).

(72) 発明者: 西 村 方 成 (**NISHIMURA, Masanari**); 〒4308650 静岡県浜松市中区中沢町１０番１号 ヤマハ株式会社内 Shizuoka (JP).

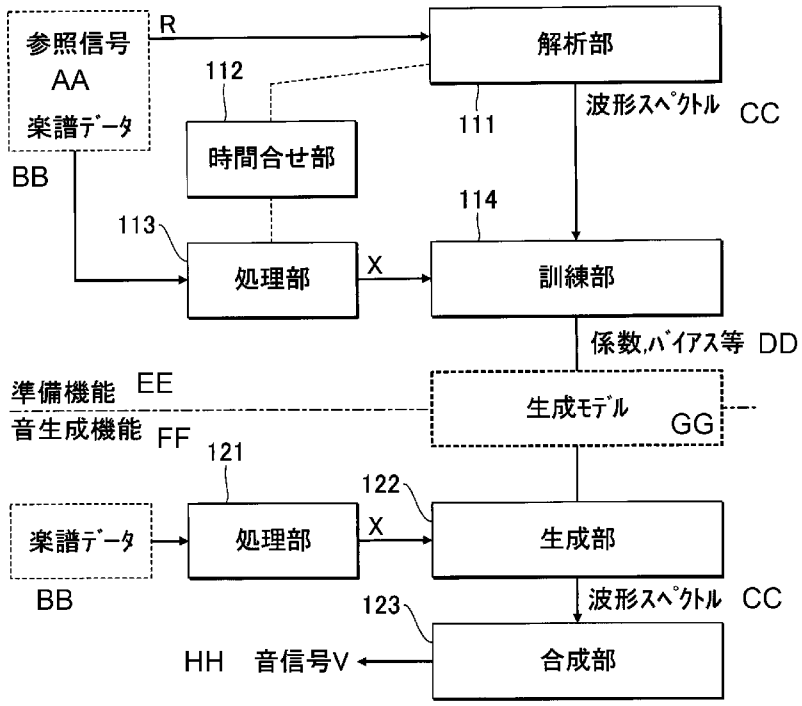
(74) 代理人: 大林 章, 外(**OHBAYASHI, Akira** et al.); 〒1130033 東京都文京区本郷２－１５－１３ お茶の水ウィングビル６階 Tokyo (JP).

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ,

WO 2020/171036 A1

(54) **Title:** SOUND SIGNAL SYNTHESIS METHOD, GENERATIVE MODEL TRAINING METHOD, SOUND SIGNAL SYNTHESIS SYSTEM, AND PROGRAM

(54) 発明の名称： 音信号合成方法、生成モデルの訓練方法、音信号合成システムおよびプログラム



111 Analysis unit
112 Time setting unit
113, 121 Processing unit
114 Training unit
122 Generation unit
123 Synthesis unit
AA Reference signal
BB Musical score data
CC Waveform spectrum
DD Coefficient, bias, etc.
EE Preparation function
FF Sound generation function
GG Generative model
HH Sound signal

(57) **Abstract:** This sound signal synthesis method to be realized by a computer entails: setting a hot value for each piece of pitch name data that is specified by a pitch name of a sound signal to be synthesized, and that is part of pitch data which includes a plurality of pieces of pitch name data corresponding to a plurality of pitch names, the hot value corresponding to a shift from the pitch name of the pitch of the sound signal; using a generative model to estimate output data indicating the sound signal according to the set pitch data; and synthesizing the sound signal in accordance with the estimated output data.

[続葉有]

NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

一 国際調査報告(条約第21条(3))

(57) 要約:コンピュータにより実現される音信号合成方法は、複数の音名に対応する複数の音名データを含む音高データの、合成すべき音信号の音名で特定される音名データに、それぞれ、その音信号の音高のその音名からのずれに応じたホット値を設定し、生成モデルを用いて、設定された音高データに応じた音信号を示す出力データを推定し、推定された出力データに応じて、音信号を合成する。

明　細　書

発明の名称 ：

音信号合成方法、生成モデルの訓練方法、音信号合成システムおよびプログラム

技術分野

[0001] 本発明は、音信号を合成する音源技術に関する。

背景技術

[0002] 特許文献１に示すNSynth、または非特許文献１に示すNPSS（Neural Parametric Singing Synthesizer）など、ニューラルネットワーク（以下、「NN」と呼ぶ）を用いて、条件入力に応じた音波形を生成する音源（以下、「DNN（Deep Neural Network）音源」と呼ぶ）が提案されている。NSynthは、エンベディング（embedding／埋込ベクトル）に応じて、サンプリング周期ごとに、音信号のサンプルを生成する。NPSSのTimbreモデルは、ピッチおよびタイミング情報に応じて、フレームごとに、音信号のスペクトルを生成する。

先行技術文献

特許文献

[0003] 特許文献1：米国特許第１００６８５５７号明細書

非特許文献

[0004] 非特許文献1：Merlijn Blaauw, Jordi Bonada, 「A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs」, Appl. Sci. 2017, 7, 1313

発明の概要

発明が解決しようとする課題

[0005] 特許文献１のNsynthまたは非特許文献１のNPSSなどのDNN音源では、所望の１個の音階を指定する音高データにより、合成する音信号の音高を制御する。ピッチ包絡またはビブラートなど、音符などで指定される音階からの音高

の動的なずれを制御することについては、考慮されていない。

[0006] DNN音源の訓練フェーズでは、音高データを入力として、音信号または波形スペクトルを表す出力データを推定するNNを訓練する。DNN音源は、ビブラートのかかった音信号で訓練をすれば、ビブラートのかかった音信号を生成し、ピッチベンドがかかった音信号で訓練すれば、ピッチベンドのかかった音信号を生成する。しかしながら、ビブラートまたはピッチベンドのような、動的に変化する音高のずれ（ピッチベンド量）を、時間変化する数値で制御することはできない。

[0007] 本開示は、合成する音信号の動的なピッチ変化を、時間変化する数値で制御することを目的とする。

課題を解決するための手段

[0008] 本開示のひとつの態様に係る音信号合成方法は、合成すべき第１音信号の音高を表す第１音高データを生成し、第２音信号の音高を表す第２音高データと前記第２音信号との関係を学習した生成モデルを用いて、前記第１音高データに応じた前記第１音信号を示す出力データを推定するコンピュータにより実現される音信号合成方法であって、前記第１音高データは、相異なる音名に対応する複数の音名データを含み、前記第１音高データの生成においては、前記複数の音名データのうち、前記第１音信号の音高に対応する音名データを、当該音名データに対応する音名の基準音高と当該第１音信号の音高との相違に応じたホット値に設定する。

[0009] 本開示のひとつの態様に係る訓練方法は、合成すべき音信号の音高を表す音高データを用意し、前記音高データの入力に対して前記音信号を示す出力データが出力されるように生成モデルを訓練するコンピュータにより実現される生成モデルの訓練方法であって、前記音高データは、相異なる音名に対応する複数の音名データを含み、前記音高データの用意においては、前記複数の音名データのうち、前記音信号の音高に対応する音名データを、当該音名データに対応する音名の基準音高と当該音信号の音高との相違に応じたホット値に設定する・

[0010] 本開示のひとつの態様に係る音信号合成システムは、１以上のプロセッサと１以上のメモリとを具備する音信号合成システムであって、前記１以上のメモリは、第２音信号の音高を表す第２音高データと前記第２音信号との関係を学習した生成モデルを記憶し、前記１以上のプロセッサは、合成すべき第１音信号の音高を表す第１音高データを生成し、第２音信号の音高を表す第２音高データと前記第２音信号との関係を学習した生成モデルを用いて、前記第１音高データに応じた前記第１音信号を示す出力データを推定し、前記第１音高データは、相異なる音名に対応する複数の音名データを含み、前記１以上のプロセッサは、第１音高データの生成において、前記複数の音名データのうち、前記第１音信号の音高に対応する音名データを、当該音名データに対応する音名の基準音高と当該第１音信号の音高との相違に応じたホット値に設定する。

[0011] 本開示のひとつの態様に係るプログラムは、合成すべき第１音信号の音高を表す第１音高データを生成する処理部、および、第２音信号の音高を表す第２音高データと前記第２音信号との関係を学習した生成モデルを用いて、前記第１音高データに応じた前記第１音信号を示す出力データを推定する生成部としてコンピュータを機能させるプログラムであって、前記第１音高データは、相異なる音名に対応する複数の音名データを含み、前記第１音高データの生成においては、前記複数の音名データのうち、前記第１音信号の音高に対応する音名データを、当該音名データに対応する音名の基準音高と当該第１音信号の音高との相違に応じたホット値に設定する。

図面の簡単な説明

[0012] [図1]音信号合成システムのハードウェア構成を示すブロック図である。

[図2]音信号合成システムの機能構成を示すブロック図である。

[図3]音高データの説明図である。

[図4]訓練部と生成部による処理の説明図である。

[図5]１ホットレベル記法に従う音高データの説明図である。

[図6]準備処理のフローチャートである。

[図7]音生成処理のフローチャートである。

[図8]２ホットレベル記法に従う音高データの説明図である。

[図9]４ホットレベル記法に従う音高データの説明図である。

[図10]各音名と音信号の音高の近接度の変形例を示す図である。

発明を実施するための形態

[0013] Ａ：第１実施形態

図１は、本開示の音信号合成システム１００の構成を例示するブロック図である。音信号合成システム１００は、制御装置１１と記憶装置１２と表示装置１３と入力装置１４と放音装置１５とを具備するコンピュータシステムで実現される。音信号合成システム１００は、例えば携帯電話機、スマートフォンまたはパーソナルコンピュータ等の情報端末である。音信号合成システム１００は、単体の装置で実現されるほか、相互に別体で構成された複数の装置（例えばサーバークライアントシステム）でも実現される。

[0014] 制御装置１１は、音信号合成システム１００を構成する各要素を制御する単数または複数のプロセッサである。具体的には、例えばＣＰＵ（Central Processing Unit）、ＳＰＵ（Sound Processing Unit）、ＤＳＰ（Digital Signal Processor）、ＦＰＧＡ（Field Programmable Gate Array）、またはＡＳＩＣ（Application Specific Integrated Circuit）等の１種類以上のプロセッサにより、制御装置１１が構成される。制御装置１１は、合成音の波形を表す時間領域の音信号Ｖを生成する。

[0015] 記憶装置１２は、制御装置１１が実行するプログラムと制御装置１１が使用する各種のデータとを記憶する単数または複数のメモリである。記憶装置１２は、例えば磁気記録媒体もしくは半導体記録媒体等の公知の記録媒体、または、複数種の記録媒体の組合せで構成される。なお、音信号合成システム１００とは別体の記憶装置１２（例えばクラウドストレージ）を用意し、移動体通信網またはインターネット等の通信網を介して制御装置１１が記憶装置１２に対する書込および読出を実行してもよい。すなわち、記憶装置１２は音信号合成システム１００から省略されてもよい。

[0016] 表示装置１３は、制御装置１１が実行したプログラムの演算結果を表示する。表示装置１３は、例えばディスプレイである。表示装置１３は音信号合成システム１００から省略されてもよい。

[0017] 入力装置１４は、ユーザの入力を受け付ける。入力装置１４は、例えばタッチパネルである。入力装置１４は音信号合成システム１００から省略されてもよい。

[0018] 放音装置１５は、制御装置１１が生成した音信号Ｖが表す音声を再生する。放音装置１５は、例えばスピーカまたはヘッドホンである。なお、制御装置１１が生成した音信号Ｖをデジタルからアナログに変換するＤ/Ａ変換器と音信号Ｖを増幅する増幅器とについては図示を便宜的に省略した。また、図１では、放音装置１５を音信号合成システム１００に搭載した構成を例示したが、音信号合成システム１００とは別体の放音装置１５を音信号合成システム１００に有線または無線で接続してもよい。

[0019] 図２は、音信号合成システム１００の機能構成を示すブロック図である。制御装置１１は、記憶装置１２に記憶されたプログラムを実行することで、生成モデルを用いて、歌手の歌唱音または楽器の演奏音などの音波形を表す時間領域の音信号Ｖを生成する音生成機能（処理部１２１、生成部１２２，および合成部１２３）を実現する。また、制御装置１１は、記憶装置１２に記憶されたプログラムを実行することで、音信号Ｖの生成に用いられる生成モデルの準備を行う準備機能（解析部１１１、時間合せ部１１２、処理部１１３、および訓練部１１４）を実現する。なお、複数の装置の集合（すなわちシステム）で制御装置１１の機能を実現してもよいし、制御装置１１の機能の一部または全部を専用の電子回路（例えば信号処理回路）で実現してもよい。

[0020] まず、音高データＸ1と、音高データＸ1に応じて出力データを生成する生成モデルと、生成モデルの訓練に用いる参照信号Ｒについて説明する。

[0021] 音高データＸ1は、音信号（参照信号Ｒまたは音信号Ｖ）の音高（以下「目標音高」という）Ｐを示すデータである。図３は、音高データＸ1の例である

。音高データX1は、相異なる音名（…”Ｇ＃３”，”Ａ３”，”Ａ＃３”，”Ｂ３”，”Ｃ４”，”Ｃ＃４”，”Ｄ４”，…）に対応する複数（Ｍ個）の音名データを有する（Ｍは２以上の自然数）。なお、音名を表す記号（Ｃ，Ｄ，Ｅ，…）が共通する音名であってもオクターブが相違する音名は別個の音名として区別する。

[0022] 音高データX1を構成するＭ個の音名データのうち、目標音高Ｐに対応する１個の音名データ（以下「有効音名データ」という）は、当該有効音名データの音名に対応する所定の音高（以下「基準音高」という）Ｑと目標音高Ｐとの音高差（ずれ）に応じたずれ値に設定される。ずれ値は、ホット値（Hot Value）の一例である。１個の音名に対応する基準音高Ｑは、当該音名に対応する標準的な音高（ピッチ）である。他方、音高データX1を構成するＭ個の音名データのうち、有効音名データ以外の(Ｍ－１)個の音名データは、目標音高Ｐに関与しないことを示す定数値（例えば０）に設定される。目標音高Ｐに関与しないことを示す定数値は、コールド値の一例である。以上の説明から理解される通り、音高データX1は、音信号（参照信号Ｒまたは音信号Ｖ）の目標音高Ｐに対応する音名と、当該音名の基準音高Ｑに対する目標音高Ｐのずれ値との双方を指定する。

[0023] 生成モデルは、音高データX1を含む制御データＸに応じて、音信号Ｖの波形スペクトル（例えば、メルスペクトログラム、または基本周波数などの特徴量）の時系列を生成する統計的モデルである。制御データＸは、合成されるべき音信号の条件を指定するデータである。生成モデルの生成特性は、記憶装置１２に記憶された複数の変数（係数およびバイアスなど）により規定される。統計的モデルは、波形スペクトルを推定するニューラルネットワークである。そのニューラルネットワークは、例えば、WaveNet(TM)のような、音信号Ｖの過去の複数のサンプルに基づいて、現在のサンプルの確率密度分布を推定する回帰的なタイプでもよい。また、そのアルゴリズムも任意であり、例えば、CNN（Convolutional Neural Network）タイプでもRNN（Recurrent Neural Network）タイプでよいし、その組み合わせでもよい。さらに、LS

TM（Long Short-Term Memory）またはATTENTIONなどの付加的要素を備えるタイプでもよい。生成モデルの複数の変数は、後述する準備機能による訓練データを用いた訓練により確立される。複数の変数が確立された生成モデルは、後述する音生成機能において音信号Ｖの生成に使用される。

[0024] 記憶装置１２は、生成モデルの訓練のために、ある楽譜をプレイヤーが演奏した時間領域の波形を示す音信号（以下、「参照信号」と呼ぶ）Ｒと、当該楽譜を表す楽譜データとが複数記録されている。各楽譜データは音符の時系列を含む。各楽譜データに対応する参照信号Ｒは、当該楽譜データが表す楽譜の音符の系列に対応する部分波形の時系列を含む。各参照信号Ｒは、サンプリング周期（例えば、48kHz）ごとのサンプルの時系列で構成され、音の波形を表す時間領域の信号である。演奏は、人間による楽器の演奏に限らず、歌手による歌唱、または楽器の自動演奏であってもよい。機械学習で良い音を生成するためには、一般的に十分な個数の訓練データが要求されるので、ターゲットとする楽器またはプレイヤーなどについて、多数の演奏の音信号を事前に収録し、参照信号Ｒとして記憶装置１２に記憶しておく。

[0025] 図２の上段に図示される準備機能について説明する。解析部１１１は、複数の楽譜にそれぞれ対応する複数の参照信号Ｒの各々について、時間軸上のフレームごとに周波数領域のスペクトル（以下「波形スペクトル」と呼ぶ）を算定する。参照信号Ｒの波形スペクトルの算定には、例えば離散フーリエ変換等の公知の周波数解析が用いられる。波形スペクトルには、基本周波数などの音響的な特徴量も含まれる。

[0026] 時間合せ部１１２は、解析部１１１で得られた波形スペクトル等の情報に基づき、各参照信号Ｒに対応する楽譜データにおける複数の発音単位の開始時点と終了時点とを、参照信号Ｒにおけるその発音単位に対応する部分波形の開始時点と終了時点とに揃える。ここで、発音単位は、例えば、音高と発音期間とが指定された１個の音符である。なお、１個の音符を、音色等の波形の特徴が変化するポイントで分割して、複数の発音単位に分けてもよい。

[0027] 処理部１１３は、各参照信号Ｒに時間が揃えられた楽譜データの各発音単

位の情報に基づき、フレームを単位とする時刻ｔごとに、参照信号Ｒのうち当該時刻ｔの部分波形に対応する制御データＸを生成する。処理部１１３が生成する制御データＸは、前述の通り、参照信号Ｒの条件を指定する。

[0028] 制御データＸは、図４に例示される通り、音高データＸ1と開始停止データＸ2とコンテキストデータＸ3とを含む。音高データＸ1は、参照信号Ｒの部分波形における目標音高Ｐを表す。開始停止データＸ2は、各部分波形の開始期間（アタック）と終了期間（リリース）とを表す。１個の音符に相当する部分波形内の１個のフレームのコンテキストデータＸ3は、当該音符と前後の音符との音高差、または楽譜内における当該音符の相対的な位置を表す情報など、複数の発音単位との関係（すなわちコンテキスト）を表す。制御データＸには、さらに、楽器、歌手または奏法など、その他の情報を含んでいてもよい。

[0029] 前述の通り、音高データＸ1を構成するＭ個の音名データのうち、音信号（参照信号Ｒまたは音信号Ｖ）の目標音高Ｐに対応する１個の有効音名データは、当該音名に対応する基準音高Ｑに対する目標音高Ｐの音高差に応じたずれ値に設定される。この記法に従う音高データＸ1を、１ホットレベル記法の音高データＸ1と呼ぶ。処理部１１３（制御装置１１）は、音高データＸ1のＭ個の音名データのうち、参照信号Ｒの目標音高Ｐに対応する１個の有効音名データを、当該音名に対応する基準音高Ｑと目標音高Ｐとの音高差に応じたずれ値に設定する。図５に、その設定例を示す。

[0030] 図５の上段には、時間軸（横軸）と音高軸（縦軸）とが設定された２次元平面に、楽譜データが表す楽譜を構成する音符の時系列と、当該楽譜を演奏した演奏音の音高（目標音高Ｐ）とが図示されている。図５の例では、音符Ｆ＃、音符Ｆ、休符、音符Ｆ、音符Ｆ＃、音符Ｆが順に演奏される。図５における目標音高Ｐは、例えば音高が連続的に変化する楽器から発音される演奏音の音高である。

[0031] 図５に例示される通り、音高軸は、相異なる音名に対応する複数の範囲（以下「単位範囲」という）Ｕに区画される。各音名に対応する基準音高Ｑは

、例えば当該音名に対応する単位範囲Ｕの中点に相当する。例えば、音名Ｆ＃に対応する基準音高Ｑ(F#)は、当該音名Ｆ＃に対応する単位範囲Ｕ(F#)の中点である。図５から理解される通り、目標音高Ｐが各音符の基準音高Ｑに近付くように楽曲が演奏される。各音名に対応する基準音高Ｑは音高軸上に離散的に設定されるのに対し、目標音高Ｐは時間の経過とともに連続的に変化する。したがって、目標音高Ｐは基準音高Ｑに対してずれている。

[0032] 図５の中段は、音高データＸ1の各音名データが表す数値の時間的な変化を表すグラフである。図５の中段における縦軸の数値０は、各音名に対応する基準音高Ｑを意味する。目標音高Ｐが１個の音名の単位範囲Ｕ内にある場合、音高データＸ1のＭ個の音名データのうち当該音名に対応する１個の音名データが有効音名データとして選択され、当該有効音名データが基準音高Ｑに対するずれ値に設定される。

[0033] 有効音名データが表すずれ値は、当該有効音名データの音名に対応する基準音高Ｑ（＝０）に対する目標音高Ｐの相対値である。１個の音名に対応する単位範囲Ｕの幅は１００セント（半音分）であるから、基準音高Ｑに対する目標音高Ｐの音高差は±５０セントの範囲に収まる。一方、有効音名データに設定されるずれ値は、０から１の範囲の何れかの値をとる。音高差とずれ値との対応付けは任意であるが、例えば、ずれ値の０から１の範囲が、音高差の－５０セントから＋５０セントに対応付けられる。例えば、ずれ値の０が音高差の－５０セントに対応し、ずれ値の０．５が音高差の０セントに対応し、ずれ値の１が音高差の＋５０セントに対応する。

[0034] 図５に例示される通り、時間軸上の時刻ｔ1において、目標音高Ｐは、音名Ｆ＃に対応する単位範囲Ｕ(F#)内にあり、かつ、当該音名Ｆ＃に対応する基準音高Ｑ(F#)との音高差は＋４０セントである。したがって、時刻ｔ1においては、音高データＸ1のＭ個の音名データのうち、音名Ｆ＃に対応する１個の有効音名データが、＋４０セントの音高差に対応するずれ値０．９に設定され、残余の(Ｍ－１)個の音名データは０（コールド値）に設定される。

[0035] また、時刻ｔ2において、目標音高Ｐは、音名Ｆに対応する単位範囲Ｕ(F)

内にあり、かつ、当該音名Ｆに対応する基準音高Ｑ(F)との音高差は＋２０セントである。したがって、時刻ｔ2においては、音高データＸ1のＭ個の音名データのうち、音名Ｆに対応する１個の有効音名データが、＋２０セントの音高差に対応するずれ値０．７に設定される。

[0036] なお、音高差とずれ値との対応は上記に限らず、例えば、ずれ値の０．２から１の範囲を、音高差の－５０セントから＋５０セントの範囲に対応付けてもよい。例えば、ずれ値の０．２が音高差の－５０セントに対応し、ずれ値の０．６が音高差の０セントに対応し、ずれ値の１が音高差の＋５０セントに対応する。或いは、ずれ値の正負と音高差の正負との関係を反転し、ずれ値の０．２から１の範囲を、音高差の＋５０セントから－５０セントに対応付けてもよい。

[0037] 解析部１１１および処理部１１３による処理の結果、複数の参照信号Ｒと複数の楽譜データとから、生成モデルを訓練するための複数の発音単位データが準備される。各発音単位データは、制御データＸと波形スペクトルとのセットである。複数の発音単位データは、訓練部１１４による訓練に先立ち、生成モデルの訓練のための訓練データと、生成モデルのテストのためのテストデータとに分けられる。複数の発音単位データの大部分を訓練データとし、一部をテストデータにする。訓練データによる訓練は、複数の発音単位データをフレームの所定個ごとにバッチとして分割し、バッチ単位で全バッチにわたり順番に行われる。

[0038] 訓練部１１４は、図４の上段に例示するように、訓練データを受け取り、その各バッチの複数の発音単位の波形スペクトルと制御データＸとを順番に用いて生成モデルを訓練する。生成モデルは、フレーム（時刻ｔ）ごとに、波形スペクトルを表す出力データを推定する。出力データは、波形スペクトルを構成する複数の成分の各々の確率密度分布を示すデータであっても良いし、各成分の値であってもよい。訓練部１１４は、１バッチ分の各発音単位データにおける制御データＸを生成モデルに入力することで、その制御データＸに対応する出力データの時系列を推定する。訓練部１１４は、推定され

た出力データと訓練データのうち対応する波形スペクトル（すなわち正解値）とに基づいて損失関数Ｌ（１バッチ分の累算値）を計算する。そして、訓練部１１４は、その損失関数Ｌが最小化されるように生成モデルの複数の変数を最適化する。例えば、損失関数Ｌとしては、出力データが確率密度分布である場合にはクロスエントロピー関数などが使用され、出力データが波形スペクトルの値である場合には二乗誤差関数などが使用される。訓練部１１４は、訓練データを利用した以上の訓練を、テストデータについて算出される損失関数Ｌの値が十分に小さくなるか、或いは、相前後する損失関数Ｌの変化が十分に小さくなるまで繰り返し行う。こうして確立された生成モデルは、各時刻tにおける制御データＸと、参照信号Ｒのうち当該時刻ｔに対応する波形スペクトルとの間に潜在する関係を習得している。この生成モデルを用いることで、生成部１２２は、未知の音信号Ｖの制御データＸについても、品質の良い波形スペクトルを生成できる。

[0039] 図６は、準備処理のフローチャートである。準備処理は、例えば音信号合成システム１００の利用者からの指示を契機として開始される。

[0040] 準備処理が開始されると、制御装置１１（解析部１１１）は、複数の参照信号Ｒの各々から各部分波形の波形スペクトルを生成する（Sa1）。次に、制御装置１１（時間合せ部１１２および処理部１１３）は、その部分波形に対応する楽譜データから、その部分波形に対応する発音単位の音高データＸ1を含む制御データＸを作成する（Sa2）。制御装置１１（訓練部１１４）は、各発音単位の各時刻tの制御データＸと当該発音単位に対応する波形スペクトルとを用いて生成モデルを訓練し、生成モデルの複数の変数を確立する（Sa3）。

[0041] ここでは、各音名の基準音高Ｑに対するずれ値を示す音高データＸ1を含む制御データＸを入力として生成モデルが訓練される。したがって、その訓練により確立された生成モデルは、制御データＸが示す音高のずれ値と、音信号（参照信号Ｒ）の波形スペクトルとの間に潜在する関係を習得する。これにより、音名とずれ値とを指定する音高データＸ1を含む制御データを生成モ

デルに入力すれば、その指定された音名とずれに応じた音高の音信号Ｖを生成できる。

[0042] 本願の発明者は、対比例として、音信号の音名を示す従来のone-hotの音高データと、当該音名の基準音高Ｑに対する音信号の音高ずれを示すベンドデータとをパラレルに含む制御データを入力として生成モデルを訓練した。しかしながら、その訓練で確立された生成モデルを用いた場合、生成される音信号の音高は、音高データが示す音高に追従するものの、ベンドデータが示すずれに安定して追従しなかった。これは、生成モデルで生成される音信号の１個の特徴量である音高を、音高データおよびベンドデータという別個の２種類のデータで制御しようとしたため、と考察される。

[0043] 続いて、図２の下段に図示される、生成モデルを用いて音信号Ｖを生成する音生成機能について説明する。処理部１２１は、処理部１１３と同様に、再生すべき楽譜データが表す一連の発音単位に基づき、制御データＸを生成して生成部１２２に出力する。制御データＸは、楽譜データの各時刻tにおける発音単位の条件（すなわち、合成されるべき音信号Ｖの条件）を表す。具体的には、制御データＸは、音高データＸ1と開始停止データＸ2とコンテキストデータＸ3とを含む。処理部１１３が生成する音高データＸ1が、参照信号Ｒの目標音高Ｐを表すのに対し、処理部１２１が生成する音高データＸ1は、合成すべき音信号Ｖの目標音高Ｐを表す。ただし、処理部１１３が実行する処理と処理部１２１が実行する処理とは実質的に共通し、処理部１１３が生成する音高データＸ1の形式と処理部１２１が生成する音高データＸ1の形式とは共通する。なお、制御データＸは、さらに、楽器、歌手または奏法など、その他の情報を含んでもよい。

[0044] 生成部１２２は、図４の下段に例示するように、複数の変数が確立された生成モデルを用いて、制御データＸに応じた波形スペクトルの時系列を生成する。生成部１２２は、生成モデルを用いて、フレームごと（時刻t）に、制御データＸに応じた波形スペクトルを示す出力データを推定する。推定される出力データが、波形スペクトルを構成する複数の成分の各々の確率密度分

布を表す場合、生成部１２２は、その成分の確率密度分布に従う乱数を生成して、当該乱数を波形スペクトルの成分値として出力する。推定される出力データが複数の成分の値を表す場合は、その成分値を出力する。

[0045] 合成部１２３は、周波数領域の波形スペクトルの時系列を受け取り、その波形スペクトルの時系列に応じた時間領域の音信号Ｖを合成する。合成部１２３は、いわゆるボコーダである。例えば、合成部１２３は、波形スペクトルから最小位相スペクトルを求めて、それら波形スペクトルと位相スペクトルとに対して逆フーリエ変換を実行することで音信号Ｖを合成する。或いは、波形スペクトルと音信号Ｖの間に潜在する関係を学習したニューラルボコーダを用いて、波形スペクトルから音信号Ｖを直接的に合成する。

[0046] 図７は、各発音単位の音生成処理のフローチャートである。この音生成処理は、例えば音信号合成システム１００の利用者からの指示を契機として、時刻ｔ毎に、当該時刻に対応するフレームの音信号Ｖを生成するために実行される。時刻tの進行速度は、実時間の進行速度と同じ位でも良いし、速いあるいは遅い、つまり、実時間と異なっていてもよい。

[0047] ある時刻tの音生成処理が開始されると、制御装置１１（処理部１２１）は、楽譜データに基づいて、その時刻tの制御データＸを生成する（Sb1）。次に、制御装置１１（生成部１２２）は、生成モデルを用いて、生成された制御データＸに応じた、その時刻tの音信号Ｖの波形スペクトルを生成する（Sb2）。次に、制御装置１１（合成部１２３）は、生成された波形スペクトルに応じて、その時刻tに対応するフレームの音信号Ｖを合成する（Sb3）。以上の処理が、楽譜データの各時刻tについて順次行われることで、楽譜データに対応する音信号Ｖが生成される。

[0048] 第１実施形態においては、１個の音高データＸ1が、合成されるべき音信号Ｖの目標音高Ｐと当該音名の基準音高Ｑとの音高差に応じたずれ値とを指定する。そして、生成部１２２は、この音高データＸ1を含む制御データＸを入力とする生成モデルを用いて、音高データＸ1で指定された音名とずれ値とに応じた音高の音信号Ｖを生成する。したがって、生成される音信号Ｖの音高

は、音高データＸ1が指定する音名と、その音名の基準音高Ｑからのずれ値との変化に良く追従する。例えば、音高データＸ1が示すずれ値を動的に変化させることにより、生成する音信号Ｖに、例えばビブラートまたはピッチベンドなどの、動的なピッチ変化を付与できる。

[0049] Ｂ：第２実施形態

第２実施形態では、生成モデルへの入力として、第１実施形態の１ホットレベル記法の音高データＸ1の代わりに、図８に例示する２ホットレベル記法の音高データＸ1を用いる。第２実施形態の音信号合成システム１００の構成、及び、制御装置１１の機能構成は、基本的に第１実施形態と同じである。

[0050] ２ホットレベル記法の音高データＸ1においては、相異なる音名に対応するＭ個の音名データのうち、音信号（参照信号Ｒまたは音信号Ｖ）の目標音高Ｐに対応する２個の有効音名データの各々に、当該有効音名データの音名に対応する基準音高Ｑと目標音高Ｐとの音高差に応じたホット値が設定される。処理部１１３または処理部１２１（制御装置１１）は、音高データＸ1のＭ個の音名データのうち、音信号（参照信号Ｒまたは音信号Ｖ）の目標音高Ｐを挟む２個の基準音高Ｑの各々に対応する音名データを有効音名データとして選択し、２個の有効音名データの各々を、目標音高Ｐと当該音名データの音名に対応する基準音高Ｑとの近接度（ホット値の一例）に設定する。すなわち、２ホットレベル記法とは、音高データＸ1を構成するＭ個の音名データのうちの２個の有効音名データをホット値（近接度）に設定し、残余の(Ｍ－２)個の音名データをコールド値（例えば０）に設定する記述方法である。

[0051] 図８の上段には、時間軸（横軸）と音高軸（縦軸）とが設定された２次元平面に、楽譜データが表す楽譜を構成する音符の時系列と、当該楽譜を演奏した演奏音の音高（目標音高Ｐ）とが図示されている。図８から分かるように、音高データＸ1を構成するＭ個の音名データのうち、目標音高Ｐに１番近い基準音高Ｑに対応する音名の音名データと、２番目に近い基準音高Ｑに対応する音名の音名データとが、有効音名データとして選択される。

[0052] 図８の中段に、各音名データの音名に対応する基準音高Ｑと目標音高Ｐと

の近接度を示す。ここで、近接度は、０から１の範囲の何れかの値を取る。具体的には、目標音高Ｐとある音名の基準音高Ｑとが一致するときに近接度は１である。また、目標音高Ｐとその音名の基準音高Ｑとの音高差がｘセントのとき、近接度は（１００－ｘ）／１００である。すなわち、目標音高Ｐと基準音高Ｑとの音高差が大きいほど近接度は小さい数値となる。例えば、目標音高Ｐがある音名の基準音高Ｑから半音以上離れると、近接度は０となる。

[0053] 図８の時刻ｔ3においては、目標音高Ｐが、音名Ｇに対応する基準音高Ｑ(G)と音名Ｆ＃に対応する基準音高Ｑ(F#)との間に位置する。したがって、音高データＸ1のＭ個の音名データのうち、音名Ｇに対応する音名データと音名Ｆ＃に対応する音名データとが有効音名データとして選択される。時刻ｔ3において、基準音高Ｑ(G)と目標音高Ｐとの音高差は５０セントであるから、音名Ｇに対応する有効音名データは、近接度０．５に設定される。また、時刻ｔ3において、基準音高Ｑ(F#)と目標音高Ｐとの音高差も５０セントであるから、音名Ｆ＃に対応する有効音名データも、近接度０．５に設定される。以上の通り、第２実施形態の処理部１１３または処理部１２１は、時刻ｔ3において、音高データＸ1を構成するＭ個の音名データのうち、音名Ｇに対応する有効音名データを０．５に設定し、音名Ｆ＃に対応する有効音名データを０．５に設定し、残余の(Ｍ－２)個の音名データを０（コールド値）に設定する。

[0054] 他方、図８の時刻ｔ4においては、目標音高Ｐが、音名Ｆに対応する基準音高Ｑ(F)と音名F#に対応する基準音高Ｑ(F#)との間に位置する。したがって、音高データＸ1のＭ個の音名データのうち音名Ｆに対応する音名データと音名Ｆ＃に対応する音名データとが有効音名データとして選択される。時刻ｔ4において、基準音高Ｑ(F)と目標音高Ｐとの音高差は８０セントであるから、音名Ｆに対応する有効音名データは、近接度０．２に設定される。また、時刻ｔ4において、基準音高Ｑ（F#）と目標音高Ｐとの音高差は２０セントであるから、音名Ｆ＃に対応する有効音名データは、近接度０．８に設定される。

以上の通り、第２実施形態の処理部１１３または処理部１２１は、時刻ｔ4において、音高データＸ1を構成するＭ個の音名データのうち、音名Ｆに対応する有効音名データを０．２に設定し、音名Ｆ＃に対応する有効音名データを０．８に設定し、残余の(Ｍ－２)個の音名データを０（コールド値）に設定する。

[0055] 訓練部１１４は、２ホットレベル表記の音高データＸ1を含む制御データＸを入力として、その制御データＸに対応する波形スペクトルを示す出力データを生成するよう、生成モデルを訓練する。複数の変数が確立された生成モデルは、複数の発音単位データにおける各制御データＸと、参照信号Ｒの波形スペクトルとの間に潜在する関係を習得している。

[0056] 生成部１２２は、確立された生成モデルを用いて、２ホットレベル表記の音高データＸ1を含む制御データＸに応じた波形スペクトルを時刻t毎に生成する。合成部１２３は、第１実施形態と同様に、生成部１２２が生成した波形スペクトルの時系列に応じた時間領域の音信号Ｖを合成する。

[0057] 第２実施形態においては、生成モデルを用いて、２ホットレベル表記の音高データＸ1が表す目標音高Ｐの変化によく追従する音信号Ｖを生成できる。

[0058] Ｃ：第３実施形態

第２実施形態においては、目標音高Ｐに対応する２個の有効音名データがホット値に設定されるが、音高データＸ1を構成するＭ個の音名データのうちホット値に設定される有効音名データの個数は任意である。第３実施形態においては、第２実施形態の２ホットレベル記法の音高データＸ1の代わりに、図９に例示する４ホットレベル記法の音高データＸ1を生成モデルに対する入力として利用する。第３実施形態の音信号合成システム１００の構成、及び、制御装置１１の機能構成は、基本的に第１実施形態および第２実施形態と同じである。

[0059] ４ホットレベル記法の音高データＸ1においては、相異なる音名に対応するＭ個の音名データのうち、音信号（参照信号Ｒまたは音信号Ｖ）の目標音高Ｐに対応する４個の音名データが有効音名データとして選択される。具体的

には、目標音高Ｐを挟む２個の基準音高Ｑの各々に対応する２個の音名データと、当該２個の音名データに隣合う２個の音名データとが、有効音名データとして選択される。目標音高Ｐに近い４個の音名データが有効音名データとして選択されると換言してもよい。４個の有効音名データの各々は、,目標音高Ｐと当該有効音名データの音名に対応する基準音高Ｑとの近接度（ホット値）に設定される。すなわち、４ホットレベル記法とは、音高データＸ1を構成するＭ個の音名データのうちの４個の音名データをホット値（近接度）に設定し、残余の(Ｍ－４)個の音名データをコールド値（例えば０）に設定する記述方法である。処理部１１３または処理部１２１（制御装置１１）は、以上に説明した音高データＸ1を生成する。

[0060] 図９の上段には、第２実施形態と同様に、時間軸（横軸）と音高軸（縦軸）とが設定された２次元平面に、楽譜データが表す楽譜を構成する音符の時系列と、当該楽譜を演奏した演奏音の音高（目標音高Ｐ）とが図示されている。図９から理解される通り、音高データＸ1を構成するＭ個の音名データのうち、目標音高Ｐに近い４個の基準音高Ｑにそれぞれ対応する４個の音名データが有効音名データとして選択される。

[0061] 図９の中段に、各音名データの音名に対応する基準音高Ｑと目標音高Ｐとの近接度を示す。ここで、近接度は、第２実施形態と同様に、０から１の範囲の何れかの値を取る。具体的には、目標音高Ｐとある音名の基準音高Ｑとが一致するときに近接度は１である。また、目標音高Ｐとその音名の基準音高Ｑとの音高差がｘセントのとき、近接度は（２００－ｘ）／２００である。すなわち、第２実施形態と同様に、目標音高Ｐと基準音高Ｑとの音高差が大きいほど近接度は小さい数値となる。例えば、目標音高Ｐがある音名の基準音高Ｑから全音以上離れると、近接度は０となる。

[0062] 図９の時刻ｔ5においては、目標音高Ｐが、音名Ｇに対応する基準音高Ｑ(G)と音名Ｆ＃に対応する基準音高Ｑ(F#)との間に位置する。したがって、音高データＸ1のＭ個の音名データのうち、音名Ｇおよび音名Ｆ＃と、音名Ｇの高位側に隣合う音名Ｇ＃と、音名Ｆ＃の低位側に隣合う音名Ｆとに対応する４

個の音名データが、有効音名データとして選択される。時刻ｔ5において、基準音高Ｑ(G)と目標音高Ｐとの音高差は５０セントであるから、音名Ｇに対応する有効音名データは近接度０．７５に設定される。同様に、基準音高Ｑ(F#)と目標音高Ｐとの音高差は５０セントであるから、音名Ｆ＃に対応する有効音名データは近接度０．７５に設定される。また、基準音高Ｑ(F)と目標音高Ｐとの音高差は１５０セントであるから、音名Ｆに対応する有効音名データは近接度０．２５に設定される。同様に、基準音高Ｑ(G#)と目標音高Ｐとの音高差は１５０セントであるから、音名Ｇ＃に対応する有効音名データは近接度０．２５に設定される。以上の通り、第３実施形態の処理部１１３または処理部１２１は、時刻ｔ5において、音高データＸ1を構成するＭ個の音名データのうち、音名Ｇおよび音名Ｆ＃に対応する２個の有効音名データを０．５に設定し、音名Ｆおよび音名Ｇ＃に対応する２個の有効音名データを０．２５に設定し、残余の(Ｍ－４)個の音名データを０（コールド値）に設定する。

[0063] 時刻ｔ6においては、目標音高Ｐが、音名Ｆ＃に対応する基準音高Ｑ(F#)と音名Ｆに対応する基準音高Ｑ(F)と音間に位置する。したがって、音高データＸ1のＭ個の音名データのうち、音名Ｆ＃および音名Ｆと、音名Ｆ＃の高位側に隣合う音名Ｇと、音名Ｆの低位側に隣合う音名Ｅとに対応する４個の音名データが、有効音名データとして選択される。時刻ｔ6において、基準音高Ｑ(F#)と目標音高Ｐとの音高差は２５セントであるから、音名Ｆ＃に対応する有効音名データは近接度０．８７５に設定される。基準音高Ｑ(F)と目標音高Ｐとの音高差は７５セントであるから、音名Ｆに対応する有効音名データは近接度０．６２５に設定される。基準音高Ｑ(G)と目標音高Ｐとの音高差は１２５セントであるから、音名Ｇに対応する有効音名データは近接度０．３７５に設定される。また、基準音高Ｑ(E)と目標音高Ｐとの音高差は１７５セントであるから、音名Ｅに対応する有効音名データは近接度０．１２５に設定される。以上の通り、第３実施形態の処理部１１３または処理部１２１は、時刻ｔ6において、音高データＸ1を構成するＭ個の音名データのうち、音名

Ｆ＃に対応する有効音名データを０．８７５に設定し、音名Ｆに対応する有効音名データを０．６２５に設定し、音名Ｇに対応する有効音名データを０．３７５に設定し、音名Ｅに対応する有効音名データを０．１２５に設定し、残余の(Ｍ－４)個の音名データを０（コールド値）に設定する。

[0064] 訓練部１１４は、４ホットレベル表記の音高データＸ1を含む制御データＸを入力として、その制御データＸに対応する波形スペクトルを示す出力データを生成するよう、生成モデルを訓練する。複数の変数が確立された生成モデルは、複数の発音単位データにおける各制御データＸと、参照信号Ｒの波形スペクトルとの間に潜在する関係を習得している。

[0065] 生成部１２２は、確立された生成モデルを用いて、２ホットレベル表記の音高データＸ1を含む制御データＸに応じた波形スペクトルを時刻t毎に生成する。合成部１２３は、第１実施形態と同様に、生成部１２２が生成した波形スペクトルの時系列に応じた時間領域の音信号Ｖを合成する。

[0066] 第３実施形態においては、生成モデルを用いて、４ホットレベル表記の音高データＸ1が表す目標音高Ｐの変化によく追従する音信号Ｖを生成できる。

[0067] 第１実施形態に例示した１ホットレベル記法と、第２実施形態に例示した２ホットレベル記法と、第３実施形態に例示した４ホットレベル記法とは、音高データＸ1内の有効音名データの個数をＮとしたＮホットレベル記法として一般化される（Ｎは１以上の自然数）。Ｎホットレベル記法においては、音高データＸ1を構成するＭ個の音名データのうち、目標音高Ｐに対応するＮ個の有効音名データが、当該音名の基準音高Ｑと目標音高Ｐとの音高差に応じたホット値（ずれ値または近接度）に設定され、残余の(Ｍ－Ｎ)個の音名データがコールド値（例えば０）に設定される。目標音高Ｐとある音名の基準音高Ｑとの音高差がｘセントであるとき、近接度は(５０×Ｎ－ｘ)／５０×Ｎと表現される。ただし、近接度を算定するための演算式は以上の例示に限定されない。以上の通り、目標音高Ｐを表現するために使用される有効音名データの個数Ｎは任意である。

[0068] Ｄ：第４実施形態

第１、第２、および第３実施形態の生成部１２２は波形スペクトルを生成したが、第４実施形態では、生成部１２２が、生成モデルを用いて音信号Ｖを生成する。第４実施形態の機能構成は、図２と基本的に同じだが、合成部１２３は不要である。訓練部１１４は、参照信号Ｒを用いて生成モデルを訓練し、生成部１２２はその生成モデルを用いて音信号Ｖを生成する。第４実施形態における訓練用の発音単位データは、各発音単位の制御データＸと参照信号Ｒの部分波形（すなわち参照信号Ｒのサンプル）のセットである。

[0069] 第４実施形態の訓練部１１４は、訓練データを受け取り、その各バッチの複数の発音単位の部分波形と制御データＸとを順番に用いて生成モデルを訓練する。生成モデルは、サンプリング周期（時刻ｔ）毎に、音信号Ｖのサンプルを表す出力データを推定する。訓練部１１４は、制御データＸから推定された出力データの時系列と訓練データのうち対応する部分波形とに基づいて損失関数Ｌ（１バッチ分の累算値）を計算し、その損失関数Ｌが最小化されるように生成モデルの複数の変数を最適化する。こうして確立された生成モデルは、複数の発音単位データにおける各制御データＸと、参照信号Ｒの部分波形との間に潜在する関係を学習している。

[0070] 第４実施形態の生成部１２２は、確立された生成モデルを用いて、制御データＸに応じた音信号Ｖを生成する。生成部１２２は、生成モデルを用いて、サンプリング周期ごと（時刻ｔ）に、制御データＸに応じた音信号Ｖのサンプルを示す出力データを推定する。出力データが複数のサンプルの各々の確率密度分布を表す場合は、生成部１２２は、その成分の確率密度分布に従う乱数を生成して、当該乱数を音信号Ｖのサンプルとして出力する。出力データがサンプルの値を表す場合には、そのサンプルの時系列が音信号Ｖとして出力される。

[0071] Ｅ：第５実施形態

図２に図示される実施形態の音生成機能では、楽譜データの一連の発音単位の情報に基づいて、音信号Ｖを生成していたが、鍵盤等から供給される発音単位の情報に基づいて、リアルタイムに音信号Ｖを生成するようにしても

よい。その場合、処理部１２１は、各時点の制御データＸを、その時点までに供給された発音単位の情報に基づいて生成する。ここでは、制御データＸに含まれるコンテキストデータＸ3には、基本的に、未来の発音単位の情報を含むことができないが、過去の情報から未来の発音単位の情報を予測して、未来の発音単位の情報を含めるようにしてもよい。

[0072] Ｆ：変形例

第１実施形態では、音高データＸ1を構成するＭ個の音名データのうち、目標音高Ｐに近い基準音高Ｑに対応する音名の音名データを有効音名データとして選択したが、目標音高Ｐから遠い基準音高Ｑに対応する音名の音名データを有効音名データとして選択してもよい。その場合、第１実施形態のずれ値は、±５０セントを超える音高差を表現できるようにスケーリングされる。

[0073] 第２および第３実施形態では、目標音高Ｐと基準音高Ｑとの近接度は、セントスケールにおける両者間の音高差に応じて０から１までの範囲内においてリニアに変化したが、代わりに、図１０のような正規分布などの確率分布またはコサインカーブ等の任意のカーブ、または、折れ線に従って１から０に減少してもよい。

[0074] 第１および第２実施形態では、セントスケールで音名の対応付けを行っていたが、ヘルツスケールなどのその他音高を表現する任意のスケールで対応付けを行っても良い。その際、ずれ値は各スケール上における適切な値を挿入する。

[0075] 第１、第２実施形態では、処理部１１３または処理部１２１において０から１にスケーリングしていたが、スケーリングについては任意の値で行っても良い。例えば、－１から＋１にスケーリングしてもよい。

[0076] なお、音信号合成システム１００が合成する音信号は、楽器音または音声に限らず、動物の鳴き声、または風音および波音のような自然界の音であっても、その音高の動的な制御を行いたい場合は、本開示を適用できる。

符号の説明

[0077] １００…音信号生成装置、１１…制御装置、１２…記憶装置、１３…表示装置、１４…入力装置、１５…放音装置、１１１…解析部、１１２…時間合わせ部、１１３…処理部、１１４…訓練部、１２１…処理部、１２２…生成部、１２３…合成部。

請求の範囲

[請求項1] 合成すべき第１音信号の音高を表す第１音高データを生成し、

第２音信号の音高を表す第２音高データと前記第２音信号との関係を学習した生成モデルを用いて、前記第１音高データに応じた前記第１音信号を示す出力データを推定する

コンピュータにより実現される音信号合成方法であって、

前記第１音高データは、相異なる音名に対応する複数の音名データを含み、

前記第１音高データの生成においては、前記複数の音名データのうち、前記第１音信号の音高に対応する音名データを、当該音名データに対応する音名の基準音高と当該第１音信号の音高との相違に応じたホット値に設定する

音信号合成方法。

[請求項2] 前記第１音高データの生成においては、前記複数の音名データのうち、前記第１音信号の音高に対応する音名データ以外の音名データを、前記第１音信号の音高に関与しないことを表すコールド値に設定する

請求項１の音信号合成方法。

[請求項3] 前記第１音信号の音高は動的に変化し、

前記第１音高データは、前記第１音信号において動的に変化する音高を表す

請求項１または請求項２の音信号合成方法。

[請求項4] 前記第２音信号の音高は動的に変化し、

前記第２音高データは、前記第２音信号において動的に変化する音高を表し、

前記生成モデルは、前記第２音信号と前記第２音高データとを用いて訓練されている

請求項１から請求項３の何れかの音信号合成方法。

[請求項5] 前記第１音信号の音高は、一の音名に対応する発音期間内において動的に変化し、前記第１音信号の音高に対応する音名データに設定されるホット値は、当該音高に応じて変化する

請求項１から請求項３の何れかの音信号合成方法。

[請求項6] 前記第１音信号の音高に対応する音名データは、相異なる音名に対応する複数の単位範囲のうち、当該音高を含む１個の単位範囲に対応する音名の音名データである

請求項１から請求項５の何れかの音信号合成方法。

[請求項7] 前記第１音信号の音高に対応する音名データは、相異なる音名に対応する複数の基準音高のうち、当該音高を挟む２個の基準音高にそれぞれ対応する２個の音名データである

請求項１から請求項５の何れかの音信号合成方法。

[請求項8] 前記第１音信号の音高に対応する音名データは、相異なる音名に対応する複数の基準音高のうち、当該音高に近いＮ個（Ｎは１以上の自然数）の基準音高にそれぞれ対応するＮ個の音名データである

請求項１から請求項５の何れかの音信号合成方法。

[請求項9] 前記推定される出力データは、前記第１音信号の波形スペクトルに関する特徴量を表す

請求項１から請求項８の何れかの音信号合成方法。

[請求項10] 前記推定される出力データは、前記第１音信号のサンプルを表す

請求項１から請求項８の何れかの音信号合成方法。

[請求項11] 合成すべき音信号の音高を表す音高データを用意し、

前記音高データの入力に対して前記音信号を示す出力データが出力されるように生成モデルを訓練する

コンピュータにより実現される生成モデルの訓練方法であって、

前記音高データは、相異なる音名に対応する複数の音名データを含み、

前記音高データの用意においては、前記複数の音名データのうち、

前記音信号の音高に対応する音名データを、当該音名データに対応する音名の基準音高と当該音信号の音高との相違に応じたホット値に設定する

訓練方法。

[請求項12] １以上のプロセッサと１以上のメモリとを具備する音信号合成システムであって、

前記１以上のメモリは、第２音信号の音高を表す第２音高データと前記第２音信号との関係を学習した生成モデルを記憶し、

前記１以上のプロセッサは、

合成すべき第１音信号の音高を表す第１音高データを生成し、

前記生成モデルを用いて、前記第１音高データに応じた前記第１音信号を示す出力データを推定し、

前記第１音高データは、相異なる音名に対応する複数の音名データを含み、

前記１以上のプロセッサは、前記第１音高データの生成において、前記複数の音名データのうち、前記第１音信号の音高に対応する音名データを、当該音名データに対応する音名の基準音高と当該第１音信号の音高との相違に応じたホット値に設定する

音信号合成システム。

[請求項13] 合成すべき第１音信号の音高を表す第１音高データを生成する処理部、および、

第２音信号の音高を表す第２音高データと前記第２音信号との関係を学習した生成モデルを用いて、前記第１音高データに応じた前記第１音信号を示す出力データを推定する生成部

としてコンピュータを機能させるプログラムであって、

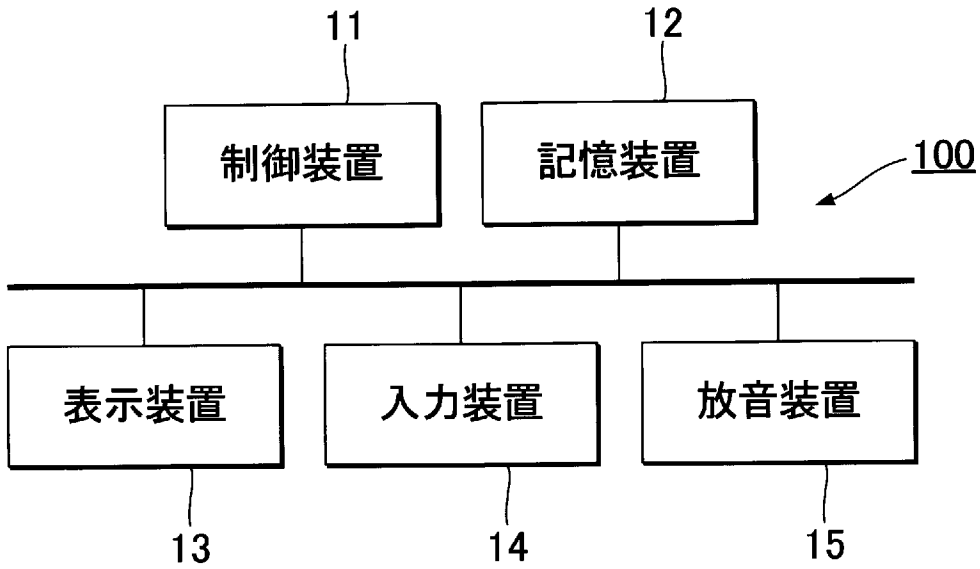
前記第１音高データは、相異なる音名に対応する複数の音名データを含み、

前記第１音高データの生成においては、前記複数の音名データのう

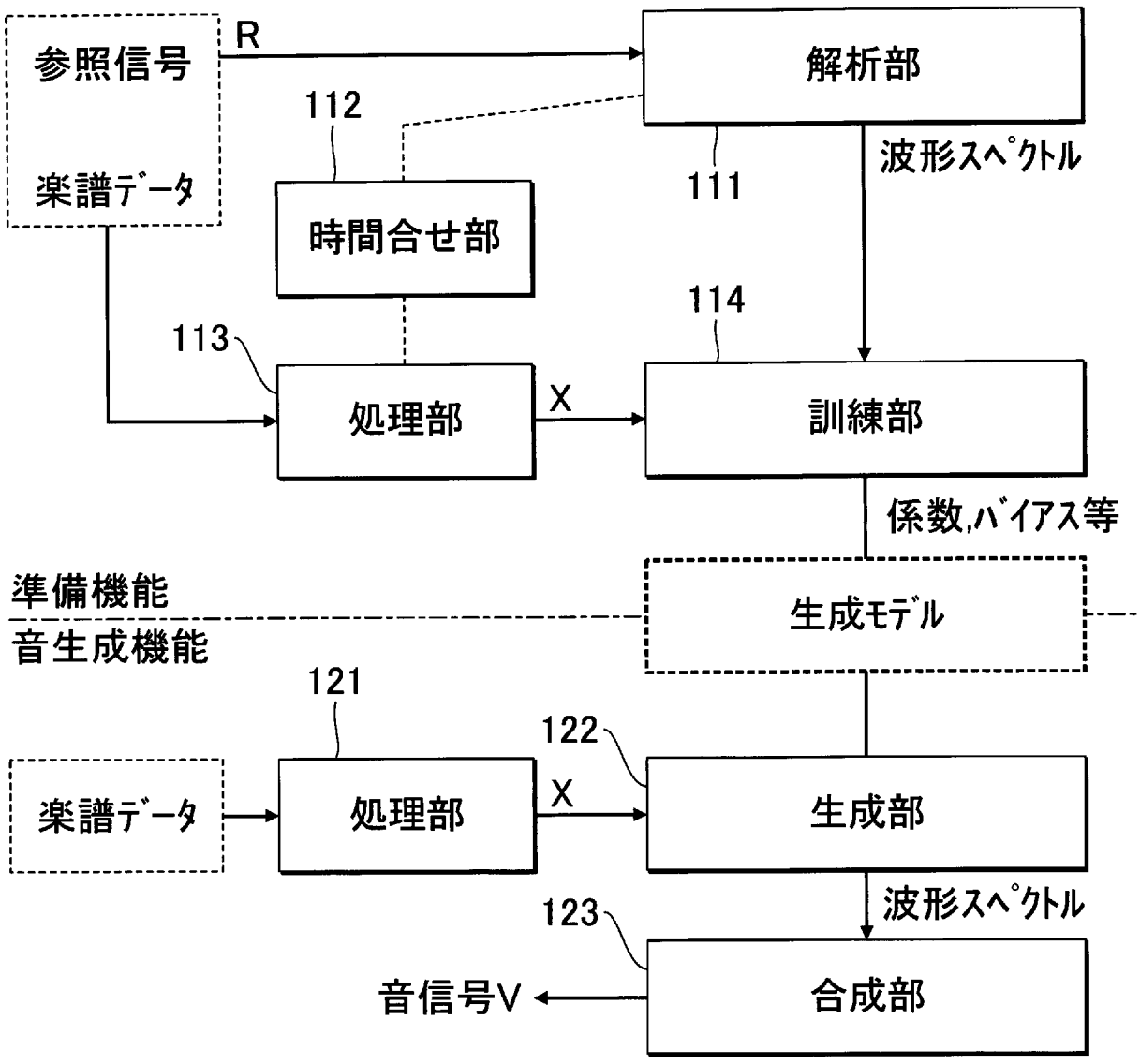
ち、前記第１音信号の音高に対応する音名データを、当該音名データに対応する音名の基準音高と当該第１音信号の音高との相違に応じたホット値に設定する

プログラム。

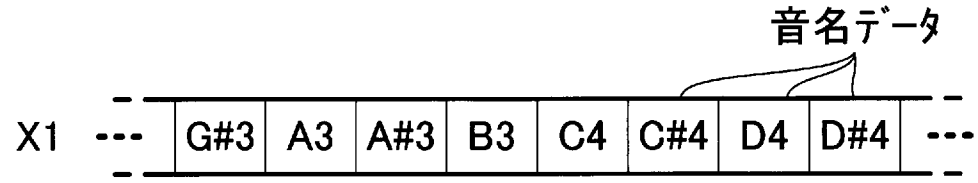
[図1]


11
12
制御装置
記憶装置
100
表示装置
入力装置
放音装置
13
14
15

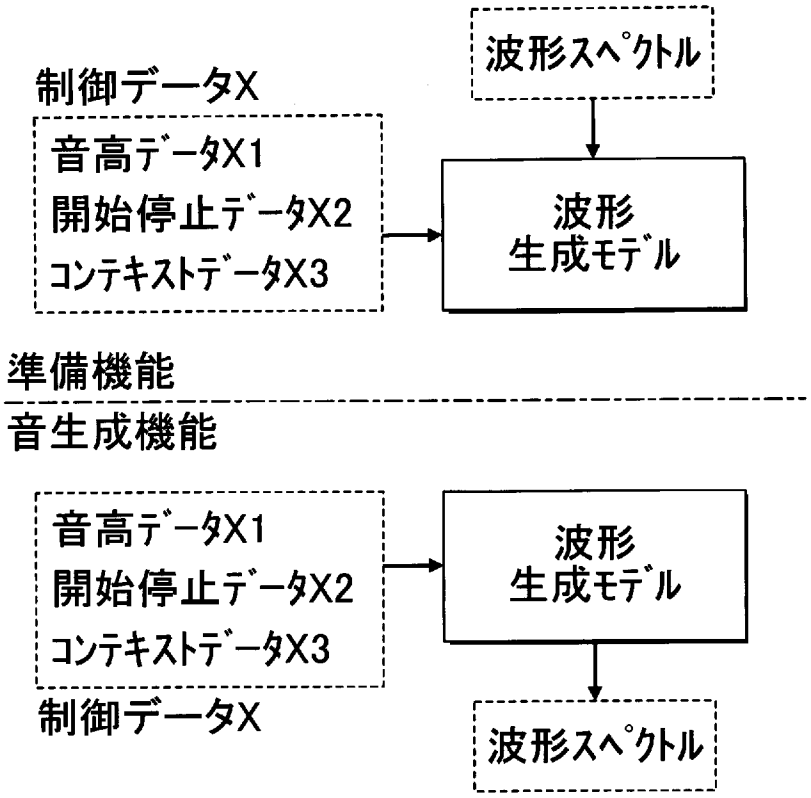
[図2]


参照信号
楽譜ﾃﾞｰﾀ
R
解析部
112
111
波形ｽﾍﾟｸﾄﾙ
時間合せ部
114
113
処理部
X
訓練部
係数,ﾊﾞｲｱｽ等
準備機能
音生成機能
生成ﾓﾃﾞﾙ
121
122
楽譜ﾃﾞｰﾀ
処理部
X
生成部
波形ｽﾍﾟｸﾄﾙ
123
音信号V
合成部

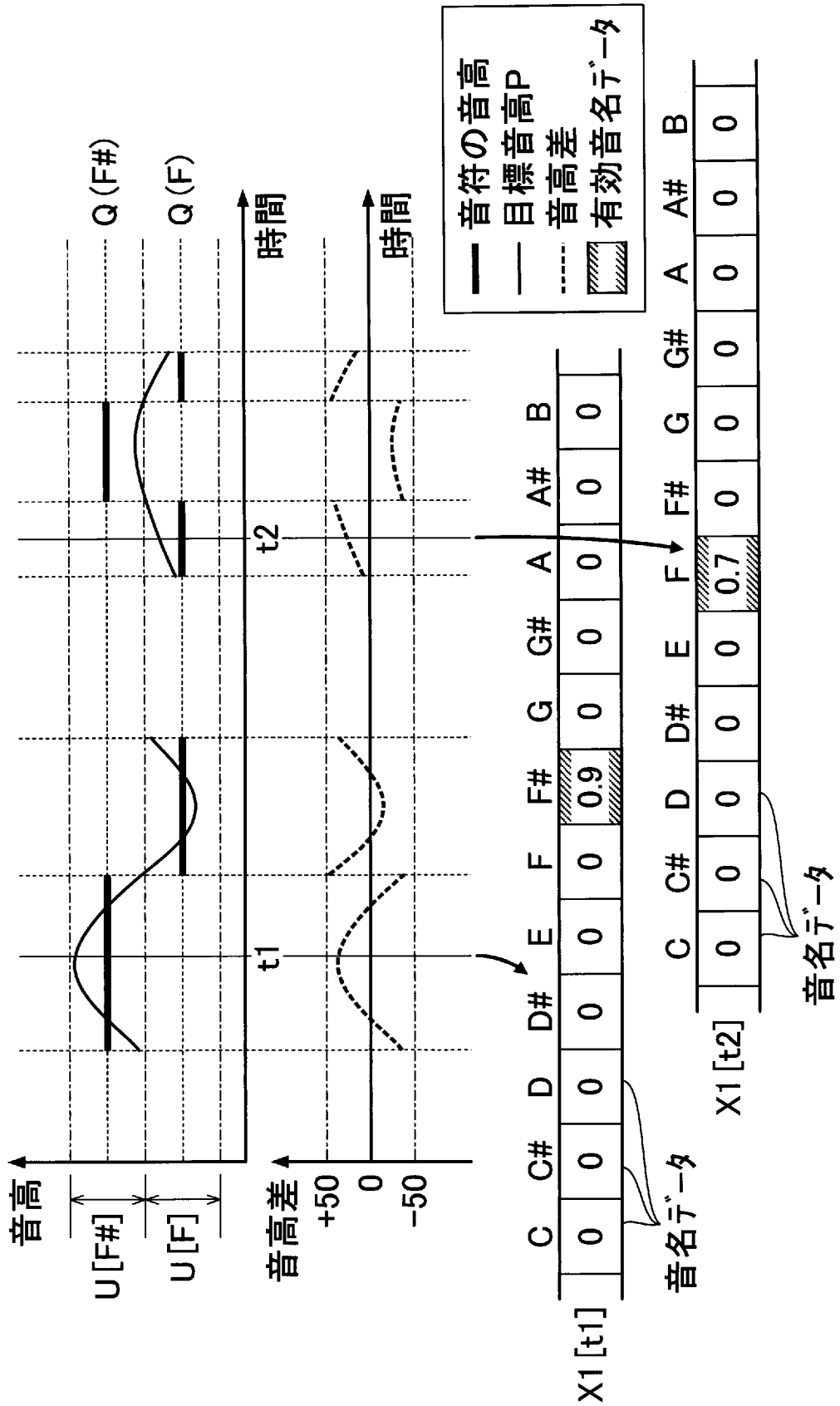
[図3]


音名データ
X1
G#3
A3
A#3
B3
C4
C#4
D4
D#4

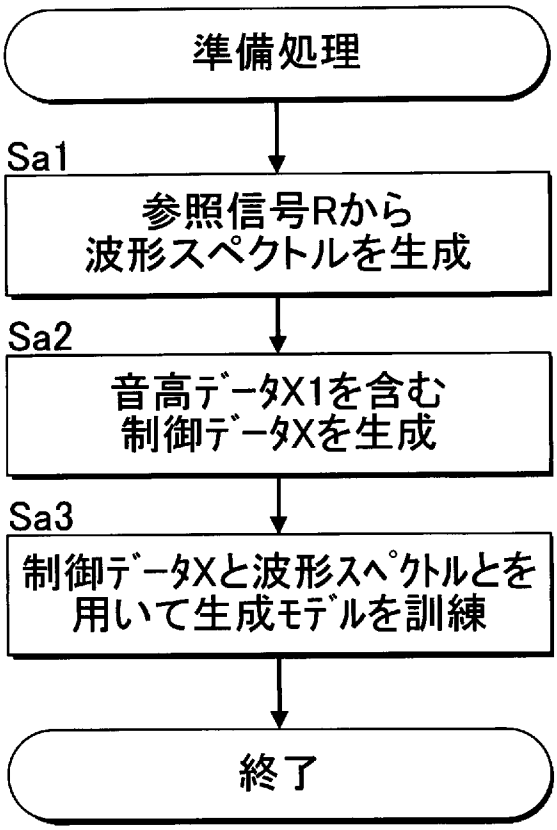
[図4]


制御データX
音高データX1
開始停止データX2
コンテキストデータX3
波形スペクトル
波形
生成モデル
準備機能
音生成機能
音高データX1
開始停止データX2
コンテキストデータX3
制御データX
波形
生成モデル
波形スペクトル

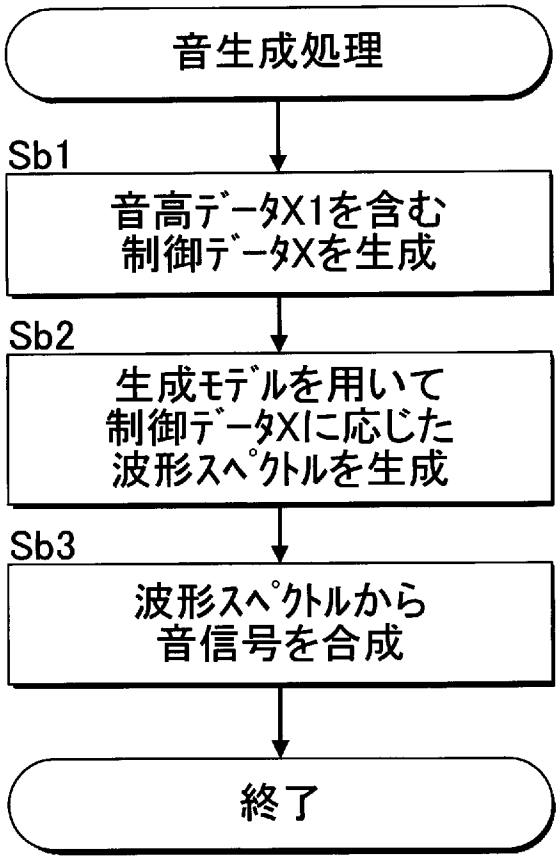
[図5]


音符の音高
目標音高P
音高差
有効音名データ
Q(F#)
Q(F)
時間
時間
音高
U[F#]
U[F]
音高差
+50
0
−50
t1
t2
C C# D D# E F F# G G# A A# B
X1[t1]
0 0 0 0 0 0 0.9 0 0 0 0 0
音名データ
C C# D D# E F F# G G# A A# B
X1[t2]
0 0 0 0 0 0 0.7 0 0 0 0 0
音名データ

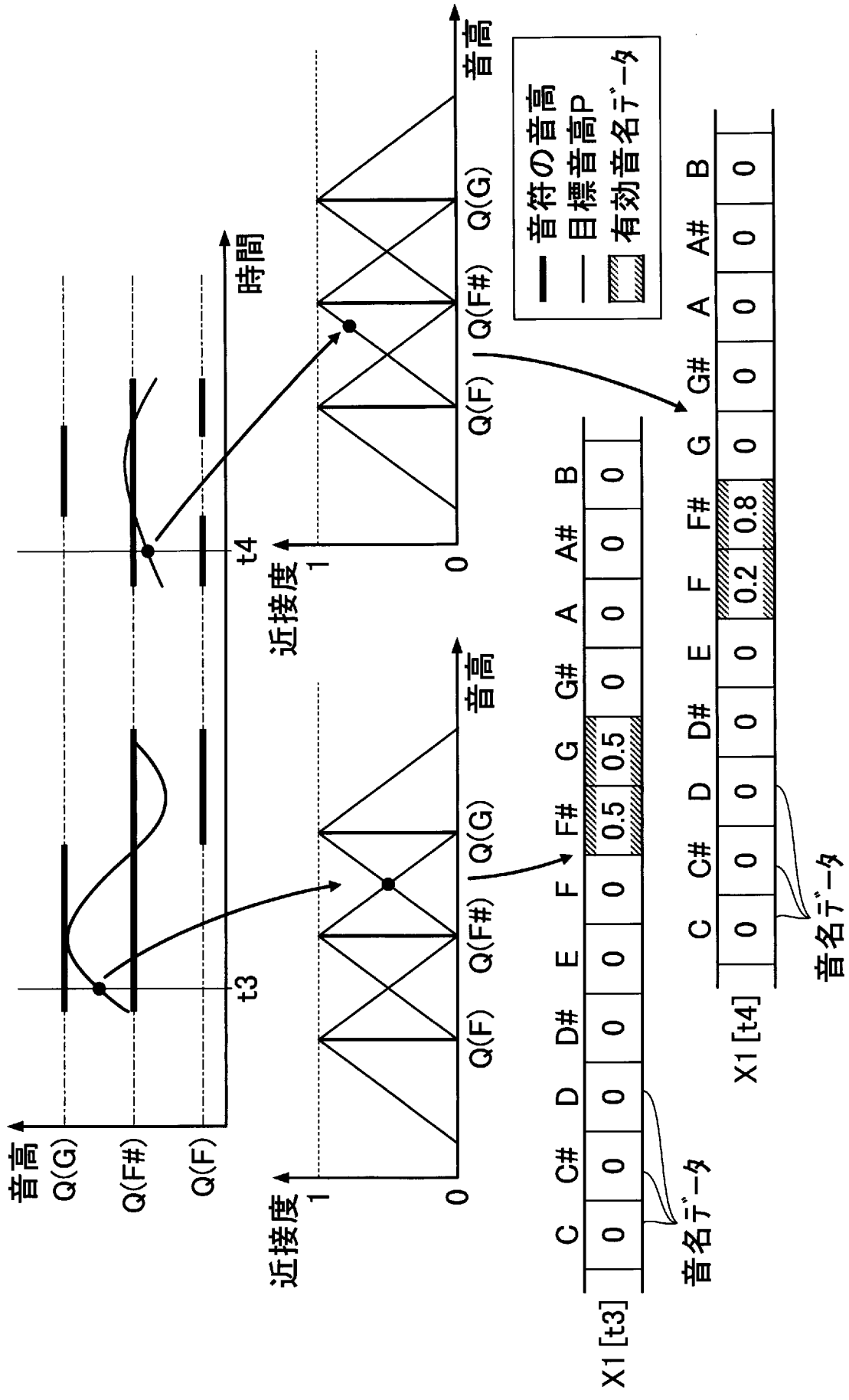
[図6]


準備処理
Sa1
参照信号Rから
波形スペクトルを生成
Sa2
音高データX1を含む
制御データXを生成
Sa3
制御データXと波形スペクトルとを
用いて生成モデルを訓練
終了

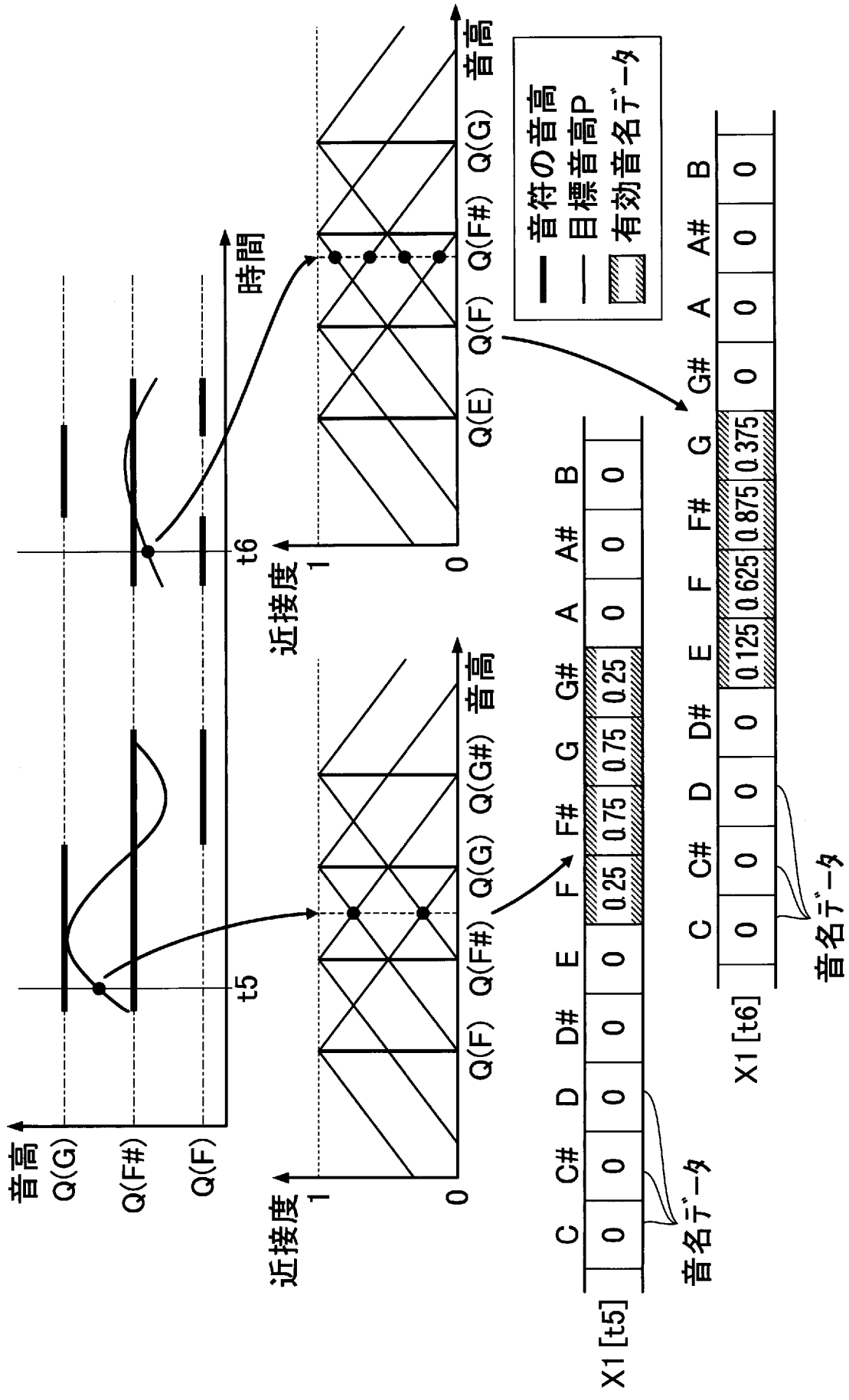
[図7]


音生成処理
Sb1
音高ﾃﾞｰﾀX1を含む
制御ﾃﾞｰﾀXを生成
Sb2
生成ﾓﾃﾞﾙを用いて
制御ﾃﾞｰﾀXに応じた
波形ｽﾍﾟｸﾄﾙを生成
Sb3
波形ｽﾍﾟｸﾄﾙから
音信号を合成
終了

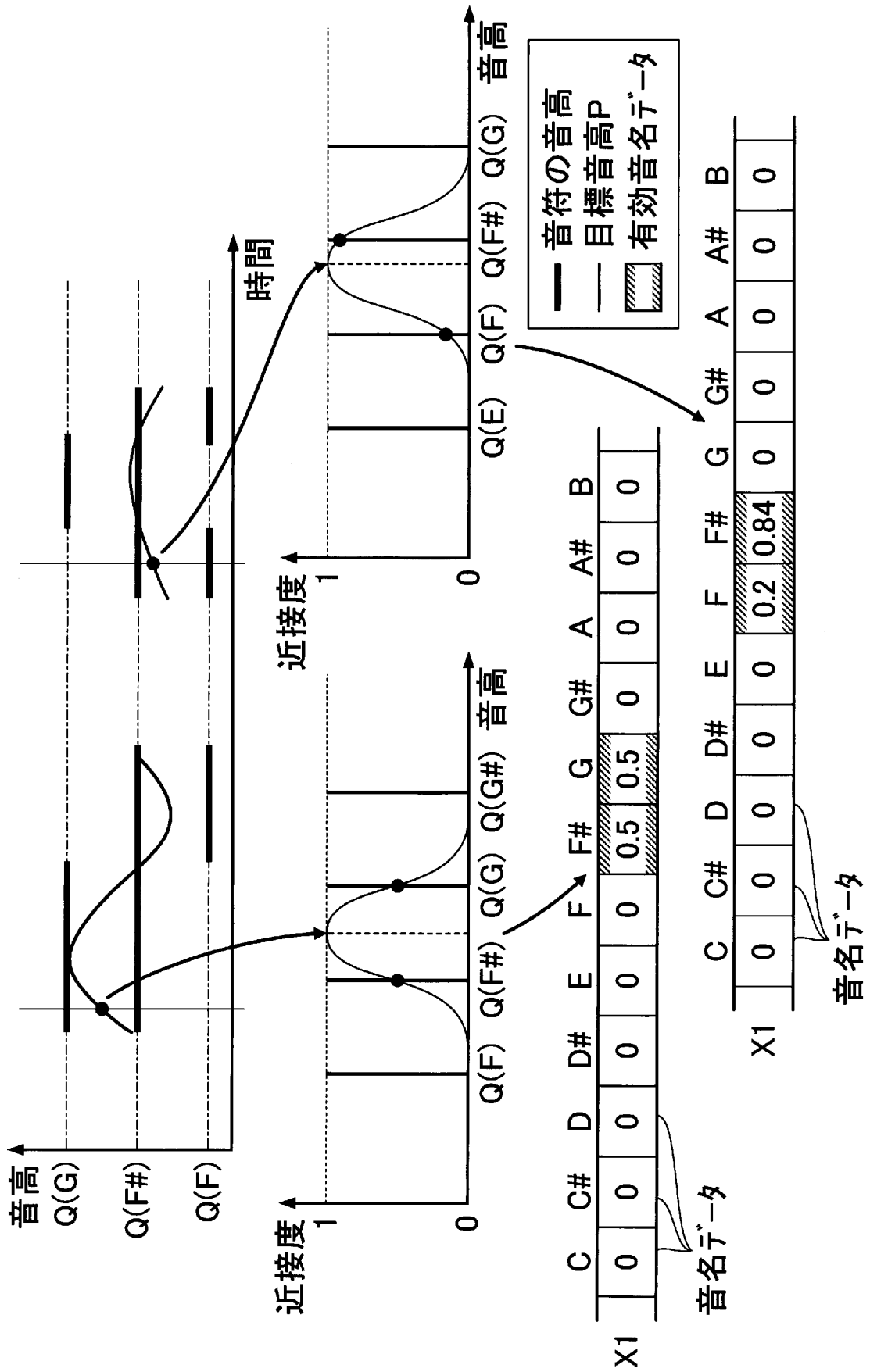
[図8]


音高
Q(G)
Q(F#)
Q(F)
時間
t3
t4
近接度
1
0
音高
Q(F)
Q(F#)
Q(G)
音符の音高
目標音高P
有効音名ﾃﾞｰﾀ
C C# D D# E F F# G G# A A# B
X1 [t3]
0 0 0 0 0 0 0.5 0.5 0 0 0 0
音名ﾃﾞｰﾀ
X1 [t4]
0 0 0 0 0 0 0.2 0.8 0 0 0 0
音名ﾃﾞｰﾀ

[図9]


音高
Q(G)
Q(F#)
Q(F)
時間
t5
t6
近接度
1
0
Q(F) Q(F#) Q(G) Q(G#) 音高
Q(E) Q(F) Q(F#) Q(G) 音高
音符の音高
目標音高P
有効音名データ
C C# D D# E F F# G G# A A# B
X1 [t5]
0 0 0 0 0 0.25 0.75 0.75 0.25 0 0 0
音名データ
X1 [t6]
0 0 0 0 0.125 0.625 0.875 0.375 0 0 0 0

[図10]


音高
Q(G)
Q(F#)
Q(F)
時間
近接度
1
0
Q(F) Q(F#) Q(G) Q(G#) 音高
Q(E) Q(F) Q(F#) Q(G) 音高
音符の音高
目標音高P
有効音名ﾃﾞｰﾀ
C C# D D# E F F# G G# A A# B
X1
0 0 0 0 0 0 0.5 0.5 0 0 0 0
0 0 0 0 0 0 0.2 0.84 0 0 0 0
音名ﾃﾞｰﾀ

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2020/006162

A. CLASSIFICATION OF SUBJECT MATTER
G10L 13/00(2006.01)i; G10H 7/08(2006.01)i
FI: G10H7/08; G10L13/00 100Y

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G10L13/00; G10H7/08

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan	1922-1996
Published unexamined utility model applications of Japan	1971-2020
Registered utility model specifications of Japan	1996-2020
Published registered utility model applications of Japan	1994-2020

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	"How to use LSTM neural network with 3D categorical input and output?", STACK OVERFLOW, [online], 22 May 2018, [retrieved on 12 May 2020], <URL:https://stackoverflow.com/questions/50474524/how-to-use-lstm-neural-network-with-3d¬categorical-input-and-output?rq=1>, in particular, see "Answer"	1-13
A	BLAAUW, Merlijn et al., "A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs", Applied Sciences, [online], December 2017, 1313, pp. 1-23, [retrieved on 12 May 2020], <URL:https://pdfs.semanticsholar.org/0376/966eldee6ead97730f2f8ce55df936a7111a.pdf>, pp. 2-9, 19-20	1-13

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 May 2020 (12.05.2020)	26 May 2020 (26.05.2020)

Name and mailing address of the ISA/	Authorized officer
Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2020/006162

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 10068557 B1 (GOOGLE LLC) 04.09.2018 (2018-09-04) column 5, line 54 to column 10, line 21	1-13
A	西村 方成，外 4 名，Deep Neural Network に基づく歌声合成の検討，日本音響学会 2016 年春季研究発表会講演論文集，[CD-ROM], 24 February 2016, pp. 213-214, entire text, all drawings, (NISHIMURA, Masanari et al., "Investigation of singing voice synthesis based on deep neural networks", Lecture proceedings of 2016 spring research conference of the acoustical Society of Japan)	1-13

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/JP2020/006162

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
US 10068557 B1	04 Sep. 2018	WO 2019/040132 A1 paragraphs [0067]-[0105]	

国際調査報告

国際出願番号
PCT/JP2020/006162

A. 発明の属する分野の分類（国際特許分類（ＩＰＣ））

G10L 13/00(2006.01)i; G10H 7/08(2006.01)i
FI: G10H7/08; G10L13/00 100Y

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

G10L13/00; G10H7/08

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０２０年
日本国実用新案登録公報	１９９６－２０２０年
日本国登録実用新案公報	１９９４－２０２０年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献のカテゴリー*	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	How to use LSTM neural network with 3D categorical input and output?, STACK OVERFLOW, [online], 2018.05.22, [retrieved on 2020.05.12], <URL: https://stackoverflow.com/questions/50474524/how-to-use-lstm-neural-network-with-3d-categorical-input-and-output?rq=1> 特に、「Answer」を参照	1-13
A	Merlijn Blaauw et al., A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs, Applied Sciences, [online], 2017.12, 1313, pp. 1-23, [retrieved on 2020.05.12], <URL: https://pdfs.semanticscholar.org/0376/966e1dee6ead97730f2f8ce55df936a7111a.pdf> 第2-9, 19-20ページ	1-13
A	US 10068557 B1 (GOOGLE LLC) 04.09.2018 (2018-09-04) 第5欄第54行 - 第10欄第21行	1-13
A	西村 方成, 外4名, Deep Neural Networkに基づく歌声合成の検討, 日本音響学会2016年春季研究発表会講演論文集, [CD-ROM], 2016.02.24, 第213-214ページ 全文 全図	1-13

☐ C欄の続きにも文献が列挙されている。

☑ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー
"A" 特に関連のある文献ではなく、一般的技術水準を示すもの
"E" 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
"L" 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
"O" 口頭による開示、使用、展示等に言及する文献
"P" 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
"T" 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの
"X" 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
"Y" 特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
"&" 同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
12.05.2020	26.05.2020

名称及びあて先

日本国特許庁(ISA/JP)
〒100-8915
日本国
東京都千代田区霞が関三丁目４番３号

権限のある職員（特許庁審査官）

岩田　淳 5Z 4052

電話番号 03-3581-1101 内線 3591

国際調査報告
パテントファミリーに関する情報

国際出願番号
PCT/JP2020/006162

引用文献	公表日	パテントファミリー文献	公表日
US 10068557 B1	04.09.2018	WO 2019/040132 A1 [0067]-[0105]	