(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
2021年1月14日(14.01.2021)

WIPO | PCT

(10) 国際公開番号
WO 2021/006117 A1

(51) 国際特許分類:
G10L 13/08 (2013.01) *G10L 25/30* (2013.01)
G10L 13/10 (2013.01)

(21) 国際出願番号： PCT/JP2020/025682

(22) 国際出願日： 2020年6月30日(30.06.2020)

(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 2019-125726 2019年7月5日(05.07.2019) JP
特願 2019-200440 2019年11月5日(05.11.2019) JP

(71) 出願人: 国立研究開発法人情報通信研究機構 (**NATIONAL INSTITUTE OF INFORMATION AND COMMUNICATIONS TECHNOLOGY**) [JP/JP]; 〒1848795 東京都小金井市貫井北町４－２－１ Tokyo (JP).

(72) 発明者: 岡本 拓磨 (**OKAMOTO Takuma**); 〒1848795 東京都小金井市貫井北町４－２－１ 国立研究開発法人情報通信研究機構内 Tokyo (JP). 戸田 智基(**TODA Tomoki**); 〒4801103 愛知県長久手市岩作三ヶ峯３８－１ Aichi (JP). 志賀 芳則(**SHIGA Yoshinori**); 〒1848795 東京都小金井市貫井北町４－２－１ 国立研究開発法人情報通信研究機構内 Tokyo (JP). 河井 恒(**KAWAI Hisashi**); 〒1848795 東京都小金井市貫井北町４－２－１ 国立研究開発法人情報通信研究機構内 Tokyo (JP).

(74) 代理人: 中西 健, 外 (**NAKANISHI Ken** et al.); 〒5300001 大阪府大阪市北区梅田二丁目５番６号 桜橋八千代ビル３Ｆ りのわ国際特許事務所 Osaka (JP).

(54) **Title:** VOICE SYNTHESIS PROCESSING DEVICE, VOICE SYNTHESIS PROCESSING METHOD, AND PROGRAM

(54) 発明の名称： 音声合成処理装置、音声合成処理方法、および、プログラム

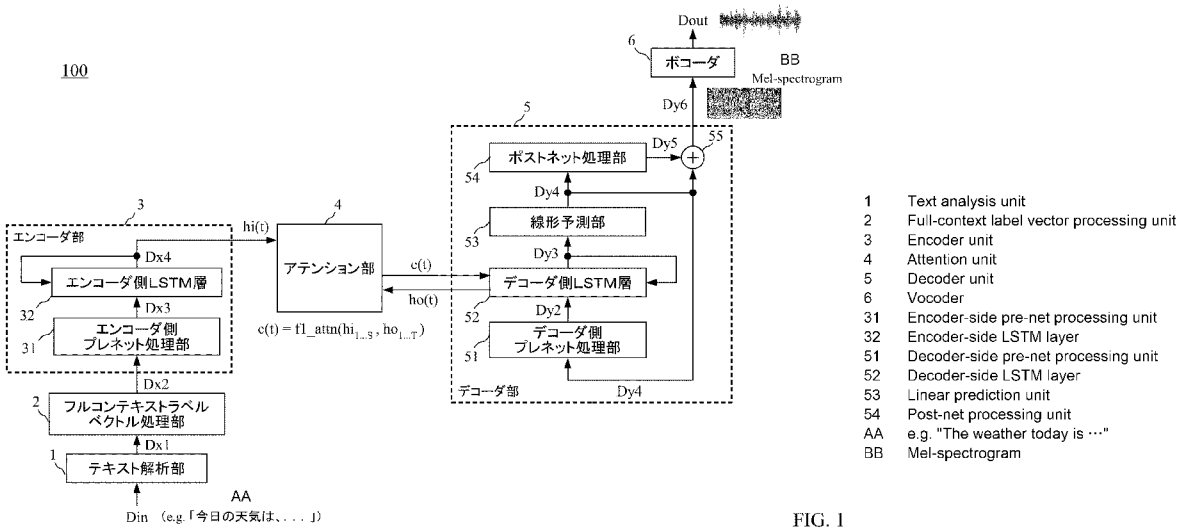


FIG. 1

(57) **Abstract:** The present invention enables a voice synthesis processing device which performs learning and optimization and enables a high-quality voice synthesis process by means of a neural network model that can turn a processing target language into an arbitrarily defined language and that is for text-to-speech synthesis using the sequence-to-sequence method. In a voice synthesis processing device (100), a text analysis process is performed in accordance with a processing target language, optimized full-context label data that is suitable for processing with a neural network model is acquired from full-context label data acquired in the text analysis process, and the acquired optimized full-context label data is used to perform a process, whereby a highly accurate voice synthesis process for an arbitrarily defined processing target language can be performed.

WO 2021/006117 A1

[続葉有]

(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, 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, 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：処理対象言語を任意の言語にできる、sequence-to-sequence方式を用いたテキスト音声合成用のニューラルネットワークのモデルにより、学習・最適化を行い、高品質な音声合成処理を実現する音声合成処理装置を実現する。音声合成処理装置(１００)では、処理対象言語に応じたテキスト解析処理を行い、当該テキスト解析処理で取得されたフルコンテキストラベルデータから、ニューラルネットワークのモデルで処理するのに適した最適化フルコンテキストラベルデータを取得し、取得した最適化フルコンテキストラベルデータを用いて処理を行うことで、任意の処理対象言語について、高精度な音声合成処理を行うことができる。

明　細　書

発明の名称 ：

音声合成処理装置、音声合成処理方法、および、プログラム

技術分野

[0001] 本発明は、音声合成処理技術に関する。特に、テキストを音声に変換するテキスト音声合成（ＴＴＳ：ｔｅｘｔ-ｔｏ-ｓｐｅｅｃｈ）技術に関する。

背景技術

[0002] テキストから自然な音声を合成するテキスト音声合成（ＴＴＳ）技術において、近年、ニューラルネットワークの導入により高品質な音声合成が可能となっている。このようなテキスト音声合成技術を用いたシステムでは、英語音声を合成する場合、音素継続長と音響モデルとを同時に学習・最適化するsequence-to-sequence方式を用いたテキスト音声合成技術により、英語テキストからメルスペクトログラムを推定し、推定したメルスペクトログラムから、ニューラルボコーダにより音声波形を取得する。このように処理することで、上記テキスト音声合成技術を用いたシステムでは、処理対象言語が英語である場合、人間の音声と同等の品質の音声合成が可能となる（例えば、非特許文献１を参照）。

先行技術文献

非特許文献

[0003] 非特許文献1：Jonathan Shen, R Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, ”Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” Proc. ICASSP, Apr. 2018, pp. 4779-4783.

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、上記のsequence-to-sequence方式を用いたテキスト音声合

成を日本語に適用するのは困難である。日本語は、漢字を使う言語であり、漢字の数が膨大であり、また、漢字の読みのバリエーションも多いので、日本語テキストを、sequence-to-sequence方式を用いたテキスト音声合成のモデルに、直接入力し、処理言語を英語としたときと同様に、当該モデルの学習・最適化を行うことは困難である。

[0005] そこで本発明は、上記課題に鑑み、日本語等の英語以外の言語を処理対象言語とする場合においても（処理対象言語を任意の言語にできる）、sequence-to-sequence方式を用いたテキスト音声合成用のニューラルネットワークのモデルにより、学習・最適化を行い、高品質な音声合成処理を実現する音声合成処理装置、音声合成処理方法、および、プログラムを実現することを目的とする。

課題を解決するための手段

[0006] 上記課題を解決するための第１の発明は、任意の言語を処理対象言語とし、エンコーダ・デコーダ方式のニューラルネットワークを用いて音声合成処理を実行する音声合成処理装置であって、テキスト解析部と、フルコンテキストラベルベクトル処理部と、エンコーダ部と、デコーダ部と、を備える。

[0007] テキスト解析部は、処理対象言語のテキストデータに対してテキスト解析処理を実行し、コンテキストラベルデータを取得する。

[0008] フルコンテキストラベルベクトル処理部は、テキスト解析部により取得されたコンテキストラベルデータから、コンテキストラベルデータを取得する処理において処理対象とされた音素である単独音素についてのコンテキストラベルを取得することで、ニューラルネットワークの学習処理に適した最適化フルコンテキストラベルデータを取得する。

[0009] エンコーダ部は、最適化フルコンテキストラベルデータに基づいて、ニューラルネットワークのエンコード処理を実行することで、隠れ状態データを取得する。

[0010] デコーダ部は、隠れ状態データに基づいて、ニューラルネットワークのデコード処理を実行することで、最適化フルコンテキストラベルデータに対応

する音響特徴量データを取得する。

[0011] ボコーダは、デコーダ部により取得された音響特徴量から音声波形データを取得する。

[0012] この音声合成処理装置では、ニューラルネットワークのモデルで処理するのに適した最適化フルコンテキストラベルデータを用いて、ニューラルネットワークによる処理（学習処理、予測処理）を実行するため、高精度な音声合成処理を実行することができる。つまり、この音声合成処理装置では、従来技術とは異なり、処理対象の音素に先行する、あるいは、後続する音素についてのデータを含まないコンテキストラベルデータを最適化フルコンテキストラベルデータとして取得し、取得した最適化フルコンテキストラベルデータにより、ニューラルネットワークのモデルの処理を行う。ニューラルネットワーク（特に、sequence-to-sequence方式のニューラルネットワーク）では、時系列のデータを用いた処理が実行されるので、従来の音声合成処理に用いるコンテキストラベルデータに含める必要があった、処理対象の音素に先行する、あるいは、後続するデータは、ニューラルネットワークのモデルの処理において冗長となり、処理効率を落とす原因となる。この音声合成処理装置１００では、最適化フルコンテキストラベルデータ（単独音素についてのデータから構成されるコンテキストラベルデータ）を用いるので、ニューラルネットワークのモデルの処理が非常に効果的に実行できる。その結果、この音声合成処理装置では、高精度の音声合成処理を実行できる。

[0013] また、この音声合成処理装置では、処理対象言語に応じたテキスト解析処理を行い、当該テキスト解析処理で取得されたフルコンテキストラベルデータから、ニューラルネットワーク（例えば、sequence-to-sequence方式を用いたニューラルネットワーク）のモデルで処理するのに適した最適化フルコンテキストラベルデータを取得し、取得した最適化フルコンテキストラベルデータを用いて処理を行うことで、任意の処理対象言語について、高精度な音声合成処理を行うことができる。

[0014] したがって、この音声合成処理装置では、日本語等の英語以外の言語を処

理対象言語とする場合においても（処理対象言語を任意の言語にできる）、例えば、sequence-to-sequence方式を用いたテキスト音声合成用のニューラルネットワークのモデルにより、学習・最適化を行い、高品質な音声合成処理を実現することができる。

[0015] なお、「単独音素」とは、テキスト解析処理においてコンテキストラベルデータを取得するときに、処理対象とした音素のことをいう。

[0016] また、「最適化」とは、厳密な意味での最適化の他に、所定の誤差範囲を許容する範囲内に収めることを含む概念である。

[0017] 第２の発明は、第１の発明であって、音響特徴量は、メルスペクトログラムのデータである。

[0018] これにより、この音声合成処理装置では、入力されたテキストに対応するメルスペクトログラムのデータにより、音声合成処理を実行できる。

[0019] 第３の発明は、第１または第２の発明であって、ボコーダは、ニューラルネットワークのモデルを用いた処理を実行することで、音響特徴量から音声波形データを取得する。

[0020] これにより、この音声合成処理装置では、ニューラルネットワーク処理ができるボコーダを用いて、音声合成処理を実行できる。

[0021] 第４の発明は、第３の発明であって、ボコーダは、可逆変換ネットワークにより構成されたニューラルネットワークのモデルを用いた処理を実行することで、音響特徴量から音声波形データを取得する。

[0022] この音声合成処理装置では、ボコーダが、可逆変換ネットワークにより構成されたニューラルネットワークのモデルを用いた処理を行うので、ボコーダの構成をシンプルにできる。その結果、この音声合成処理装置では、ボコーダでの処理を高速化でき、音声合成処理をリアルタイムで実行できる。

[0023] 第５の発明は、第１から第４のいずれかの発明であって、音素単位のコンテキストラベルデータから音素継続長を推定する音素継続長推定部をさらに備える。

[0024] フルコンテキストラベルベクトル処理部は、音素継続長推定部により推定

された音素継続長である推定音素継続長に対応する期間において、当該推定音素継続長に対応する音素の最適化フルコンテキストラベルデータを継続してエンコーダ部に出力する。

[0025] この音声合成処理装置では、エンコーダ部への入力データ（最適化フルコンテキストラベルデータ）を、音素継続長推定部により取得（推定）した音素ごとの音素継続長に基づいて、引き延ばす処理（音素ph_kの音素継続長$dur(ph_k)$に相当する期間、音素ph_kの最適化フルコンテキストラベルデータを、繰り返しエンコーダ部３に入力する処理）を実行する。つまり、この音声合成処理装置では、安定して音素継続長を適切に推定することができる、隠れマルコフモデル等のモデルを用いた推定処理を実行して取得した音素継続長を用いて予測処理を実行するので、注意機構予測が失敗することに起因する、合成発話が途中で止まってしまう、同じフレーズを何回も繰り返してしまう、等の問題が発生することはない。

[0026] すなわち、この音声合成処理装置では、（１）音素継続長については、安定して音素継続長を適切に推定することができる、隠れマルコフモデル等のモデルを用いた推定処理（音素継続長推定部による処理）により取得し、（２）音響特徴量については、sequence-to-sequence方式を用いたニューラルネットワークのモデルで処理することにより取得する。

[0027] したがって、この音声合成処理装置では、注意機構予測が失敗することに起因する、合成発話が途中で止まってしまう、同じフレーズを何回も繰り返してしまう、等の問題が発生することを適切に防止するとともに、高精度な音声合成処理を実行することができる。

[0028] 第６の発明は、任意の言語を処理対象言語とし、エンコーダ・デコーダ方式のニューラルネットワークを用いて音声合成処理を実行する音声合成処理方法であって、テキスト解析ステップと、フルコンテキストラベルベクトル処理ステップと、エンコード処理ステップと、デコード処理ステップと、ボコーダ処理ステップと、を備える。

[0029] テキスト解析ステップは、処理対象言語のテキストデータに対してテキス

ト解析処理を実行し、コンテキストラベルデータを取得する。

[0030] フルコンテキストラベルベクトル処理ステップは、テキスト解析ステップにより取得されたコンテキストラベルデータから、コンテキストラベルデータを取得する処理において処理対象とされた音素である単独音素についてのコンテキストラベルを取得することで、ニューラルネットワークの学習処理に適した最適化フルコンテキストラベルデータを取得する。

[0031] エンコード処理ステップは、最適化フルコンテキストラベルデータに基づいて、ニューラルネットワークのエンコード処理を実行することで、隠れ状態データを取得する。

[0032] デコード処理ステップは、隠れ状態データに基づいて、ニューラルネットワークのデコード処理を実行することで、最適化フルコンテキストラベルデータに対応する音響特徴量データを取得する。

[0033] ボコーダ処理ステップは、デコード処理ステップにより取得された音響特徴量から音声波形データを取得する。

[0034] これにより、第１の発明と同様の効果を奏する音声合成処理方法を実現することができる。

[0035] 第７の発明は、第６の発明である音声合成処理方法をコンピュータに実行させるためのプログラムである。

[0036] これにより、第１の発明と同様の効果を奏する音声合成処理方法をコンピュータに実行させるためのプログラムを実現することができる。

[0037] 第８の発明は、任意の言語を処理対象言語とし、エンコーダ・デコーダ方式のニューラルネットワークを用いて音声合成処理を実行する音声合成処理装置であって、テキスト解析部と、フルコンテキストラベルベクトル処理部と、エンコーダ部と、音素継続長推定部と、強制アテンション部と、内分処理部と、コンテキスト算出部と、デコーダ部と、ボコーダと、を備える。

[0038] テキスト解析部は、処理対象言語のテキストデータに対してテキスト解析処理を実行し、コンテキストラベルデータを取得する。

[0039] フルコンテキストラベルベクトル処理部は、テキスト解析部により取得さ

れたコンテキストラベルデータから、コンテキストラベルデータを取得する処理において処理対象とされた音素である単独音素についてのコンテキストラベルを取得することで、ニューラルネットワークの学習処理に適した最適化フルコンテキストラベルデータを取得する。

[0040] エンコーダ部は、最適化フルコンテキストラベルデータに基づいて、ニューラルネットワークのエンコード処理を実行することで、隠れ状態データを取得する。

[0041] 音素継続長推定部は、音素単位のコンテキストラベルデータから音素継続長を推定する。

[0042] 強制アテンション部は、音素継続長推定部により推定された音素継続長に基づいて、第１重み付け係数データを取得する。

[0043] アテンション部は、エンコーダ部により取得された隠れ状態データに基づいて、第２重み付け係数データを取得する。

[0044] 内分処理部は、第１重み付け係数データと第２重み付け係数データとに対して内分処理を行うことで、合成重み付け係数データを取得する。

[0045] コンテキスト算出部は、合成重み付け係数データにより、エンコーダ部により取得された隠れ状態データに対して重み付け合成処理を実行することで、コンテキスト状態データを取得する。

[0046] デコーダ部は、コンテキスト状態データに基づいて、ニューラルネットワークのデコード処理を実行することで、最適化フルコンテキストラベルデータに対応する音響特徴量データを取得する。

[0047] ボコーダは、デコーダ部により取得された音響特徴量から音声波形データを取得する。

[0048] この音声合成処理装置では、音素継続長については、安定して音素継続長を適切に推定することができる、隠れマルコフモデル等のモデルを用いた推定処理（音素継続長推定部による処理）により取得した音素継続長を用いて処理することで、音素継続長の予測精度を保証する。つまり、この音声合成処理装置では、安定して音素継続長を適切に推定することができる、隠れマ

ルコフモデル等のモデルを用いた推定処理（音素継続長推定部による処理）により取得した音素継続長を用いて強制アテンション部により取得した重み付け係数データと、アテンション部により取得された重み付け係数データとを適度に合成した重み付け係数データにより生成したコンテキスト状態データを用いて予測処理を実行する。したがって、この音声合成処理装置では、注意機構の予測が失敗する場合（アテンション部により適切な重み付け係数データが取得できない場合）であっても、強制アテンション部により取得した重み付け係数データによる重み分の重み付け係数データが取得できるため、注意機構の予測の失敗が音声合成処理に影響を及ぼさないようにできる。

[0049] さらに、この音声合成処理装置では、音響特徴量については、 sequence-to-sequence方式を用いたニューラルネットワークのモデルで処理することにより取得できるので、高精度な音響特徴量の予測処理が実現できる。

[0050] したがって、この音声合成処理装置では、注意機構予測が失敗することに起因する、合成発話が途中で止まってしまう、同じフレーズを何回も繰り返してしまう、等の問題が発生することを適切に防止するとともに、高精度な音声合成処理を実行することができる。

[0051] なお、この音声合成処理装置において、内分処理を実行するときの内分比は、固定値であってもよいし、動的に変化する（更新される）値であってもよい。

発明の効果

[0052] 本発明によれば、日本語等の英語以外の言語を処理対象言語とする場合においても（処理対象言語を任意の言語にできる）、sequence-to-sequence方式を用いたテキスト音声合成用のニューラルネットワークのモデルにより、学習・最適化を行い、高品質な音声合成処理を実現する音声合成処理装置、音声合成処理方法、および、プログラムを実現することができる。

図面の簡単な説明

[0053] [図1]第１実施形態に係る音声合成処理装置１００の概略構成図。

[図2]処理対象言語を日本語とした場合のテキスト解析処理により取得される

フルコンテキストラベルデータに含まれる情報（パラメータ）（一例）を示す図。

[図3]最適化フルコンテキストラベルデータに含まれる情報（パラメータ）（一例）を示す図。

[図4]第１実施形態の第１変形例の音声合成処理装置のボコーダ６の概略構成を示す図。

[図5]第１実施形態の第１変形例の音声合成処理装置のボコーダ６の概略構成を示す図。

[図6]第１実施形態の第１変形例の音声合成処理装置によりＴＴＳ処理（処理対象言語：日本語）実行し、取得した音声波形データのメルスペクトログラム（予測データ）と、入力テキストの実際の音声波形データのメルスペクトログラム（オリジナルデータ）とを示す図。

[図7]第２実施形態に係る音声合成処理装置２００の概略構成図

[図8]推定された音素継続長に基づいて、エンコーダ部３に入力するデータＤｘ２を生成する処理を説明するための図。

[図9]第３実施形態に係る音声合成処理装置３００の概略構成図。

[図10]アテンション部４Ａにより取得された重み付け係数データｗａｔｔ（ｔ）と、強制アテンション部８により取得された重み付け係数データｗｆ（ｔ）とから取得した合成重み付け係数データｗ（ｔ）を用いてコンテキスト状態データｃ（ｔ）を取得する処理について説明するための図。

[図11]アテンション部４Ａにより取得された重み付け係数データｗａｔｔ（ｔ）と、強制アテンション部８により取得された重み付け係数データｗｆ（ｔ）とから取得した合成重み付け係数データｗ（ｔ）を用いてコンテキスト状態データｃ（ｔ）を取得する処理について説明するための図（時刻ｔ２の処理）。

[図12]アテンション部４Ａにより取得された重み付け係数データｗａｔｔ（ｔ）と、強制アテンション部８により取得された重み付け係数データｗｆ（ｔ）とから取得した合成重み付け係数データｗ（ｔ）を用いてコンテキスト

状態データｃ（ｔ）を取得する処理について説明するための図（時刻ｔ３の処理）。

[図13]時刻ｔ２においての処理で、注意機構の予測が失敗している場合を説明するための図。

[図14]本発明に係る音声合成処理装置を実現するコンピュータのハードウェア構成を示すブロック図。

発明を実施するための形態

[0054] ［第１実施形態］

第１実施形態について、図面を参照しながら、以下説明する。

[0055] ＜１．１：音声合成処理装置の構成＞

図１は、第１実施形態に係る音声合成処理装置１００の概略構成図である。

[0056] 音声合成処理装置１００は、図１に示すように、テキスト解析部１と、フルコンテキストラベルベクトル処理部２と、エンコーダ部３と、アテンション部４と、デコーダ部５と、ボコーダ６とを備える。

[0057] テキスト解析部１は、処理対象言語のテキストデータＤｉｎを入力とし、入力されたテキストデータＤｉｎに対して、テキスト解析処理を実行し、様々な言語情報からなるコンテキストを含む音素ラベルであるコンテキストラベルの系列を取得する。なお、日本語のように、アクセントやピッチによって、同じ文字（例えば、漢字）であっても、発音されたときの音声波形が異なる言語では、当該音素（処理対象の音素）の前後の音素についての言語情報も、コンテキストラベルに含める必要がある。テキスト解析部１は、上記のように、テキストが発音されたときの音声波形を特定するためのコンテキストラベル（処理対象言語によって必要となる先行する音素、および／または、後続する音素のデータを含めたコンテキストラベル）をフルコンテキストラベルデータＤｘ１として、フルコンテキストラベルベクトル処理部２に出力する。

[0058] フルコンテキストラベルベクトル処理部２は、テキスト解析部１から出力

されるデータＤｘ１（フルコンテキストラベルのデータ）を入力する。フルコンテキストラベルベクトル処理部２は、入力されたフルコンテキストラベルデータＤｘ１から、sequence-to-sequence方式のニューラルネットワークのモデルの学習処理に適したフルコンテキストラベルデータを取得するためのフルコンテキストラベルベクトル処理を実行する。そして、フルコンテキストラベルベクトル処理部２は、フルコンテキストラベルベクトル処理により取得したデータをデータＤｘ２（最適化フルコンテキストラベルデータＤｘ２）として、エンコーダ部３のエンコーダ側プレネット処理部３１に出力する。

[0059] エンコーダ部３は、図１に示すように、エンコーダ側プレネット処理部３１と、エンコーダ側ＬＳＴＭ層３２（ＬＳＴＭ：Ｌｏｎｇ　ｓｈｏｒｔ－ｔｅｒｍ　ｍｅｍｏｒｙ）とを備える。

[0060] エンコーダ側プレネット処理部３１は、フルコンテキストラベルベクトル処理部２から出力されるデータＤｘ２を入力する。エンコーダ側プレネット処理部３１は、入力したデータＤｘ２に対して、コンボリューション処理（コンボリューションフィルタによる処理）、データの正規化処理、活性化関数による処理（例えば、ＲｅＬＵ関数（ＲｅＬＵ：Ｒｅｃｔｉｆｉｅｄ　Ｌｉｎｅａｒ　Ｕｎｉｔ）による処理）を実行し、エンコーダ側ＬＳＴＭ層３２に入力可能なデータを取得する。そして、エンコーダ側プレネット処理部３１は、上記処理（プレネット処理）により取得したデータをデータＤｘ３としてエンコーダ側ＬＳＴＭ層３２に出力する。

[0061] エンコーダ側ＬＳＴＭ層３２は、リカーレントニューラルネットワークの隠れ層（ＬＳＴＭ層）に対応する層であり、エンコーダ側プレネット処理部３１から、現時刻ｔにおいて出力されるデータＤｘ３（これをデータＤｘ３（ｔ）と表記する）と、１つ前の時間ステップにおいて、エンコーダ側ＬＳＴＭ層３２から出力されたデータＤｘ４（これをデータＤｘ４（ｔ－１）と表記する）とを入力する。エンコーダ側ＬＳＴＭ層３２は、入力されたデータＤｘ３（ｔ）、データＤｘ４（ｔ－１）に対して、ＬＳＴＭ層による処理

を実行し、処理後のデータをデータＤｘ４（データＤｘ４（ｔ））としてアテンション部４に出力する。

[0062] アテンション部４は、エンコーダ部３から出力されるデータＤｘ４と、デコーダ部５のデコーダ側ＬＳＴＭ層５２から出力されるデータｈｏ（出力側隠れ状態データｈｏ）とを入力する。アテンション部４は、エンコーダ部３から出力されるデータＤｘ４、すなわち、入力側隠れ状態データ（これをデータｈｉという。また、時刻ｔの入力側隠れ状態データをデータｈｉ（ｔ）と表記する。）を所定の時間ステップ分記憶保持する。時間ステップｔ＝１からｔ＝Ｓ（Ｓ：自然数）の期間において、エンコーダ部３により取得され、アテンション部４に出力されたデータＤｘ４（＝ｈｉ）の集合を、$hi_{1...S}$と表記する。つまり、アテンション部４は、下記に相当するデータ$hi_{1...S}$を記憶保持する。

$hi_{1...S}$＝｛Ｄｘ４（１），Ｄｘ４（２），・・・，Ｄｘ４（Ｓ）｝

また、アテンション部４は、デコーダ部５のデコーダ側ＬＳＴＭ層５２から出力されるデータＤｙ３、すなわち、出力側隠れ状態データ（これをデータｈｏという）を所定の時間ステップ分記憶保持する。時間ステップｔ＝１からｔ＝Ｔ（Ｔ：自然数）の期間において、デコーダ側ＬＳＴＭ層５２により取得され、アテンション部４に出力されたデータＤｙ３（＝ｈｏ）の集合を、$ho_{1...T}$と表記する。つまり、アテンション部４は、下記に相当するデータ$ho_{1...T}$を記憶保持する。

$ho_{1...T}$＝｛Ｄｙ３（１），Ｄｙ３（２），・・・，Ｄｙ３（Ｔ）｝

そして、アテンション部４は、入力側隠れ状態データの集合データ$hi_{1...S}$と、出力側隠れ状態データの集合データ$ho_{1...T}$と、に基づいて、例えば、

ｃ（ｔ）＝ｆ１＿ａｔｔｎ（$hi_{1...S}$，$ho_{1...T}$）

ｆ１＿ａｔｔｎ（）：コンテキスト状態データを取得する関数

に相当する処理を実行して、現時刻ｔのコンテキスト状態データｃ（ｔ）を取得する。そして、アテンション部４は、取得したコンテキスト状態データ

ｃ（ｔ）をデコーダ側ＬＳＴＭ層５２に出力する。

デコーダ部５は、図１に示すように、デコーダ側プレネット処理部５１と、デコーダ側ＬＳＴＭ層５２と、線形予測部５３と、ポストネット処理部５４と、加算器５５と、を備える。

[0063] デコーダ側プレネット処理部５１は、線形予測部５３から出力される、１時間ステップ前のデータＤｙ４（これをＤｙ４（ｔ－１）という）を入力する。デコーダ側プレネット処理部５１は、例えば、複数層（例えば、２層）の全結合層を有しており、データの正規化処理（例えば、線形予測部５３から出力されるデータ（ベクトルデータ）の次元数が２Ｎであり、デコーダ側ＬＳＴＭ層に入力されるデータ（ベクトルデータ）の次元数がＮである場合、データの次元数をＮにするように、例えば、ドロップアウト処理を行うことを含む）、活性化関数による処理（例えば、ＲｅＬＵ関数（ＲｅＬＵ：Ｒｅｃｔｉｆｉｅｄ Ｌｉｎｅａｒ Ｕｎｉｔ）による処理）を実行し、デコーダ側ＬＳＴＭ層５２に入力可能なデータを取得する。そして、デコーダ側プレネット処理部５１は、上記処理（プレネット処理）により取得したデータをデータＤｙ２としてデコーダ側ＬＳＴＭ層５２に出力する。

[0064] デコーダ側ＬＳＴＭ層５２は、リカーレントニューラルネットワークの隠れ層（ＬＳＴＭ層）に対応する層である。デコーダ側ＬＳＴＭ層５２は、デコーダ側プレネット処理部５１から、現時刻ｔにおいて出力されるデータＤｙ２（これをデータＤｙ２（ｔ）と表記する）と、１つ前の時間ステップにおいて、デコーダ側ＬＳＴＭ層５２から出力されたデータＤｙ３（これをデータＤｙ３（ｔ－１）と表記する）と、アテンション部４から出力される時刻ｔのコンテキスト状態データｃ（ｔ）とを入力する。

[0065] デコーダ側ＬＳＴＭ層５２は、入力されたデータＤｙ２（ｔ）、データＤｙ３（ｔ－１）、および、コンテキスト状態データｃ（ｔ）を用いて、ＬＳＴＭ層による処理を実行し、処理後のデータをデータＤｙ３（データＤｙ３（ｔ））として線形予測部５３に出力する。また、デコーダ側ＬＳＴＭ層５２は、データＤｙ３（ｔ）、すなわち、時刻ｔの出力側隠れ状態データｈｏ

（ｔ）をアテンション部４に出力する。

[0066] 線形予測部５３は、デコーダ側ＬＳＴＭ層５２から出力されるデータＤｙ３を入力する。線形予測部５３は、所定の期間（例えば、メルスペクトログラムを取得するための１フレーム期間に相当する期間）内に、デコーダ側ＬＳＴＭ層５２から出力されるデータＤｙ３（複数のデータＤｙ３）を記憶保持し、当該複数のデータＤｙ３を用いて線形変換することで、所定期間におけるメルスペクトログラムの予測データＤｙ４を取得する。そして、線形予測部５３は、取得したデータＤｙ４をポストネット処理部５４、加算器５５、および、デコーダ側プレネット処理部５１に出力する。

[0067] ポストネット処理部５４は、例えば、複数層（例えば、５層）のコンボリューション層を有しており、コンボリューション処理（コンボリューションフィルタによる処理）、データの正規化処理、活性化関数による処理（例えば、ＲｅＬＵ関数（ＲｅＬＵ：Ｒｅｃｔｉｆｉｅｄ Ｌｉｎｅａｒ Ｕｎｉｔ）による処理やｔａｎｈ関数による処理）を実行し、予測データ（予測メルスペクトログラム）の残差データ（residual）を取得し、取得した残差データをデータＤｙ５として加算器５５に出力する。

[0068] 加算器５５は、線形予測部５３から出力される予測データＤｙ４（予測メルスペクトログラムのデータ）と、ポストネット処理部５４から出力される残差データＤｙ５（予測メルスペクトログラムの残差データ）とを入力する。加算器５５は、予測データＤｙ４（予測メルスペクトログラムのデータ）と、残差データＤｙ５（予測メルスペクトログラムの残差データ）とに対して加算処理を実行し、加算処理後のデータ（予測メルスペクトログラムのデータ）をデータＤｙ６としてボコーダ６に出力する。

[0069] ボコーダ６は、音響特徴量のデータを入力とし、入力された音響特徴量のデータから、当該音響特徴量に対応する音声信号波形を出力する。本実施形態において、ボコーダ６は、ニューラルネットワークによるモデルを用いたボコーダを採用する。ボコーダ６は、入力される音響特徴量を、メルスペクトログラムのデータとし、出力を当該メルスペクトログラムに対応する音声

信号波形とする。ボコーダ６は、学習時において、メルスペクトログラムと、当該メルスペクトログラムにより実現される音声信号波形（教師データ）として、ニューラルネットワークのモデルを学習させ、当該ニューラルネットワークのパラメータの最適化パラメータを取得することで、当該ニューラルネットワークのモデルを最適化する処理を行う。そして、ボコーダ６は、予測時において、最適化したニューラルネットワークのモデルを用いて、処理を行うことで、入力されるメルスペクトログラムのデータ（例えば、デコーダ部５から出力されるデータＤｙ６）から、当該メルスペクトログラムに対応する音声信号波形を予測し、予測した音声信号波形のデータをデータＤｏｕｔとして出力する。

[0070] ＜１．２：音声合成処理装置の動作＞

以上のように構成された音声合成処理装置１００の動作について以下説明する。

[0071] 以下では、音声合成処理装置１００の動作を、（１）学習処理（学習時の処理）と、（２）予測処理（予測時の処理）とに分けて説明する。

[0072] （１．２．１：学習処理）

まず、音声合成処理装置１００による学習処理について、説明する。なお、説明便宜のため、処理対象言語を日本語として、以下、説明する。

[0073] 処理対象言語である日本語のテキストデータＤｉｎをテキスト解析部１に入力する。また、当該テキストデータＤｉｎに対応するメルスペクトログラム（音響特徴量）のデータを教師データとして用意する。

[0074] テキスト解析部１は、入力されたテキストデータＤｉｎに対して、テキスト解析処理を実行し、様々な言語情報からなるコンテキストを含む音素ラベルであるコンテキストラベルの系列を取得する。

[0075] 日本語は、アクセントやピッチによって、同じ文字（例えば、漢字）であっても、発音されたときの音声波形が異なる言語であるので、当該音素（処理対象の音素）の前後の音素についての言語情報も、コンテキストラベルに含める必要がある。テキスト解析部１は、処理対象を日本語とする場合、テ

キストデータＤｉｎに対して、日本語用のテキスト解析処理を実行し、テキストが発音されたときの音声波形を特定するためのパラメータについて、必要に応じて、（１）当該音素のみのデータ、（２）先行する音素、および／または、後続する音素についてのデータを取得し、取得したデータをまとめてフルコンテキストラベルデータを取得する。

[0076] 図２は、処理対象言語を日本語とした場合のテキスト解析処理により取得されるフルコンテキストラベルデータに含まれる情報（パラメータ）（一例）を示す図である。

[0077] 図２に示す場合では、フルコンテキストラベルデータの各パラメータは、図２の「概要」に示した内容を特定するためのデータであり、図２の表に示した次元数、音素数分のデータである。

[0078] 図２に示すように、テキスト解析部１は、図２の表の全てのパラメータのデータをまとめて、フルコンテキストラベルデータ（ベクトルのデータ）として、取得する。図２の場合、フルコンテキストラベルデータは、４７８次元のベクトルデータとなる。

[0079] 上記のようにして取得されたフルコンテキストラベルデータＤｘ１は、テキスト解析部１からフルコンテキストラベルベクトル処理部２に出力される。

[0080] フルコンテキストラベルベクトル処理部２は、入力されたフルコンテキストラベルデータＤｘ１から、sequence-to-sequence方式のニューラルネットワークのモデルの学習処理に適したフルコンテキストラベルデータを取得するためのフルコンテキストラベルベクトル処理を実行する。具体的には、フルコンテキストラベルベクトル処理部２は、先行する音素についてのパラメータ（データ）、後続する音素についてのパラメータ（データ）を削除することで、最適化フルコンテキストラベルデータＤｘ２を取得する。例えば、フルコンテキストラベルデータＤｘ１が図２に示すパラメータを含むデータである場合、先行する音素についてのパラメータ（データ）、後続する音素についてのパラメータ（データ）を削除することで、最適化フルコンテキス

トラベルデータＤｘ２を取得する。

[0081] 図３は、上記のようにして取得した最適化フルコンテキストラベルデータに含まれる情報（パラメータ）（一例）を示す図である。

[0082] 図３の場合、最適化フルコンテキストラベルデータは、１３０次元のベクトルデータとなり、４７８次元のベクトルデータであるフルコンテキストラベルデータＤｘ１と比べると、次元数が著しく低減されていることが分かる。

[0083] 音声合成処理装置１００で用いられているニューラルネットワークのモデルが、sequence-to-sequence方式のニューラルネットワーク（リカーレントニューラルネットワーク）のモデルであり、エンコーダ側ＬＳＴＭ層３２、デコーダ側ＬＳＴＭ層５２を有しているので、入力されるデータ列について、時系列の関係を考慮した学習処理、予測処理ができるため、従来技術で必要とされていた先行する音素、後続する音素のデータは、冗長となり、学習処理の効率、予測処理の精度を悪化させる原因となる。そのため、音声合成処理装置１００では、上記のように、当該音素についてのパラメータ（データ）のみを残して取得した最適化フルコンテキストラベルデータＤｘ２を取得し、取得した最適化フルコンテキストラベルデータＤｘ２を用いて、学習処理、予測処理を行うことで、高速かつ高精度に処理を実行することができる。

[0084] 上記により取得されたデータＤｘ２（最適化フルコンテキストラベルデータＤｘ２）は、フルコンテキストラベルベクトル処理部２からのエンコーダ部３のエンコーダ側プレネット処理部３１に出力される。

[0085] エンコーダ側プレネット処理部３１は、フルコンテキストラベルベクトル処理部２から入力したデータＤｘ２に対して、コンボリューション処理（コンボリューションフィルタによる処理）、データの正規化処理、活性化関数による処理（例えば、ＲｅＬＵ関数（ＲｅＬＵ：Ｒｅｃｔｉｆｉｅｄ Ｌｉｎｅａｒ Ｕｎｉｔ）による処理）を実行し、エンコーダ側ＬＳＴＭ層３２に入力可能なデータを取得する。そして、エンコーダ側プレネット処理部３

１は、上記処理（プレネット処理）により取得したデータをデータＤｘ３としてエンコーダ側ＬＳＴＭ層３２に出力する。

[0086] エンコーダ側ＬＳＴＭ層３２は、エンコーダ側プレネット処理部３１から、現時刻ｔにおいて出力されるデータＤｘ３（ｔ）と、１つ前の時間ステップにおいて、エンコーダ側ＬＳＴＭ層３２から出力されたデータＤｘ４（ｔ－１）とを入力する。そして、エンコーダ側ＬＳＴＭ層３２は、入力されたデータＤｘ３（ｔ）、データＤｘ４（ｔ－１）に対して、ＬＳＴＭ層による処理を実行し、処理後のデータをデータＤｘ４（データＤｘ４（ｔ））としてアテンション部４に出力する。

[0087] アテンション部４は、エンコーダ部３から出力されるデータＤｘ４と、デコーダ部５のデコーダ側ＬＳＴＭ層５２から出力されるデータｈｏ（出力側隠れ状態データｈｏ）とを入力する。アテンション部４は、エンコーダ部３から出力されるデータＤｘ４、すなわち、入力側隠れ状態データｈｉを所定の時間ステップ分記憶保持する。例えば、アテンション部４は、時間ステップｔ＝１からｔ＝Ｓ（Ｓ：自然数）の期間において、エンコーダ部３により取得され、アテンション部４に出力されたデータＤｘ４（＝ｈｉ）の集合を、$hi_{1...S}$（＝｛Ｄｘ４（１），Ｄｘ４（２），・・・，Ｄｘ４（Ｓ）｝）として記憶保持する。

[0088] また、アテンション部４は、デコーダ部５のデコーダ側ＬＳＴＭ層５２から出力されるデータＤｙ３、すなわち、出力側隠れ状態データｈｏを所定の時間ステップ分記憶保持する。例えば、アテンション部４は、時間ステップｔ＝１からｔ＝Ｔ（Ｔ：自然数）の期間において、デコーダ側ＬＳＴＭ層５２により取得され、アテンション部４に出力されたデータＤｙ３（＝ｈｏ）の集合を、$ho_{1...T}$（＝｛Ｄｙ３（１），Ｄｙ３（２），・・・，Ｄｙ３（Ｔ）｝）として記憶保持する。

[0089] そして、アテンション部４は、入力側隠れ状態データの集合データ$hi_{1...S}$と、出力側隠れ状態データの集合データ$ho_{1...T}$と、に基づいて、例えば、

ｃ（ｔ）＝ｆ１＿ａｔｔｎ（ｈｉ$_{1...S}$，ｈｏ$_{1...T}$）

ｆ１＿ａｔｔｎ（）：コンテキスト状態データを取得する関数

に相当する処理を実行して、現時刻ｔのコンテキスト状態データｃ（ｔ）を取得する。

[0090] そして、アテンション部４は、取得したコンテキスト状態データｃ（ｔ）をデコーダ側ＬＳＴＭ層５２に出力する。

デコーダ側プレネット処理部５１は、線形予測部５３から出力される、１時間ステップ前のデータＤｙ４（ｔ－１）を入力する。デコーダ側プレネット処理部５１は、例えば、複数層（例えば、２層）の全結合層を有しており、データの正規化処理（例えば、線形予測部５３から出力されるデータ（ベクトルデータ）の次元数が２Ｎであり、デコーダ側ＬＳＴＭ層に入力されるデータ（ベクトルデータ）の次元数がＮである場合、データの次元数をＮにするように、例えば、ドロップアウト処理を行うことを含む）、活性化関数による処理（例えば、ＲｅＬＵ関数（ＲｅＬＵ：Ｒｅｃｔｉｆｉｅｄ Ｌｉｎｅａｒ Ｕｎｉｔ）による処理）を実行し、デコーダ側ＬＳＴＭ層５２に入力可能なデータを取得する。そして、デコーダ側プレネット処理部５１は、上記処理（プレネット処理）により取得したデータをデータＤｙ２としてデコーダ側ＬＳＴＭ層５２に出力する。

[0091] デコーダ側ＬＳＴＭ層５２は、デコーダ側プレネット処理部５１から、現時刻ｔにおいて出力されるデータＤｙ２（ｔ）と、１つ前の時間ステップにおいて、デコーダ側ＬＳＴＭ層５２から出力されたデータＤｙ３（ｔ－１）と、アテンション部４から出力される時刻ｔのコンテキスト状態データｃ（ｔ）とを入力する。

[0092] デコーダ側ＬＳＴＭ層５２は、入力されたデータＤｙ２（ｔ）、データＤｙ３（ｔ－１）、および、コンテキスト状態データｃ（ｔ）を用いて、ＬＳＴＭ層による処理を実行し、処理後のデータをデータＤｙ３（ｔ）として線形予測部５３に出力する。また、デコーダ側ＬＳＴＭ層５２は、データＤｙ３（ｔ）、すなわち、時刻ｔの出力側隠れ状態データｈｏ（ｔ）をアテンシ

ョン部４に出力する。

[0093] 線形予測部５３は、デコーダ側ＬＳＴＭ層５２から出力されるデータＤｙ３を入力する。線形予測部５３は、所定の期間（例えば、メルスペクトログラムを取得するための１フレーム期間に相当する期間）内に、デコーダ側ＬＳＴＭ層５２から出力されるデータＤｙ３（複数のデータＤｙ３）を記憶保持し、当該複数のデータＤｙ３を用いて線形変換することで、所定期間におけるメルスペクトログラムの予測データＤｙ４を取得する。そして、線形予測部５３は、取得したデータＤｙ４をポストネット処理部５４、加算器５５、および、デコーダ側プレネット処理部５１に出力する。

[0094] ポストネット処理部５４は、例えば、コンボリューション処理（コンボリューションフィルタによる処理）、データの正規化処理、活性化関数による処理（例えば、ＲｅＬＵ関数（ＲｅＬＵ：Ｒｅｃｔｉｆｉｅｄ Ｌｉｎｅａｒ Ｕｎｉｔ）による処理やｔａｎｈ関数による処理）を実行し、予測データ（予測メルスペクトログラム）の残差データ（residual）を取得し、取得した残差データをデータＤｙ５として加算器５５に出力する。

[0095] 加算器５５は、線形予測部５３から出力される予測データＤｙ４（予測メルスペクトログラムのデータ）と、ポストネット処理部５４から出力される残差データＤｙ５（予測メルスペクトログラムの残差データ）とを入力する。加算器５５は、予測データＤｙ４（予測メルスペクトログラムのデータ）と、残差データＤｙ５（予測メルスペクトログラムの残差データ）とに対して加算処理を実行し、加算処理後のデータ（予測メルスペクトログラムのデータ）をデータＤｙ６として出力する。

[0096] そして、音声合成処理装置１００では、上記のように取得されたデータＤｙ６（予測メルスペクトログラムのデータ）と、テキストデータＤｉｎに対応するメルスペクトログラム（音響特徴量）の教師データ（正解のメルスペクトログラム）とを比較し、両者の差（比較結果）（例えば、差分ベクトルのノルムやユークリッド距離により表現する差）が小さくなるように、エンコーダ部３、デコーダ部５のニューラルネットワークのモデルのパラメータ

を更新する。音声合成処理装置１００では、このパラメータ更新処理を繰り返し実行し、データＤｙ６（予測メルスペクトログラムのデータ）と、テキストデータＤｉｎに対応するメルスペクトログラム（音響特徴量）の教師データ（正解のメルスペクトログラム）との差が十分小さくなる（所定の誤差範囲におさまる）、ニューラルネットワークのモデルのパラメータを最適化パラメータとして取得する。

[0097] 音声合成処理装置１００では、上記のようにして取得した最適化パラメータに基づいて、エンコーダ部３、デコーダ部５のニューラルネットワークのモデルの各層に含まれるシナプス間の結合係数（重み係数）を設定することで、エンコーダ部３、デコーダ部５のニューラルネットワークのモデルを最適化モデル（学習済みモデル）とすることができる。

[0098] 以上により、音声合成処理装置１００において、入力をテキストデータとし、出力をメルスペクトログラムとするニューラルネットワークの学習済みモデル（最適化モデル）を構築できる。

[0099] また、ボコーダ６として、ニューラルネットワークによるモデルを用いたボコーダを採用する場合、入力される音響特徴量を、メルスペクトログラムのデータとし、出力を当該メルスペクトログラムに対応する音声信号波形として学習処理を実行する。つまり、ボコーダ６において、メルスペクトログラムのデータを入力し、音声合成処理をニューラルネットワークによるモデルを用いた処理により実行し、音声波形データを出力させる。ボコーダ６から出力される当該音声波形データと、ボコーダに入力したメルスペクトログラムに対応する音声波形データ（正解の音声波形データ）とを比較し、両者の差（比較結果）（例えば、差分ベクトルのノルムやユークリッド距離により表現する差）が小さくなるように、ボコーダ６のニューラルネットワークのモデルのパラメータを更新する。ボコーダ６では、このパラメータ更新処理を繰り返し実行し、ボコーダの入力データ（メルスペクトログラムのデータ）と、ボコーダ６に入力されたメルスペクトログラムに対応する音声波形データ（正解の音声波形データ）との差が十分小さくなる（所定の誤差範囲

におさまる）、ニューラルネットワークのモデルのパラメータを最適化パラメータとして取得する。

[0100] ボコーダ６では、上記のようにして取得した最適化パラメータに基づいて、ボコーダ６のニューラルネットワークのモデルの各層に含まれるシナプス間の結合係数（重み係数）を設定することで、ボコーダ６のニューラルネットワークのモデルの最適化モデル（学習済みモデル）とすることができる。

[0101] 以上により、ボコーダ６において、入力をテキストデータとし、出力をメルスペクトログラムとするニューラルネットワークの学習済みモデル（最適化モデル）を構築できる。

[0102] なお、音声合成処理装置１００において、（１）エンコーダ部３、デコーダ部５の学習処理と、（２）ボコーダ６の学習処理とを連携させて学習処理を実行してもよいし、上記のように、個別に学習処理を実行してもよい。音声合成処理装置１００において、（１）エンコーダ部３、デコーダ部５の学習処理と、（２）ボコーダ６の学習処理とを連携させて学習処理を実行する場合、入力をテキストデータとし、当該テキストデータに対応する音声波形データ（正解の音声波形データ）とを用いて、（１）エンコーダ部３、デコーダ部５のニューラルネットワークのモデルと、（２）ボコーダ６のニューラルネットワークのモデルの最適化パラメータを取得することで学習処理を実行すればよい。

[0103] （１．２．２：予測処理）

次に、音声合成処理装置１００による予測処理について、説明する。なお、予測処理においても、説明便宜のため、処理対象言語を日本語として、以下、説明する。

[0104] 予測処理を実行する場合、音声合成処理装置１００では、上記の学習処理により取得された学習済みモデル、すなわち、エンコーダ部３、デコーダ部５のニューラルネットワークの最適化モデル（最適化パラメータが設定されているモデル）、および、ボコーダ６のニューラルネットワークの最適化モデル（最適化パラメータが設定されているモデル）が構築されている。そし

て、音声合成処理装置１００では、当該学習済みモデルを用いて予測処理が実行される。

[0105] 音声合成処理の対象とする日本語のテキストデータＤｉｎをテキスト解析部１に入力する。

[0106] テキスト解析部１は、入力されたテキストデータＤｉｎに対して、日本語用のテキスト解析処理を実行し、例えば、図２に示すパラメータを含む４７８次元のベクトルデータとして、フルコンテキストラベルデータＤｘ１を取得する。

[0107] そして、取得されたフルコンテキストラベルデータＤｘ１は、テキスト解析部１からフルコンテキストラベルベクトル処理部２に出力される。

[0108] フルコンテキストラベルベクトル処理部２は、入力されたフルコンテキストラベルデータＤｘ１に対して、フルコンテキストラベルベクトル処理を実行し、最適化フルコンテキストラベルＤｘ２を取得する。なお、ここで取得される最適化フルコンテキストラベルＤｘ２は、エンコーダ部３、デコーダ部５のsequence-to-sequence方式のニューラルネットワークのモデルの学習処理を行うときに設定した最適化フルコンテキストラベルデータＤｘ２と同じ次元数を有し、かつ、同じパラメータ（情報）を有するデータである。

[0109] 上記により取得されたデータＤｘ２（最適化フルコンテキストラベルデータＤｘ２）は、フルコンテキストラベルベクトル処理部２からエンコーダ部３のエンコーダ側プレネット処理部３１に出力される。

[0110] エンコーダ側プレネット処理部３１は、フルコンテキストラベルベクトル処理部２から入力したデータＤｘ２に対して、コンボリューション処理（コンボリューションフィルタによる処理）、データの正規化処理、活性化関数による処理（例えば、ＲｅＬＵ関数（ＲｅＬＵ：Ｒｅｃｔｉｆｉｅｄ Ｌｉｎｅａｒ Ｕｎｉｔ）による処理）を実行し、エンコーダ側ＬＳＴＭ層３２に入力可能なデータを取得する。そして、エンコーダ側プレネット処理部３１は、上記処理（プレネット処理）により取得したデータをデータＤｘ３としてエンコーダ側ＬＳＴＭ層３２に出力する。

[0111] エンコーダ側ＬＳＴＭ層３２は、エンコーダ側プレネット処理部３１から、現時刻ｔにおいて出力されるデータＤｘ３（ｔ）と、１つ前の時間ステップにおいて、エンコーダ側ＬＳＴＭ層３２から出力されたデータＤｘ４（ｔ－１）とを入力する。そして、エンコーダ側ＬＳＴＭ層３２は、入力されたデータＤｘ３（ｔ）、データＤｘ４（ｔ－１）に対して、ＬＳＴＭ層による処理（ニューラルネットワーク処理）を実行し、処理後のデータをデータＤｘ４（データＤｘ４（ｔ））としてアテンション部４に出力する。

[0112] アテンション部４は、エンコーダ部３から出力されるデータＤｘ４と、デコーダ部５のデコーダ側ＬＳＴＭ層５２から出力されるデータｈｏ（出力側隠れ状態データｈｏ）とを入力する。アテンション部４は、エンコーダ部３から出力されるデータＤｘ４、すなわち、入力側隠れ状態データｈｉを所定の時間ステップ分記憶保持する。例えば、アテンション部４は、時間ステップｔ＝１からｔ＝Ｓ（Ｓ：自然数）の期間において、エンコーダ部３により取得され、アテンション部４に出力されたデータＤｘ４（＝ｈｉ）の集合を、$hi_{1...S}$（＝｛Ｄｘ４（１），Ｄｘ４（２），・・・，Ｄｘ４（Ｓ）｝）として記憶保持する。

[0113] また、アテンション部４は、デコーダ部５のデコーダ側ＬＳＴＭ層５２から出力されるデータＤｙ３、すなわち、出力側隠れ状態データｈｏを所定の時間ステップ分記憶保持する。例えば、アテンション部４は、時間ステップｔ＝１からｔ＝Ｔ（Ｔ：自然数）の期間において、デコーダ側ＬＳＴＭ層５２により取得され、アテンション部４に出力されたデータＤｙ３（＝ｈｏ）の集合を、$ho_{1...T}$（＝｛Ｄｙ３（１），Ｄｙ３（２），・・・，Ｄｙ３（Ｔ）｝）として記憶保持する。

[0114] そして、アテンション部４は、入力側隠れ状態データの集合データ$hi_{1...S}$と、出力側隠れ状態データの集合データ$ho_{1...T}$と、に基づいて、例えば、

$$c(t) = f1\_attn(hi_{1...S}, ho_{1...T})$$

ｆ１＿ａｔｔｎ（）：コンテキスト状態データを取得する関数

に相当する処理を実行して、現時刻ｔのコンテキスト状態データｃ（ｔ）を取得する。

[0115] そして、アテンション部４は、取得したコンテキスト状態データｃ（ｔ）をデコーダ側ＬＳＴＭ層５２に出力する。

デコーダ側プレネット処理部５１は、線形予測部５３から出力される、１時間ステップ前のデータＤｙ４（ｔ－１）を入力する。デコーダ側プレネット処理部５１は、例えば、複数層（例えば、２層）の全結合層を有しており、データの正規化処理（例えば、線形予測部５３から出力されるデータ（ベクトルデータ）の次元数が２Ｎであり、デコーダ側ＬＳＴＭ層に入力されるデータ（ベクトルデータ）の次元数がＮである場合、データの次元数をＮにするように、例えば、ドロップアウト処理を行うことを含む）、活性化関数による処理（例えば、ＲｅＬＵ関数（ＲｅＬＵ：Ｒｅｃｔｉｆｉｅｄ Ｌｉｎｅａｒ Ｕｎｉｔ）による処理）を実行し、デコーダ側ＬＳＴＭ層５２に入力可能なデータを取得する。そして、デコーダ側プレネット処理部５１は、上記処理（プレネット処理）により取得したデータをデータＤｙ２としてデコーダ側ＬＳＴＭ層５２に出力する。

[0116] デコーダ側ＬＳＴＭ層５２は、デコーダ側プレネット処理部５１から、現時刻ｔにおいて出力されるデータＤｙ２（ｔ）と、１つ前の時間ステップにおいて、デコーダ側ＬＳＴＭ層５２から出力されたデータＤｙ３（ｔ－１）と、アテンション部４から出力される時刻ｔのコンテキスト状態データｃ（ｔ）とを入力する。

[0117] デコーダ側ＬＳＴＭ層５２は、入力されたデータＤｙ２（ｔ）、データＤｙ３（ｔ－１）、および、コンテキスト状態データｃ（ｔ）を用いて、ＬＳＴＭ層による処理を実行し、処理後のデータをデータＤｙ３（ｔ）として線形予測部５３に出力する。また、デコーダ側ＬＳＴＭ層５２は、データＤｙ３（ｔ）、すなわち、時刻ｔの出力側隠れ状態データｈｏ（ｔ）をアテンション部４に出力する。

[0118] 線形予測部５３は、デコーダ側ＬＳＴＭ層５２から出力されるデータＤｙ

３を入力する。線形予測部５３は、所定の期間（例えば、メルスペクトログラムを取得するための１フレーム期間に相当する期間）内に、デコーダ側ＬＳＴＭ層５２から出力されるデータＤｙ３（複数のデータＤｙ３）を記憶保持し、当該複数のデータＤｙ３を用いて線形変換することで、所定期間におけるメルスペクトログラムの予測データＤｙ４を取得する。そして、線形予測部５３は、取得したデータＤｙ４をポストネット処理部５４、加算器５５、および、デコーダ側プレネット処理部５１に出力する。

[0119] ポストネット処理部５４は、例えば、コンボリューション処理（コンボリューションフィルタによる処理）、データの正規化処理、活性化関数による処理（例えば、ＲｅＬＵ関数（ＲｅＬＵ：Ｒｅｃｔｉｆｉｅｄ Ｌｉｎｅａｒ Ｕｎｉｔ）による処理やｔａｎｈ関数による処理）を実行し、予測データ（予測メルスペクトログラム）の残差データ（residual）を取得し、取得した残差データをデータＤｙ５として加算器５５に出力する。

[0120] 加算器５５は、線形予測部５３から出力される予測データＤｙ４（予測メルスペクトログラムのデータ）と、ポストネット処理部５４から出力される残差データＤｙ５（予測メルスペクトログラムの残差データ）とを入力する。加算器５５は、予測データＤｙ４（予測メルスペクトログラムのデータ）と、残差データＤｙ５（予測メルスペクトログラムの残差データ）とに対して加算処理を実行し、加算処理後のデータ（予測メルスペクトログラムのデータ）をデータＤｙ６として、ボコーダ６に出力する。

[0121] ボコーダ６は、デコーダ部５の加算器５５から出力されるデータＤｙ６（予測メルスペクトログラムのデータ（音響特徴量のデータ））を入力とし、入力されたデータＤｙ６に対して、学習済みモデルを用いたニューラルネットワーク処理による音声合成処理を実行し、データＤｙ６（予測メルスペクトログラム）に対応する音声信号波形データを取得する。そして、ボコーダ６は、取得した音声信号波形データを、データＤｏｕｔとして出力する。

[0122] このように、音声合成処理装置１００では、入力されたテキストデータＤｉｎに対応する音声波形データＤｏｕｔを取得することができる。

[0123] 以上のように、音声合成処理装置１００では、処理対象言語（上記では日本語）のテキストを入力とし、当該処理対象言語に応じたテキスト解析処理により、フルコンテキストラベルデータを取得し、取得したフルコンテキストラベルデータからsequence-to-sequence方式を用いたニューラルネットワークのモデルで処理（学習処理、および／または、予測処理）を実行するのに適したデータである最適化フルコンテキストラベルデータを取得する。そして、音声合成処理装置１００では、入力を最適化フルコンテキストラベルデータとし、出力をメルスペクトログラム（音響特徴量の一例）として、エンコーダ部３、アテンション部４、および、デコーダ部５において、ニューラルネットワークのモデルを用いた処理（学習処理、予測処理）を実行することで、高精度な処理を実現できる。さらに、音声合成処理装置１００では、ボコーダ６により、上記により取得したメルスペクトログラム（音響特徴量の一例）から、当該メルスペクトログラムに対応する音声信号波形データを取得し、取得したデータを出力することで、音声波形データ（データＤｏｕｔ）を取得する。これにより、音声合成処理装置１００では、入力されたテキストに相当する音声波形データを取得することができる。

[0124] つまり、音声合成処理装置１００では、sequence-to-sequence方式を用いたニューラルネットワークのモデルで処理するのに適した最適化フルコンテキストラベルデータを用いて、ニューラルネットワークによる処理が実行されるため、高精度な音声合成処理を実行することができる。また、音声合成処理装置１００では、処理対象言語に応じたテキスト解析処理を行い、当該テキスト解析処理で取得されたフルコンテキストラベルデータから、sequence-to-sequence方式を用いたニューラルネットワークのモデルで処理するのに適した最適化フルコンテキストラベルデータを取得し、取得した最適化フルコンテキストラベルデータを用いて処理を行うことで、任意の処理対象言語について、高精度な音声合成処理を行うことができる。

[0125] したがって、音声合成処理装置１００では、日本語等の英語以外の言語を処理対象言語とする場合においても（処理対象言語を任意の言語にできる）

、sequence-to-sequence方式を用いたテキスト音声合成用のニューラルネットワークのモデルにより、学習・最適化を行い、高品質な音声合成処理を実現することができる。

[0126] ≪第１変形例≫

次に、第１実施形態の第１変形例について、説明する。なお、上記実施形態と同様の部分については、同一符号を付し、詳細な説明を省略する。

[0127] 本変形例の音声合成処理装置では、ボコーダ６が、例えば、下記先行技術文献に開示されているような、可逆変換が可能なニューラルネットワークのモデルを用いた処理を行う。この点が第１実施形態と相違し、それ以外については、本変形例の音声合成処理装置は、第１実施形態の音声合成処理装置１００と同様である。

（先行技術文献Ａ）：

R. Prenger, R. Valle, and B. Catanzaro, “WaveGlow: A flowbased generative network for speech synthesis,” in Proc. ICASSP, May 2019.

図４は、第１実施形態の第１変形例の音声合成処理装置のボコーダ６の概略構成を示す図であり、学習処理時においけるデータの流れを明示した図である。

[0128] 図５は、第１実施形態の第１変形例の音声合成処理装置のボコーダ６の概略構成を示す図であり、予測処理時においけるデータの流れを明示した図である。

[0129] 本変形例のボコーダ６は、図４に示すように、ベクトル処理部６１と、アップサンプリング処理部６２と、ｍ個（ｍ：自然数）の可逆処理部６３ａ～６３ｘとを備える。

[0130] まず、本変形例のボコーダ６の学習処理について、説明する。

[0131] 本変形例のボコーダ６は、学習処理において、音響特徴量としてメルスペクトログラム（これをデータｈとする）と、当該メルスペクトログラムに対応する音声信号波形データ（正解データ）（これをデータｘとする）とを入力し、ガウス白色ノイズ（これをデータｚとする）を出力する。

[0132] ベクトル処理部６１は、学習処理時において、音声信号波形データｘを入力し、入力したデータｘに対して、例えば、コンボリューション処理を施して、可逆処理部６３ａ（学習処理時において最初にデータ入力される可逆処理部）に入力可能な次元数のベクトルデータＤｘ１に変換する。そして、ベクトル処理部６１は、変換したベクトルデータDx_1を可逆処理部６３ａに出力する。

[0133] アップサンプリング処理部６２は、音響特徴量としてメルスペクトログラムのデータｈを入力し、入力されたメルスペクトログラムのデータｈに対して、アップサンプリング処理を実行し、処理後のデータ（アップサンプリングされたメルスペクトログラムのデータ）をデータｈ１として、可逆処理部６３ａ～６３ｘのそれぞれのＷＮ変換部６３２に出力する。

[0134] 可逆処理部６３ａは、図４に示すように、可逆１×１畳み込み層と、アフィンカップリング層とを備える。

[0135] 可逆１×１畳み込み層は、ベクトル処理部６１から出力されるデータDx_1を入力とし、入力されたデータに対して、重み係数行列W_k（ｋ＝１）（シナプス間の結合係数（重み係数）を規定する行列）により、ニューラルネットワーク処理を実行する、つまり、

$$DxA_1 = W_1 \times Dx_1$$

に相当する処理を実行して、データDxA_1を取得する。

[0136] なお、重み係数行列W_kは、直行行列となるように設定されており、したがって、逆変換が可能となる。

[0137] このようにして取得されたデータDxA_1は、可逆１×１畳み込み層からアフィンカップリング層に出力される。

[0138] アフィンカップリング層では、データ分割部６３１により、

$$x = DxA_1$$

$$x_a, x_b = split(x)$$

ｓｐｌｉｔ（）：データ分割をする関数

に相当する処理を実行し、入力データｘを２分割し、分割データx_aとx_bを

取得する。例えば、ｘがｎ１×２（ｎ１：自然数）のビット数のデータである場合、x_aは、ｘの上位ｎ１ビット分のデータであり、x_bは、ｘの下位ｎ１ビット分のデータである。

[0139] そして、データx_aは、ＭＮ変換部６３２およびデータ合成部６３４に出力される。また、データx_bは、アフィン変換部６３３に出力される。

[0140] ＭＮ変換部６３２は、データ分割部６３１から出力されるデータx_aと、アップサンプリング処理部６２から出力されるアップサンプリングされたメルスペクトログラムのデータｈ１とを入力する。そして、ＭＮ変換部６３２は、データx_aと、データｈ１とに対して、任意の変換であるＭＮ変換（例えば、ＷａｖｅＮｅｔによる変換）を実行し、アフィン変換のパラメータとするデータs_j，t_j（s_j：アフィン変換用の行列、t_j：アフィン変換用のオフセット）を取得する。取得されたアフィン変換のパラメータとするデータs_j，t_jは、ＷＮ変換部６３２からアフィン変換部６３３に出力される。

[0141] アフィン変換部６３３は、ＭＮ変換部６３２により取得されたデータs_j，t_jを用いて、データ分割部６３１から入力されるデータx_bに対して、アフィン変換を行う。つまり、アフィン変換部６３３は、

$$x_b' = Affin(s_j, t_j, x_b)$$
$$= s_j \times x_b + t_j$$

に相当する処理を実行することで、データx_bのアフィン変換後のデータx_b'を取得し、取得したデータx_b'をデータ合成部６３４に出力する。

[0142] データ合成部６３４では、データ分割部６３１から出力されるデータx_aと、アフィン変換部６３３から出力されるデータx_b'とを入力し、データx_aと、データx_b'とを合成する処理、すなわち、

$$Dx_2 = concat(x_a, x_b')$$

に相当する処理を実行し、データDx_2を取得する。なお、データ合成部６３４でのデータ合成処理は、例えば、x_a、x_b'が、それぞれ、ｎ１ビットのデータである場合、上位ｎ１ビットがx_aとなり、下位ｎ１ビットがx_b'となるｎ１×２ビットのデータを取得する処理である。

[0143] このようにして取得されたデータＤｘ$_2$は、可逆処理部６３ａから、可逆処理部６３ｂ（２番目の可逆処理部）に出力される。

[0144] 可逆処理部６３ｂ～６３ｘでは、可逆処理部６３ａと同様の処理が実行される。つまり、本変形例のボコーダ６では、図４に示すように、可逆処理部６３ａの処理がｍ回繰り返し実行される。そして、最終段の可逆処理部６３ｘからのデータｚが出力される。なお、本変形例のボコーダ６は、ｍ個の可逆処理部を備えるものとする。

[0145] そして、本変形例のボコーダ６では、出力データｚが、ガウス白色ノイズとなるように、ニューラルネットワークのモデルの学習を行う。つまり、ｘを入力としたときのｚをｚ（ｘ）とすると、ｚ（ｘ）がガウス分布Ｎ（μ，σ）（μは平均値でありμ＝０、σは標準偏差）に従うガウス確率変数となるように、本変形例のボコーダ６のニューラルネットワークのモデルのパラメータを設定する。なお、σは、例えば、入力される音響特徴量としてメルスペクトログラムのデータの情報量Ｉに相関のあるデータとする。

[0146] つまり、本変形例のボコーダ６では、ｘが入力されたときの尤度（θ：ニューラルネットワークのパラメータ）p_θ（ｘ）を、下記数式により規定することができ、当該尤度p_θ（ｘ）を最大にするパラメータθ_{opt}を取得することで、学習処理を実行する。

[数1]

$$-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\boldsymbol{z}(\boldsymbol{x})^T \boldsymbol{z}(\boldsymbol{x})}{2\sigma_{WG}^2} - \sum_{j=1}^{m1} \log \boldsymbol{s}_j(\boldsymbol{x}, \boldsymbol{h}) - \sum_{k=1}^{m2} \log |\det(\boldsymbol{W}_k)|$$

p_θ（ｘ）：ｘが入力されたときの尤度（θ：ニューラルネットワークのパラメータ）

s_j（ｘ，ｈ）：ｘ、ｈが入力されたときのｊ番目のアフィンカップリング層の出力係数ベクトル

W_k：ｋ番目の可逆１×１畳み込み層の係数行列（重み付け係数の行列）

ｚ（ｘ）：ｘが入力されたときの出力値（出力ベクトル）。

ｈ：音響特徴量（ここでは、メルスペクトログラム）

σ_{WG}^{2}：ガウス分布の予測分散値

なお、ｚ（ｘ）は、ガウス分布Ｎ（μ，σ）（μは平均値でありμ＝０、σは標準偏差）に従うガウス確率変数に相当するものである。すなわち、ｚ～Ｎ（μ，σ）＝Ｎ（０，σ）である。また、ｍ１は、アフィンカップリング層の処理の回数、ｍ２は、可逆１×１畳み込み層の処理の回数であり、本変形例のボコーダ６では、ｍ１＝ｍ２＝ｍである。

[0147] 本変形例のボコーダ６では、下記数式に相当する処理を実行することで、本変形例のボコーダ６のニューラルネットワークのモデルの最適化パラメータθ_{opt}を取得する。

[数2]

$$\boldsymbol{\theta_{opt}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p_{\boldsymbol{\theta}}(\boldsymbol{x})$$

本変形例のボコーダ６では、上記の学習処理により取得した最適化パラメータθ_{opt}により、ニューラルネットワークのモデルのパラメータが設定され（各可逆処理部６３ｂ～６３ｘのアフィンカップリング層、可逆１×１畳み込み層のパラメータが設定され）、学習済みモデルが構築される。

[0148] 次に、本変形例のボコーダ６の予測処理について、説明する。

[0149] 本変形例のボコーダ６は、予測処理において、音響特徴量としてメルスペクトログラム（これをデータｈとする）と、当該メルスペクトログラムの情報量Ｉに相関のあるデータを標準偏差σとし、平均値を「０」とするガウス白色ノイズｚとを入力とする。

[0150] 本変形例のボコーダ６では、予測処理時において、図５に示すように、学習処理時とは、逆の処理が実行される。

[0151] メルスペクトログラムのデータ（例えば、デコーダ部５から出力されるデータＤｙ６）がアップサンプリング処理部６２に入力される。

[0152] アップサンプリング処理部６２は、音響特徴量としてメルスペクトログラムのデータｈを入力し、入力されたメルスペクトログラムのデータｈに対して、アップサンプリング処理を実行し、処理後のデータ（アップサンプリングされたメルスペクトログラムのデータ）をデータｈ１として、可逆処理部６３ａ～６３ｘのそれぞれのＷＮ変換部６３２に出力する。

[0153] また、ガウス白色ノイズｚ（データｚという）が可逆処理部６３ｘに入力される。

[0154] そして、可逆処理部６３ｘにおいて、入力されたデータｚに対して、アフィンカップリング層の処理、可逆１×１畳み込み層の層の処理が実行される。この処理が、図５に示すように、ｍ回繰り返し実行される。各処理は、同様であるので、可逆処理部６３ａでの処理について、説明する。

[0155] データ合成部６３４では、可逆処理部６３ｂから出力されるデータDx'_2を入力し、学習処理時とは逆の処理、すなわち、データ分割処理を実行する。つまり、データ合成部６３４では、

$x = Dx'_2$

$x_a, x_b' = split(x)$

ｓｐｌｉｔ（）：データ分割をする関数

に相当する処理を実行し、入力データｘを２分割し、分割データx_aとx_b'を取得する。

[0156] そして、データ合成部６３４は、取得したデータx_aをＭＮ変換部６３２およびデータ分割部６３１に出力し、データx_b'をアフィン変換部６３３に出力する。

[0157] ＭＮ変換部６３２は、データ合成部６３４から出力されるデータx_aと、アップサンプリング処理部６２から出力されるアップサンプリングされたメルスペクトログラムのデータｈ１とを入力する。そして、ＭＮ変換部６３２は、データx_aと、データｈ１とに対して、任意の変換であるＭＮ変換（例えば、ＷａｖｅＮｅｔによる変換）を実行し、アフィン変換のパラメータとするデータs_j，t_j（s_j：アフィン変換用の行列、t_j：アフィン変換用のオフ

セット）を取得する。取得されたアフィン変換のパラメータとするデータs_j，t_jは、ＷＮ変換部６３２からアフィン変換部６３３に出力される。

[0158] アフィン変換部６３３は、ＭＮ変換部６３２により取得されたデータs_j，t_jを用いて、データ合成部６３４から入力されるデータx'_bに対して、アフィン逆変換（学習処理時に行ったアフィン変換の逆変換）を行う。つまり、アフィン変換部６３３は、

$$x_b = Affine^{-1}(s_j, t_j, x_b')$$

に相当する処理を実行することで、データx_b'のアフィン逆変換後のデータx_bを取得し、取得したデータx_bをデータ分割部６３１に出力する。

[0159] データ分割部６３１は、データ合成部６３４から出力されるデータx_aと、アフィン変換部６３３から出力されるデータx_bとを入力し、データx_aと、データx_bとを合成する処理、すなわち、

$$Dx'_1 = concat(x_a, x_b)$$

に相当する処理を実行し、データDx'_1を取得する。そして、データ分割部６３１は、取得したデータDx'_1を出力する。

[0160] 上記のようにして可逆処理部６３ｘ～６３ａにより処理されることで取得されたデータDx'_1が、ベクトル処理部６１に入力される。

[0161] ベクトル処理部６１は、学習処理時と逆の処理を実行することで、データDx'_1から、予測音声信号波形データｘを取得し、出力する。

[0162] 以上のように処理することで、本変形例のボコーダ６では、入力ｚ（ガウス白色ノイズｚ）と、メルスペクトログラムのデータｈから、予測音声信号波形データｘを取得することができる。

[0163] 本変形例のボコーダ６では、ニューラルネットワークを可逆変換できる構成を採用している。このため、本変形例のボコーダ６では、（１）ガウス白色ノイズが入力されたときに出力される音声波形データの尤度と、（２）音声波形データが入力されたときに出力されるガウス白色ノイズの尤度とを等価にし、学習処理を行いやすい（計算が容易である）後者（音声波形データが入力されたときに出力されるガウス白色ノイズの尤度）により、学習処理

を行うことで、効率良く学習処理を行うことができる。

[0164] そして、本変形例のボコーダ６では、ニューラルネットワークを可逆変換できる構成を有しているので、上記学習処理により取得した学習済みモデルにより、予測処理を、学習処理時とは逆の処理（逆変換）により実現できる。

[0165] このように、本変形例のボコーダ６では、音響特徴量としてメルスペクトログラムのデータから音声波形データを直接予測（取得）できる構成をシンプルな構成で実現できる。そして、本変形例のボコーダ６では、このようなシンプルな構成を有しているので、処理精度を保ちながら、予測処理を高速に行うことができ、音声合成処理をリアルタイムで実行することが可能になる。

[0166] 図６は、本変形例の音声合成処理装置によりＴＴＳ処理（処理対象言語：日本語）実行し、取得した音声波形データのメルスペクトログラム（予測データ）と、入力テキストの実際の音声波形データのメルスペクトログラム（オリジナルデータ）とを示す図である。

[0167] 図６から分かるように、本変形例の音声合成処理装置によりＴＴＳ処理では、非常に高精度な音声波形データが予測（取得）できる。

[0168] ［第２実施形態］

次に、第２実施形態について、説明する。なお、上記実施形態と同様の部分については、同一符号を付し、詳細な説明を省略する。

[0169] 第１実施形態では、エンコーダ・デコーダ方式（sequence-to-sequence方式）を用いたの音声合成処理装置１００について、説明した。第１実施形態の音声合成処理装置１００は、注意機構（アテンション部４）を備えており、音素継続長と音響モデルとを注意機構を用いて同時に最適化するニューラル音声合成処理を実現することができる。これにより、第１実施形態の音声合成処理装置１００では、自然音声クラスの高音質なテキスト音声合成を実現できる。しかしながら、第１実施形態の音声合成処理装置１００では、推論時（予測処理時）に、まれに注意機構予測が失敗することがあり、これに

より合成発話が途中で止まってしまう、同じフレーズを何回も繰り返してしまう、等の問題がある。

[0170] 第２実施形態では、上記問題を解決するための技術について、説明する。

[0171] ＜２．１：音声合成処理装置の構成＞

図７は、第２実施形態に係る音声合成処理装置２００の概略構成図である。

[0172] 第２実施形態に係る音声合成処理装置２００は、第１実施形態の音声合成処理装置１００において、アテンション部４を削除し、音素継続長推定部７を追加した構成を有している。そして、第２実施形態に係る音声合成処理装置２００は、第１実施形態の音声合成処理装置１００において、テキスト解析部１をテキスト解析部１Ａに置換し、フルコンテキストラベルベクトル処理部２をフルコンテキストラベルベクトル処理部２Ａに置換し、デコーダ部５をデコーダ部５Ａに置換した構成を有している。

[0173] テキスト解析部１Ａは、第１実施形態のテキスト解析部１と同様の機能を有しており、さらに、音素のコンテキストラベルを取得する機能を有している。テキスト解析部１Ａは、処理対象言語のテキストデータＤｉｎから音素のコンテキストラベルを取得し、取得した音素のコンテキストラベルのデータをデータＤｘ０１として、音素継続長推定部７に出力する。

[0174] 音素継続長推定部７は、テキスト解析部１Ａから出力されるデータＤｘ０１（音素のコンテキストラベルのデータ）を入力する。音素継続長推定部７は、データＤｘ０１（音素のコンテキストラベルのデータ）から、データＤｘ０１に対応する音素の音素継続長を推定（取得）する音素継続長推定処理を実行する。具体的には、音素継続長推定部７は、例えば、隠れマルコフモデル（ＨＭＭ：Ｈｉｄｄｅｎ　Ｍａｒｋｏｖ　Ｍｏｄｅｌ）、ニューラルネットワークモデル等を用いた、音素のコンテキストラベルから当該音素の音素継続長を推定（予測）するモデル（処理システム）により、音素継続長推定処理を実行する。

[0175] そして、音素継続長推定部７は、音素継続長推定処理により取得（推定）

した音素継続長のデータをデータＤｘ０２として、フルコンテキストラベルベクトル処理部２Ａに出力する。

[0176] フルコンテキストラベルベクトル処理部２Ａは、第１実施形態のフルコンテキストラベルベクトル処理部２と同様の機能を有しており、さらに、音素継続長推定部７により推定された音素継続長に相当する期間において、当該音素継続長に対応する音素についての最適化フルコンテキストラベルデータをエンコーダ部３に継続して出力する機能を有する。

[0177] フルコンテキストラベルベクトル処理部２Ａは、テキスト解析部１から出力されるデータＤｘ１（フルコンテキストラベルのデータ）と、音素継続長推定部７から出力されるデータＤｘ０２（音素の音素継続長のデータ）とを入力する。

[0178] フルコンテキストラベルベクトル処理部２Ａは、テキスト解析部１Ａから出力されるデータＤｘ１（フルコンテキストラベルのデータ）を入力する。フルコンテキストラベルベクトル処理部２Ａは、入力されたフルコンテキストラベルデータＤｘ１から、sequence-to-sequence方式のニューラルネットワークのモデルの学習処理に適したフルコンテキストラベルデータを取得するためのフルコンテキストラベルベクトル処理を実行する。そして、フルコンテキストラベルベクトル処理部２Ａは、フルコンテキストラベルベクトル処理により取得したデータをデータＤｘ２（最適化フルコンテキストラベルデータＤｘ２）として、エンコーダ部３のエンコーダ側プレネット処理部３１に出力する。このとき、フルコンテキストラベルベクトル処理部２Ａは、音素継続長推定部７により推定された音素継続長に相当する期間において、当該音素継続長に対応する音素についての最適化フルコンテキストラベルデータをエンコーダ部３に継続して出力する。

[0179] デコーダ部５Ａは、第１実施形態のデコーダ部５において、デコーダ側ＬＳＴＭ層５２をデコーダ側ＬＳＴＭ層５２Ａに置換した構成を有している。それ以外は、デコーダ部５Ａは、第１実施形態のデコーダ部５と同様である。

[0180] デコーダ側ＬＳＴＭ層５２Ａは、デコーダ側ＬＳＴＭ層５２と同様の機能を有している。デコーダ側ＬＳＴＭ層５２Ａは、デコーダ側プレネット処理部５１から、現時刻ｔにおいて出力されるデータＤｙ２（これをデータＤｙ２（ｔ）と表記する）と、１つ前の時間ステップにおいて、デコーダ側ＬＳＴＭ層５２Ａから出力されたデータＤｙ３（これをデータＤｙ３（ｔ－１）と表記する）と、エンコーダ部３から出力される時刻ｔの入力側隠れ状態データｈｉ（ｔ）を入力する。

[0181] デコーダ側ＬＳＴＭ層５２Ａは、入力されたデータＤｙ２（ｔ）、データＤｙ３（ｔ－１）、および、入力側隠れ状態データｈｉ（ｔ）を用いて、ＬＳＴＭ層による処理を実行し、処理後のデータをデータＤｙ３（データＤｙ３（ｔ））として線形予測部５３に出力する。

[0182] ＜２．２：音声合成処理装置の動作＞

以上のように構成された音声合成処理装置２００の動作について以下説明する。

[0183] 図８は、推定された音素継続長に基づいて、エンコーダ部３に入力するデータＤｘ２を生成する処理を説明するための図である。

[0184] 以下では、音声合成処理装置２００の動作を、（１）学習処理（学習時の処理）と、（２）予測処理（予測時の処理）とに分けて説明する。

[0185] （２．２．１：学習処理）

まず、音声合成処理装置２００による学習処理について、説明する。なお、説明便宜のため、処理対象言語を日本語として、以下、説明する。

[0186] 処理対象言語である日本語のテキストデータＤｉｎをテキスト解析部１Ａに入力する。また、当該テキストデータＤｉｎに対応するメルスペクトログラム（音響特徴量）のデータを教師データとして用意する。

[0187] テキスト解析部１Ａは、第１実施形態と同様に、入力されたテキストデータＤｉｎに対して、テキスト解析処理を実行し、様々な言語情報からなるコンテキストを含む音素ラベルであるコンテキストラベルの系列を取得する。

[0188] テキスト解析部１Ａは、第１実施形態と同様に、取得したフルコンテキス

トラベルデータをデータＤｘ１としてフルコンテキストラベルベクトル処理部２に出力する。

[0189] また、テキスト解析部１Ａは、処理対象言語のテキストデータＤｉｎから音素のコンテキストラベルを取得し、取得した音素のコンテキストラベルのデータをデータＤｘ０１として、音素継続長推定部７に出力する。

[0190] 音素継続長推定部７は、テキスト解析部１Ａから出力されるデータＤｘ０１（音素のコンテキストラベルのデータ）から、データＤｘ０１に対応する音素の音素継続長を推定（取得）する音素継続長推定処理を実行する。具体的には、音素継続長推定部７は、例えば、隠れマルコフモデル（ＨＭＭ：Ｈｉｄｄｅｎ　Ｍａｒｋｏｖ　Ｍｏｄｅｌ）、ニューラルネットワークモデル等を用いた、音素のコンテキストラベルから当該音素の音素継続長を推定（予測）するモデル（処理システム）により、音素継続長推定処理を実行する。

[0191] そして、音素継続長推定部７は、音素継続長推定処理により取得（推定）した音素継続長のデータをデータＤｘ０２として、フルコンテキストラベルベクトル処理部２Ａに出力する。

[0192] フルコンテキストラベルベクトル処理部２Ａは、テキスト解析部１Ａから出力されるデータＤｘ１（フルコンテキストラベルのデータ）から、sequence-to-sequence方式のニューラルネットワークのモデルの学習処理に適したフルコンテキストラベルデータを取得するためのフルコンテキストラベルベクトル処理（第１実施形態と同様のフルコンテキストラベルベクトル処理）を実行する。そして、フルコンテキストラベルベクトル処理部２Ａは、フルコンテキストラベルベクトル処理により取得したデータをデータＤｘ２（最適化フルコンテキストラベルデータＤｘ２）として、エンコーダ部３のエンコーダ側プレネット処理部３１に出力する。このとき、フルコンテキストラベルベクトル処理部２Ａは、音素継続長推定部７により推定された音素継続長に相当する期間において、当該音素継続長に対応する音素についての最適化フルコンテキストラベルデータをエンコーダ部３に継続して出力する。

[0193] フルコンテキストラベルベクトル処理部２Ａにより取得されたデータＤｘ２（最適化フルコンテキストラベルデータＤｘ２）は、フルコンテキストラベルベクトル処理部２からのエンコーダ部３のエンコーダ側プレネット処理部３１に出力される。

[0194] エンコーダ側プレネット処理部３１は、フルコンテキストラベルベクトル処理部２から入力したデータＤｘ２に対して、コンボリューション処理（コンボリューションフィルタによる処理）、データの正規化処理、活性化関数による処理（例えば、ＲｅＬＵ関数（ＲｅＬＵ：Ｒｅｃｔｉｆｉｅｄ Ｌｉｎｅａｒ Ｕｎｉｔ）による処理）を実行し、エンコーダ側ＬＳＴＭ層３２に入力可能なデータを取得する。そして、エンコーダ側プレネット処理部３１は、上記処理（プレネット処理）により取得したデータをデータＤｘ３としてエンコーダ側ＬＳＴＭ層３２に出力する。

[0195] エンコーダ側ＬＳＴＭ層３２は、エンコーダ側プレネット処理部３１から、現時刻ｔにおいて出力されるデータＤｘ３（ｔ）と、１つ前の時間ステップにおいて、エンコーダ側ＬＳＴＭ層３２から出力されたデータＤｘ４（ｔ－１）とを入力する。そして、エンコーダ側ＬＳＴＭ層３２は、入力されたデータＤｘ３（ｔ）、データＤｘ４（ｔ－１）に対して、ＬＳＴＭ層による処理を実行し、処理後のデータをデータＤｘ４（データＤｘ４（ｔ）（＝入力側隠れ状態データｈｉ（ｔ））としてデコーダ部５Ａのデコーダ側ＬＳＴＭ層５２Ａに出力する。

[0196] デコーダ側プレネット処理部５１は、線形予測部５３から出力される、１時間ステップ前のデータＤｙ４（ｔ－１）を入力する。デコーダ側プレネット処理部５１は、例えば、複数層（例えば、２層）の全結合層を有しており、データの正規化処理（例えば、線形予測部５３から出力されるデータ（ベクトルデータ）の次元数が２Ｎであり、デコーダ側ＬＳＴＭ層に入力されるデータ（ベクトルデータ）の次元数がＮである場合、データの次元数をＮにするように、例えば、ドロップアウト処理を行うことを含む）、活性化関数による処理（例えば、ＲｅＬＵ関数（ＲｅＬＵ：Ｒｅｃｔｉｆｉｅｄ Ｌｉ

ｎｅａｒ　Ｕｎｉｔ）による処理）を実行し、デコーダ側ＬＳＴＭ層５２に入力可能なデータを取得する。そして、デコーダ側プレネット処理部５１は、上記処理（プレネット処理）により取得したデータをデータＤｙ２としてデコーダ側ＬＳＴＭ層５２に出力する。

[0197] デコーダ側ＬＳＴＭ層５２Ａは、デコーダ側プレネット処理部５１から、現時刻ｔにおいて出力されるデータＤｙ２（ｔ）と、１つ前の時間ステップにおいて、デコーダ側ＬＳＴＭ層５２から出力されたデータＤｙ３（ｔ－１）と、エンコーダ部３から出力される時刻ｔの入力側隠れ状態データｈｉ（ｔ）（＝Ｄｘ４（ｔ））とを入力する。

[0198] デコーダ側ＬＳＴＭ層５２Ａは、入力されたデータＤｙ２（ｔ）、データＤｙ３（ｔ－１）、および、入力側隠れ状態データｈｉ（ｔ）を用いて、ＬＳＴＭ層による処理を実行し、処理後のデータをデータＤｙ３（ｔ）として線形予測部５３に出力する。

線形予測部５３、ポストネット処理部５４、および、加算器５５では、第１実施形態と同様の処理が実行される。

[0199] そして、音声合成処理装置２００では、上記のように取得されたデータＤｙ６（予測メルスペクトログラムのデータ）と、テキストデータＤｉｎに対応するメルスペクトログラム（音響特徴量）の教師データ（正解のメルスペクトログラム）とを比較し、両者の差（比較結果）（例えば、差分ベクトルのノルムやユークリッド距離により表現する差）が小さくなるように、エンコーダ部３、デコーダ部５Ａのニューラルネットワークのモデルのパラメータを更新する。音声合成処理装置１００では、このパラメータ更新処理を繰り返し実行し、データＤｙ６（予測メルスペクトログラムのデータ）と、テキストデータＤｉｎに対応するメルスペクトログラム（音響特徴量）の教師データ（正解のメルスペクトログラム）との差が十分小さくなる（所定の誤差範囲におさまる）、ニューラルネットワークのモデルのパラメータを最適化パラメータとして取得する。

[0200] 音声合成処理装置２００では、上記のようにして取得した最適化パラメー

タに基づいて、エンコーダ部３、デコーダ部５Ａのニューラルネットワークのモデルの各層に含まれるシナプス間の結合係数（重み係数）を設定することで、エンコーダ部３、デコーダ部５Ａのニューラルネットワークのモデルを最適化モデル（学習済みモデル）とすることができる。

[0201] 以上により、音声合成処理装置２００において、入力をテキストデータとし、出力をメルスペクトログラムとするニューラルネットワークの学習済みモデル（最適化モデル）を構築できる。

[0202] なお、音声合成処理装置２００において、第１実施形態の音声合成処理装置１００における学習処理により取得したニューラルネットワークの学習済みモデル（最適化モデル）を用いてもよい。つまり、音声合成処理装置２００において、第１実施形態の音声合成処理装置１００における学習処理により取得したニューラルネットワークの学習済みモデルのエンコーダ部３およびデコーダ部５の最適パラメータを用いて、音声合成処理装置２００のエンコーダ部３およびデコーダ部５Ａのパラメータを設定することで、音声合成処理装置２００において、学習済みモデルを構築するようにしてもよい。

[0203] また、ボコーダ６として、ニューラルネットワークによるモデルを用いたボコーダを採用する場合、その学習処理は、第１実施形態と同様である。

[0204] これにより、第１実施形態と同様に、ボコーダ６において、入力をテキストデータとし、出力をメルスペクトログラムとするニューラルネットワークの学習済みモデル（最適化モデル）を構築できる。

[0205] なお、音声合成処理装置２００において、（１）エンコーダ部３、デコーダ部５Ａの学習処理と、（２）ボコーダ６の学習処理とを連携させて学習処理を実行してもよいし、上記のように、個別に学習処理を実行してもよい。音声合成処理装置２００において、（１）エンコーダ部３、デコーダ部５Ａの学習処理と、（２）ボコーダ６の学習処理とを連携させて学習処理を実行する場合、入力をテキストデータとし、当該テキストデータに対応する音声波形データ（正解の音声波形データ）とを用いて、（１）エンコーダ部３、デコーダ部５Ａのニューラルネットワークのモデルと、（２）ボコーダ６の

ニューラルネットワークのモデルの最適化パラメータを取得することで学習処理を実行すればよい。

[0206] （２．２．２：予測処理）

次に、音声合成処理装置２００による予測処理について、説明する。なお、予測処理においても、説明便宜のため、処理対象言語を日本語として、以下、説明する。

[0207] 予測処理を実行する場合、音声合成処理装置２００では、上記の学習処理により取得された学習済みモデル、すなわち、エンコーダ部３、デコーダ部５Ａのニューラルネットワークの最適化モデル（最適化パラメータが設定されているモデル）、および、ボコーダ６のニューラルネットワークの最適化モデル（最適化パラメータが設定されているモデル）が構築されている。そして、音声合成処理装置２００では、当該学習済みモデルを用いて予測処理が実行される。

[0208] 音声合成処理の対象とする日本語のテキストデータＤｉｎをテキスト解析部１Ａに入力する。

[0209] テキスト解析部１Ａは、入力されたテキストデータＤｉｎに対して、日本語用のテキスト解析処理を実行し、例えば、図２に示すパラメータを含む４７８次元のベクトルデータとして、フルコンテキストラベルデータＤｘ１を取得する。

[0210] そして、取得されたフルコンテキストラベルデータＤｘ１は、テキスト解析部１Ａからフルコンテキストラベルベクトル処理部２Ａに出力される。

[0211] また、テキスト解析部１Ａは、処理対象言語のテキストデータＤｉｎから音素のコンテキストラベルを取得し、取得した音素のコンテキストラベルのデータをデータＤｘ０１として、音素継続長推定部７に出力する。

[0212] 音素継続長推定部７は、テキスト解析部１Ａから出力されるデータＤｘ０１（音素のコンテキストラベルのデータ）から、データＤｘ０１に対応する音素の音素継続長を推定（取得）する音素継続長推定処理を実行する。具体的には、音素継続長推定部７は、例えば、隠れマルコフモデル（ＨＭＭ：Ｈ

ｉｄｄｅｎ　Ｍａｒｋｏｖ　Ｍｏｄｅｌ）、ニューラルネットワークモデル等を用いた、音素のコンテキストラベルから当該音素の音素継続長を推定（予測）するモデル（処理システム）により、音素継続長推定処理を実行する。

[0213] 例えば、図８に示すように、入力データＤｉｎが「今日の天気は．．．」である場合、データＤｘ０１に含まれる各音素のデータを、

（１）ph_0＝「ｋ」、（２）ph_1＝「ｙ」、（３）ph_2＝「ｏｕ」、（４）ph_3＝「ｎ」、（５）ph_{04}＝「ｏ」、（６）ph_{sil}＝無音状態、（７）ph_5＝「ｔ」、（８）ph_6＝「ｅ」、（９）ph_{07}＝「ｎ」、・・・

とし、音素ph_k（ｋ：整数）の推定された音素継続長を$dur(ph_k)$とすると、音素継続長推定部７は、音素ph_k（ｋ：整数）のコンテキストラベルを用いて、音素継続長推定処理を実行することで、音素ph_kの推定された音素継続長$dur(ph_k)$を取得する。例えば、上記の各音素（音素ph_k）について、音素継続長推定部７により取得（推定）された音素継続長$dur(ph_k)$が、図８に示す時間の長さ（継続長）を有するものとする。

[0214] そして、音素継続長推定部７は、音素継続長推定処理により取得（推定）した音素継続長のデータ（図８の場合、$dur(ph_k)$）をデータＤｘ０２として、フルコンテキストラベルベクトル処理部２Ａに出力する。

[0215] フルコンテキストラベルベクトル処理部２Ａは、入力されたフルコンテキストラベルデータＤｘ１に対して、フルコンテキストラベルベクトル処理を実行し、最適化フルコンテキストラベルＤｘ２を取得する。なお、ここで取得される最適化フルコンテキストラベルＤｘ２は、エンコーダ部３、デコーダ部５Ａのsequence-to-sequence方式のニューラルネットワークのモデルの学習処理を行うときに設定した最適化フルコンテキストラベルデータＤｘ２と同じ次元数を有し、かつ、同じパラメータ（情報）を有するデータである。

[0216] 上記により取得されたデータＤｘ２（最適化フルコンテキストラベルデータＤｘ２）は、フルコンテキストラベルベクトル処理部２からエンコーダ部

３のエンコーダ側プレネット処理部３１に出力される。このとき、フルコンテキストラベルベクトル処理部２Ａは、音素継続長推定部７により推定された音素継続長に相当する期間において、当該音素継続長に対応する音素についての最適化フルコンテキストラベルデータをエンコーダ部３に継続して出力する。例えば、図８に示すように、音素ph_kについての最適化フルコンテキストラベルデータをデータＤｘ２（ph_k）とすると、フルコンテキストラベルベクトル処理部２Ａは、音素ph_kについての最適化フルコンテキストラベルデータＤｘ２（ph_k）を、当該音素ph_kの推定された音素継続長ｄｕｒ（ph_k）に相当する期間において、継続してエンコーダ部３に出力する。

[0217] つまり、音素ph_kについての最適化フルコンテキストラベルデータＤｘ２（ph_k）は、推定された音素継続長ｄｕｒ（ph_k）に相当する期間、繰り返しエンコーダ部３に出力される。すなわち、フルコンテキストラベルベクトル処理部２Ａでは、推定された音素継続長ｄｕｒ（ph_k）に基づいて、エンコーダ部３へ入力するデータ（最適化フルコンテキストラベルデータＤｘ２（ph_k））の時間引き延ばし処理が実行される。

[0218] エンコーダ側プレネット処理部３１は、フルコンテキストラベルベクトル処理部２Ａから入力したデータＤｘ２に対して、コンボリューション処理（コンボリューションフィルタによる処理）、データの正規化処理、活性化関数による処理（例えば、ＲｅＬＵ関数（ＲｅＬＵ：Ｒｅｃｔｉｆｉｅｄ Ｌｉｎｅａｒ Ｕｎｉｔ）による処理）を実行し、エンコーダ側ＬＳＴＭ層３２に入力可能なデータを取得する。そして、エンコーダ側プレネット処理部３１は、上記処理（プレネット処理）により取得したデータをデータＤｘ３としてエンコーダ側ＬＳＴＭ層３２に出力する。

[0219] エンコーダ側ＬＳＴＭ層３２は、エンコーダ側プレネット処理部３１から、現時刻ｔにおいて出力されるデータＤｘ３（ｔ）と、１つ前の時間ステップにおいて、エンコーダ側ＬＳＴＭ層３２から出力されたデータＤｘ４（ｔ－１）とを入力する。そして、エンコーダ側ＬＳＴＭ層３２は、入力されたデータＤｘ３（ｔ）、データＤｘ４（ｔ－１）に対して、ＬＳＴＭ層による

処理（ニューラルネットワーク処理）を実行し、処理後のデータをデータＤｘ４（データＤｘ４（ｔ）（＝入力側隠れ状態データｈｉ（ｔ））としてデコーダ部５Ａのデコーダ側ＬＳＴＭ層５２Ａに出力する。

[0220] デコーダ側ＬＳＴＭ層５２Ａは、入力されたデータＤｙ２（ｔ）、データＤｙ３（ｔ－１）、および、入力側隠れ状態データｈｉ（ｔ）を用いて、ＬＳＴＭ層による処理を実行し、処理後のデータをデータＤｙ３（ｔ）として線形予測部５３に出力する。

線形予測部５３、ポストネット処理部５４、および、加算器５５では、第１実施形態と同様の処理が実行される。

[0221] ボコーダ６は、デコーダ部５Ａの加算器５５から出力されるデータＤｙ６（予測メルスペクトログラムのデータ（音響特徴量のデータ））を入力とし、入力されたデータＤｙ６に対して、学習済みモデルを用いたニューラルネットワーク処理による音声合成処理を実行し、データＤｙ６（予測メルスペクトログラム）に対応する音声信号波形データを取得する。そして、ボコーダ６は、取得した音声信号波形データを、データＤｏｕｔとして出力する。

[0222] このように、音声合成処理装置２００では、入力されたテキストデータＤｉｎに対応する音声波形データＤｏｕｔを取得することができる。

[0223] 以上のように、音声合成処理装置２００では、処理対象言語（上記では日本語）のテキストを入力とし、当該処理対象言語に応じたテキスト解析処理により、フルコンテキストラベルデータを取得し、取得したフルコンテキストラベルデータからsequence-to-sequence方式を用いたニューラルネットワークのモデルで処理（学習処理、および／または、予測処理）を実行するのに適したデータである最適化フルコンテキストラベルデータを取得する。そして、音声合成処理装置２００では、入力を最適化フルコンテキストラベルデータとし、出力をメルスペクトログラム（音響特徴量の一例）として、エンコーダ部３、および、デコーダ部５Ａにおいて、ニューラルネットワークのモデルを用いた処理（学習処理、予測処理）を実行することで、高精度な処理を実現できる。さらに、音声合成処理装置２００では、ボコーダ６によ

り、上記により取得したメルスペクトログラム（音響特徴量の一例）から、当該メルスペクトログラムに対応する音声信号波形データを取得し、取得したデータを出力することで、音声波形データ（データＤｏｕｔ）を取得する。これにより、音声合成処理装置２００では、入力されたテキストに相当する音声波形データを取得することができる。

[0224] さらに、音声合成処理装置２００では、エンコーダ部３への入力データ（最適化フルコンテキストラベルデータ）を、音素継続長推定部７により取得（推定）した音素ごとの音素継続長に基づいて、引き延ばす処理（音素ph_kの音素継続長ｄｕｒ（ph_k）に相当する期間、音素ph_kの最適化フルコンテキストラベルデータを、繰り返しエンコーダ部３に入力する処理）を実行する。つまり、音声合成処理装置２００では、安定して音素継続長を適切に推定することができる、隠れマルコフモデル等のモデルを用いた推定処理を実行して取得した音素継続長を用いて予測処理を実行するので、注意機構予測が失敗することに起因する、合成発話が途中で止まってしまう、同じフレーズを何回も繰り返してしまう、等の問題が発生することはない。

[0225] すなわち、音声合成処理装置２００では、（１）音素継続長については、安定して音素継続長を適切に推定することができる、隠れマルコフモデル等のモデルを用いた推定処理（音素継続長推定部７による処理）により取得し、（２）音響特徴量については、sequence-to-sequence方式を用いたニューラルネットワークのモデルで処理することにより取得する。

[0226] したがって、音声合成処理装置２００では、注意機構予測が失敗することに起因する、合成発話が途中で止まってしまう、同じフレーズを何回も繰り返してしまう、等の問題が発生することを適切に防止するとともに、高精度な音声合成処理を実行することができる。

[0227] ［第３実施形態］

次に、第３実施形態について、説明する。なお、上記実施形態と同様の部分については、同一符号を付し、詳細な説明を省略する。

[0228] ＜３．１：音声合成処理装置の構成＞

図９は、第３実施形態に係る音声合成処理装置３００の概略構成図である。

[0229] 第３実施形態に係る音声合成処理装置３００は、第１実施形態の音声合成処理装置１００において、テキスト解析部１をテキスト解析部１Ａに置換し、アテンション部４をアテンション部４Ａに置換し、デコーダ部５をデコーダ部５Ｂに置換した構成を有している。そして、音声合成処理装置３００は、音声合成処理装置１００において、音素継続長推定部７と、強制アテンション部８と、内分処理部９と、コンテキスト算出部１０とを追加した構成を有している。

[0230] テキスト解析部１Ａ、および、音素継続長推定部７は、第２実施形態のテキスト解析部１Ａと同様の構成、機能を有している。

[0231] なお、音素継続長推定部７は、音素継続長推定処理により取得（推定）した音素継続長のデータをデータＤｘ０２として、強制アテンション部８に出力する。

[0232] アテンション部４Ａは、エンコーダ部３から出力されるデータＤｘ４と、デコーダ部５Ｂのデコーダ側ＬＳＴＭ層５２Ｂから出力されるデータｈｏ（出力側隠れ状態データｈｏ）とを入力する。アテンション部４Ａは、エンコーダ部３から出力されるデータＤｘ４、すなわち、入力側隠れ状態データｈｉを所定の時間ステップ分記憶保持する。時間ステップｔ＝１からｔ＝Ｓ（Ｓ：自然数）の期間において、エンコーダ部３により取得され、アテンション部４Ａに出力されたデータＤｘ４（＝ｈｉ）の集合を、$hi_{1...S}$と表記する。つまり、アテンション部４Ａは、下記に相当するデータ$hi_{1...S}$を記憶保持する。

$hi_{1...S}$＝｛Ｄｘ４（１），Ｄｘ４（２），・・・，Ｄｘ４（Ｓ）｝

また、アテンション部４Ａは、デコーダ部５Ｂのデコーダ側ＬＳＴＭ層５２Ｂから出力されるデータＤｙ３、すなわち、出力側隠れ状態データｈｏを所定の時間ステップ分記憶保持する。時間ステップｔ＝１からｔ＝Ｔ（Ｔ：自然数）の期間において、デコーダ側ＬＳＴＭ層５２Ｂにより取得され、ア

テンション部４Ａに出力されたデータＤｙ３（＝ｈｏ）の集合を、$ho_{1...T}$と表記する。つまり、アテンション部４Ａは、下記に相当するデータ$ho_{1...T}$を記憶保持する。

$ho_{1...T}$＝｛Ｄｙ３（１），Ｄｙ３（２），・・・，Ｄｙ３（Ｔ）｝

そして、アテンション部４Ａは、入力側隠れ状態データの集合データ$hi_{1...S}$と、出力側隠れ状態データの集合データ$ho_{1...T}$と、に基づいて、例えば、

$w_{att}(t)_{1...S}$＝ｆ２＿ａｔｔｎ（$hi_{1...S}$，$ho_{1...T}$）

ｆ２＿ａｔｔｎ（）：重み付け係数データを取得する関数

に相当する処理を実行して、現時刻ｔの重み付け係数データ$w_{att}(t)_{1...S}$を取得する。そして、アテンション部４Ａは、取得した重み付け係数データ$w_{att}(t)_{1...S}$を内分処理部９に出力する。なお、入力側隠れ状態データの集合データ$hi_{1...S}$の各要素データに対する重み付け係数データの集合データを重み付け係数データ$w_{att}(t)_{1...S}$と表記する。

[0233] また、アテンション部４Ａは、データＤｘ４（＝ｈｉ）の集合データ$hi_{1...S}$をコンテキスト算出部１０に出力する。

[0234] 強制アテンション部８は、音素継続長推定部７から出力される推定された音素継続長のデータＤｘ０２を入力する。強制アテンション部８は、音素継続長データＤｘ０２に対応する音素についてのエンコーダ部３により処理されたデータが出力されるとき、当該音素の推定された音素継続長（音素継続長データＤｘ０２）に相当する期間、重み付け係数を強制的に所定の値（例えば、「１」）にした重み付け係数データ$w_f(t)$を生成する。なお、入力側隠れ状態データの集合データ$hi_{1...S}$の各要素データに対する重み付け係数データと対応づけるために、時刻ｔを中心として、Ｓ個にデータを拡張（同一データを複製して拡張）した重み付け係数データ$w_f(t)$を重み付け係数データ$w_f(t)_{1...S}$と表記する。

[0235] 強制アテンション部８は、上記により生成した重み付け係数データ$w_f(t)_{1...S}$を内分処理部９に出力する。

[0236] 内分処理部９は、アテンション部４Ａから出力される重み付け係数データ$w_{att}(t)_{1...S}$と、強制アテンション部８から出力される重み付け係数データ$w_f(t)_{1...S}$とを入力する。そして、内分処理部９は、重み付け係数データ$w_{att}(t)_{1...S}$と、重み付け係数データ$w_f(t)_{1...S}$とに対して、内分処理を実行することで、合成重み付け係数データｗ（ｔ）を取得する。具体的には、内分処理部９は、

$$w(t)_{1...S}=(1-\alpha)\times w_{att}(t)_{1...S}+\alpha\times w_f(t)_{1...S}$$

$$0\leqq\alpha\leqq1$$

に相当する処理を実行することで、合成重み付け係数データｗ（ｔ）を取得する。なお、上記数式（内分処理）は、それぞれ対応する要素ごとに、内分処理を実行することを表している。つまり、ｊ番目（１≦ｊ≦Ｓ）のデータについては、

$$w(t)_j=(1-\alpha)\times w_{att}(t)_j+\alpha\times w_f(t)_j$$

に相当する処理が実行されることで、ｊ番目の合成重み付け係数データｗ（ｔ）$_j$が取得される。

[0237] そして、内分処理部９は、取得した合成重み付け係数データｗ（ｔ）$_{1...S}$をコンテキスト算出部１０に出力する。

[0238] コンテキスト算出部１０は、アテンション部４Ａから出力されるデータＤｘ４（＝ｈｉ）の集合データ$hi_{1...S}$と、内分処理部９から出力される合成重み付け係数データｗ（ｔ）$_{1...S}$とを入力する。そして、コンテキスト算出部１０は、合成重み付け係数データｗ（ｔ）$_{1...S}$に基づいて、データＤｘ４（＝ｈｉ）の集合データ$hi_{1...S}$に対して、重み付け加算処理を実行することで、コンテキスト状態データｃ（ｔ）を取得する。そして、コンテキスト算出部１０は、取得したコンテキスト状態データｃ（ｔ）をデコーダ部５Ｂのデコーダ側ＬＳＴＭ層５２Ｂに出力する。

[0239] デコーダ部５Ｂは、第１実施形態のデコーダ部５において、デコーダ側ＬＳＴＭ層５２をデコーダ側ＬＳＴＭ層５２Ｂに置換した構成を有している。

それ以外は、デコーダ部５Ｂは、第１実施形態のデコーダ部５と同様である。

[0240] デコーダ側ＬＳＴＭ層５２Ｂは、デコーダ側ＬＳＴＭ層５２と同様の機能を有している。デコーダ側ＬＳＴＭ層５２Ｂは、デコーダ側プレネット処理部５１から、現時刻ｔにおいて出力されるデータＤｙ２（これをデータＤｙ２（ｔ）と表記する）と、１つ前の時間ステップにおいて、デコーダ側ＬＳＴＭ層５２Ｂから出力されたデータＤｙ３（これをデータＤｙ３（ｔ－１）と表記する）と、コンテキスト算出部１０から出力される時刻ｔのコンテキスト状態データｃ（ｔ）とを入力する。

[0241] デコーダ側ＬＳＴＭ層５２Ｂは、入力されたデータＤｙ２（ｔ）、データＤｙ３（ｔ－１）、および、コンテキスト状態データｃ（ｔ）を用いて、ＬＳＴＭ層による処理を実行し、処理後のデータをデータＤｙ３（データＤｙ３（ｔ））として線形予測部５３に出力する。また、デコーダ側ＬＳＴＭ層５２Ｂは、データＤｙ３（ｔ）、すなわち、時刻ｔの出力側隠れ状態データｈｏ（ｔ）をアテンション部４Ａに出力する。

[0242] ＜３．２：音声合成処理装置の動作＞

以上のように構成された音声合成処理装置３００の動作について以下説明する。

[0243] 図１０～図１２は、アテンション部４Ａにより取得された重み付け係数データw_{att}（ｔ）と、強制アテンション部８により取得された重み付け係数データw_{f}（ｔ）とから取得した合成重み付け係数データｗ（ｔ）を用いてコンテキスト状態データｃ（ｔ）を取得する処理について説明するための図である。

[0244] （３．２．１：学習処理）

まず、音声合成処理装置３００による学習処理について、説明する。なお、説明便宜のため、処理対象言語を日本語として、以下、説明する。

[0245] 処理対象言語である日本語のテキストデータＤｉｎをテキスト解析部１Ａに入力する。また、当該テキストデータＤｉｎに対応するメルスペクトログ

ラム（音響特徴量）のデータを教師データとして用意する。

[0246] テキスト解析部１Ａは、第１実施形態と同様に、入力されたテキストデータＤｉｎに対して、テキスト解析処理を実行し、様々な言語情報からなるコンテキストを含む音素ラベルであるコンテキストラベルの系列を取得する。

[0247] テキスト解析部１Ａは、第１実施形態と同様に、取得したフルコンテキストラベルデータをデータＤｘ１としてフルコンテキストラベルベクトル処理部２に出力する。

[0248] また、テキスト解析部１Ａは、処理対象言語のテキストデータＤｉｎから音素のコンテキストラベルを取得し、取得した音素のコンテキストラベルのデータをデータＤｘ０１として、音素継続長推定部７に出力する。

[0249] 音素継続長推定部７は、テキスト解析部１Ａから出力されるデータＤｘ０１（音素のコンテキストラベルのデータ）から、データＤｘ０１に対応する音素の音素継続長を推定（取得）する音素継続長推定処理を実行する。具体的には、音素継続長推定部７は、例えば、隠れマルコフモデル（ＨＭＭ：Ｈｉｄｄｅｎ　Ｍａｒｋｏｖ　Ｍｏｄｅｌ）、ニューラルネットワークモデル等を用いた、音素のコンテキストラベルから当該音素の音素継続長を推定（予測）するモデル（処理システム）により、音素継続長推定処理を実行する。

[0250] そして、音素継続長推定部７は、音素継続長推定処理により取得（推定）した音素継続長のデータをデータＤｘ０２として、強制アテンション部８に出力する。

[0251] フルコンテキストラベルベクトル処理部２Ａは、テキスト解析部１Ａから出力されるデータＤｘ１（フルコンテキストラベルのデータ）から、sequence-to-sequence方式のニューラルネットワークのモデルの学習処理に適したフルコンテキストラベルデータを取得するためのフルコンテキストラベルベクトル処理（第１実施形態と同様のフルコンテキストラベルベクトル処理）を実行する。そして、フルコンテキストラベルベクトル処理部２Ａは、フルコンテキストラベルベクトル処理により取得したデータをデータＤｘ２（最適

化フルコンテキストラベルデータＤｘ２）として、エンコーダ部３のエンコーダ側プレネット処理部３１に出力する。

[0252] フルコンテキストラベルベクトル処理部２Ａにより取得されたデータＤｘ２（最適化フルコンテキストラベルデータＤｘ２）は、フルコンテキストラベルベクトル処理部２からのエンコーダ部３のエンコーダ側プレネット処理部３１に出力される。

[0253] エンコーダ側プレネット処理部３１は、フルコンテキストラベルベクトル処理部２から入力したデータＤｘ２に対して、コンボリューション処理（コンボリューションフィルタによる処理）、データの正規化処理、活性化関数による処理（例えば、ＲｅＬＵ関数（ＲｅＬＵ：Ｒｅｃｔｉｆｉｅｄ Ｌｉｎｅａｒ Ｕｎｉｔ）による処理）を実行し、エンコーダ側ＬＳＴＭ層３２に入力可能なデータを取得する。そして、エンコーダ側プレネット処理部３１は、上記処理（プレネット処理）により取得したデータをデータＤｘ３としてエンコーダ側ＬＳＴＭ層３２に出力する。

[0254] エンコーダ側ＬＳＴＭ層３２は、エンコーダ側プレネット処理部３１から、現時刻ｔにおいて出力されるデータＤｘ３（ｔ）と、１つ前の時間ステップにおいて、エンコーダ側ＬＳＴＭ層３２から出力されたデータＤｘ４（ｔ－１）とを入力する。そして、エンコーダ側ＬＳＴＭ層３２は、入力されたデータＤｘ３（ｔ）、データＤｘ４（ｔ－１）に対して、ＬＳＴＭ層による処理を実行し、処理後のデータをデータＤｘ４（データＤｘ４（ｔ）（＝入力側隠れ状態データｈｉ（ｔ））としてアテンション部４Ａに出力する。

[0255] アテンション部４Ａは、エンコーダ部３から出力されるデータＤｘ４と、デコーダ部５Ｂのデコーダ側ＬＳＴＭ層５２Ｂから出力されるデータｈｏ（出力側隠れ状態データｈｏ）とを入力する。アテンション部４Ａは、エンコーダ部３から出力されるデータＤｘ４、すなわち、入力側隠れ状態データｈｉを所定の時間ステップ分記憶保持する。例えば、アテンション部４Ａは、時間ステップｔ＝１からｔ＝Ｓ（Ｓ：自然数）の期間において、エンコーダ部３により取得され、アテンション部４Ａに出力されたデータＤｘ４（＝ｈｉ）

の集合を、$hi_{1...S}$（＝｛Ｄｘ４（１），Ｄｘ４（２），・・・，Ｄｘ４（Ｓ）｝）として記憶保持する。

[0256] また、アテンション部４Ａは、デコーダ部５Ｂのデコーダ側ＬＳＴＭ層５２Ｂから出力されるデータＤｙ３、すなわち、出力側隠れ状態データｈｏを所定の時間ステップ分記憶保持する。例えば、アテンション部４Ａは、時間ステップｔ＝１からｔ＝Ｔ（Ｔ：自然数）の期間において、デコーダ側ＬＳＴＭ層５２Ｂにより取得され、アテンション部４Ａに出力されたデータＤｙ３（＝ｈｏ）の集合を、$ho_{1...T}$（＝｛Ｄｙ３（１），Ｄｙ３（２），・・・，Ｄｙ３（Ｔ）｝）として記憶保持する。

[0257] そして、アテンション部４Ａは、入力側隠れ状態データの集合データ$hi_{1...S}$と、出力側隠れ状態データの集合データ$ho_{1...T}$と、に基づいて、例えば、

$w_{att}(t)_{1...S}$＝ｆ２＿ａｔｔｎ（$hi_{1...S}$，$ho_{1...T}$）

ｆ２＿ａｔｔｎ（）：重み付け係数データを取得する関数

に相当する処理を実行して、現時刻ｔの重み付け係数データ$w_{att}(t)_{1...S}$を取得する。

[0258] そして、アテンション部４Ａは、取得した重み付け係数データ$w_{att}(t)_{1...S}$を内分処理部９に出力する。また、アテンション部４Ａは、データＤｘ４（＝ｈｉ）の集合データ$hi_{1...S}$をコンテキスト算出部１０に出力する。

[0259] 強制アテンション部８は、音素継続長データＤｘ０２に対応する音素についてのエンコーダ部３により処理されたデータが出力されるとき、当該音素の推定された音素継続長（音素継続長データＤｘ０２）に相当する期間、重み付け係数を強制的に所定の値（例えば、「１」）にした重み付け係数データ$w_f(t)$を生成する。そして、強制アテンション部８は、入力側隠れ状態データの集合データ$hi_{1...S}$の各要素データに対する重み付け係数データと対応づけるために（内分処理ができるようにするために）、時刻ｔを中心として、Ｓ個にデータを拡張（同一データを複製して拡張）した重み付け係数データ$w_f(t)_{1...S}$を生成する。

[0260] 強制アテンション部８は、上記により生成した重み付け係数データ$w_f(t)_{1...S}$を内分処理部９に出力する。

[0261] 内分処理部９は、アテンション部４Ａから出力される重み付け係数データ$w_{att}(t)_{1...S}$と、強制アテンション部８から出力される重み付け係数データ$w_f(t)_{1...S}$とを入力する。そして、内分処理部９は、重み付け係数データ$w_{att}(t)_{1...S}$と、重み付け係数データ$w_f(t)_{1...S}$とに対して、内分処理を実行することで、合成重み付け係数データｗ（ｔ）を取得する。具体的には、内分処理部９は、

$$w(t)_{1...S}=(1-\alpha)\times w_{att}(t)_{1...S}+\alpha\times w_f(t)_{1...S}$$

$$0\leqq\alpha\leqq1$$

に相当する処理を実行することで、合成重み付け係数データ$w(t)_{1...S}$を取得する。そして、内分処理部９は、取得した合成重み付け係数データ$w(t)_{1...S}$をコンテキスト算出部１０に出力する。

[0262] コンテキスト算出部１０は、アテンション部４Ａから出力されるデータＤｘ４（＝ｈｉ）の集合データ$hi_{1...S}$と、内分処理部９から出力される合成重み付け係数データ$w(t)_{1...S}$とを入力する。そして、コンテキスト算出部１０は、合成重み付け係数データ$w(t)_{1...S}$に基づいて、データＤｘ４（＝ｈｉ）の集合データ$hi_{1...S}$に対して、重み付け加算処理を実行することで、コンテキスト状態データｃ（ｔ）を取得する。そして、コンテキスト算出部１０は、取得したコンテキスト状態データｃ（ｔ）をデコーダ部５Ｂのデコーダ側ＬＳＴＭ層５２Ｂに出力する。

[0263] なお、学習処理時において、内分比αを「０」に固定してもよい。この場合（内分比αを「０」に固定した場合）、音声合成処理装置３００では、第１実施形態と同様の構成により学習処理が実行されることになる。また、学習処理時において、内分比αを所定の値（例えば、０．５）に固定して、音声合成処理装置３００において、学習処理を実行してもよい。

[0264] ここで、学習処理時において、内分比αを所定の値に固定する場合につい

て、図１０～図１２を用いて説明する。なお、説明便宜のため、内分比αを「０．５」に固定する場合について、説明する。以下では、（１）音素に対応する音声が出力される期間内の処理（図１１の場合）と、（２）無音状態である期間内の処理（図１２の場合）とについて説明する。

[0265] まず、「（１）音素に対応する音声が出力される期間内の処理（図１１の場合）」について、説明する。

[0266] 例えば、図１０に示すように、入力データＤｉｎが「今日の天気は．．．」である場合、データＤｘ０１に含まれる各音素のデータを、

（１）ph_0＝「ｋ」、（２）ph_1＝「ｙ」、（３）ph_2＝「ｏｕ」、（４）ph_3＝「ｎ」、（５）ph_{04}＝「ｏ」、（６）ph_{sil}＝無音状態、（７）ph_5＝「ｔ」、（８）ph_6＝「ｅ」、（９）ph_{07}＝「ｎ」、・・・

とし、音素ph_k（ｋ：整数）の推定された音素継続長をｄｕｒ（ph_k）とすると、音素継続長推定部７は、音素ph_k（ｋ：整数）のコンテキストラベルを用いて、音素継続長推定処理を実行することで、音素ph_kの推定された音素継続長ｄｕｒ（ph_k）を取得する。例えば、上記の各音素（音素ph_k）について、音素継続長推定部７により取得（推定）された音素継続長ｄｕｒ（ph_k）が、図１０に示す時間の長さ（継続長）を有するものとする。

[0267] 強制アテンション部８は、音素継続長データＤｘ０２に対応する音素についてのエンコーダ部３により処理されたデータが出力されるとき、当該音素の推定された音素継続長（音素継続長データＤｘ０２）に相当する期間、重み付け係数を強制的に所定の値（例えば、「１」）にした重み付け係数データw_f（ｔ）を生成する。図１０の場合、強制アテンション部８は、音素ph_kについてのエンコーダ部３により処理されたデータが出力されるとき、音素ph_kの音素継続長ｄｕｒ（ph_k）に相当する期間、重み付け係数を強制的に所定の値（例えば、「１」）にした重み付け係数データw_f（ｔ）を内分処理部９に出力し続ける（図１０において、w_f（ｔ）［ph_k］と表記した部分に相当）。

[0268] また、図１０において、処理対象の音素に対応付けて、アテンション部４

Ａにより取得された重み付け係数データw_{att}（ｔ）を示している。具体的には、図１０において、音素ph_kに対応する、アテンション部４Ａにより取得された重み付け係数データw_{att}（ｔ）が出力される期間を「w_{att}（ｔ）［ph_k］」として示している。なお、説明便宜のため、図１０では、アテンション部４Ａによる音素継続長の予測が正しくなされた場合を示している。

[0269] また、図１０において、音素ph_kに対応する合成重み付け係数データｗ（ｔ）を「ｗ（ｔ）［ph_k］」として示している。

[0270] 図１１は、時刻ｔ２（時間ステップｔ２）における処理を説明するための図であり、図１０において処理対象音素が「ｏｕ」であるときの期間の一部を時間軸方向に拡大して示した図である。なお、説明便宜のため、音声合成処理装置３００において、データＤｘ４（＝ｈｉ）の集合データ$hi_{1...S}$は、９個のデータ（すなわち、Ｓ＝９）（図１１において、期間Ｔ（ｔ２）において取得され、記憶保持されているデータ）であるものとする（以下、同様）。

[0271] ここで、時刻ｔ２における処理について、説明する。

[0272] 強制アテンション部８は、時刻ｔ２において、音素継続長Ｄｘ０２から、音素「ｏｕ」に相当する音声が出力継続される期間であることを認識し、時刻ｔ２の重み付け係数データw_f（ｔ）を「１」に設定する。さらに、強制アテンション部８は、入力側隠れ状態データの集合データ$hi_{1...S}$の各要素データに対する重み付け係数データと対応づけるために（内分処理ができるようにするために）、時刻ｔ２を中心として、Ｓ（＝９）個にデータを拡張（同一データを複製して拡張）した重み付け係数データw_f（ｔ）$_{1...S}$を生成する。なお、w_f（ｔ）$_{1...S}$は、

$$w_f(t)_{1...S} = \{w_{01}, w_{02}, w_{03}, w_{04}, w_{05}, w_{06}, w_{07}, w_{08}, w_{09}\}$$

$$0 \leqq w_{0j} \leqq 1 \quad (1 \leqq j \leqq S)$$

$$t = t2$$

であるものとし、w_f（ｔ２）$_{1...S}$において、w_{01}～w_{09}は、すべて「１」

に設定される（図１１参照）。

[0273] 強制アテンション部８は、上記により生成した重み付け係数データw_f（ｔ２）$_{1...S}$を内分処理部９に出力する。

[0274] アテンション部４Ａは、入力側隠れ状態データの集合データ$hi_{1...S}$と、出力側隠れ状態データの集合データ$ho_{1...T}$と、に基づいて、例えば、

w_{att}（ｔ）$_{1...S}$＝ｆ２＿ａｔｔｎ（$hi_{1...S}$，$ho_{1...T}$）

ｆ２＿ａｔｔｎ（）：重み付け係数データを取得する関数

に相当する処理を実行して、時刻ｔ２の重み付け係数データw_{att}（ｔ２）$_{1...S}$を取得する。時刻ｔ２の重み付け係数データw_{att}（ｔ）$_{1...S}$が図１１に示すデータ（一例）であるものとする。なお、w_{att}（ｔ）$_{1...S}$は、

w_{att}（ｔ）$_{1...S}$＝｛w_{11}，w_{12}，w_{13}，w_{14}，w_{15}，w_{16}，w_{17}，w_{18}，w_{19}｝

$0 \leqq w_{1j} \leqq 1$（$1 \leqq j \leqq S$）

ｔ＝ｔ２

であるものとし、w_{11}～w_{19}は、例えば、アテンション部４Ａにより、以下の値として、取得されたものとする（図１１参照）。

w_{11}＝０．０、w_{12}＝０．２、w_{13}＝０．４、w_{14}＝０．８、w_{15}＝１．０

w_{16}＝０．８、w_{17}＝０．４、w_{18}＝０．２、w_{19}＝０．０

アテンション部４Ａは、上記により取得された重み付け係数データw_{att}（ｔ２）$_{1...S}$を内分処理部９に出力する。

[0275] 内分処理部９は、アテンション部４Ａから出力される重み付け係数データw_{att}（ｔ２）$_{1...S}$と、強制アテンション部８から出力される重み付け係数データw_f（ｔ２）$_{1...S}$とを入力する。そして、内分処理部９は、重み付け係数データw_{att}（ｔ２）$_{1...S}$と、重み付け係数データw_f（ｔ２）$_{1...S}$とに対して、内分処理を実行することで、合成重み付け係数データｗ（ｔ２）$_{1...S}$を取得する。具体的には、内分処理部９は、

ｗ（ｔ２）$_{1...S}$＝（１－α）×w_{att}（ｔ２）$_{1...S}$＋α×w_f（ｔ２

$)_{1...S}$

$0 \leqq \alpha \leqq 1$

に相当する処理を実行することで、合成重み付け係数データｗ（ｔ２）$_{1...S}$を取得する。

[0276] ここでは、$\alpha = 0.5$であるので、$w_{att}(t2)_{1...S}$と、$w_f(t2)_{1...S}$との平均値が合成重み付け係数データｗ（ｔ）$_{1...S}$となる。なお、ｗ（ｔ）$_{1...S}$は、

$w(t)_{1...S} = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9\}$

$0 \leqq w_{1j} \leqq 1$ （$1 \leqq j \leqq S$）

ｔ＝ｔ２

であるものとすると、w_1～w_9は、内分処理部９により、以下の値として、取得される（図１１参照）。

$w_1 = 0.5 \times w_{01} + 0.5 \times w_{11} = 0.5 + 0 = 0.5$

$w_2 = 0.5 \times w_{02} + 0.5 \times w_{12} = 0.5 + 0.1 = 0.6$

$w_3 = 0.5 \times w_{03} + 0.5 \times w_{13} = 0.5 + 0.2 = 0.7$

$w_4 = 0.5 \times w_{04} + 0.5 \times w_{14} = 0.5 + 0.4 = 0.9$

$w_5 = 0.5 \times w_{05} + 0.5 \times w_{15} = 0.5 + 0.5 = 1.0$

$w_6 = 0.5 \times w_{06} + 0.5 \times w_{16} = 0.5 + 0.4 = 0.9$

$w_7 = 0.5 \times w_{07} + 0.5 \times w_{17} = 0.5 + 0.2 = 0.7$

$w_8 = 0.5 \times w_{08} + 0.5 \times w_{18} = 0.5 + 0.1 = 0.6$

$w_9 = 0.5 \times w_{09} + 0.5 \times w_{19} = 0.5 + 0 = 0.5$

そして、内分処理部９は、取得した合成重み付け係数データｗ（ｔ２）$_{1...S}$をコンテキスト算出部１０に出力する。

[0277] コンテキスト算出部１０は、アテンション部４Ａから出力されるデータＤｘ４（＝ｈｉ）の集合データｈｉ$_{1...S}$と、内分処理部９から出力される合成重み付け係数データｗ（ｔ２）$_{1...S}$とを入力する。そして、コンテキスト算出部１０は、合成重み付け係数データｗ（ｔ２）$_{1...S}$に基づいて、データＤｘ４（＝ｈｉ）の集合データｈｉ$_{1...S}$に対して、重み付け加算処理を実行す

ることで、コンテキスト状態データｃ（ｔ）を取得する。つまり、コンテキスト算出部１０は、以下の数式に相当する処理を実行することで、コンテキスト状態データｃ（ｔ）を取得する。

[数3]

$$c(t) = \sum_{j=1}^{S} (w_j \times hi_j)$$

ｔ＝ｔ２

w_j：合成重み付け係数データｗ（ｔ２）$_{1...S}$のｊ番目の要素データ（１≦ｊ≦Ｓ）

そして、コンテキスト算出部１０は、取得したコンテキスト状態データｃ（ｔ２）をデコーダ部５Ｂのデコーダ側ＬＳＴＭ層５２Ｂに出力する。

[0278] 次に、「（２）無音状態である期間内の処理（図１２の場合）」について、説明する。

[0279] 図１２は、時刻ｔ３（時間ステップｔ３）における処理を説明するための図であり、図１０において無音状態の期間（図１０において、「ｓｉｌｅｎｔ（無音状態）」で示した期間）の一部を時間軸方向に拡大して示した図である。

[0280] ここで、時刻ｔ３における処理について、説明する。

[0281] 強制アテンション部８は、時刻ｔ３において、音素継続長Ｄｘ０２から、無音状態（発声すべき音素がない状態）の期間であることを認識し、時刻ｔ３の重み付け係数データw_f（ｔ）を「０」に設定する。さらに、強制アテンション部８は、入力側隠れ状態データの集合データｈｉ$_{1...S}$の各要素データに対する重み付け係数データと対応づけるために（内分処理ができるようにするために）、時刻ｔ２を中心として、Ｓ（＝９）個にデータを拡張（同一データを複製して拡張）した重み付け係数データw_f（ｔ）$_{1...S}$を生成する

。なお、$w_f(t)_{1...S}$は、

$w_f(t)_{1...S}=\{w_{01}, w_{02}, w_{03}, w_{04}, w_{05}, w_{06}, w_{07}, w_{08}, w_{09}\}$

$0 \leqq w_{0j} \leqq 1$ （$1 \leqq j \leqq S$）

$t=t2$

であるものとし、$w_f(t3)_{1...S}$において、w_{01}～w_{09}は、すべて「０」に設定される（図１２参照）。

[0282] 強制アテンション部８は、上記により生成した重み付け係数データ$w_f(t3)_{1...S}$を内分処理部９に出力する。

[0283] アテンション部４Ａは、入力側隠れ状態データの集合データ$hi_{1...S}$と、出力側隠れ状態データの集合データ$ho_{1...T}$と、に基づいて、例えば、

$w_{att}(t)_{1...S}=f2\_attn(hi_{1...S}, ho_{1...T})$

f２_ａｔｔｎ（）：重み付け係数データを取得する関数

に相当する処理を実行して、時刻ｔ３の重み付け係数データ$w_{att}(t3)_{1...S}$を取得する。時刻ｔ３の重み付け係数データ$w_{att}(t)_{1...S}$が図１２に示すデータ（一例）であるものとする。なお、$w_{att}(t)_{1...S}$は、

$w_{att}(t)_{1...S}=\{w_{11}, w_{12}, w_{13}, w_{14}, w_{15}, w_{16}, w_{17}, w_{18}, w_{19}\}$

$0 \leqq w_{1j} \leqq 1$ （$1 \leqq j \leqq S$）

$t=t2$

であるものとし、w_{11}～w_{19}は、例えば、アテンション部４Ａにより、すべて値が「０」として、取得されたものとする（図１２参照）。

[0284] アテンション部４Ａは、上記により取得された重み付け係数データ$w_{att}(t3)_{1...S}$を内分処理部９に出力する。

[0285] 内分処理部９は、アテンション部４Ａから出力される重み付け係数データ$w_{att}(t3)_{1...S}$と、強制アテンション部８から出力される重み付け係数データ$w_f(t3)_{1...S}$とを入力する。そして、内分処理部９は、重み付け係数データ$w_{att}(t3)_{1...S}$と、重み付け係数データ$w_f(t3)_{1...S}$

とに対して、内分処理を実行することで、合成重み付け係数データｗ（ｔ３）$_{1...S}$を取得する。具体的には、内分処理部９は、

$$w(t3)_{1...S}=(1-\alpha)\times w_{att}(t3)_{1...S}+\alpha\times w_{f}(t3)_{1...S}$$

$$0\leqq\alpha\leqq1$$

に相当する処理を実行することで、合成重み付け係数データｗ（ｔ３）$_{1...S}$を取得する。

[0286] ここでは、$\alpha=0.5$であるので、$w_{att}(t3)_{1...S}$と、$w_{f}(t3)_{1...S}$との平均値が合成重み付け係数データｗ（ｔ）$_{1...S}$となる。なお、ｗ（ｔ）$_{1...S}$は、

$$w(t)_{1...S}=\{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9\}$$

$$0\leqq w_{1j}\leqq1\ (1\leqq j\leqq S)$$

$$t=t2$$

であるものとすると、w_1～w_9は、内分処理部９により、すべて値が「０」として、取得される（図１２参照）。

[0287] そして、内分処理部９は、取得した合成重み付け係数データｗ（ｔ２）$_{1...S}$をコンテキスト算出部１０に出力する。

[0288] コンテキスト算出部１０は、アテンション部４Ａから出力されるデータＤｘ４（＝ｈｉ）の集合データｈｉ$_{1...S}$と、内分処理部９から出力される合成重み付け係数データｗ（ｔ３）$_{1...S}$とを入力する。そして、コンテキスト算出部１０は、合成重み付け係数データｗ（ｔ２）$_{1...S}$に基づいて、データＤｘ４（＝ｈｉ）の集合データｈｉ$_{1...S}$に対して、重み付け加算処理を実行することで、コンテキスト状態データｃ（ｔ）を取得する。つまり、コンテキスト算出部１０は、以下の数式に相当する処理を実行することで、コンテキスト状態データｃ（ｔ）を取得する。

[数4]

$$c(t) = \sum_{j=1}^{S} (w_j \times hi_j)$$

ｔ＝ｔ２

w_j：合成重み付け係数データｗ（ｔ３）$_{1...S}$のｊ番目の要素データ（１≦ｊ≦Ｓ）

そして、コンテキスト算出部１０は、取得したコンテキスト状態データｃ（ｔ３）をデコーダ部５Ｂのデコーダ側ＬＳＴＭ層５２Ｂに出力する。

[0289] 図１２の場合、無音状態であるので、アテンション部４Ａ、および、強制アテンション部８により取得される重み付け係数データがすべて０であるので、コンテキスト状態データｃ（ｔ３）も「０」となる。つまり、上記により、無音状態であることを適切に示すコンテキスト状態データｃ（ｔ３）が取得される。

[0290] 上記のように取得されたコンテキスト状態データｃ（ｔ）がデコーダ部５Ｂのデコーダ側ＬＳＴＭ層５２Ｂに出力される。

[0291] そして、アテンション部４Ａは、取得した重み付け係数データw_{att}（ｔ）$_{1...S}$を内分処理部９に出力する。また、アテンション部４Ａは、データＤｘ４（＝ｈｉ）の集合データｈｉ$_{1...S}$をコンテキスト算出部１０に出力する。

[0292] デコーダ側プレネット処理部５１での処理は、第１実施形態と同様である。

[0293] デコーダ側ＬＳＴＭ層５２Ｂは、デコーダ側プレネット処理部５１から、現時刻ｔにおいて出力されるデータＤｙ２（ｔ）と、１つ前の時間ステップにおいて、デコーダ側ＬＳＴＭ層５２から出力されたデータＤｙ３（ｔ－１）と、コンテキスト算出部１０から出力される時刻ｔのコンテキスト状態データｃ（ｔ）とを入力する。

[0294] デコーダ側ＬＳＴＭ層５２Ａは、入力されたデータＤｙ２（ｔ）、データＤｙ３（ｔ－１）、および、コンテキスト状態データｃ（ｔ）を用いて、ＬＳＴＭ層による処理を実行し、処理後のデータをデータＤｙ３（ｔ）として線形予測部５３に出力する。

線形予測部５３、ポストネット処理部５４、および、加算器５５では、第１実施形態と同様の処理が実行される。

[0295] そして、音声合成処理装置２００では、上記のように取得されたデータＤｙ６（予測メルスペクトログラムのデータ）と、テキストデータＤｉｎに対応するメルスペクトログラム（音響特徴量）の教師データ（正解のメルスペクトログラム）とを比較し、両者の差（比較結果）（例えば、差分ベクトルのノルムやユークリッド距離により表現する差）が小さくなるように、エンコーダ部３、デコーダ部５Ｂのニューラルネットワークのモデルのパラメータを更新する。音声合成処理装置１００では、このパラメータ更新処理を繰り返し実行し、データＤｙ６（予測メルスペクトログラムのデータ）と、テキストデータＤｉｎに対応するメルスペクトログラム（音響特徴量）の教師データ（正解のメルスペクトログラム）との差が十分小さくなる（所定の誤差範囲におさまる）、ニューラルネットワークのモデルのパラメータを最適化パラメータとして取得する。

[0296] 音声合成処理装置３００では、上記のようにして取得した最適化パラメータに基づいて、エンコーダ部３、デコーダ部５Ｂのニューラルネットワークのモデルの各層に含まれるシナプス間の結合係数（重み係数）を設定することで、エンコーダ部３、デコーダ部５Ａのニューラルネットワークのモデルを最適化モデル（学習済みモデル）とすることができる。

[0297] 以上により、音声合成処理装置３００において、入力をテキストデータとし、出力をメルスペクトログラムとするニューラルネットワークの学習済みモデル（最適化モデル）を構築できる。

[0298] なお、音声合成処理装置３００において、第１実施形態の音声合成処理装置１００における学習処理により取得したニューラルネットワークの学習済

みモデル（最適化モデル）を用いてもよい。つまり、音声合成処理装置２００において、第１実施形態の音声合成処理装置１００における学習処理により取得したニューラルネットワークの学習済みモデルのエンコーダ部３およびデコーダ部５の最適パラメータを用いて、音声合成処理装置２００のエンコーダ部３およびデコーダ部５Ｂのパラメータを設定することで、音声合成処理装置３００において、学習済みモデルを構築するようにしてもよい。

[0299] また、ボコーダ６として、ニューラルネットワークによるモデルを用いたボコーダを採用する場合、その学習処理は、第１実施形態と同様である。

[0300] これにより、第１実施形態と同様に、ボコーダ６において、入力をテキストデータとし、出力をメルスペクトログラムとするニューラルネットワークの学習済みモデル（最適化モデル）を構築できる。

[0301] なお、音声合成処理装置３００において、（１）エンコーダ部３、デコーダ部５Ｂの学習処理と、（２）ボコーダ６の学習処理とを連携させて学習処理を実行してもよいし、上記のように、個別に学習処理を実行してもよい。音声合成処理装置３００において、（１）エンコーダ部３、デコーダ部５Ｂの学習処理と、（２）ボコーダ６の学習処理とを連携させて学習処理を実行する場合、入力をテキストデータとし、当該テキストデータに対応する音声波形データ（正解の音声波形データ）とを用いて、（１）エンコーダ部３、デコーダ部５Ｂのニューラルネットワークのモデルと、（２）ボコーダ６のニューラルネットワークのモデルの最適化パラメータを取得することで学習処理を実行すればよい。

[0302] （３．２．２：予測処理）

次に、音声合成処理装置３００による予測処理について、説明する。なお、予測処理においても、説明便宜のため、処理対象言語を日本語として、以下、説明する。

[0303] 予測処理を実行する場合、音声合成処理装置３００では、上記の学習処理により取得された学習済みモデル、すなわち、エンコーダ部３、デコーダ部５Ｂのニューラルネットワークの最適化モデル（最適化パラメータが設定さ

れているモデル）、および、ボコーダ６のニューラルネットワークの最適化モデル（最適化パラメータが設定されているモデル）が構築されている。そして、音声合成処理装置３００では、当該学習済みモデルを用いて予測処理が実行される。

[0304] 音声合成処理の対象とする日本語のテキストデータＤｉｎをテキスト解析部１Ａに入力する。

[0305] テキスト解析部１Ａは、入力されたテキストデータＤｉｎに対して、日本語用のテキスト解析処理を実行し、例えば、図２に示すパラメータを含む４７８次元のベクトルデータとして、フルコンテキストラベルデータＤｘ１を取得する。

[0306] そして、取得されたフルコンテキストラベルデータＤｘ１は、テキスト解析部１Ａからフルコンテキストラベルベクトル処理部２に出力される。

[0307] また、テキスト解析部１Ａは、処理対象言語のテキストデータＤｉｎから音素のコンテキストラベルを取得し、取得した音素のコンテキストラベルのデータをデータＤｘ０１として、音素継続長推定部７に出力する。

[0308] 音素継続長推定部７は、テキスト解析部１Ａから出力されるデータＤｘ０１（音素のコンテキストラベルのデータ）から、データＤｘ０１に対応する音素の音素継続長を推定（取得）する音素継続長推定処理を実行する。具体的には、音素継続長推定部７は、例えば、隠れマルコフモデル（ＨＭＭ：Ｈｉｄｄｅｎ　Ｍａｒｋｏｖ　Ｍｏｄｅｌ）、ニューラルネットワークモデル等を用いた、音素のコンテキストラベルから当該音素の音素継続長を推定（予測）するモデル（処理システム）により、音素継続長推定処理を実行する。

[0309] 例えば、図８に示すように、入力データＤｉｎが「今日の天気は．．．」である場合、データＤｘ０１に含まれる各音素のデータを、

（１）ph_0＝「ｋ」、（２）ph_1＝「ｙ」、（３）ph_2＝「ｏｕ」、（４）ph_3＝「ｎ」、（５）ph_{04}＝「ｏ」、（６）ph_{sil}＝無音状態、（７）ph_5＝「ｔ」、（８）ph_6＝「ｅ」、（９）ph_{07}＝「ｎ」、・・・

とし、音素ph_k（ｋ：整数）の推定された音素継続長をｄｕｒ（ph_k）とすると、音素継続長推定部７は、音素ph_k（ｋ：整数）のコンテキストラベルを用いて、音素継続長推定処理を実行することで、音素ph_kの推定された音素継続長ｄｕｒ（ph_k）を取得する。例えば、上記の各音素（音素ph_k）について、音素継続長推定部７により取得（推定）された音素継続長ｄｕｒ（ph_k）が、図８に示す時間の長さ（継続長）を有するものとする。

[0310] そして、音素継続長推定部７は、音素継続長推定処理により取得（推定）した音素継続長のデータ（図８の場合、ｄｕｒ（ph_k））をデータＤｘ０２として、強制アテンション部８に出力する。

[0311] 強制アテンション部８は、音素継続長データＤｘ０２に対応する音素についてのエンコーダ部３により処理されたデータが出力されるとき、当該音素の推定された音素継続長（音素継続長データＤｘ０２）に相当する期間、重み付け係数を強制的に所定の値（例えば、「１」）にした重み付け係数データw_f（ｔ）を生成する。そして、強制アテンション部８は、入力側隠れ状態データの集合データ$hi_{1...S}$の各要素データに対する重み付け係数データと対応づけるために（内分処理ができるようにするために）、時刻ｔを中心として、Ｓ個にデータを拡張（同一データを複製して拡張）した重み付け係数データ$w_f(t)_{1...S}$を生成する。

[0312] 強制アテンション部８は、上記により生成した重み付け係数データ$w_f(t)_{1...S}$を内分処理部９に出力する。

[0313] エンコーダ部３では、第１実施形態と同様の処理が実行される。

[0314] アテンション部４Ａは、エンコーダ部３から出力されるデータＤｘ４と、デコーダ部５Ｂのデコーダ側ＬＳＴＭ層５２Ｂから出力されるデータho（出力側隠れ状態データho）とを入力する。アテンション部４Ａは、エンコーダ部３から出力されるデータＤｘ４、すなわち、入力側隠れ状態データｈｉを所定の時間ステップ分記憶保持する。例えば、アテンション部４Ａは、時間ステップｔ＝１からｔ＝Ｓ（Ｓ：自然数）の期間において、エンコーダ部３により取得され、アテンション部４Ａに出力されたデータＤｘ４（＝ｈｉ）

の集合を、$hi_{1...S}$（＝｛Ｄｘ４（１），Ｄｘ４（２），・・・，Ｄｘ４（Ｓ）｝）として記憶保持する。

[0315] また、アテンション部４Ａは、デコーダ部５Ｂのデコーダ側ＬＳＴＭ層５２Ｂから出力されるデータＤｙ３、すなわち、出力側隠れ状態データｈｏを所定の時間ステップ分記憶保持する。例えば、アテンション部４Ａは、時間ステップｔ＝１からｔ＝Ｔ（Ｔ：自然数）の期間において、デコーダ側ＬＳＴＭ層５２Ｂにより取得され、アテンション部４Ａに出力されたデータＤｙ３（＝ｈｏ）の集合を、$ho_{1...T}$（＝｛Ｄｙ３（１），Ｄｙ３（２），・・・，Ｄｙ３（Ｔ）｝）として記憶保持する。

[0316] そして、アテンション部４Ａは、入力側隠れ状態データの集合データ$hi_{1...S}$と、出力側隠れ状態データの集合データ$ho_{1...T}$と、に基づいて、例えば、

$w_{att}(t)_{1...S}$＝ｆ２＿ａｔｔｎ（$hi_{1...S}$，$ho_{1...T}$）

ｆ２＿ａｔｔｎ（）：重み付け係数データを取得する関数

に相当する処理を実行して、現時刻ｔの重み付け係数データ$w_{att}(t)_{1...S}$を取得する。

[0317] そして、アテンション部４Ａは、取得した重み付け係数データ$w_{att}(t)_{1...S}$を内分処理部９に出力する。また、アテンション部４Ａは、データＤｘ４（＝ｈｉ）の集合データ$hi_{1...S}$をコンテキスト算出部１０に出力する。

[0318] 内分処理部９は、アテンション部４Ａから出力される重み付け係数データ$w_{att}(t)_{1...S}$と、強制アテンション部８から出力される重み付け係数データ$w_f(t)_{1...S}$とを入力する。そして、内分処理部９は、重み付け係数データ$w_{att}(t)_{1...S}$と、重み付け係数データ$w_f(t)_{1...S}$とに対して、内分処理を実行することで、合成重み付け係数データｗ（ｔ）を取得する。具体的には、内分処理部９は、

$w(t)_{1...S}=(1-\alpha)\times w_{att}(t)_{1...S}+\alpha\times w_f(t)_{1...S}$

$0 \leqq \alpha \leqq 1$

に相当する処理を実行することで、合成重み付け係数データｗ（ｔ）$_{1...S}$を取得する。そして、内分処理部９は、取得した合成重み付け係数データｗ（ｔ）$_{1...S}$をコンテキスト算出部１０に出力する。

[0319] コンテキスト算出部１０は、アテンション部４Ａから出力されるデータＤｘ４（＝ｈｉ）の集合データｈｉ$_{1...S}$と、内分処理部９から出力される合成重み付け係数データｗ（ｔ）$_{1...S}$とを入力する。そして、コンテキスト算出部１０は、合成重み付け係数データｗ（ｔ）$_{1...S}$に基づいて、データＤｘ４（＝ｈｉ）の集合データｈｉ$_{1...S}$に対して、重み付け加算処理を実行することで、コンテキスト状態データｃ（ｔ）を取得する。そして、コンテキスト算出部１０は、取得したコンテキスト状態データｃ（ｔ）をデコーダ部５Ｂのデコーダ側ＬＳＴＭ層５２Ｂに出力する。

[0320] デコーダ側ＬＳＴＭ層５２Ｂは、入力されたデータＤｙ２（ｔ）、データＤｙ３（ｔ－１）、および、コンテキスト状態データｃ（ｔ）を用いて、ＬＳＴＭ層による処理を実行し、処理後のデータをデータＤｙ３（ｔ）として線形予測部５３に出力する。

[0321] 線形予測部５３、ポストネット処理部５４、および、加算器５５では、第１実施形態と同様の処理が実行される。

[0322] ボコーダ６は、デコーダ部５Ｂの加算器５５から出力されるデータＤｙ６（予測メルスペクトログラムのデータ（音響特徴量のデータ））を入力とし、入力されたデータＤｙ６に対して、学習済みモデルを用いたニューラルネットワーク処理による音声合成処理を実行し、データＤｙ６（予測メルスペクトログラム）に対応する音声信号波形データを取得する。そして、ボコーダ６は、取得した音声信号波形データを、データＤｏｕｔとして出力する。

[0323] このように、音声合成処理装置３００では、入力されたテキストデータＤｉｎに対応する音声波形データＤｏｕｔを取得することができる。

[0324] 音声合成処理装置３００では、図１０～図１２を用いて説明したのと同様に、予測処理時においても、アテンション部４Ａにより取得された重み付け係数データｗ$_{att}$（ｔ）と、強制アテンション部８により取得された重み付け

係数データw_f（ｔ）とを内分処理により合成した重み付け係数データを用いて、コンテキスト状態データｃ（ｔ）を生成する。そして、音声合成処理装置３００では、上記のようにして生成されたコンテキスト状態データｃ（ｔ）を用いて、デコーダ部５Ｂ、ボコーダ６による処理が実行されるため、注意機構予測が失敗することに起因する、合成発話が途中で止まってしまう、同じフレーズを何回も繰り返してしまう、等の問題が発生することを適切に防止できる。

[0325] 例えば、図１３に示すように、時刻ｔ２においての処理で、注意機構の予測が失敗している場合、すなわち、図１３に示すように、アテンション部４により取得された重み付け係数データが「０」（あるいは所定の値以下）である場合（w_{att}（ｔ）$_{1...S}$のすべての要素データの値が「０」（あるいは所定の値以下）である場合）であっても、音声合成処理装置３００では、強制アテンション部８により取得された重み付け係数データw_f（ｔ）の重みにより、注意機構の予測の失敗が音声合成処理に影響を及ぼさないようにできる合成重み付け係数データｗ（ｔ）$_{1...S}$を取得することができる（図１３の場合。合成重み付け係数データｗ（ｔ）$_{1...S}$の各要素データの値は、すべて「０．５」）。

[0326] このように、音声合成処理装置３００では、音素継続長については、安定して音素継続長を適切に推定することができる、隠れマルコフモデル等のモデルを用いた推定処理（音素継続長推定部７による処理）により取得した音素継続長を用いて処理することで、音素継続長の予測精度を保証する。つまり、音声合成処理装置３００では、安定して音素継続長を適切に推定することができる、隠れマルコフモデル等のモデルを用いた推定処理（音素継続長推定部７による処理）により取得した音素継続長を用いて強制アテンション部８により取得した重み付け係数データと、アテンション部４Ａにより取得された重み付け係数データとを適度に合成した重み付け係数データにより生成したコンテキスト状態データｃ（ｔ）を用いて予測処理を実行する。したがって、音声合成処理装置３００では、注意機構の予測が失敗する場合（ア

テンション部４により適切な重み付け係数データが取得できない場合）であっても、強制アテンション部８により取得した重み付け係数データによる重み分の重み付け係数データが取得できるため、注意機構の予測の失敗が音声合成処理に影響を及ぼさないようにできる。

[0327] さらに、音声合成処理装置３００では、音響特徴量については、 sequence-to-sequence方式を用いたニューラルネットワークのモデルで処理することにより取得できるので、高精度な音響特徴量の予測処理が実現できる。

[0328] したがって、音声合成処理装置３００では、注意機構予測が失敗することに起因する、合成発話が途中で止まってしまう、同じフレーズを何回も繰り返してしまう、等の問題が発生することを適切に防止するとともに、高精度な音声合成処理を実行することができる。

[0329] なお、上記では、内分比αを固定値（例えば、０．５）に設定した場合について、説明したが、これに限定されることはなく、内分比αは動的に更新されるものであってもよい。例えば、内分処理部９において、アテンション部４Ａから入力される重み付け係数データ$w_{att}(t)_{1...S}$が所定の期間、継続して、所定の値よりも小さい、あるいは、略０であり、かつ、強制アテンション部８から入力される重み付け係数データ$w_f(t)_{1...S}$が「１」である場合、アテンション部４による処理が失敗している（注意機構予測が失敗している）と判定し、αの値をより大きな値（重み付け係数データ$w_f(t)_{1...S}$の重みが大きくなる値）に調整（更新）するようにしてもよい。

[0330] また、音声合成処理装置３００において、エンコーダ部３、デコーダ部５は、上記の構成に限定されるものではなく、他の構成のものであってよい。例えば、下記文献Ａに開示されているトランスフォーマーモデルのアーキテクチャによるエンコーダ、デコーダの構成を採用して、エンコーダ部３、デコーダ部５を構成するようにしてもよい。この場合、トランスフォーマーモデルのアーキテクチャによるエンコーダとデコーダの間に設置されるアテンション機構を、本実施形態で説明した機構、すなわち、アテンション部４、強制アテンション部８、内分処理部９、コンテキスト算出部１０により、ア

テンション機構が取得した重み付け係数データと、強制アテンション部８が取得した重み付け係数データとを内分処理により合成し、合成した重み付け係数データによりコンテキスト状態データを取得する機構に置換する構成を採用すればよい。

（文献Ａ）：A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, AN. Gomez, L. Kaiser, I. Polosukhin, “Attention is all you need” 31st Conference on Nural information Processing System (NIPS 2017), Long Beach, CA, USA.

［他の実施形態］

上記実施形態（変形例を含む）の音声合成処理装置において、エンコーダ側ＬＳＴＭ層３２、デコーダ側ＬＳＴＭ層５２は、それぞれ、複数のＬＳＴＭ層を備えるものであってもよい。また、エンコーダ側ＬＳＴＭ層３２、デコーダ側ＬＳＴＭ層５２は、それぞれ、双方向ＬＳＴＭ層（順伝搬、逆伝搬をＬＳＴＭ層）で構成されるものであってもよい。

[0331] また、上記実施形態（変形例を含む）では、音声合成処理装置が、テキスト解析部１と、フルコンテキストラベルベクトル処理部２とを備え、テキスト解析部１で取得したフルコンテキストラベルデータから、フルコンテキストラベルベクトル処理部２により、最適化フルコンテキストラベルデータを取得する場合について説明したが、これに限定されることはなく、例えば、音声合成処理装置において、最適化フルコンテキストラベルデータを取得する、テキスト解析部を設け、フルコンテキストラベルベクトル処理部を省略する構成としてもよい。

[0332] また、上記実施形態（変形例を含む）を適宜組み合わせてもよい。

[0333] また上記実施形態（変形例を含む）で説明した音声合成処理装置において、各ブロックは、ＬＳＩなどの半導体装置により個別に１チップ化されても良いし、一部または全部を含むように１チップ化されても良い。

[0334] なおここではＬＳＩとしたが、集積度の違いにより、ＩＣ、システムＬＳＩ、スーパーＬＳＩ、ウルトラＬＳＩと呼称されることもある。

[0335] また集積回路化の手法はＬＳＩに限るものではなく、専用回路または汎用プロセサで実現してもよい。ＬＳＩ製造後にプログラムすることが可能なＦＰＧＡ（Ｆｉｅｌｄ　Ｐｒｏｇｒａｍｍａｂｌｅ　Ｇａｔｅ　Ａｒｒａｙ）や、ＬＳＩ内部の回路セルの接続や設定を再構成可能なリコンフィギュラブル・プロセッサーを利用しても良い。

[0336] また上記各実施形態の各機能ブロックの処理の一部または全部は、プログラムにより実現されるものであってもよい。そして上記各実施形態の各機能ブロックの処理の一部または全部は、コンピュータにおいて、中央演算装置（ＣＰＵ）により行われる。また、それぞれの処理を行うためのプログラムは、ハードディスク、ＲＯＭなどの記憶装置に格納されており、ＲＯＭにおいて、あるいはＲＡＭに読み出されて実行される。

[0337] また上記実施形態の各処理をハードウェアにより実現してもよいし、ソフトウェア（ＯＳ（オペレーティングシステム）、ミドルウェア、あるいは所定のライブラリとともに実現される場合を含む。）により実現してもよい。さらにソフトウェアおよびハードウェアの混在処理により実現しても良い。

[0338] 例えば上記実施形態の各機能部をソフトウェアにより実現する場合、図１４に示したハードウェア構成（例えばＣＰＵ、ＧＰＵ、ＲＯＭ、ＲＡＭ、入力部、出力部、通信部、記憶部（例えば、ＨＤＤ、ＳＳＤ等により実現される記憶部）、外部メディア用ドライブ等をバスＢｕｓにより接続したハードウェア構成）を用いて各機能部をソフトウェア処理により実現するようにしてもよい。

[0339] また上記実施形態の各機能部をソフトウェアにより実現する場合、当該ソフトウェアは、図１４に示したハードウェア構成を有する単独のコンピュータを用いて実現されるものであってもよいし、複数のコンピュータを用いて分散処理により実現されるものであってもよい。

[0340] また上記実施形態における処理方法の実行順序は、必ずしも上記実施形態の記載に制限されるものではなく、発明の要旨を逸脱しない範囲で、実行順序を入れ替えることができるものである。

[0341] 前述した方法をコンピュータに実行させるコンピュータプログラム、及びそのプログラムを記録したコンピュータ読み取り可能な記録媒体は、本発明の範囲に含まれる。ここでコンピュータ読み取り可能な記録媒体としては、例えば、フレキシブルディスク、ハードディスク、ＣＤ－ＲＯＭ、ＭＯ、ＤＶＤ、ＤＶＤ－ＲＯＭ、ＤＶＤ－ＲＡＭ、大容量ＤＶＤ、次世代ＤＶＤ、半導体メモリを挙げることができる。

[0342] 上記コンピュータプログラムは、上記記録媒体に記録されたものに限らず、電気通信回線、無線または有線通信回線、インターネットを代表とするネットワーク等を経由して伝送されるものであってもよい。

[0343] なお本発明の具体的な構成は、前述の実施形態に限られるものではなく、発明の要旨を逸脱しない範囲で種々の変更および修正が可能である。

符号の説明

[0344] １００、２００、３００　音声合成処理装置

１　テキスト解析部

２、２Ａ　フルコンテキストラベルベクトル処理部

３　エンコーダ部

４、４Ａ　アテンション部

５　デコーダ部

６　ボコーダ

７　音素継続長推定部

８　強制アテンション部

９　内分処理部

１０　コンテキスト算出部

請求の範囲

[請求項1] 任意の言語を処理対象言語とし、エンコーダ・デコーダ方式のニューラルネットワークを用いて音声合成処理を実行する音声合成処理装置であって、

前記処理対象言語のテキストデータに対してテキスト解析処理を実行し、コンテキストラベルデータを取得するテキスト解析部と、

前記テキスト解析部により取得された前記コンテキストラベルデータから、コンテキストラベルデータを取得する処理において処理対象とされた音素である単独音素についてのコンテキストラベルを取得することで、前記ニューラルネットワークの学習処理に適した最適化フルコンテキストラベルデータを取得するフルコンテキストラベルベクトル処理部と、

前記最適化フルコンテキストラベルデータに基づいて、ニューラルネットワークのエンコード処理を実行することで、隠れ状態データを取得するエンコーダ部と、

前記隠れ状態データに基づいて、ニューラルネットワークのデコード処理を実行することで、前記最適化フルコンテキストラベルデータに対応する音響特徴量データを取得するデコーダ部と、

前記デコーダ部により取得された音響特徴量から音声波形データを取得するボコーダと、

を備える音声合成処理装置。

[請求項2] 前記音響特徴量は、メルスペクトログラムのデータである、

請求項１に記載の音声合成処理装置。

[請求項3] 前記ボコーダは、

ニューラルネットワークのモデルを用いた処理を実行することで、音響特徴量から音声波形データを取得する、

請求項１または２に記載の音声合成処理装置。

[請求項4] 前記ボコーダは、

可逆変換ネットワークにより構成されたニューラルネットワークのモデルを用いた処理を実行することで、音響特徴量から音声波形データを取得する、

請求項３に記載の音声合成処理装置。

[請求項5] 音素単位のコンテキストラベルデータから音素継続長を推定する音素継続長推定部をさらに備え、

前記フルコンテキストラベルベクトル処理部は、前記音素継続長推定部により推定された音素継続長である推定音素継続長に対応する期間において、当該推定音素継続長に対応する音素の前記最適化フルコンテキストラベルデータを継続して前記エンコーダ部に出力する、

請求項１から４のいずれかに記載の音声合成処理装置。

[請求項6] 任意の言語を処理対象言語とし、エンコーダ・デコーダ方式のニューラルネットワークを用いて音声合成処理を実行する音声合成処理方法であって、

前記処理対象言語のテキストデータに対してテキスト解析処理を実行し、コンテキストラベルデータを取得するテキスト解析ステップと、

前記テキスト解析ステップにより取得された前記コンテキストラベルデータから、コンテキストラベルデータを取得する処理において処理対象とされた音素である単独音素についてのコンテキストラベルを取得することで、前記ニューラルネットワークの学習処理に適した最適化フルコンテキストラベルデータを取得するフルコンテキストラベルベクトル処理ステップと、

前記最適化フルコンテキストラベルデータに基づいて、ニューラルネットワークのエンコード処理を実行することで、隠れ状態データを取得するエンコード処理ステップと、

前記隠れ状態データに基づいて、ニューラルネットワークのデコード処理を実行することで、前記最適化フルコンテキストラベルデータ

に対応する音響特徴量データを取得するデコード処理ステップと、

前記デコード処理ステップにより取得された音響特徴量から音声波形データを取得するボコーダ処理ステップと、

を備える音声合成処理方法。

[請求項7] 請求項６に記載の音声合成処理方法をコンピュータに実行させるためのプログラム。

[請求項8] 任意の言語を処理対象言語とし、エンコーダ・デコーダ方式のニューラルネットワークを用いて音声合成処理を実行する音声合成処理装置であって、

前記処理対象言語のテキストデータに対してテキスト解析処理を実行し、コンテキストラベルデータを取得するテキスト解析部と、

前記テキスト解析部により取得された前記コンテキストラベルデータから、コンテキストラベルデータを取得する処理において処理対象とされた音素である単独音素についてのコンテキストラベルを取得することで、前記ニューラルネットワークの学習処理に適した最適化フルコンテキストラベルデータを取得するフルコンテキストラベルベクトル処理部と、

前記最適化フルコンテキストラベルデータに基づいて、ニューラルネットワークのエンコード処理を実行することで、隠れ状態データを取得するエンコーダ部と、

音素単位のコンテキストラベルデータから音素継続長を推定する音素継続長推定部と、

前記音素継続長推定部により推定された音素継続長に基づいて、第１重み付け係数データを取得する強制アテンション部と、

前記エンコーダ部により取得された隠れ状態データに基づいて、第２重み付け係数データを取得するアテンション部と、

前記第１重み付け係数データと第２重み付け係数データとに対して内分処理を行うことで、合成重み付け係数データを取得する内分処理

部と、

前記合成重み付け係数データにより、前記エンコーダ部により取得された前記隠れ状態データに対して重み付け合成処理を実行することで、コンテキスト状態データを取得するコンテキスト算出部と、

前記コンテキスト状態データに基づいて、ニューラルネットワークのデコード処理を実行することで、前記最適化フルコンテキストラベルデータに対応する音響特徴量データを取得するデコーダ部と、

前記デコーダ部により取得された音響特徴量から音声波形データを取得するボコーダと、

を備える音声合成処理装置。

[図1]

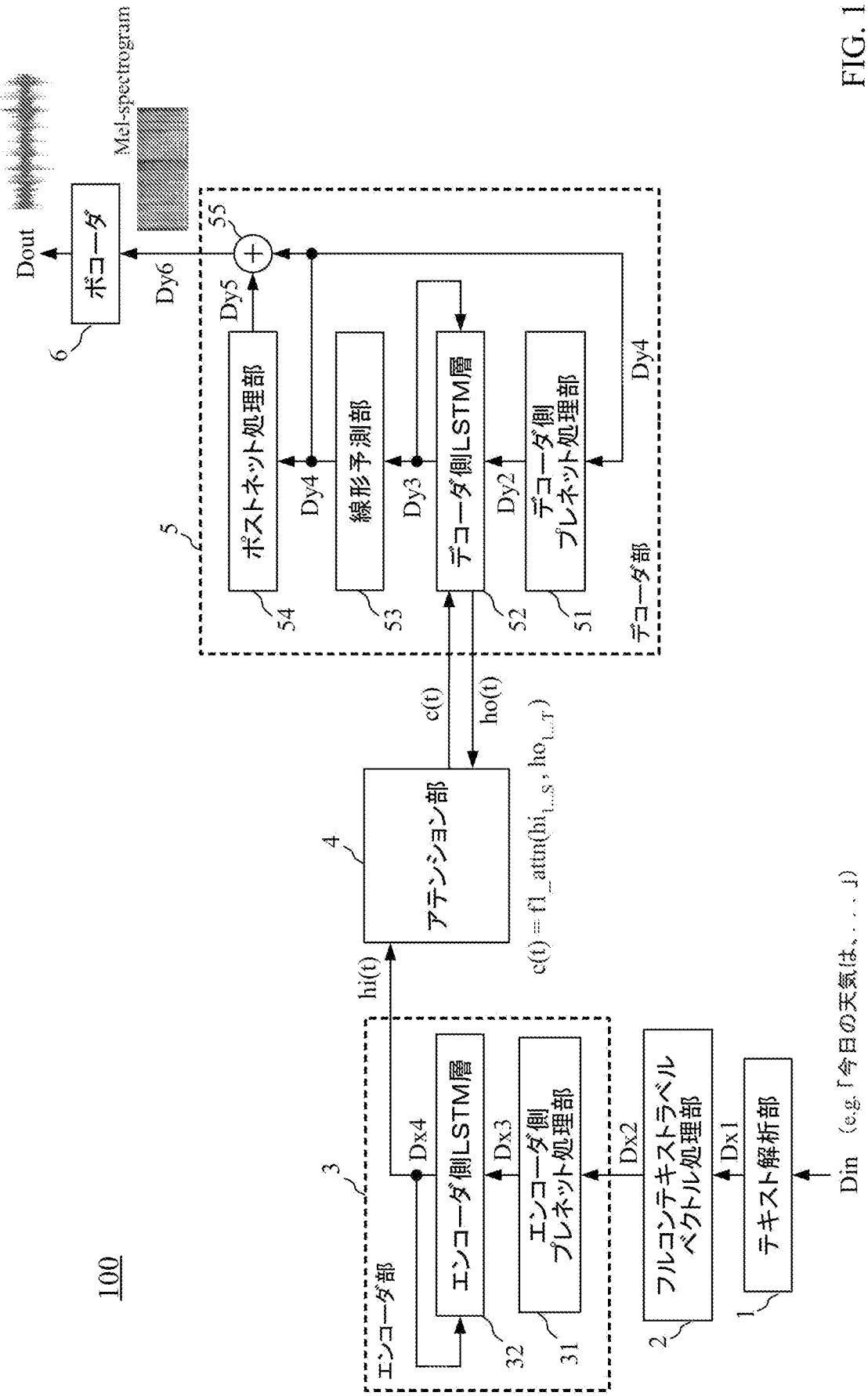

100
Din （e.g.「今日の天気は、・・・」）
テキスト解析部
1
フルコンテキストラベル
ベクトル処理部
2
エンコーダ側
プレネット処理部
31
エンコーダ側LSTM層
32
エンコーダ部
3
アテンション部
4
デコーダ側
プレネット処理部
51
デコーダ側LSTM層
52
線形予測部
53
ポストネット処理部
54
デコーダ部
5
ボコーダ
6
Dx1
Dx2
Dx3
Dx4
Dy2
Dy3
Dy4
Dy5
Dy6
55
hi(t)
c(t)
ho(t)
c(t) = fl_attn(hi1...s, ho1...r)
Mel-spectrogram
Dout

FIG. 1

[図2]

フルコンテキストラベルに含まれる情報（一例）

次元	音素数	総計	概要
43	5	215	[2つ先行/1つ先行/当該/1つ後続/2つ後続]の音素インデックス
1	1	1	前から数えた当該アクセント句内のアクセント位置
2	1	2	[前/後ろ]から数えた当該アクセント句内のモーラ位置
14	1	14	[1つ先行]のPOS(品詞)
15	1	15	[1つ先行]の名詞のカテゴリ
7	1	7	[1つ先行]のPOS(品詞)活用型
14	1	14	[当該]のPOS
15	1	15	[当該]の名詞のカテゴリ
7	1	7	[当該]のPOS活用型
14	1	14	[1つ後続]のPOS
15	1	15	[1つ後続]の名詞のカテゴリ
7	1	7	[1つ後続]のPOS活用型
1	1	1	[1つ先行]のアクセント句のモーラ数
1	1	1	[1つ先行]のアクセント句のアクセント位置
1	1	1	[1つ先行]のアクセント句の語尾上げ判定
1	1	1	[1つ先行]のアクセント句のポーズ有無
1	1	1	[当該]のアクセント句のモーラ数
1	1	1	[当該]のアクセント句のアクセント位置
1	1	1	[当該]のアクセント句の語尾上げ判定
2	1	2	[前/後ろ]から数えた当該呼気段落内のアクセント句位置（アクセント句単位）
2	1	2	[前/後ろ]から数えた当該呼気段落内のアクセント句位置（モーラ単位）
1	1	1	[1つ後続]のアクセント句のモーラ数
1	1	1	[1つ後続]のアクセント句のアクセント位置
1	1	1	[1つ後続]のアクセント句の語尾上げ判定
1	1	1	[1つ後続]のアクセント句のポーズ有無
2	1	2	先行呼気段落の[アクセント句数/モーラ数]
1	1	1	文頭判定
2	1	2	当該呼気段落の[アクセント句数/モーラ数]
2	1	2	[前/後ろ]から数えた当該呼気段落の文中位置（呼気段落単位）
2	1	2	[前/後ろ]から数えた当該呼気段落の文中位置（アクセント句単位）
2	1	2	[前/後ろ]から数えた当該呼気段落の文中位置（モーラ単位）
2	1	2	後続呼気段落の[アクセント句数/モーラ数]
1	1	1	文末判定
1	1	1	文章内の呼気段落数
1	1	1	文章内のアクセント句数
1	1	1	文章内のモーラ数
22	5	110	[2つ先行/1つ先行/当該/1つ後続/2つ後続]の音素カテゴリ
1	5	5	[2つ先行/1つ先行/当該/1つ後続/2つ後続]音素の調音位置によるカテゴライズ
1	5	5	[2つ先行/1つ先行/当該/1つ後続/2つ後続]音素の調音方法によるカテゴライズ

合計： 478 次元

FIG. 2

[図3]

最適化フルコンテキストラベルに含まれる情報（一例）

次元	音素数	総計	概要
43	1	43	[当該]の音素インデックス
1	1	1	前から数えた当該アクセント句内のアクセント位置
2	1	2	[前/後ろ]から数えた当該アクセント句内のモーラ位置
14	1	14	[当該]のPOS
15	1	15	[当該]の名詞のカテゴリ
7	1	7	[当該]のPOS活用型
1	1	1	[当該]のアクセント句のモーラ数
1	1	1	[当該]のアクセント句のアクセント位置
1	1	1	[当該]のアクセント句の語尾上げ判定
2	1	2	[前/後ろ]から数えた当該呼気段落内のアクセント句位置（アクセント句単位）
2	1	2	[前/後ろ]から数えた当該呼気段落内のアクセント句位置（モーラ単位）
2	1	2	先行呼気段落の[アクセント句数/モーラ数]
1	1	1	文頭判定
2	1	2	当該呼気段落の[アクセント句数/モーラ数]
2	1	2	[前/後ろ]から数えた当該呼気段落の文中位置（呼気段落単位）
2	1	2	[前/後ろ]から数えた当該呼気段落の文中位置（アクセント句単位）
2	1	2	[前/後ろ]から数えた当該呼気段落の文中位置（モーラ単位）
2	1	2	後続呼気段落の[アクセント句数/モーラ数]
1	1	1	文末判定
1	1	1	文章内の呼気段落数
1	1	1	文章内のアクセント句数
1	1	1	文章内のモーラ数
22	1	22	[当該]の音素カテゴリ
1	1	1	[当該]音素の調音位置によるカテゴライズ
1	1	1	[当該]音素の調音方法によるカテゴライズ

合計： 130 次元

FIG. 3

[図4]

学習処理時

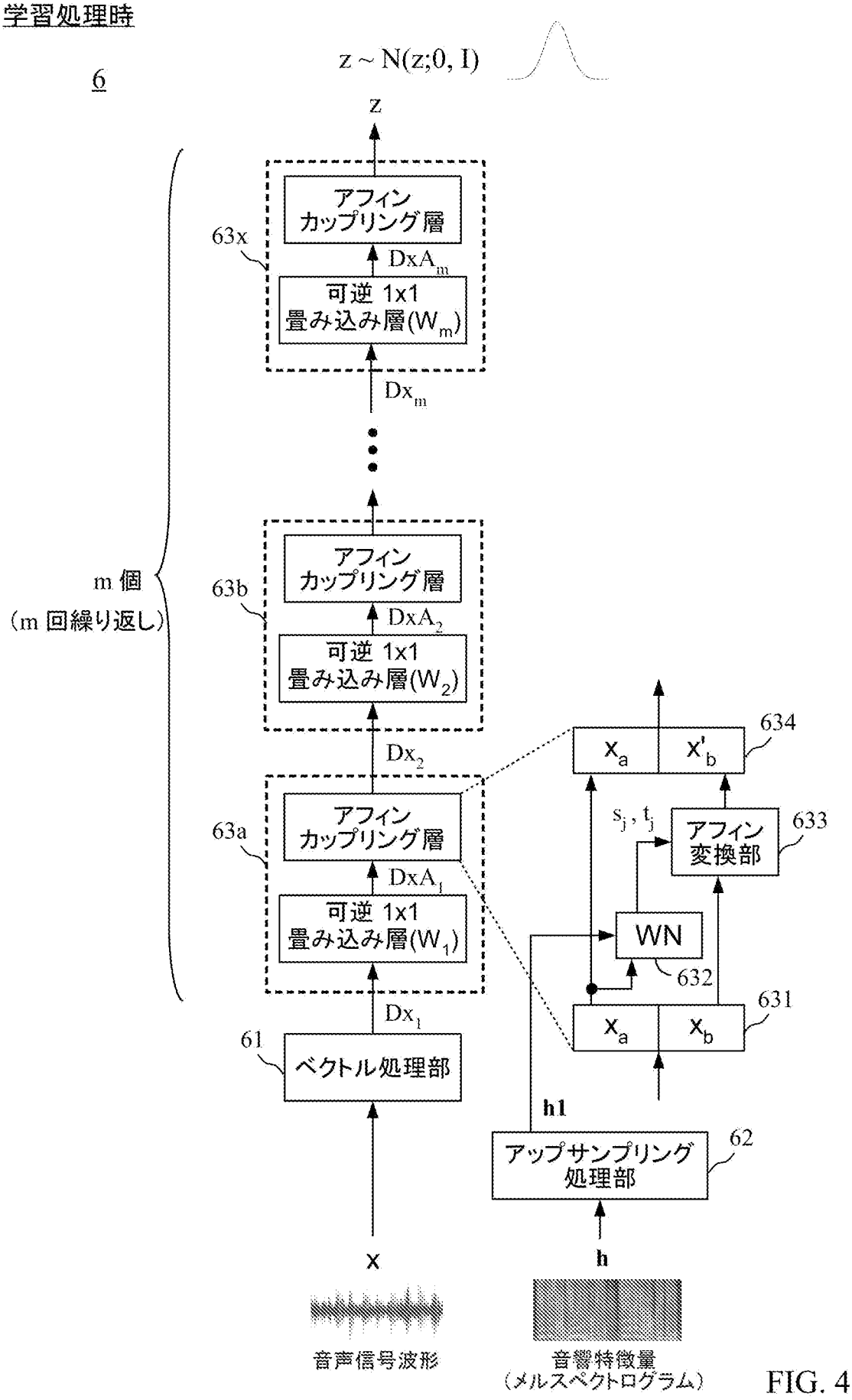


FIG. 4

[図5]

予測処理時

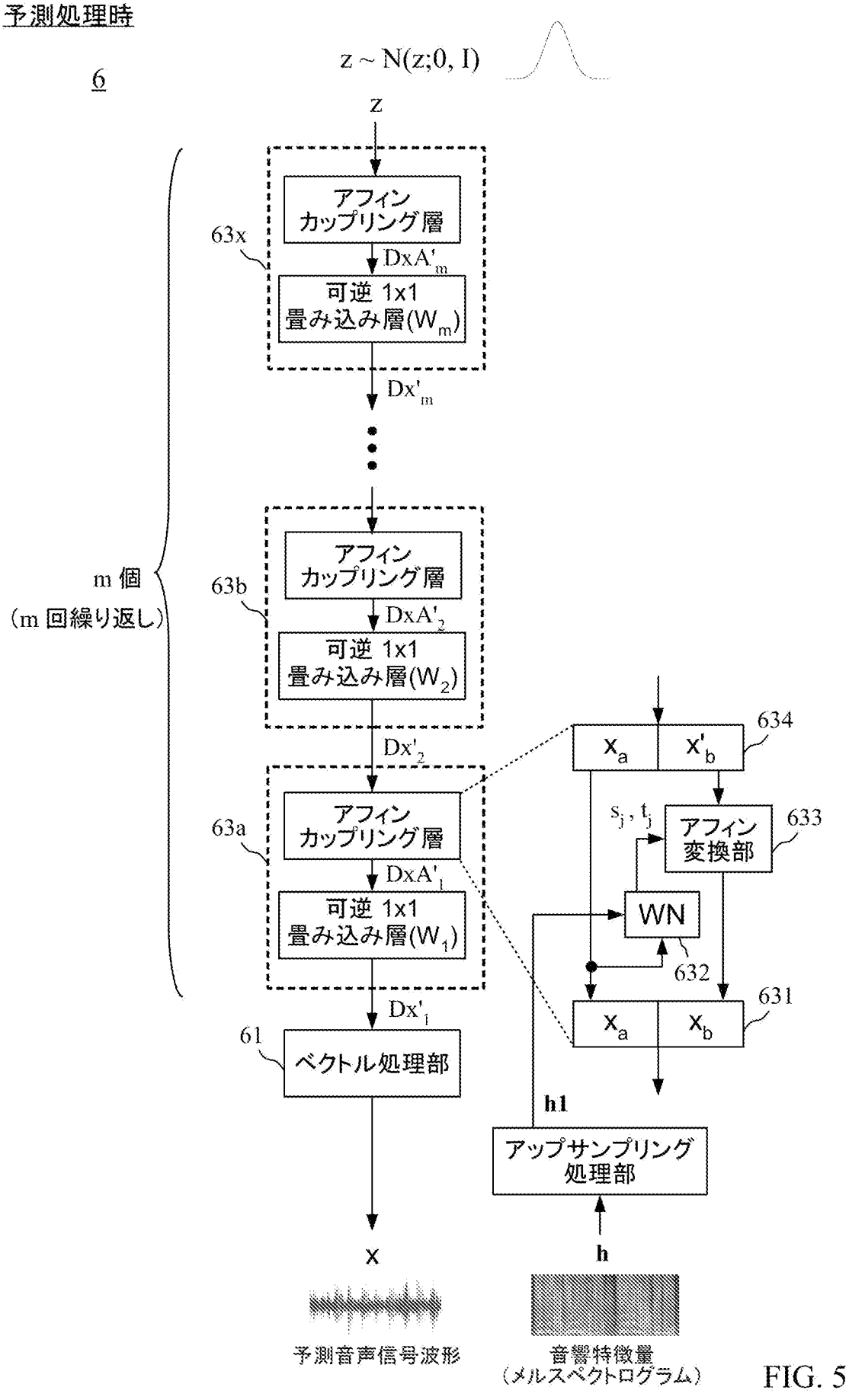


FIG. 5

[図6]

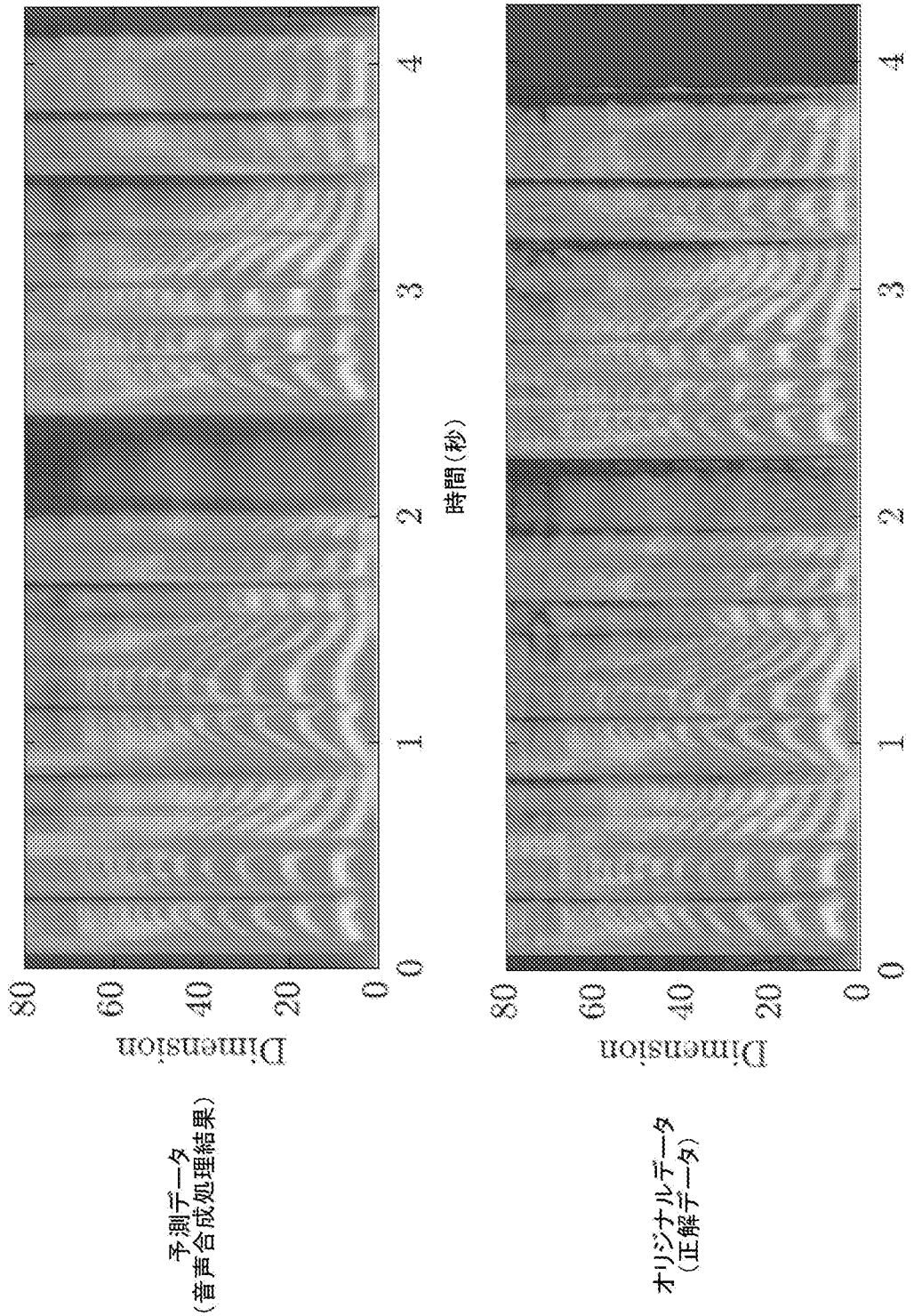

予測データ
（音声合成処理結果）
Dimension
0
20
40
60
80
0
1
2
3
4
時間（秒）
オリジナルデータ
（正解データ）
Dimension
0
20
40
60
80
0
1
2
3
4
時間（秒）

FIG. 6

[図7]

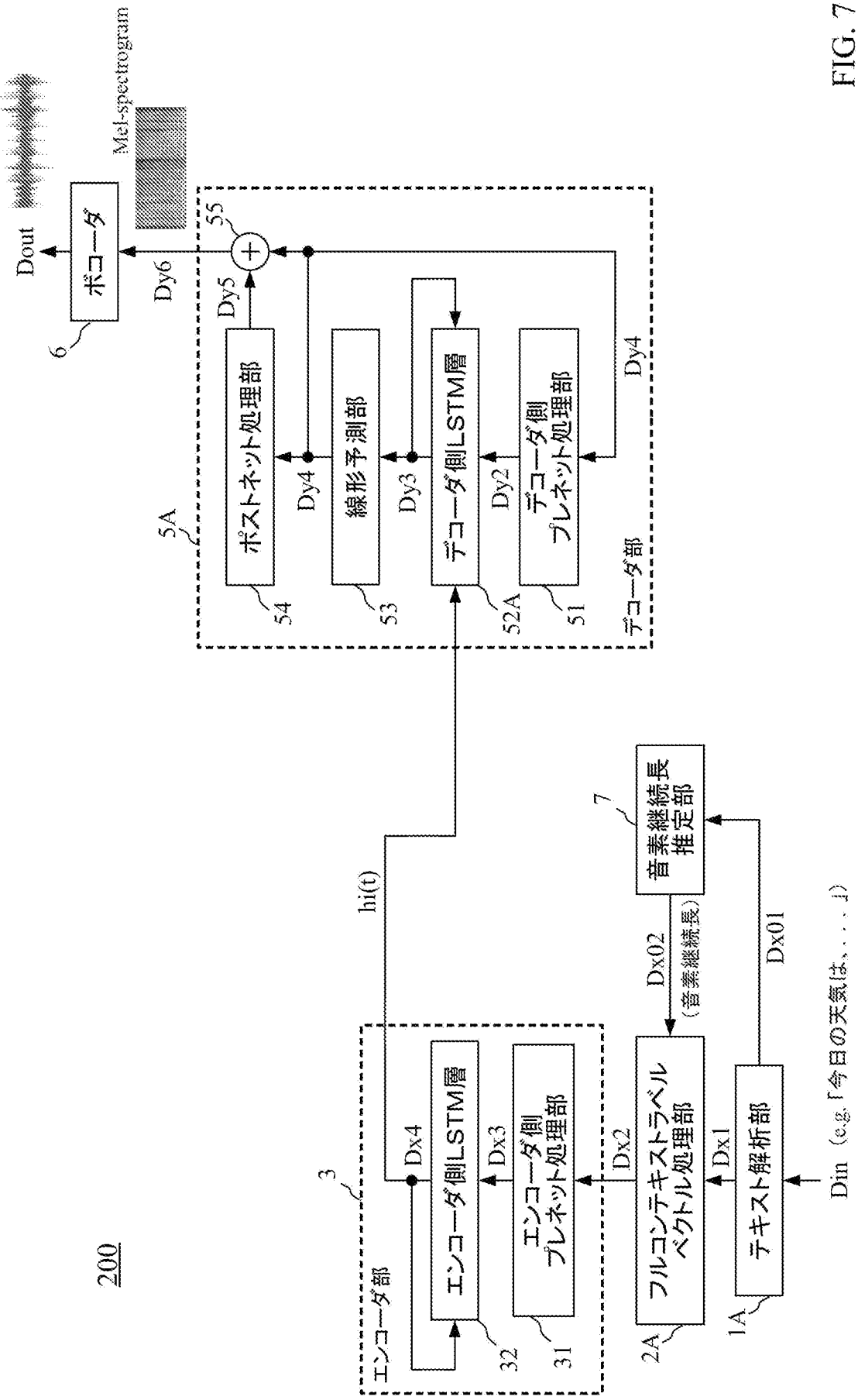

200
エンコーダ部
3
エンコーダ側LSTM層
32
エンコーダ側
プレネット処理部
31
Dx4
Dx3
Dx2
hi(t)
フルコンテキストラベル
ベクトル処理部
2A
テキスト解析部
1A
Din（e.g.「今日の天気は、...」）
Dx1
Dx01
音素継続長
推定部
7
Dx02
（音素継続長）
デコーダ部
5A
ポストネット処理部
54
線形予測部
53
デコーダ側LSTM層
52A
デコーダ側
プレネット処理部
51
Dy4
Dy3
Dy2
Dy4
Dy5
55
Dy6
ボコーダ
6
Dout
Mel-spectrogram

FIG. 7

[図8]

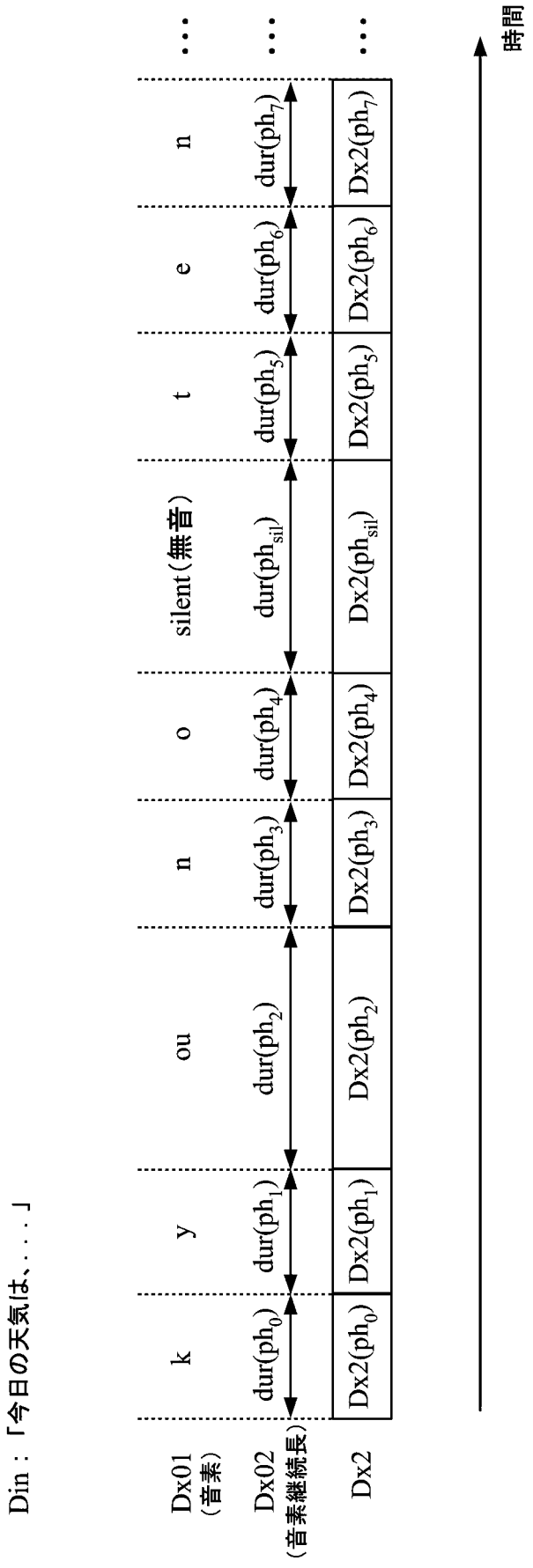

Din：「今日の天気は、・・・」
Dx01 (音素)
Dx02 (音素継続長)
Dx2
k
y
ou
n
o
silent（無音）
t
e
n
・・・
・・・
・・・
dur(ph0)
dur(ph1)
dur(ph2)
dur(ph3)
dur(ph4)
dur(phsil)
dur(ph5)
dur(ph6)
dur(ph7)
Dx2(ph0)
Dx2(ph1)
Dx2(ph2)
Dx2(ph3)
Dx2(ph4)
Dx2(phsil)
Dx2(ph5)
Dx2(ph6)
Dx2(ph7)
時間

FIG. 8

[図9]

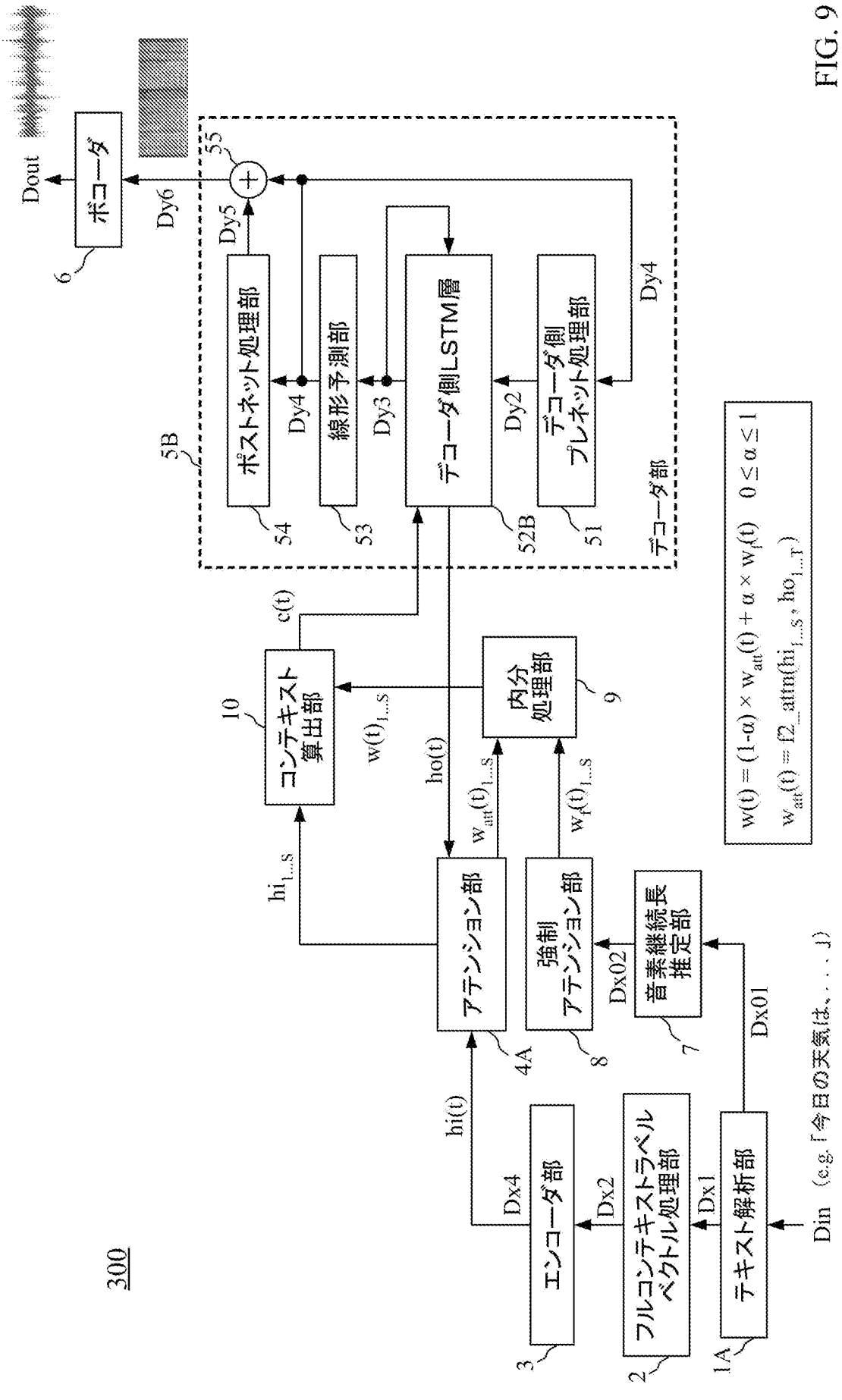

300
Din (e.g.「今日の天気は、・・・」)
テキスト解析部
1A
フルコンテキストラベル
ベクトル処理部
2
エンコーダ部
3
アテンション部
4A
強制
アテンション部
8
音素継続長
推定部
7
内分
処理部
9
コンテキスト
算出部
10
デコーダ部
5B
ポストネット処理部
54
線形予測部
53
デコーダ側LSTM層
52B
デコーダ側
プレネット処理部
51
ボコーダ
6
Dout
Dy6
Dy5
Dy4
Dy3
Dy2
Dy4
55
Dx01
Dx02
Dx1
Dx2
Dx4
hi(t)
hi1...S
c(t)
w(t)1...S
ho(t)
watt(t)1...S
wf(t)1...S
w(t) = (1-α) × watt(t) + α × wf(t) 0 ≤ α ≤ 1
watt(t) = f2_attn(hi1...S, ho1...T)

FIG. 9

[図10]

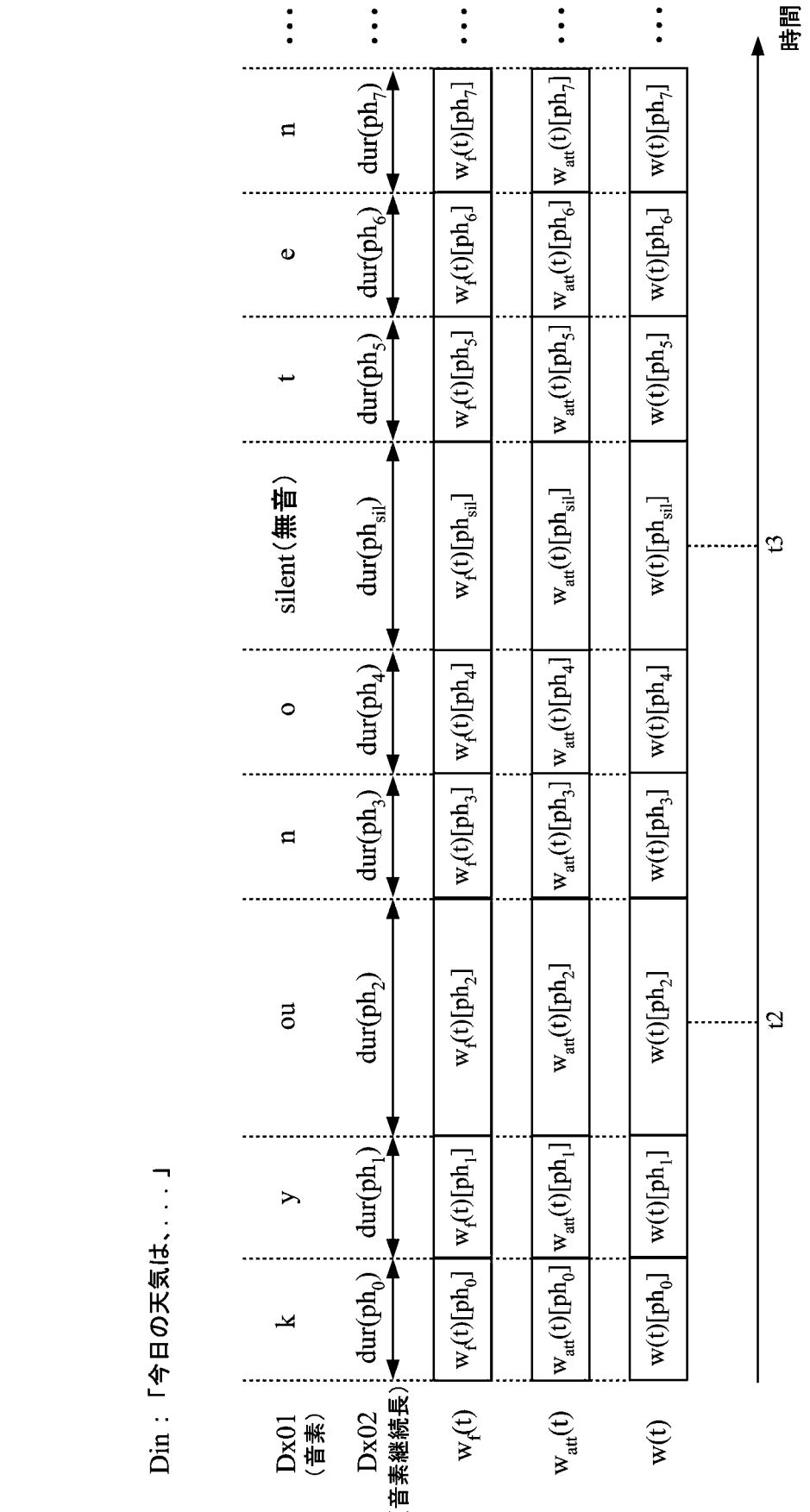

Din：「今日の天気は、・・・」
Dx01（音素）
Dx02（音素継続長）
k
y
ou
n
o
silent（無音）
t
e
n
・・・
dur(ph0)
dur(ph1)
dur(ph2)
dur(ph3)
dur(ph4)
dur(phsil)
dur(ph5)
dur(ph6)
dur(ph7)
wf(t)[ph0]
wf(t)[ph1]
wf(t)[ph2]
wf(t)[ph3]
wf(t)[ph4]
wf(t)[phsil]
wf(t)[ph5]
wf(t)[ph6]
wf(t)[ph7]
watt(t)[ph0]
watt(t)[ph1]
watt(t)[ph2]
watt(t)[ph3]
watt(t)[ph4]
watt(t)[phsil]
watt(t)[ph5]
watt(t)[ph6]
watt(t)[ph7]
wf(t)
watt(t)
w(t)
w(t)[ph0]
w(t)[ph1]
w(t)[ph2]
w(t)[ph3]
w(t)[ph4]
w(t)[phsil]
w(t)[ph5]
w(t)[ph6]
w(t)[ph7]
t2
t3
時間

FIG. 10

[図11]

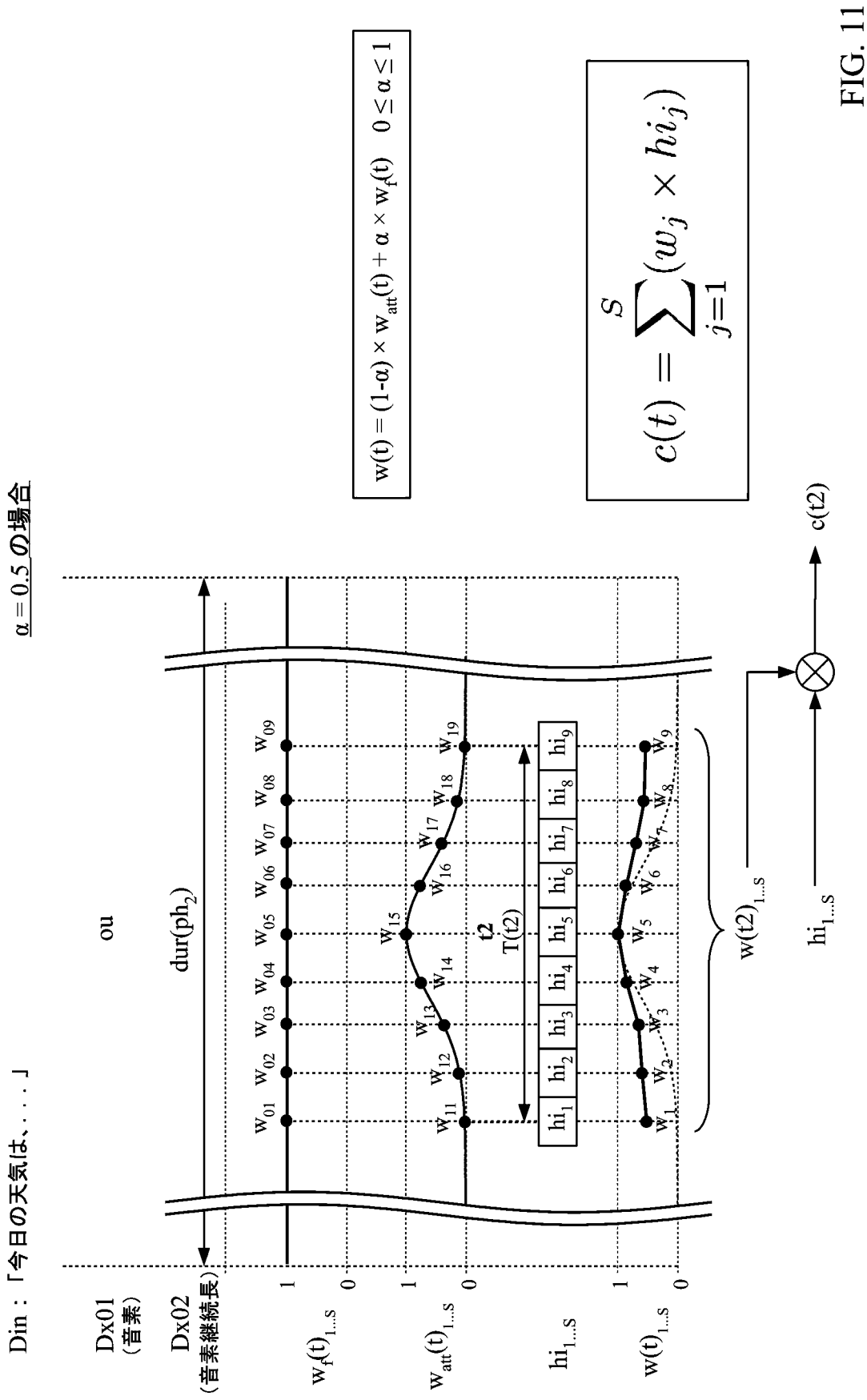

α = 0.5 の場合
Din：「今日の天気は、．．．」
Dx01
(音素)
Dx02
(音素継続長)
ou
dur(ph2)
w(t) = (1-α) × w_att(t) + α × w_f(t) 0 ≤ α ≤ 1
c(t) = Σ_{j=1}^{S} (w_j × hi_j)
t2
T(t2)
hi1 hi2 hi3 hi4 hi5 hi6 hi7 hi8 hi9
w01 w02 w03 w04 w05 w06 w07 w08 w09
w11 w12 w13 w14 w15 w16 w17 w18 w19
w1 w2 w3 w4 w5 w6 w7 w8 w9
w(t2)1...S
hi1...S
c(t2)
wf(t)1...S
watt(t)1...S
hi1...S
w(t)1...S
1
0

FIG. 11

[図12]

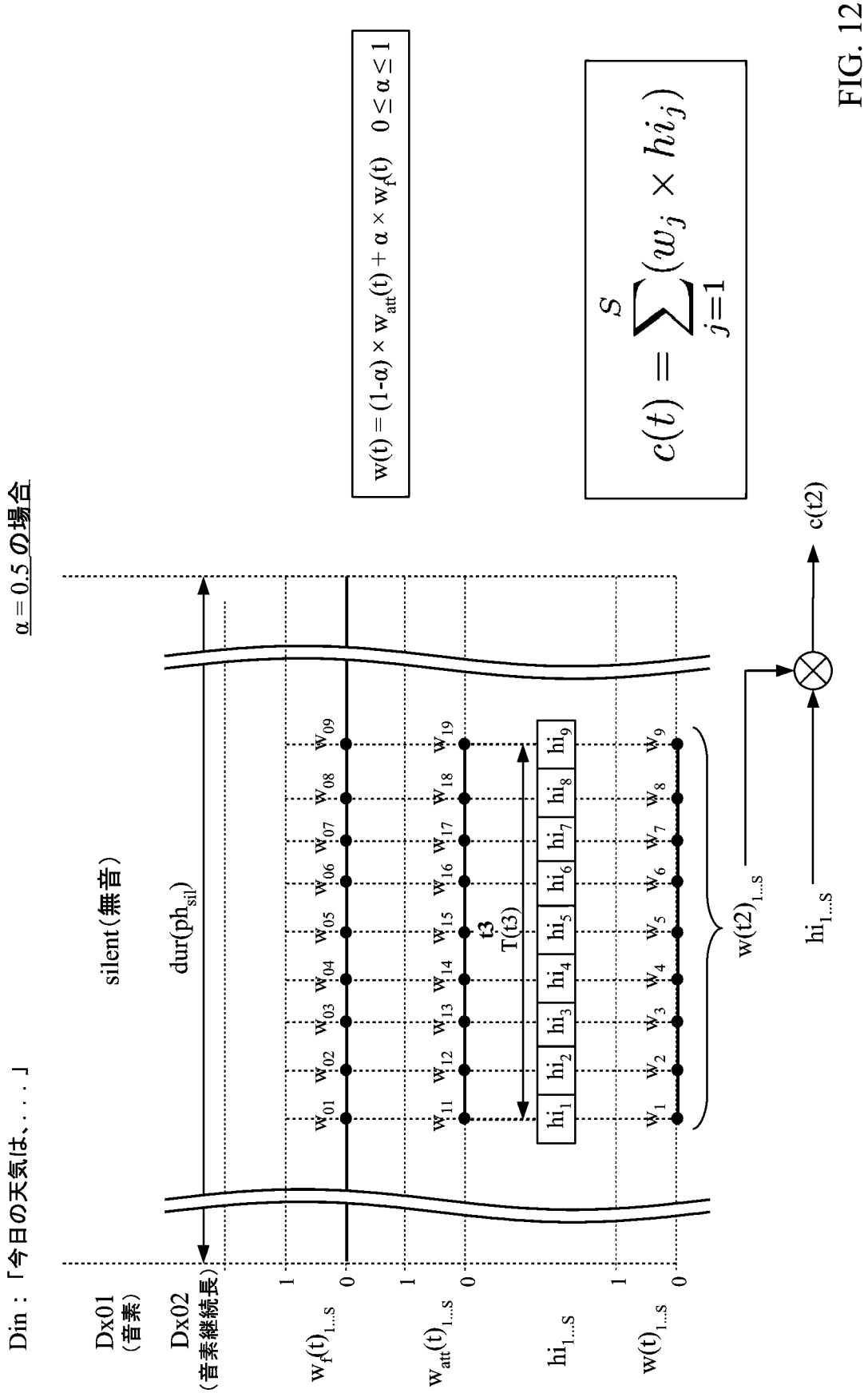

α = 0.5 の場合
Din：「今日の天気は、．．．」
Dx01 (音素)
Dx02 (音素継続長)
silent(無音)
dur(ph_sil)
w(t) = (1-α) × w_att(t) + α × w_f(t)　0 ≤ α ≤ 1
c(t) = Σ_{j=1}^{S} (w_j × hi_j)
t3
T(t3)
w(t2)_1..S
hi_1..S
c(t2)
w_f(t)_1..S
w_att(t)_1..S
hi_1..S
w(t)_1..S

FIG. 12

[図13]

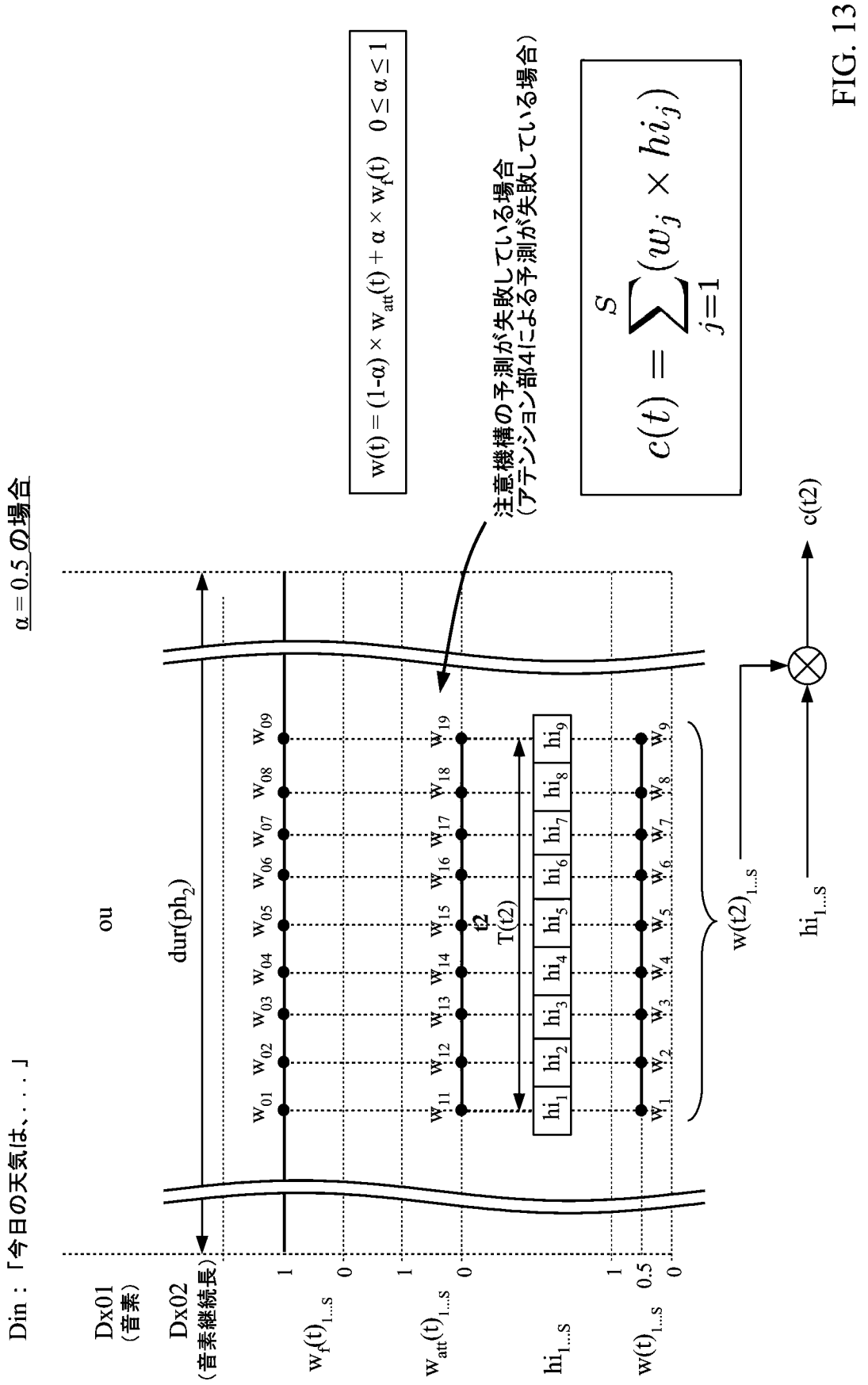
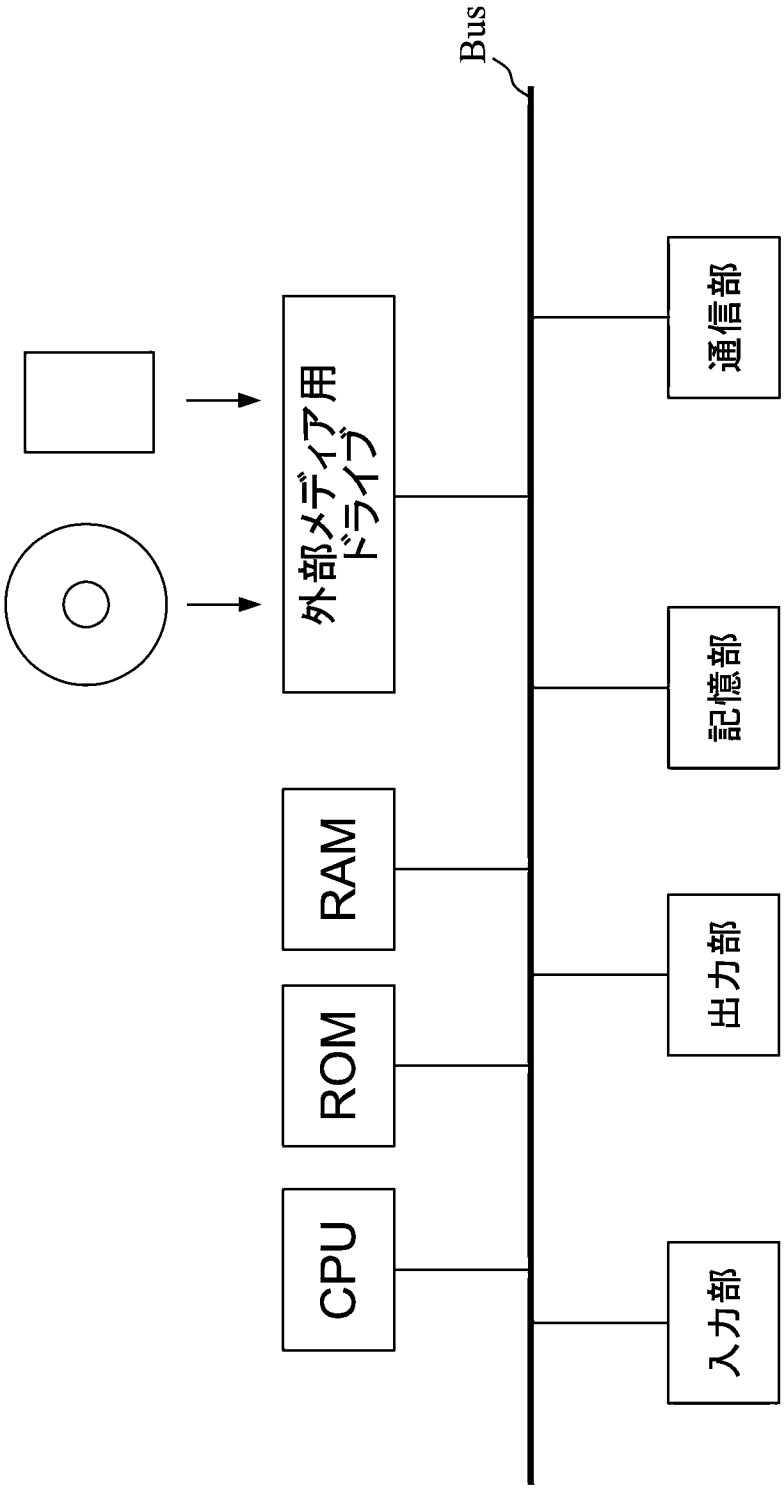

α = 0.5 の場合
Din：「今日の天気は、．．．」
Dx01
（音素）
Dx02
（音素継続長）
ou
dur(ph2)
w(t) = (1-α) × w_att(t) + α × w_f(t)　0 ≤ α ≤ 1
注意機構の予測が失敗している場合
（アテンション部4による予測が失敗している場合）
c(t) = Σ_{j=1}^{S} (w_j × hi_j)
t2
T(t2)
w(t2)1..S
hi1..S
c(t2)
w_f(t)1..S
w_att(t)1..S
hi1..S
w(t)1..S

FIG. 13

[図14]

FIG. 14



INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2020/025682

A. CLASSIFICATION OF SUBJECT MATTER
G10L 13/08(2013.01)i; G10L 13/10(2013.01)i; G10L 25/30(2013.01)i
FI: G10L13/08 110Z; G10L25/30; G10L13/08 150Z; G10L13/10 111F

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)
G10L13/08; G10L13/10; G10L25/30

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	藤本崇人、外 4 名，日本語 End-to-End 音声合成における入力言語特徴量の影響，日本音響学会 2019 年春季研究発表会講演論文集 CD-ROM, 19 February 2019, pp. 1061-1062, pp. 1061-1062, non-official translation (FUJIMOTO, Takahito et al., "Effect of input language feature values on Japanese End-to-End speech synthesis", Lecture proceedings of 2019 spring research conference of the Acoustical Society of Japan CD-ROM)	1-8
A	安田裕介、外 3 名，日本語エンドツーエンド音声合成へむけて－日本語 Tacotron の初期的検討，日本音響学会 2018 年秋季研究発表会講演論文集, 29 August 2018, pp. 1167-1168, pp. 1167-1168, non-official translation (YASUDA, Yusuke et al., "Towards end-to-end Japanese speech synthesis -An initial consideration of Japanese Tacotron", Lecture Proceedings of 2018 Autumn Research Conference of the Acoustical Society of Japan)	1-8

☒ 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
26 August 2020 (26.08.2020)	08 September 2020 (08.09.2020)

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/JP2020/025682

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	栗原清、外３名，読み仮名と韻律記号を入力とする日本語 end-to-end 音声合成の音質評価，電子情報通信学会技術研究報告，03 December 2018, vol. 118, no. 354, pp. 89-94, pp. 89-94, (KURIHARA, Kiyoshi et al., "Evaluation of Japanese end-to-end speech synthesis method inputting kana and prosodic symbols", IEICE technical report)	1-8
A	SHEN, Jonathan et al., "NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS", Proc. ICASSP 2018, April 2018, pp. 4779-4783, pp. 4779-4783	1-8
A	WO 2018/183650 A2 (GOOGLE LLC) 04.10.2018 (2018-10-04) entire text, all drawings	1-8
A	PRENGER, Ryan et al., "WAVEGLOW: A FLOW-BASED GENERATIVE NETWORK FOR SPEECH SYNTHESIS", Proc. ICASSP 2019, May 2019, pp. 3617-3621, pp. 3617-3621	4

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/JP2020/025682

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
WO 2018/183650 A2	04 Oct. 2018	US 2019/0311708 A1 EP 3533594 A2 AU 2018244917 A CA 3058433 A1 CN 110476206 A KR 10-2019-0130634 A	

国際調査報告

国際出願番号
PCT/JP2020/025682

A. 発明の属する分野の分類（国際特許分類（ＩＰＣ））

G10L 13/08(2013.01)i; G10L 13/10(2013.01)i; G10L 25/30(2013.01)i
FI: G10L13/08 110Z; G10L25/30; G10L13/08 150Z; G10L13/10 111F

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））

G10L13/08; G10L13/10; G10L25/30

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	藤本 崇人 、外４名，日本語End-to-End 音声合成における入力言語特徴量の影響，日本音響学会 ２０１９年 春季研究発表会講演論文集ＣＤ－ＲＯＭ，2019.02.19，pp. 1061-1062 pp. 1061-1062	1-8
A	安田 裕介、外３名，日本語エンドツーエンド音声合成へむけて－日本語Tacotronの初期的検討，日本音響学会 ２０１８年 秋季研究発表会講演論文集，2018.08.29，pp. 1167-1168 pp. 1167-1168	1-8
A	栗原 清、外３名，読み仮名と韻律記号を入力とする日本語ｅｎｄ－ｔｏ－ｅｎｄ音声合成の音質評価，電子情報通信学会技術研究報告，2018.12.03，Vol. 118，No. 354，pp. 89-94 pp. 89-94	1-8
A	SHEN, Jonathan, et al., NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS, Proc. ICASSP 2018, 2018.04, pp. 4779-4783 pp. 4779-4783	1-8
A	WO 2018/183650 A2 (GOOGLE LLC) 04.10.2018 (2018-10-04) 全文，全図	1-8

☑ 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
26.08.2020	08.09.2020

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）

上田　雄 5Z 5095

電話番号 03-3581-1101 内線 3591

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	PRENGER, Ryan, et al., WAVEGLOW: A FLOW-BASED GENERATIVE NETWORK FOR SPEECH SYNTHESIS, Proc. ICASSP 2019, 2019.05, pp. 3617-3621 pp. 3617-3621	4

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/JP2020/025682

引用文献	公表日	パテントファミリー文献	公表日
WO 2018/183650 A2	04.10.2018	US 2019/0311708 A1	
		EP 3583594 A2	
		AU 2018244917 A	
		CA 3058433 A1	
		CN 110476206 A	
		KR 10-2019-0130634 A	