

(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
2018年9月7日(07.09.2018)



(10) 国際公開番号

WO 2018/159403 A1

(51) 国際特許分類:

G10L 13/06 (2013.01) G10L 19/06 (2013.01)
G10L 13/10 (2013.01)

(21) 国際出願番号: PCT/JP2018/006166

(22) 国際出願日: 2018年2月21日(21.02.2018)

(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 2017-037220 2017年2月28日(28.02.2017) 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: 橘 健太郎 (TACHIBANA, Kentaro); 〒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

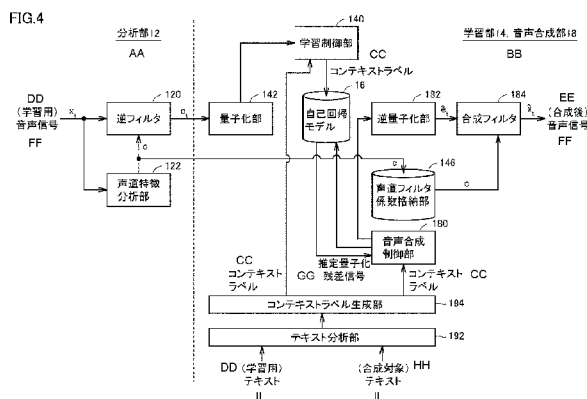
(JP). 戸田 智基(TODA, Tomoki); 〒4801103 愛知県長久手市岩作三ヶ峯3-8-1 Aichi (JP).

(74) 代理人: 特許業務法人深見特許事務所(FUKAMI PATENT OFFICE, P.C.); 〒5300005 大阪府大阪市北区中之島三丁目2番4号 中之島フェスティバルタワー・ウエスト Osaka (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, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(54) Title: LEARNING DEVICE, SPEECH SYNTHESIS SYSTEM, AND SPEECH SYNTHESIS METHOD

(54) 発明の名称: 学習装置、音声合成システムおよび音声合成方法



16 Autoregressive model
120 Inverse filter
122 Vocal tract feature analysis unit
140 Learning control unit
142 Quantization unit
146 Vocal tract filter coefficient storage unit
180 Speech synthesis control unit
182 Inverse quantization unit
184 Synthesis filter
192 Text analysis unit
194 Context label generation unit
AA Analysis unit 12
BB Learning unit 14, speech synthesis unit 18
CC Context label
DD (For learning)
EE (After synthesis)
FF Speech signal
GG Estimation quantized residual signal
HH (Synthesis object)
II Text

(57) Abstract: This learning device includes: an autoregressive model that estimates a present signal from a past signal sequence and a present context label; a vocal tract feature analysis unit that analyzes an input speech signal to determine a vocal tract filter coefficient indicating a vocal tract feature; a residual signal generation unit that outputs a residual signal between the input speech signal and a speech signal predicted on the basis of the vocal tract filter coefficient; a quantization unit that quantizes the residual signal outputted from the residual signal generation unit to generate a quantized residual signal; and a learning control unit that, regarding the input speech signal corresponding to a known input text, provides, as a condition, a context label of the known input text to the autoregressive model, and associates the present context label and the past quantized residual signal sequence regarding the input speech signal with the present quantized residual signal, thereby performing learning of the autoregressive model.



(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:

— 国際調査報告 (条約第21条(3))

(57) 要約: 学習装置は、過去の信号系列および現在のコンテキストラベルから現在の信号を推定する自己回帰モデルと、入力音声信号を分析して声道特徴を示す声道フィルタ係数を決定する声道特徴分析部と、声道フィルタ係数に基づいて予測される音声信号と入力音声信号との間の残差信号を出力する残差信号生成部と、残差信号生成部から出力される残差信号を量子化して量子化残差信号を生成する量子化部と、既知の入力テキストに対応する入力音声信号について、既知の入力テキストのコンテキストラベルを自己回帰モデルに対して条件として与えるとともに、入力音声信号についての量子化残差信号の過去の系列および現在のコンテキストラベルと、量子化残差信号の現在の信号とを対応付けることで自己回帰モデルの学習を実施する学習制御部とを含む。

明 細 書

発明の名称：学習装置、音声合成システムおよび音声合成方法 技術分野

[0001] 本発明は、入力テキストに応じた音声を作成して出力する音声合成技術に関する。

背景技術

[0002] 従来から、音声合成分野においては、統計モデルに基づいて音声信号を生成する枠組みである、統計的パラメトリック音声合成 (statistical parametric speech synthesis: 以下、「SPSS」とも略称する。) が盛んに研究されてきた。SPSSにおいては、入力されたテキストと当該テキストに対応する音声信号との対応関係を統計的にモデル化する。このような関係を直接モデル化することは容易ではないので、それぞれを特徴量系列として表現することで、統計モデルを構築する。具体的には、入力テキストは言語特徴量を表すコンテキストラベル系列で表現され、音声信号は音響特徴量系列で表現される。

[0003] このような音響特徴量系列から音声信号を推定する方法に代えて、コンテキストラベル系列から音声信号を直接的に推定する、“WaveNet”と称される手法が提案されている（例えば、非特許文献1参照）。このWaveNetは、既存の最新手法を凌ぐ性能を示すことが報告されている。

先行技術文献

非特許文献

- [0004] 非特許文献1: A. van den Oord et al., “Wavenet: A Generative Model for Raw Audio”, arXiv preprint arXiv:1609.03499, 2016
非特許文献2: A. van den Oord et al., “Pixel Recurrent Neural Networks”, arXiv preprint arXiv:1601.06759v3, 2016.08.19

発明の概要

発明が解決しようとする課題

[0005] 非特許文献 1 に開示される WaveNet において推定出力される信号は、 $\mu - l$ a w 方式により量子化された音声信号である。 $\mu - l$ a w 方式により量子化された音声信号を推定する場合、その信号を復元した際の推定誤差は帯域全体に広がってしまうため、特に高周波帯域のノイズが知覚されやすくなるという課題がある。

[0006] 本発明は、上述したような課題を考慮するものであり、入力テキストに基づくコンテキストラベルから音声信号を直接的に推定する際の音声品質を高めることを目的としている。

課題を解決するための手段

[0007] 本発明のある局面に従えば、音声合成システムに向けられた学習装置が提供される。学習装置は、過去の信号系列および現在のコンテキストラベルから現在の信号を推定する自己回帰モデルを含む。自己回帰モデルは、統計的データモデリングが可能なネットワーク構造を有している。学習装置は、入力音声信号を分析して声道特徴を示す声道フィルタ係数を決定する声道特徴分析部と、声道フィルタ係数に基づいて予測される音声信号と入力音声信号との間の残差信号を出力する残差信号生成部と、残差信号生成部から出力される残差信号を量子化して量子化残差信号を生成する量子化部と、既知の入力テキストに対応する入力音声信号について、既知の入力テキストのコンテキストラベルを自己回帰モデルに対して条件として与えるとともに、入力音声信号についての量子化残差信号の過去の系列および現在のコンテキストラベルと、量子化残差信号の現在の信号とを、対応付けることで自己回帰モデルの学習を実施する学習制御部とを含む。

[0008] 本発明の別の局面に従えば、入力テキストに応じた音声を合成して出力する音声合成システムが提供される。音声合成システムは、未知の入力テキストが入力されると、未知の入力テキストのコンテキストラベルを自己回帰モデルに対して条件として与えるとともに、過去の推定量子化残差信号を請求項 1 に記載の学習装置によって構築された自己回帰モデルを用いて、現在の量子化残差信号を出力する音声合成制御部を含む。

[0009] 好ましくは、音声合成システムは、量子化部から出力される過去の量子化残差信号および現在のコンテキストラベルから推定された推定量子化残差信号を逆量子化して推定残差信号を生成する逆量子化部と、逆量子化部から出力される推定残差信号を声道フィルタ係数に基づいてフィルタリングした結果を音声信号として出力する合成フィルタと、入力音声信号についての声道フィルタ係数を格納する格納部とをさらに含む。

[0010] 好ましくは、声道フィルタ係数は、聴覚重み付け係数により調整可能になっている。

好ましくは、音声合成システムは、入力テキストを分析して文脈情報を生成するテキスト分析部と、テキスト分析部からの文脈情報に基づいて、入力テキストのコンテキストラベルを生成するコンテキストラベル生成部とをさらに含む。

[0011] 本発明のさらに別の局面に従えば、入力テキストに応じた音声を作成して出力する音声合成方法が提供される。音声合成方法は、既知の入力テキストに対応する入力音声信号を分析して声道特徴を示す声道フィルタ係数を決定するステップと、声道フィルタ係数に基づいて予測される音声信号と入力音声信号との間の残差信号を生成するステップと、残差信号を量子化して量子化残差信号を生成するステップと、既知の入力テキストのコンテキストラベルを自己回帰モデルに対して条件として与えるとともに、過去の量子化残差信号および現在のコンテキストラベルから現在の量子化残差信号を推定する自己回帰モデルを学習するステップとを含む。自己回帰モデルは、過去の信号系列および現在のコンテキストラベルから現在値を推定するためのパラメータを格納しており、統計的データモデリングが可能なネットワーク構造を有している。

発明の効果

[0012] 本発明によれば、入力テキストに基づくコンテキストラベルから音声信号を直接的に推定する際の音声品質を高めることができる。

図面の簡単な説明

[0013] [図1]本実施の形態に従う音声合成システムを用いた多言語翻訳システムの概要を示す模式図である。

[図2]本実施の形態に従うサービス提供装置のハードウェア構成例を示す模式図である。

[図3]本実施の形態に従う音声合成システムにて採用する予測量子化の概要を説明するためのブロック図である。

[図4]本実施の形態に従う音声合成システムにおける要部の処理を説明するためのブロック図である。

[図5]本実施の形態に従う音声合成システムに用いられる自己回帰モデルの概要を説明するためのブロック図である。

[図6]本実施の形態に従う音声合成システムにおける処理手順の一例を示すフローチャートである。

[図7]本実施の形態に従う音声合成システムにおける処理手順の一例を示すフローチャートである。

[図8]本実施の形態に従う音声合成システムについてのノイズシェーピング効果を評価した結果の一例を示す図である。

[図9]本実施の形態に従う音声合成システムについての対比較実験の評価結果例を示す図である。

発明を実施するための形態

[0014] 本発明の実施の形態について、図面を参照しながら詳細に説明する。なお、図中の同一または相当部分については、同一符号を付してその説明は繰返さない。

[0015] [A. 応用例]

まず、本実施の形態に従う音声合成システムの一つの応用例について説明する。より具体的には、本実施の形態に従う音声合成システムを用いた多言語翻訳システムについて説明する。

[0016] 図1は、本実施の形態に従う音声合成システムを用いた多言語翻訳システム1の概要を示す模式図である。図1を参照して、多言語翻訳システム1は

、サービス提供装置 10 を含む。サービス提供装置 10 は、ネットワーク 2 を介して接続される携帯端末 30 からの入力音声（第 1 言語で発せられたなんらかのことば）に対して音声認識、多言語翻訳などを行なって、第 2 言語での対応することばを合成して、その合成結果を出力音声として携帯端末 30 へ出力する。

[0017] 例えば、ユーザ 4 は、携帯端末 30 に対して、「Where is the station ?」という英語のことばを発すると、携帯端末 30 は、その発せられたことばからマイクロフォンなどにより入力音声を生成し、生成した入力音声をサービス提供装置 10 へ送信する。サービス提供装置 10 は、「Where is the station ?」に対応する、日本語の「駅はどこですか？」ということばを示す出力音声を合成する。携帯端末 30 は、サービス提供装置 10 から出力音声を受信すると、その受信した出力音声を再生する。これによって、ユーザ 4 の対話相手には、日本語の「駅はどこですか？」とのことばが聞こえる。

[0018] 図示していないが、ユーザ 4 の対話相手も同様の携帯端末 30 を有していてもよく、例えば、ユーザ 4 からの質問に対して、「まっすぐ行って左です」との回答を自身の携帯端末に向かって発すると、上述したような処理が実行されて、ユーザ 4 に対して、対応する英語の「Go straight and turn left」ということばが回答される。

[0019] このように、多言語翻訳システム 1 においては、第 1 言語のことば（音声）と第 2 言語のことば（音声）との間で自在に翻訳が可能である。なお、2 つの言語に限らず、任意の数の言語間で相互に自動翻訳できるようにしてもよい。

[0020] このような自動音声翻訳の機能を利用することで、外国旅行や外国人とのコミュニケーションを容易化できる。

[0021] サービス提供装置 10 に含まれる本実施の形態に従う音声合成システムは、後述するように、自己回帰モデルを用いて、入力テキストから生成されるコンテキストラベル系列から対応する音声信号を再構成する。サービス提供

装置 10 は、音声合成システムに関するコンポーネントとして、分析部 12 と、学習部 14 と、自己回帰モデル 16 と、音声合成部 18 とを含む。

[0022] サービス提供装置 10 は、自動翻訳に関するコンポーネントとして、音声認識部 20 と、翻訳部 22 とを含む。サービス提供装置 10 は、さらに、携帯端末 30 との間で通信処理を行なうための通信処理部 24 を含む。

[0023] より具体的には、分析部 12 および学習部 14 は、自己回帰モデル 16 を構築するための機械学習を担当する。すなわち、分析部 12 および学習部 14 は、音声合成システムに向けられた学習装置として機能し、自己回帰モデル 16 を構築する。分析部 12 および学習部 14（学習装置）の機能および処理の詳細については、後述する。自己回帰モデル 16 は、分析部 12 および学習部 14 による機械学習の結果に相当する。

[0024] 音声認識部 20 は、通信処理部 24 を介して受信した携帯端末 30 からの入力音声に対して、音声認識処理を実行することで音声認識テキストを出力する。翻訳部 22 は、音声認識部 20 からの音声認識テキストから、指定された言語のテキスト（説明の便宜上、「翻訳テキスト」とも記す。）を生成する。音声認識部 20 および翻訳部 22 については、公知の任意の方法を採用できる。

[0025] 音声合成部 18 は、翻訳部 22 からの翻訳テキストに対して、自己回帰モデル 16 を参照して音声合成を行ない、その結果得られる出力音声を、通信処理部 24 を介して携帯端末 30 へ送信する。

[0026] 図 1 には、説明の便宜上、自己回帰モデル 16 を構築するための機械学習を担当するコンポーネント（主として、分析部 12 および学習部 14）と、生成された自己回帰モデル 16 を用いて多言語翻訳を担当するコンポーネント（主として、音声認識部 20、翻訳部 22、および音声合成部 18）が同一のサービス提供装置 10 に実装されている例を示すが、これらの機能をそれぞれ別の装置に実装してもよい。この場合、第 1 の装置において、機械学習を実施することで自己回帰モデル 16 を構築し、第 2 の装置において、当該生成された自己回帰モデル 16 を用いて音声合成および当該音声合成を利

用したサービスを提供するようにしてもよい。

[0027] 上述したような多言語翻訳サービスにおいては、音声認識部 20 および翻訳部 22 の少なくとも一部の機能を携帯端末 30 で実行されるアプリケーションが担当するようにしてもよい。また、音声合成を担当するコンポーネント（自己回帰モデル 16 および音声合成部 18）の機能を携帯端末 30 で実行されるアプリケーションが担当するようにしてもよい。

[0028] このように、サービス提供装置 10 および携帯端末 30 が任意の形態で協働することで、多言語翻訳システム 1 およびその一部である音声合成システムを実現できる。このとき、それぞれの装置が分担する機能については、状況に応じて適宜決定すればよく、図 1 に示される多言語翻訳システム 1 に限定されるようなものではない。

[0029] [B. サービス提供装置のハードウェア構成]

次に、サービス提供装置のハードウェア構成の一例について説明する。図 2 は、本実施の形態に従うサービス提供装置 10 のハードウェア構成例を示す模式図である。サービス提供装置 10 は、典型的には、汎用コンピュータを用いて実現される。

[0030] 図 2 を参照して、サービス提供装置 10 は、主要なハードウェアコンポーネントとして、プロセッサ 100 と、主メモリ 102 と、ディスプレイ 104 と、入力デバイス 106 と、ネットワークインターフェイス（I/F : interface）108 と、光学ドライブ 134 と、二次記憶装置 112 とを含む。これらのコンポーネントは、内部バス 110 を介して互いに接続される。

[0031] プロセッサ 100 は、後述するような各種プログラムを実行することで、本実施の形態に従うサービス提供装置 10 の実現に必要な処理を実行する演算主体であり、例えば、1 または複数の CPU（central processing unit）や GPU（graphics processing unit）などで構成される。複数のコアを有するような CPU または GPU を用いてもよい。

[0032] 主メモリ 102 は、プロセッサ 100 がプログラムを実行するにあたって、プログラムコードやワークメモリなどを一時的に格納する記憶領域であり

、例えば、D R A M (dynamic random access memory) や S R A M (static random access memory) などの揮発性メモリデバイスなどで構成される。

[0033] ディスプレイ 1 0 4 は、処理に係るユーザインターフェイスや処理結果などを出力する表示部であり、例えば、L C D (liquid crystal display) や有機 E L (electroluminescence) ディスプレイなどで構成される。

[0034] 入力デバイス 1 0 6 は、ユーザからの指示や操作などを受付けるデバイスであり、例えば、キーボード、マウス、タッチパネル、ペンなどで構成される。また、入力デバイス 1 0 6 としては、機械学習に必要な音声を収集するためのマイクロフォンを含んでいてもよいし、機械学習に必要な音声を収集した集音デバイスと接続するためのインターフェイスを含んでいてもよい。

[0035] ネットワークインターフェイス 1 0 8 は、インターネット上またはイントラネット上の携帯端末 3 0 や任意の情報処理装置などとの間でデータを遣り取りする。ネットワークインターフェイス 1 0 8 としては、例えば、イーサネット（登録商標）、無線 L A N (local area network)、B l u e t o o t h（登録商標）などの任意の通信方式を採用できる。

[0036] 光学ドライブ 1 3 4 は、C D - R O M (compact disc read only memory)、D V D (digital versatile disc) などの光学ディスク 1 3 6 に格納されている情報を読み出して、内部バス 1 1 0 を介して他のコンポーネントへ出力する。光学ディスク 1 3 6 は、非一過的 (non-transitory) な記録媒体の一例であり、任意のプログラムを不揮発的に格納した状態で流通する。光学ドライブ 1 3 4 が光学ディスク 1 3 6 からプログラムを読み出して、二次記憶装置 1 1 2 などにインストールすることで、汎用コンピュータがサービス提供装置 1 0（または、音声合成装置）として機能するようになる。したがって、本発明の主題は、二次記憶装置 1 1 2 などにインストールされたプログラム自体、または、本実施の形態に従う機能や処理を実現するためのプログラムを格納した光学ディスク 1 3 6 などの記録媒体でもあり得る。

[0037] 図 2 には、非一過的な記録媒体の一例として、光学ディスク 1 3 6 などの

光学記録媒体を示すが、これに限らず、フラッシュメモリなどの半導体記録媒体、ハードディスクまたはストレージテープなどの磁気記録媒体、MO (magneto-optical disk) などの光磁気記録媒体を用いてもよい。

[0038] 二次記憶装置 112 は、プロセッサ 100 にて実行されるプログラム、プログラムが処理対象とする入力データ（学習用の入力音声およびテキスト、ならびに、携帯端末 30 からの入力音声などを含む）、および、プログラムの実行により生成される出力データ（携帯端末 30 へ送信される出力音声などを含む）などを格納するコンポーネントであり、例えば、ハードディスク、SSD (solid state drive) などの不揮発性記憶装置で構成される。

[0039] より具体的には、二次記憶装置 112 は、典型的には、図示しない OS (operating system) の他、分析部 12 を実現するための分析プログラム 121 と、学習部 14 を実現するための学習プログラム 141 と、音声認識部 20 を実現するための音声認識プログラム 201 と、翻訳部 22 を実現するための翻訳プログラム 221 と、音声合成部 18 を実現するための音声合成プログラム 181 とを格納している。

[0040] これらのプログラムをプロセッサ 100 で実行する際に必要となるライブラリや機能モジュールの一部を、OS が標準で提供するライブラリまたは機能モジュールを用いて代替するようにしてもよい。この場合には、各プログラム単体では、対応する機能を実現するために必要なプログラムモジュールのすべてを含むものにはならないが、OS の実行環境下にインストールされることで、必要な機能を実現できる。このような一部のライブラリまたは機能モジュールを含まないプログラムであっても、本発明の技術的範囲に含まれ得る。

[0041] また、これらのプログラムは、上述したようないずれかの記録媒体に格納されて流通するだけでなく、インターネットまたはイントラネットを介してサーバ装置などからダウンロードすることで配布されてもよい。

[0042] なお、実際には、音声認識部 20 および翻訳部 22 を実現するためのデータベースが必要となるが、説明の便宜上、それらのデータベースについては

描いていない。

[0043] 二次記憶装置 112 は、自己回帰モデル 16 に加えて、自己回帰モデル 16 を構築するための、機械学習用の入力音声 130 および対応するテキスト 132 を格納していてもよい。

[0044] 図 2 には、単一のコンピュータがサービス提供装置 10 を構成する例を示すが、これに限らず、ネットワークを介して接続された複数のコンピュータが明示的または黙示的に連携して、多言語翻訳システム 1 およびその一部である音声合成システムを実現するようにしてもよい。

[0045] コンピュータ（プロセッサ 100）がプログラムを実行することで実現される機能の全部または一部を、集積回路などのハードワイヤード回路（hard-wired circuit）を用いて実現してもよい。例えば、ASIC（application specific integrated circuit）やFPGA（field-programmable gate array）などを用いて実現してもよい。

[0046] 当業者であれば、本発明が実施される時代に応じた技術を適宜用いて、本実施の形態に従う多言語翻訳システムを実現できるであろう。

[0047] [C. 概要]

本実施の形態に従う音声合成システムは、入力テキストに応じた音声を合成して出力するシステムであって、入力音声に対する予測量子化と上述の非特許文献 1 に開示される自己回帰モデルとを組み合わせることで、合成音声に生じる聴覚上のノイズを低減させる。

[0048] 予測量子化は、入力信号を直接量子化するのではなく、予測係数に基づいて生成される予測値と入力信号との残差信号を量子化する手法である。予測量子化により、入力信号は予測係数と残差信号とに分離される。予測量子化を音声信号に適用した場合、予測係数は声道フィルタを示すパラメータに相当し、残差信号は励振源に相当する。本実施の形態に従う音声合成システムにおいては、自己回帰モデルを用いて残差信号を推定する。典型的には、上述の非特許文献 1 に開示されるWaveNetと称される方式を採用してもよい。

[0049] このような、予測量子化を採用することで、推定誤差から生じるノイズの

スペクトル形状をシェーピングしてノイズをパワーの大きい帯域へ集中させることで、聴覚マスキング効果により、ノイズを知覚しにくくできる。

[0050] また、上述の非特許文献 1 に開示される WaveNet とは異なり、残差信号を推定するため、音声信号を直接推定する場合と比較して、必要なダイナミックレンジを低減できる。そのため、同一の量子化ビット数であれば、高精度に量子化することができ、音声品質を向上できる。

[0051] まず、予測量子化の概要について説明する。図 3 は、本実施の形態に従う音声合成システムにて採用する予測量子化の概要を説明するためのブロック図である。

[0052] 図 3 を参照して、予測量子化は、基本構成として、エンコーダ 200 およびデコーダ 210 を含む。エンコーダ 200 は、入力信号を予測係数および残差信号に分離する。デコーダ 210 は、残差信号から入力信号を再構成する。

[0053] より具体的には、エンコーダ 200 は、演算器 202、206 と、量子化器 204 と、予測器 208 とを含む。エンコーダ 200 においては、入力信号 x_t と予測器 208 が過去のサンプルに基づいて生成した $\hat{x}_t$ との残差信号 e_t が演算器 202 により算出され、その算出された残差信号 e_t が量子化器 204 により量子化されることで、量子化残差信号 $\hat{e}_t$ が算出される。なお、本来は、「 $\hat{e}$ 」は「 e 」の上に配置されるものであるが、使用可能な文字コードの都合上、「 e 」と「 $\hat{e}$ 」とを並べて記載する。また、「 $\sim$ 」についても同様である。

[0054] さらに、量子化残差信号 $\hat{e}_t$ と $\hat{x}_t$ とが演算器 206 により加算され、その加算結果が予測信号 $\hat{x}_t$ として予測器 208 へ与えられる。

[0055] そして、時刻 t における予測信号 $\hat{x}_t$ ($t=t$) に予測器 208 を適用することで時刻 $t+1$ における予測信号 $\hat{x}_t$ ($t=t+1$) が算出される。このようにして、エンコーダ 200 の内部では予測信号 $\hat{x}_t$ がサイクル毎に算出されるとともに、入力信号 x_t と算出される予測信号 $\hat{x}_t$ との差が量子化された上で、量子化残差信号 $\hat{e}_t$ として出力される。

[0056] デコーダ210は、エンコーダ200の予測器208と同様の動きをする予測器212と、演算器214とを含む。演算器214は、サイクル毎に入力される量子化残差信号 e_t^q と、量子化残差信号 e_t^q に対して予測器208から出力される予測結果とを加算することで、入力信号 x_t に対応する予測信号 $\hat{x}_t$ を再構成する。

[0057] 上述したような手順によって、エンコーダ200は入力信号 x_t についての量子化残差信号 e_t^q をサイクル毎に出力するとともに、デコーダ210は、量子化残差信号 e_t^q に基づいて入力信号 x_t を復元する。

[0058] 本実施の形態に従う音声合成システムにおいては、量子化残差信号 e_t^q を学習することで、コンテキストラベル系列に応じた、量子化残差についての自己回帰モデル16を構築する。

[0059] [D. 学習処理および音声合成処理]

次に、本実施の形態に従う音声合成システムにおける学習処理および音声合成処理の詳細について説明する。図4は、本実施の形態に従う音声合成システムにおける要部の処理を説明するためのブロック図である。

[0060] 図4を参照して、音声合成システムは、自己回帰モデル16を構築するための分析部12および学習部14と、自己回帰モデル16を用いて音声信号を出力する音声合成部18とを含む。以下、これらの各部の処理および機能について詳述する。

[0061] (d1: 分析部12)

まず、分析部12における処理および機能について説明する。分析部12は、音声分析を担当する部分であり、学習用の入力音声 x_t を、声道フィルタ係数 c と、励振源に相当する残差信号 e_t とに分離する。本実施の形態において、声道フィルタ係数 c は時不変とする。

[0062] より具体的には、分析部12は、逆フィルタ120と、声道特徴分析部122とを含む。声道特徴分析部122は、入力される音声信号 x_t を分析して、声道特徴を示す声道フィルタ係数 c を出力する。声道特徴分析部122は、声道フィルタ係数 c を、逆フィルタ120へ出力するとともに、声道フィ

ルタ係数格納部 146 に格納する。ここで、フィルタ係数は、LSP (line spectral pair)、LPC (linear prediction coefficients)、メルケプストラム係数のいずれを採用してもよい。以降 [G. 実験的評価] では、メルケプストラム係数を用いた例を示す。

[0063] 逆フィルタ 120 は、声道フィルタ係数 c に基づいて予測される音声信号と入力される音声信号との間の残差信号を出力する残差信号生成部に相当する。より具体的には、逆フィルタ 120 は、声道特徴分析部 122 からの声道フィルタ係数 c に基づいて、音声信号を内部的に予測するとともに、入力される音声信号 x_t とその予測した音声信号との残差信号 e_t を出力する。逆フィルタ 120 から出力される残差信号 e_t は、学習部 14 へ与えられる。

[0064] (d2 : 学習部 14)

次に、学習部 14 における処理および機能について説明する。学習部 14 は、分析部 12 から与えられる残差信号 e_t を量子化することで得られる量子化残差信号を自己回帰モデル 16 に入力する。ここで、誤差として、量子化残差信号と推定量子化残差信号との数値上での距離、または、量子化ビットに応じた one-hot ベクトルの cross-entropy を採用してもよい。学習部 14 は、量子化残差誤差と量子化推定誤差との間の差分（推定誤差）が最小化するように、自己回帰モデル 16 を構築する。

[0065] 学習部 14 は、各サンプルに対応するコンテキストラベル毎と、過去の入力された音声信号とに基づいて、自己回帰モデル 16 を構築する。本質的には、自己回帰モデル 16 は、過去の信号系列および現在のコンテキストラベルから現在値を推定するためのパラメータを格納する。より具体的には、学習部 14 は、学習制御部 140 と、量子化部 142 と、声道フィルタ係数格納部 146 とを含む。

[0066] なお、本実施の形態においては、量子化した信号同士の誤差を最小化しているが、推定残差信号 $\hat{e}_t$ と残差信号 e_t との誤差を最小化するようにしてもよい。

[0067] 図 4 に示す構成においては、コンテキストラベル系列を生成するコンポー

ネットとして、テキスト分析部 192 およびコンテキストラベル生成部 194 を含む。テキスト分析部 192 およびコンテキストラベル生成部 194 は、既知のテキストの文脈情報に基づくコンテキストラベルを生成する。

[0068] コンテキストラベルは、学習部 14 および音声合成部 18 の両方で用いるため、学習部 14 および音声合成部 18 が共通に利用する構成例を示している。しかしながら、学習部 14 および音声合成部 18 の各々に、コンテキストラベルを生成するためのコンポーネントをそれぞれ実装するようにしてもよい。

[0069] テキスト分析部 192 は、入力される学習用または合成対象のテキストを分析して、その文脈情報をコンテキストラベル生成部 194 へ出力する。コンテキストラベル生成部 194 は、テキスト分析部 192 からの分脈情報に基づいて、入力される学習用または合成対象のテキストのコンテキストラベルを決定して学習制御部 140 および音声合成制御部 180 へ出力する。

[0070] 量子化部 142 は、逆フィルタ 120（残差信号生成部）から出力される残差信号を量子化して量子化残差信号を生成する。ここで、量子化方式は μ -law 方式であってもよいし、学習データから統計的に、または、線形に、量子化幅を決定してもよい。また、量子化ビット数は一般的に音声で用いられる 16 ビットでもよいし、任意に設定してもかまわない。

[0071] 学習制御部 140 は、コンテキストラベル生成部 194 から与えられるコンテキストラベルを条件として、自己回帰モデル 16 を学習させる。すなわち、学習制御部 140 は、既知の入力テキストに対応する入力音声信号について、既知の入力テキストのコンテキストラベルを自己回帰モデル 16 に対して条件として与えるとともに、入力音声信号についての量子化残差信号を学習制御部 140 に入力することで自己回帰モデル 16 の学習を実施する。自己回帰モデル 16 の詳細およびその構築方法については、後述する。

[0072] 声道フィルタ係数格納部 146 は、入力音声信号についての声道フィルタ係数 c を格納する格納部に相当する。

[0073] (d3：音声合成部 18)

次に、音声合成部 18 における処理および機能について説明する。音声合成部 18 は、合成対象のテキストから生成されるサンプル毎のコンテキストラベルを生成し、生成したサンプル毎のコンテキストラベルに応じた自己回帰モデル 16 に過去の推定量量子化残差信号を入力することで、現在の推定量量子化残差信号を得る。

[0074] より具体的には、音声合成部 18 は、音声合成制御部 180 と、逆量子化部 182 と、合成フィルタ 184 とを含む。

[0075] 何らかの合成対象のテキストが入力されると、テキスト分析部 192 が入力されたテキストを分析して文脈情報を出し、コンテキストラベル生成部 194 が文脈情報に基づいてコンテキストラベルを生成する。すなわち、テキスト分析部 192 およびコンテキストラベル生成部 194 は、任意のテキストの入力に応答して、当該テキストの文脈情報に基づくコンテキストラベルを決定する。

[0076] 音声合成制御部 180 は、未知の入力テキストが入力されると、当該未知の入力テキストのコンテキストラベルを自己回帰モデル 16 に対して条件として与えると同時に、過去の推定量量子化残差信号を自己回帰モデル 16 に入力し、現在の推定量量子化残差信号を得る。現在の推定量量子化残差信号を過去の系列の 1 点として追加入力し、次の 1 時刻先の推定量量子化残差信号を推定する。この推定を再帰的に最終点まで繰り返す。

[0077] 逆量子化部 182 は、音声合成制御部 180 が最終点まで推定した推定量量子化残差信号を逆量子化して、推定残差信号 $\hat{e}_t$ を生成する。

[0078] 合成フィルタ 184 は、声道フィルタ係数格納部 146 から読み出した声道フィルタ係数 c に基づいて、逆量子化部 182 からの推定残差信号をフィルタリングすることで、合成後の音声信号を出力する。すなわち、合成フィルタ 184 は、逆量子化部 182 から出力される推定残差信号を声道フィルタ係数 c に基づいてフィルタリングした結果を音声信号として出力する。

[0079] 本実施の形態に従う音声合成システムにおいては、量子化された残差信号 e_t を自己回帰モデル 16 および音声合成制御部 180 にて再帰的に推定し、

その推定結果を逆量子化することで、推定残差信号 $\hat{e}_t$ が生成される。このとき、推定誤差 ($|e_t - \hat{e}_t|$) は、量子化される各ビットに対して一様に分布する。合成フィルタ 184 にて推定残差信号 $\hat{e}_t$ が声道フィルタ係数 c でフィルタリングされることにより、生成される音声信号の音声スペクトルは聴覚的に重み付けされる。その結果、推定残差信号 $\hat{e}_t$ に含まれる推定誤差は、音声スペクトルの形状に応じたパワーの大きい帯域に集中させることができる。これにより、聴覚マスキング効果が発揮され、合成音声に含まれるノイズを低減できる。

[0080] [E. 自己回帰モデル]

次に、本実施の形態に従う音声合成システムに用いられる自己回帰モデル 16 について説明する。自己回帰モデル 16 として、上述の非特許文献 1 に開示される WaveNet と同様の構成を想定する。

[0081] WaveNet は、PixelCNN (例えば、非特許文献 2 参照) に類似した生成モデルであり、過去の信号系列から現在のサンプル (現在値) を推定する自己回帰モデルとして表現される。一般化すれば、音声信号 $x = \{x_1, \dots, x_T\}$ の結合確率は、条件付き確率の積として以下の (1) 式のように表現できる。

[0082] [数1]

$$P(\mathbf{x}) = \prod_{t=1}^T p(x_t | x_1, \dots, x_{t-1}) \quad \dots(1)$$

[0083] 本実施の形態に従う音声合成システムのように、テキスト音声合成を行う場合は、コンテキストラベル h を条件として加え、条件付き確率 $p(x | h)$ として以下の (2) 式のようにモデル化できる。

[0084] [数2]

$$P(\mathbf{x} | \mathbf{h}) = \prod_{t=1}^T p(x_t | x_1, \dots, x_{t-1}, \mathbf{h}) \quad \dots(2)$$

[0085] 図 5 は、本実施の形態に従う音声合成システムに用いられる自己回帰モデル 16 の概要を説明するためのブロック図である。図 5 に示すように、本実施の形態に従う自己回帰モデル 16 は、統計的データモデリングが可能なネ

ットワーク構造を有している。

[0086] 具体的には、WaveNetのネットワーク構成は、図5に示すように、入力に対して、causal convolution (161) を経て、複数の layer_k (162) が積層されている。最終的に、それぞれの layer (162) からskip connectionの要素として出力される $s_1, \dots, s_k$ が結合要素 (163) にて結合される。

[0087] 結合された出力は、ReLU (activation関数164, 166) および畳み込み (165, 167) が適用された後、softmax関数 (168) に入力される。softmax関数 (168) から出力される信号が推定値として出力される。

[0088] 各 layer (162) においては、dilated causal convolution (1621) を用いて、過去の信号系列を効率的にモデル化する。dilated causal convolution (1621) は、入力信号をあるサンプルずつスキップして畳み込むことにより、計算量の削減および大域的な変化を学習することを実現する。

[0089] dilated causal convolution (1621) の出力は、gated activation unit (1622) に入力される。gated activation unit (1622) においては、dilated causal convolution (1621) の出力は、双曲線関数1623およびシグモイド関数1624に入力される。双曲線関数1623およびシグモイド関数1624のそれぞれからの出力は、アダマール要素1625にて要素毎の積が計算される。

[0090] layer_k (162) への入力ベクトル x_k に対して、gated activation unit (1622) の出力ベクトル z_k は、以下の(3)式に従って算出できる。

[0091] [数3]

$$z_k = \tanh(W_{\text{filter},k} * x_k + b_{\text{filter},k}) \circ \sigma(W_{\text{gate},k} * x_k + b_{\text{gate},k}) \quad \dots(3)$$

[0092] ここで、 $*$ は畳み込み演算を示し、 $\circ$ はelement-wise (要素毎) の掛け算を示し、 $\sigma()$ はシグモイド関数を示し、 k は layer のインデックを示す。また、 $W_{\text{filter},k}$ および $W_{\text{gate},k}$ は、 layer_k の畳み込みフィルタを示し、 $b_{\text{filter},k}$ および $b_{\text{gate},k}$ は、 layer_k の畳み込みバイアス項を示す。

[0093] 出力ベクトル z_k に対してresidual connection (1628) を適用した後

に、次のlayerへの入力として与える。residual connection (1628)においては、出力ベクトル z_k に対して入力ベクトル x_k が加算される。また、出力ベクトル z_k は、 1×1 の畳み込みが適用された後、skip connectionの要素 s_k として出力される。

[0094] 図5に示す自己回帰モデルにおいて、誤差関数として、二乗誤差ではなく、softmaxに基づくcross entropyが採用された場合、ガウス分布を仮定した平均ベクトルの誤差最小問題ではなく、音声信号の振幅値の多クラス分類問題へと置き換えられる。このような問題置き換えにより、入力に仮説を設けることなく、より柔軟でかつ曖昧な分布をモデル化できる。

[0095] 非特許文献1に開示されるWaveNetにおいては、量子化に μ -law方式が採用されており、量子化された信号は各量子化ビットにおいて一様な確率で分布する。多クラス分類問題であるため、WaveNetにより生じる推定誤差も一様に分布し、再構成された信号には推定誤差が一様に分布してしまう。その結果、信号パワーの小さい帯域（特に、高周波帯域）にてノイズが相対的に大きくなり、ノイズが知覚されやすくなるという課題がある。これに対して、本実施の形態に従う音声合成システムにおいては、予測量子化を組み合わせることで、このような課題を解決する。

[0096] 本実施の形態に従う音声合成システムにおいては、非特許文献1に開示されるWaveNetに限らず、自己回帰モデルに従って音声信号を直接的に再構成できるものであり、どのようなネットワーク構成にも適用可能である。

[0097] [F. 処理手順]

図6および図7は、本実施の形態に従う音声合成システムにおける処理手順の一例を示すフローチャートである。より具体的には、図6には、自己回帰モデル16を構築するための事前の学習処理に係る手順を示し、図7には、自己回帰モデル16を用いた音声合成処理に係る手順を示す。図6および図7に示す各ステップは、1または複数のプロセッサ（例えば、図2に示すプロセッサ100）が1または複数のプログラムを実行することで実現されてもよい。

- [0098] 図6を参照して、プロセッサ100は、既知のテキストおよび当該テキストに対応する音声信号が入力されると（ステップS100）、入力されたテキストを分析して文脈情報を生成し（ステップS102）、当該生成された文脈情報に基づいてコンテキストラベルを決定する（ステップS104）。
- [0099] 続いて、プロセッサ100は、入力された音声信号を分析して声道フィルタ係数を決定し（ステップS106）、決定した声道フィルタ係数に基づいて、入力された音声信号から残差信号を生成する（ステップS108）。さらに、プロセッサ100は、生成した残差信号を量子化して量子化残差信号を生成する（ステップS110）。
- [0100] そして、プロセッサ100は、自己回帰モデル16の内部パラメータをランダムに決定するとともに、決定したコンテキストラベルを条件として設定する（ステップS112）とともに、過去の量子化残差信号と現在のコンテキストラベルとを、現在の量子化残差信号に対応付けて自己回帰モデル16を学習する（ステップS114, S116）。
- [0101] プロセッサ100は、自己回帰モデル16に従って推定量子化残差信号を算出する（ステップS116）。そして、プロセッサ100は、学習の終了条件が満たされたか否かを判断する（ステップS118）。学習の終了条件としては、例えば、入力された音声信号の回数が規定値に到達したこと、自己回帰モデルからの推定値についての推定誤差が予め定められたしきい値以下になったことなどが想定される。
- [0102] 学習の終了条件が満たされていない場合（ステップS118においてNOの場合）には、ステップS114以下の処理が繰返される。このステップS114～S118の処理が繰返されることで、自己回帰モデルに入力された残差信号と推定残差信号との間の差分（推定誤差）が最小化するように、自己回帰モデル16が構築される。
- [0103] このように、既知の入力テキストのコンテキストラベルを自己回帰モデル16に対して条件として与えるとともに、量子化残差信号を自己回帰モデル16に対して入力することで、自己回帰モデル16の学習が実施される。

- [0104] 一方、学習の終了条件が満たされている場合（ステップS 1 1 8においてY E Sの場合）には、学習処理は終了する。
- [0105] 次に、図7を参照して、プロセッサ1 0 0は、合成対象のテキストが入力されると（ステップS 2 0 0）、入力されたテキストを分析して文脈情報を生成し（ステップS 2 0 2）、当該生成された文脈情報に基づいて、対応するフレームについてのコンテキストラベルを決定する（ステップS 2 0 4）。そして、プロセッサ1 0 0は、ステップS 2 0 4において決定したコンテキストラベルを自己回帰モデル1 6に条件として設定する（ステップS 2 0 6）。
- [0106] 続いて、プロセッサ1 0 0は、過去の推定量子化残差信号およびコンテキストラベルを自己回帰モデル1 6に入力し（ステップS 2 0 8）、自己回帰モデル1 6に従って、入力に対する推定量子化残差信号を算出する（ステップS 2 1 0）。そして、プロセッサ1 0 0は、自己回帰モデル1 6の最終点まで処理が完了したか否かを判断する（ステップS 2 1 2）。最終点まで処理が完了していない場合（ステップS 2 1 2においてN Oの場合）、ステップS 2 0 8以下の処理が繰返される。
- [0107] 最終点まで処理が完了している場合（ステップS 2 1 2においてY E Sの場合）には、プロセッサ1 0 0は、再帰的に推定された推定量子化残差信号を逆量子化した推定残差信号を生成し（ステップS 2 1 4）、声道フィルタ係数を合成フィルタ（声道フィルタ）に設定し（ステップS 2 1 6）、生成される推定残差信号を声道フィルタ係数が設定された合成フィルタでフィルタリングすることで、合成後の音声信号を出力する（ステップS 2 1 8）。そして、入力されたテキストに対する音声合成処理は終了する。
- [0108] このように、未知の入力テキストのコンテキストラベルを自己回帰モデル1 6に対して条件として与えるとともに、過去の推定量子化残差信号を自己回帰モデル1 6を用いて、現在の量子化残差信号を再帰的に推定し、その推定された現在の量子化残差信号から音声信号を再構成する。
- [0109] [G. 実験的評価]

次に、本実施の形態に従う音声合成システムにおいて、ノイズ低減の有効性について実施した実験的評価について説明する。そのため、条件としてのコンテキストラベルを与えず、正解の音声波形のみを入力とする。

[0110] (g 1 : 実験条件)

本実施の形態に係る実施例の比較対象となる比較例は、上述の非特許文献 1 に開示される WaveNet を用いる方式とした。

[0111] 音声データとして、日本語女性話者 1 名により発声された A T R 音素バランス文および旅行会話文からなる 7 3 9 5 文を用いた。このうち、7 3 6 5 文を学習データとして用いるとともに、残り 3 0 文をテストデータとして用いた。

[0112] 音声データは、サンプリング周波数が 4 8 k H z であったものを 1 6 k H z にダウンサンプリングした上で、ハイパスフィルタを適用して 5 0 H z 以下の成分を除去したものをを用いた。量子化方式は、8 b i t の μ -l a w 方式とし、自己回帰モデル (WaveNet) への入力は one-hot ベクトルとした。

[0113] 実施例における声道フィルタ係数 c (合成フィルタ) としては、1 1 9 次メルケプストラム係数 (0 次を含めて 1 2 0 次) を用いた。本評価においては、学習データから算出した時不変のメルケプストラム係数を用いてフィルタリングして残差信号を生成した。残差信号は、- 1 から 1 の範囲で正規化を行なった。

[0114] また、声道フィルタによる聴覚重み付けについて、聴覚重み付け係数を採用した。具体的には、メルケプストラム係数のうち 0 次を除く各係数を定数倍することにより、聴覚重み付けフィルタのダイナミックレンジを変化させて、聴覚重み付けの強度を調整した。

[0115] このように、声道フィルタ係数は、聴覚重み付け係数により調整可能になっていてもよい。

[0116] 自己回帰モデル (WaveNet) のネットワーク構成として、causal convolution のフィルタ長を 3 2 とし、skip connection の要素数を 5 0 とし、layer を 1, 2, 4, 8 から 5 1 2 サンプルまでの dilation をもった 1 0 個の dilat

ed causal convolution layerを5個積層されたものを採用した。すなわち、本評価に用いたネットワーク構成においては、receptive fieldは320ms（5120サンプル）となる。Residual connectionのフィルタ長を2とし、出力チャンネル数を32とし、skip connectionのフィルタ長を2とし、出力チャンネル数を512とした。

[0117] モデル学習のoptimizerはAdamとし、学習係数は 1.0×10^{-3} とし、バッチサイズは100,000サンプルとし、試行回数は10万回とした。

[0118] プログラムの実行には、3台のGPUを用いて、バッチサイズを等分割して、並列学習を行なった。

[0119] 「比較例」は、上述の非特許文献1に開示されるWaveNetを用いる方式を用いるものであり、 μ -law方式により量子化された過去の音声信号系列から現在のサンプルを推定する。

[0120] 「実施例」は、上述したような本実施の形態に従う音声合成システムに対応するものであり、予測量子化に自己回帰モデルを提供したものである。過去の残差信号系列からWaveNetを用いて現在のサンプルを推定する。推定された残差信号から声道フィルタでフィルタリングすることで、推定された音声信号を得る。

[0121] なお、「比較例」および「実施例」のいずれについても、音声生成時には、源信号を入力として用いた。

[0122] （g2：実験結果：ノイズシェーピング）

まず、聴覚重み付けによるノイズシェーピング効果についての実験結果の評価について説明する。具体的には、以下に説明するような方法で生成された音声信号と源信号との誤差の周波数特性を分析した。

[0123] 図8は、本実施の形態に従う音声合成システムについてのノイズシェーピング効果を評価した結果の一例を示す図である。図8には、テストデータから10文をサンプリングして、平均した結果を示す。図8中の凡例は、それぞれ以下のような意味を示す。なお、 α は、実施例において採用する聴覚重み付け係数を示す。

- [0124] 「源信号」：源信号を μ -law方式により量子化した後、さらに逆量子化して再構成した信号と源信号との誤差を意味する。すなわち、 μ -law方式による量子化により生じる誤差の周波数特性を示す。
- [0125] 「残差 ($\alpha = 0.5$)」, 「残差 ($\alpha = 1.0$)」：実施例で用いる残差信号を μ -law方式により量子化した後、さらに逆量子化して再構成した信号を声道フィルタでフィルタリングすることで得られた信号と源信号との誤差を意味する。すなわち、自己回帰モデルを用いた推定に誤りがないと仮定した場合に生じる誤差の周波数特性を示す。
- [0126] 「比較例」：源信号を μ -law方式により量子化した信号をWaveNetにより推定した後、逆量子化して再構成した信号と源信号との誤差を意味する。すなわち、比較例において生じる誤差の周波数特性を示す。
- [0127] 「実施例 ($\alpha = 0.5$)」, 「実施例 ($\alpha = 1.0$)」：実施例で用いる残差信号を μ -law方式により量子化した信号を自己回帰モデルにより推定した後、逆量子化して再構成した信号と源信号との誤差を意味する。すなわち、実施例において生じる誤差の周波数特性を示す。
- [0128] 図8に示す実験結果によれば、「源信号」および「比較例」は、想定のとおり、帯域全体にわたって一様に残差が分布していることがわかる。これに対して、「残差」および「実施例」は、200Hz~300Hz付近にピークが存在しており、また高周波帯域では、「源信号」および「比較例」よりもパワーが小さくなっている。さらに、「残差」および「実施例」のパワースペクトラムの形状からも、聴覚特性に応じてシェーピングされていることが確認できる。
- [0129] また、聴覚重み付け係数 α の大きさに追従してパワースペクトラムの形状も変化しており、聴覚重み付け係数 α を調整することで、シェーピング度合いを制御できることも確認できる。
- [0130] 図8において、「源信号」および「残差」は量子化誤差を含むものであり、「比較例」および「実施例」は推定誤差および量子化誤差を含むものとなる。これらを比較することで、推定誤差は、量子化誤差を大きく上回っている。

ることが確認できる。

[0131] (g 3 : 客観評価)

次に、実施例と比較例との客観評価指標として、源信号と推定された音声信号との間のSNR (S/N比) を用いた。その結果を以下の表に示す。

[0132] 実施例は、聴覚重み付け係数 α が0.1のときにわずかながら比較例より良好な結果を示したが、他の場合は劣っている。この結果から、音声スペクトルのシェーピングは、源信号に対するSNRという面では改善方向にあまり寄与しないと考えられる。

[0133] [表1]

	SNR (dB)
実施例 ($\alpha=0.1$)	17.3
実施例 ($\alpha=0.5$)	16.8
実施例 ($\alpha=1.0$)	16.1
比較例	17.2

[0134] (g 4 : 主観評価)

次に、実施例と比較例との間における合成音声の自然性を対比較実験により比較した。評価音声は、テストデータから抽出した30文とした。実施例および比較例のそれぞれによって生成された合成音声を被験者（内訳：男性3名、女性2名）に聞いてもらい、より自然性である（音声品質が高い）と感じたものを選択してもらった。但し、提示音声対に差が感じられない際には、「どちらでもない」という選択肢を認めた。

[0135] 図9は、本実施の形態に従う音声合成システムについての対比較実験の評価結果例を示す図である。図9において、聴覚重み付け係数 $\alpha=0.1, 0.5, 1.0$ におけるp値は、それぞれ、 2.0×10^{-3} , 7.2×10^{-10} , 0.8×10^{-3} となった。聴覚重み付け係数 $\alpha=1.0$ では、比較例が実施例に対して有意差 ($p < 0.01$) を示したが、聴覚重み付け係数 $\alpha=0.1, 0.5$ では、実施例が比較例に対して有意差を示した。

[0136] [H. まとめ]

本実施の形態に従う音声合成システムにおいては、過去の信号系列から現在値を推定する自己回帰モデルに予測量子化の手法を組み合わせることで、再構成された音声信号の帯域全体にわたって存在していたノイズを聴覚マスキングを考慮した形状に変化させることができる。これによって、入力テキストに基づくコンテキストラベルから音声信号を直接的に推定する際の音声品質を高めることができる。

[0137] 今回開示された実施の形態は、すべての点で例示であって制限的なものではないと考えられるべきである。本発明の範囲は、上記した実施の形態の説明ではなくて請求の範囲によって示され、請求の範囲と均等の意味および範囲内でのすべての変更が含まれることが意図される。

符号の説明

[0138] 1 多言語翻訳システム、2 ネットワーク、4 ユーザ、10 サービス提供装置、12 分析部、14 学習部、16 自己回帰モデル、18 音声合成部、20 音声認識部、22 翻訳部、24 通信処理部、30 携帯端末、100 プロセッサ、102 主メモリ、104 ディスプレイ、106 入力デバイス、108 ネットワークインターフェイス、110 内部バス、112 二次記憶装置、120 逆フィルタ、121 分析プログラム、122 声道特徴分析部、130 入力音声、132 テキスト、134 光学ドライブ、136 光学ディスク、140 学習制御部、141 学習プログラム、142 量子化部、146 声道フィルタ係数格納部、180 音声合成制御部、181 音声合成プログラム、182 逆量子化部、184 合成フィルタ、192 テキスト分析部、194 コンテキストラベル生成部、200 エンコーダ、201 音声認識プログラム、202, 206, 214 演算器、204 量子化器、208, 212 予測器、210 デコーダ、221 翻訳プログラム。

請求の範囲

[請求項1]

音声合成システムに向けられた学習装置であって、

過去の信号系列および現在のコンテキストラベルから現在の信号を推定する自己回帰モデルを備え、前記自己回帰モデルは、統計的データモデリングが可能なネットワーク構造を有しており、

入力音声信号を分析して声道特徴を示す声道フィルタ係数を決定する声道特徴分析部と、

前記声道フィルタ係数に基づいて予測される音声信号と前記入力音声信号との間の残差信号を出力する残差信号生成部と、

前記残差信号生成部から出力される前記残差信号を量子化して量子化残差信号を生成する量子化部と、

既知の入力テキストに対応する入力音声信号について、前記既知の入力テキストのコンテキストラベルを前記自己回帰モデルに対して条件として与えるとともに、前記入力音声信号についての前記量子化残差信号の過去の系列および現在のコンテキストラベルと、前記量子化残差信号の現在の信号とを対応付けることで前記自己回帰モデルの学習を実施する学習制御部とを備える、学習装置。

[請求項2]

入力テキストに応じた音声合成して出力する音声合成システムであって、

未知の入力テキストが入力されると、前記未知の入力テキストのコンテキストラベルを前記自己回帰モデルに対して条件として与えるとともに、過去の推定量子化残差信号を請求項1に記載の学習装置によって構築された前記自己回帰モデルを用いて、現在の量子化残差信号を出力する音声合成制御部を備える、音声合成システム。

[請求項3]

前記量子化部から出力される過去の量子化残差信号および現在のコンテキストラベルから推定された推定量子化残差信号を逆量子化して推定残差信号を生成する逆量子化部と、

前記逆量子化部から出力される推定残差信号を前記声道フィルタ係

数に基づいてフィルタリングした結果を音声信号として出力する合成フィルタと、

前記入力音声信号についての声道フィルタ係数を格納する格納部とをさらに備える、請求項2に記載の音声合成システム。

[請求項4] 前記声道フィルタ係数は、聴覚重み付け係数により調整可能になっている、請求項2または3に記載の音声合成システム。

[請求項5] 入力テキストを分析して文脈情報を生成するテキスト分析部と、
前記テキスト分析部からの文脈情報に基づいて、入力テキストのコンテキストラベルを生成するコンテキストラベル生成部とをさらに備える、請求項2～4のいずれか1項に記載の音声合成システム。

[請求項6] 入力テキストに応じた音声を合成して出力する音声合成方法であって、

既知の入力テキストに対応する入力音声信号を分析して声道特徴を示す声道フィルタ係数を決定するステップと、

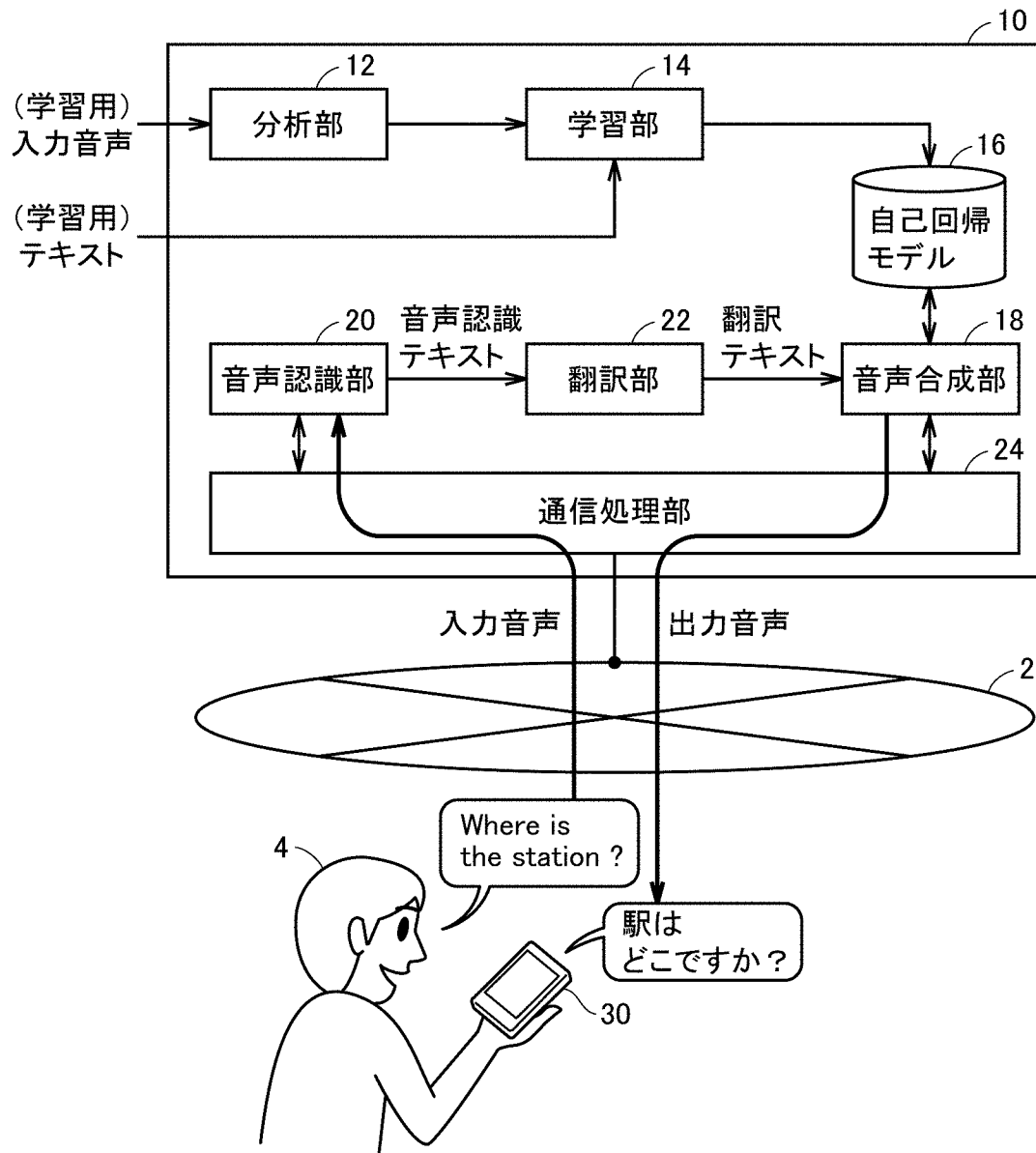
前記声道フィルタ係数に基づいて予測される音声信号と前記入力音声信号との間の残差信号を生成するステップと、

前記残差信号を量子化して量子化残差信号を生成するステップと、

前記既知の入力テキストのコンテキストラベルを自己回帰モデルに対して条件として与えるとともに、過去の前記量子化残差信号および現在のコンテキストラベルから現在の前記量子化残差信号を推定する前記自己回帰モデルを学習するステップとを備え、前記自己回帰モデルは、過去の信号系列および現在のコンテキストラベルから現在値を推定するためのパラメータを格納しており、統計的データモデリングが可能なネットワーク構造を有している、音声合成方法。

[図1]

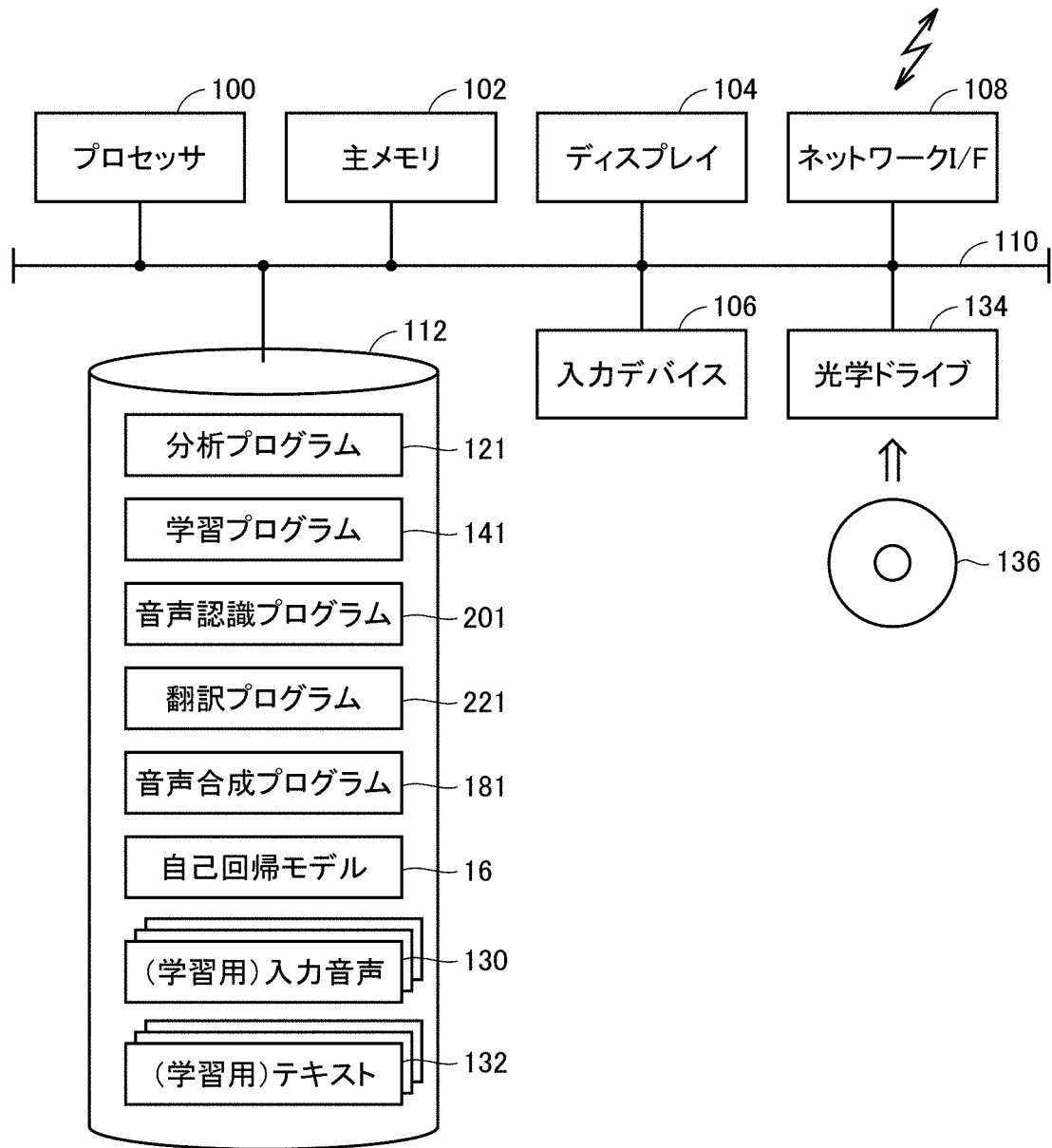
FIG.1

1

[図2]

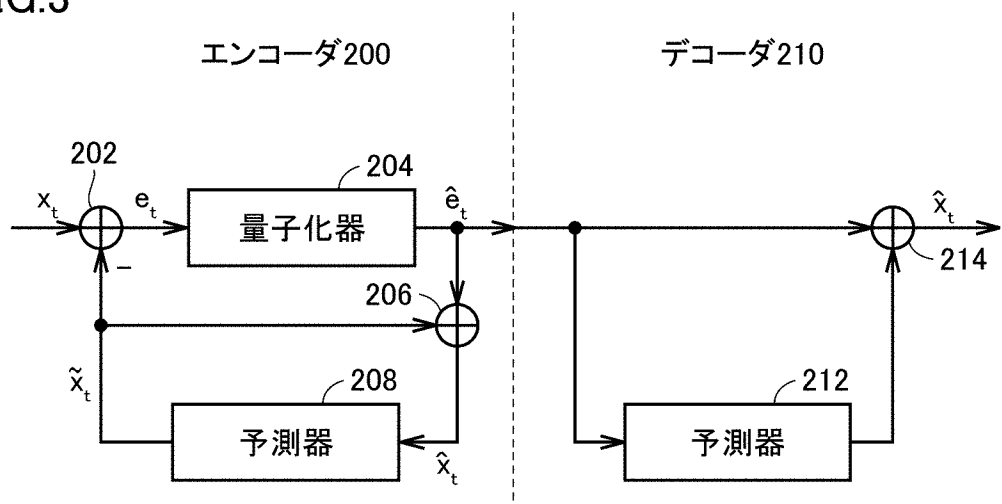
FIG.2

10



[図3]

FIG.3



[図5]

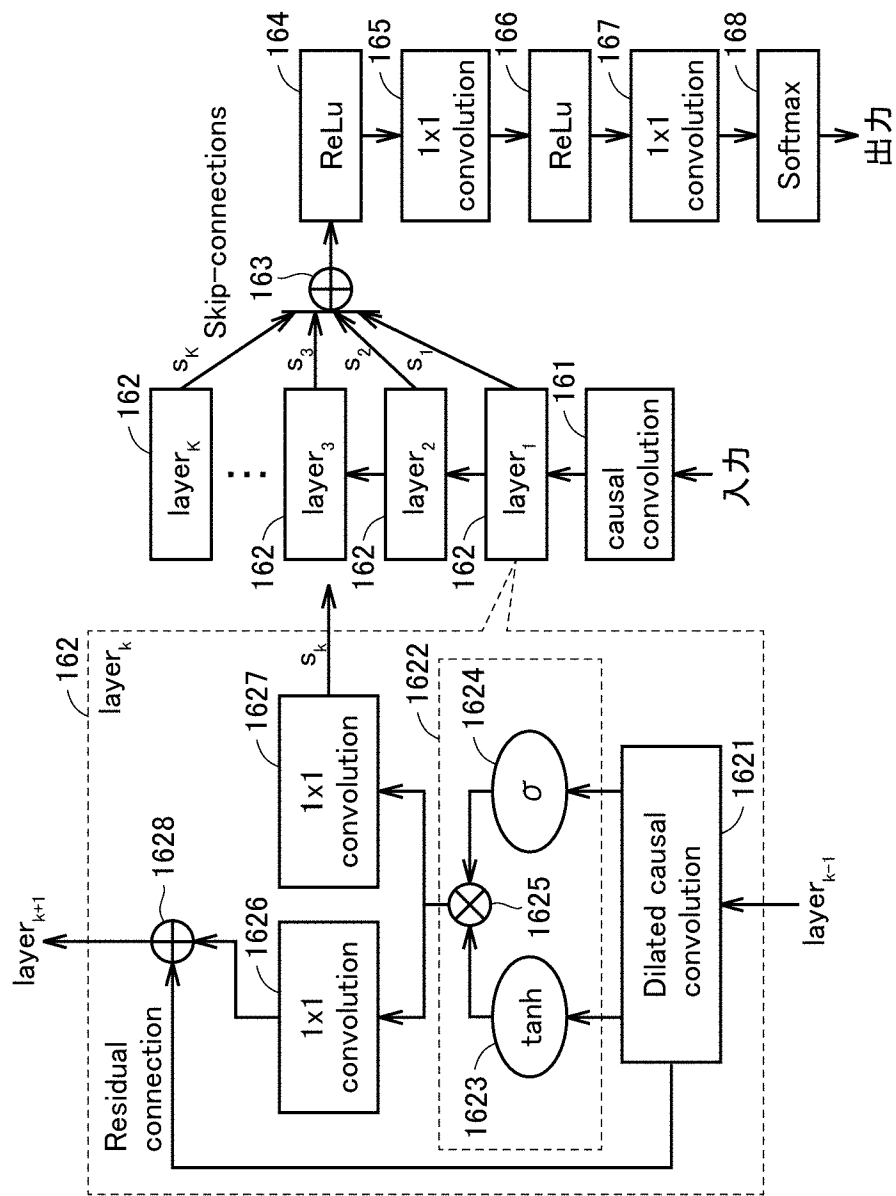
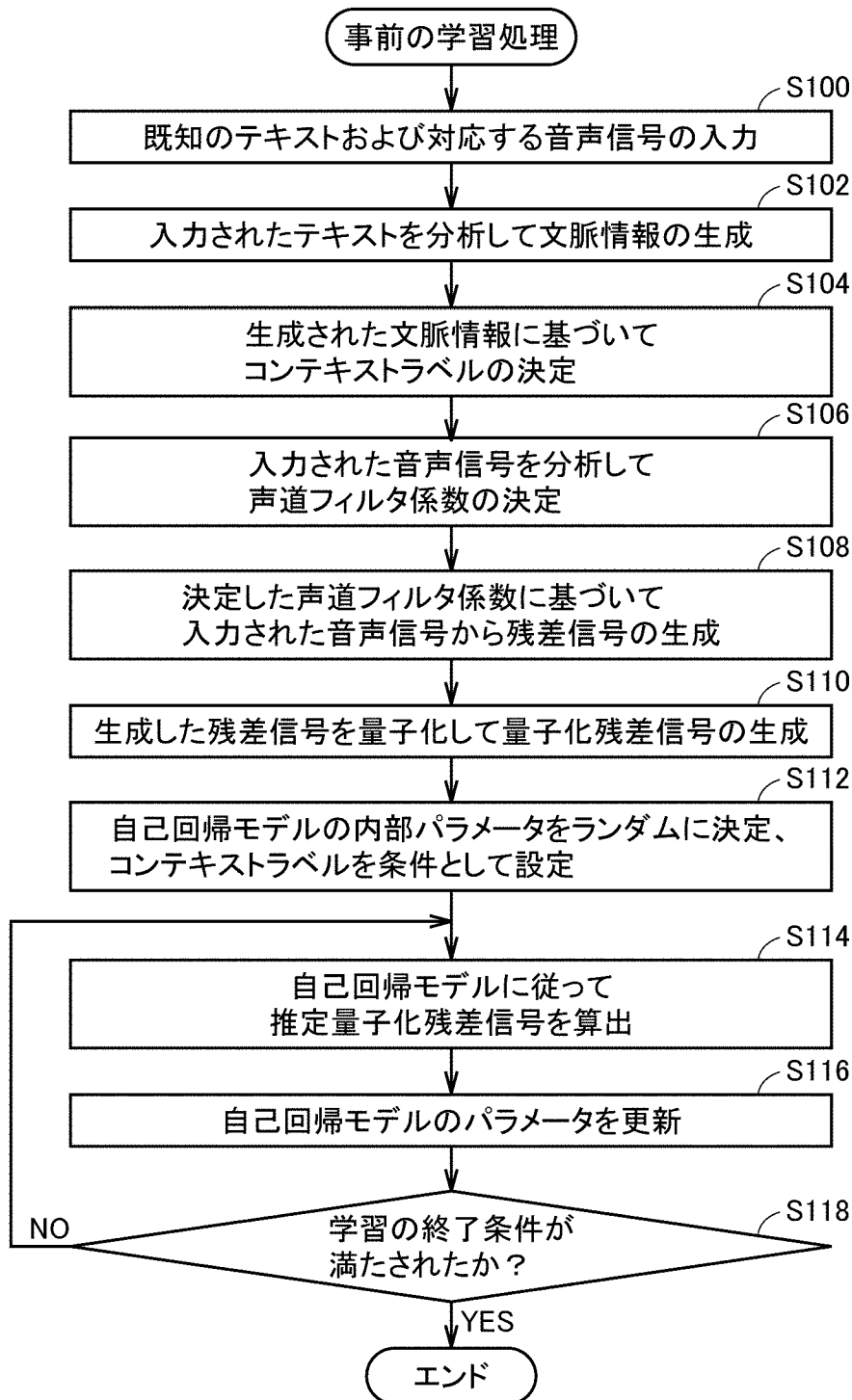


FIG. 5

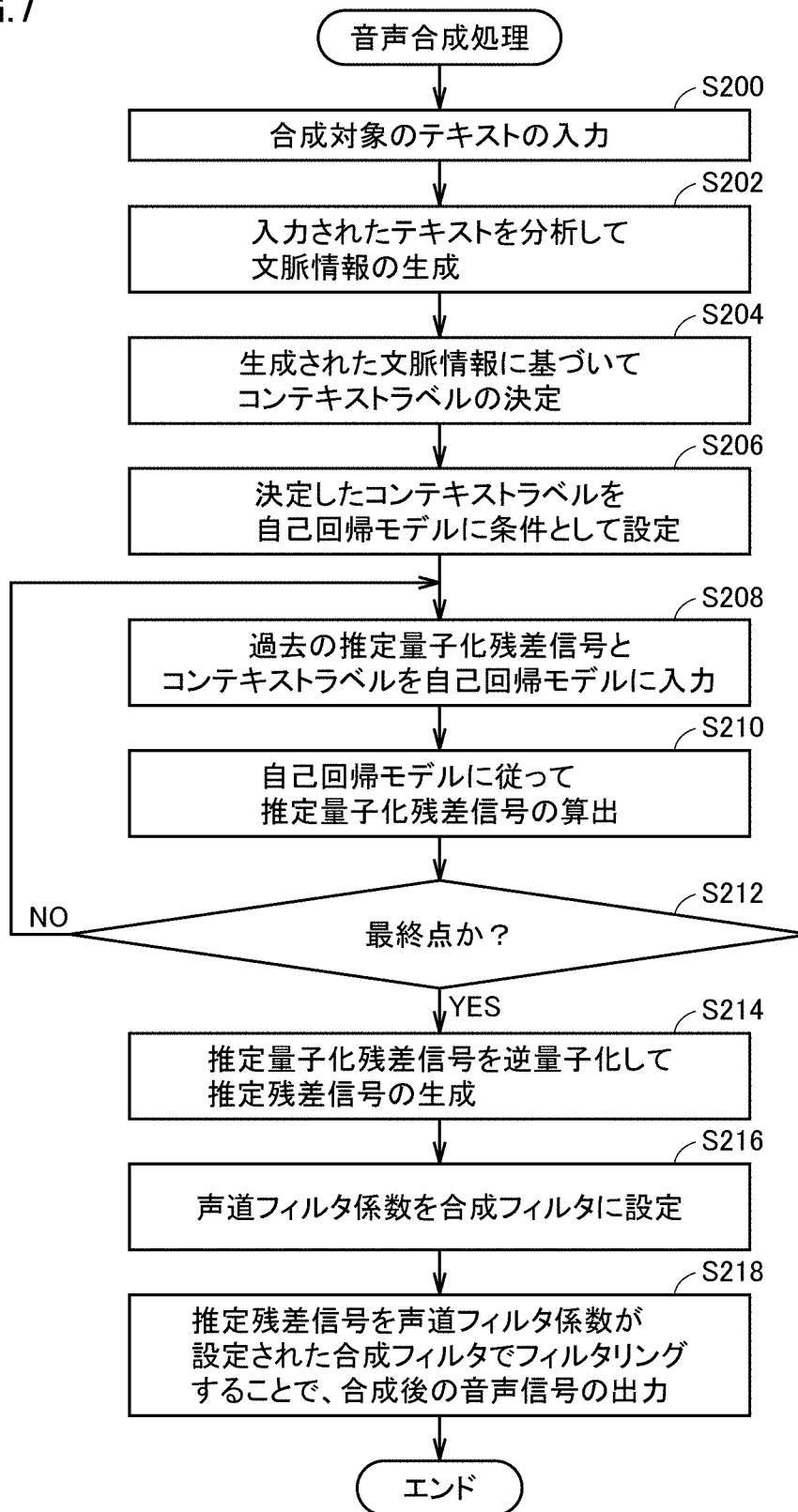
[図6]

FIG.6

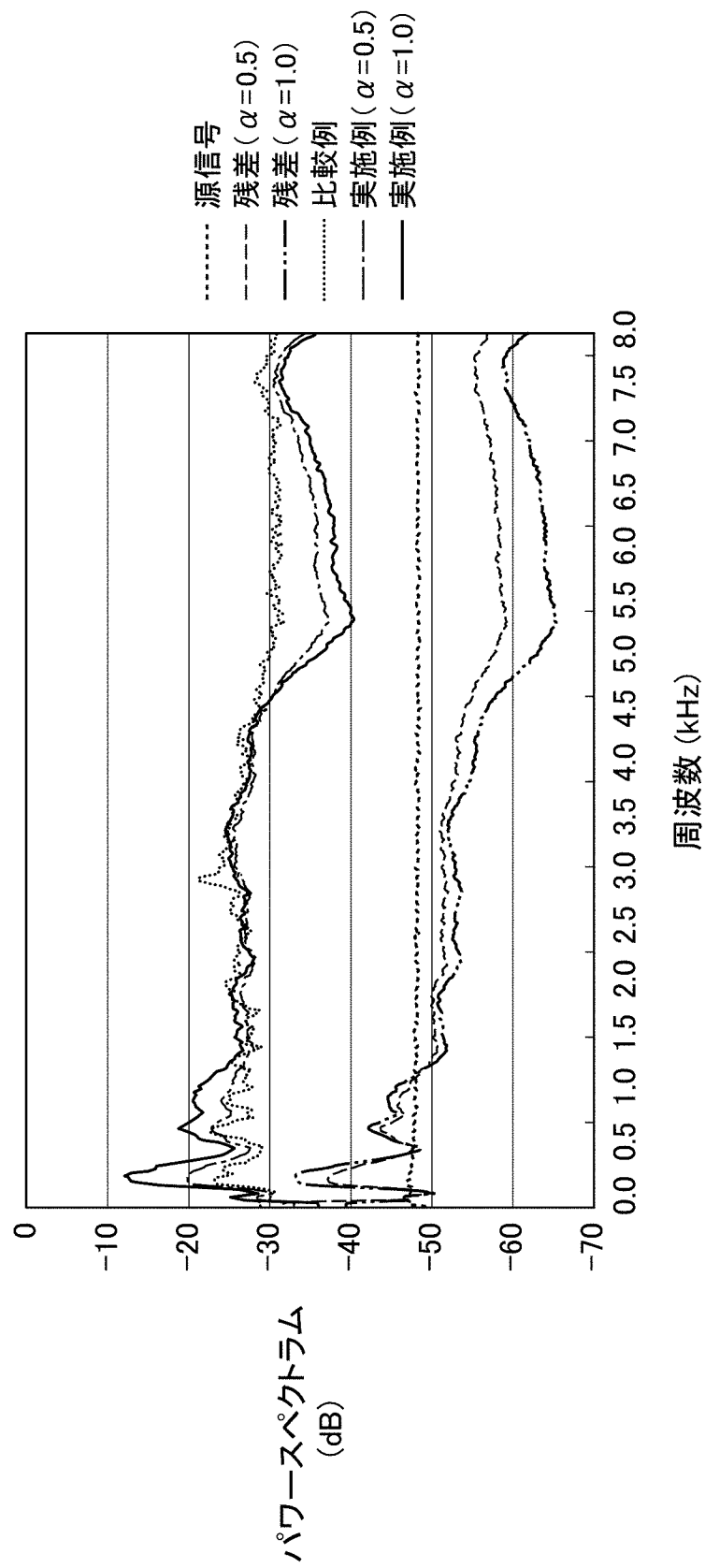


[図7]

FIG.7

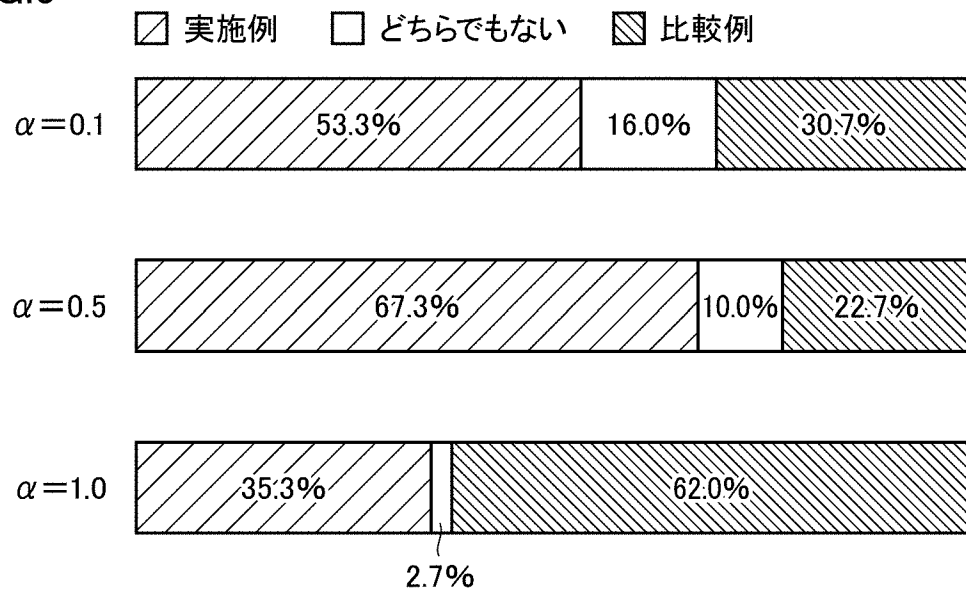


[図8]



[図9]

FIG.9



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2018/006166

A. CLASSIFICATION OF SUBJECT MATTER

Int. Cl. G10L13/06(2013.01)i, G10L13/10(2013.01)i, G10L19/06(2013.01)i

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)

Int. Cl. G10L13/06, G10L13/10, G10L19/06

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 1971-2018

Registered utility model specifications of Japan 1996-2018

Published registered utility model applications of Japan 1994-2018

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

IEEE Xplore

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	金子卓弘ほか, 統計的パラメトリック音声合成のための敵対的学習に基づくポストフィルタリング, 電子情報通信学会技術研究報告 IEICE Technical Report, 13 December 2016, vol. 116, no. 378, pp. 89-94, ISSN:0913-5685 (KANEKO, Takuhiro et al. Generative Adversarial Network-based Postfiltering for Statistical Parametric Speech Synthesis. IEICE Technical Report.)	1-6
A	WO 2009/022454 A1 (PANASONIC CORP.) 19 February 2009, entire text, all drawings & JP 4294724 B2 & US 2010/0004934 A1 & CN 101589430 A	1-6
A	JP 3-269599 A (SONY CORP.) 02 December 1991, entire text, all drawings (Family: none)	1-6

☐ 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
01.05.2018

Date of mailing of the international search report
15.05.2018

Name and mailing address of the ISA/
Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

A. 発明の属する分野の分類（国際特許分類（IPC））

Int.Cl. G10L13/06(2013.01)i, G10L13/10(2013.01)i, G10L19/06(2013.01)i

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））

Int.Cl. G10L13/06, G10L13/10, G10L19/06

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	1971-2018年
日本国実用新案登録公報	1996-2018年
日本国登録実用新案公報	1994-2018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）

IEEE Xplore

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	金子卓弘 ほか, 統計的パラメトリック音声合成のための敵対的学習に基づくポストフィルタリング, 電子情報通信学会技術研究報告 Vol. 116 No. 378 IEICE Technical Report, 2016.12.13, 第116巻、第378号, p.89-94, ISSN:0913-5685	1-6
A	WO 2009/022454 A1 (パナソニック株式会社) 2009.02.19, 全文、全図 & JP 4294724 B2 & US 2010/0004934 A1 & CN 101589430 A	1-6

☑ 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矛盾するものではなく、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

01.05.2018

国際調査報告の発送日

15.05.2018

国際調査機関の名称及びあて先

日本国特許庁（ISA/J P）

郵便番号100-8915

東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）

富澤 直樹

5Z

4188

電話番号 03-3581-1101 内線 3591

C (続き) . 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	JP 3-269599 A (ソニー株式会社) 1991. 12. 02, 全文、全図 (ファミリーなし)	1-6