(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6911208号
(P6911208)

(45)発行日　令和3年7月28日(2021.7.28)　(24)登録日　令和3年7月9日(2021.7.9)

(51)Int.Cl.		F I	
G10L 21/007	(2013.01)	G10L	21/007
G10L 25/30	(2013.01)	G10L	25/30
G10L 25/90	(2013.01)	G10L	25/90

請求項の数 27　(全 37 頁)

(21)出願番号	特願2020-542648 (P2020-542648)
(86)(22)出願日	平成31年2月14日 (2019.2.14)
(65)公表番号	特表2021-508859 (P2021-508859A)
(43)公表日	令和3年3月11日 (2021.3.11)
(86)国際出願番号	PCT/US2019/017941
(87)国際公開番号	WO2019/161011
(87)国際公開日	令和1年8月22日 (2019.8.22)
審査請求日	令和2年8月6日 (2020.8.6)
(31)優先権主張番号	18157080.5
(32)優先日	平成30年2月16日 (2018.2.16)
(33)優先権主張国・地域又は機関	欧州特許庁 (EP)
(31)優先権主張番号	62/710,501
(32)優先日	平成30年2月16日 (2018.2.16)
(33)優先権主張国・地域又は機関	米国 (US)

(73)特許権者　507236292
ドルビー　ラボラトリーズ　ライセンシング　コーポレイション
アメリカ合衆国　９４１０３　カリフォルニア州　サンフランシスコ　マーケット　ストリート　１２７５
(74)代理人　100107766
弁理士　伊東　忠重
(74)代理人　100070150
弁理士　伊東　忠彦
(74)代理人　100135079
弁理士　宮崎　修

最終頁に続く

(54)【発明の名称】発話スタイル転移

(57)【特許請求の範囲】

【請求項１】

コンピュータで実装されるオーディオ処理方法であって：

発話合成器をトレーニングすることを含み、該トレーニングは：

（ａ）一つまたは複数のプロセッサおよび一つまたは複数の非一時的記憶媒体を有する制御システムを介して実装される内容抽出プロセスによって、第一の人物の第一の発話に対応する第一のオーディオ・データを受領する段階と；

（ｂ）前記内容抽出プロセスによって、前記第一の発話に対応する第一のタイムスタンプ付けされた音素シーケンスおよび第一のピッチ輪郭データを生成する段階と；

（ｃ）前記制御システムを介して実装される第一のニューラルネットワークによって、前記第一のタイムスタンプ付けされた音素シーケンスおよび前記第一のピッチ輪郭データを受領する段階と；

（ｄ）前記第一のニューラルネットワークによって、前記第一のタイムスタンプ付けされた音素シーケンスおよび前記第一のピッチ輪郭データに対応する第一のニューラルネットワーク出力を生成する段階であって、前記第一のニューラルネットワーク出力は複数のフレーム・サイズを含む、段階と；

（ｅ）前記制御システムを介して実装される第二のニューラルネットワークによって、前記第一のニューラルネットワーク出力を受領する段階であって、前記第二のニューラルネットワークはモジュールの階層構成を含み、各モジュールは異なる時間分解能で動作し、前記第一のニューラルネットワークは、前記第一のニューラルネットワーク出力の前記

複数のフレーム・サイズのそれぞれが前記第二のニューラルネットワークのあるモジュールの時間分解能に対応するよう、前記第一のニューラルネットワーク出力を生成している、段階と；

（ｆ）前記第二のニューラルネットワークによって、第一の予測されたオーディオ信号を生成する段階と；

（ｇ）前記制御システムを介して、前記第一の予測されたオーディオ信号を第一の試験データと比較する段階であって、前記試験データは、前記第一の人物の発話に対応するオーディオ・データである、段階と；

（ｈ）前記制御システムを介して、前記第一の予測されたオーディオ信号についての損失関数値を決定する段階と；

（ｉ）前記第一の予測されたオーディオ信号についての現在の損失関数値と前記第一の予測されたオーディオ信号についての以前の損失関数値との間の差が所定の値以下になるまで（ａ）ないし（ｈ）を繰り返す段階であって、（ｆ）を繰り返すことは、前記第二のニューラルネットワークの少なくとも一つの重みに対応する少なくとも一つの非一時的記憶媒体位置の物理的状態を変更することを含む、段階とを含む、
オーディオ処理方法。

【請求項２】

（ａ）は、前記第一の人物の前記第一の発話に対応する第一のタイムスタンプ付けされたテキストを受領することをさらに含む、請求項１記載のオーディオ処理方法。

【請求項３】

（ａ）は、前記第一の人物に対応する第一の識別データを受領することをさらに含む、請求項１または２記載のオーディオ処理方法。

【請求項４】

前記発話合成器を発話生成のために制御することをさらに含み、該発話生成は：

（ｊ）前記内容抽出プロセスによって、第二の人物の第二の発話に対応する第二のオーディオ・データ、前記第二の発話に対応する第二のタイムスタンプ付けされたテキストおよび前記第一の人物の発話に対応する第一の識別データを受領する段階と；

（ｋ）前記内容抽出プロセスによって、前記第二の発話に対応する第二のタイムスタンプ付けされた音素シーケンスおよび第二のピッチ輪郭データを生成する段階と；

（ｌ）前記第一のニューラルネットワークによって、（ｋ）の前記第二のタイムスタンプ付けされた音素シーケンスおよび前記第二のピッチ輪郭データを受領する段階と；

（ｍ）前記第一のニューラルネットワークによって、（ｋ）の前記第二のタイムスタンプ付けされた音素シーケンスおよび前記第二のピッチ輪郭データに対応する第一のニューラルネットワーク出力を生成する段階と；

（ｎ）前記第二のニューラルネットワークによって、（ｍ）の前記第一のニューラルネットワーク出力および前記第一の識別データを受領する段階と；

（ｏ）前記第二のニューラルネットワークによって、（ｍ）の前記第一のニューラルネットワーク出力および前記第一の識別データに対応する合成されたオーディオ・データを生成する段階とを含む、
請求項３記載のオーディオ処理方法。

【請求項５】

合成されたオーディオ・データは、前記第一の人物の発話特性に従って前記第二の人物によって発声される単語に対応する、請求項４記載のオーディオ処理方法。

【請求項６】

前記トレーニングは、第一の言語で前記第一のオーディオ・データを受領することに関わり、前記合成されたオーディオ・データは、第二の言語で前記第二の人物によって発声された単語に対応する、請求項５記載のオーディオ処理方法。

【請求項７】

一つまたは複数のトランスデューサに、前記合成されたオーディオ・データを再生させることをさらに含む、請求項４ないし６のうちいずれか一項記載のオーディオ処理方法。

【請求項８】
前記トレーニングは：
第三のニューラルネットワークによって、前記第一のオーディオ・データを受領する段階と；
前記第一の人物の発話に対応する第一の発話特性を決定し、エンコードされたオーディオ・データを出力するよう前記第三のニューラルネットワークをトレーニングする段階とをさらに含む、
請求項４ないし７のうちいずれか一項記載のオーディオ処理方法。
【請求項９】
前記トレーニングは、前記エンコードされたオーディオ・データが前記第一の人物の発話に対応するかどうかを判定するよう第四のニューラルネットワークをトレーニングする段階をさらに含む、請求項８記載のオーディオ処理方法。
【請求項１０】
前記発話生成は：
前記第三のニューラルネットワークによって、前記第二のオーディオ・データを受領する段階と；
前記第三のニューラルネットワークによって、前記第二のオーディオ・データに対応する第二のエンコードされたオーディオ・データを生成する段階と；
前記第四のニューラルネットワークによって、前記第二のエンコードされたオーディオ・データを受領する段階と；
前記第四のニューラルネットワークが、修正された第二のエンコードされたオーディオ・データが前記第一の人物の発話に対応すると判定するまで、対話的プロセスを介して、修正された第二のエンコードされたオーディオ・データを生成し、前記第四のニューラルネットワークが、修正された第二のエンコードされたオーディオ・データが前記第一の人物の発話に対応すると判定した後、該修正された第二のエンコードされたオーディオ・データを前記第二のニューラルネットワークに提供する段階とをさらに含む、
請求項９記載のオーディオ処理方法。
【請求項１１】
（ａ）ないし（ｈ）を繰り返すことは、前記第一のニューラルネットワークまたは前記第二のニューラルネットワークの少なくとも一方を、現在の損失関数値に基づく逆方向伝搬を介してトレーニングすることに関わる、請求項１ないし１０のうちいずれか一項記載のオーディオ処理方法。
【請求項１２】
前記第一のニューラルネットワークは、双方向再帰型ニューラルネットワークを含む、請求項１ないし１１のうちいずれか一項記載のオーディオ処理方法。
【請求項１３】
インターフェース・システムと；一つまたは複数のプロセッサおよび該一つまたは複数のプロセッサに動作上結合された一つまたは複数の非一時的記憶媒体を有する制御システムとを有する発話合成装置であって、前記制御システムは、発話合成器を実装するよう構成されており、前記発話合成器は、内容抽出器、第一のニューラルネットワークおよび第二のニューラルネットワークを含み、前記第一のニューラルネットワークは双方向再帰型ニューラルネットワークを含み、前記第二のニューラルネットワークは階層構成をなすモジュールを含み、各モジュールは異なる時間分解能で動作し、前記第一のニューラルネットワークおよび前記第二のニューラルネットワークは：
（ａ）前記インターフェース・システムを介して前記内容抽出器によって、第一の人物の第一の発話に対応する第一のオーディオ・データを受領する段階と；
（ｂ）前記内容抽出器によって、前記第一の発話に対応する第一のタイムスタンプ付けされた音素シーケンスおよび第一のピッチ輪郭データを生成する段階と；
（ｃ）前記第一のニューラルネットワークによって、前記第一のタイムスタンプ付けされた音素シーケンスおよび前記第一のピッチ輪郭データを受領する段階と；

（ｄ）前記第一のニューラルネットワークによって、前記第一のタイムスタンプ付けされた音素シーケンスおよび前記第一のピッチ輪郭データに対応する第一のニューラルネットワーク出力を生成する段階であって、前記第一のニューラルネットワーク出力は複数のフレーム・サイズを含み、各フレーム・サイズは前記第二のニューラルネットワークのあるモジュールの時間分解能に対応する、段階と；
（ｅ）前記第二のニューラルネットワークによって、前記第一のニューラルネットワーク出力を受領する段階と；
（ｆ）前記第二のニューラルネットワークによって、第一の予測されたオーディオ信号を生成する段階と；
（ｇ）前記第一の予測されたオーディオ信号を第一の試験データと比較する段階であって、前記試験データは前記第一の人物の発話に対応するオーディオ・データである、段階と；
（ｈ）前記第一の予測されたオーディオ信号についての損失関数値を決定する段階と；
（ｉ）前記第一の予測されたオーディオ信号についての現在の損失関数値と前記第一の予測されたオーディオ信号についての以前の損失関数値との間の差が所定の値以下になるまで（ａ）ないし（ｈ）を繰り返す段階とを含むプロセスに従ってトレーニングされており、
前記制御システムは、発話生成のために前記発話合成器モジュールを制御するよう構成されており、前記発話生成は：
（ｊ）前記内容抽出器によって、前記インターフェース・システムを介して、第二の人物の第二の発話に対応する第二のオーディオ・データ、前記第二の発話に対応する第二のタイムスタンプ付けされたテキストおよび前記第一の人物の発話に対応する第一の識別データを受領する段階と；
（ｋ）前記内容抽出器によって、前記第二の発話に対応する第二のタイムスタンプ付けされた音素シーケンスおよび第二のピッチ輪郭データを生成する段階と；
（ｌ）前記第一のニューラルネットワークによって、（ｋ）の前記第二のタイムスタンプ付けされた音素シーケンスおよび前記第二のピッチ輪郭データを受領する段階と；
（ｍ）前記第一のニューラルネットワークによって、（ｋ）の前記第二のタイムスタンプ付けされた音素シーケンスおよび前記第二のピッチ輪郭データに対応する第一のニューラルネットワーク出力を生成する段階と；
（ｎ）前記第二のニューラルネットワークによって、（ｍ）の前記第一のニューラルネットワーク出力および前記第一の識別データを受領する段階と；
（ｏ）前記第二のニューラルネットワークによって、（ｍ）の前記第一のニューラルネットワーク出力および前記第一の識別データに対応する合成されたオーディオ・データを生成する段階とを含む、
発話合成装置。
【請求項１４】
前記合成されたオーディオ・データは、前記第一の人物の発話特性に従って前記第二の人物によって発声される単語に対応する、請求項１３記載の発話合成装置。
【請求項１５】
前記トレーニングは、第一の言語で前記第一のオーディオ・データを受領することに関わり、合成されたオーディオ・データは、第二の言語で前記第二の人物によって発声された単語に対応する、請求項１４記載の発話合成装置。
【請求項１６】
前記制御システムは、一つまたは複数のトランスデューサに、第二の合成されたオーディオ・データを再生させるよう構成される、請求項１３ないし１５のうちいずれか一項記載の発話合成装置。
【請求項１７】
合成されたオーディオ・データを生成することは、前記第二のニューラルネットワークの少なくとも一つの重みに対応する少なくとも一つの非一時的記憶媒体位置の物理的状態

を変更することを含む、請求項１３ないし１６のうちいずれか一項記載の発話合成装置。
【請求項１８】
インターフェース・システムと；
一つまたは複数のプロセッサおよび該一つまたは複数のプロセッサに動作上結合された一つまたは複数の非一時的記憶媒体を有する制御システムとを有する発話合成装置であって、
前記制御システムは、発話合成器を実装するよう構成されており、前記発話合成器は、内容抽出器、第一のニューラルネットワークおよび第二のニューラルネットワークを含み、前記第一のニューラルネットワークは双方向再帰型ニューラルネットワークを含み、前記第二のニューラルネットワークは階層構成をなすモジュールを含み、各モジュールは異なる時間分解能で動作しし、前記第二のニューラルネットワークは、請求項１ないし１２のうちいずれか一項記載のオーディオ処理方法によって第一の話者の第一の発話に対応する第一の合成されたオーディオ・データを生成するようトレーニングされており、前記制御システムは：
（ａ）前記内容抽出器によって前記インターフェース・システムを介して、第二の人物の第二の発話に対応する第二のオーディオ・データ、前記第二の発話に対応する第二のタイムスタンプ付けされたテキストおよび前記第一の人物の発話に対応する第一の識別データを受領する段階と；
（ｂ）前記内容抽出器によって、前記第二の発話に対応する第二のタイムスタンプ付けされた音素シーケンスおよび第二のピッチ輪郭データを生成する段階と；
（ｃ）前記第一のニューラルネットワークによって、（ｂ）の前記第二のタイムスタンプ付けされた音素シーケンスおよび前記第二のピッチ輪郭データを受領する段階と；
（ｄ）前記第一のニューラルネットワークによって、（ｂ）の前記第二のタイムスタンプ付けされた音素シーケンスおよび前記第二のピッチ輪郭データに対応する第一のニューラルネットワーク出力を生成する段階であって、前記第一のニューラルネットワーク出力は複数のフレーム・サイズを含み、各フレーム・サイズは前記第二のニューラルネットワークのあるモジュールの時間分解能に対応する、段階と；
（ｅ）前記第二のニューラルネットワークによって、（ｄ）の前記第一のニューラルネットワーク出力および前記第一の識別データを受領する段階と；
（ｆ）前記第二のニューラルネットワークによって、（ｄ）の前記第一のニューラルネットワーク出力および前記第一の識別データに対応する第二の合成されたオーディオ・データを生成する段階とを実行するよう前記発話合成器を制御するよう構成されている、
発話合成装置。
【請求項１９】
前記第二の合成されたオーディオ・データは、前記第一の人物の発話特性に従って前記第二の人物によって発声される単語に対応する、請求項１８記載の発話合成装置。
【請求項２０】
前記トレーニングは、第一の言語で前記第一のオーディオ・データを受領することに関わり、前記第二の合成されたオーディオ・データは、第二の言語で前記第二の人物によって発声された単語に対応する、請求項１９記載の発話合成装置。
【請求項２１】
一つまたは複数のトランスデューサに、前記第二の合成されたオーディオ・データを再生させることをさらに含む、請求項１８ないし２０のうちいずれか一項記載の発話合成装置。
【請求項２２】
前記合成されたオーディオ・データを生成することは、前記第二のニューラルネットワークの少なくとも一つの重みに対応する少なくとも一つの非一時的記憶媒体位置の物理的状態を変更することを含む、請求項１８ないし２１のうちいずれか一項記載の発話合成装置。
【請求項２３】

インターフェース・システムと；一つまたは複数のプロセッサおよび該一つまたは複数のプロセッサに動作上結合された一つまたは複数の非一時的記憶媒体を有する制御システムとを有する発話合成装置であって、前記制御システムは、発話合成器を実装するよう構成されており、前記発話合成器は、内容抽出器、第一のニューラルネットワークおよび第二のニューラルネットワークを含み、前記第一のニューラルネットワークは双方向再帰型ニューラルネットワークを含み、前記第二のニューラルネットワークは階層構成をなすモジュールを含み、各モジュールは異なる時間分解能で動作し、前記第一のニューラルネットワークおよび前記第二のニューラルネットワークは：

（ａ）前記インターフェース・システムを介して前記内容抽出器によって、目標話者の第一の発話に対応する第一のオーディオ・データを受領する段階と；

（ｂ）前記内容抽出器によって、前記第一の発話に対応する第一のタイムスタンプ付けされた音素シーケンスおよび第一のピッチ輪郭データを生成する段階と；

（ｃ）前記第一のニューラルネットワークによって、前記第一のタイムスタンプ付けされた音素シーケンスおよび前記第一のピッチ輪郭データを受領する段階と；

（ｄ）前記第一のニューラルネットワークによって、前記第一のタイムスタンプ付けされた音素シーケンスおよび前記第一のピッチ輪郭データに対応する第一のニューラルネットワーク出力を生成する段階であって、前記第一のニューラルネットワーク出力は複数のフレーム・サイズを含み、各フレーム・サイズは前記第二のニューラルネットワークのあるモジュールの時間分解能に対応する、段階と；

（ｅ）前記第二のニューラルネットワークによって、前記第一のニューラルネットワーク出力を受領する段階と；

（ｆ）前記第二のニューラルネットワークによって、第一の予測されたオーディオ信号を生成する段階と；

（ｇ）前記第一の予測されたオーディオ信号を第一の試験データと比較する段階と；

（ｈ）前記第一の予測されたオーディオ信号についての損失関数値を決定する段階と；

（ｉ）前記第一の予測されたオーディオ信号についての現在の損失関数値と前記第一の予測されたオーディオ信号についての以前の損失関数値との間の差が所定の値以下になるまで（ａ）ないし（ｈ）を繰り返す段階とを含むプロセスに従ってトレーニングされており、

前記制御システムは、発話生成のために前記発話合成器モジュールを制御するよう構成されており、前記発話生成は：

（ｊ）前記内容抽出器によって、前記インターフェース・システムを介して、源話者の第二の発話に対応する第二のオーディオ・データ、前記第二の発話に対応する第二のタイムスタンプ付けされたテキストおよび前記目標話者の発話に対応する第一の識別データを受領する段階と；

（ｋ）前記内容抽出器によって、前記第二の発話に対応する第二のタイムスタンプ付けされた音素シーケンスおよび第二のピッチ輪郭データを生成する段階と；

（ｌ）前記第一のニューラルネットワークによって、（ｋ）の前記第二のタイムスタンプ付けされた音素シーケンスおよび前記第二のピッチ輪郭データを受領する段階と；

（ｍ）前記第一のニューラルネットワークによって、（ｋ）の前記第二のタイムスタンプ付けされた音素シーケンスおよび前記第二のピッチ輪郭データに対応する第一のニューラルネットワーク出力を生成する段階と；

（ｎ）前記第二のニューラルネットワークによって、（ｍ）の前記第一のニューラルネットワーク出力および前記第一の識別データを受領する段階と；

（ｏ）前記第二のニューラルネットワークによって、（ｍ）の前記第一のニューラルネットワーク出力および前記第一の識別データに対応する合成されたオーディオ・データを生成する段階とを含む、

発話合成装置。

【請求項２４】

前記合成されたオーディオ・データが、前記目標話者の発話特性に従って前記源話者に

よって発声された単語に対応する、請求項２３記載の発話合成装置。
【請求項２５】
前記目標話者および前記源話者が、異なる年齢における同じ人物である、請求項２３記載の発話合成装置。
【請求項２６】
前記目標話者の前記第一の発話は、第一の年齢におけるまたは該第一の年齢を含む年齢範囲の間の、ある人物の発話に対応し、前記源話者の前記第二の発話は、第二の年齢におけるその人物の発話に対応する、請求項２３記載の発話合成装置。
【請求項２７】
前記第一の年齢が前記第二の年齢より若い年齢である、請求項２６記載の発話合成装置。
【発明の詳細な説明】
【技術分野】
【０００１】
関連出願への相互参照
本願は2018年2月16日に出願された米国仮特許出願第62/710,501号および2019年1月28日に出願された第62/797,864号ならびに2019年2月28日に出願された欧州特許出願第18157080.5号の利益を主張するものである。各出願の内容は参照によってその全体において組み込まれる。
【０００２】
技術分野
本開示はオーディオ信号の処理に関する。特に、本開示は発話スタイル転移実装のためのオーディオ信号の処理に関する。
【背景技術】
【０００３】
人物Aの発話を人物Bのスタイルで現実的に提示することは困難である。人物Aと人物Bが異なる言語を話す場合に、困難はさらに増す。たとえば、英語の映画を標準中国語で吹き替える声優を考える。キャラクターAの声を当てる声優は、あたかもキャラクターAが標準中国語で話しているかのように発話を発生する必要がある。この場合、発話スタイル転移の目標は、標準中国語での声優の声を入力として使って、あたかもキャラクターAが流暢な標準中国語を話しているかのように聞こえる声を生成することであろう。非特許文献１は、声変換のための、深双方向長短期メモリ・ベースの再帰型ニューラルネットワーク（Deep Bidirectional Long Short-Term Memory based Recurrent Neural Networks (DBLSTM-RNNs)）の使用を記載している。
【非特許文献１】Lifa Sun et al.、“Voice Conversion Using Deep Bidirectional Long Short-Term Memory Based Recurrent Neural Networks”、2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 19 April 2015, pp.4869-4873
【発明の概要】
【発明が解決しようとする課題】
【０００４】
本稿ではさまざまなオーディオ処理方法が開示される。いくつかのそのような方法は、発話合成器をトレーニングすることに関わってもよい。いくつかの例では、方法がコンピュータ実装されてもよい。たとえば、方法は、少なくとも部分的には、一つまたは複数のプロセッサおよび一つまたは複数の非一時的記憶媒体を有する制御システムを介して実装されてもよい。いくつかのそのような例では、トレーニングは：（ａ）一つまたは複数のプロセッサおよび一つまたは複数の非一時的記憶媒体を有する制御システムを介して実装される内容抽出プロセスによって、第一の人物の第一の発話に対応する第一のオーディオ・データを受領する段階と；（ｂ）前記内容抽出プロセスによって、前記第一の発話に対応する第一のタイムスタンプ付けされた音素シーケンスおよび第一のピッチ輪郭データを

生成する段階と；（ｃ）前記制御システムを介して実装される第一のニューラルネットワークによって、前記第一のタイムスタンプ付けされた音素シーケンスおよび前記第一のピッチ輪郭データを受領する段階と；（ｄ）前記第一のニューラルネットワークによって、前記第一のタイムスタンプ付けされた音素シーケンスおよび前記第一のピッチ輪郭データに対応する第一のニューラルネットワーク出力を生成する段階であって、前記第一のニューラルネットワーク出力は複数のフレーム・サイズを含む、段階と；（ｅ）前記制御システムを介して実装される第二のニューラルネットワークによって、前記第一のニューラルネットワーク出力を受領する段階であって、前記第二のニューラルネットワークはモジュールの階層構成を含み、各モジュールは異なる時間分解能で動作する、段階と；（ｆ）前記第二のニューラルネットワークによって、第一の予測されたオーディオ信号を生成する段階と；（ｇ）前記制御システムを介して、前記第一の予測されたオーディオ信号を第一の試験データと比較する段階と；（ｈ）前記制御システムを介して、前記第一の予測されたオーディオ信号についての損失関数値を決定する段階と；（ｉ）前記第一の予測されたオーディオ信号についての現在の損失関数値と前記第一の予測されたオーディオ信号についての以前の損失関数値との間の差が所定の値以下になるまで（ａ）ないし（ｈ）を繰り返す段階とに関わってもよい。

【０００５】

いくつかの実装によれば、当該方法の少なくともいくつかの動作は、少なくとも一つの非一時的記憶媒体位置の物理的状態を変更することに関わってもよい。たとえば、（ｆ）を繰り返すことは、前記第二のニューラルネットワークの少なくとも一つの重みと対応する少なくとも一つの非一時的記憶媒体位置の物理的状態を変更することに関わってもよい。

【０００６】

いくつかの例では、（ａ）は、第一の人物の第一の発話に対応する第一のタイムスタンプ付けされたテキストを受領することに関わってもよい。代替的または追加的に、（ａ）は、第一の人物に対応する第一の識別データを受領することに関わってもよい。

【０００７】

いくつかの実装では、本方法は、発話合成器を発話生成のために制御することに関わってもよい。いくつかのそのような実施形態では、発話生成は：（ｊ）前記内容抽出プロセスによって、第二の人物の第二の発話に対応する第二のオーディオ・データ、前記第二の発話に対応する第二のタイムスタンプ付けされたテキストおよび第一の人物の発話に対応する第一の識別データを受領する段階と；（ｋ）前記内容抽出プロセスによって、前記第二の発話に対応する第二のタイムスタンプ付けされた音素シーケンスおよび第二のピッチ輪郭データを生成する段階と；（ｌ）前記第一のニューラルネットワークによって、（ｋ）の前記第二のタイムスタンプ付けされた音素シーケンスおよび前記第二のピッチ輪郭データを受領する段階と；（ｍ）前記第一のニューラルネットワークによって、（ｋ）の前記第二のタイムスタンプ付けされた音素シーケンスおよび前記第二のピッチ輪郭データに対応する第一のニューラルネットワーク出力を生成する段階と；（ｎ）前記第二のニューラルネットワークによって、（ｍ）の前記第一のニューラルネットワーク出力および前記第一の識別データを受領する段階と；（ｏ）前記第二のニューラルネットワークによって、（ｍ）の前記第一のニューラルネットワーク出力および前記第一の識別データに対応する合成されたオーディオ・データを生成する段階とに関わってもよい。いくつかの例によれば、前記第二のオーディオ・データは、第一の人物が異なる年齢であったときの第一の人物の発話に対応してもよい。いくつかのそのような例では、前記第二のオーディオ・データは、現時点または最近の時点において第一の人物から受領される発話であってもよい。第一の人物の前記第一の発話は、たとえば、第一の人物がもっと若かったときの第一の人物の発話に対応してもよい。

【０００８】

いくつかの例によれば、合成されたオーディオ・データは、第一の人物の発話特性に従って第二の人物によって発音される単語に対応してもよい。いくつかのそのような例では

、トレーニングは、第一の言語で前記第一のオーディオ・データを受領することに関わってもよく、合成されたオーディオ・データは、第二の言語で第二の人物によって発音された単語に対応してもよい。しかしながら、いくつかの代替例では、合成されたオーディオ・データは、第一の年齢の第一の人物の発話特性に従って第二の年齢で、または第一の人物が第一の年齢範囲内であった時の間に、第一の人物によって発音された単語に対応してもよい。第一の年齢は、たとえば、第二の年齢より若い年齢であってもよい。いくつかの例は、一つまたは複数のトランスデューサに、合成されたオーディオ・データを再生させることに関わってもよい。

【０００９】

いくつかの実装によれば、前記トレーニングは、第三のニューラルネットワークによって、前記第一のオーディオ・データを受領することに関わってもよい。いくつかのそのような実装では、トレーニングは、第一の人物の発話に対応する第一の発話特性を決定し、エンコードされたオーディオ・データを出力するよう第三のニューラルネットワークをトレーニングすることに関わってもよい。

【００１０】

いくつかの例では、前記トレーニングは、エンコードされたオーディオ・データが第一の人物の発話に対応するかどうかを判定するよう第四のニューラルネットワークをトレーニングすることに関わってもよい。いくつかのそのような例では、発話生成は：第三のニューラルネットワークによって、前記第二のオーディオ・データを受領する段階と；第三のニューラルネットワークによって、前記第二のオーディオ・データに対応する第二のエンコードされたオーディオ・データを生成する段階と；第四のニューラルネットワークによって、前記第二のエンコードされたオーディオ・データを受領する段階と；第四のニューラルネットワークが修正された第二のエンコードされたオーディオ・データが第一の人物の発話に対応すると判定するまで、対話的プロセスを介して、修正された第二のエンコードされたオーディオ・データを生成する段階とに関わってもよい。いくつかの実装によれば、第四のニューラルネットワークが、修正された第二のエンコードされたオーディオ・データが第一の人物の発話に対応すると判定した後、本方法は、修正された第二のエンコードされたオーディオ・データを第二のニューラルネットワークに提供することに関わってもよい。

【００１１】

いくつかの実装では、（ａ）ないし（ｈ）を繰り返すことは、第一のニューラルネットワークまたは第二のニューラルネットワークの少なくとも一方を、現在の損失関数値に基づく後方伝搬を介してトレーニングすることに関わってもよい。いくつかの例によれば、第一のニューラルネットワークは、双方向再帰型ニューラルネットワークを含んでいてもよい。

【００１２】

本稿に記載される方法の一部または全部は、一つまたは複数の非一時的媒体上に記憶される命令（たとえばソフトウェア）に従って一つまたは複数の装置によって実行されてもよい。そのような非一時的な媒体は、ランダムアクセスメモリ（RAM）デバイス、読み出し専用メモリ（ROM）デバイスなどを含むがそれに限られない、本稿に記載されるもののようなメモリ・デバイスを含んでいてもよい。よって、本開示に記載される主題のさまざまな革新的な側面は、ソフトウェアが記憶されている非一時的媒体において実装されることができる。ソフトウェアは、たとえば、オーディオ・データを処理するよう少なくとも一つの装置を制御するための命令を含んでいてもよい。ソフトウェアは、たとえば、本稿に開示されるもののような制御システムの一つまたは複数のコンポーネントによって実行可能であってもよい。ソフトウェアは、たとえば、本稿に開示される方法のうちの一つまたは複数を実行するための命令を含んでいてもよい。

【００１３】

本開示の少なくともいくつかの側面は、装置により実装されてもよい。たとえば、一つまたは複数のデバイスが、本稿に開示される方法を少なくとも部分的に実行するよう構成

されてもよい。いくつかの実装では、装置は、インターフェース・システムおよび制御システムを含んでいてもよい。インターフェース・システムは、一つまたは複数のネットワーク・インターフェース、前記制御システムとメモリ・システムとの間の一つまたは複数のインターフェース、前記制御システムと別のデバイスとの間の一つまたは複数のインターフェースおよび／または一つまたは複数の外部デバイス・インターフェースを含んでいてもよい。制御システムは、汎用単一チップもしくは複数チップ・プロセッサ、デジタル信号プロセッサ（DSP）、特定用途向け集積回路（ASIC）、フィールドプログラマブルゲートアレイ（FPGA）または他のプログラム可能な論理デバイス、離散的なゲートもしくはトランジスタ論理または離散的なハードウェア・コンポーネントのうちの少なくとも一つを含みうる。よって、いくつかの実装では、制御システムは、一つまたは複数のプロセッサと、前記一つまたは複数のプロセッサに動作上結合された一つまたは複数の非一時的記憶媒体を含んでいてもよい。

【００１４】

いくつかのそのような例によれば、本装置は、インターフェース・システムおよび制御システムを含んでいてもよい。制御システムは、たとえば、発話合成器を実装するよう構成されてもよい。いくつかの実装では、発話合成器は、内容抽出器、第一のニューラルネットワークおよび第二のニューラルネットワークを含んでいてもよい。第一のニューラルネットワークはたとえば、双方向再帰型ニューラルネットワークを含んでいてもよい。いくつかの実装によれば、第二のニューラルネットワークは、複数のモジュールを含んでいてもよく、該モジュールはいくつかの事例では階層構成をなすモジュールであってもよい。いくつかのそのような例では、各モジュールは、異なる時間分解能で動作してもよい。

【００１５】

いくつかのそのような例によれば、第一のニューラルネットワークおよび第二のニューラルネットワークは：（ａ）前記インターフェース・システムを介して前記内容抽出器によって、第一の人物の第一の発話に対応する第一のオーディオ・データを受領する段階と；（ｂ）前記内容抽出器によって、前記第一の発話に対応する第一のタイムスタンプ付けされた音素シーケンスおよび第一のピッチ輪郭データを生成する段階と；（ｃ）前記第一のニューラルネットワークによって、前記第一のタイムスタンプ付けされた音素シーケンスおよび前記第一のピッチ輪郭データを受領する段階と；（ｄ）前記第一のニューラルネットワークによって、前記第一のタイムスタンプ付けされた音素シーケンスおよび前記第一のピッチ輪郭データに対応する第一のニューラルネットワーク出力を生成する段階であって、前記第一のニューラルネットワーク出力は複数のフレーム・サイズを含み、各フレーム・サイズは前記第二のニューラルネットワークのモジュールのある時間分解能に対応する、段階と；（ｅ）前記第二のニューラルネットワークによって、前記第一のニューラルネットワーク出力を受領する段階と；（ｆ）前記第二のニューラルネットワークによって、第一の予測されたオーディオ信号を生成する段階と；（ｇ）前記第一の予測されたオーディオ信号を第一の試験データと比較する段階と；（ｈ）前記第一の予測されたオーディオ信号についての損失関数値を決定する段階と；（ｉ）前記第一の予測されたオーディオ信号についての現在の損失関数値と前記第一の予測されたオーディオ信号についての以前の損失関数値との間の差が所定の値以下になるまで（ａ）ないし（ｈ）を繰り返す段階とを含むプロセスに従ってトレーニングされていてもよい。

【００１６】

いくつかの実装では、前記制御システムは、発話生成のために前記発話合成器モジュールを制御するよう構成されてもよい。前記発話生成は、たとえば：（ｊ）前記内容抽出器によって、前記インターフェース・システムを介して、第二の人物の第二の発話に対応する第二のオーディオ・データ、前記第二の発話に対応する第二のタイムスタンプ付けされたテキストおよび第一の人物の発話に対応する第一の識別データを受領する段階と；（ｋ）前記内容抽出器によって、前記第二の発話に対応する第二のタイムスタンプ付けされた音素シーケンスおよび第二のピッチ輪郭データを生成する段階と；（ｌ）前記第一のニューラルネットワークによって、（ｋ）の前記第二のタイムスタンプ付けされた音素シーケ

ンスおよび前記第二のピッチ輪郭データを受領する段階と；（ｍ）前記第一のニューラルネットワークによって、（ｋ）の前記第二のタイムスタンプ付けされた音素シーケンスおよび前記第二のピッチ輪郭データに対応する第一のニューラルネットワーク出力を生成する段階と；（ｎ）前記第二のニューラルネットワークによって、（ｍ）の前記第一のニューラルネットワーク出力および前記第一の識別データを受領する段階と；（ｏ）前記第二のニューラルネットワークによって、（ｍ）の前記第一のニューラルネットワーク出力および前記第一の識別データに対応する合成されたオーディオ・データを生成する段階とに関わってもよい。いくつかの例によれば、前記第二のオーディオ・データは、第一の人物が異なる年齢または異なる年齢範囲内であったときの第一の人物の発話に対応してもよい。いくつかのそのような例では、前記第二のオーディオ・データは、現時点または最近の時点において第一の人物から受領される発話であってもよい。第一の人物の前記第一の発話は、たとえば、第一の人物がもっと若かったときの第一の人物の発話に対応してもよい。

【００１７】

いくつかの例によれば、合成されたオーディオ・データは、第一の人物の発話特性に従って第二の人物によって発音される単語に対応してもよい。いくつかのそのような実装では、トレーニングは、第一の言語で前記第一のオーディオ・データを受領することに関わってもよく、合成されたオーディオ・データは、第二の言語で第二の人物によって発音された単語に対応してもよい。しかしながら、いくつかの代替例では、合成されたオーディオ・データは、第一の年齢での第一の人物の発話特性に従って第二の年齢で、または第一の年齢を含む年齢範囲（たとえば21ないし25歳、26ないし30歳など）の間に、第一の人物によって発音された単語に対応してもよい。第一の年齢は、たとえば、第二の年齢より若い年齢であってもよい。いくつかの例は、制御システムは、一つまたは複数のトランスデューサに、第二の合成されたオーディオ・データを再生させるよう構成されてもよい。いくつかの実装によれば、合成されたオーディオ・データを生成することは、前記第二のニューラルネットワークの少なくとも一つの重みと対応する少なくとも一つの非一時的記憶媒体位置の物理的状態を変更することに関わってもよい。

【００１８】

本開示の少なくともいくつかの代替的な側面は、装置により実装されてもよい。いくつかのそのような例によれば、装置は、インターフェース・システムおよび制御システムを含んでいてもよい。制御システムは、たとえば、発話合成器を実装するよう構成されてもよい。いくつかの実装では、発話合成器は、内容抽出器、第一のニューラルネットワークおよび第二のニューラルネットワークを含んでいてもよい。第一のニューラルネットワークはたとえば、双方向再帰型ニューラルネットワークを含んでいてもよい。いくつかの実装によれば、第二のニューラルネットワークは、複数のモジュールを含んでいてもよく、該モジュールはいくつかの事例では階層構成をなすモジュールであってもよい。いくつかのそのような例では、各モジュールは、異なる時間分解能で動作してもよい。

【００１９】

いくつかのそのような例によれば、第二のニューラルネットワークは、第一の話者の第一の発話に対応する第一の合成されたオーディオ・データを生成するようトレーニングされてもよい。いくつかの実装では、制御システムは：（ａ）前記内容抽出器によって前記インターフェース・システムを介して、第二の人物の第二の発話に対応する第二のオーディオ・データ、前記第二の発話に対応する第二のタイムスタンプ付けされたテキストおよび第一の人物の発話に対応する第一の識別データを受領する段階と；（ｂ）前記内容抽出器によって、第二の発話に対応する第二のタイムスタンプ付けされた音素シーケンスおよび第二のピッチ輪郭データを生成する段階と；（ｃ）第一のニューラルネットワークによって、（ｂ）の前記第二のタイムスタンプ付けされた音素シーケンスおよび前記第二のピッチ輪郭データを受領する段階と；（ｄ）前記第一のニューラルネットワークによって、（ｂ）の前記第二のタイムスタンプ付けされた音素シーケンスおよび前記第二のピッチ輪郭データに対応する第一のニューラルネットワーク出力を生成する段階であって、前記第

一のニューラルネットワーク出力は複数のフレーム・サイズを含み、各フレーム・サイズは前記第二のニューラルネットワークのあるモジュールの時間分解能に対応する、段階と；（ｅ）前記第二のニューラルネットワークによって、（ｄ）の前記第一のニューラルネットワーク出力および前記第一の識別データを受領する段階と；（ｆ）前記第二のニューラルネットワークによって、（ｄ）の前記第一のニューラルネットワーク出力および前記第一の識別データに対応する第二の合成されたオーディオ・データを生成する段階とを実行するよう前記発話合成器を制御するよう構成されてもよい。

【００２０】

いくつかの実装では、前記第二の合成されたオーディオ・データは、第一の人物の発話特性に従って第二の人物によって発音される単語に対応してもよい。いくつかのそのような例では、トレーニングは、第一の言語で前記第一のオーディオ・データを受領することに関わってもよく、合成されたオーディオ・データは、第二の言語で第二の人物によって発音された単語に対応してもよい。いくつかのそのような例は、一つまたは複数のトランスデューサに、合成されたオーディオ・データを再生させることに関わってもよい。いくつかの例では、合成されたオーディオ・データを生成することは、前記第二のニューラルネットワークの少なくとも一つの重みと対応する少なくとも一つの非一時的記憶媒体位置の物理的状態を変更することに関わってもよい。

【００２１】

本明細書に記載される主題の一つまたは複数の実装の詳細は、付属の図面および下記の記述において記載される。他の特徴、側面および利点は、該記述、図面および請求項から明白になるであろう。下記の図面の相対的な寸法は同縮尺で描かれていないことがあることに注意されたい。さまざまな図面における同様の参照符号および記号は、一般に、同様の要素を示す。

【図面の簡単な説明】

【００２２】

【図１】本稿に開示されるいくつかの実装による、発話スタイル転移のための一つまたは複数のニューラルネットワークをトレーニングするプロセスを示す。

【００２３】

【図２】本稿に開示されるいくつかの実装による、発話スタイル転移のための一つまたは複数のトレーニングされたニューラルネットワークを使用するプロセスを示す。

【００２４】

【図３】本稿に開示される方法の少なくともいくつかを実行するよう構成されうる装置のコンポーネントの例を示すブロック図である。

【００２５】

【図４】一例による、発話合成器をトレーニングする方法の諸ブロックを概説する流れ図である。

【００２６】

【図５】いくつかの例による、発話合成器トレーニング・システムの諸ブロックを示す。

【００２７】

【図６】トレーニングされた発話合成器に合成されたオーディオ・データを生成させる例を示す。

【００２８】

【図７】双方向再帰型ニューラルネットワークの一例を示す。

【００２９】

【図８】声モデリング・ニューラルネットワークの例示的な諸ブロックおよび該声モデリング・ニューラルネットワークにトレーニングの間に提供されうる入力の例を示す。

【００３０】

【図９】声モデリング・ニューラルネットワークの例示的な諸ブロックおよび該声モデリング・ニューラルネットワークに発話生成プロセスの間に提供されうる入力の例を示す。

【００３１】

【図１０】オートエンコーダの諸ブロックの例を示す。
【００３２】
【図１１】オートエンコーダを含む発話合成器をトレーニングするプロセスのための例示的な諸ブロックを示す。
【００３３】
【図１２Ａ】話者分類器をトレーニングするプロセスの間に使用されうる諸ブロックの例を示す。
【００３４】
【図１２Ｂ】話者分類器の一例を示す。
【００３５】
【図１３】発話合成のために話者分類器およびオートエンコーダを使用することの例を与える。
【００３６】
【図１４】一例による、整形ニューラルネットワークおよび声モデリング・ニューラルネットワークの諸ブロックを示す。
【発明を実施するための形態】
【００３７】
下記の記述は、本開示のいくつかの革新的な側面およびこれらの革新的な側面が実装されうるコンテキストの例を記述する目的のためのある種の実装に向けられる。しかしながら、本稿の教示は、さまざまな異なる仕方で適用できる。さらに、記載される実施形態は、多様なハードウェア、ソフトウェア、ファームウェアなどで実装されうる。たとえば、本願の諸側面は、少なくとも部分的には、装置、二つ以上のデバイスを含むシステム、方法、コンピュータ・プログラム・プロダクトなどで具現されうる。よって、本願の諸側面は、ハードウェア実施形態、ソフトウェア実施形態（ファームウェア、常駐ソフトウェア、マイクロコードなどを含む）および／またはソフトウェアおよびハードウェアの両側面を組み合わせる実施形態の形を取りうる。そのような実施形態は、本稿において「回路」、「モジュール」または「エンジン」と称されうる。本願のいくつかの側面は、コンピュータ可読プログラムコードが具現されている一つまたは複数の非一時的媒体において具現されるコンピュータ・プログラム・プロダクトの形を取りうる。そのような非一時的媒体は、たとえば、ハードディスク、ランダムアクセスメモリ（RAM）、読み出し専用メモリ（ROM）、消去可能なプログラム可能な読み出し専用メモリ（EPROMまたはフラッシュメモリ）、ポータブルなコンパクトディスク読み出し専用メモリ（CD-ROM）、光記憶デバイス、磁気記憶デバイスまたは上記の任意の好適な組み合わせを含みうる。よって、本開示の教示は、図面に示されるおよび／または本稿に記載される実装に限定されることは意図されておらず、広い適用可能性をもつ。
【００３８】
発話スタイル転移（speech style transfer）は、時に、「声変成」または「声変換」と称され、よって、これらの用語は本稿では交換可能に使われることがある。図１は、本稿に開示されるいくつかの実装に従って、発話スタイル転移のために一つまたは複数のニューラルネットワークをトレーニングするプロセスを示している。図１に示される例では、トレーニング・プロセスは、内容抽出ブロックに入力される、ある人物（話者A）の発話に対応するオーディオ・データを提供することに関わる。話者Aは、いくつかの開示される例では、「目標話者（target speaker）」と称されてもよい。
【００３９】
この例によれば、内容抽出ブロックは、テキスト入力を要求しない。ニューラルネットワーク・ベースの音素分類器が、たとえば、入力発話に対応するタイムスタンプ付けされた音素シーケンスを得るために、内容抽出ブロックにおいてトレーニングされ、使用されてもよい。この例では、話者Aの発話に対応するオーディオ・データだけが、入力として与えられるが、代替例では、人物Aの発話に対応するテキストが、入力発話波形と一緒に入力されてもよい。いくつかのそのような実装によれば、テキストは、タイムスタンプと

一緒に与えられてもよい。

【００４０】

この例によれば、内容抽出ブロックの出力は、入力発話のピッチ輪郭（pitch contour）（これは本稿では時に、「ピッチ輪郭データ」または「イントネーション」と称されることがある）と、入力発話に対応するタイムスタンプ付けされた音素シーケンス（これは時に発話テンポと称されることがある）とを含むデータである。ピッチ輪郭データは、たとえば、入力発話のピッチを時間を追って追跡する関数または曲線でありうる。いくつかの実装によれば、ある特定の時点におけるピッチ輪郭データは、その時点における入力発話の基本周波数、入力発話の基本周波数の対数値、入力発話の基本周波数の正規化された対数値またはその時点における入力発話のピーク（最高エネルギー）周波数に対応してもよい。しかしながら、代替例では、ピッチ輪郭データは、所与の時点における複数のピッチに対応してもよい。

【００４１】

この例によれば、ピッチ輪郭データおよびタイムスタンプ付けされた音素シーケンスは、入力発話について使用されている単語（「トレーニング・データセット」）の同じ系列を生成するよう一つまたは複数のボーカル・モデル・ニューラルネットワークをトレーニングするために使用される。好適なニューラルネットワークのいくつかの例は、後述する。いくつかの実装によれば、トレーニング・プロセスは、二つ以上のニューラルネットワークをトレーニングすることに関わってもよい。いくつかのそのような実装によれば、トレーニング・プロセスは、一つまたは複数のニューラルネットワークの出力を第二のニューラルネットワークに提供することに関わってもよい。第二のニューラルネットワークは、ボーカル・モデル・ニューラルネットワークであってもよい。ニューラルネットワーク（単数または複数）がトレーニングされた後は、発話生成のために使用されうる。発話生成は、本稿では「発話合成」とも称される。

【００４２】

図２は、本稿に開示されるいくつかの実装による発話スタイル転移のために一つまたは複数のトレーニングされたニューラルネットワークを使うプロセスを示している。この例では、別の人物（話者B）の発話が、図１を参照して上述した同じ内容抽出ブロックに入力される。内容抽出ブロックによって出力される、第二の人物の発話の、ピッチ輪郭データおよびタイムスタンプ付けされた音素シーケンスは、話者Aの声のためにトレーニングされたボーカル・モデル・ニューラルネットワークに提供される。いくつかの代替例では、ボーカル・モデルニューラルネットワークは、第一の年齢の話者Aの声のためにトレーニングされている。別の人物の話者を内容抽出ブロックに提供する代わりに、いくつかの実装は、第二の年齢の話者Aの発話を提供することに関わる。第一の年齢は、たとえば、第二の年齢よりも若い年齢であってもよい。

【００４３】

しかしながら、話者Aの声のためにトレーニングされたボーカル・モデル・ニューラルネットワークは、話者Aに対応する、または話者Aの発話に対応する識別データ（これは単純な「ID」、またはより複雑な識別データでありうる）をも提供される。よって、この例によれば、ボーカル・モデル・ニューラルネットワークは、話者Bの単語を、話者Bの発話テンポおよびイントネーションをもって、話者Aの声で出力する。別の言い方をすると、この例では、ボーカル・モデル・ニューラルネットワーク出力の合成されたオーディオ・データは、ボーカル・モデル・ニューラルネットワークによって学習された話者Aの発話特性に従って話者Bによって発音された単語を含む。いくつかの代替例では、第一の年齢の話者Aの声のためにトレーニングされたボーカル・モデル・ニューラルネットワークが、第一の年齢の話者Aに対応する識別データを提供される。第二の年齢における話者Aの入力単語は、第一の年齢の話者Aの声で出力されうる。第一の年齢は、たとえば、第二の年齢よりも若い年齢であってもよい。

【００４４】

いくつかの実装では、トレーニング・プロセスは、第一のオーディオ・データを第一の

言語で受領することに関わってもよく、合成されたオーディオ・データは、第二の言語で第二の人物によって発音される単語に対応してもよい。たとえば、話者Aが英語を話すことが知られており、話者Bが標準中国語を話すことが知られている場合、話者Aについての識別データは英語（言語１）と関連付けられてもよく、話者Bについての識別データは標準中国語（言語２）と関連付けられることができる。生成フェーズでは、ボーカル・モデル・ニューラルネットワーク（単数または複数）は、人物Bからの発話を与えられるが、人物Aについての識別データを与えられる。結果は、人物Aのスタイルでの、言語２での発話である。

【００４５】

よって、いくつかの実装は、言語２（たとえば英語）に存在しない言語１（たとえば標準中国語）における音について、ボーカル・モデル・ニューラルネットワーク（単数または複数）をトレーニングすることに関わってもよい。いくつかのトレーニング実装は、英語について一つ、標準中国語について一つではなく、両言語についての合同音素集合（「上位集合」）を使用してもよい。いくつかの例は、通常は言語２で話す目標話者が、言語２に対応する音がない言語１の音素（たとえば標準中国語の音素）と対応する音を出すよう促されるトレーニング・プロセスに関わってもよい。目標話者は、たとえば、与えられる音をマイクロフォンに向かって繰り返してもよい。音は、目標話者の声を入力するために使われるマイクロフォンに拾われないよう、ヘッドフォンを介して与えられる。

【００４６】

いくつかの代替例によれば、トレーニング・プロセスの間、目標話者は、自分の母語（言語１）の音素のみを含む音を与えてもよい。いくつかのそのような例では、発話生成プロセスは、言語２の発話を生成するとき、言語１からの最も類似した音素を使うことに関わってもよい。

【００４７】

いくつかの代替的な実装では、トレーニング・プロセスは、原子的発声の上位集合の表現を生成することに関わってもよい。原子的発声は、人間が出すことのできる、音素および非発話発声の両方を含み、摩擦音および声門音を含む。原子的発声は、発声の基本単位と見なす者もいる。

【００４８】

満足のいく発話スタイル転移方法およびデバイスを開発しようとさまざまな以前の試みがなされてきた。たとえば、「SampleRNN」は、モントリオール学習プログラム協会（Montreal Institute for Learning Algorithms、MILA）によって開発されたエンドツーエンドのニューラル・オーディオ生成モデルであり、これは一時に一つのオーディオ・サンプルを生成する（非特許文献２参照）。このモデルは、長い時間期間にわたる時間シーケンスにおける変動の根底にある源を捕捉するために、自己回帰型（autoregressive）深層ニューラルネットワークおよびステートフルなニューラルネットワークを階層構造において組み合わせる。

【非特許文献２】SAMPLERNN: An Unconditional End-To-End Neural Audio Generation Model、会議論文として公開、International Conference on Learning Representations, Toulon, France, April 24-26, 2017

【００４９】

SampleRNNは声変換分野における一歩前進であったが、SampleRNNはいくつかの欠点をもつ。たとえば、SampleRNNによって生成される発話信号は、もごもごした声であり、了解可能な発話ではない。さらに、俳優によって伝えられる感情は、自然に声にされることはできない：人は同じ感情を種々の仕方で声にできるので、テキストの感情を抽出する意味解析を使用しても、十分ではない。さらに、SampleRNNは、複数の目標話者を扱う機構を提供しない。

【００５０】

いくつかの開示される実装は、次の潜在的な技術的改善のうちの一つまたは複数を提供する。たとえば、さまざまな開示される実装に従って生成される発話信号は、もごもごし

た声ではなく、了解可能な発話である。本稿に開示されるいくつかの実装によれば、複数の話者の声スタイルが、トレーニング・プロセスを通じて学習されてもよい。いくつかの実施形態は、源話者と目標話者からのパラレル発話をトレーニングのために要求しない。いくつかの実装は、生成される発話信号の韻律を改善するために、声優の入力発話信号（または他の語り手の入力発話信号）の内容的意味および／または入力発話信号の特性を考慮に入れる。いくつかの実装は、より了解可能な、自然に聞こえる発話信号を生成するために、複数の目標話者の学習可能な高次元表現を提供する。

【００５１】

図３は、本稿に開示される方法の少なくともいくつかを実行するよう構成されうる装置のコンポーネントの例を示すブロック図である。いくつかの例では、装置３０５は、パーソナルコンピュータ、デスクトップコンピュータまたはオーディオ処理を提供するよう構成されている他のローカル装置であってもよく、またはそれを含んでいてもよい。いくつかの例では、装置３０５は、サーバーであってもよく、またはサーバーを含んでいてもよい。いくつかの例によれば、装置３０５は、ネットワーク・インターフェースを介したサーバーとの通信のために構成されたクライアント装置であってもよい。装置３０５のコンポーネントは、ハードウェアを介して、非一時的媒体に記憶されたソフトウェアを介して、ファームウェアを介しておよび／またはそれらの組み合わせによって、実装されてもよい。図３および本願で開示される他の図面に示されるコンポーネントの型および数は、単に例として示されている。代替的な実装は、より多数の、より少数のおよび／または異なるコンポーネントを含んでいてもよい。

【００５２】

この例では、装置３０５は、インターフェース・システム３１０および制御システム３１５を含む。インターフェース・システム３１０は、一つまたは複数のネットワーク・インターフェース、制御システム３１５とメモリ・システムとの間の一つまたは複数のインターフェースおよび／または一つまたは複数の外部装置インターフェース（一つまたは複数のユニバーサルシリアルバス（USB）インターフェースのような）を含んでいてもよい。いくつかの実装では、インターフェース・システム３１０は、ユーザー・インターフェース・システムを含んでいてもよい。ユーザー・インターフェース・システムは、ユーザーから入力を受け取るよう構成されてもよい。いくつかの実装では、ユーザー・インターフェース・システムは、ユーザーにフィードバックを提供するよう構成されてもよい。たとえば、ユーザー・インターフェース・システムは、対応するタッチおよび／またはジェスチャー検出システムとともに一つまたは複数のディスプレイを含んでいてもよい。いくつかの例では、ユーザー・インターフェース・システムは、一つまたは複数のマイクロフォンおよび／またはスピーカーを含んでいてもよい。いくつかの例によれば、ユーザー・インターフェース・システムは、モーター、バイブレーターなどといった、触覚フィードバックを提供する装置を含んでいてもよい。制御システム３１５は、たとえば、汎用単一チップもしくは複数チップ・プロセッサ、デジタル信号プロセッサ（DSP）、特定用途向け集積回路（ASIC）、フィールドプログラマブルゲートアレイ（FPGA）または他のプログラム可能な論理デバイス、離散的なゲートもしくはトランジスタ論理および／または離散的なハードウェア・コンポーネントを含みうる。

【００５３】

いくつか例では、装置３０５は、単一の装置において実装されてもよい。しかしながら、いくつかの実装では、装置３０５は、二つ以上の装置において実装されてもよい。いくつかのそのような実装では、制御システム３１５の機能は、二つ以上の装置に含められてもよい。いくつかの例では、装置３０５は別の装置のコンポーネントであってもよい。

【００５４】

図４は、一例による、発話合成器をトレーニングする方法の諸ブロックを概説する流れ図である。方法は、いくつかの事例では、図３の装置または別の型の装置によって実行されてもよい。いくつかの例では、方法４００のブロックは、一つまたは複数の非一時的媒体に記憶されたソフトウェアを介して実装されてもよい。方法４００のブロックは、本稿

に記載される他の方法と同様に、必ずしも示される順序で実行されない。さらに、そのような方法は、図示および／または記載されたものより多数または少数のブロックを含んでいてもよい。

【００５５】

ここで、ブロック４０５は、第一の人物の発話に対応するデータを受領することに関わる。この例では、「第一の人物」は、目標話者であり、実際にその人の声のために発話合成器がトレーニングされている第一の人物であってもなくてもよい。一貫した用語を維持するために、ブロック４０５で受領されるオーディオ・データは本稿では「第一のオーディオ・データ」と称されてもよく、第一の人物の発話は本稿では「第一の発話」と称されてもよい。

【００５６】

いくつかの例では、第一の人物からの第一のオーディオ・データのみが入力として提供されてもよい。そのような実装によれば、第一の発話からテキストが得られてもよい。しかしながら、そのような方法は、発話‐テキスト変換方法における潜在的な不正確さのために、最適ではないことがある。よって、代替的実装では、ブロック４０５は、第一の発話に対応するテキストを受領することに関わる。受領されるテキストは、第一のオーディオ・データの諸時刻に対応する諸タイムスタンプを含んでいてもよい。

【００５７】

いくつかの例では、ブロック４０５は、第一の人物に対応する識別データを受領することに関わってもよい。該識別データは、本稿では「第一の識別データ」と称されてもよい。いくつかの例によれば、第一の識別データは、本質的には単に「これは話者Aである」ということを示しうる。しかしながら、代替例では、第一の識別データは、第一の発話の一つまたは複数の属性に関する情報を含んでいてもよい。下記にいくつかの例を記載する。トレーニングのコンテキストでは、識別データを提供することは、一つまたは複数のニューラルネットワークが、複数の目標話者からの発話を用いてトレーニングされることを許容できる。代替的または追加的に、トレーニングのコンテキストにおいて識別データを提供することは、その人が異なる年齢の時の同じ人からの発話を区別できる。単純な例では、第一の年齢（または第一の年齢を含む年齢範囲）の話者Aの発話はA1と指定されてもよく、第二の年齢（または第二の年齢を含む年齢範囲）の話者Aの発話はA2と指定されてもよい。

【００５８】

この例によれば、ブロック４０５は、一つまたは複数のプロセッサおよび一つまたは複数の有体の記憶媒体を有する制御システムを介して実装される内容抽出プロセスによって、前記第一のオーディオ・データを受領することに関わる。内容抽出プロセスは、たとえば、一つまたは複数の非一時的記憶媒体に記憶されたソフトウェアに従って制御システムの一つまたは複数のプロセッサによって実装されてもよい。

【００５９】

この例では、内容抽出プロセスは、ブロック４１０において、第一の発話に対応するタイムスタンプ付けされた音素シーケンスおよびピッチ輪郭データを生成するよう構成される。一貫した用語を維持するために、ブロック４１０で生成されるタイムスタンプ付けされた音素シーケンスは、本稿では、「第一のタイムスタンプ付けされた音素シーケンス」と称されてもよく、ブロック４１０で生成されたピッチ輪郭データは本稿では「第一のピッチ輪郭データ」と称されてもよい。

【００６０】

この実装によれば、ブロック４０５は、内容抽出プロセスによって生成された第一のタイムスタンプ付けされた音素シーケンスおよび第一のピッチ輪郭データを、ニューラルネットワークによって、受領することに関わる。第一のニューラルネットワークは、本稿では「整形（conditioning）ネットワーク」と称されてもよい。いくつかの実施形態では、第一のニューラルネットワークは、第一のタイムスタンプ付けされた音素シーケンスおよび第一のピッチ輪郭データを、声モデリング・ニューラルネットワークに提供される前に

前処理するまたは整えることがあるからである。

【００６１】

ブロック４２０は、第一のニューラルネットワークがどのようにして第一のタイムスタンプ付けされた音素シーケンスおよび第一のピッチ輪郭データを整えうるかの一例を与えている。この例では、第一のニューラルネットワークは、第一のタイムスタンプ付けされた音素シーケンスおよび第一のピッチ輪郭データに対応する第一のニューラルネットワーク出力を生成するよう構成される。ここで、該第一のニューラルネットワーク出力は、二つ以上（「複数」）のフレーム・サイズで生成される。第一のニューラルネットワーク出力の諸フレーム・サイズは、いくつかの例では、第二のニューラルネットワークの好適な諸入力フレーム・サイズに対応しうる。

【００６２】

いくつかの例では、第一のニューラルネットワークは、第一のニューラルネットワーク出力を第二のニューラルネットワークに渡す前に、複数のフレームを処理してもよい。しかしながら、代替的実装では、第一のニューラルネットワークは、第一のニューラルネットワーク出力の複数のフレームが処理される間、第一のニューラルネットワーク出力を第二のニューラルネットワークに渡すのを遅らせなくてもよい。

【００６３】

よって、この例では、ブロック４２５は、制御システムを介して実装される第二のニューラルネットワークによって、第一のニューラルネットワーク出力を受領することに関わる。この例では、第二のニューラルネットワークは、階層構造をなすモジュールを含み、各モジュールは異なる時間分解能で動作する。第一のニューラルネットワーク出力のフレーム・サイズは、第二のニューラルネットワークのモジュールの時間分解能に対応する。いくつかの詳細な例を後述する。

【００６４】

この実装によれば、ブロック４３０は、第二のニューラルネットワークによって、第一の予測されたオーディオ信号を生成することに関わる。この例では、ブロック４３５は、制御システムを介して、第一の予測されたオーディオ信号を第一の試験データと比較することに関わり、ブロック４４０は、制御システムを介して、第一の予測されたオーディオ信号についての損失関数値を決定することに関わる。当業者によって理解されるであろうように、損失関数値は、第二のニューラルネットワークをトレーニングするために使用されうる。いくつかの実装によれば、損失関数値は、第一のニューラルネットワークをトレーニングするためにも使用されてもよい。

【００６５】

いくつかの例によれば、第二のニューラルネットワークをトレーニングすること（およびいくつかの事例では第一のニューラルネットワークをトレーニングすること）は、損失関数が比較的「平坦」になり、現在の損失関数値と前の損失関数値（たとえば直前の損失関数値）との間の差が閾値以下になるまで続けられてもよい。図４に示される例では、ブロック４４５は、第一の予測されたオーディオ信号についての現在の損失関数値と第一の予測されたオーディオ信号についての前の損失関数値との間の差が所定の値、たとえば1.90、1.92、1.94、1.96、1.98、2.00など以下になるまで、ブロック４０５ないし４４０のうちの少なくともいくつかを反復することに関わる。後述するように、いくつかのブロックを反復すること（たとえばブロック４２０の反復および／またはブロック４３０の反復）は、第二のニューラルネットワークの少なくとも一つの重みと対応する少なくとも一つの有体な記憶媒体位置の物理的な状態を変化させることに関わってもよい。

【００６６】

上記のように、識別データを提供することは、一つまたは複数のニューラルネットワークが複数の目標話者からの発話を用いてトレーニングされることを許容できる。図４を参照して上記したプロセスは、各目標話者について繰り返されてもよく、トレーニング結果が保存されて、各目標話者に対応する識別データと関連付けられてもよい。そのようなトレーニング結果は、たとえば、選択された目標話者に基づく発話生成のために使用されて

もよい。
【００６７】
図５は、いくつかの例による発話合成器トレーニング・システムのブロックを示している。システム５００は、いくつかの事例では、図３の装置または別の型の装置によって実行されてもよい。いくつかの例では、方法５００のブロックは、一つまたは複数の非一時的媒体に記憶されたソフトウェアを介して実装されてもよい。代替的実装は、図示および／または記載されたものより多数または少数のブロックを含んでいてもよい。
【００６８】
ここで、システム５００は、目標話者の発話に対応する入力オーディオ・データを内容抽出ブロック５０５に提供するよう構成される。目標話者は、いくつかの事例では、本稿の他所で言及される「第一の人物の第一の発話」に対応してもよい。いくつかの例では、入力オーディオ・データは、図３に示され上述したインターフェース・システム３１０のようなインターフェース・システムを介して提供されてもよい。いくつかの例では、入力オーディオ・データは、目標話者のその時の発話からのマイクロフォン信号を介して提供されてもよい。しかしながら、他の例では、入力オーディオ・データは、目標話者の前に記録された発話を介して提供されてもよい。いくつかのそのような例では、前に記録された発話は、入力オーディオ・データが内容抽出ブロック５０５に提供される時点より数年、あるいはさらには数十年前に記録されたものであってもよい。たとえば、前に記録された発話は、人の経歴のより早い段階の間に、たとえば俳優の経歴、音楽家の経歴、政治家の経歴などのより早い段階の特定の時間期間の間に記録されたものであってもよい。そのような以前に記録された発話はたとえば、映画またはテレビジョンのサウンドトラック、インタビュー録音などの一部に対応していてもよい。
【００６９】
この例では、内容抽出ブロック５０５は、音素シーケンス整列推定器ブロック５１０およびピッチ輪郭推定器ブロック５１５を含む。この例では、音素シーケンス整列推定器ブロック５１０は、入力オーディオ・データおよび該入力オーディオ・データに対応するテキストを受領する。受領されたテキストは、いくつかの例では、入力オーディオ・データの諸時刻に対応するタイムスタンプを含んでいてもよい。いくつかの例によれば、受領されたテキストは、インタビュー文字起こしテキスト、スクリプト・テキスト、文字起こしテキストなどを含みうる。この例によれば、ピッチ輪郭推定器ブロック５１５は、入力オーディオ・データを受領するが、入力オーディオ・データに対応するテキストは受領しない。いくつかの例では、内容抽出ブロック５０５は、目標話者に対応する識別データをも受領してもよい。
【００７０】
この例では、音素シーケンス整列推定器ブロック５１０は、入力オーディオ・データに対応するタイムスタンプ付けされた音素シーケンス５１２を生成するよう構成される。いくつかの実装では、音素シーケンス整列推定器ブロック５１０は、入力オーディオ・データをメル周波数ケプストラル係数（mel-frequency cepstral coefficient、MFCC）に変換してもよい。MFCCは、音の短期パワースペクトルの表現であり、周波数の非線形なメル・スケール上での対数パワースペクトルの線形コサイン変換に基づく。音素シーケンス整列推定器ブロック５１０は、辞書を参照してテキストを既知の音素に変換するよう構成されてもよい。音素シーケンス整列推定器ブロック５１０は、MFCC特徴と音素との間の整列を取るよう構成されてもよい。音素シーケンス整列推定器ブロック５１０は、いくつかの例では、Kaldi発話認識システムのような発話認識システムに基づく強制される整列器であってもよい。しかしながら、他の実装は、他の型の発話認識システムを用いてもよい。
【００７１】
ここで、ピッチ輪郭推定器ブロック５１５は、入力オーディオ・データに対応するピッチ輪郭データ５１７を生成するよう構成される。いくつかの例では、ピッチ輪郭データ５１７は、各オーディオ・フレームについて、ピッチの正規化された対数値を示してもよい。いくつかの例では、オーディオ・フレームは5ミリ秒（ms）の継続時間であってもよい

。しかしながら、代替例は、4msのフレーム、6msのフレーム、8msのフレーム、10msのフレームなどといった、より小さな、またはより大きなオーディオ・フレームを実装してもよい。いくつかの例では、ピッチ輪郭推定器ブロック５１５は、ピッチ絶対値を表わす、たとえば200個の浮動小数点数をシーケンスで生成するよう構成されてもよい。ピッチ輪郭推定器ブロック５１５は、これらの浮動小数点数に対して対数演算を実行し、各目標話者について、結果として得られる値を正規化して、出力が絶対的なピッチ値（たとえば200.0Hz）ではなく、0.0のまわりの輪郭（たとえば0.5）となるようにするよう構成されてもよい。いくつかの例では、ピッチ輪郭推定器ブロック５１５は、発話時間特性を使用するよう構成されてもよい。いくつかの例によれば、ピッチ輪郭推定器ブロック５１５は、まず、異なるカットオフ周波数をもつ、いくつかの低域通過フィルタを使用してもよい。一例では、低域通過フィルタは、50Hzのカットオフ周波数を有していてもよく、よって、0から50Hzまでの間の信号を通過させる。他の実装は他のカットオフ周波数を有していてもよい。いくつかのそのような実装によれば、これらの低域通過フィルタのカットオフ周波数は、50Hzから500Hzの間であってもよい。フィルタ処理された信号が基本周波数のみからなる場合、ピッチ輪郭推定器ブロック５１５は、正弦波を形成しうる。基本周波数は、この正弦波の周期に基づいて得られてもよい。いくつかの事例では、ピッチ輪郭推定器ブロック５１５は、最良の基本周波数候補を選ぶために零交差およびピーク谷間隔を使用してもよい。いくつかの例では、ピッチ輪郭推定器ブロック５１５は、Githubで提供されているWorldピッチ推定器のようなピッチ推定器を含んでいてもよい。しかしながら、他の例では、ピッチ輪郭推定器ブロック５１５は、別の型のピッチ推定器を含んでいてもよい。

【００７２】

いくつかの例では、内容抽出ブロック５０５は、現在のオーディオ・フレームが有声であるか無声であるかを示す指標を出力するよう構成される。いくつかの実装によれば、内容抽出ブロック５０５は、現在の音素、最も最近の諸音素のうちの一つまたは複数および一つまたは複数の将来の音素（たとえば次の音素および可能性としてはさらなる音素）を含む複数の音素を出力するよう構成される。あるそのような例によれば、内容抽出ブロック５０５は、現在の音素、二つの前の音素および次の二つの音素を含む五つの音素を出力するよう構成される。

【００７３】

この実装によれば、タイムスタンプ付けされた音素シーケンス５１２およびピッチ輪郭データ５１７は、整形ニューラルネットワーク５２０によって受領される。整形ニューラルネットワーク５２０は、図４を参照して上述した「第一のニューラルネットワーク」の事例である。この例では、整形ニューラルネットワーク５２０は、タイムスタンプ付けされた音素シーケンス５１２およびピッチ輪郭データ５１７を、これらのデータが声モデリング・ニューラルネットワーク５２５に提供される前に、前処理するまたは整えるよう構成される。いくつかの例では、声モデリング・ニューラルネットワーク５２５は複数のモジュールを含んでいてもよく、各モジュールは異なるレートまたは時間分解能で動作する。整形ニューラルネットワーク５２０によって出力される諸フレーム・サイズは、声モデリング・ニューラルネットワーク５２５のモジュールの時間分解能に対応してもよい。

【００７４】

この例では、声モデリング・ニューラルネットワーク５２５は、予測されたオーディオ信号５３０を生成し、該予測されたオーディオ信号５３０を損失関数決定ブロック５３５に提供するよう構成される。ここで、損失関数決定ブロック５３５は、予測されたオーディオ信号を試験データ５４０と比較し、予測されたオーディオ信号についての損失関数値を決定するよう構成される。この例によれば、試験データ５４０は目標話者の発話に対応するオーディオ・データである。いくつかの例では、試験データ５４０は、内容抽出ブロック５０５に以前に提供されていない、目標話者の発話に対応するオーディオ・データを含む。この例では、損失関数値は、声モデリング・ニューラルネットワーク５２５をトレーニングするために使用される。いくつかの実装によれば、損失関数値は、整形ニューラ

ルネットワーク５２０をトレーニングするためにも使用されてもよい。
【００７５】
いくつかの例によれば、声モデリング・ニューラルネットワーク５２５をトレーニングすること（およびいくつかの事例では、整形ニューラルネットワーク５２０をトレーニングすること）は、損失関数が比較的「平坦」になり、現在の損失関数値と前の損失関数値との間の差が閾値以下になるまで続けられてもよい。のちに詳述するように、声モデリング・ニューラルネットワーク５２５をトレーニングすることは、声モデリング・ニューラルネットワーク５２５の重みおよび／または活性化関数と対応する有体な記憶媒体位置の物理的な状態を変化させることに関わってもよい。
【００７６】
図６は、トレーニングされた発話合成器に、合成されたオーディオ・データを生成させる例を示している。この例では、声モデリング・ニューラルネットワーク５２５は、目標話者の声に対応するオーディオ・データを合成するようすでにトレーニングされている。この例によれば、源話者（source　speaker）の発話に対応する入力オーディオ・データが内容抽出ブロック５０５に提供される。源話者の発話は、いくつかの事例では、本稿の他所で言及されている「第二の人物の第二の発話」に対応してもよい。しかしながら、他の例では、声モデリング・ニューラルネットワーク５２５は、すでに、話者が第一の年齢であったときまたは話者が第一の年齢を含む年齢範囲（たとえば21歳ないし25歳、26歳ないし30歳など）であった時間の間の目標話者の声に対応するオーディオ・データを合成するようトレーニングされていてもよい。いくつかのそのような例によれば、源話者の発話は、目標話者が異なる年齢であった時の間、たとえば目標話者がより高い年齢であったときの目標話者の発話に対応してもよい。
【００７７】
この例では、音素シーケンス整列推定器ブロック５１０は、入力オーディオ・データおよび該入力オーディオ・データに対応するテキストを受領し、ピッチ輪郭推定器５１５は入力オーディオ・データを受領するが、入力オーディオ・データに対応するテキストは受領しない。受領されたテキストは、入力オーディオ・データの諸時刻に対応するタイムスタンプを含んでいてもよい。入力オーディオ・データが今は源話者に対応しているが、いくつかの実装では、内容抽出ブロック５０５は、目標話者に対応する識別データを受領する。源話者および目標話者が異なる年齢の同じ人物であるいくつかの実装では、所望される年齢または年齢範囲の目標話者についての識別データが、ブロック５０５に提供されてもよい。一つの単純な例では、第一の年齢（または年齢範囲）での話者Aの発話がA1と表わされ、第二の年齢（または年齢範囲）での話者Aの発話がA2と表わされてもよい。よって、そのような例では、システム５００は、目標話者の発話特性に従って、源話者によって発声された単語に対応する合成されたオーディオ・データを生成する。
【００７８】
この例では、音素シーケンス整列推定器ブロック５１０は、源話者からの入力オーディオ・データに対応するタイムスタンプ付けされた音素シーケンス５１２を生成するよう構成される。ここで、ピッチ輪郭推定器ブロック５１５は、源話者からの入力オーディオ・データに対応するピッチ輪郭データ５１７を生成するよう構成される。
【００７９】
この実装によれば、整形ニューラルネットワーク５２０は、タイムスタンプ付けされた音素シーケンス５１２およびピッチ輪郭データ５１７を、これらのデータの整形されたバージョン（これが、本稿の他所で言及される「第一のニューラルネットワーク出力」の例である）が声モデリング・ニューラルネットワーク５２５に提供される前に前処理するまたは整えるように構成される。
【００８０】
この例では、源話者からの入力オーディオ・データに対応するタイムスタンプ付けされた音素シーケンス５１２およびピッチ輪郭データ５１７の整形されたバージョンを受領することに加えて、声モデリング・ニューラルネットワーク５２５は、目標話者に対応する

識別データをも受領する。よって、この例では、声モデリング・ニューラルネットワーク５２５は、第一のニューラルネットワーク出力および第一の識別データに対応する合成されたオーディオ・データを含む予測されたオーディオ信号５３０（「合成されたオーディオ・データ」ともいう）を生成するよう構成される。そのような事例では、合成されたオーディオ・データは、目標話者の発話特性に従って、源話者によって発声された単語に対応する。トレーニング・プロセスが第一の言語で目標話者に対応するオーディオ・データを受領することに関わるいくつかの例によれば、合成されたオーディオ・データは、第二の言語で源話者によって発声された単語に対応してもよい。

【００８１】

この例では、声モデリング・ニューラルネットワーク５２５は、目標話者の発話に対応する発話を合成するようすでにトレーニングされているが、いくつかの実装では、予測されたオーディオ信号５３０は、記憶のため、一つまたは複数のトランスデューサを介した再生のためなに出力される前に、評価され、洗練されてもよい。いくつかのそのような実装では、予測されるオーディオ信号５３０は損失関数決定ブロック５３５に提供されてもよい。損失関数決定ブロック５３５は、予測されたオーディオ信号５３０を試験データと比較してもよく、予測されたオーディオ信号５３０についての損失関数値を決定してもよい。損失関数値は、予測されたオーディオ信号５３０をさらに洗練するために使用されてもよい。いくつかの実装によれば、損失関数値は、整形ニューラルネットワーク５２０をトレーニングするために使用されてもよい。

【００８２】

いくつかの実装によれば、第一のニューラルネットワークおよび／または第二のニューラルネットワークは再帰型ニューラルネットワークであってもよい。当業者に知られているように、再帰型ニューラルネットワークは、個々のユニットまたは「ニューロン」の間の接続が有向サイクルをなすニューラルネットワークの類型である。この特徴のため、再帰型ニューラルネットワークは、動的な時間的挙動を示す。フォードフォワード型ニューラルネットワークと異なり、再帰型ニューラルネットワークは、その内部メモリを、任意の入力シーケンスを処理するために使うことができる。この能力のため、再帰型ニューラルネットワークは、手書き認識または発話認識といったタスクに適用可能になる。

【００８３】

基本的な再帰型ニューラルネットワークは、しばしば「ニューロン」と称されるネットワーク・ノードを含む。各ニューロンは、他のニューロンへの有向の（一方向の）接続をもつ。各ニューロンは、入力または入力の集合を与えられて、そのニューロンの出力を定義する、一般に「活性化」と称される、時間変化する、実数値の活性化関数をもつ。ニューロン間の各接続（「シナプス」とも称される）は、修正可能な実数値の重みをもつ。ニューロンは、入力ニューロン（ネットワーク外部からデータを受領する）、出力ニューロンまたは入力ニューロンから出力ニューロンへの途上でデータを修正する隠れニューロンでありうる。いくつかの再帰型ニューラルネットワークは、入力ニューロンの層と出力ニューロンの層の間に、隠れニューロンのいくつかの層を含んでいてもよい。

【００８４】

ニューラルネットワークは、図３を参照して上記した制御システム３１５のような制御システムによって実装されてもよい。よって、第一のニューラルネットワークまたは第二のニューラルネットワークが再帰型ニューラルネットワークである実装については、第一のニューラルネットワークまたは第二のニューラルネットワークをトレーニングすることは、再帰型ニューラルネットワークにおける重みに対応する非一時的記憶媒体位置の物理的な状態を変更することに関わってもよい。前記記憶媒体位置は、制御システムによってアクセス可能なまたは制御システムの一部である一つまたは複数の記憶媒体の一部でありうる。上記の重みは、ニューロン間の接続に対応する。第一のニューラルネットワークまたは第二のニューラルネットワークをトレーニングすることも、ニューロンの活性化関数の値に対応する非一時的記憶媒体位置の物理的状態を変更することに関わってもよい。

【００８５】

第一のニューラルネットワークは、いくつかの例では、双方向再帰型ニューラルネットワークであってもよい。標準的な再帰型ニューラルネットワークでは、将来の時刻に対応する入力は、現在の状態から到達できない。対照的に、双方向再帰型ニューラルネットワークは、その入力データが固定されることを要求しない。さらに、双方向再帰型ニューラルネットワークの将来の入力情報は、現在の状態から到達可能である。双方向再帰型ニューラルネットワークの基本的なプロセスは、反対の時間的方向に対応する二つの隠れ層を同じ入力および出力に接続することである。この型の構造を実装することにより、双方向再帰型ニューラルネットワークの出力層におけるニューロンは、過去および将来の状態から情報を受領できる。双方向再帰型ニューラルネットワークは、入力のコンテキストが必要とされるときに特に有用である。たとえば、手書き認識アプリケーションでは、現在の文字の前後の文字の知識によって、パフォーマンスが向上されうる。

【００８６】

図７は、双方向再帰型ニューラルネットワークの一例を示している。図７に示される層の数、各層におけるニューロンの数などは、単に例である。他の実装は、より多数またはより少数の層、各層内のニューロンなどを含んでいてもよい。

【００８７】

図７では、ニューロン７０１は丸で表わされている。層７０５の「x」ニューロンは入力ニューロンであり、双方向再帰型ニューラルネットワーク７００の外部からデータを受領するよう構成される。層７３０の「y」ニューロンは出力ニューロンであり、双方向再帰型ニューラルネットワーク７００からデータを出力するよう構成される。層７１０～７２５におけるニューロンは、入力ニューロンから出力ニューロンへの途上でデータを修正する隠れニューロンである。いくつかの実装では、双方向再帰型ニューラルネットワーク７００のニューロンは、シグモイド活性化関数、tanh活性化関数またはシグモイドおよびtanh活性化関数の両方を用いてもよい。四つの隠れ層が図７に示されているが、いくつかの実装は、より多数またはより少数の隠れ層を含んでいてもよい。いくつかの実装は、ずっと多くの隠れ層、たとえば数百または数千の隠れ層を含んでいてもよい。たとえば、いくつかの実装は、128、256、512、1024、2048またはより多くの隠れ層を含んでいてもよい。

【００８８】

この例では、双方向再帰型ニューラルネットワーク７００は、三列のニューロンを含み、各列は異なる時間に対応する。それらの時間はたとえば、入力データが双方向再帰型ニューラルネットワーク７００に提供される時間区間に対応してもよい。中央の列７０４は、時間tに対応し、左の列７０２は時間t－1に対応し、右の列７０６は時間t＋1に対応する。時間t－1は、たとえば、時間tの直前の時間に取られたデータ・サンプルに対応してもよく、時間t＋1は、たとえば、時間tの直後の時間に取られたデータ・サンプルに対応してもよい。

【００８９】

この例では、隠れ層７１０および７１５は反対の時間方向に対応する。隠れ層７１０のニューロン７０１はデータを時間的に前方に渡す。一方、隠れ層７１５のニューロンはデータを時間的に逆方向に渡す。しかしながら、隠れ層７１０は隠れ層７１５に入力を与えず、隠れ層７１５は隠れ層７１０に入力を与えない。

【００９０】

特定の時間に対応する層７０５の入力ニューロン――たとえば時間tに対応する列７０４の入力ニューロン――は、隠れ層７１０のニューロンおよび隠れ層７１５のニューロンに情報を提供する。隠れ層７１０のニューロンおよび隠れ層７１５のニューロンは、同じ時間に対応する層７２０の単一のニュウーロンに情報を提供する。

【００９１】

ニューラルネットワークの情報および処理の流れは入力ニューロンから出力ニューロンに進むが、図７は、点線矢印７４０によって描かれる、逆方向の「逆方向伝搬」（「逆伝搬」としても知られる）をも示している。逆伝搬は、データのバッチが処理された後に各

ニューロンの誤差寄与を計算するためにニューラルネットワークにおいて使用される方法である。逆伝搬は、損失関数の勾配を計算することによって、ニューロンの重みを調整するよう勾配降下最適化アルゴリズムを適用することに関わってもよい。

【００９２】

図７に示した例では、層７３０の出力ニューロンは、損失関数決定ブロックに出力を提供してもよく、損失関数決定ブロックは層７３０の出力ニューロンに現在の損失関数値を提供してもよい。誤差が出力において計算され、点線矢印７４０によって示されるようにニューラルネットワークの諸層を通じて戻り方向に分配されるからである。

【００９３】

この例では、逆伝搬は双方向再帰型ニューラルネットワークのコンテキストで図示され、記述されたが、逆伝搬技法は、他の型の再帰型ニューラルネットワークを含むがそれに限られない他の型のニューラルネットワークに適用されてもよい。たとえば、逆伝搬技法は、本稿の他所で記載される声モデリング・ニューラルネットワーク（「第二のニューラルネットワーク」）、オートエンコーダおよび／または発話分類器ニューラルネットワークに適用されてもよい。

【００９４】

図８は、声モデリング・ニューラルネットワークの例示的なブロックおよびトレーニング中に声モデリング・ニューラルネットワークに与えられうる入力の例を示している。いくつかの例によれば、声モデリング・ニューラルネットワーク５２５のニューロンは、シグモイド活性化関数および／またはtanh活性化関数を用いてもよい。代替的または追加的に、声モデリング・ニューラルネットワーク５２５のニューロンは、整流線形ユニット（ReLU）活性化関数を用いてもよい。この例では、S_t、P_t、Q_tおよびF_tが声モデリング・ニューラルネットワーク５２５に提供される。ここで、S_tは目標話者識別データを表わし、P_tは目標話者について声モデリング・ニューラルネットワーク５２５によって生成された前の予測されたオーディオ信号を表わし、Q_tは、目標話者の声に対応する入力時間整列された音素シーケンスを表わし、F_tは、目標話者の声に対応する基本周波数輪郭データを表わす。

【００９５】

この例では、声モデリング・ニューラルネットワーク５２５は、モジュール８０５、８１０、８１５を含み、そのそれぞれは異なる時間分解能で動作する。この例では、モジュール８０５は、モジュール８１０よりもフレーム当たり、より多くのサンプルを処理し、モジュール８１５はモジュール８０５よりもフレーム当たり、より多くのサンプルを処理する。いくつかのそのような例では、モジュール８０５は、フレーム当たりモジュール８１０の10倍のサンプルを処理し、モジュール８１５は、フレーム当たりモジュール８１０の80倍のサンプルを処理する。いくつかのそのような実装によれば、声モデリング・ニューラルネットワーク５２５は、上記のような、SampleRNNニューラルネットワークの修正されたバージョンを含んでいてもよい。SampleRNNニューラルネットワークはたとえば、複数の目標話者についてトレーニングされ、該複数の目標話者のうちの選択された一人に対応する合成されたオーディオ・データを生成するよう修正されてもよい。しかしながら、これらは単に例である。他の実装では、声モデリング・ニューラルネットワーク５２５は、異なる数のモジュールを含んでいてもよく、および／またはモジュールは異なるフレーム・サイズを処理するよう構成されてもよい。

【００９６】

よって、異なるフレーム・サイズの入力データは、たとえば整形ニューラルネットワーク５２０（図８には示さず）によって声モデリング・ニューラルネットワーク５２５に入力されてもよい。一つのそのような例では、オーディオ・データは16kHzでサンプリングされ、よって、整形ニューラルネットワーク５２０は、モジュール８０５の各フレームについて、5msのオーディオ・データに相当する80サンプル（「大フレーム」サイズ）を提供してもよい。一つのそのような実装では、整形ニューラルネットワーク５２０は、モジュール８１０の各フレームについて、0.5msのオーディオ・データに相当する8サンプル（

「小フレーム」サイズ）を提供してもよい。一つのそのような実装では、整形ニューラルネットワーク５２０は、モジュール８１５の各フレームについて、40msのオーディオ・データに相当する640サンプル（「サンプル予測器」フレーム・サイズ）を提供してもよい。いくつかのそのような例では、モジュール８１０は、モジュール８０５の10倍の速さ、モジュール815の80倍の速さで動作してもよい。

【００９７】

いくつかの実装では、整形ニューラルネットワーク５２０は、モジュール８１０に提供される同じ8個のサンプルを10回繰り返して、モジュール805への入力のための80サンプルを生成してもよい。いくつかのそのような実装によれば、整形ニューラルネットワーク５２０は、モジュール８１０に提供される同じ8個のサンプルを80回繰り返して、モジュール815への入力のための640サンプルを生成してもよい。しかしながら、代替的な実装では、整形ニューラルネットワーク５２０は、他の方法に従って声モデリング・ニューラルネットワーク５２５に入力を提供してもよい。たとえば、入力オーディオ・データは、16kHz以外の周波数でサンプリングされてもよく、モジュール８０５、８１０、８１５は異なるフレーム・サイズに対して動作してもよい、などである。ある代替的な実装では、モジュール８０５は、フレーム当たり20サンプルを受領してもよく、モジュール８１５は以前の20サンプルを履歴として使ってもよい。モジュール８０５、８１０、８１５のいくつかの詳細な例は、図１４を参照して後述する。

【００９８】

この例では、C_tは、現在の目標話者について声モデリング・ニューラルネットワーク５２５によって出力される合成されたオーディオ・データを表わす。図８には示されていないが、多くの実装において、トレーニング・プロセスは、C_tを損失関数決定ブロックに提供し、損失関数決定ブロックから損失関数値を受領し、損失関数値に従って少なくとも声モデリング・ニューラルネットワーク５２５をトレーニングすることに関わる。いくつかの実装は、損失関数値に従って整形ニューラルネットワークをトレーニングすることにも関わる。

【００９９】

図９は、声モデリング・ニューラルネットワークの例示的なブロックおよび発話生成プロセスの間に声モデリング・ニューラルネットワーク提供されうる入力の例を示している。この例では、S_t、P_s→t、Q_s、F_sが声モデリング・ニューラルネットワーク５２５に提供される。ここで、S_tは目標話者識別データを表わし、P_s→tは声モデリング・ニューラルネットワーク５２５によって生成された、前の予測された（源から目標へのスタイル転移された（source-to-target　style-transferred））オーディオ信号を表わし、Q_sは源話者の声に対応する、入力の時間整列された音素シーケンスを表わし、F_sは源話者の声に対応する基本的な周波数輪郭データを表わす。

【０１００】

図８を参照して上記されるように、声モデリング・ニューラルネットワーク５２５は、この例において、モジュール８０５、８１０、８１５を含み、そのそれぞれは異なる時間分解能で動作する。よって、異なるフレーム・サイズの入力データが、たとえば整形ニューラルネットワーク５２０（図９には示さず）によって声モデリング・ニューラルネットワーク５２５に入力されてもよい。

【０１０１】

いくつかの実装は、一つまたは複数の追加的なニューラルネットワークに関わってもよい。いくつかのそのような実装では、第三のニューラルネットワークは、トレーニング・プロセス中に入力オーディオ・データ（たとえば図４を参照して上記した「第一の発話」）を受領してもよい。トレーニング・プロセスは、第一の人物の発話に対応する第一の発話特性を決定し、エンコードされたオーディオ・データを出力するよう第三のニューラルネットワークをトレーニングすることに関わってもよい。

【０１０２】

いくつかのそのような例によれば、第三のニューラルネットワークは、オートエンコー

ダであってもよく、あるいはオートエンコーダを含んでいてもよい。オートエンコーダは、効率的な符号化の教師なし学習のために使用されうるニューラルネットワークである。一般に、オートエンコーダの目標は、典型的には次元削減の目的のために、一組のデータについての表現または「エンコード」を学習することである。

【０１０３】

図１０は、オートエンコーダのブロックの例を示す。オートエンコーダ１００５は、たとえば、図３を参照して上記した制御システム３１５のような制御システムによって実装されてもよい。オートエンコーダ１００５はたとえば、一つまたは複数の非一時的記憶媒体に記憶されたソフトウェアに従って、制御システムの一つまたは複数のプロセッサによって実装されてもよい。図１０に示される要素の数および型は単に例である。オートエンコーダ１００５の他の実装は、より多数、より少数または異なる要素を含んでいてもよい。

【０１０４】

この例において、オートエンコーダ１００５は、三層のニューロンをもつ再帰型ニューラルネットワーク（recurrent neural network、RNN）を含む。いくつかの例によれば、オートエンコーダ１００５のニューロンは、シグモイド活性化関数および／またはtanh活性化関数を用いてもよい。RNN層１～３におけるニューロンは、N次元入力データを、そのN次元状態を維持しつつ処理する。層１０１０は、RUN層３の出力を受領して、プーリング・アルゴリズムを適用するよう構成される。プーリングは、非線形ダウンサンプリングの一つの形である。この例によれば、層１０１０は、RNN層３の出力をM個の重ならない部分または「部分領域」の集合に分割して、そのような各部分領域について最大値を出力するmaxプーリング関数を適用するよう構成される。

【０１０５】

図１１は、オートエンコーダを含む発話合成器をトレーニングするプロセスについての例示的なブロックを示している。この例では、トレーニング・プロセスの大半の側面が、図８を参照して上記したようにして実装されてもよい。S_t、F_t、Q_t、P_tは図８を参照して上記されている。

【０１０６】

しかしながら、この例では、オートエンコーダ１００５も、声モデリング・ニューラルネットワーク５２５に入力を提供する。この例によれば、オートエンコーダ１００５は、トレーニング・プロセス中に目標話者から入力オーディオ・データC_tを受領し、声モデリング・ニューラルネットワーク５２５にZ_tを出力する。この実装において、Z_tは、入力オーディオ・データC_tに比べて次元が削減されたオーディオ・データを出力する。

【０１０７】

この例において、C_t’は、現在の目標話者について声モデリング・ニューラルネットワーク５２５によって出力される合成されたオーディオ・データを表わす。この実装において、トレーニング・プロセスは、C_t’および「確固とした真実」オーディオ・データ――この例では入力オーディオ・データC_tである――を損失関数決定ブロック５３５に提供し、損失関数決定ブロックからの損失関数値を受領し、損失関数値に従って少なくとも声モデリング・ニューラルネットワーク５２５をトレーニングすることに関わる。この例は、損失関数値に従ってオートエンコーダ１００５をトレーニングすることにも関わる。無用な混雑を避けるため、図１０は、損失関数値が損失関数決定ブロック５３５によって声モデリング・ニューラルネットワーク５２５およびオートエンコーダ１００５に提供されることを示す矢印を含んでいない。いくつかの実装は、損失関数値に従って整形ニューラルネットワークをトレーニングすることにも関わる。すべての場合において、トレーニング・プロセスは、トレーニングされるニューラルネットワークの少なくとも一つの重みに対応する少なくとも一つの有体な記憶媒体位置の物理的な状態を変更することに関わる。

【０１０８】

いくつかのそのような例では、トレーニングは、第三のニューラルネットワークによっ

て生成されたエンコードされたオーディオ・データが第一の人物の発話に対応するかどうかを判定するよう第四のニューラルネットワークをトレーニングすることにも関わってもよい。第四のニューラルネットワークは、本稿では「話者素性分類器」または単に「話者分類器」と称されてもよい。いくつかのそのような実装では、発話生成プロセスは、第三のニューラルネットワークによって、源話者の発話に対応するオーディオ・データを受領することに関わってもよい。源話者は、いくつかの事例では、本稿の他所で（たとえば図６の記述において）言及される「第二の人物」に対応してもよい。したがって、受領されたオーディオ・データは、本稿の他所で言及される「第二の人物の第二の発話に対応する第二のオーディオ・データ」に対応してもよい。

【０１０９】

いくつかのそのような例では、発話生成プロセスは、第三のニューラルネットワークによって、第二のオーディオ・データに対応する第二のエンコードされたオーディオ・データを生成することに関わってもよい。発話生成プロセスは、第四のニューラルネットワークによって、第二のエンコードされたオーディオ・データを受領することに関わってもよい。いくつかの例では、発話生成プロセスは、第四のニューラルネットワークが修正された第二のエンコードされたオーディオ・データが第一の人物の発話に対応すると判定するまで、逐次反復プロセスを介して、修正された第二のエンコードされたオーディオ・データを生成し、第四のニューラルネットワークが修正された第二のエンコードされたオーディオ・データが第一の人物の発話に対応すると判定した後、修正された第二のエンコードされたオーディオ・データを第二のニューラルネットワークに（たとえば声モデリング・ニューラルネットワーク５２５）提供することに関わってもよい。

【０１１０】

図１２Aは、話者分類器をトレーニングするプロセスの間に使用されうるブロックの例を示している。この例では、話者分類器１２０５は、ニューラルネットワークの一つの型であり、オートエンコーダ１００５からの入力および損失関数決定ブロック５３５からのフィードバックに従ってトレーニングされる。話者分類器１２０５の、より詳細な例が、図１２Bに示されており、下記に記述される。この実装によれば、話者分類器１２０５がトレーニングされる時点において、オートエンコーダ１００５はすでにトレーニングされており、オートエンコーダ１００５の重みは固定されている。

【０１１１】

この例によれば、オートエンコーダ１００５は、トレーニング・プロセスの間に目標話者から入力オーディオ・データC_tを受領し、話者分類器１２０５にZ_tを出力する。この実装では、Z_tは、入力オーディオ・データC_tに比べて次元が削減されている、エンコードされたオーディオ・データを含む。

【０１１２】

この実装によれば、話者分類器１２０５は、目標話者についての予測された話者識別データであるS_t^を、損失関数決定ブロック５３５に対して出力する。この例において目標話者についての「確固とした真実」話者識別データS_tも損失関数決定ブロック５３５に入力される。

【０１１３】

「話者識別データ」に含まれるデータの型および量は、具体的な実装によって変わりうる。単純な場合には、話者識別データは、単に、特定の話者（たとえば「話者A」）が誰であるかを示してもよい。いくつかのそのような事例では、話者分類器１２０５は、単に、たとえば話者が話者Aであるか、話者Aでないかを判定するようトレーニングされてもよい。いくつかのそのような実装によれば、話者分類器１２０５は、たとえば、損失関数決定ブロック５３５がS_t^がS_tに一致すると判定するまでトレーニングされてもよい。

【０１１４】

しかしながら、いくつかの実装では、話者識別データは、より複雑であってもよい。いくつかのそのような実装によれば、話者識別データは、話者分類器１２０５によっておよび／またはオートエンコーダ１００５によって学習された目標話者の発話特性を示しうる

。いくつかのそのような実装では、話者識別データは、目標話者の発話特性を表わす多次元ベクトルであってもよい。いくつかの実装では、ベクトルの次元は8、16、32、64または128であってもよい。いくつかのそのような実装によれば、話者分類器１２０５は、損失関数決定ブロック５３５がS_t^とS_tの差が閾値以下であると判定するまで、トレーニングされてもよい。トレーニング・プロセスは、話者分類器１２０５の少なくとも一つの重みと対応する少なくとも一つの有体な記憶媒体位置の物理的な状態を変更することに関わる。

【０１１５】

図１２Bは、話者分類器の一例を示している。この例では、話者分類器１２０５は、畳み込みニューラルネットワークを含む。この例によれば、話者分類器１２０５は、オートエンコーダ１００５の出力を入力として受領し、この入力に基づいて話者分類を行なう。いくつかの例では、話者分類器１２０５への入力は、M×Nの特徴を含む。ここで、Nは入力フレーム数、Mは特徴次元である。

【０１１６】

この例において、畳み込み層１２１０は64個のフィルタを含む。しかしながら、他の実装では、畳み込み層１２１０は30フィルタ、40フィルタ、50フィルタ、60フィルタ、70フィルタ、80フィルタ、90フィルタ、100フィルタなど異なる数のフィルタを含んでいてもよい。ここで、各フィルタ・カーネルは、16×1のフィルタ・サイズをもつ。この例によれば、畳み込み層１２１０は、ステップ・サイズ、または「ストライド」が4の畳み込み演算を実行する。ストライドは、フィルタをスライドさせるときに何個の特徴を通過するかを示す。よって、入力データに沿ったスライド・フィルタは、畳み込み層１２１０がこの例において実行する畳み込み演算の型である。この例におけるM×N入力を与えられると、出力サイズはC1×floor((N－16)/4＋1)となる。ここで、floor(x)はi≦xとなるような最大の整数iを取る演算を表わす。

【０１１７】

この実装によれば、ニューラルネットワーク層１２１５は、畳み込み層１２１０からの出力を受領し、ReLU活性化関数を適用する。ここで、maxプール・ブロック１２２０はニューラルネットワーク層１２１５の出力に、maxプール演算を適用する。この例では、maxプール・ブロック１２２０は、8特徴毎から最大値を取ることによって、ニューラルネットワーク層１２１５の出力の次元を削減する。ここで、maxプール・ブロック１２２０は、8×1のカーネル・サイズをもち、ストライド8を適用する。

【０１１８】

この例において、畳み込み層１２２５は100個のフィルタを含む。しかしながら、他の実装では、畳み込み層１２２５は30フィルタ、40フィルタ、50フィルタ、60フィルタ、70フィルタ、80フィルタ、90フィルタ、110フィルタ、120フィルタ、130フィルタなど異なる数のフィルタを含んでいてもよい。ここで、各フィルタ・カーネルは、5×1のフィルタ・サイズをもつ。この例によれば、畳み込み層１２２５は、ステップ・サイズ、または「ストライド」が1の畳み込み演算を実行する。

【０１１９】

この実装によれば、ニューラルネットワーク層１２３０は、畳み込み層１２２５からの出力を受領し、ReLU活性化関数を適用する。ここで、maxプール・ブロック１２３５はニューラルネットワーク層１２３０の出力に、maxプール演算を適用する。この例では、maxプール・ブロック１２３５は、6特徴毎から最大値を取ることによって、ニューラルネットワーク層１２３０の出力の次元を削減する。ここで、maxプール・ブロック１２２０は、6×1のカーネル・サイズをもち、ストライド1を適用する。

【０１２０】

この例では、線形層１２４０は、maxプール・ブロック１２３５の出力を受領し、行列乗算を通じて線形変換を適用する。一つのそのような例によれば、線形層１２４０は

$$y = Ax + b \qquad (式1)$$

いよる線形変換を適用する。

【０１２１】
式1において、xは入力を表わし、Aは学習可能な重み行列を表わし、bは学習可能なバイアスを表わす。この実装によれば、層１２４５は線形層１２４０の出力にsoftmax関数を適用する。正規化指数関数としても知られるsoftmax関数は、任意の実数値のK次元ベクトルzを、合計すると1になる範囲[0,1]内の実数値のK次元ベクトルσ(z)に還元するロジスティック関数の一般化である。
【０１２２】
話者分類器１２０５の出力１２５０は、話者識別情報である。この例によれば、出力１２５０は、総数の話者素性クラスのうちの各話者素性クラスについて話者分類確率分布を含む。たとえば、出力１２５０は、話者が話者Aである確率P1、話者が話者Bである確率P2などを含んでいてもよい。
【０１２３】
図１３は、発話合成のために話者分類器およびオートエンコーダを使用することの例を与えている。このプロセスのいくつかの側面は、C_s、S_t、P_s→t、F_sが何を表わすかも含め、上記してあり（たとえば図９を参照して）、ここでは繰り返さない。この例では、C_sはオートエンコーダ１００５に提供される源話者についての入力オーディオ・データを表わす。Z_sは、話者分類器１２０５に入力されるオートエンコーダ１００５の出力を表わす。
【０１２４】
図１３に示される時点では、オートエンコーダ１００５および話者分類器１２０５は、トレーニング済みであり、その重みは記憶され、固定されている。この例によれば、損失関数決定ブロック５３５からのフィードバックはオートエンコーダ１００５または話者分類器１２０５の重みを変更するためには使用されず、その代わり、話者分類器１２０５が修正された源話者の発話を目標話者の発話であると分類するようになるまでZ_sの値を修正するために使われる。
【０１２５】
Z_sの値を修正するプロセスは、いくつかの実装では、確率的勾配降下プロセスのような逆最適化プロセスを含んでいてもよい。一例では、確率的勾配降下プロセスは、次のモデル関数F：

$$y=F(x,w) \qquad (式2)$$

に基づいていてもよい。
【０１２６】
式2において、Fは、パラメータw、入力xおよび出力yをもつモデル関数を表わす。式2に基づき、損失関数Lを：

$$Loss=L(F(x,w),Y) \qquad (式3)$$

のように構築してもよい。
【０１２７】
式3において、Yは確固とした真実のラベルを表わす。ニューラルネットワークをトレーニングする通常のプロセスにおいては、wの値を更新しようとする。しかしながら、この例では、ニューラルネットワークはトレーニング済みであり、wの値は固定されている。よって、この例では、逆最適化プロセスは、式3におけるLの値を最小化するためにxの値を更新または最適化することに関わってもよい。擬似コードでは、このプロセスの一例は、次のように記述できる。次のプロセスを、L(F(x,w),Y)が適切な最小に達するまで繰り返す：

$$For \quad i=1,2,\cdots n, \quad do \quad x=x-\eta \nabla Li(x) \qquad (式4)$$

【０１２８】
式4において、ηは学習率を表わす。この例において、xは、オートエンコーダ１００５から受領されたエンコードされたオーディオ・データを表わす。
【０１２９】
図１４は、一例による、整列ニューラルネットワークおよび声モデリング・ニューラル

ネットワークのブロックを示す。図１４のブロックは、たとえば、図３に示され、上記で述べた制御システム３１５のような制御システムを介して実装されてもよい。いくつかのそのような例では、図１４のブロックは、一つまたは複数の非一時的媒体に記憶されたソフトウェア命令に従って制御システムによって実施されてもよい。図１４に示されるフレーム・サイズ、フレーム・レート、要素数および要素の型は単に例である。

【０１３０】

この例によれば、整形ニューラルネットワーク５２０は、双方向RNN　１４１５を含む。この例では、整形ニューラルネットワーク５２０は、複数の話者（この例では109の話者）の発話特性データが記憶されているブロック１４０７をも含む。この例によれば、発話特性データは、各話者について整形ニューラルネットワーク５２０をトレーニングするプロセスの間に整形ニューラルネットワーク５２０によって学習された目標話者の発話特性に対応する。いくつかの実装では、ブロック１４０７は、発話特性データが記憶されるメモリ位置へのポインタを含んでいてもよい。この例では、発話特性データは32次元ベクトルによって表わされているが、他の例では、発話特性データは、他の次元のベクトルによってなど、他の仕方で表わされてもよい。

【０１３１】

この例では、ブロック１４０５は、たとえばユーザーから受領された入力に従って選択された特定の目標話者を表わす。ブロック１４０５からの話者識別データは、ブロック１４０７に提供され、ブロック１４０７は話者特性データ１４１０を連結ブロック１４１４に提供する。この実装では、連結ブロック１４１４は音素特徴１４１２をも（たとえば、上記の内容抽出ブロック５０５のような内容抽出ブロックから）受領する。

【０１３２】

この例では、連結ブロック１４１４は、発話特性データ１４１０を音素特徴１４１２と連結して、出力を、双方向RNN　１４１５に提供するよう構成される。双方向RNN　１４１５は、たとえば、（たとえば図７を参照して）上記したように機能するよう構成されてもよい。

【０１３３】

この例では、声モデリング・ニューラルネットワーク５２５は、モジュール８０５、８１０、８１５を含み、そのそれぞれは異なる時間分解能で動作する。この例では、モジュール８０５は、モジュール８１０よりもフレーム当たり、より多くのサンプルを処理し、モジュール８１５はモジュール８０５よりもフレーム当たり、より多くのサンプルを処理する。いくつかのそのような例では、モジュール８０５は、フレーム当たりモジュール８１０の10倍のサンプルを処理し、モジュール８１５は、フレーム当たりモジュール８１０の80倍のサンプルを処理する。いくつかのそのような実装によれば、声モデリング・ニューラルネットワーク５２５は、上記のような、SampleRNNニューラルネットワークの修正されたバージョンを含んでいてもよい。しかしながら、これらは単に例である。他の実装では、声モデリング・ニューラルネットワーク５２５は、異なる数のモジュールを含んでいてもよく、および／またはモジュールは異なるフレーム・サイズを処理するよう構成されてもよい。

【０１３４】

モジュール８０５、８１０、８１５がそれぞれ異なる時間分解能で動作するので、この例では、整形ニューラルネットワーク５２０は、異なるフレーム・サイズの出力を、モジュール８０５、８１０、８１５のそれぞれに提供する。この例によれば、整形ニューラルネットワーク５２０は、50フレームに対応する時間区間、たとえば整形ニューラルネットワーク５２０が50個のフレームを生成した時間区間の間に1024サンプルのサイズをもつ50フレームをモジュール８０５に提供する。ここで、整形ニューラルネットワーク５２０は、1024サンプルのサイズをもつ50フレームを反復テンソル・ブロック１４３０に提供し、を反復テンソル・ブロック１４３０が該50フレームを10回反復し、モジュール８１０に、1024サンプルのサイズをもつ500個のフレームを提供する。この例において、整形ニューラルネットワーク５２０は、1024サンプルのサイズをもつ50フレームを反復テンソル・ブ

ロック１４４５に提供し、を反復テンソル・ブロック１４４５が該50フレームを80回反復し、モジュール８１５に、1024サンプルのサイズをもつフレーム4000個を提供する。

【０１３５】

この例によれば、モジュール８０５は、4000個のオーディオ・サンプルをもつ単一のフレームをそれぞれ80サンプルをもつ50個のフレームに再編するよう構成された再編ブロック１４１８を含む。この例では、線形演算ブロック１４２０は、再編ブロック１４１８によって出力されたフレーム当たり80サンプルから、フレーム当たり1024個のサンプルが生成されるよう、各フレームに対して線形演算を実行するよう構成される。いくつかのそのような例では、線形演算ブロック１４２０は、次式：

Y＝X*W　　(式5)

に従った行列乗算を介して各フレームに対して線形演算を実行するよう構成される。

【０１３６】

式5において、Xは入力行列を表わし、それはこの例では50かける80の次元をもち、再編ブロック１４１８の出力に対応する。式5では、Yは次元50かける1024の出力行列を表わし、Wは次元80かける1024の行列を表わす。よって、線形演算ブロック１４２０の出力は、双方向RNN　１４１５の出力と同数のフレームおよび同じフレーム・サイズをもち、それにより、線形演算ブロック１４２０の出力および双方向RNN　１４１５の出力は、合計され、RNN　１４２２に提供される。

【０１３７】

線形演算ブロック１４２５は、フレーム毎の1024サンプルが、RNN　１４２２によって出力されたフレーム毎の1024サンプルから生成されるよう、各フレームに対して線形演算を実行するよう構成される。いくつかの例では、線形演算ブロック１４２５は、行列乗算を介して各フレームに対して線形演算を実行するよう構成される。再編ブロック１４２７は、この例では、50フレームの10240個のオーディオ・サンプルを、それぞれ1024サンプルをもつ500個のフレームに再編するよう構成される。よって、再編ブロック１４２７の出力は、反復テンソル・ブロック１４３０の出力と合計されうる。

【０１３８】

この例において、再編ブロック１４３２は、4000サンプルをもつ入力オーディオ・データの一つのフレームを、それぞれ8サンプルをもつ500個のフレームに再編するよう構成される。線形演算ブロック１４２５は、再編ブロック１４３２によって出力されるフレーム当たり8個のサンプルから1024個のサンプルが生成されるよう、各フレームに対して線形演算を実行するよう構成される。いくつかの例では、線形演算ブロック１４３５は、行列乗算を介して各フレームに対して線形演算を実行するよう構成される。これらの演算後、線形演算ブロック１４３５の出力は、再編ブロック１４２７の出力および反復テンソル・ブロック１４３０の出力と合計されうる。

【０１３９】

この合計は、この例ではRNN　１４３７に提供される。ここで、RNN　１４３７の出力は線形演算ブロック１４４０に提供され、線形演算ブロック１４４０はこの例では、RNN　１４３７によって出力されるフレーム当たり1024サンプルから、フレーム当たり8192サンプルが生成されるよう、各フレームに対して線形演算を実行するよう構成される。いくつかの例では、線形演算ブロック１４４０は、行列乗算を介して各フレームに対して線形演算を実行するよう構成される。この例では、再編ブロック１４４２は、それぞれ8192サンプルをもつデータの500フレームを、それぞれ1024サンプルをもつ4000フレームに再編するよう構成される。すると、再編ブロック１４４２の出力は、反復テンソル・ブロック１４４５の出力と同じ次元をもち、よって、反復テンソル・ブロック１４４５の出力と合計されうる。

【０１４０】

この例によれば、モジュール８１５は、最近傍ブロック１４４７を含む。最近傍ブロック１４４７は、前の7個のサンプルを、オーディオ・データの現在のサンプルと一緒に、線形演算ブロック１４５０に提供するよう構成される。この実装では、線形演算ブロック

１４５０は、最近傍ブロック１４４７によって出力されるフレーム当たり8個のサンプルからフレーム当たり1024個のサンプルが生成されるよう、各フレームに対して線形演算を実行するよう構成される。すると、線形演算ブロック１４５０の出力は、反復テンソル・ブロック１４４５の出力と同じ次元をもち、よって、反復テンソル・ブロック１４４５の出力および再編ブロック１４４２の出力と合計されうる。いくつかの代替的な実装では、ブロック１４４７および１４５０は、たとえば全1024個のフィルタをもち各フィルタのサイズが8かける1である単一の畳み込み層によって、置き換えられてもよい。8×1のフィルタ・サイズでは、畳み込みフィルタは前の7個のサンプルおよび現在のサンプルに対して作用することができる。

【０１４１】

結果として得られる合計は、線形演算ブロック１４５２に提供される。この例では、線形演算ブロック１４５２は、線形演算ブロック１４５７および線形演算ブロック１４６２をも含む多層パーセプトロンの一部であるよう構成される。線形演算ブロック１４５２の出力はReLUブロック１４５５に提供され、ReLUブロック１４５５はその出力を線形演算ブロック１４５７に提供する。線形演算ブロック１４５７の出力はReLUブロック１４６０に提供され、ReLUブロック１４６０はその出力を線形演算ブロック１４６２に提供する。この例において、線形演算ブロック１４６２は、ReLUブロック１４６０によって出力されたフレーム当たり1024サンプルからフレーム当たり256サンプルが生成されるよう、各フレームに対して線形演算を実行するよう構成される。この実装では、オーディオ・データは8ビット・オーディオ・データであり、よって、フレーム当たり256サンプルは、入力オーディオ・データの可能なオーディオ・サンプル値の数に対応する。

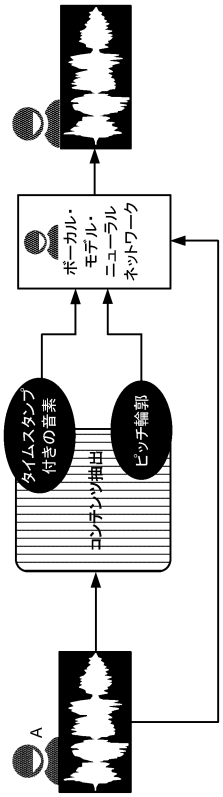
【０１４２】

この実装によれば、ブロック１４６５は線形演算ブロック１４６２の出力にsoftmax関数を適用する。この例では、softmax関数は、256個の値のそれぞれについての分類またはフレーム毎のクラスを提供する。この例では、声モデリング・ニューラルネットワーク５２５によって出力される出力データ１４７０は、オーディオ・サンプル分布を含み、それが、256個の値のそれぞれについての確率またはフレーム毎のクラスを示す。

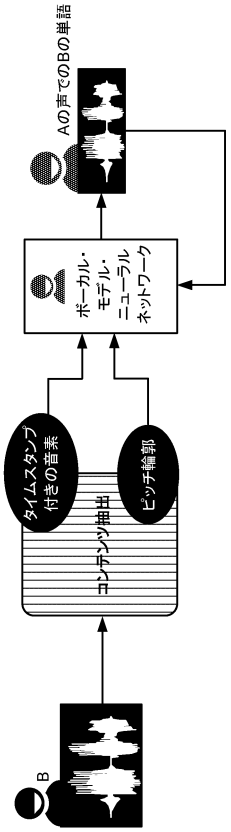
【０１４３】

本稿で定義される一般原理は、付属の請求項の範囲から外れることなく他の実装に適用されてもよい。よって、請求項は、本稿に示される実装に限定されることは意図されておらず、本開示および本願で開示される原理および新規な特徴と整合する最も広い範囲を与えられるべきである。

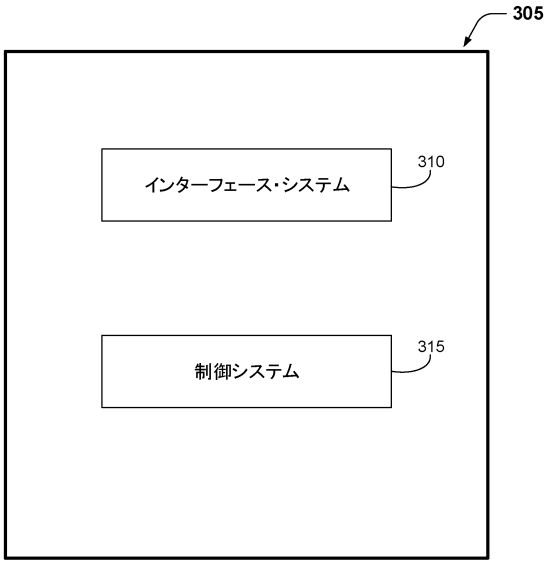
【図１】


ボーカル・モデル・ニューラルネットワーク
タイムスタンプ付きの音素
ピッチ輪郭
コンテンツ抽出
A

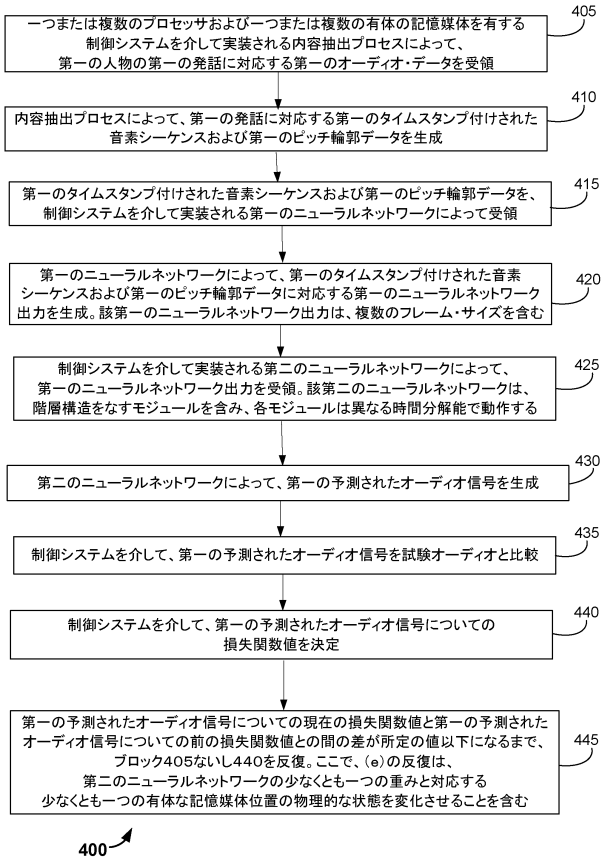
【図２】


Aの声でのBの単語
ボーカル・モデル・ニューラルネットワーク
タイムスタンプ付きの音素
ピッチ輪郭
コンテンツ抽出
B

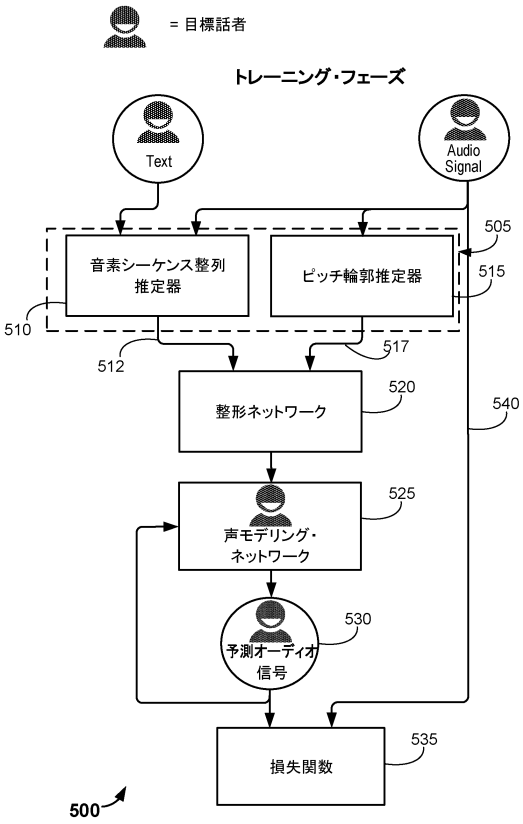
【図３】


305
インターフェース・システム
310
制御システム
315

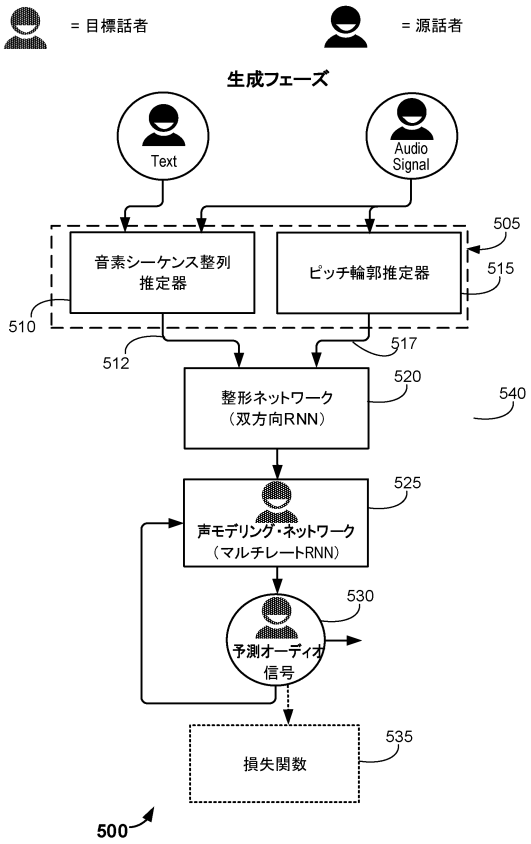
【図４】


一つまたは複数のプロセッサおよび一つまたは複数の有体の記憶媒体を有する制御システムを介して実装される内容抽出プロセスによって、第一の人物の第一の発話に対応する第一のオーディオ・データを受領
405
内容抽出プロセスによって、第一の発話に対応する第一のタイムスタンプ付けされた音素シーケンスおよび第一のピッチ輪郭データを生成
410
第一のタイムスタンプ付けされた音素シーケンスおよび第一のピッチ輪郭データを、制御システムを介して実装される第一のニューラルネットワークによって受領
415
第一のニューラルネットワークによって、第一のタイムスタンプ付けされた音素シーケンスおよび第一のピッチ輪郭データに対応する第一のニューラルネットワーク出力を生成。該第一のニューラルネットワーク出力は、複数のフレーム・サイズを含む
420
制御システムを介して実装される第二のニューラルネットワークによって、第一のニューラルネットワーク出力を受領。該第二のニューラルネットワークは、階層構造をなすモジュールを含み、各モジュールは異なる時間分解能で動作する
425
第二のニューラルネットワークによって、第一の予測されたオーディオ信号を生成
430
制御システムを介して、第一の予測されたオーディオ信号を試験オーディオと比較
435
制御システムを介して、第一の予測されたオーディオ信号についての損失関数値を決定
440
第一の予測されたオーディオ信号についての現在の損失関数値と第一の予測されたオーディオ信号についての前の損失関数値との間の差が所定の値以下になるまで、ブロック405ないし440を反復。ここで、(e)の反復は、第二のニューラルネットワークの少なくとも一つの重みと対応する少なくとも一つの有体な記憶媒体位置の物理的な状態を変化させることを含む
445
400

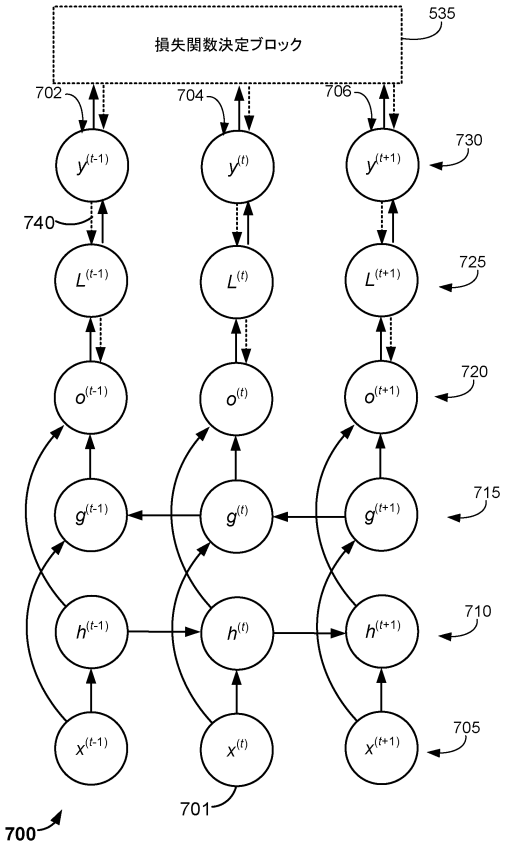
【図５】


= 目標話者
トレーニング・フェーズ
Text
Audio Signal
音素シーケンス整列推定器
ピッチ輪郭推定器
505
515
510
512
517
整形ネットワーク
520
540
声モデリング・ネットワーク
525
予測オーディオ信号
530
損失関数
535
500

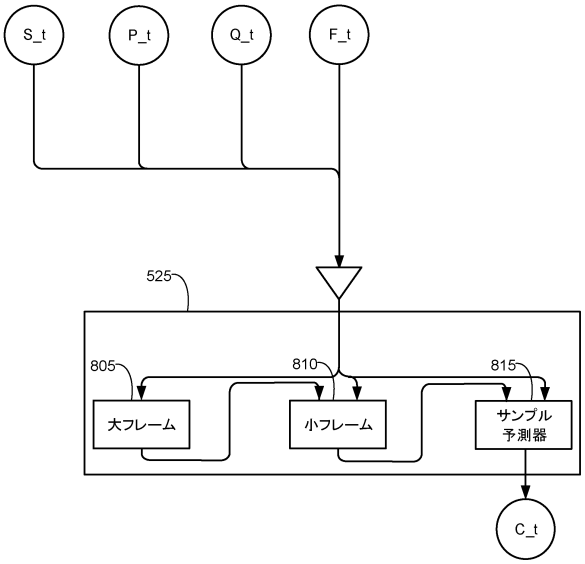
【図６】


= 目標話者
= 源話者
生成フェーズ
Text
Audio Signal
音素シーケンス整列推定器
ピッチ輪郭推定器
505
515
510
512
517
整形ネットワーク（双方向RNN）
520
540
声モデリング・ネットワーク（マルチレートRNN）
525
予測オーディオ信号
530
損失関数
535
500

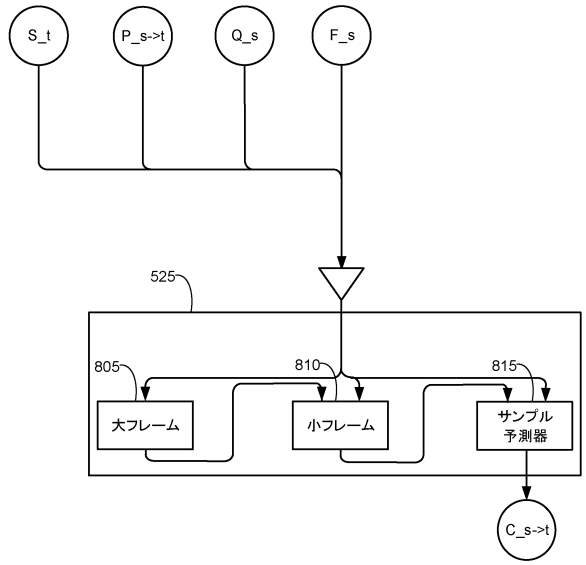
【図７】


損失関数決定ブロック
535
702
704
706
730
740
725
720
715
710
705
701
700

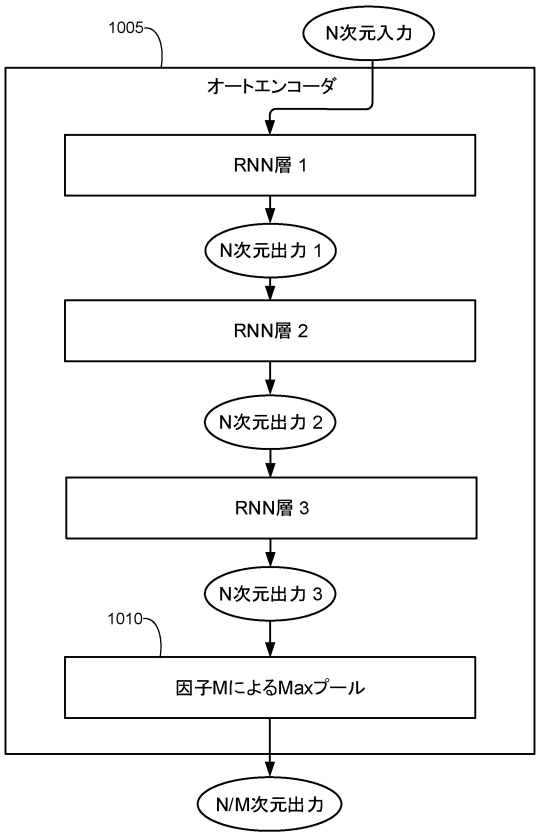
【図８】


S_t
P_t
Q_t
F_t
525
805
810
815
大フレーム
小フレーム
サンプル予測器
C_t

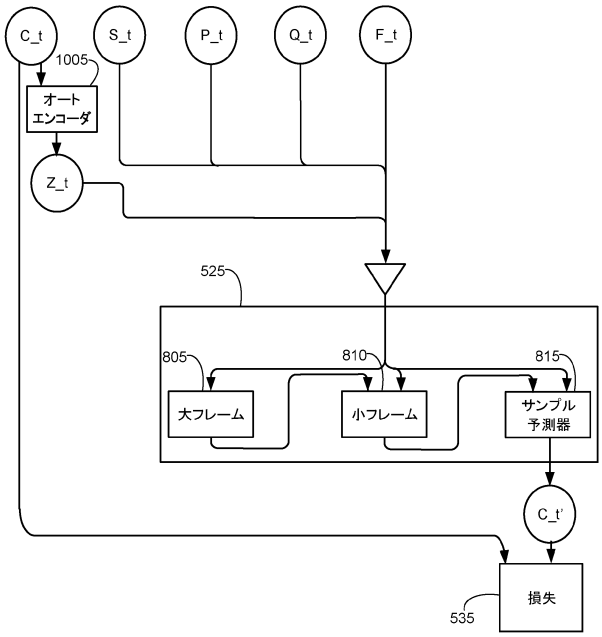
【図９】



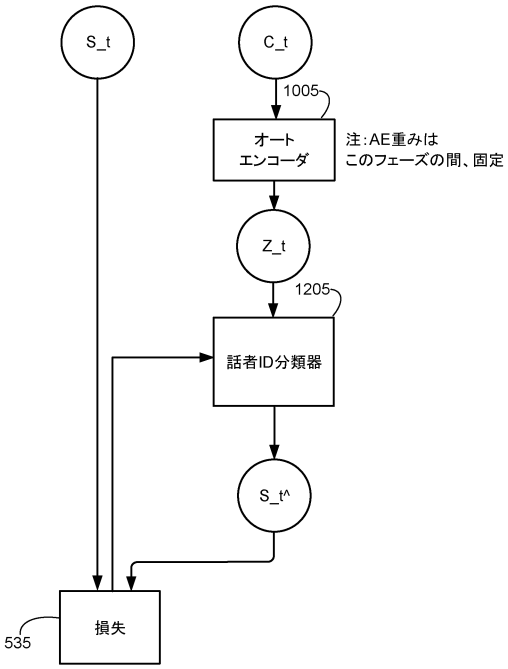
【図１０】



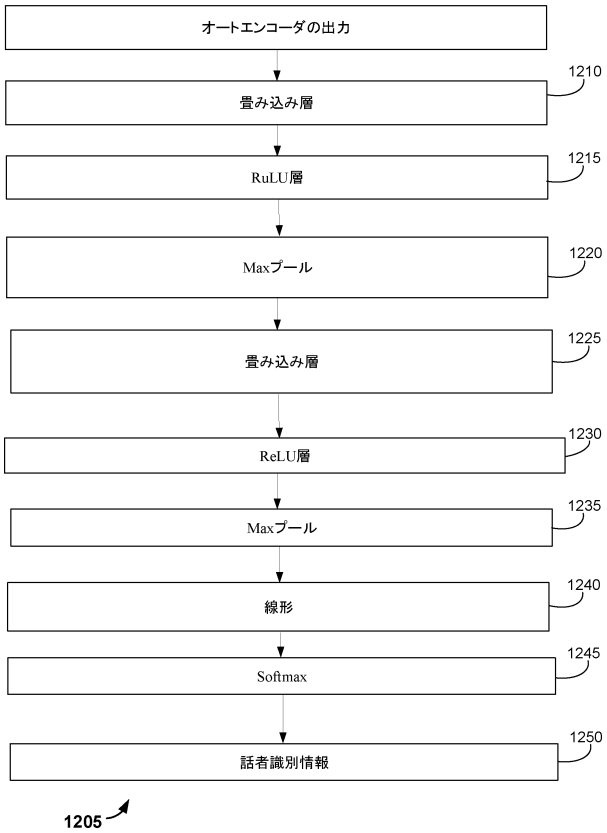
【図１１】



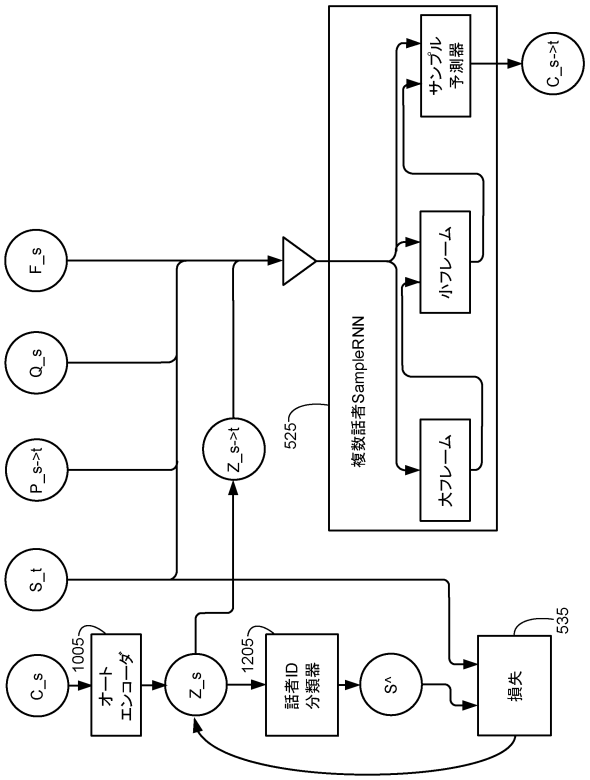
【図１２Ａ】



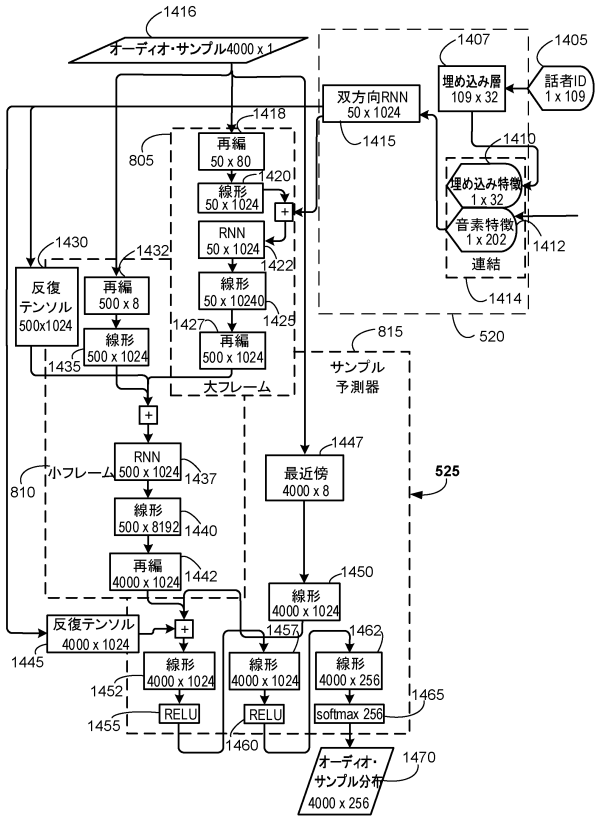
【図１２Ｂ】


オートエンコーダの出力
畳み込み層
1210
RuLU層
1215
Maxプール
1220
畳み込み層
1225
ReLU層
1230
Maxプール
1235
線形
1240
Softmax
1245
話者識別情報
1250
1205

【図１３】


F_s
Q_s
P_s->t
S_t
C_s
1005
オートエンコーダ
Z_s
Z_s->t
525
1205
話者ID分類器
S^
損失
535
複数話者SampleRNN
サンプル予測器
C_s->t
小フレーム
大フレーム

【図１４】


1416
オーディオ・サンプル4000 x 1
1407
埋め込み層 109 x 32
1405
話者ID 1 x 109
双方向RNN 50 x 1024
1415
1418
805
再編 50 x 80
線形 50 x 1024
1420
RNN 50 x 1024
1422
線形 50 x 10240
1425
1410
埋め込み特徴 1 x 32
音素特徴 1 x 202
連結
1412
1414
520
1430
反復テンソル 500x1024
1432
再編 500 x 8
線形 500 x 1024
1435
1427
再編 500 x 1024
大フレーム
815
サンプル予測器
RNN 500 x 1024
1437
小フレーム
810
線形 500 x 8192
1440
再編 4000 x 1024
1442
最近傍 4000 x 8
525
1447
線形 4000 x 1024
1450
反復テンソル 4000 x 1024
1445
1457
1462
線形 4000 x 1024
線形 4000 x 1024
線形 4000 x 256
1452
1465
RELU
RELU
softmax 256
1455
1460
オーディオ・サンプル分布 4000 x 256
1470

フロントページの続き

(31)優先権主張番号 62/797,864
(32)優先日 平成31年1月28日(2019.1.28)
(33)優先権主張国・地域又は機関
米国(US)

早期審査対象出願

(72)発明者 ジョウ, ツオーン
アメリカ合衆国 ９４１０３ カリフォルニア州 サンフランシスコ マーケット ストリート １２７５ ドルビー ラボラトリーズ インコーポレイテッド 内
(72)発明者 ホーガン, マイケル ゲッティ
アメリカ合衆国 ９４１０３ カリフォルニア州 サンフランシスコ マーケット ストリート １２７５ ドルビー ラボラトリーズ インコーポレイテッド 内
(72)発明者 クマール, ヴィヴェク
アメリカ合衆国 ９４１０３ カリフォルニア州 サンフランシスコ マーケット ストリート １２７５ ドルビー ラボラトリーズ インコーポレイテッド 内
(72)発明者 モラレス, ジェイミー エイチ.
アメリカ合衆国 ９４５０１ カリフォルニア州 アラメダ ブロードウェイ ３３９ ユニット ナンバー２０５
(72)発明者 ヴァスコ, クリスティーナ ミシェル
アメリカ合衆国 ９４１１２ カリフォルニア州 サンフランシスコ ミッション ストリート ３９７５ アパートメント １

審査官 上田 雄

(56)参考文献 米国特許出願公開第２０１８／００１２６１３（ＵＳ，Ａ１）
特開２０１７－００３６２２（ＪＰ，Ａ）
米国特許出願公開第２０１６／０３７９６２２（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ１０Ｌ ２１／００－２５／９３