AI 特許紹介(21)

AI 特許を学ぶ！究める！

～コンテキスト依存音響モデル特許～

2020 年 10 月 9 日

河野特許事務所

所長　弁理士　河野英仁

「AI 特許紹介」シリーズは、注目すべき AI 特許のポイントを紹介します。熾烈な競争となっている第４次産業革命下では AI 技術がキーとなり、この AI 技術・ソリューションを特許として適切に権利化しておくことが重要であることは言うまでもありません。

AI 技術は Google, Microsoft, Amazon を始めとした IT プラットフォーマ、研究機関及び大学から毎週のように新たな手法が提案されており、また AI 技術を活用した新たなソリューションも次々とリリースされています。

本稿では米国先進 IT 企業を中心に、これらの企業から出願された AI 特許に記載された AI テクノロジー・ソリューションのポイントをわかりやすく解説致します。

1.概要

特許権者　Google

出願日　2015 年 10 月 7 日

登録日　2017 年 11 月 14 日

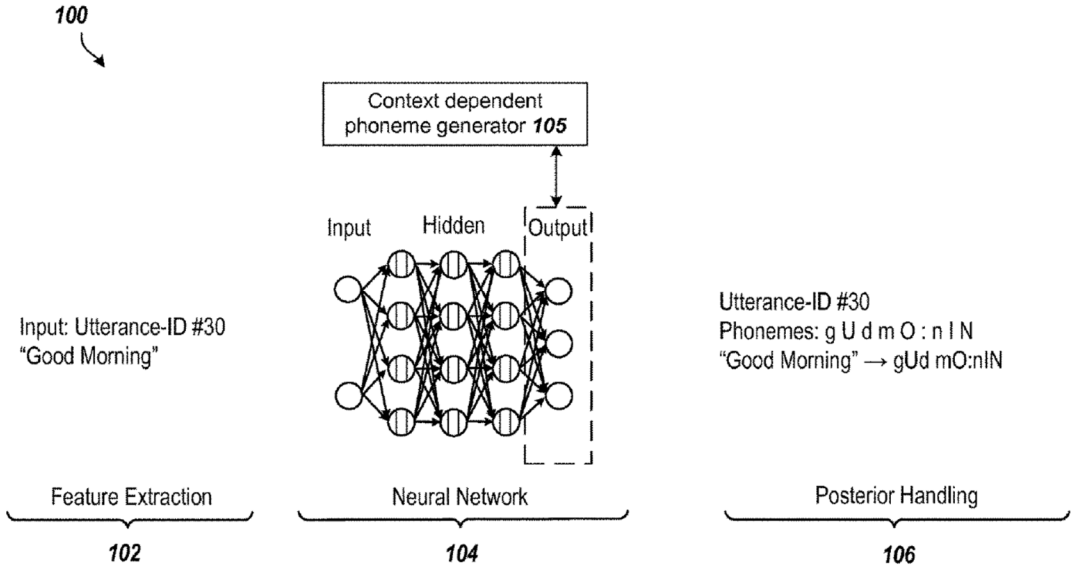
登録番号 US9818409

発明の名称　音素のコンテキスト依存モデリング

409 特許は、自動音声認識システムの音響モデリングシステムによって、音響シーケンスのコンテキスト依存音素表現を決定するアイデアに関する。

2.特許内容の説明

図１は、音響モデリングシステム１００の例を示す。



音響モデリングシステム 100 は、特徴抽出モジュール 102、ニューラルネットワーク 104、コンテキスト依存音素ジェネレータ 105、および事後処理モジュール 106 を含む。

特徴抽出モジュール 102 は、音響シーケンスを受信し、音響波形から、音響シーケンスにおける音響データのフレームの特徴表現を生成する。例えば、音響モデリングシステム 100 は、データの連続ストリームとして発声のデジタル表現を受け取り、ストリームを一連の時間ステップ（例えば 10ms）に対応する一連の複数のデータフレームに分割する。

特徴抽出モジュール 102 は、各フレームを分析して、フレームの特徴値を決定し、対応する音響特徴表現を生成する。例えば、特徴抽出モジュール 102 は、フレームの特徴値を決定し、隣接する特徴ベクトルの左右のコンテキストを使用して積み重ねることができる特徴表現ベクトルに特徴値を配置して、対応するタイムステップでの発話を特徴付けるより大きな特徴表現ベクトルを作成する。

ニューラルネットワーク 104 は、一組の時間ステップのそれぞれの特徴表現を受け取る。ニューラルネットワーク 104 は、特徴表現を処理し、各時間ステップの音素スコアのセットを生成するように訓練されている。各時間ステップの音素スコアのセットには、コンテキスト依存の音素スコアが含まれる。

コンテキスト依存音素ジェネレータ 105 は、コンテキスト依存音素のセットを生成し、ニューラルネットワーク 104 の出力層を構成して、コンテキスト依存音素のコンテキスト依存音素スコアのセットを生成する。

例えば、英語には、/ a /などの少なくとも 41 の語彙音素が含まれている。しかし、/ a /の前に/ k /があり、その後に/ t /が続く場合は、「cat」という単語のように、/ a /の前に/ b /があり、その後に/t/が続く「bat」という単語とは異なり、音は異なる。

事後処理モジュール 106 は、音素スコアを処理し、音響特徴表現のシーケンスの音素表現を生成する。

図 1 に示すように、音響モデリングシステム 100 は、音声のデジタル表現が「Good Morning」という発話を表すデータを含む、時間ウィンドウに対する音声のデジタル表現を受け取る。音響モデリングシステム 100 は、ウィンドウをいくつかのフレームに分割する。特徴抽出モジュール 102 は、各フレームの特徴表現を決定し、例えば、各フレームの特徴ベクトルを決定し、各フレームの特徴表現をニューラルネットワーク 104 に提供する。

ニューラルネットワーク 104 は、特徴表現を分析し、特徴表現のそれぞれについて、一組の音素スコアを生成する。ニューラルネットワーク 104 は、各フレームの音素スコアのセットを事後処理モジュール 106 に提供する。事後処理モジュール 106 は、フレームの音素スコアを組み合わせて、「Good Morning」という発話の音素表現を生成する。例えば、図 1 に示すように事後処理モジュール 106 は、音素表現「gUd mO:nIN」を生成する。

3.クレーム

409 特許のクレーム 1 は以下の通りである。

1. 音響モデリングシステムと言語モデリングシステムを含む自動音声認識システムによって、複数のコンテキスト依存語彙音素を生成し、

トレーニングデータを使用して一連の語彙音素クラスを生成し、

音声の質問を使用して、各語彙音素クラスを 1 つ以上のサブクラスに分割し、

複数のコンテキスト依存語彙音素を生成するために、状態結合アルゴリズムを使用して類似のコンテキストをクラスタリングし、

自動音声認識システムの音響モデリングシステムによって、音響シーケンスを受け取り、音響シーケンスは発話を表し、音響シーケンスは複数の時間ステップのそれぞれで各音響特徴表現を含み、

複数のタイムステップのそれぞれについて：

自動音声認識システムの音響モデリングシステムによって、時間ステップのリカレ

ント出力を生成すべく、1 つ以上のリカレントニューラルネットワーク層のそれぞれを介した時間ステップの音響特徴表現を処理し、

自動音声認識システムの音響モデリングシステムにより、時間ステップのスコアのセットを生成すべく、softmax 出力レイヤーを使用して時間ステップのリカレント出力を処理し、時間ステップのスコアのセットは、複数のコンテキスト依存語彙音素のそれぞれの各スコアを含み、各コンテキスト依存語彙音素のスコアは、コンテキスト依存語彙音素が時間ステップでの発話を表す可能性を表し、

自動音声認識システムの音響モデリングシステムによって、複数の時間ステップのスコアから、音響シーケンスのコンテキスト依存音素表現を決定し、

音響シーケンスの音声認識結果を生成するために、自動音声認識システムの言語モデリングシステムを使用して、自動音声認識システムの音響モデリングシステムによって決定された音響シーケンスのコンテキスト依存音素表現を処理する方法。

4. コンテキスト依存音響モデルに関する論文

Google の Andrew Senior 氏らにより本特許に関連する論文[1]が公表されている。論文ではコンテキスト依存音響モデルにおける WER(Word Error Rate)が示されている。

Model	WER (%)
126-state CI model	16.5
42-state phone model	20.0
42-state phone model with minimum duration 3	16.4

Duration	WER
1 state	12.3
3 state	10.4
4 state	10.2
5 state	10.3
Per-phone	10.1
Per-CD phone	10.0

上側のテーブルは非コンテキスト依存モデルについての WER を示し、下側はコンテキスト依存 LSTM 音響モデルを使用した場合の WER を示している。最高で 10.0WER を達成している。

以上

[1] Andrew Senior, Has¸im Sak, Izhak Shafran “CONTEXT DEPENDENT PHONE MODELS FOR LSTM RNN ACOUSTIC MODELLING”

著者紹介

河野英仁

河野特許事務所、所長弁理士。立命館大学情報システム学博士前期課程修了、米国フランクリンピアースローセンター知的財産権法修士修了、中国清華大学法学院知的財産夏季セミナー修了、MIT(マサチューセッツ工科大学)コンピュータ科学・AI 研究所 AI コース修了。

AI 特許コンサルティング、医療 AI 特許コンサルティングの他、米国・中国特許の権利化・侵害訴訟を専門としている。著書に「世界のソフトウエア特許(共著)」、「FinTech 特許入門」、「AI/IoT 特許入門 2.0」がある。