様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月　 ２ 日現在

機関番号：１２６０１
研究種目：挑戦的萌芽研究
研究期間：2013～2015
課題番号：２５５４００６４

研究課題名（和文）語彙レベル韻律情報の高精度予測に基づく大語彙連続音声認識の高精度化

研究課題名（英文）Improvement of large vocabulary speech recognition performance based on high-precision lexical prosody prediction

研究代表者
峯松　信明（Minematsu, Nobuaki）

東京大学・工学（系）研究科（研究院）・教授

研究者番号：９０２７３３３３

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：日本語は語彙レベルの韻律情報（単語アクセント）が，孤立発声時と文音声発声時とでは異なる。複数出力される音声認識仮説の妥当性を再評価するリランキング処理において，予測される（変形後の）語彙レベル韻律と，実際に観測された韻律とを比較することで，精度向上が期待できる。種々の検討の結果，任意テキスト（認識仮説）に対して語彙韻律を予測するモジュール，及び，リランキング処理の実装は完了したが，観測された韻律に対して正しくアクセント核位置を検出する処理系の実装が極めて困難であることがわかった。最終的には準韻律的特徴と解釈できる音声の構造的表象に基づくリランキングを実装し，音声認識精度向上を実現した。

研究成果の概要（英文）：Japanese has unique characteristics where lexical prosody often vary when words are combined together. In speech recognition research, re-ranking is often used to re-evaluate multiple recognition hypotheses generated from a recognizer and determine the final one. In re-ranking, it is expected that, by comparing lexical prosody predicted from each of the hypotheses and that estimated from an input utterance, better re-ranking is made possible. We implemented successfully 1) lexical prosody prediction from hypotheses and 2) re-ranking of hypotheses based on lexical prosody but it was found to be extremely difficult to build a module that can estimate lexical prosody information precisely only from an utterance. Then, we turned into another strategy of applying quasi-prosody to re-ranking. In the new strategy, structural features are predicted from hypotheses and are also estimated from an input utterance. Experiments showed a high effectiveness of structural re-ranking.

研究分野：音声科学・音声工学

キーワード：音声認識　韻律的特徴　アクセント句境界　アクセント核位置　リランキング　Average perceptron　CRF　構造的表象

１．研究開始当初の背景

連続音声認識技術は，スマホやカーナビシステムなど，日常的に使われるに至っている。しかし，雑音，方言，更には話者の年齢など，様々な要因によって容易に性能が劣化することが報告されている。種々の観点からの解決方法が模索されているが，一つのアプローチとして，音声認識処理の音声特徴抽出モジュールの高機能化が上げられる。従来，音声を認識するためには，音声波形の位相特性や，基本周波数などは，音声の文字化とは独立であると考え，積極的に排除した上で認識処理を行っていた。これに対して，位相特性や韻律を有効利用することによる精度向上が，幾つかの研究において報告されている。

韻律を利用する場合，認識仮説が主張する韻律的情報と，入力音声に観測される韻律的情報とに不一致があれば，その認識仮説は棄却することが期待できる。この場合，認識仮説に基づく韻律予測（生成）技術が必要となるが，これは音声合成技術の一部を用いることで可能となる。申請者らはこれまで，テキスト読み上げ技術における（テキストからの）韻律予測の高精度化を検討しており，この技術を用い，認識仮説群を韻律的に再評価することで，認識精度の向上を検討した。

２．研究の目的

音声認識は，一般に，認識仮説を複数生成し，それらを後処理的に再評価し，最終的な結果を出力することが広く行なわれている（リランキング処理）。当然再評価の際に参照される音響特徴や言語特徴は，認識仮説生成時の時とは異なるものが望ましい。既に示したように，音声認識処理（仮説生成処理）では，韻律的特徴は積極的に排除されるため，この仮説群を韻律的に再評価することで精度向上が見込める。この場合，以下の技術を構築する必要がある。

A) 各仮説を読み上げた場合に想定される韻律を，仮説のみから（テキストのみから）推定する技術。本研究では韻律的特徴として基本周波数（ピッチ）に着眼するため，それと関連するテキスト情報である，アクセント句境界位置，アクセント核位置の推定技術。
B) 入力音声から抽出された基本周波数パターンより，仮定されるアクセント句境界位置，アクセント核位置を推定する技術。

即ち，アクセント句境界や，アクセント核位置などの韻律情報をテキストから予測する技術と，実際に観測された音声波形から予測する技術が必要になる。これらの技術を効果的に用いて認識仮説の再評価を行ない，音声認識の精度向上を狙う。

３．研究の方法

テキストからアクセント句境界，アクセント核位置を求める手法であるが，日本語の場合，アクセント核位置は，孤立単語発声時と，文発声時とでは異なってくる（東京＋大学→東京大学)。これはアクセント変形と呼ばれるが，この変形を高精度に予測する必要がある。テキスト音声合成の分野では，従来この処理は規則で記述することが多かったが，例外が多く，規則が煩雑となる。この問題に対し，大規模な事例集（コーパス）を収集し，機械学習を用いて，入力（孤立発声時の単語アクセント情報）から出力（文発声時の各単語のアクセント情報）を予測する方法を検討した。この場合，アクセント情報は，話者（世代）によって揺れることが想定されるため，特定の話者に，約 6000 文の日本語テキストに対し，個々の単語の孤立発声時のアクセントと，文発声時のアクセントを記載させ，また，アクセント句境界位置についてもラベリングさせた。構築されたラベル付きテキストコーパスを使って，多種多様な素性を定義し，Conditional Random Field (CRF)を用いて入力・出力の対応付けを学習させ，韻律予測の精度向上を試みた。

韻律予測に関する話題として，推定されたアクセント句境界位置，アクセント核位置より，そのテキストを読み上げた場合に観測されると想定される，基本周波数パターンの生成・推定についても検討した。これは，従来より広く用いられている基本周波数パターン生成過程モデルを用いることで実装できるが，ここでは，実測された基本周波数パターンより，生成過程モデルパラメータ（アクセント句境界位置やアクセント核位置とは異なる）を推定する問題（逆問題）に着手した。ここでは，安定して基本周波数が検出できる母音部のみに着眼し（有声子音による基本周波数を除外して）モデルパラメータを推定することによる，精度向上（安定性向上）を検討した。

認識仮説に対して想定されるアクセント句境界位置やアクセント核位置（更には，それらに対する基本周波数パターン）は上記技術の構築によって推定されるが，これらとは別に，入力音声に対して推定されるアクセント句境界位置や，アクセント核位置を推定する必要がある。これについては，従来より広く使われている読み上げ音声コーパス（ATR503 文コーパス）に対して，アクセント句境界，アクセント核位置をラベル化し，各種韻律的特徴より，どのモーラ境界に句境界やアクセント核が存在するのかを検出する技術が必要となる。ここでは，句境界やアクセント核が存在する・しないの二値問題として考え，SVM による実装を検討した。

認識仮説から推定される各種韻律情報，入力音声から推定される各種韻律情報を比較することで，認識仮説の再評価を行うが，こ

の処理については，識別的なリランキングが可能な，average perceptron 法を採用した。

４．研究成果

認識仮説（テキスト）からのアクセント句境界位置，アクセント核位置の推定に関しては，形態素解析結果，文節解析結果，更には，助数詞の特性など様々な解析結果より素性値を取得し，これに対して CRF を適用することで韻律情報の推定を試みた。従来の規則処理と比較して，アクセント句境界位置推定では，88.9%から 93.8%へとＦ値が向上した。また，推定されたアクセント句に対するアクセント核位置推定においては，87.6%から 94.7%と大幅に精度向上を実現することができた（詳細は雑誌論文６に譲る)。

これら（テキストからの）韻律情報の推定に関連する話題として検討した，音声からの基本周波数パターン生成過程モデルパラメータの推定（逆問題）に関しても，母音部分へ積極的に着眼することによる（有声子音部を無視する）有意な精度向上が実現できた（詳細は雑誌論文２に譲る)。

しかしながら，入力された音声に対して，何ら言語的な制約を入れずに，アクセント句境界位置，アクセント核位置を推定する技術に関しては，困難を極めることとなった。これは，アクセント句境界位置や，アクセント核位置と関連する基本周波数の変動パターンが，コンテキストに大きく依存するからである。このコンテキストを事前に与えようとすれば，それはその音声の認識することとなり，音声認識精度の向上のために，認識結果を与えることとなる。これらの検討の結果，当初の予定であった，韻律的な認識仮説の再評価という方針を，大きく方向転換する必要が生じた。このために研究期間を延長した。

最終的には，準韻律的特徴と解釈できる，音声の構造的表象に基づく認識仮説の再評価を検討した。音声の構造的表象は，メロディーに対する相対音感（階名知覚）にヒントを得て提唱された音声表象であり，分節的特徴の軌跡（トラジェクトリ），即ち音系列に対して，音と音の相対的な位置関係だけを保存する表象である（通常，メロディーは基本周波数という一次元の物理量の軌跡となるが，音声を多次元の物理量の軌跡と捉え，音と音との相対関係だけを捉える)。アクセント句境界位置やアクセント核位置といった韻律情報に基づく仮説再評価を，音声の構造的表象に基づく仮設再評価として実装を試みる訳だが，それ以外の検討事項（仮説から想定される構造的情報の推定，実際の入力音声から推定される構造的情報の推定，及び，両者の比較による仮説の再評価）は，ほぼ同じ枠組みのものが利用可能である。

認識仮説から得られる（音素を単位とした）構造的表象は，事前に大規模コーパスより，音素-音素の構造的表象として構築しておくことができる。一方，実際の入力音声についても仮説生成時に想定される音素アライメント結果より，音素-音素の構造的表象を抽出できる。これら二種類の構造的表象を average perceptron により識別的に再評価し，最終仮説を決定した。なお，タスクとしては，大語彙連続音声認識以外にも，連続数字認識についても検討した。前者の場合，誤り削減率 25%，後者の場合 50%を達成することができた。前者の場合の精度が落ちる理由としては，日本語には同音異義表現が多いことが理由であると考えている（詳細は雑誌論文１に譲る)。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 6 件）


1. M. Suzuki, G. Kurata, M. Nishimura, N. Minematsu, "Discriminative re-ranking for automatic recognition by leveraging invariant structures," Speech Communication, 72, 208-217 (2015)
2. 橋本浩弥，齋藤大輔，峯松信明，広瀬啓吉，"HMM 音声合成を目的とした基本周波数パターン生成過程モデルのモデルパラメータ推定"，電子情報通信学会和文論文誌, J98-D, 3, 481-491 (2015)
3. C. Zhang, M. Suzuki, G. Kurata, M. Nishimura, N. Minematsu, "Leveraging phonetic context dependent invariant structure for continous speech recognition," Proc. IEEE China Summit & Int. Conf. on Signal and Information Processing, 52-56 (2014)
4. Y. Luan, D. Saito, Y. Kashiwagi, N. Minematsu, K. Hirose, "Semi-supervised noise dictionary adaptation for exemplar-based noise robust speech recognition," Proc. ICASSP, 1764-1767 (2014)
5. Y. Kashiwagi, D. Saito, N. Minematsu, K. Hirose, "Discriminative piecewise linear transformation based on deep learning for noise robust automatic speech recognition," Proc. ASRU, pp.350-355 (2013)
6. 鈴木雅之，黒岩龍，印南佳祐，小林俊平，清水信哉，峯松信明，広瀬啓吉，"条件付き確率場を用いた日本語東京方言のアクセント結合自動推定"，電子情報通信学会論文誌，J96-D，3，644-654 (2013)


〔学会発表〕（計 4 件）

1. 柏木陽祐，齋藤大輔，峯松信明，広瀬啓吉，″識別的アプローチによる分布間距離推定の検討とその言語識別への応用″，電子情報通信学会音声研究会資料，SP2015-38，pp.13-18（2015）
2. 柏木陽佑，齋藤大輔，峯松信明，広瀬啓吉，″制約付き話者コードの同時推定によるニューラルネット音響モデルの話者正規化学習″，日本音響学会秋季講演論文集，1-8-3，pp.7-10（2014）
3. 柏木陽佑，齋藤大輔，峯松信明，広瀬啓吉，″Deep Learningに基づくクリーン音声状態識別による雑音環境下音声認識″，日本音響学会秋季講演論文集，1-8-3，pp.9-12（2013）
4. 橋本浩弥，広瀬啓吉，峯松信明，″文節を基本単位とした基本周波数パターン生成過程モデルのパラメータ自動抽出″，日本音響学会秋季講演論文集，1-P-5a，pp.327-328（2013）

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
峯松　信明（MINEMATSU, Nobuaki）
東京大学・大学院工学系研究科・教授
研究者番号：９０２７３３３３