6U-5

言語的・音響的コンテキストが音声の聴取および認識に及ぼす影響の考察

榎並 大介† 山本 一公† 北岡 教英†† 中川 聖一†
†豊橋技術科学大学 ††名古屋大学

1 はじめに

大語彙連続音声認識 (Large Vocabulary Continuous Speech Recognition; LVCSR) においては，隠れマルコフモデル (Hidden Markov model; HMM) と N-gram が音響モデルおよび言語モデルとしてよく用いられる．

本稿では，人間の音響的知覚能力と言語的単語予測能力を，局所的なコンテキストを教示して音声を聞かせ理解させることで調査し，音響モデルと N-gram 言語モデルによる音声認識システムと比較，人間と機械との違いを各モデルについて検討する．文献 [2][3] ではユニグラムとトライグラムを比較対象としたが，本報告ではバイグラムも比較対象に加えた．

2 人間による聴取実験

本節では，人間へのコンテキストの教示と人間の音響的知覚能力を組み合わせ，その能力を調べる．

2.1 聴取実験の設定

何種類かの単語コンテキストが与えられた場合の，人間の単語聴取能力を調査した．テストセットとして，読み上げ音声および自由発話音声の 2 種類の音声データからそれぞれ 100 単語ランダムに選定，聴取対象単語とし，各聴取対象単語はコンテキストを含めて人手により切り出した．被験者はすべて，音声処理に関係があり，講演音声の分野についてある程度の知識のある修士の学生である．

2.2 聴取実験と単語予測実験の結果

2.1 節の設定による聴取実験の結果を，図 1 に結果を示す．切出した音声区間が前 2 単語のコンテキストありの場合の聴取率が，コンテキストなしの場合を大きく上回っている．特に短い単語 (助詞など) は，それのもつ音響的情報が少ないため，それだけでは聴取が難しいが，コンテキストを加えると言語的制約が効率的に働くと考えられる．

人間が音響的情報を用いずに，単語のコンテキスト (文字列) 情報のみを与えられた場合に対象単語を予測する能力についても調査した．すなわち，テキストで

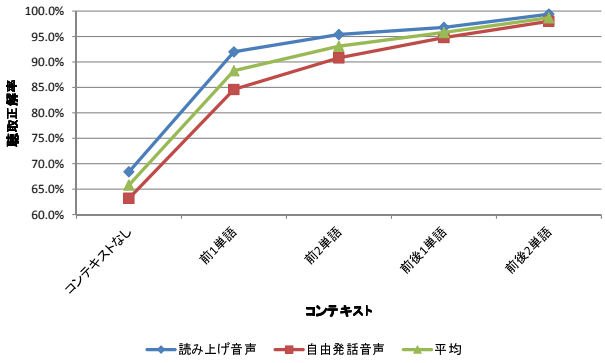

図 1: 人間の聴取実験結果

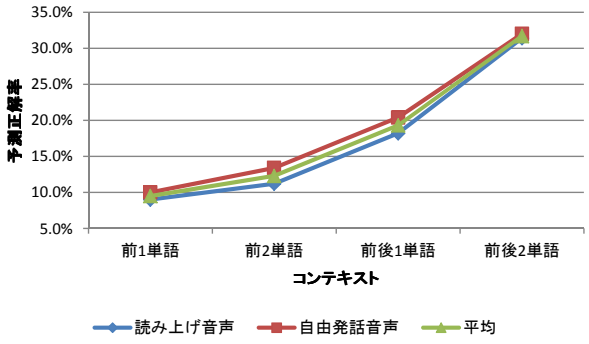

図 2: 人間の単語予測実験結果

与えられるコンテキストのみから対象単語を予測する実験である．結果を図 2 に示す．前 2 単語はトライグラム言語モデルに対応する．

人間は，より多くのコンテキストを与えることによって正確に単語が予測できることがわかる．

3 音声認識システムによる認識実験

本節では，2 節で行った聴取実験と同様の条件下での音声認識システムによる単語認識実験を行う．

3.1 認識実験の設定

音声認識システムによる認識実験においても 2 節と同じ単語を対象単語とした．既知である直前 1 単語および直後 1 単語の HMM は固定し，中心単語に対応する HMM を差し替えて 3 単語の区間とのマッチングを行って得られる 3 単語分の尤度を中心単語の尤度として扱い，これを認識対象語彙すべてに対して行った．こうして求めた音響尤度に，対数領域において言語スコアを適切な重みで加えることによりトータルのスコ

Consideration for Effects of Linguistic and Acoustic Contexts on Speech Perception and Recognition
Daisuke ENAMI†, Kazumasa YAMAMOTO†, Norihide KITAOKA ††, Seiichi NAKAGAWA†
† Toyohashi University of Technology
†† Nagoya University

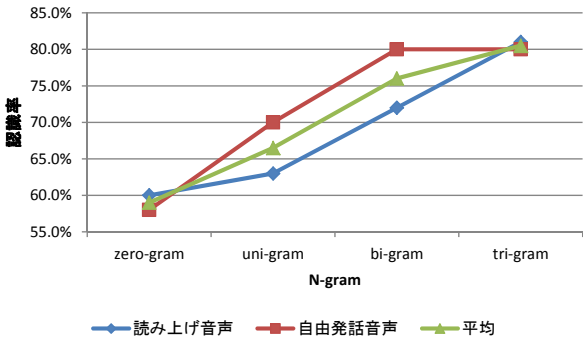

図 3: 音声認識システムの単語認識実験結果

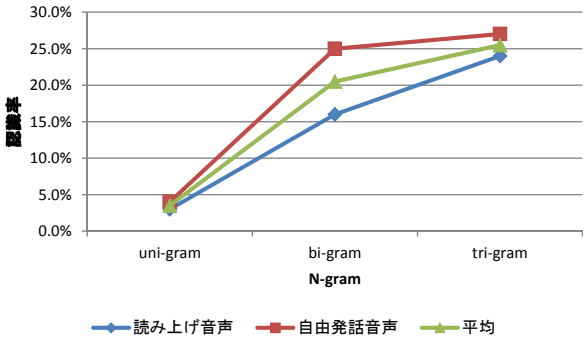

図 4: 言語スコアによる単語認識実験結果

アを求めた．

音声データに窓長 25ms のハミング窓を用いて窓掛けしたフレームからシフト長 10ms で 38 次元 (12 次元の MFCC、およびその Δ，ΔΔ，対数パワーの Δ，ΔΔ) の特徴パラメータを求めた．928 個の左コンテキスト依存型音節 HMM は各々5 状態 4 出力分布を持ち，各出力分布は 32 混合の対角共分散 GMM からなる．

学習データには，読み上げ音声の認識のために，日本音響学会音声データベース (30 話者，4518 発声) および新聞読み上げコーパス JNAS(145 話者，23474 発声) を，自由発話音声のために，日本語話し言葉コーパス CSJ の講演のうち男性話者 814 講演を用いた．認識デコーダには大語彙連続音声認識システム SPOJUS++[4] を用いた．

言語モデルは，読み上げ音声のために毎日新聞 75ヶ月分を，自由発話音声のために CSJ の学会講演 970 講演を用いて作成した 2 万語のトライグラムを用いた．

3.2 認識実験の結果

トータルスコアによる単語認識実験の結果を図 3 に，言語スコアのみによる認識実験の結果を図 4 示す．

図 3 において，言語モデルを用いずに音響モデルのみを用いた場合 (zero-gram) は平均で認識率 59.0%となった．ユニグラムを用いた場合 (uni.-gram) は 66.5%に改善した．さらにバイグラム (bi-gram)，トライグラム (tri-gram) により，それぞれ 76.0%，80.5%に改善した．図 4 の音響モデルを用いない場合は，人間による視察におけるコンテキストからの単語予測実験に相当し，言語モデルの予測能力を示していると言える．トライグラム (tri-gram) により，全対象単語のうち，平均で 25.5%が正しく予測されている．

これらの音声認識結果は，人間による知覚実験結果における図 1，2 と比較できる．認識システムの平均性能が，人間の聴取と比べて明らかに劣っているといえる．一方，図 4 のトライグラム言語モデルのみによる予測性能 (tri-gram) は，図 2 の前 2 単語を与えた場合の人間による予測能力よりも良い結果である．図 2 によれば，前後 2 単語のコンテキストを用いる条件における人間の予測能力は前 2 単語や前後 1 単語の場合のそれよりもかなり良いが，より大きな N を用いることによる N-gram の予測性能の向上はそれほどは見込めない．人間とシステムの言語コンテキストの利用に違いがあると言える．このことより，N-gram 言語モデルを用いるという考えのもとで，$N = 3$ は十分であると示唆される．

4 まとめ

人間の知覚実験においては，より長いコンテキストを与えることによって単語予測能力は向上する．一方，音声認識システムの認識実験において，トライグラム言語モデルによる認識 (予測) は人間が前 2 単語や前後 1 単語のコンテキストを用いて行う予測よりも優れている．これらのことから，局所的な言語知識を用いる N-gram モデルと HMM のような音響モデルとの組合せによる音声認識においては，トライグラムモデル化は十分に強力な表現能力を持っているが，言語モデルのこれ以上の改善による認識率向上は難しい一方で，音響モデルはまだ改善すべき点が多く存在すると考えられる．

これらの結果は，音響情報が言語情報よりもより多くの情報を持つという知見・考察 [1] を明確に支持するものとなった．

参考文献

[1] 中川 聖一, “音声認識研究の課題”, 信学技法, SP99–93, 1999.

[2] 北岡 教英, 新宮 将久, 中川 聖一,“言語的・音響的コンテキストが講演音声の聴取および認識に及ぼす効果”, 信学技法, SP2003-33, pp.95-96, 2003.

[3] 榎並 大介, 山本 一公, 北岡 教英, 中川 聖一,“言語的・音響的コンテキストが音声の聴収および認識に及ぼす効果の再評価”, WiNF2011 第 9 回情報学ワークショップ論文集, 2011.

[4] 藤井 康寿, 山本 一公, 中川 聖一,“大語彙連続音声認識システムの改善：SPOJUS++”, 第 4 回音声ドキュメント処理ワークショップ論文集, 2010.