定型的表現部分の発話検証法を用いた大語彙音声認識

足立 賢一郎† 甲斐 充彦‡
† 静岡大学大学院 理工学研究科 ‡ 静岡大学 工学部

1 はじめに

音声認識システムでは、音声の音響的特徴を表現する音響モデルと、言語的特徴を表現する言語モデルが使われる。大語彙音声認識では、言語モデルとして単語連鎖の統計的モデルである N-gram 言語モデルが一般に用いられる。このような統計的手法は、扱う対象が書き言葉を読み上げた音声であるものに対して有効であったが、講演音声のような変動の要素が大きい自然な話し言葉の音声では、認識精度が極端に低下してしまう問題がある。N-gram 言語モデル (N=1,2,3,…) において、長い単語単位を用いることは広い言語コンテキストをモデル化出来ることになるが、学習データが有限であるためモデルの正確な推定は難しくなる。また実験的に、短い単語は誤認識の原因になりやすいことが知られている [1]。本研究では、ある程度の長さ (N>>2) の言語コンテキストを言語情報として取り入れることを目的とする。そのため、まず長い単語単位を用いた単語スポッティングを実現し、その定型的表現部分の検証法と単語 N-gram を用いた大語彙音声認識とを併用する手法を検討して、講演音声認識タスクで評価を行う。

2 定型的表現部分の検証法の概要

N-gram 言語モデルによる認識において、典型的な誤り易い言語表現 (定型表現) へ対処する方法として、長い単語の登録が考えられる。しかしこの場合、N>>3 の N-gram 統計量の推定が困難であるのと同様の問題が生じる。これに対して可変長 N-gram の考え方があるが、これはデコーダのアルゴリズムを複雑化する。

そこで、認識率が低く、出現頻度の高い単語列部分を一つの長い単語 (検証する定型表現) とみなし、それらの表現部分をスポッターにより発話検証 (検出・同定) して、既存のデコーダと併用・統合するアプローチを考える。つまり、N-gram で認識が困難な表現はより長い単語単位の検証結果と比較し、式 (1) で示すように、スコアの差がある閾値 σ よりも高ければスポッターによる結果を、低ければ既存の単語 N-gram を用いたデコーダによる結果を選択する。図 1 において、$t >$ τ であり、τ はスポッターのバックトレースにおいて求まる、定型的表現（■■■の部分）とその先行単語との間の最適な単語境界である。

$$\frac{P(Y_1^t|W'\oplus W_k)P(W'\oplus W_k)}{P(Y_1^t|W)P(W)} > \sigma \qquad (1)$$

但し、W_k は定型的表現部分、W' は W_k に先行する単語列であり、$W'\oplus W_k$ は単語列 W' と W_k の連結を表す。また、後述の評価実験では、定型的表現部分を含む言語確率 $W'\oplus W_k$ は $P(W')$ で近似するか、または $P(W_k)$ の部分を単語数に比例した値で近似する。ここで使用するスポッターは、連続 DP 法の原理で始端点フリーの照合を行うことで発話中に存在する定型的表現部分の区間の候補とスコアの出力を行う。しかし、発話先頭からスポッティング区間の先行部分までは、既存のデコーダの途中結果第一位の仮説、すなわち単語列 W' のスコアを仮定し、区間抽出誤りを抑える [2]。よって発話中のある時点で定型表現で終わっているかどうかの検証は、スポッターで求まる仮説のスコアが式 (1) の分子に直接対応するため、既存のデコーダの中間結果 (分母) と直接比較することができる。

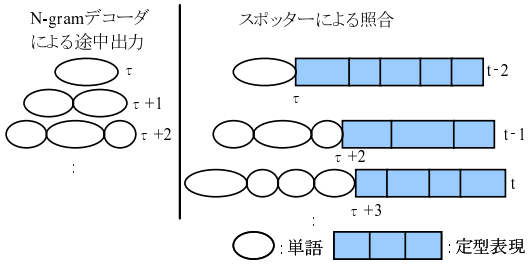

図 1: 定型的表現部分のスポッターの併用

上記の考え方に基づく処理の流れは次のようになる。まず、評価タスクに類似した開発用 (定型表現の抽出用) データの音声認識結果の評価に基づいて、誤りが多く出現頻度が高い定型表現を抽出する。次に、評価データに対しては、一般的な単語 N-gram を用いたデコーダと並行して、与えられた定型表現のスポッティングを行う。デコーダとスポッターの結果から総合的に式 (1) に基づいて定型表現の有無を考慮した仮説 ($W'\oplus W_k$ か W) の評価を行い、最終的な認識結果の出力を行う。

3 評価実験

3.1 誤り易い定型的表現の抽出

検証する定型的表現部分を選出するために「日本語話し言葉コーパス (CSJ:Corpus of Spontaneous Japanese)」のうち、評価データで用いた以外の 50 講演を使用した。音響モデルは 5 状態 4 ループの音節 HMM (116 音節)、言語モデルは単語 bigram(単語 N-gram:N=2) を用いた。またデコーダには SPOJUS を使用し、言語重み、単語挿入ペナルティの値は前実験で使用された最適な値 (言語重み 15、単語挿入ペナルティ－ 10) を利用した。

認識結果で単語正解精度が 50%以下の発話から、3 つ組み以上の頻出定型表現を求める (頻度は 5 回以上)。また 3 つ組みの定型表現については 5 モーラ以上の定型表現を選出した。デコーダ及びスポッターで扱う発話の単位は、CSJ コーパスで決められている時間タグ付与の最小単位の基準と同じく、ポーズ長 300ms 以上で囲まれた範囲を一発話としている。定型表現のうち、実際に単語 bigram を用いた認識で誤認識している定型表現のみを、検証する定型表現とした。その結果、そのような条件を満たす定型表現は全部で 268 種類となった。

評価用データの中で検証する定型表現を含む発話は、全 4054 発話に対し 979 発話 (24.2%) あり、979 発話中 565 発話 (57.7%:全体の 13.9%) が発話末に検証する定型表現を含んでいた。このことから、まず発話末に着目してスポッターによる定型表現部分の検証を含んだ音声認識の評価を行う。

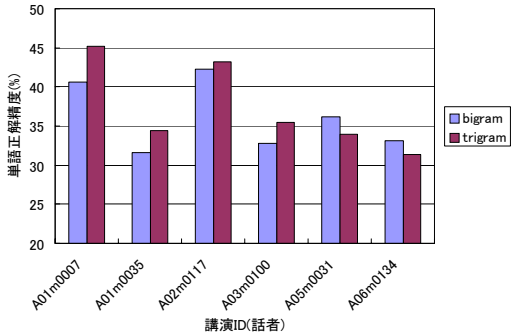

図 2: trigram 言語モデルの効果

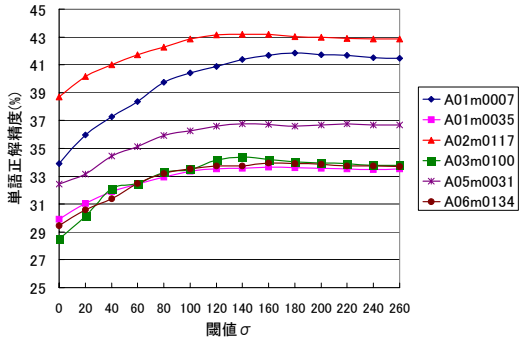

図 3: 閾値の違いによる単語正解精度の変化

表 1: スポッターの併用による認識性能の比較 (%)

講演 ID(話者)	bigram	bigram+Spotter
A01m0007	40.7	41.4
A01m0035	32.6	33.6
A02m0117	42.3	43.2
A03m0100	32.8	34.4
A05m0031	36.2	36.8
A06m0134	33.1	33.8

3.2 定型表現検証法の評価

まずベースラインの性能として、既存の単語 N-gram を用いたデコーダの結果を図 2 に示す。但し、単語 trigram を用いた結果は、ツリー構造辞書を一本化して 1-best 近似照合を行う 1 パス・デコーダにおいて、trigram の再計算を行うように改良した 1 パス探索法を用いた場合の結果である。この方法は、従来のリスコアリングによる 2 パスの方法より良い認識性能が得られている。結果は一部の講演（話者:A05m0031,A06m0134）では、高次の単語 N-gram 言語モデルが有効に働いていないことが分かる。また、認識精度が全体として低いが、言語モデル作成において形態素解析・読み付与誤りが多く含まれるのが一因となっている。

次に、式 (1) で σ を変化させたときの話者ごとの単語正解精度を図 3 に示す。この結果より σ=140 のとき話者ごとの単語正解精度の平均が最も良くなった。また、この閾値において既存のデコーダによる認識性能 (N-gram) と比較した結果は表 1 のようになり、スポッター併用による改善の効果がみられた。

3.3 定型的表現部分における認識性能

3.2 節では図 1 の時刻 t を発話末に限定した話であったため、ベースラインの性能改善に寄与するのは発話末の部分の認識結果の変化だけである。そこで、検証する単語列部分を発話末に持つような発話に対して、その部分のみの単語正解精度を、N-gram を用いた結果（N-gram）、スポッターを併用した場合の結果 (N+S) で比較する。閾値 σ は 40 で単語正解精度の平均が最高となった。発話末の検証する定型表現部分に対する N-gram を用いた単語正解精度と、スポッターを併用したもので σ=40、σ=140 での単語正解精度を図 4 に示す。

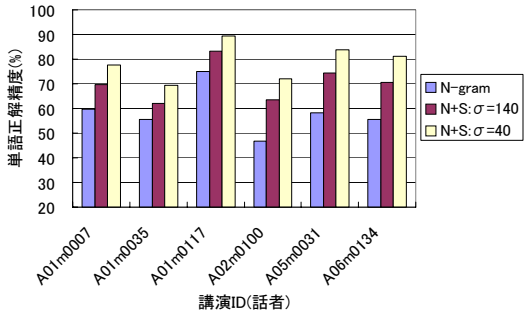

図 4: 検証する単語列を含んだ部分での認識性能

表 2: スコア評価式の改良後の認識性能 (%)

講演 ID(話者)	bigram	bigram+Spotter(ペナルティ / 閾値)
A01m0007	40.7	41.7 (60 / -60)
A01m0035	32.6	33.7 (80 / -120)
A02m0117	42.3	43.3 (40 / 0)
A03m0100	32.8	34.4 (20 / 80)
A05m0031	36.2	36.8 (60 / -90)
A06m0134	33.1	33.9 (80 / -120)

3.4 スポッターにおけるスコア評価式の改良

前述の方法では、スポッターにおけるスコア評価においてその単語間の確率 (N-gram) を考慮していない。そのため検証する単語列が長くなるにつれ、式 (1) の比較ではスポッターによる仮説のほうが有利に働きやすくなる。そこで式 (1) の分子において、検証する単語列の長さに対応したペナルティの値を与える。ペナルティの値は定型的表現部分の長さに比例した値とする。様々なペナルティに応じた閾値を事後的に最適に調整したときの単語正解精度の最大値と bigram による単語正解精度の比較を表 2 に示す。

4 まとめ

スポッターによる発話末の定型的表現部分の検証法を提案し、大語彙音声認識の精度改善の可能性を示した。

本研究では、調整すべきパラメータを事後的に調整しているため、パラメータ決定を含めた評価・検討を行うことが今後の課題である。また、この処理を発話末以外の定型的表現部分にも適用した場合での効果も評価を行う。その際、現在の方法では検証する定型表現とその前後の検証部分以外との接続における連鎖確率を考慮していないため、それらを考慮して更に評価を行う。

謝辞

本研究では開放的融合研究「話し言葉工学」による「日本語話し言葉コーパス (モニター版 2001)」を利用した。

参考文献

[1] 篠崎隆宏、古井貞熙 ”話し言葉認識における決定木を用いた誤り要因の分析,”音講論,1-1-9,pp.17-18,(2001-10)

[2] 渡辺隆夫、吉田和永、畑崎香一朗 ”バンドルサーチ法を用いた連続音声認識の高速化,”信学論,Vol.J75-D- ,No.11,pp.1761-1769,(1992)