

音声認識の信頼度に着目した文境界検出に関する検討

畑 昇 吾^{†1} 西田 昌 史^{†2}
堀 内 靖 雄^{†1} 黒 岩 眞 吾^{†1}

自然言語処理では処理単位として文などの意味的なまとまりがある単位を用いるため、音声認識結果に対して文境界を示す必要がある。本研究では、まず SVM を用いた文境界検出において文境界直前における語の出現しやすさを考慮することによって文境界検出に適した特徴空間の作成方法を提案する。さらに、音声認識時に認識結果と共に出力される単語信頼度を素性として文境界検出に利用することを検討する。文境界検出においては『日本語話し言葉コーパス (CSJ)』を対象として SVM を用いて評価実験を行った。

Sentence Boundary Detection Focused on Confidence Measure of Automatic Speech Recognition

SHOGO HATA,^{†1} MASAFUMI NISHIDA,^{†2} YASUO HORIUCHI^{†1}
and SHINGO KUROIWA^{†1}

Since the units of processing for Natural Language Processing(NLP) are based on syntactic structure, for example sentence, it is necessary to detect the sentence boundary for the Automatic Speech Recognition(ASR) outputs. In this paper, at first, we propose the feature space that is applied to detecting sentence boundary with Support Vector Machine(SVM) by considering the frequency of the word immediately before sentence boundary. At second, we examine using confidence measure of ASR outputs for sentence boundary detection with SVM. We evaluated our methods on the Corpus of Spontaneous Japanese(CSJ).

^{†1} 千葉大学
Chiba University

^{†2} 同志社大学
Doshisha University

1. はじめに

近年、音声認識と自然言語処理を組み合わせた音声要約、音声理解、音声翻訳などの研究が進められている。現在の自然言語処理では、書き言葉の「節」や「文」といった意味的なまとまりを処理単位として用いることが多い。その一方で、音声認識は一般的に一定のポーズ長で分割した音声を処理単位として用いている。ポーズは「節」や「文」とは無関係に出現することがあるため、音声認識結果が「節」や「文」と一致するとは限らない。そのため、音声認識結果に対して既存の自然言語処理を組み合わせることは困難といえる。また、音声認識結果の可読性についても、意味的なまとまりとして出力されないため十分とはいえない。

これらの問題点から、入力音声を文単位に分割或いは音声認識結果に対して文境界を推定する研究が進められている。音声認識過程での文境界の推定に関する研究¹⁾が日本語を対象として行われている。また、日本語では明らかな文末表現(「です」「ます」など)が存在することから音声認識結果に対して言語的な特徴を用いて文境界を推定する手法も多く提案されている。例えば、統計的言語モデルおよびサポートベクターマシン(SVM)を用いた研究²⁾や係り受け関係を利用した研究³⁾などが挙げられる。文献³⁾では、文境界直前の語の係り受け関係が他の語のそれとは異なる特徴を示しやすいという言語的手がかりを利用して文境界の推定を行っている。また、係り受け関係を隣接文節間に限定した研究⁴⁾も行われている。文献⁴⁾では、係り受けを隣接文節間に限定することで話し言葉の非定型性や音声認識誤りに対して頑健な特徴の抽出を行っている。一方で音響的な素性としては韻律を用いた研究が行われている⁵⁾。文献⁵⁾では、基本周波数(F0)が文あるいは句単位ごとに立ち上がった後に下降調になる等の特徴的なパターンを示す韻律的特徴を文境界推定に用いている。

本稿では、文境界検出において SVM に与える素性を文境界検出に適した特徴空間上へ数値化(プロット)する手法を提案する。これは文境界直前で出現しやすい語、出現しにくい語を考慮することで実現する。また、新たな素性として音声認識結果と共に出力される単語信頼度の導入についても検討する。これは、音声認識における探索時の展開候補単語数が少なくなれば単語信頼度の値が大きくなるという特徴を利用するものである。これらを SVM に対して利用した文境界検出の評価を『日本語話し言葉コーパス (CSJ)』を対象として行う。

表 1 CSJ の節境界ラベルの例

Table 1 Example of ClauseBoundaryLabel of CSJ

境界の種類	節境界ラベル	使用例
絶対境界	文末 文末候補 と文末	思います
強境界	並列節「ケド」「ガ」など	思うけど
弱境界	タリ節 理由節「カラ」 テ節 条件節「ナラ」「レバ」「ト」など	思っ

2. CSJ における節境界

CSJ は主に学会講演や模擬講演などを対象として収集・構築されたコーパスである。CSJ では統語的・意味的な妥当性を備えた単位として「節単位」を定義している。節単位はその性質上、音声言語処理に有用であると考えられる。この節単位は、節境界によって推定され、人手による修正を経て認定される⁶⁾。節境界は、直後の切れ目の大きさによって絶対境界・強境界・弱境界の 3 種類に分けられている。各境界の例を表 1 に示す。

絶対境界は形式上明示的な文末表現（「です」「ます」「た」など）に相当する。強境界、弱境界はいわゆる文末ではない。強境界は発話の大きな切れ目として考えられており、弱境界は通常発話の切れ目になることはないと考えられている節境界である。本稿では絶対境界を文境界として扱う。節境界ラベルの推定にはプログラム CBAP-CSJ⁷⁾ が用いられている。CBAP-CSJ はルールに基づいて節境界を判定しているため、音声認識結果に対しては認識誤りの存在により精度が低下することが知られている。

3. 文境界での単語の出現頻度を考慮した特徴空間に基づく文境界検出

文境界検出において SVM を用いる手法が提案されている²⁾⁻⁵⁾。これらは SVM ベースのテキストチャンカ YamCha^{*1} を用いて文境界検出をテキストチャンキングの問題として扱っている。YamCha は学習データ・評価データ共に文字列を直接与えることができるといった利点がある反面、使用可能なカーネルが多項式カーネルに限定されている。本章では、SVM に与える素性を特徴空間上で扱う際、文境界検出に適した特徴空間へのプロット方法とその評価について述べる。

表 2 文境界に出現しやすい語

Table 2 High frequency word before sentence boundary

出現しやすさ	表層表現	読み	品詞情報
1 位	ます	マス	助動詞/終止形
2 位	た	タ	助詞/終助詞
3 位	です	デス	動詞/終止形
4 位	ね	ネ	助詞/格助詞
5 位	か	カ	動詞/命令形

3.1 文境界検出に適した特徴空間の作成

SVM は与えられた学習データの中からクラス境界近傍に位置するサポートベクターと識別面の距離であるマージンが最大となるようにクラス判別するための分離超平面を構築し判別を行う。

SVM に与える素性を特徴空間上で扱うには、与える素性を数値化する必要がある。文境界直前の語の出現頻度を基にした特徴空間上に、与えられた素性をプロットすることを考える。まず、文境界直前に出現しやすい語と出現しにくい語を分類した特徴空間を作成する。次に、作成した特徴空間に SVM に与える素性をプロットする。すると、文境界直前に出現しやすい語と出現しにくい語の距離が離れる。これにより、文境界検出に適した特徴量が得られると考えられる。文境界直前に出現しやすい「表層表現」「読み」「品詞情報」それぞれの一例を表 2 に示す。ここでは文境界検出に適した特徴空間を作成するための手順について述べる。

提案手法では、SVM に与える素性の「表層表現」「読み」「品詞情報」のそれぞれを特徴空間上で扱うために数値と対応させた辞書の作成を行う。ただし、「sil」「sp」については、「読み」「品詞情報」をそれぞれ「sil」「sp」とする。つまり、それぞれ『(sil)+sil+sil』、『(sp)+sp+sp』となる。一方、フィラーについては「表層表現」「読み」「品詞情報」を全て Filler として扱う。つまり、『えーっと+エーっと+フィラー』や『あー+あー+フィラー』などは全て『Filler+Filler+Filler』に統一される。このような制約を考慮に入れた上で、辞書作成の手順を以下に示す。

- (1) 各形態素を「表層表現」「読み」「品詞情報」に分割
- (2) 「表層表現」「読み」「品詞情報」ごとに文境界直前における出現率を算出
- (3) sil, sp を 1 位, 2 位とした後 2 で求めた確率を基に昇順に並べ替えランキングを作成
- (4) 3 で作成したランキングに順に番号を付与
- (5) 4 で付与した番号を 0~1 の値になるように正規化

*1 <http://chasen.org/taku/software/yamcha/>

表 3 表層表現の数値化例
Table 3 Example of numbering morphemes

表層表現	数値
sil	0.000000
sp	0.000070
常連	0.000139
踏み板	0.000209
...	...
です	0.999861
た	0.999930
ます	1.000000

手順(2)において、文境界直前に出現する語 w に対して求める出現率を $p(w)$ 、文境界数を N 、文境界直前における出現回数を $g(w)$ とすると出現率 $p(w)$ は以下のようにして求めることができる。

$$p(w) = \frac{g(w)}{N} \quad (1)$$

表 3 は手順に従って作成された「表層表現」と数値を対応させた辞書の例である。「読み」「品詞情報」についても「表層表現」と同様に手順に従って表 3 のような辞書を作成する。学習データ、評価データの形態素列に対して作成した辞書を基に各形態素の「表層表現」「読み」「品詞情報」をそれぞれ数値化する。これにより、得られる 3 次元の特徴量の例を以下に示す。

例)

た+タ+助動詞/終止形	(0.999930, 0.999929, 1.000000)
た+タ+助動詞/連体形	(0.999930, 0.999929, 0.695652)
か+カ+助詞/終助詞	(0.999721, 0.999715, 0.956522)
か+カ+助詞/副助詞	(0.999721, 0.999715, 0.760870)
ください+クダサイ+動詞/命令形	(0.999303, 0.999288, 0.978261)
だせ+ダセ+動詞/命令形	(0.361686, 0.044196, 0.978261)

3.2 評価実験

本実験では、CSJ 公開版の音声認識テストセット 30 講演(学会講演 20 講演, 模擬講演 10 講演)を評価に用い、これらを除くコア 168 講演(学会講演 64 講演, 模擬講演 104 講演)を学習セットとした。なお、単語正解精度は 30 講演の平均で 62.88%である。SVM に与える素性は、現在注目している形態素とその前後 3 形態素の「表層表現」「読み」「品詞

表 4 文境界検出精度

Table 4 Accuracy of sentence boundary detection

対象	手法	文境界検出精度 (F 値)
書き起こし	従来	0.958
	提案	0.835
音声認識結果	従来	0.759
	提案	0.751

情報」を提案手法に従ってそれぞれ数値化した 21 次元の特徴量に文献 3) や文献 5) 等で使用される各講演で正規化したポーズ長を加えた計 22 次元の特徴量である。現在注目している形態素が文頭に位置する場合はポジティブのラベルを付与し、その他はネガティブのラベルを与えた。また、従来手法としては、文献 5) の YamCha を用いた文境界検出を挙げる。SVM の設定は、とりあげた従来手法と同じ 3 次の多項式カーネルに設定した。結果を表 4 に示す。表 4 を見ると音声認識結果に対しては同程度の精度となっているが、書き起こしに対しては従来よりも精度が低下している。この要因として、辞書に登録されていない語の影響が考えられる。なぜなら、作成した辞書は学習データに出現した語を対象としており、評価データにて初めて出現する語は対応する数値が存在しないため適切な値を与えることができないからである。このような箇所はテストセットの全形態素 78653 個の内「表層表現」では 1946 箇所、「読み」では 2031 箇所、「品詞情報」では 0 箇所である。このような箇所には、中間値である 0.500000 という値を使用している。

今回作成した辞書は各語の値がほぼ等間隔(表層表現においてはおよそ 0.000070)となっている。しかし、文境界直前に出現する語の出現率にはばらつきがある。例えば、表層表現においては今回用いた学習データ内では文中を含めて全部で 14356 種類あるが、文境界直前に 1 回以上出現した語は 196 種類である。また、「ます」「た」「です」「ね」の 4 種類で 80%以上の出現率を占める。出現率に応じて、間隔が広くなるように重み付けを行うことで文境界直前で出現しやすい語と出現しにくい語が明確に分かれるため、より高精度な文境界検出が可能になるのではないかと考えられる。また、今回の実験では文境界直前での出現率にのみ注目していたが、文境界直前に出現する語であっても文中で出現しやすい語であった場合には文境界検出精度が低下してしまう可能性がある。そのため、この問題を解決するには文中での語の出現率や文末での語の出現率を素性として加えるといった方法が考えられる。

表 5 各データ
Table 5 Each data

	学習データ		評価データ	
	全体	ポーズ出現位置	全体	ポーズ出現位置
形態素数	484827	52858	78653	8332
絶対境界数	10446	8858	1796	1582

4. 単語信頼度を利用した文境界検出

本章では SVM に与える文境界検出のためのパラメータとして、発話末の単語信頼度を利用することについて述べる。ここで述べる単語信頼度とは、大語彙連続音声認識エンジン Julius^{*1} を使用して音声認識を行う際に音声認識結果と共に出力される単語信頼度のことである。

表 5 に全体を対象とした場合と 200ms 以上のポーズ出現位置を対象とした場合の書き起こしのデータ数を示す。表 5 より全体を対象とした場合と比較するとポーズ出現位置のみでは形態素数が約 10% にまで削減されているにもかかわらず、絶対境界数は全体の約 90% が 200ms 以上のポーズと共に出現することがわかる。このことより、絶対境界検出の高精度化を目指すにあたって重要であることは、200ms 以上のポーズと共に出現する絶対境界を高精度に検出することであると考えられる。よって、本章では 200ms 以上のポーズ出現位置を対象とする。また、200ms 以上のポーズを伴わない絶対境界に関しては今後の課題とする。

4.1 単語信頼度

単語信頼度とは、Julius を用いた音声認識における 2 パス探索アルゴリズムにおける高速な単語事後確率に基づく信頼度である⁸⁾。この単語信頼度はシステムがどれだけの確信を持って結果を出力したかを示す尺度と捉えることができ、この確信度を考慮することで誤認識の問題を緩和できると考えられている。このような確信度を応用したシステムとしては、音声対話システムにおけるタスク外発話の検証⁹⁾、対話管理¹⁰⁾などが挙げられる。以下、文献 8) に沿って単語信頼度の算出方法を説明する。

単語信頼度の計算は第 2 パス探索中に行われるが、その計算を行うにあたって第 1 パスの探索で求めた尤度を使用する。第 1 パスの展開時に木構造化辞書を用いたビーム探索を

*1 <http://julius.sourceforge.jp/>

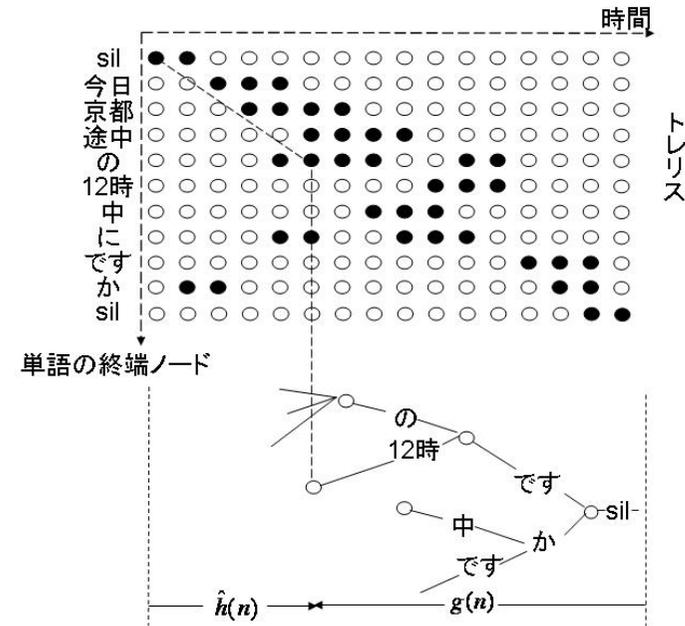


図 1 単語トレリスを用いたトレートリス探索
Fig. 1 A tree-trellis search based on word trellis

行う。各入力フレームにおいてビーム幅内に終端単語が残った単語について、その始端・終端時刻と入力先頭からの累積尤度を保存する。これを単語トレリスという。第 2 パス探索時には、この単語トレリスを展開単語の絞り込みと未探索部分の推定スコアとして用いる。

図 1 に単語トレリスと第 2 パスの探索の様子を示す。図の上部は第 1 パスの探索で得た単語トレリスを表しており、図中の黒丸は各時刻において終端ノードの残った単語である。図の下部は第 2 パスの探索の様子を表している。第 2 パスは第 1 パスとは逆向きに探索を行う。第 2 パスは探索時に部分文仮説の最終単語に対応するトレリス上の単語を参照し、その尤度を未探索部分のスコアとして用いる。また、仮説展開時の次単語集合は、該当時刻において単語トレリス上に存在する単語のみに限定される。第 2 パスにおける具体的な探索方法として、まず、ある部分文仮説 $w_1^{n-1} = w_1, w_2, \dots, w_{n-1}$ に対して単語 w_n を新たに接続することを考える。時刻 t における第 2 パスの部分文仮説 w_1^{n-1} の先頭部分の尤度を $g(w_1^{n-1})$ 、時刻 t において接続する単語 w_n の単語トレリス上での尤度を $\hat{h}(w_n, t)$ で表す

と、単語 w_n を部分文仮説に接続した際の新たな仮説のスコア $f(w_1^n)$ は以下のようにして求めることができる。

$$f(w_1^{n-1}, [w_n; \tau, t]) = g(w_1^{n-1}, t) + \hat{h}(w_n, t) \quad (2)$$

$$f(w_1^n) = \max_{0 \leq t < T} f(w_1^{n-1}, [w_n; t]) \quad (3)$$

式(2)式(3)を用いて近似的な事後確率 $\hat{p}(w_n|X)$ を計算すると以下ようになる。

$$W_c = [w; \tau, t] : \tau \leq t_n \leq t \quad (4)$$

$$\hat{p}(w_n|X) = \frac{e^{f(w_1^n)}}{\sum_{W_c} e^{f(w_1^{n-1}, [w; \tau, t])}} \quad (5)$$

ただし、 t_n は式(3)において最大値をとる時刻 t であり、 $[w; \tau, t]$ はある単語仮説 w が入力フレーム τ から t に存在するときの単語仮説、 W_c は展開単語 w_n と同じフレーム上に展開される全てのトリス上の展開候補単語、 X は入力音声系列である。この近似により、探索過程において単語事後確率を用いた単語信頼度の計算を行う。

4.2 発話末の単語信頼度を利用した文境界検出

CSJ を用いて学習した講演音声認識のための標準的な言語モデル(以下、ベースモデル)の学習データは 1000ms 以上のポーズで分割されて 1 つの発話となっている。これに対して、絶対境界で分割された学習データについて考える。絶対境界で分割を行うと発話末が文末表現となり、これにより作成した言語モデル(以下、絶対境界モデル)を用いて音声認識を行うと言語的な制約から発話末が文末表現に認識されやすくなる。そのため、発話末が文末表現のときは認識精度の向上が見込まれ、それに伴って文境界検出精度の向上も見込まれる。前章と同じ条件で 200ms 以上のポーズ出現位置について文境界検出を行った。音声認識エンジンは Julius を用い、入力音声は予め 200ms 以上のポーズ出現箇所を分割したものを使用した。また、言語モデルの学習データは前章で使用した 168 講演を用いている。全体の単語正解精度は 30 講演の平均で 61.66% であり 3 章で述べたベースモデルにおける単語正解精度と同程度の精度である。文境界検出結果を表 6 に示す。

表 6 より絶対境界モデルを用いることで再現率は上昇するが適合率が大きく低下した。これは、発話末が文末以外の場合にも文末表現に誤認識しやすかったためである。

前節で述べた単語信頼度は式(5)から展開候補となる単語が多ければ値は 0 に近づきやすく、展開候補となる単語が少なければ 1 に近づきやすいということがわかる。そこで、絶対境界モデルを使用した際の単語信頼度を利用することを考える。絶対境界モデルは前述のように発話末が文末表現になりやすいという特徴から発話末の単語信頼度に影響を与え

表 6 ベースモデルと絶対境界モデルの認識結果を用いた文境界検出精度

Table 6 Accuracy of detecting sentence boundary with ASR output by base model and absolute boundary model

対象	再現率	適合率	F 値
書き起こし	88.9%	93.7%	0.913
ベースモデル	85.3%	86.7%	0.860
絶対境界モデル	89.6%	57.6%	0.702

表 7 単語信頼度のみを用いた文境界検出精度

Table 7 Accuracy of detecting sentence boundary with only CMscore

使用モデル	再現率	適合率	F 値
ベースモデル	32.4%	87.7%	0.473
絶対境界モデル	68.0%	89.1%	0.771
両方	69.7%	91.6%	0.792

ると考えられる。つまり、元々発話末が文末表現の箇所に対しては絶対境界モデルの特徴に合致しているため、展開候補単語数は少なくなり高い単語信頼度が得られると考えられる。発話末が文末以外の箇所では音響的に大きく離れるものであれば展開候補単語が多くなり単語信頼度は低下すると考えられる。このような特徴を利用して文境界検出を行う。

4.3 評価実験

本実験では、前章と同じ CSJ テストセット 30 講演を評価データとし、168 講演を学習データとした。SVM のカーネルも前章と同じ 3 次の多項式カーネルを用いた。まず、単語信頼度のみによる予備実験を行った。ベースモデルを用いた際の単語信頼度(1次元)、絶対境界モデルを用いた際の単語信頼度(1次元)、両方(2次元)の 3 種類で実験を行った。結果を表 7 に示す。表 7 より絶対境界モデルを用いることでより高い精度の文境界検出が行えることがわかるが、前後 3 形態素と各講演で正規化したポーズ長を素性とした従来手法(表 6 のベースモデル)と比較すると単語信頼度単独の場合、適合率が高くなるが再現率、F 値共に精度が不十分である。そのため、従来用いられてきた素性と組み合わせる必要がある。つまり、ポーズ出現箇所の前後 3 形態素と各講演で正規化したポーズ長、単語信頼度を素性として用いるということである。ベースモデルの認識結果のそれぞれに対する結果を表 8 に示す。表 8 を見ると F 値が最も変化したもので上昇幅は 0.006 であり、単語信頼度を加えることによりわずかではあるが改善した。わずかな改善しか見られなかった要因としては、形態素情報とポーズの組み合わせで検出可能な文境界と単語信頼度で検出可能な文

表 8 単語信頼度を利用した文境界検出精度
Table 8 Accuracy of detecting sentence boundary with CMscore

対象	素性	再現率	適合率	F 値
認識結果	形態素・ポーズ長	85.3%(1349/1582)	86.7%(1349/1556)	0.860
	+ 単語信頼度 (ベースモデル)	85.6%(1354/1582)	87.1%(1354/1555)	0.863
	+ 単語信頼度 (絶対境界モデル)	85.8%(1357/1582)	87.4%(1357/1552)	0.866
	+ 単語信頼度 (両方)	85.5%(1353/1582)	87.2%(1353/1551)	0.864

境界がほとんど一致しており、検出できなかった文境界もほぼ一致していたためだと考えられる。なお、単語信頼度単体では高い適合率を示したため、さらに新たな素性との組み合わせや複数の識別器による多数決などを用いることで文境界検出精度のさらなる改善に期待できると考えられる。

5. おわりに

本稿では、文境界検出に用いる SVM の素性を文境界検出に適した特徴空間にプロットする手法と発話末の単語信頼度に着目し、文境界検出に利用する方法を提案・評価した。CSJ の講演音声を用いた評価実験において文境界検出に適した特徴空間に素性をプロットする手法では従来法とほぼ同精度であり、単語信頼度を用いた手法ではわずかながらではあるが改善が見られた。今後の課題としては、ポーズを伴わない文境界の検出法の検討、特徴空間を等間隔の数値化ではなく文境界直前での出現頻度によって間隔を変動させること、発話末の単語信頼度を複数の識別器による多数決に用いることなどが挙げられる。

参 考 文 献

- 1) 中嶋秀治, 山本博: 音声認識過程での発話分割のための統計的言語モデル, 情報処理学会論文誌, Vol.42, No.11, pp.2681-2688 (2001).
- 2) Akita, Y., Saikou, M. and Kawahara, T.: Sentence Boundary Detection of Spontaneous Japanese using Statistical Language Model and Support Vector Machines, In *Proc. ICSLP* (2006)
- 3) 下岡和也, 内元清貴, 河原達也, 井佐原均: 日本語話し言葉の係り受け解析と文境界推定の相互作用による高精度化, 自然言語処理, Vol.12, No.3, pp.3-17 (2005).
- 4) 西光雅弘, 河原達也, 高梨克也: 隣接文節間の係り受け情報に着目した話し言葉のチャンキングの評価, 情報処理学会研究報告, SLP-61, pp.19-24 (2006).
- 5) 秋田祐哉, 尾嶋憲治, 河原達也: 隣接文節間の係り受けと韻律を用いた SVM による話し言葉の節・文境界推定, 日本音響学会秋季研究発表会講演論文集 (2007).

- 6) 高梨克也, 丸山岳彦, 内元清貴, 井佐原均: 話し言葉の文境界-CSJ コーパスにおける文境界の定義と半自動認定-, 言語処理学会第 9 回年次大会, pp.521-524 (2003).
- 7) 丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝: 日本語節境界プログラム cbap の開発と評価, 自然言語処理, Vol.11, No.3, pp.39-68 (2004).
- 8) 李晃伸, 河原達也, 鹿野清宏: 2 パス探索アルゴリズムにおける高速な単語事後確率に基づく信頼度算出法, 情報処理学会研究報告, SLP-49, pp.281-286 (2003).
- 9) Ananth Sankar and Su-Lin Wu.: Utterance verification based on statistics of phone-level confidence scores, In *Proc. ICASSP*, Vol.1, pp.584-587 (2003).
- 10) 駒谷和範, 河原達也: 音声認識結果の信頼度を用いた効率的な確認・誘導を行う対話管理, 情報処理学会論文誌, Vol.43, No.10, pp.3078-3086 (2002).