音声言語情報処理 29－30
(1999. 12. 21)

認識結果の正解確率に基づく信頼度とリジェクション

北岡教英 †‡ 赤堀一郎 † 中川聖一 ‡
† (株) デンソー
〒 448-8661 愛知県刈谷市昭和町 1–1
‡ 豊橋技術科学大学 情報工学系
〒 441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1
kitaoka@jo1.denso.co.jp

あらまし 音声認識の認識結果の信頼度に応じた対話戦略を用いる音声対話システムのために、認識結果に正解確率の意味付けのある信頼度を設けることを考えた。認識結果に付随する特徴量に対して正解確率を表現する関数を仮定し、正解 (1) と誤認識 (0) との自乗誤差を評価関数としてそのパラメータを推定する。認識単語と音節連接モデルとの尤度比、単語内の音節継続時間の分散という二つの特徴量に対して、正解確率をシグモイド関数で表現できた。また、これら二つの組み合わせに対しても正解確率を表現できた。正解と誤認識の分離度をリジェクション実験で評価した結果、二つの特徴量を単独で用いるよりも組み合わせた方が分離度が高いことが示された。

キーワード 正解確率 信頼度 尤度比 継続時間の分散 リジェクション

Confidence Measure and Rejection based on Correct Probability of Recognition Candidate

Norihide KITAOKA†‡ Ichiro AKAHORI† Seiichi NAKAGAWA‡
†DENSO CORPORATON
1-1, Showa-cho, Kariya-shi, Aichi, 448-8661, Japan
‡Department of Information and Computer Sciences, Toyohashi University of Technology
1-1, Hibarigaoka, Tenpaku-cho, Toyohashi-shi, Aichi, 441-8580, Japan
kitaoka@jo1.denso.co.jp

Abstract We attempted to make a confidence measure describing correct probability to develop a dialog system which has dialog strategy depending on it. We assume that the correct probability led from a feature value of a recognition candidate follows a function of the feature, and then estimate the parameters of the function to minimize the sum of square errors from the correct(1) or incorrect(0) data. We used the likelihood ratio between the recognized word model and a concatenated syllable model and the variance of syllables' durations in the recognized word. The probability functions are well described as sigmoid functions. We could also estimate a function on the combination of two features. We evaluated the ability to separate the correct and incorrect candidates, the function based on the combination of two features is superior to that base on each feature.

keywords correct probability confidence measure likelihood ratio variance of syllable's durations rejection

1 はじめに

音声対話システムにおいて、よく問題とされるのは、システムの誤認識によって、システムとユーザの間で食い違いが生じ、対話が破綻してしまうことである。これを避けるために、誤認識を考慮に入れた対話戦略があるが、常に誤認識の可能性を考慮しながらの対話は、確認のための発話などが増えることにより、一般には冗長で煩わしいものとなる。

その問題に対して、信頼度を用いた対話戦略が考えられる。認識結果の信頼度に応じて、応答を変えることなどがある [1]。このような対話システムを構築する場合を考えた時、信頼度には直観的に理解できる意味付けがあることが望ましい。そこで、「正解確率」の意味をもつ信頼度を考える。正解確率が 95%の認識結果であれば、確信して対話を進めるが、50%ならば、確認を求めるなど誤認識の場合を考慮することが重要となろう。

本稿では、単語と音節連接モデルの尤度比や、単語内の音節の継続時間の分散に基づいて、認識結果の正解確率を得る方法を述べる。それぞれ単独に用いる場合と、両方を同時に用いた場合を考える。

信頼度はその閾値処理によって正解と誤認識がよく分離できる方が精度がよいといえる。そこで正解確率に閾値処理をして、誤認識のリジェクションを試みることで分離度を評価する。また、語彙外単語のリジェクションに応用した場合の精度も評価する。

2 正解確率に基づく信頼度

2.1 正解確率

ある認識結果 W が正解である場合を $C(W)=1$、誤認識である場合を $C(W)=0$ と表現する。また W に対するある特徴量 x の値を x_W とする。このとき、認識結果 W の特徴量 x の値が x_W であった場合に W が正解である確率、すなわち

$$p(C(W)=1|x=x_W)$$

を考える。この値は、特徴量 $x=x_W$ が得られた場合に結果 W がどの程度信頼できるかを直感的にわかりやすく表現できていると考える。

2.2 正解確率の表現とその推定

特徴量 x の値 x_W が得られた場合に、W の正解確率を知る方法を考える。一般的には、多くの正解/誤認識のサンプルから、あらゆる特徴量 x の値に対する正解確率を事前に調べておくことになる。しかし、x が連続値である場合、特定の値に対して多くのサンプルを得て確率を求めることは不可能である。

そこで、正解確率は x の関数 $f(x)$ に従っている、すなわち、

$$p(C(W)=1|x=x_W)\equiv f(x)$$

であると仮定する。この場合、正解確率を事前に推定する問題は、$f(x)$ のパラメータを推定する問題となる。

パラメータ推定に用いる認識結果のサンプルを $\{W_1, W_2, \ldots, W_N\}$ としたとき、それらに対応する特徴量を $\{x_{W_1}, x_{W_2}, \ldots, x_{W_N}\}$ とし、それらが正解か誤認識かを表現した列を $\{C(W_1), C(W_2), \ldots, C(W_N)\}\in\{0,1\}$ (1: 正解、0: 誤認識) とする。このとき、

$$E=\sum_{n=1}^{N}(C(W_n)-f(x_{W_n}))^2 \quad (1)$$

つまり、関数とサンプルの正解/誤認識の自乗誤差により、その関数を評価する。そして E を最小化するようにパラメータを推定する。

こうして推定された関数 $f(x)$ は、事後確率

$$p(C(W)=1|x=x_W) = \frac{p(x=x_W, C(W)=1)}{p(x=x_W, C(W)=1)+p(x=x_W, C(W)=0)} \quad (2)$$

を表すことになる。ここで

$$p(x=x_W, C(W)=1) = p(x=x_W|C(W)=1)\cdot p(C(W)=1) \quad (3)$$

$$p(x=x_W, C(W)=0) = p(x=x_W|C(W)=0)\cdot p(C(W)=0) \quad (4)$$

であり、正解と誤認識に関して正規分布を仮定して式 (3) 右辺および式 (4) 右辺の第 1 項の条件確率を求め、式 (3) 右辺および式 (4) 右辺の第 2 項にそれぞれ対応する認識率および誤認識率の事前確率を用いて、式 (2) によって事後確率すなわち正解確率を求める方法も考えられるが、本稿の方法ではその仮定を必要とせずに直接事後確率を求めることができる。

2.3 正解確率の推定例

大語彙単語認識実験結果に対して、正解確率の推定を行った。認識実験は大語彙認識システム [9] を用いた。タスクはカーナビゲーションシステムで、認識語彙数は、全国の地名・施設名約 18 万とカーナビゲーションシステム操作用のコマンド約 200 である。音声サンプルは自動車内で録音した地名やカーナビゲーションシステムのコマンドの発声 (1162 サンプル) である。認識率は 80.0%であった。

推定例 1: 尤度比に基づく正解確率

認識された単語の正解確率を求めることを考える。ある単語の音声認識結果の尤度と、別に用意した競合モデルの尤度との比（対数尤度における差）を信頼度の尺度とすることがよく行われる [2, 7]。尤度比は、正解と、誤認識もしくは語彙外単語発声の認識結果をよく分離できる尺度である。ここでは競合モデルとして、日本語の音節が自由に接続できる音節連接モデル [3] を用いた。すなわち、認識結果の対数尤度 l_w と、音節連接モデルの対数尤度 l_{sc}、単語の継続時間 T として、

$$x_W = (l_w - l_{sc})/T$$

のように、対数尤度の差を時間正規化したものとする。以降、これを LLR と呼ぶ。

図 1上図に、LLR 軸方向に 10 区間に分割した認識結果のヒストグラムを示す。各区間の正解数 (N_C)、誤認識数 (N_I) をそれぞれ実線、破線で示した。図 1下図の棒グラフは、各区間における正解確率 ($\frac{N_C}{N_C+N_I}$) である。2.2節における正解確率の関数推定は、この棒グラフを連続的に表現できる関数を推定することに対応する。この図から、LLR に対して単調増加な関数 $f(x)$ を仮定できると考えれる。

そこで、関数 $f(x)$ をシグモイド関数

$$f(x) = \frac{1}{1+\exp(-ax+b)} \qquad (5)$$

と仮定し、$f(x_n)$ と $C(W_n)$ との誤差を式 (1) として、これを最小化するパラメータ a, b を推定する。図 1下図に、$(x_{W_n}, C(W_n))$ を○で、推定された関数 $f(x)$ を実線でプロットした。関数によって棒グラフで示された値がおおよそ求められることがわかり、シグモイド関数による表現は適切であると考えられる。

これまでにも、LLR などの特徴量を、そのダイナミックレンジを小さく押えたり [6]、確率的に用いるため [7] に、シグモイド関数で変換する方法は提案されている。本稿ではパラメータ推定の評価関数として式 (1) を用いることによって、正解確率として意味付ける。

推定例 2: 音節継続時間の単語内分散に基づく正解確率

日本語では、短い時間範囲（例えば単語中）において、その中に含まれる音節の継続時間はほぼ等しい。つまり、単語中の音節の継続時間の分散は小さい。

しかし、一般的な HMM による認識においては、継続時間について考慮し難い。継続時間を考慮して認識精度を向上させる方法は多くあるが、一般には音節や音韻などのサブワード単位の最短・最長継続時間で打ち切ったり、継続時間の分布を事前に求めておいて、その確率

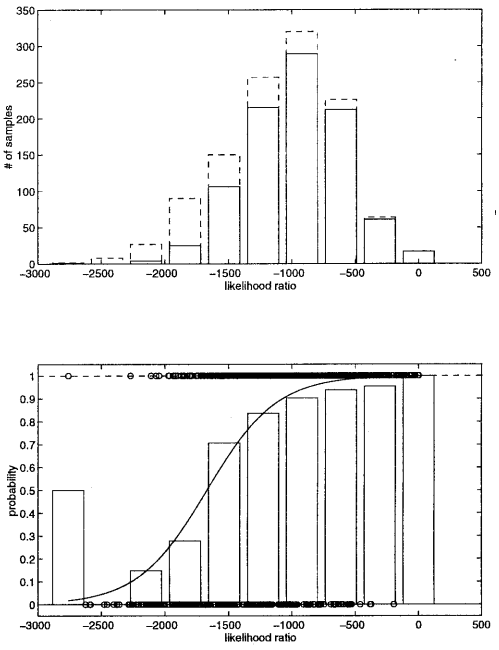

図 1. (上図)LLR に対する認識結果の分布 (実線: 正解, 破線: 誤認識) (下図) LLR に対する認識結果のプロット (正解: 1, 誤認識: 0) および正解確率とその関数表現

値を尤度計算時にマージするなどの方法であり、単語内などのサブワード間の継続時間の関係は考慮されない。そのため、モデルが部分的に伸縮することによって音声と比較的よくマッチングしてしまい、誤認識となることがある。

そこで、単語内の音節 $s_1, s_2, \ldots, s_N$ の継続時間 $l_1, l_2, \ldots, l_N$ としたときの

$$x_W = \sqrt{\frac{1}{N}\sum_{n=1}^{N} l_n^2 - (\frac{1}{N}\sum_{n=1}^{N} l_n)^2}$$

すなわち音節継続時間の単語内分散 (実際の特徴量としてはその平方根である標準偏差) を、正解と誤認識の分離の尺度とすることを考える。以降、これを VSD と呼ぶ。

認識結果の Viterbi パスを調べることによって VSD を求め、LLR の場合と同様に、VSD 軸方向に分割して、ヒストグラムと区間毎の正解確率を、それぞれ図 2上図および下図に、棒グラフで示した。

図 2下の棒グラフより、VSD に対して単調減少な関数 $f(x)$ を仮定できると考えられる。この場合にも式 (5) のシグモイド関数を仮定し、式 (1) で評価して a, b を推

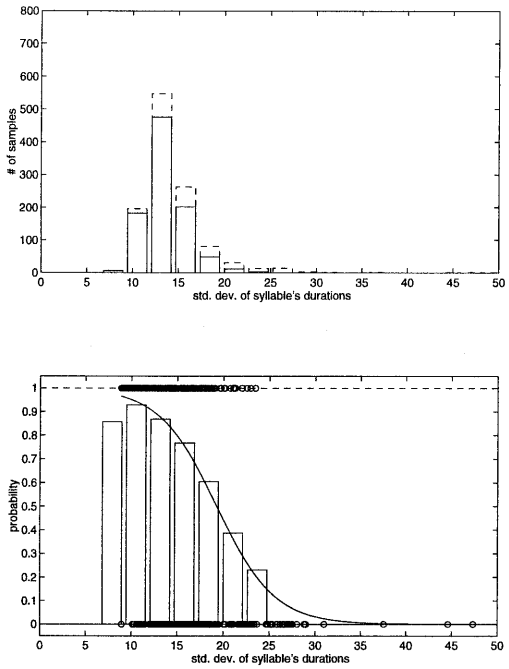

図 2. (上図)VSD に対する認識結果の分布 (実線: 正解, 破線: 誤認識) (下図) VSD に対する認識結果のプロット (正解: 1, 誤認識: 0) および正解確率とその関数近似

定した。図 2下図に、$(x_{W_n}, C(W_n))$ を○で、推定された関数 $f(x)$ を実線でプロットした。正解確率をよく表現できていると考えられる。

3 複数の特徴量を用いた正解確率

2節では正解確率を LLR もしくは VSD に基づいて定義したが、これら二つを同時に用いたほうが分離度が高くなることが考えられる。そこで、この二つの特徴量を用いて正解確率を定義することを考える。

二つの特徴量で張られる平面をメッシュで区切り、各メッシュにおける正解数および誤認識数を元に正解確率を求めてプロットしたものが図 3である。

この図を関数で表現することを考える。2 つの特徴量の値を x_1, x_2 としたとき、正解確率の関数として、

$$f(x_1, x_2) = \frac{1}{1 + \exp(g(x_1, x_2))} \qquad (6)$$

を仮定する。ここで、$g(x_1, x_2)$ として、次のような簡単な関数を用いた。

線形結合 $g(x_1, x_2) = a_1 x_1 + a_2 x_2 + a_3$

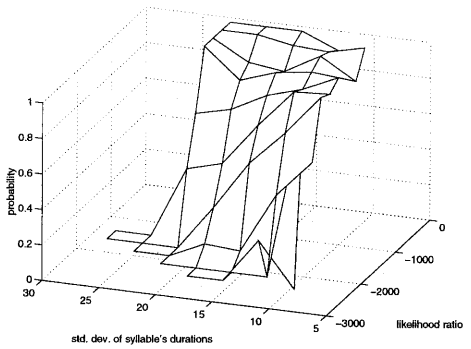

図 3. LLR と VSD に対する正解確率

双一次結合 $g(x_1, x_2) = a_1 x_1 x_2 + a_2 x_1 + a_3 x_2 + a_4$

これらの係数 a_k を式 (1) を最小化する基準で推定した。それぞれの場合について推定された曲面を図 4に示す。

認識結果が正解である場合、それぞれの単独の特徴量から得られる正解確率は両方高くなると考えられる。そこで、2 つの特徴量単独の正解確率 $f_1(x_1), f_2(x_2)$ の AND 条件である積で表せることが考えられる。ここでは、重み付き相乗平均

$$f(x_1, x_2) = f_1(x_1)^w \cdot f_2(x_2)^{1-w}$$

によって、二つの特徴量に基づく正解確率を表現することを試みる。$w = 0.5$ の場合を図 5に示す。複数の特徴量を組み合わせて信頼度を得る方法はいくつか提案されている [4, 5, 8] が、これらにおいても信頼度の積や最大値などの、AND 条件に近い結合を用いるほうが和を用いるよりもよい結果であるとされている。

以上の正解確率の表現法によって、正解確率をどの程度正確に表現できているかを知るため、式 (1) をサンプル数で割り、1 サンプルあたりの誤差を、各方法毎に求めた。結果を表 1に示す。相乗平均の重みが 1.0:0.0 もしくは 0.0:1.0 となっているものは、LLR もしくは VSD 単独の場合に相当する。両方を用いた場合のほうが単独より正確であり、この重みは等しい場合に最も正確であった。また、特徴量の線形結合や双一次結合を用いた場合の方が正確に表現できていることが分かる。

4 正解確率に基づくリジェクション

4.1 誤認識検出実験

複数の特徴量から正解確率という信頼度を得る方法を述べた。信頼度としては、正解と誤認識の分離度合い

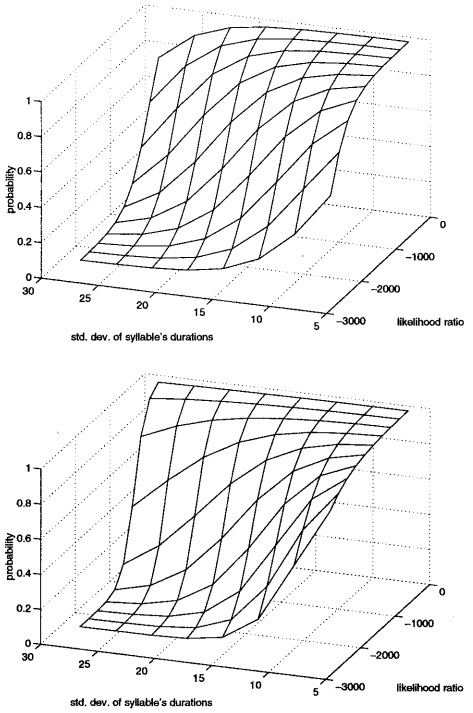

図 4. LLR と VSD の (上図) 線形結合、(下図) 双一次結合に基づく正解確率の表現

が高いものがよいと考えられる。そこで、本節では、これまでに定義した正解確率に閾値処理をすることによって、誤認識を検出しリジェクトする実験を行い、信頼度としての評価を行った。

サンプルは 2.3節と同条件で、推定用とは別に録音した 1655 発声である。認識率は 89.7%であった。

LLR と VSD を線形もしくは双一次に結合したものによる正解確率と、単独の特徴量による正解確率の重み付き相乗平均を閾値処理した。閾値を変化させた場合の、正解を誤ってリジェクトする率に対する誤認識のリジェクト率のプロットを図 6に示す。図中に示した比は、LLR と VSD による正解確率の重みを示す。プロットは、図中左上に近付くほど分離度が高いと考えられる。

LLR 単独 (1.0:0.0 に対応) と VSD(0.0:1.0 に対応) を比較した場合、LLR の方が性能がよい。しかし、それらの重み付け相乗平均を用いると、単独の場合よりもよい結果となることがわかる。同比率で乗ずるのが最もよい結果となった。

また、2 つの特徴量を線形結合もしくは双一次結合した場合にもよい結果であったが、それぞれを別々に求めておいて相乗平均を求めた場合と同程度に留まった。

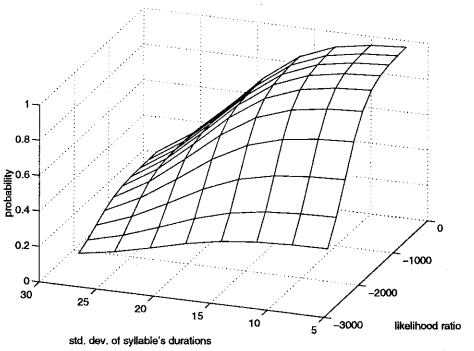

図 5. LLR と VSD による正解確率の相乗平均による正解確率の表現

表 1. 関数毎の正解確率との誤差 (1 サンプルあたり)

表現法		平均自乗誤差
	$LLR : VSD$	
重み付け相乗平均	0.0 : 1.0	0.1296
	0.25 : 0.75	0.1123
	0.5 : 0.5	0.1059
	0.75 : 0.25	0.1064
	1.0 : 0.0	0.1152
線形結合		0.0961
双一次結合		0.0957

4.2 語彙外単語発声のリジェクション実験

語彙外 (Out-of-Vocabulary; OOV) 単語を認識した場合、一般に語彙にある単語にマッチングして何らかの結果を返してしまう。LLR はこのような場合の分離に有効であるとされている [2]。また、音節継続時間も、誤認識時と同様に伸縮されている場合が多く、その分散は大きくなるため、VSD も分離に有効であると考えられる。そこで、3節の正解確率表現を用いて OOV 単語のリジェクション実験を行った。ただし、OOV 単語の検出と正解確率とは、直接の関係はない。

実験は、認識語彙からナビゲーションのコマンドの単語を除き、コマンド発声 (800 サンプル) を認識した結果に対して行った。閾値を変化させた場合の、正解を誤ってリジェクトする率 (4.1節の結果を利用) に対する語彙外単語のリジェクト率のプロットを図 7に示す。同様に、2 つの特徴量を組み合わせるとよい結果となった。

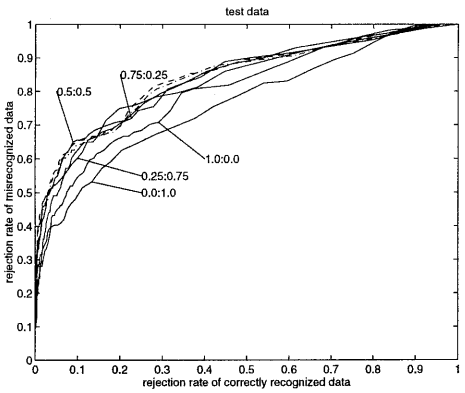

図 6. *LLR* と *VSD* に基づく正解確率を用いたリジェクションの結果
実線: 2 特徴量の正解確率の重み付き相乗平均 (*LLR*:*VSD*)
破線: 2 特徴量の線形結合に基づく正解確率
点破線: 2 特徴量の双一次結合に基づく正解確率

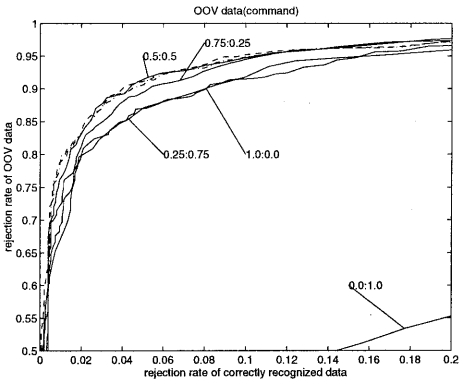

図 7. 語彙外単語を用いたリジェクション実験結果
コマンドを辞書から除外してコマンド発声を認識

5 まとめ

認識結果に対して正解確率の意味をもつ信頼度を得る方法を提案した。単語と音節連接モデルの尤度比 (*LLR*) に基づく場合、単語内の音節継続時間の分散 (*VSD*) に基づく場合、両方を用いる場合について、シグモイド関数を用いて正解確率を表現する方法を述べた。また、正解確率に基づく正解と誤認識の分離度をリジェクション実験によって評価した。また、語彙外単語のリジェクション実験も行った。その結果、複数の特徴量を本方法によって組み合わせると単独よりも精度がよくなることが確認できた。

今後は、より分離度を向上させるとともに、正解確率を用いたアプリケーションを開発する予定である。

参考文献

[1] 新美康永, 小林豊. "音声認識誤りを考慮した対話制御方式のモデル化", 情報処理学会研究報告, 95-SLP-5-7, 1995.

[2] R. A. Sukkar and C.-H. Lee. "Vocabulary independent discriminative utterance verihication for nonkeyword rejection in subword based speech recognition", IEEE Trans. on Speech and Audio Processing, Vol.4, No.6, pp.420-429, 1996.

[3] 加藤正治, 堀 貴明, 伊藤彰則, 好田正紀. "音素連接 HMM を用いた尤度正規化に基づくワードスポッティングの検討", 電子情報通信学会技術報告, SP97-77, pp.9-14, 1997.

[4] B. T. Tan, Y. Gu and T. Thomas. "Evaluation and implementation of a voice activated dialing system with utterance verification", ICSLP-98, pp.1671-1674, 1998.

[5] K. Kirchhoff and J. A. Bilmes. "Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values", IEEE ICASSP-98, pp.693-696, 1998.

[6] M.-W. Koo, C.-H. Lee and B.-H. Juang. "A new hibrid decoding algorithm for speech recognition and utterance verification", IEEE ICASSP-98, pp.213-216, 1998.

[7] 實廣貴敏, 高橋 敏, 相川清明. "対立音素間の尤度差に基づく信頼度尺度によるリジェクション", 電子情報通信学会技術報告, SP97-76, pp.1-7, 1997.

[8] E. Lleida and R. C. Rose. "Efficient decoding and training procedures for utterance verification in continuous speech recognition", IEEE ICASSP-96, pp.507-600, 1996.

[9] 赤堀一郎, 加藤利文, 北岡教英. "地名認識システムとその応用", 情報処理学会研究報告, 95-SLP-7-9, pp.55-60, 1995.