

TR-I-0248

HMM-LR 連続音声認識における適応的探索手法の検討
A Study on the Adaptive Search Methods in HMM-LR
Continuous Speech Recognition

岡本匡史[†] 山口耕市 嵯峨山茂樹 北研二

Tadashi OKAMOTO, Kouichi YAMAGUCHI, Shigeki SAGAYAMA and
Kenji KITA

概要

HMM-LR 連続音声認識において、現在用いられているビームサーチは最適性が保証されていない上、高い認識率を得るためにはビーム幅を大きくしなければならず、多くの音素照合回数を必要とする。そこで今回、探索空間の縮小(=音素照合回数の削減)をめざし、動的なビームサーチ方法と、ビームサーチとベストファーストサーチを組み合わせた方法を開発し、認識率、音素照合回数の比較・検討を行った。前者は各深さでの尤度の分布をもとに、ビーム幅を変化させるものである。音素照合回数は1/2以下に削減できたが、若干の認識率低下を招いた。後者は最初はビームサーチを行い、ある程度候補が絞られてきた時点でベストファーストに切り替えるものである。ベストファーストサーチ単独に比べ認識率は少し改善されるが音素照合回数は増加した。

[†]豊橋技術科学大学
Toyohashi University of Technology
ATR 自動翻訳電話研究所
ATR Interpreting Telephony Research Labs.

1 はじめに

われわれは混合連続分布型 HMM を用いた高精度な音素モデルを HMM-LR 連続音声認識に組み込み、特定話者の文節認識実験を行い、良好な結果を得ている [1]。ところが現在用いているビームサーチは最適性が保証されていない上、高い認識率を得るためにはビーム幅を大きくしなければならず、多くの音素照合回数を必要としていた。そこで今回、探索空間の縮小 (= 音素照合回数の削減) をめざし、動的なビームサーチ方法と、ビームサーチとベストファーストサーチを組み合わせたサーチ方法 (以降ハイブリッド法という) を開発し、認識率、音素照合回数の比較・検討を行ったので報告する。

2 混合連続分布型 HMM-LR の概要

2.1 HMM-LR システムの概要

システム構成を図 1 に示す。文法には 1407 ルールからなる一般的文法を用いている。HMM-LR 法は文献 [3] に詳しく説明されているのでここでは省略する。

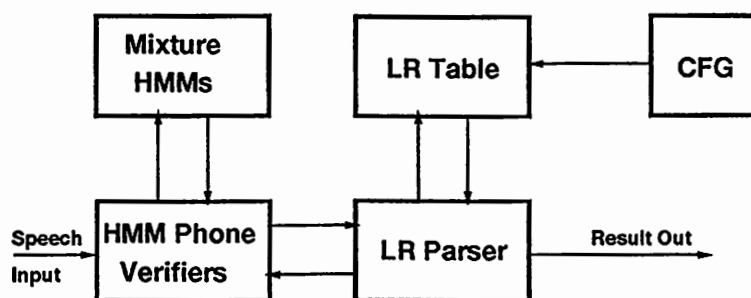


図 1: 混合連続分布型 HMM-LR

2.2 混合連続分布型 HMM

用いた連続出力分布 HMM は 4 状態 3 ループ、対角共分散行列の混合ガウス分布 HMM で、その構造を図 2 に示す。パラメータは、log power, cepstrum(16 次), Δ log power, Δ cepstrum(16 次) から成る 34 次元ベクトルで、分析条件は、サンプリング周波数 12kHz, 16bit 量子化, ハミング窓 (20ms), フレーム周期 5ms である。実験は男性 1 名 (MAU) の特定話者で行った。

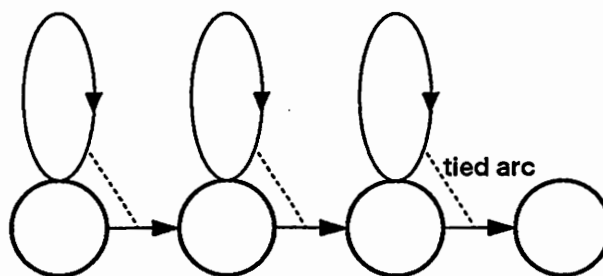


図 2: HMM の構造

HMM の学習は重要語 5240 単語および音素バランス 216 単語を用いた。

音素モデルの種類は花沢 [4] に準じた。混合数は学習サンプル数に応じて 3~15 とした。平均混合数は約 11 である。学習サンプル数が少ないと HMM の分散が極度に小さくなり、HMM の出現確率が不安定になることがある。これを回避するため、HMM の分散にフロアリングを施している。

3 各種探索手法の比較

3.1 ビームサーチ

HMM-LR システムでは従来から採用してきた探索手法である。横型探索ではあるがフレーム同期ではなく、音素同期である。ある部分仮説 n の評価値 $f(n)$ は、探索が進んでいる時点までの入力音声の対数尤度 $g_n(t)$ を入力フレー

ム数 t で正規化したもののうち、各音素の継続時間情報によって定められた照合区間 $[t_1, t_2]$ で最大のものを用いている。

$$f(n) = \max_{t_1 \leq t \leq t_2} \frac{g_n(t)}{t} \quad (1)$$

探索空間の大きさは音素照合回数であり、ビーム幅で決まる。現在のタスクでは継続時間制御なしのとき、600 くらいにしておけば最適解が枝刈りされることはほぼない。

3.2 ベストファースト探索法

評価値がもっとも高い部分仮説のノードを展開して先に進む。現在用いている連続 HMM では、確率密度関数を積分せずに出力値そのままを使っているため、各部分仮説の評価値を比較する場合、長い部分仮説の方が有利になる。その結果、盲目的な縦型探索になってしまう。そこでビームサーチのときと同様、評価値の正規化が必要となり、評価値 $f(n)$ は式 (1) と同じになる。音素認識が良好に動作しているとき、この探索法は極めて有効となる。逆に、入力音声の先頭付近で正解音素列の評価値が低いと、探索の効率が非常に悪くなる可能性がある。

3.3 可変ビームサーチ法

北方式 ビーム幅を動的に変化させるものには従来、探索の深さによってビーム幅を一意的に決めるという北氏的方式 [5] があった。これは深さに対する正解候補の順位の分布を調べ、深さ x のときビーム幅 y を次式によって近似したものである。

$$y = A \exp(B \times (x - C)^2) + D \quad (2)$$

文献 [5] のシステムにおいては、計算量が約 1.5 に削減された。現在のタスクではビーム幅を 600 一定にしたときの深さに対する正解候補の順位の分布は図 (3) のようになった。そこで $A = 450, B = 0.1, C = 5, D = 50$ として、認識実験を行った。

階段方式 図 (3) に示したように、ある程度の深さ以上探索が進んだ場合、正解仮説は上位にあることが多い。このことを利用し、ある深さまではビーム幅を一定にしておき、その後階段状にビーム幅を削減していくという方法もある [2]。図 (3) よりビーム幅 600 一定の場合は深さ 7 をピークにそれ以降は順位は良くなっている。そこで深さ 8 以降のビーム幅を表 1 のようにして認識実験を行った。

表 1: 階段方式のビーム幅

深さ	ビーム幅
1~ 7	600
8	300
9	250
10	200
11	150
12	150
13	100
14	100
15~	50

3.4 適応的ビームサーチ法

今回開発した動的ビームサーチ法は、探索の各深さにおいて、尤度の分布をもとにビーム幅を変化させるものである。ビーム幅の決定には回帰直線の利用と尤度の変化の利用の 2 種類を考えた。

回帰直線を利用する方は、各深さにおいて尤度を大きいものから N 個を回帰直線で近似を行い、その傾き、 y 切片と正解候補の順位の関係を調べる。回帰直線近似による直線の方程式は $(x_i, y_i) (1 \leq i \leq N)$ が与えられたとき次で与えられる。

$$y = \frac{S_{xy}}{S_x} \times x + (\bar{y} - \bar{x} \times \frac{S_{xy}}{S_x}) \quad (3)$$

$$S_x = \sum_{i=1}^N (x_i - \bar{x})^2 \quad (4)$$

$$S_{xy} = \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \quad (5)$$

$$\text{傾き: } a = \frac{S_{xy}}{S_x} \quad (6)$$

$$y\text{切片: } b = \bar{y} - a\bar{x} \quad (7)$$

$N = 10, 20$ としたときの傾き a 、 y 切片 b と正解候補の順位の関係は図 (4)~(7) のようになった。これらの図より y 切片 b と最適候補の順位の間には相関関係がみられなかったが、傾き a と正解候補の順位の場合は、傾きの絶対値が小さい程順位が悪くなるサンプルが多いという傾向が現れた。さらに $N = 10$ と $N = 20$ の場合の傾きと正解候補の順位の間を比べると、傾き a が -0.5 から 0 の間には同じような傾向になっていることがわかった。そこで $N = 10$ として傾きの大きさ a に対してビーム幅 W を次の式によって決定し、認識実験を行った。

$$W = 500a + E \quad (\text{ただし最低ビーム幅 } 50) \quad (8)$$

E は可変パラメータで、 $E = 550, 450, 350, 250$ のそれぞれの場合について実験を行った。(図 8 参照) なお図 (6) において、 $a = -3.00$ 付近で正解仮説が上位に集中している。これらは探索木の深さが 1 のとき、仮説数か 20 未満にもかかわらず $N = 20$ としたために生じた分布である。

一方、尤度の変化を利用する方は、各深さ毎に尤度の 1 階差分および 2 階差分を求め、正解候補の順位との関係を調べたが、ビーム幅を決定する十分な根拠が得られず、満足な結果を得ることはできなかった。

この他、第一候補の尤度との差がある一定値以上である候補を枝刈りする方式も考えられるが、今回の実習では検討していない。

3.5 ハイブリッド法

ベストファーストサーチは、音素認識がうまく働いているときは極めて有効であるが、入力音声の先頭付近で正解音素列の評価値が低いと、探索効率が非常に悪くなる。そこで今回は最初はビームサーチを行い、ある程度候補が絞られてきた時点でベストファーストに切り替えるという方法考えた。

適応的ビームサーチ法と同様に、尤度分布の傾きを考える。ここで探索が進につれ傾きの絶対値が大きくなるということは、候補が絞られてきたと推定して、傾きの絶対値が前段階より大きくなった時点でベストファーストに切り替えることにした。ただし、第 2 音素までは候補数自体が少ないため、正しく評価できないと考えビームサーチとした。ビーム幅は 600 で固定した。尤度分布の傾きは式 (6) を用い、回帰分析のサンプル数 $N = 10, 20, 30$ のそれぞれの場合について実験を行った。

4 文節音声認識実験

評価データは複合語を許さない文節発声データ (DSB3) の 279 文節である。ビーム幅 600 一定にした場合の認識結果を表 2 に示す。正解候補の枝刈りはほとんど生じなく、良好な認識率を得ている。しかし 1 文節当たりの音素照合回数は平均で 16000 以上と多くなっている。

表 2: ビームサーチの文節認識率 (ビーム幅:600)

順位	認識率 (%)
1 位	88.89
~ 2 位	97.13
~ 3 位	98.57
~ 4 位	98.92
~ 5 位	99.28
平均音素照合回数	16469.4

北方式および階段方式による可変ビームサーチの認識結果を表3に示す。両方式とも認識率は少し低下したものの、北方式では音素照合回数は1/2以下に削減され、階段方式でも10000程度になった。

表 3: 北方式、階段方式による文節認識率

順位	北方式 (%)	階段方式 (%)
1 位	87.10	87.81
～2 位	95.70	96.42
～3 位	97.13	97.85
～4 位	97.49	98.21
～5 位	97.85	98.57
平均音素照合回数	7153.9	10395.6

今回開発した適応的ビームサーチの認識結果を表4に示す。これもビーム幅600一定のものに比べて認識率が低下している。特にパラメータ $E = 250, 350$ の場合は実用的でない。 $E = 450$ の場合と階段方式を組み合わせたものは、認識率は若干低下したものの音素照合回数は1/2以下に削減されている。

表 4: 適応的ビームサーチによる文節認識率 (%)

順位	E=250	E=350	E=450	E=550	E=450 と階段方式との併用
1 位	82.08	85.66	87.81	87.46	87.46
～2 位	91.04	94.27	96.42	96.06	96.06
～3 位	92.47	95.70	97.85	97.49	97.49
～4 位	92.47	96.06	98.21	97.85	97.85
～5 位	92.83	96.42	98.57	98.21	98.21
平均音素照合回数	5410.7	8230.5	11097.7	13844.2	7625.1

次にベストファーストのみの場合と、ハイブリッド方式の場合の認識実験結果を表5に示す。ベストファーストのみの場合は語頭付近の認識が良好に動作しないと、最適解の展開が遅れ音素照合回数が増え、最悪の場合スタックから漏れ、刈られてしまう。このためビームサーチに比べ認識率が低くなっている。ハイブリッド方式は回帰分析のサンプル数 $N = 10, 20, 30$ のいずれの場合でも認識率はベストファーストのみの場合より高くなっているが、音素照合回数はかなり増えている。

なお今回の実験では計算時間の便宜上、ベストファースト探索における仮説の受理数 (N -best に相当) を30に制限している。このため認識率そのものはビームサーチに比べ低くなっている。

図(9)は本実験における4つの探索手法別に音素照合回数の分布を示したものである。ベストファーストサーチはビームサーチに比べて音素照合回数をかなり削減できている。ビームサーチは横型探索であるため、入力音声長に応じて一定数以上の音素照合を必要とする。これに対し、ベストファーストサーチの音素照合回数は入力音声長に直接は関係しない。音素認識性能に強く依存すると考えられる。個々に内容を調べてみると、ビームサーチに比べ1/10～1/100の音素照合で済んでいる場合もあった。よって正解候補の順位に関する情報を何らかの手段で得ることができれば、ベストファースト型の探索の利点を活用でき、処理の高速化が期待できる。とりわけ音素認識性能が高ければ高いほど、この高速化は顕著になると予想される。

表 5: ベストファーストによる文節認識率 (%)

順位	ベストファーストのみ	ハイブリッド方式		
		N=10	N=20	N=30
1 位	82.80	84.23	84.95	85.66
～2 位	90.68	92.83	93.91	94.27
～3 位	92.11	94.27	95.34	95.70
～4 位	92.11	94.27	95.70	96.06
～5 位	92.11	94.62	96.06	96.42
平均音素照合回数	2302.6	5005.4	5831.5	6674.7

5. おわりに

混合連続分布型 HMM-LR に適応的ビームサーチ、ビームサーチとベストファーストサーチを組み合わせたハイブリッド方式を導入し、文節認識実験を行い、認識性能、計算量について比較した。

適応的ビームサーチは音素照合数は $1/2$ 以下に削減できたが、若干の認識率の低下を招いた。

ハイブリッド方式では認識率はベストファーストのみの場合より高くなったが、ビームサーチの場合よりは低下している。音素照合回数もかなり増加し、動的ビームサーチとほぼ同等となり、ベストファーストの利点を生かすことができなかった。しかし、ベストファースト型の探索はビームサーチに比べて音素照合回数を 1～2 桁抑えられる場合もあり、このことがうまく利用できれば処理の高速化が期待できる。

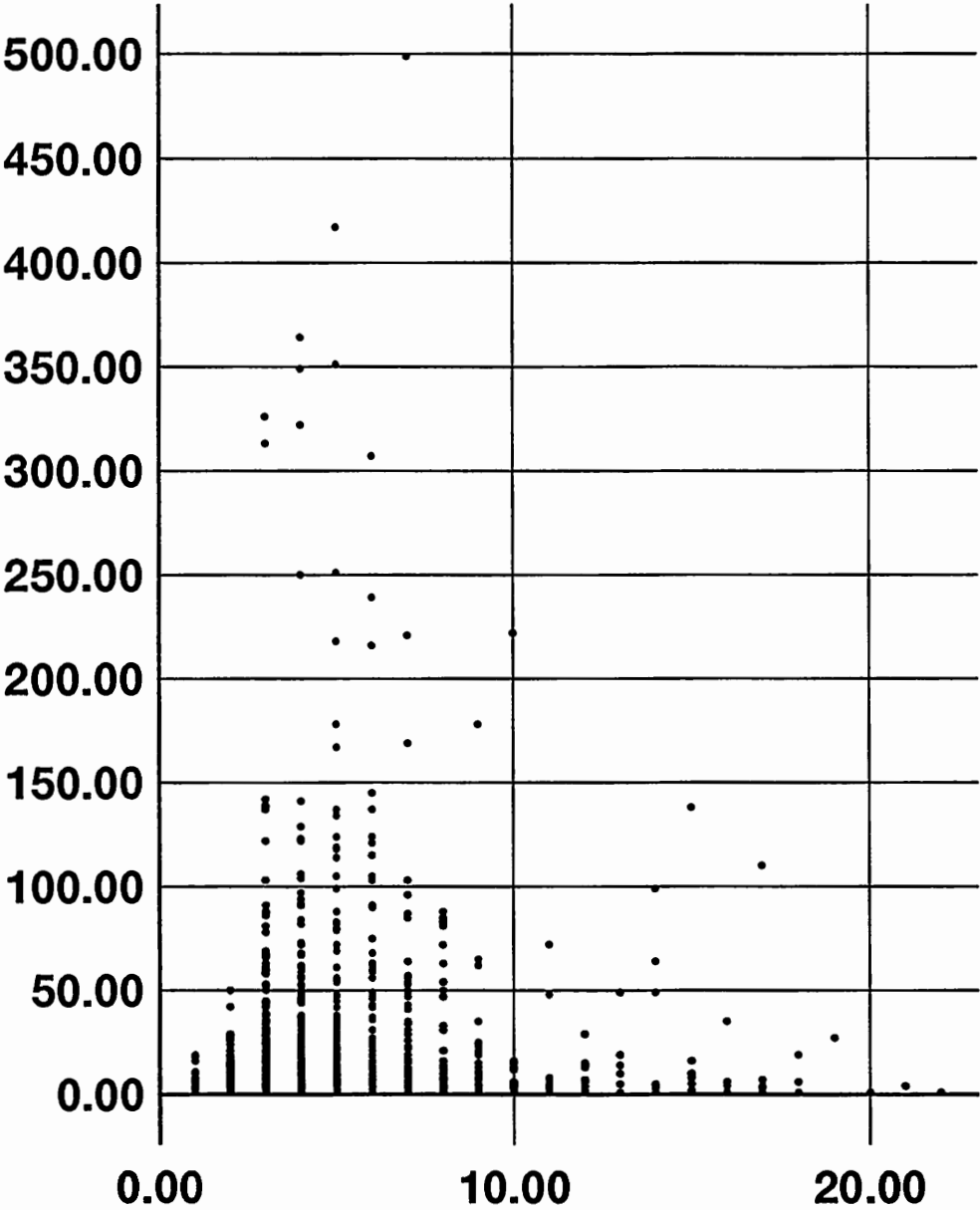
謝辞 研究の機会を与えて頂いた、ATR 自動翻訳電話研究所榎松明社長に感謝いたします。親切にご指導して下さいました山口耕市さんに感謝いたします。また、熱心に討論頂く ATR 研究員の皆様に感謝いたします。

参考文献

- [1] 山口、嵯峨山 : “混合連続分布型 HMM を用いた HMM-LR 連続音声認識”, 音講論集, 1-P-5, pp.113-114 (1992.3).
- [2] 山口、嵯峨山、北、F. K. Soong : “HMM-LR 連続音声認識における A* アルゴリズムを用いた探索手法の検討”, 音講論集, 3-1-9, pp.87-88 (1992.3).
- [3] 北、川端、斉藤 : “HMM 音韻認識と拡張 LR 構文解析法を用いた連続音声認識”, 情報処理学会論文誌, Vol.31, No.3, pp.472-479 (1990.3).
- [4] 花沢、北、中村、川端、鹿野 : “HMM-LR 音声認識システムの性能評価”, 音響学会論文誌, Vol.46, No.10, pp.817-823 (1990.10).
- [5] 北、川端、森元 : “HMM-LR 連続音声認識システムにおける計算量削減のための一検討”, 音講論集, 3-6-4, pp.77-78 (1989.3).

Fig.3 Distribution of Correct Hypotheses

Beam Width



Depth of Tree

図 3. 探索木の深さに対する正解仮説の分布

Fig.4 Distribution of Correct Hypotheses(N=10)
Beam Width

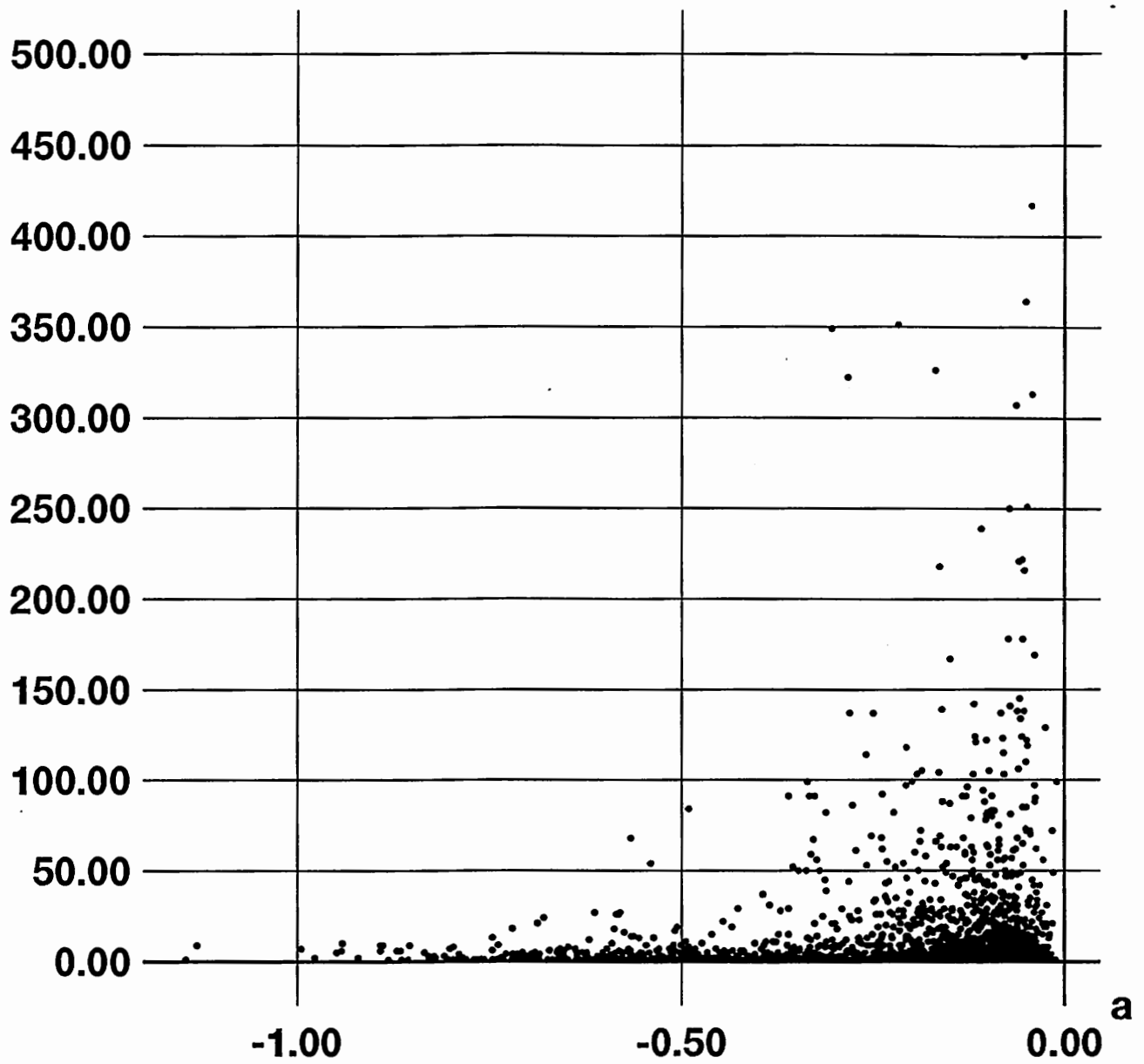


図4. 回帰係数 a に対する正解仮説の分布 ($N = 10$)

Fig.5 Distribution of Correct Hypotheses(N=10)

Beam Width

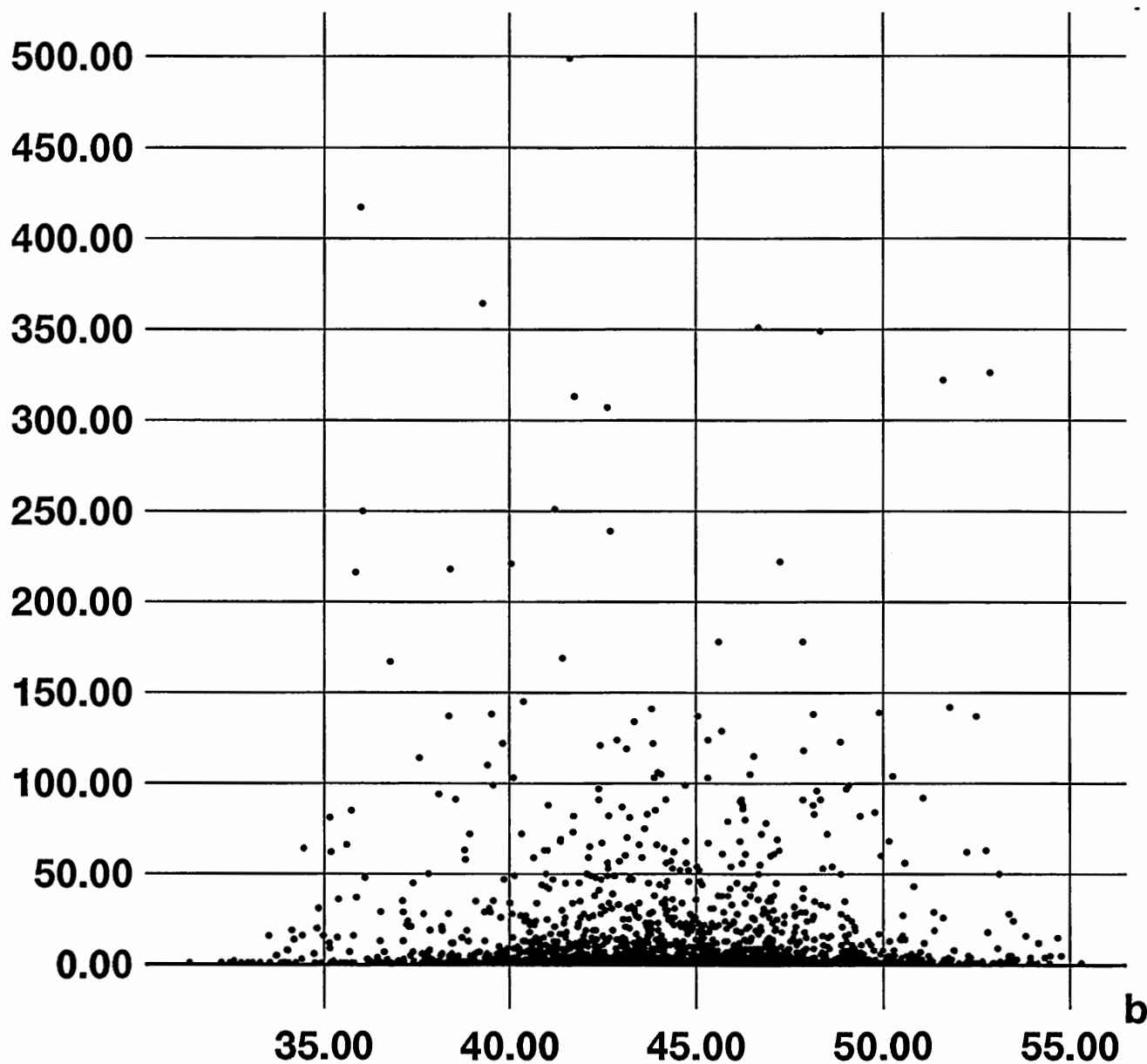


図 5. 回帰係数の y 切片 b に対する正解仮説の分布 ($N = 10$)

Fig.6 Distribution of Correct Hypotheses(N=20)
Beam Width

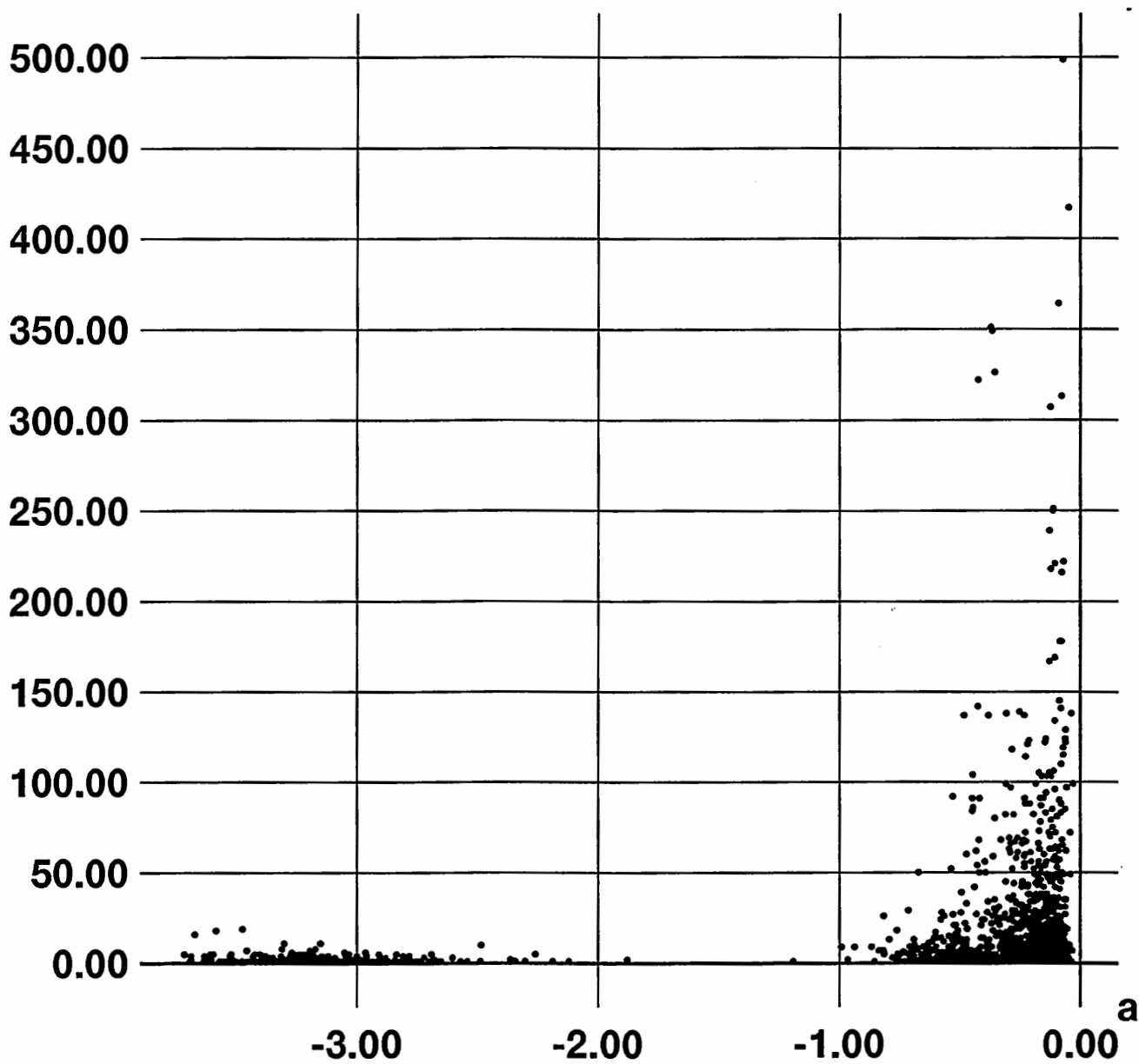


図 6. 回帰係数 a に対する正解仮説の分布 ($N = 20$)

Fig.7 Distribution of Correct Hypotheses(N=20)

Beam Width

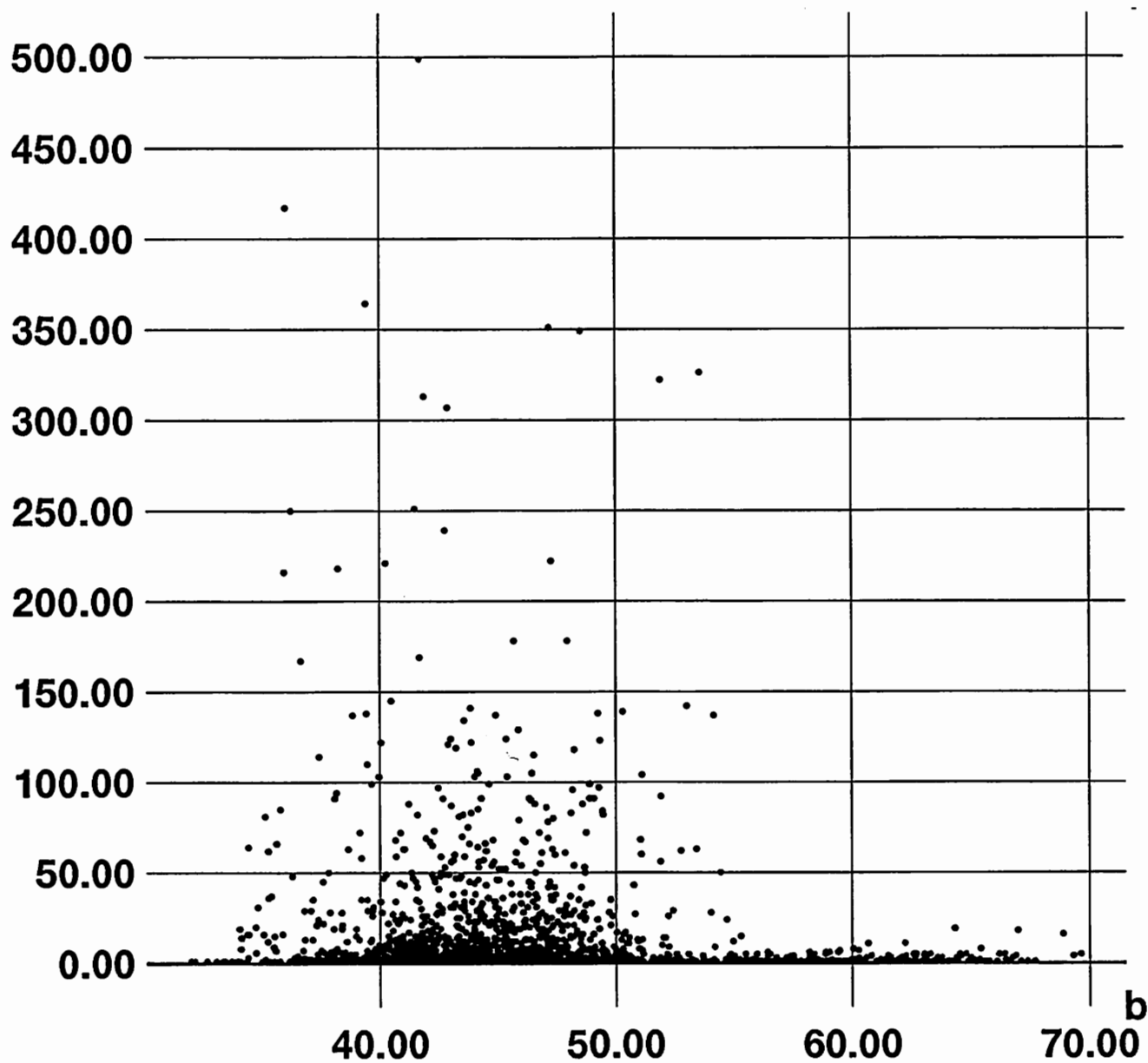


図 7. 回帰係数の y 切片 b に対する正解仮説の分布 ($N = 20$)

Fig.8 Distribution of Correct Hypotheses(N=10)
Beam Width

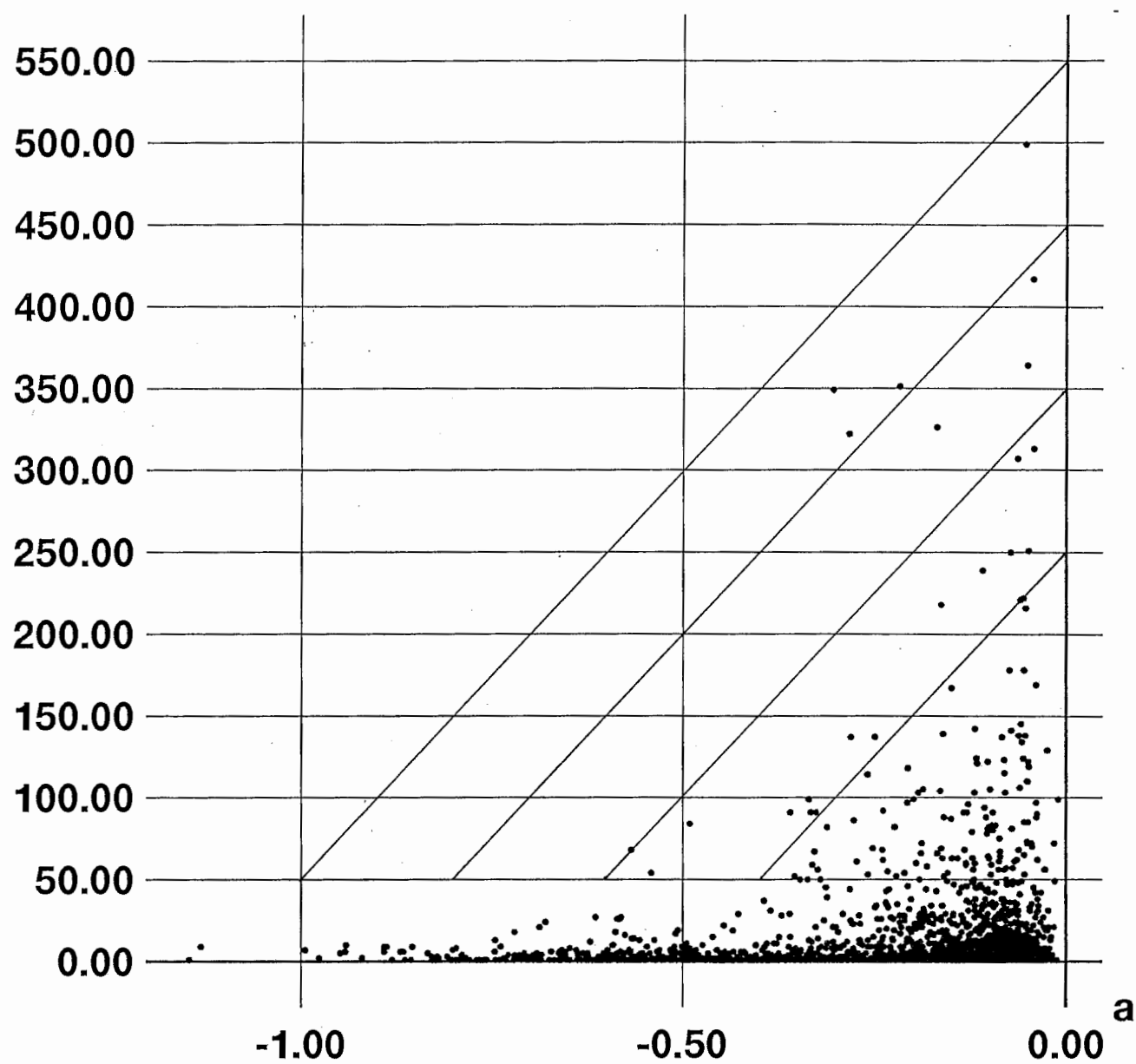


図8. 回帰係数 a に対する正解仮説の分布と適応ビーム幅
(図中、斜め線は上から順に $E = 550, 450, 350, 250$ としたときの適応ビーム幅を示す。)

Fig.9 Comparison of Verification Times

Frequency

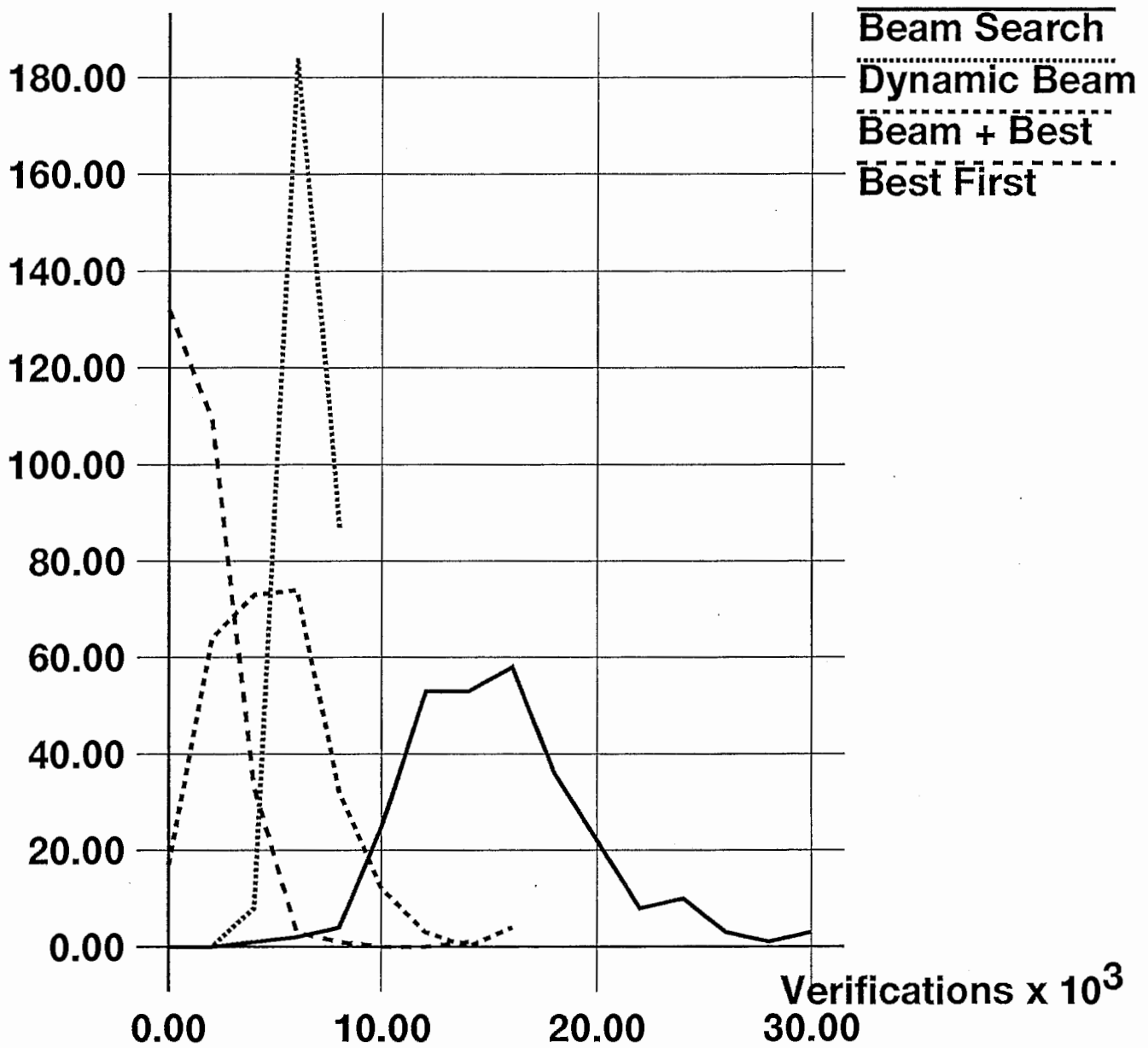


図 9. 各種探索法における音素照合回数の分布