TR-IT-0042

音素マトリックスの分析
Analysis of Phoneme Matrices

伊藤敏彦+　ローケン・キム　キュンホ　北川美宏++
Toshihiko ITOH　Kyung-ho LOKEN-KIM　Yoshihiro KITAGAWA

1994. 3. 11

概要

信頼性の高い音素とはどのような音素なのか、また信頼性の高い音素をたすためにはどのような方法を用いらば良いのかを調べるために音響マトリックスの分析をおこなった。まず音響マトリックスで認識スコアの高い音素が信頼性の高い音素か調べた。次に各音素ごとの音響マトリックスで認識スコアの島を出し、その認識スコアの島のデータ（投影面積、体積、最大高さ等）の大きな音素が信頼性の高い音素か調べた。最後に認識スコアの高い音素を集め作った音響マトリックスで同様に島を出し、島のデータの大きいものが信頼性の高い音素であるか調べた。

ATR音声翻訳通信研究所
ATR Interpreting Telecommunications Research Laboratories

+豊橋技術科学大学
++国際電気通信基礎技術研究所

目次

1　まえがき

信頼性の高い音素とはどのような音素なのか、また信頼性の高い音素を出すためにはどのような方法を用いれば良いのかを調べるために音響マトリックスの分析を行なった。まず音響マトリックスの認識Ｓｃｏｒｅの高い音素が信頼性の高い音素か調べた。次に各音素ごとの音響マトリックスで認識Ｓｃｏｒｅの島を出し、その認識Ｓｃｏｒｅの島のデータ（投影面積、体積、最大高さなど）の大きな音素が信頼性の高い音素か調べた。最後に認識Ｓｃｏｒｅの高い音素を集め作った音響マトリックスで同様に島を出し、島のデータの大きいものが信頼性の高い音素であるか調べた。

2　音素の島について

「もしもし」、「そちらは」などの音声認識用データの文節４０個それぞれで各音素ごとに認識Ｓｃｏｒｅをつける。つまり４０文節中のそれぞれの文節で、あるＦｒａｍｅ（Ｓｔａｒｔ－Ｆｒａｍｅ）からあるＦｒａｍｅ（Ｅｎｄ－Ｆｒａｍｅ）までの各音素の認識Ｓｃｏｒｅを発声開始時間から発声終了時間まで、Ｓｔａｒｔ－ＦｒａｍｅとＥｎｄ－Ｆｒａｍｅの組合せをいろいろ変化させつけていく。ただしＥｎｄ－ＦｒａｍｅとＳｔａｒｔ－Ｆｒａｍｅの差は最大５０Ｆｒａｍｅ、最小５Ｆｒａｍｅとする。ここでX軸、Y軸にＳｔａｒｔ－Ｆｒａｍｅ、Ｅｎｄ－Ｆｒａｍｅを取れば、各音素の認識Ｓｃｏｒｅは、マトリックス状に並ぶ。このマトリックス状になった認識Ｓｃｏｒｅの値をＺ軸としてとる。これを音響マトリックスとする。そうして認識Ｓｃｏｒｅは正負の値を取っているので、認識Ｓｃｏｒｅの０以上の部分だけを取り出す。そうした時、この音響マトリックスで認識Ｓｃｏｒｅが０より大きい位置が隣あっている（つまり音響マトリックス上で、ある認識Ｓｃｏｒｅが０より大きい場所の上、右上、右、右下、下、左下、左、左上の位置に同様に認識Ｓｃｏｒｅの０より大きい所がある）場合、それらの認識Ｓｃｏｒｅの部分はつながっていると考えると、いくつかの認識Ｓｃｏｒｅの島ができる。この島のデータ、つまり島の投影面積、体積、最大高さ、Ｂｅｔａ（島のなだらかさ）などを調べた。

また、今は各音素ごとに同じＳｔａｒｔ－ＦｒａｍｅとＥｎｄ－Ｆｒａｍｅの組合せでそれぞれ認識Ｓｃｏｒｅがついているが、次にそのＳｔａｒｔ－ＦｒａｍｅとＥｎｄ－Ｆｒａｍｅの組合せで認識Ｓｃｏｒｅが最大の音素とその認識Ｓｃｏｒｅを、そのＳｔａｒｔ－ＦｒａｍｅとＥｎｄ－Ｆｒａｍｅの組合せにおける音素と認識Ｓｃｏｒｅとすると、最大認識Ｓｃｏｒｅの音素だけが集まった音響マトリックスができる。そうして今度はこの最大認識Ｓｃｏｒｅの音素を集めて作った音響マトリックスを使用し、島のデータを調べる。つまり同じ音素が隣あった場合、その部分はつながっていると考えると音素ごとのさっきと同様な島ができる。この島の投影面積、体積、最大高さ、Ｂｅｔａなどを調べた。

3　使用データ

まず各Ｓｔａｒｔ－ＦｒａｍｅとＥｎｄ－Ｆｒａｍｅの組合せにおける認識Ｓｃｏｒｅは、ＨＨＭの誘導計算から求めた認識Ｓｃｏｒｅを使用した。この認識Ｓｃｏｒｅのデータを基に各文節の最高認識Ｓｃｏｒｅが１０００になるように正規化を行い使用した。また正解認識かどうか判定するためにオートラベリングの結果を使用した。オートラベリングとは、実際に発声された文節の各音素が音声信号のどの位置で発声されたか自動的に示してくれるものである。例えば２８５ｍｓｅｃから３３５ｍｓｅｃまでは／ｍ／の音素が、３３５

ｍｓｅｃから４０５ｍｓｅｃまでは／ｏ／の音素が発声されているといったものである。また認識Ｓｃｏｒｅは Ｆｒａｍｅ単位で出されており、オートラベリングの結果はｍｓｅｃ単位であるために単位の変換が必要となる。そのため、実際に発声された音素、文節発声開始時間、文節発声終了時間のデータが書いてある音声認識用ラベルを用いて、オートラベリングの結果をＦｒａｍｅ単位へ変換し使用した。

4　結果の計算方法

まず島のデータとして用いる投影面積は、島とした０より大きい認識Ｓｃｏｒｅの数である。体積は島とした認識Ｓｃｏｒｅの値の合計である。（認識Ｓｃｏｒｅの値は正規化したものを使用）、最大高さは島とした認識Ｓｃｏｒｅでもっとも大きい認識Ｓｃｏｒｅの値である。Ｂｅｔａ（Ｓｈａｐｅ Ｐａｒａｍｅｔｅｒ）はその島のなだらかさ（傾斜）を示す値であり、０（島の傾斜はなだらか）から１（島の傾斜は急）までの値を取る。

まず、４０文節すべて各音素ごとに島のデータをとり、島の投影面積、体積、最大高さ、Ｂｅｔａを出した。それから文節ごとに、現れたすべての音素の島のデータから、投影面積、体積、最大高さ、Ｂｅｔａの大きいもの（Ｂｅｔａのみ小さいものも）をそれぞれ上位から１０個取り出した。そしてこの上位１０個に入っている音素を各音素ごとに数えた。またＴｏｐ１０に入った１０個の音素の島のそれぞれもっとも認識Ｓｃｏｒｅの高い位置、つまり島の最大高さの位置のＳｔａｒｔ－ＦｒａｍｅとＥｎｄ－Ｆｒａｍｅの組合せを、その音素の島が位置するＳｔａｒｔ－ＦｒａｍｅとＥｎｄ－Ｆｒａｍｅとした。この時の音素のＳｔａｒｔ－ＦｒａｍｅとＥｎｄ－Ｆｒａｍｅが、オートラベリングによってラベリングされ、示された実際に発声された音素のＦｒａｍｅの範囲に入っていれば正しく認識されたとして、正解認識とした。この正解認識とした音素のＦｒａｍｅの範囲は、Ｓｔａｒｔ－Ｆｒａｍｅに－５Ｆｒａｍｅ、Ｅｎｄ－Ｆｒａｍｅに＋５Ｆｒａｍｅの余裕を与えた。また複数の同じ音素の島が、同一のオートラベリングによって示された音素の範囲に入り、正解認識とされることがある。その場合も、正解認識とした。しかしながら正解認識数とは別に正解音素認識数を使用し、これらの区別をつけた。つまり正解音素認識数は、実際に発声された文節の音素のうちどれだけの数の音素が認識できたかを示している。

次に各Ｓｔａｒｔ－ＦｒａｍｅとＥｎｄ－Ｆｒａｍｅの組合せで、もっとも認識Ｓｃｏｒｅの高い音素を集めて作った音響マトリックスで同様なことを行なった。また今度は正解認識とするオートラベリングの音素の範囲の余裕をＳｔａｒｔ－Ｆｒａｍｅ、Ｅｎｄ－Ｆｒａｍｅに

±５Ｆｒａｍｅ、
±音素長（（Ｅｎｄ－Ｆｒａｍｅ）－（Ｓｔａｒｔ－Ｆｒａｍｅ））＊５％
±音素長（（Ｅｎｄ－Ｆｒａｍｅ）－（Ｓｔａｒｔ－Ｆｒａｍｅ））＊１０％
±音素長（（Ｅｎｄ－Ｆｒａｍｅ）－（Ｓｔａｒｔ－Ｆｒａｍｅ））＊１５％

といった４種類を与え正解認識数、正解音素数などの変化のデータを取った。

また、認識Ｓｃｏｒｅにのみ注目したデータも取った。これは文節ごとに、認識Ｓｃｏｒｅの大きいものを１０個取りだし、正解認識数、正解認識率、正解音素数、正解音素率などを調べた。

5　結果

表１から表８までに前章で述べた方法での結果を示す。

出現回数は実際に発声された４０文節中に何回その音素が出現したかを示している。括弧の中の値はオートラベリングに失敗した２文節を除いた出現回数である。Ｔｏｐ１０出現回数は４０文節で認識Ｓｃｏｒｅ、または島のデータの投影面積、体積、最大高さ、Ｂｅｔａの大きい順に並び換えた時、どの音素が何回Ｔｏｐ１０に出現したかを示している。正解認識数は、認識ＳｃｏｒｅＴｏｐ１０、または島のデータの投影面積、体積、最大高さ、ＢｅｔａのＴｏｐ１０に出現した音素とその位置が、実際の発声された文節の音素と位置に何回一致したかを示している。実際には認識ＳｃｏｒｅＴｏｐ１０、または島のデータの投影

表 1: ＳＣＯＲＥ－Ｔｏｐ１０における結果（±５Ｆｒａｍｅ）

音素	出現回数	Ｔｏｐ１０出現回数	正解認識数	正解率（%）	正解音素数	正解率（%）
-	0	0	0	0	0	0
a	63(61)	30	30	100.0	3	4.8(4.9)
b	0	0	0	0	0	0
ch	4	3	0	0	0	0
d	15(14)	10	10	100.0	1	6.7(7.1)
e	24(22)	1	1	100.0	1	4.2(4.5)
g	9(8)	0	0	0	0	0
h	1	10	10	100.0	1	100.0
i	52(51)	33	32	97.0	4	7.7(7.8)
j	14(13)	4	4	100.0	2	14.3(15.4)
k	24	7	6	85.7	1	4.2
m	23	0	0	0	0	0
n	10(7)	10	10	100.0	1	10.0(14.3)
ng	0	0	0	0	0	0
o	53	0	0	0	0	0
p	0	46	0	0	0	0
q	0	24	0	0	0	0
r	15	0	0	0	0	0
s	22(20)	77	77	100.0	8	36.4(40.0)
sh	20	43	43	100.0	6	30.0
t	15(14)	73(63)	30	41.1(47.6)	6	40.0(42.9)
ts	3	12	10	83.3	1	33.3
u	36(33)	0	0	0	0	0
w	10	0	0	0	0	0
z	3	7	7	100.0	1	33.3
zh	4(3)	10(0)	0	0(0)	0	0

表 2: ＡＲＥＡ－Ｔｏｐ１０における結果（±５Ｆｒａｍｅ）

音素	出現回数	Ｔｏｐ１０出現回数	正解認識数	正解率（%）	正解音素数	音素認識率（%）
-	0	0	0	0	0	0
a	63(61)	41(39)	35	85.4(89.7)	35	55.5(57.4)
b	0	2	0	0	0	0
ch	4	4	2	50.0	2	50.0
d	15(14)	2	2	100.0	2	13.3(14.3)
e	24(22)	21(20)	14	66.7(70.0)	14	58.3(63.6)
g	9(8)	48(44)	3	6.3(6.8)	3	33.3(37.5)
h	1	19(17)	1	5.3(5.9)	1	100.0
i	52(51)	26(25)	23	88.5(92.0)	23	44.2(45.1)
j	14(13)	16(15)	8	50.0(53.3)	8	57.1(61.5)
k	24	45	19	42.2	18	75.0
m	23	2	2	100.0	2	8.7
n	10(7)	0	0	0	0	0
ng	0	25(23)	0	0	0	0
o	53	27	24	88.9	24	45.3
p	0	2	0	0	0	0
q	0	2	0	0	0	0
r	15	22(20)	5	22.7(25.0)	5	33.3
s	22(20)	22(21)	18	81.8(85.7)	18	81.8(90.0)
sh	20	19	17	89.5	17	85.0
t	15(14)	15(14)	6	40.0(42.9)	6	40.0(42.9)
ts	3	4	2	50.0	2	66.7
u	36(33)	27(25)	12	44.4(48.0)	12	33.3(36.4)
w	10	6	3	50.0	3	30.0
z	3	1	0	0	0	0
zh	4(3)	2(1)	1	50.0(100.0)	1	25.0(33.3)

表 3: ＶＯＬＵＭＥ－Ｔｏｐ１０における結果（±５Ｆｒａｍｅ）

音素	出現回数	Ｔｏｐ１０出現回数	正解認識数	正解率（%）	正解音素数	音素認識率（%）
-	0	0	0	0	0	0
a	63(61)	36(34)	31	86.1(91.2)	31	49.2(50.8)
b	0	3	0	0	0	0
ch	4	5	3	60.0	3	75.0
d	15(14)	2	2	100	2	13.3(14.3)
e	24(22)	21(20)	15	71.4(75.0)	15	62.5(68.2)
g	9(8)	48(44)	3	6.3(6.8)	3	33.3(37.5)
h	1	17(15)	1	5.9(6.7)	1	100.0
i	52(51)	26(25)	23	88.5(92.0)	23	44.2(45.1)
j	14(13)	16(15)	8	50.0(53.3)	8	57.1(61.5)
k	24	46	19	41.3	18	75.0
m	23	2	2	100	2	8.7
n	10(7)	1	1	100	1	10.0(14.3)
ng	0	25(23)	0	0	0	0
o	53	26	23	88.5	23	43.4
p	0	3	0	0	0	0
q	0	3	0	0	0	0
r	15	19(17)	4	21.1(23.5)	4	26.7
s	22(20)	22(21)	18	81.8(85.7)	18	81.8(90.0)
sh	20	19	17	89.5	17	85.0
t	15(14)	17(16)	8	47.1(50.0)	8	53.3(57.1)
ts	3	5	2	40.0	2	66.7
u	36(33)	26(24)	12	46.2(50.0)	12	33.3(36.4)
w	10	6	4	66.7	4	40.0
z	3	3	2	66.7	2	66.7
zh	4(3)	3(2)	1	33.3(50.0)	1	25.0(33.3)

表4: ＡＬＴＩＴＵＤＥ－Ｔｏｐ１０における結果（±５Ｆｒａｍｅ）

音素	出現回数	Ｔｏｐ１０出現回数	正解認識数	正解率（%）	正解音素数	音素認識率（%）
-	0	0	0	0	0	0
a	63(61)	26(24)	22	84.6(91.7)	22	34.9(36.1)
b	0	10(8)	0	0	0	0
ch	4	7	3	42.9	3	75.0
d	15(14)	10	7	70	7	46.7(50.0)
e	24(22)	16(15)	12	75.0(80.0)	12	50.0(54.5)
g	9(8)	8	3	37.5	2	22.2(25.0)
h	1	11	1	9.0	1	100.0
i	52(51)	23	16	69.6	16	30.8(31.4)
j	14(13)	11(10)	5	45.5(50.0)	5	35.7(38.5)
k	24	27	6	22.2	6	25.0
m	23	9(7)	7	77.8(100.0)	7	29.2
n	10(7)	4	2	50	2	20.0(28.6)
ng	0	22(20)	0	0	0	0
o	53	12	11	91.7	11	20.8
p	0	25	0	0	0	0
q	0	30	0	0	0	0
r	15	14(13)	6	42.9(46.2)	6	40.0
s	22(20)	27(24)	17	63.0(70.8)	17	77.3(85.0)
sh	20	20	17	85.0	17	85.0
t	15(14)	37(36)	14	37.8(38.9)	12	80.0(85.7)
ts	3	18(17)	2	11.1(11.8)	2	66.7
u	36(33)	16(14)	6	37.5(42.9)	6	16.7(18.2)
w	10	10	7	70.0	7	70.0
z	3	3	3	100.0	3	100.0
zh	4(3)	4(3)	2	50.0(66.7)	2	50.0(66.7)

表5: ＢＥＴＡ－Ｔｏｐ１０における結果（±５Ｆｒａｍｅ）

音素	出現回数	Ｔｏｐ１０出現回数	正解認識数	正解率（%）	正解音素数	音素認識率（%）
-	0	0	0	0	0	0
a	63(61)	32(30)	26	81.3(86.7)	26	41.3(42.6)
b	0	10(9)	0	0	0	0
ch	4	5	3	60.0	3	75.0
d	15(14)	3	2	66.7	2	13.3(14.3)
e	24(22)	14(12)	9	64.3(75.0)	9	37.5(40.9)
g	9(8)	51(47)	3	5.9(6.4)	3	33.3(37.5)
h	1	13(12)	1	7.7(8.3)	1	100.0
i	52(51)	9	8	88.9	8	15.4(15.7)
j	14(13)	16	8	50.0	8	57.1(61.5)
k	24	66(65)	24	36.4(36.9)	19	79.2
m	23	2	2	100.0	2	8.7
n	10(7)	2	2	100.0	2	20.0(28.6)
ng	0	6(4)	0	0	0	0
o	53	24	19	79.2	19	35.8
p	0	10	0	0	0	0
q	0	5	0	0	0	0
r	15	45(43)	7	15.6(16.3)	7	46.7
s	22(20)	14(13)	10	71.4(76.9)	10	45.5(50.0)
sh	20	16	15	93.8	15	75.0
t	15(14)	29(27)	10	34.5(37.0)	10	66.7(71.4)
ts	3	1	1	100.0	1	33.3
u	36(33)	10	5	50.0	5	13.9(15.2)
w	10	9	7	77.8	7	70.0
z	3	5(4)	2	40.0(50.0)	2	66.7
zh	4(3)	3(2)	1	33.3(50.0)	1	25.0(33.3)

表 6: ＡＲＥＡ＆ＶＯＬＵＭＥ－Ｔｏｐ１０における結果（±５Ｆｒａｍｅ）

音素	出現回数	Ｔｏｐ１０出現回数	正解認識数	正解率（％）	正解音素数	音素認識率（％）
-	0	0	0	0	0	0
a	63(61)	36(34)	31	86.1(91.2)	31	49.2(50.8)
b	0	2	0	0	0	0
ch	4	4	2	50.0	2	50.0
d	15(14)	2	2	100.0	2	13.3(14.3)
e	24(22)	20(19)	14	70.0(73.7)	14	58.3(63.6)
g	9(8)	48(44)	3	6.3(6.8)	3	33.3(37.5)
h	1	17(15)	1	5.9(6.7)	1	100.0
i	52(51)	25(24)	22	88.0(91.7)	22	42.3(43.1)
j	14(13)	15(14)	8	53.3(57.1)	8	57.1(61.5)
k	24	45	19	42.2	18	75.0
m	23	2	2	100.0	2	8.7
n	19(7)	0	0	0	0	0
ng	0	24(22)	0	0	0	0
o	53	26	23	88.5	23	43.4
p	0	2	0	0	0	0
q	0	2	0	0	0	0
r	15	19(17)	4	21.1(23.5)	4	26.7
s	22(20)	22(21)	18	81.8(85.7)	18	81.8(90.0)
sh	20	19	17	89.5	17	85.0
t	15(14)	15(14)	6	40.0(42.9)	6	40.0(42.9)
ts	3	4	2	50.0	2	66.7
u	36(33)	26(24)	12	46.2(50.0)	12	33.3(36.4)
w	10	5	3	60.0	3	30.0
z	3	1	0	0	0	0
zh	4(3)	2(1)	1	50.0(100.0)	1	25.0(33.3)

表7: ＡＲＥＡ&ＡＬＴＩＴＵＤＥ－Ｔｏｐ１０における結果（±５Ｆｒａｍｅ）

音素	出現回数	Ｔｏｐ１０出現回数	正解認識数	正解率（%）	正解音素数	音素認識率（%）
-	0	0	0	0	0	0
a	63(61)	22(20)	19	86.4(95.0)	19	30.2(31.1)
b	0	2	0	0	0	0
ch	4	4	2	50.0	2	50.0
d	15(14)	2	2	100.0	2	13.3(14.3)
e	24(22)	15(14)	12	80.0(85.7)	12	50.0(54.5)
g	9(8)	4	2	50.0	2	22.2(25.0)
h	1	5	1	20.0	1	100.0
i	52(51)	18	16	88.9	16	30.8(31.4)
j	14(13)	8(7)	5	62.5(71.4)	5	35.7(38.5)
k	24	13	6	46.2	6	25.0
f	23	2	2	100.0	2	8.7
n	10(7)	0	0	0	0	0
ng	0	8(6)	0	0	0	0
o	53	12	11	91.7	11	20.8
p	0	2	0	0	0	0
q	0	1	0	0	0	0
r	15	11(10)	4	36.4(40.0)	4	26.7
s	22(20)	21(20)	17	81.0(85.0)	17	77.3(85.0)
sh	20	19	17	89.5	17	85.0
t	15(14)	10(9)	6	60.0(66.7)	6	40.0(42.9)
ts	3	4	2	50.0	2	66.7
u	36(33)	9(7)	6	66.7(85.7)	6	16.7(18.2)
w	10	4	3	75.0	3	30.0
z	3	0	0	0	0	0
zh	4(3)	2(1)	1	50.0(100.0)	1	25.0(33.3)

表8: ＶＯＬＵＭＥ＆ＡＬＴＩＴＵＤＥ－Ｔｏｐ１０における結果（±５Ｆｒａｍｅ）

音素	出現回数	Ｔｏｐ１０出現回数	正解認識数	正解率（%）	正解音素数	音素認識率（%）
-	0	0	0	0	0	0
a	63(61)	21(19)	18	85.7(94.7)	18	28.6(29.5)
b	0	3	0	0	0	0
ch	4	5	3	60.0	3	75.0
d	15(14)	2	2	100.0	2	13.3(14.3)
e	24(22)	15(14)	12	80.0(85.7)	12	50.0(54.5)
g	9(8)	4	2	50.0	2	22.2(25.0)
h	1	5	1	20.0	1	100.0
i	52(51)	18	16	88.9	16	30.8(31.4)
j	14(13)	8(7)	5	62.5(71.4)	5	35.7(38.5)
k	24	13	6	46.2	6	25.0
m	23	2	2	100.0	2	8.7
n	10(7)	1	1	100.0	1	14.3
ng	0	8(6)	0	0	0	0
o	53	12	11	91.7	11	20.8
p	0	3	0	0	0	0
q	0	2	0	0	0	0
r	15	11(10)	4	36.4(40.0)	4	26.7
s	22(20)	21(20)	17	81.0(85.0)	17	77.3(85.0)
sh	20	19	17	89.5	17	85.0
t	15(14)	12(11)	8	66.7(72.7)	8	53.3(57.1)
ts	3	5	2	40.0	2	66.7
u	36(33)	9(7)	6	66.7(85.7)	6	16.7(18.2)
w	10	5	4	80.0	4	40.0
z	3	2	2	100.0	2	66.7
zh	4(3)	2(1)	1	50.0(100.0)	1	25.0(33.3)

表9: ＡＲＥＡ&ＶＯＬＵＭＥ&ＡＬＴＩＴＵＤＥにおける結果（±５Ｆｒａｍｅ）

音素	出現回数	Ｔｏｐ１０出現回数	正解認識数	正解率（%）	正解音素数	音素認識率（%）
-	0	0	0	0	0	0
a	63(61)	21(19)	18	85.7(94.7)	18	28.6(29.5)
b	0	2	0	0	0	0
ch	4	4	2	50.0	2	50.0
d	15(14)	2	2	100.0	2	13.3(14.3)
e	24(22)	15(14)	12	80.0(85.7)	12	50.0(54.5)
g	9(8)	4	2	50.0	2	22.2(25.0)
h	1	5	1	20.0	1	100.0
i	52(51)	18	16	88.9	16	30.8(31.4)
j	14(13)	8(7)	5	62.5(71.4)	5	35.7(38.5)
k	24	13	6	46.2	6	25.0
m	23	2	2	100.0	2	8.7
n	10(7)	0	0	0	0	0
ng	0	8(6)	0	0	0	0
o	53	12	11	91.7	11	20.8
p	0	2	0	0	0	0
q	0	1	0	0	0	0
r	15	11(10)	4	36.4(40.0)	4	26.7
s	22(20)	21(20)	17	81.0(85.0)	17	77.3(85.0)
sh	20	19	17	89.5	17	85.0
t	15(14)	10(9)	6	60.0(66.7)	6	40.0(42.9)
ts	3	4	2	50.0	2	66.7
u	36(33)	9(7)	6	66.7(85.7)	6	16.7(18.2)
w	10	4	3	75.0	3	30.0
z	3	0	0	0	0	0
zh	4(3)	2(1)	1	50.0(100.0)	1	25.0(33.3)

面積、体積、最大高さ、ＢｅｔａのＴｏｐ１０に出現した音素の島の最大高さの位置のＳｔａｒｔ－ＦｒａｍｅとＥｎｄ－Ｆｒａｍｅが、オートラベリングによって示された実際に発声された文節の各音素の位置のＳｔａｒｔ－ＦｒａｍｅとＥｎｄ－Ｆｒａｍｅの範囲に入っていれば正解とした。正解率は正解認識数をＴｏｐ１０出現回数で割ったものである。正解音素数は、オートラベリングによって示された音素の範囲に複数個の島の音素が正解と認識されることがあるが、これを一つとして数えたのが正解音素数である。音素正解率は正解音素数を出現回数で割ったものである。

最初に認識Ｓｃｏｒｅにのみに注目し、認識Ｓｃｏｒｅの大きいものＴｏｐ１０による結果を見てみると、Ｔｏｐ１０出現回数は／ａ／、／ｉ／、／ｐ／、／ｑ／、／ｓ／、／ｓｈ／、／ｔ／などが多く出現している。正解率では／ｐ／、／ｑ／、／ｔ／を除いてＴｏｐ１０出現回数の多いものが高い正解率を示している。しかしながら正解音素数や音素正解率を見ると、どれも大変低くなっている。これは、認識Ｓｃｏｒｅの高い音素の周辺は、やはり同じ音素の認識Ｓｃｏｒｅが同様に高く、Ｔｏｐ１０内に周辺の同じ音素がいくつも入ってしまうためである。また／ｐ／、／ｑ／、／ｔ／などは、最初の音素が発声されるまでの無音区間に良く出現し、その認識Ｓｃｏｒｅは高い。

次に音素ごとに認識Ｓｃｏｒｅの島のデータを出した場合の結果だが、これは投影面積、体積、最大高さ、ＢｅｔａのＴｏｐ１０すべて認識率が大変悪く、ほとんど正解認識とされたものが無かった。特に／ｇ／、／ｈ／、／ｋ／、／ｔ／、／ｒ／などの音素がほとんど毎文節ごとに上位に出現するが、正解認識とされることはほとんどなかった。これはこれらの音素は認識Ｓｃｏｒｅはあまり高くないが、認識Ｓｃｏｒｅが０より高くなりやすい音素であるために音響マトリックスの広範囲に出現し、さらに音素ごとに認識Ｓｃｏｒｅの島データを出したために投影面積、体積などが全体として大きくなりたびたびＴｏｐ１０内に出現したものと考えられる。

次に最大の認識Ｓｃｏｒｅの音素を集めた音響マトリックスの場合の結果では、Ｔｏｐ１０出現回数では、投影面積Ｔｏｐ１０、体積Ｔｏｐ１０、ＢｅｔａＴｏｐ１０においては、母音（ＢｅｔａＴｏｐ１０においては／ａ／、／ｏ／のみ）、／ｇ／、／ｈ／、／ｊ／、／ｋ／、／ｎｇ／、／ｒ／、／ｓ／、／ｓｈ／、／ｔ／などが多く出現している。最大高さｔｏｐ１０では、母音、／ｋ／、／ｎｇ／、／ｐ／、／ｑ／、／ｒ／、／ｓ／、／ｓｈ／、／ｔ／、／ｔｓ／ などが多く出現している。特に／ｇ／、／ｈ／、／ｎｇ／、／ｐ／、／ｑ／などは、実際の４０文節中にはほとんど出現していないのに、Ｔｏｐ１０出現回数は多い。つまりこれらの音素は認識Ｓｃｏｒｅが高くなりやすく、誤認識されやすいことを示している。

また認識率は投影面積、体積、最大高さ、Ｂｅｔａのどれも／ｕ／を除いた母音で７０％近くかそれ以上が出ている。また／ｕ／でも５０％近くは出ている。正解率で７０％を越えているものは、投影面積Ｔｏｐ１０と体積Ｔｏｐ１０で、／ａ／、／ｄ／、／ｅ／、／ｉ／、／ｍ／、／ｏ／、／ｓ／、／ｓｈ／、最大高さＴｏｐ１０で、／ａ／、／ｄ／、／ｅ／、／ｍ／、／ｏ／、／ｓ／、／ｓｈ／、／ｗ／、／ｚ／、ＢｅｔａＴｏｐ１０で、／ａ／、／ｅ／、／ｉ／、／ｍ／、／ｎ／、／ｏ／、／ｓ／、／ｓｈ／、／ｔｓ／、／ｗ／である。投影面積Ｔｏｐ１０、体積Ｔｏｐ１０、ＢｅｔａＴｏｐ１０では認識率が同じ傾向を取ってい

る。

正解音素数は認識Ｓｃｏｒｅにのみ注目した場合に比べ、ほとんど正解認識数と違いはなかった。ただ／ｋ／が良く同じ音素の範囲に複数個出現していた。

次に正解認識の範囲の余裕を±５Ｆｒａｍｅ、±音素長＊５％、±音素長＊１０％、±音素長＊１５％、とした場合の結果だが、これは投影面積Ｔｏｐ１０、体積Ｔｏｐ１０、最大高さＴｏｐ１０、偏差Ｔｏｐ１０どれも正解認識数は、±音素長＊５％＜±音素長＊１０％＜±音素長＊１５％＜±５Ｆｒａｍｅの順に多くなった。ここで音素長＊１５％が５Ｆｒａｍｅより長くなるためには音素長が３３Ｆｒａｍｅより長くならなければいけないが、オートラベリングの結果を見ると３３Ｆｒａｍｅより長いＦｒａｍｅは３８文節中（４０文節の内オートラベリングに失敗した２文節を除いたもの）で３音素だけであり、そのことから、この正解認識の音素の余裕を変えることによる特別変わった点は見られなかった。しかしながら、正解認識の範囲の余裕が±５Ｆｒａｍｅの場合と±音素長＊５％の場合を比べた時、正解認識の範囲の余裕が±音素長＊５％の時は母音や／ｋ／などの音素の正解認識数の減少が目立った。このことから母音や／ｋ／はオートラベリングに示されている音素の範囲からずれやすいことがわかる。

次に最大認識Ｓｃｏｒｅの音素を集めた音響マトリックスの島データで、投影面積、体積、最大高さのＴｏｐ１０に出現した各音素の内、投影面積＆体積（投影面積と体積のＴｏｐ１０の両方に出現した音素の島データ）、投影面積＆最大高さ、体積＆最大高さの島データで同様に正解認識率などを出して見た。結果は、投影面積＆体積ではほとんど投影面積、体積Ｔｏｐ１０それぞれの島データのみを使用して正解認識率などを出した結果とほとんど変化はなかった。これは投影面積Ｔｏｐ１０と体積Ｔｏｐ１０に出現している音素の島が非常に似ているためである。投影面積＆最大高さ、体積＆最大高さにおいては正解率は投影面積Ｔｏｐ１０、体積Ｔｏｐ１０、最大高さＴｏｐ１０のみの島データに比べ上がっている。しかしながら実際に発声された音素の内、どれだけの音素が認識できたかを示す音素正解率は、投影面積、体積、最大高さのみの島データに比べ下がっている。

最後に、Ｃｏｎｆｕｓｉｏｎ－ｍａｔｒｉｘを示す。これは各Ｓｔａｒｔ－ＦｒａｍｅとＥｎｄ－Ｆｒａｍｅの組合せにおいて、最高の認識Ｓｃｏｒｅを持つ音素をそのＳｔａｒｔ－ＦｒａｍｅとＥｎｄ－Ｆｒａｍｅの組合せにおける音素として作った音響マトリックスにおいて、オートラベリングによって示された文節の各音素のＳｔａｒｔ－ＦｒａｍｅとＥｎｄ－Ｆｒａｍｅの位置にどのような音素が出現しているかを示したものである。縦に並んでいる音素は実際に発声された文節に出てきた音素を示している。横に並んでいる音素は実際に出てきた音素の位置では、どの音素がもっとも認識Ｓｃｏｒｅが高かったかを示している。これを見ると実際に発声された音素の位置で認識Ｓｃｏｒｅのもっとも高い音素が、実際に発声された音素と同じになったのは８０．１％であり、かなり良い結果となっている。つまり投影面積、体積、最大高さ、ＢｅｔａのＴｏｐ１０内には入っていないが、実際に発声された位置ではその音素の認識Ｓｃｏｒｅは高いことを意味している。またここでよく間違えられた音素としては、

／ａ／ －> ／ｈ／

表10: Confusion Matrix

	-	a	b	ch	d	e	g	h	i	j	k	m	n	ng	o	p	q	r	s	sh	t	ts	u	w	z	zh
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a	0	51	0	0	0	1	0	5	0	0	0	0	0	3	0	0	0	1	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ch	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	0	0	1	0	9	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	0	0	0	0
e	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
g	0	0	1	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
h	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
i	1	0	0	0	0	0	0	2	31	9	2	0	0	1	1	0	1	0	0	0	3	0	0	0	0	0
j	0	0	0	0	0	0	0	0	1	11	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
m	0	0	5	0	0	0	1	0	0	0	0	12	0	0	0	0	0	4	0	0	0	0	1	0	0	0
n	0	0	1	0	0	0	1	0	0	0	0	0	4	0	0	0	0	1	0	0	0	0	0	0	0	0
ng	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
o	0	6	0	0	0	0	0	0	0	0	0	0	0	0	45	0	0	0	0	0	0	0	1	1	0	0
p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0
s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	2	0
sh	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0
t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0
ts	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
u	0	0	0	0	0	2	1	0	0	0	3	0	0	0	0	0	0	1	7	0	0	0	19	0	0	0
w	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	9	0	0
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
zh	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2

／i／ −> ／j／
／m／ −> ／b／
／o／ −> ／a／
／u／ −> ／s／

などが上げられる。特に／i／は／j／に多く間違えられていることがわかる。また間違えた数の多いものは、

／a／、／d／、／i／、／m／、／o／、／u／

などがあげられる。これも／i／の音素が一番多く間違えている。

6 評価

前章の結果では、出現回数、Top10出現回数、正解認識数、正解認識率、正解音素数、音素正解率などを示した。ここで問題となるのは出現回数が少ない音素だと正解音素率がどうしても高くなり信頼性の高低を一概に比べられない。そこで次のような値、Hit RateとFalse Alarm Rateを使い評価してみた。

Correct Answer (CA)
：出現回数
Top Ten Frequency (TF)
：Top10出現回数
Number Correctly Recognized (NCR)
：正解認識数
False Alarm (FA) =TF−NCR
False Alarm Rate (FAR)
=False Alarm／Hour／Phoneme

$$= (FA * 3600) / (CA * 32.17)$$

$$Hit\ Rate = NCR / CA$$

ＨＲは実際に発声された文節の音素の内、どれだけの音素が正しく認識されたかを示している。ＦＡＲは一時間会話をした場合、その音素が誤認識されるであろう数を示している。次に最大認識Ｓｃｏｒｅの音素を集め作った音響マトリックスでの結果のＦＡＲのｌｏｇを取ったものとＨＲを取ったグラフを図１～図３２に示す。このグラフでは左上にその音素が来た場合、実際に発声された音素のほとんどが正しく認識され、時間当たりの誤認識数も少ないことを意味している。

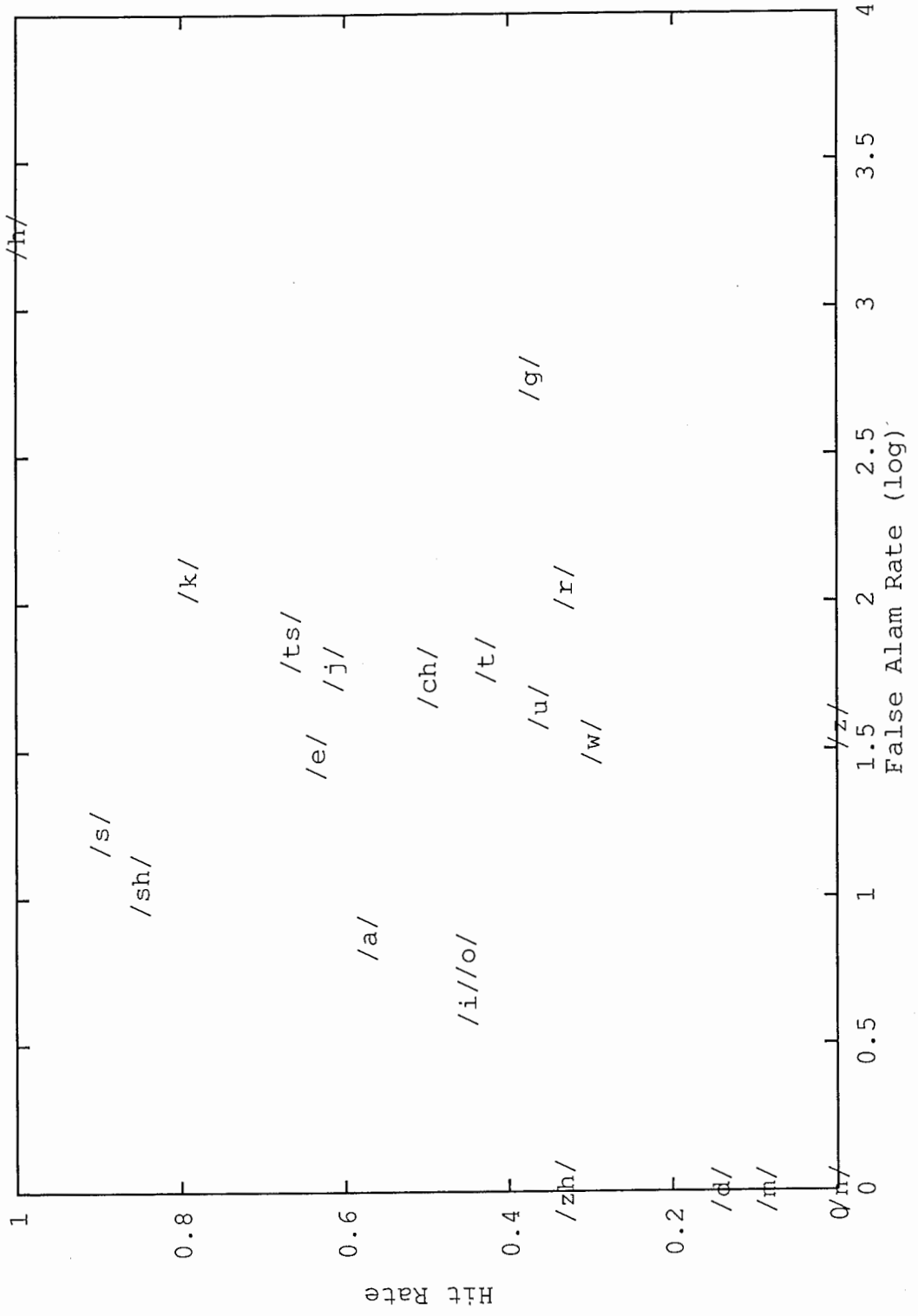

図1: HR&FAR−AREA (±5Frame)

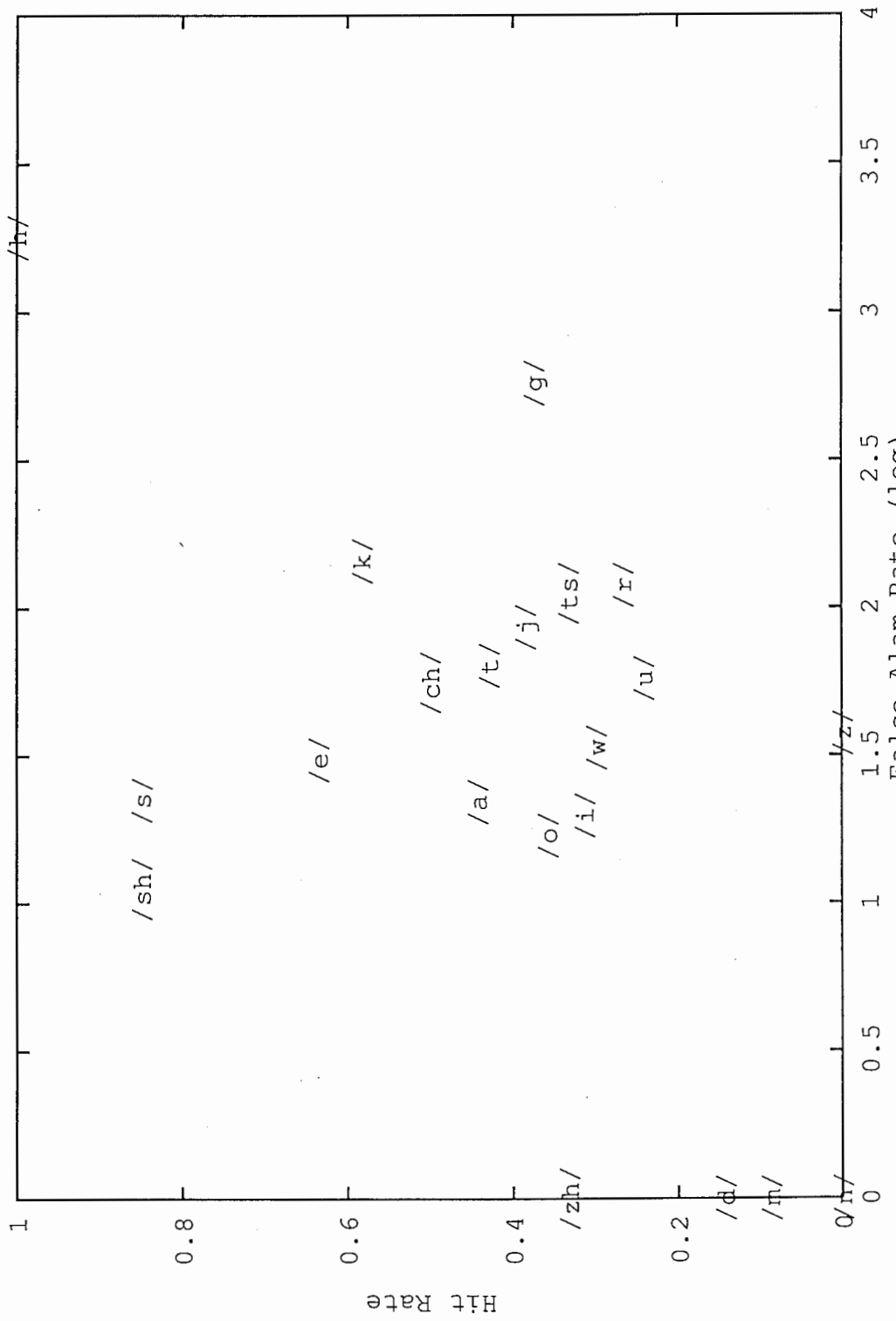

図 2: HR&FAR－AREA（±音素長＊5％）

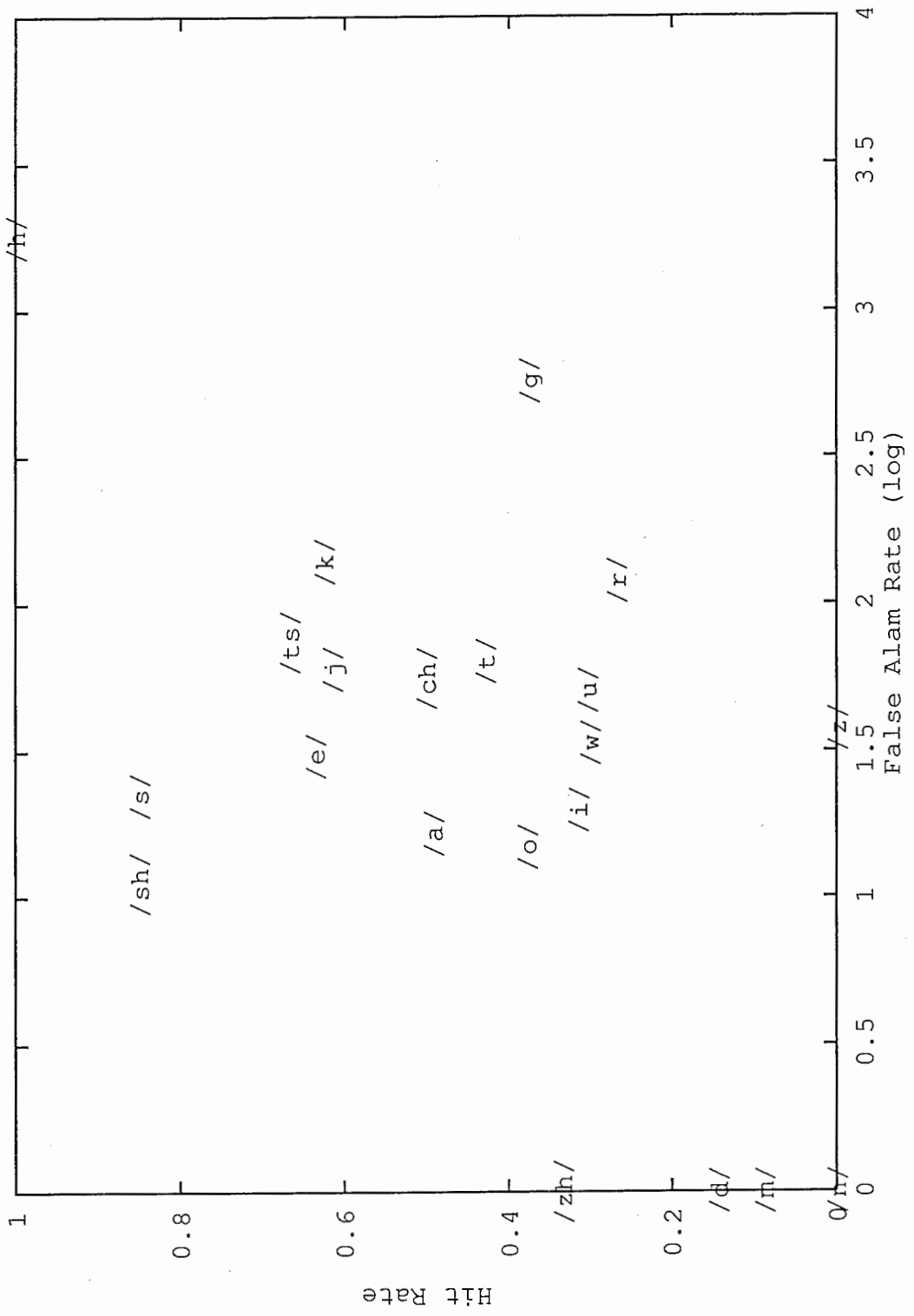

図 3: HR&FAR－AREA（±音素長＊10％）

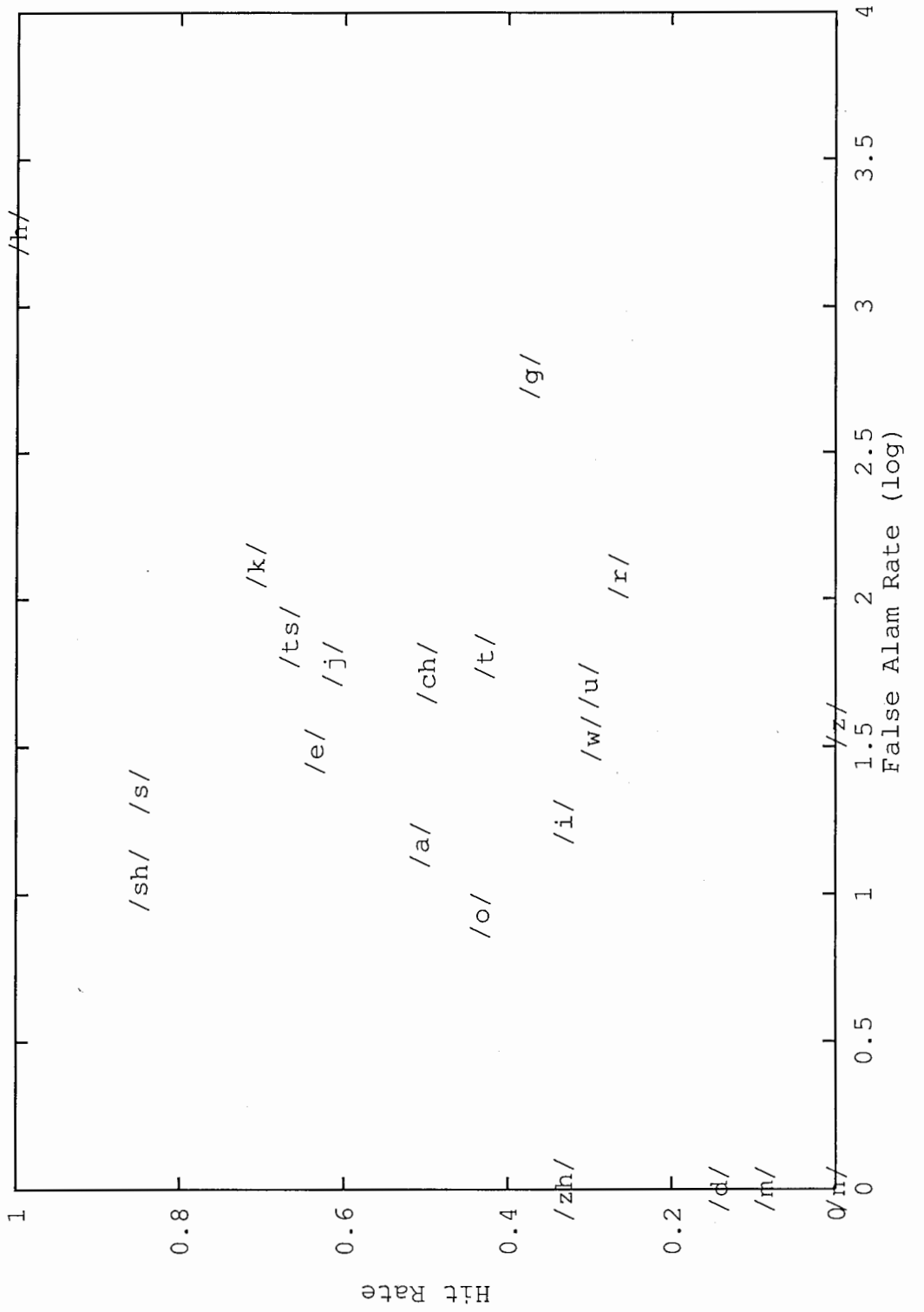

図 4: HR&FAR－AREA (±音素長＊15%)

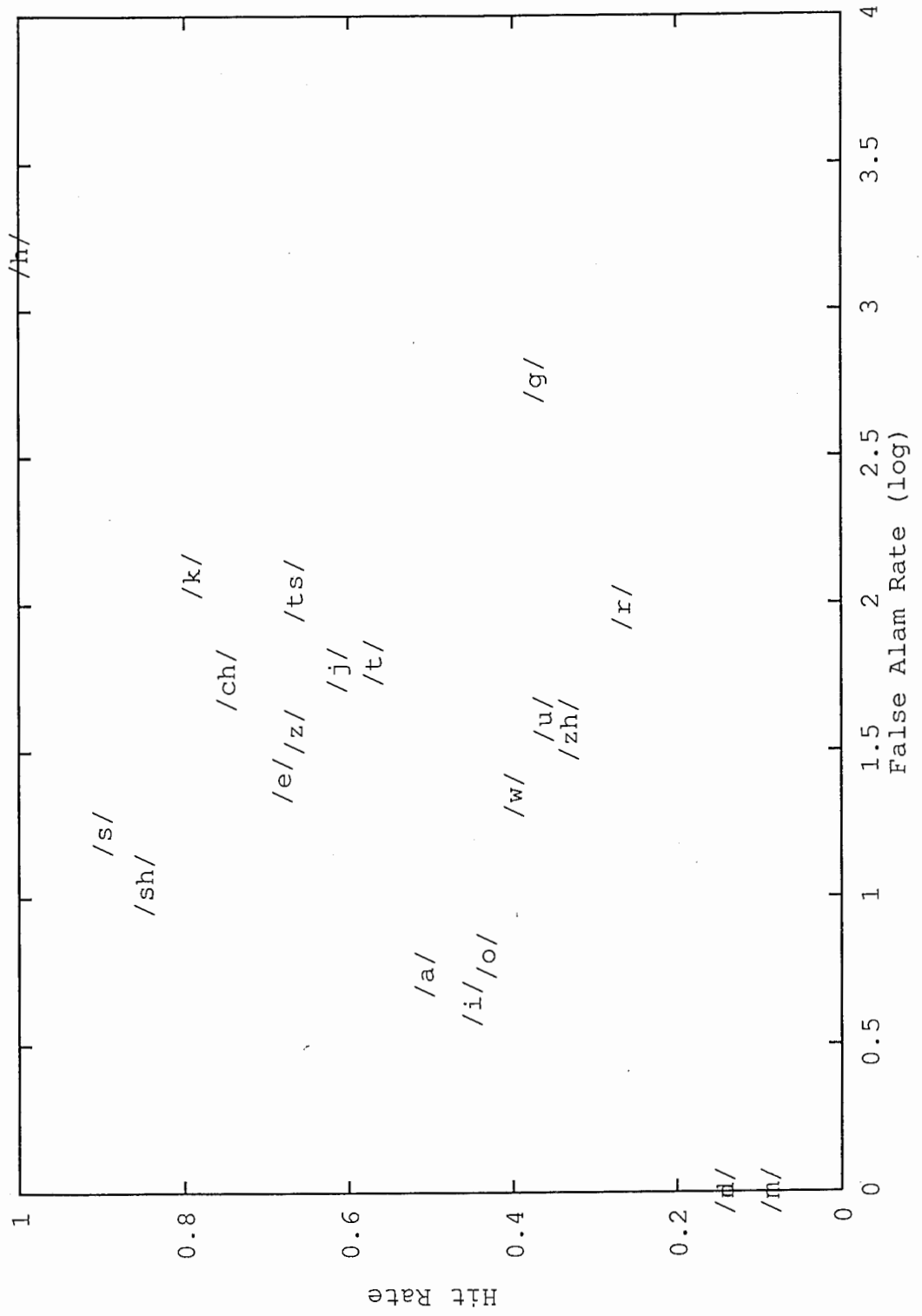

図5: HR&FAR−VOLUME (±5Frame)

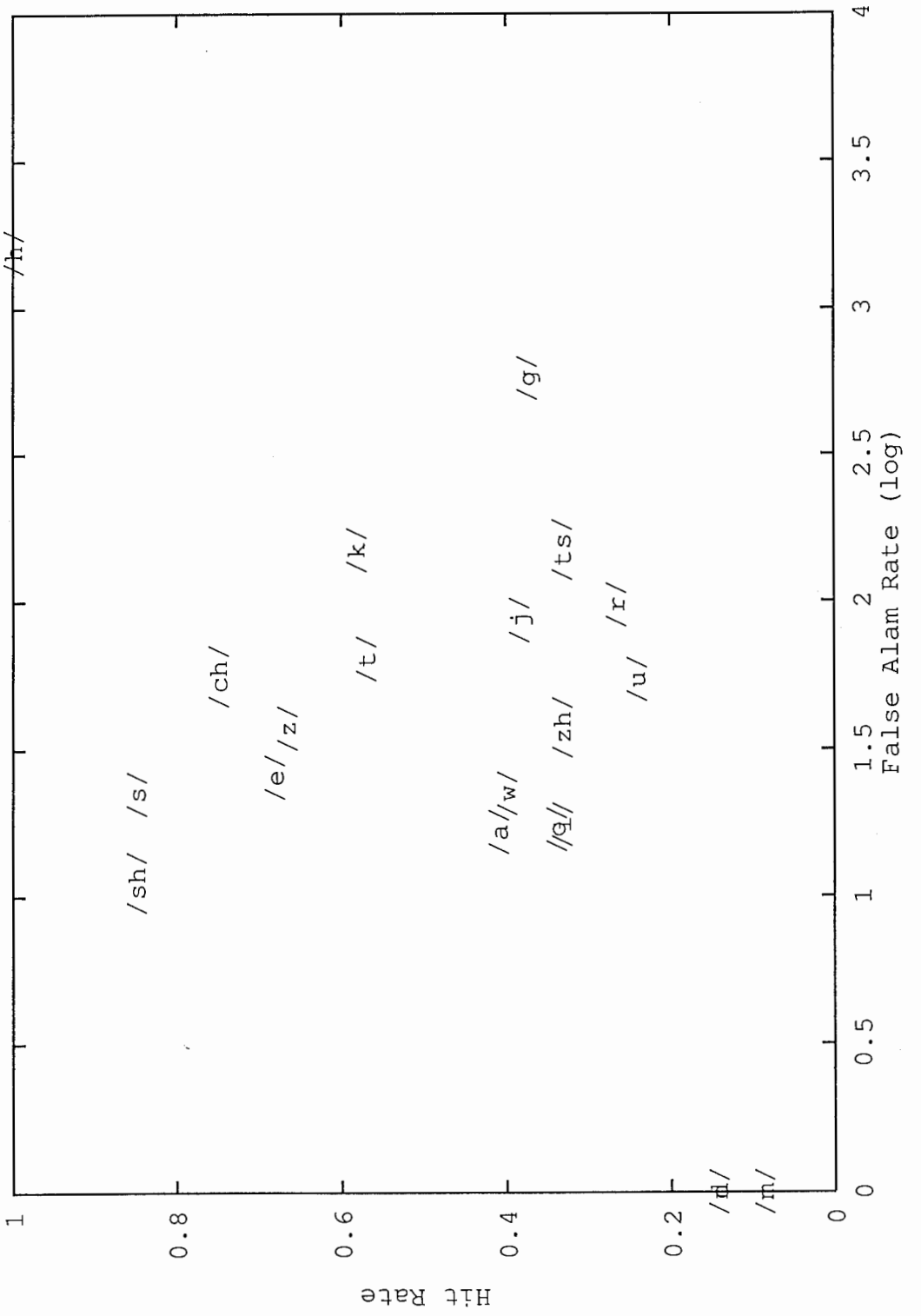

図 6: HR&FAR－VOLUME（±音素長＊5％）

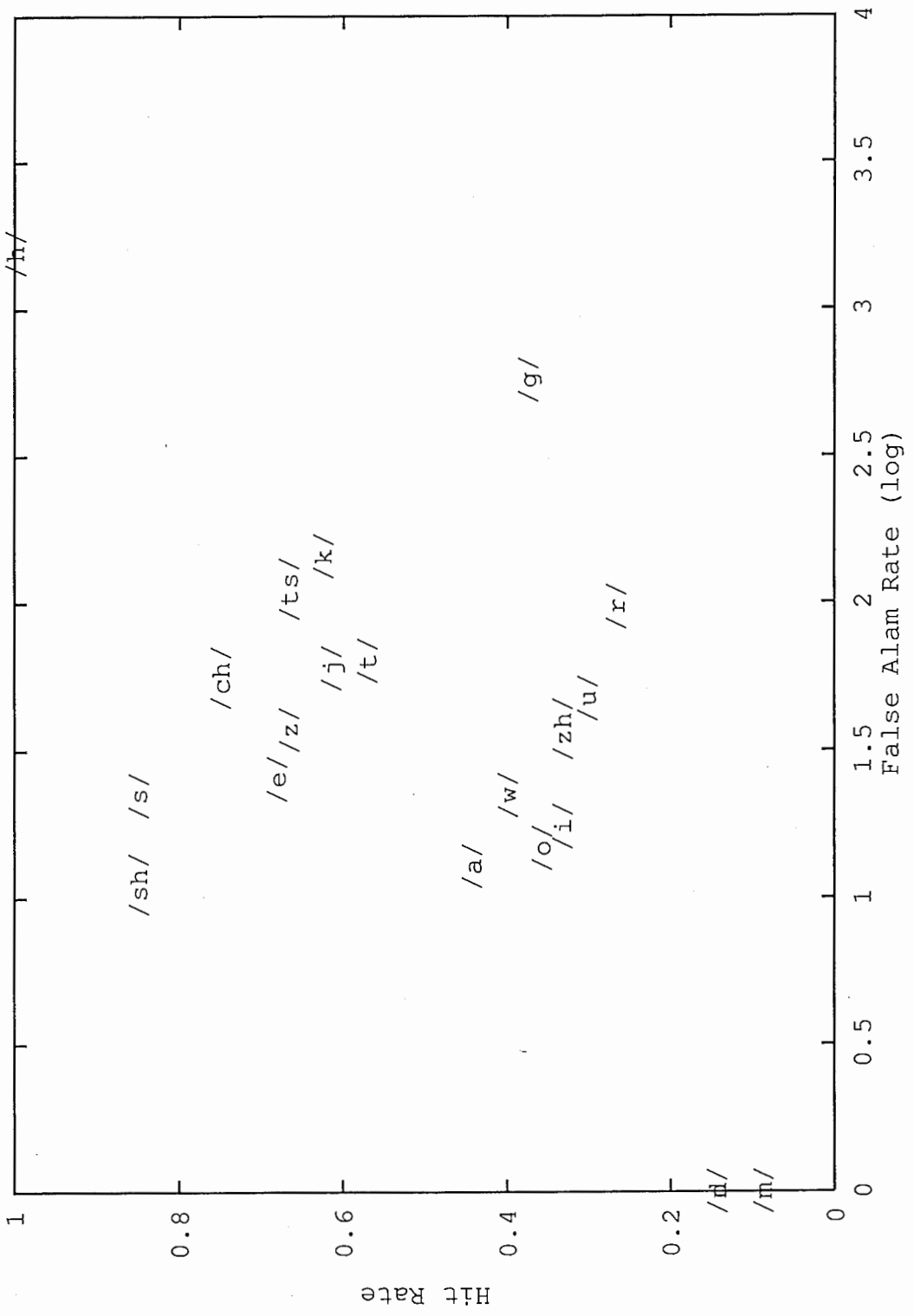

図 7: HR&FAR−VOLUME (±音素長＊10%)

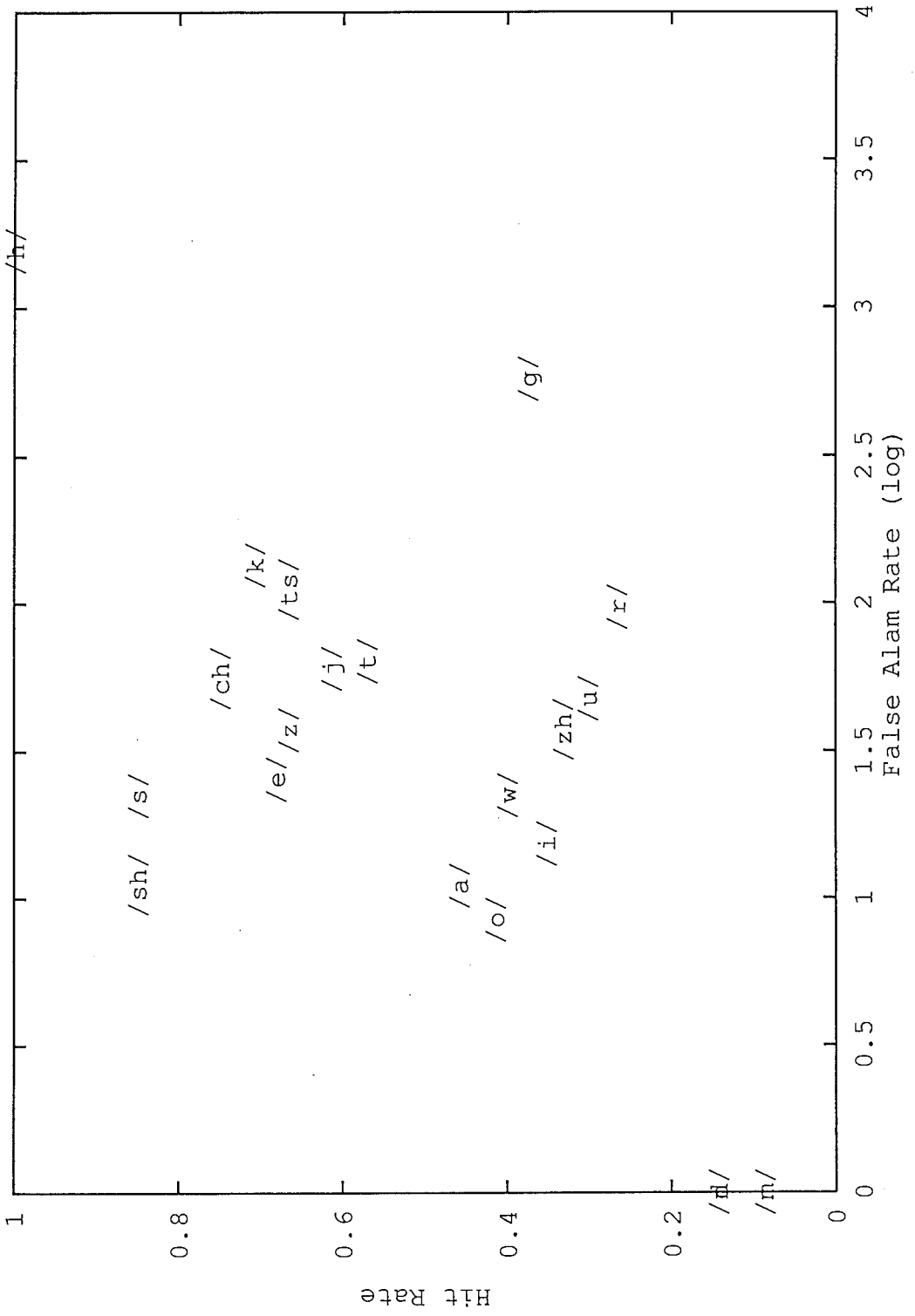

図 8: HR&FAR－VOLUME（±音素長＊１５％）

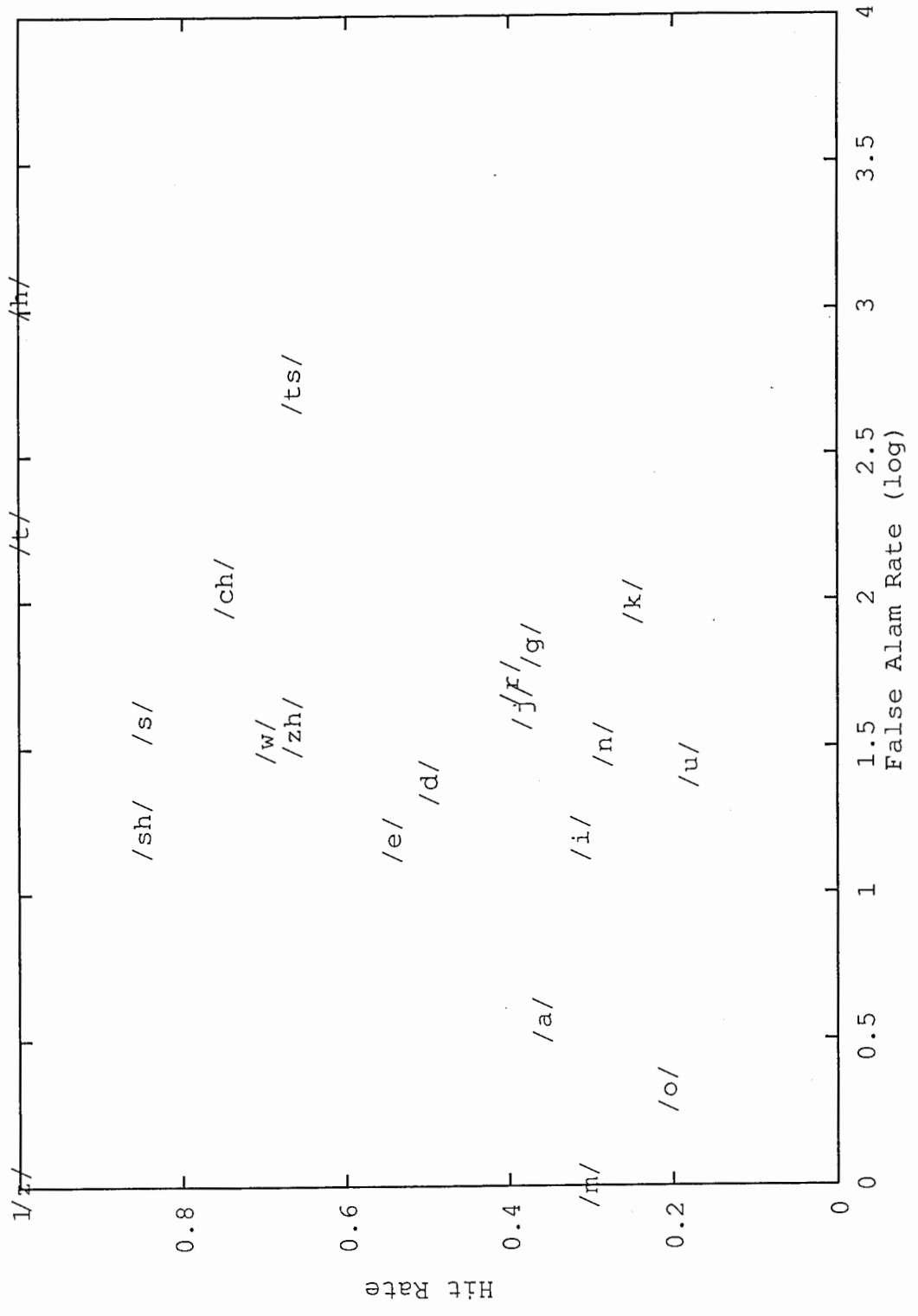

図 9: HR&FAR−ALTITUDE (±5Frame)

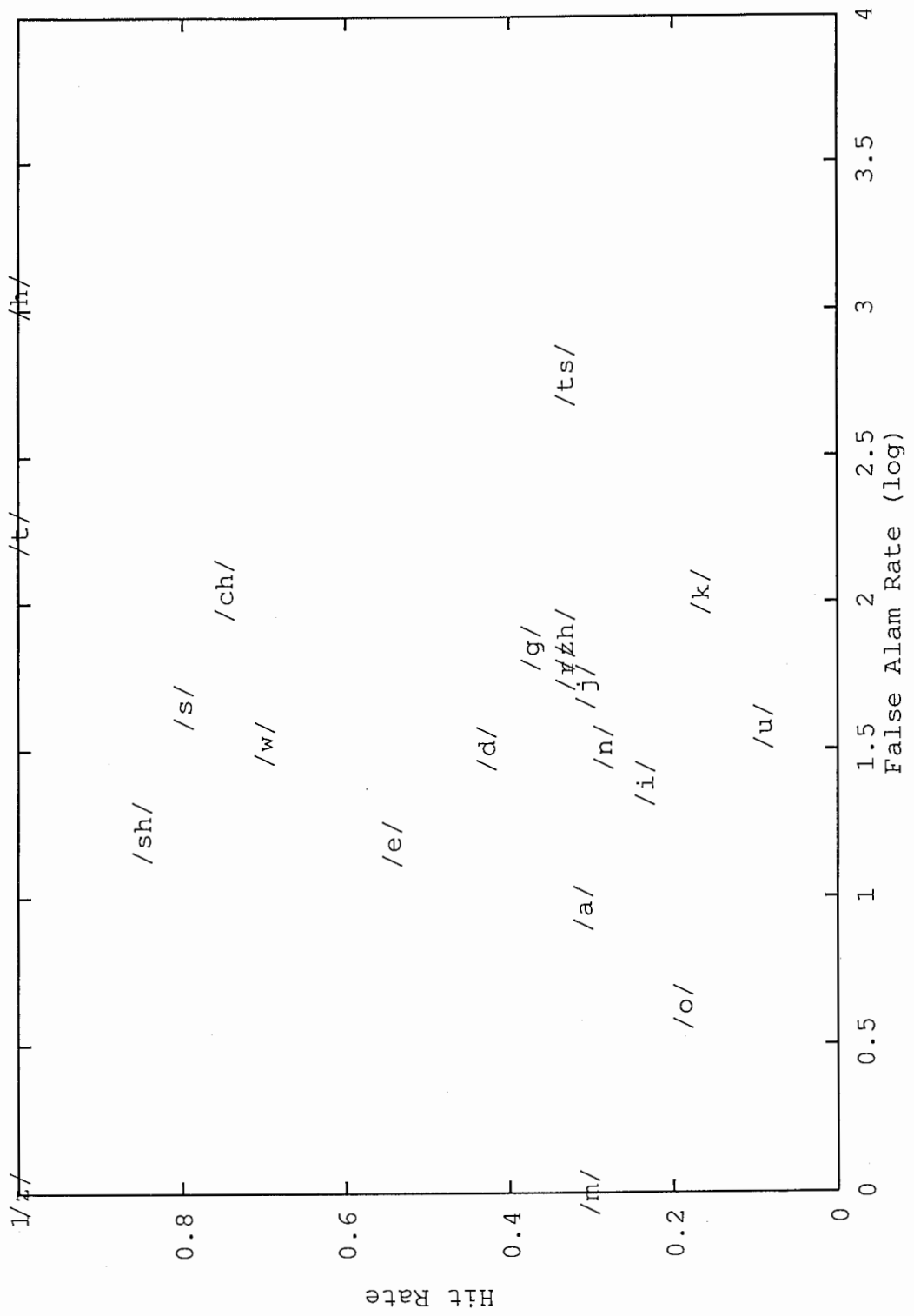

図 10: HR&FAR－ALTITUDE（±音素長＊5％）

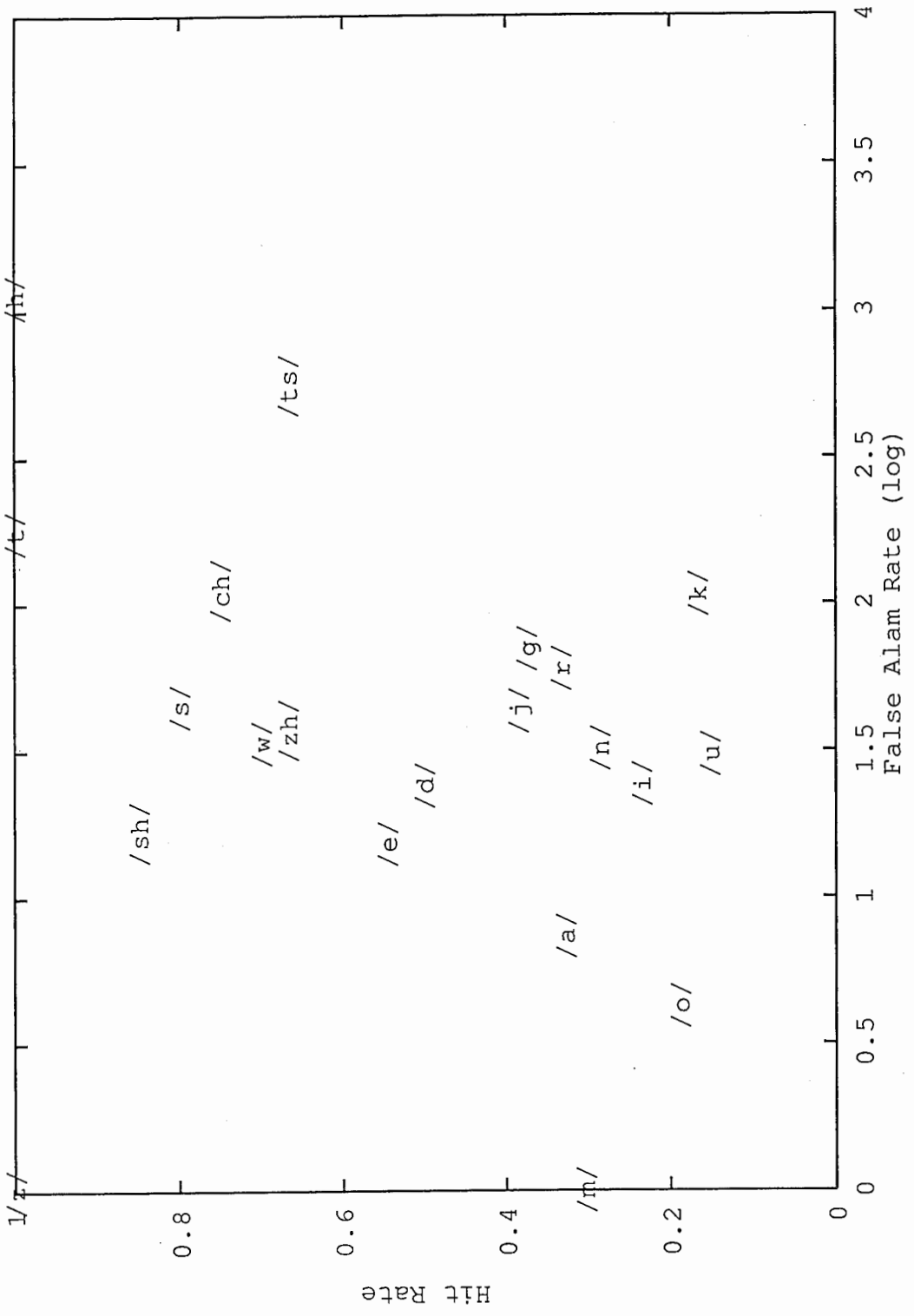

図 11: HR&FAR−ALTITUDE (±音素長＊10%)

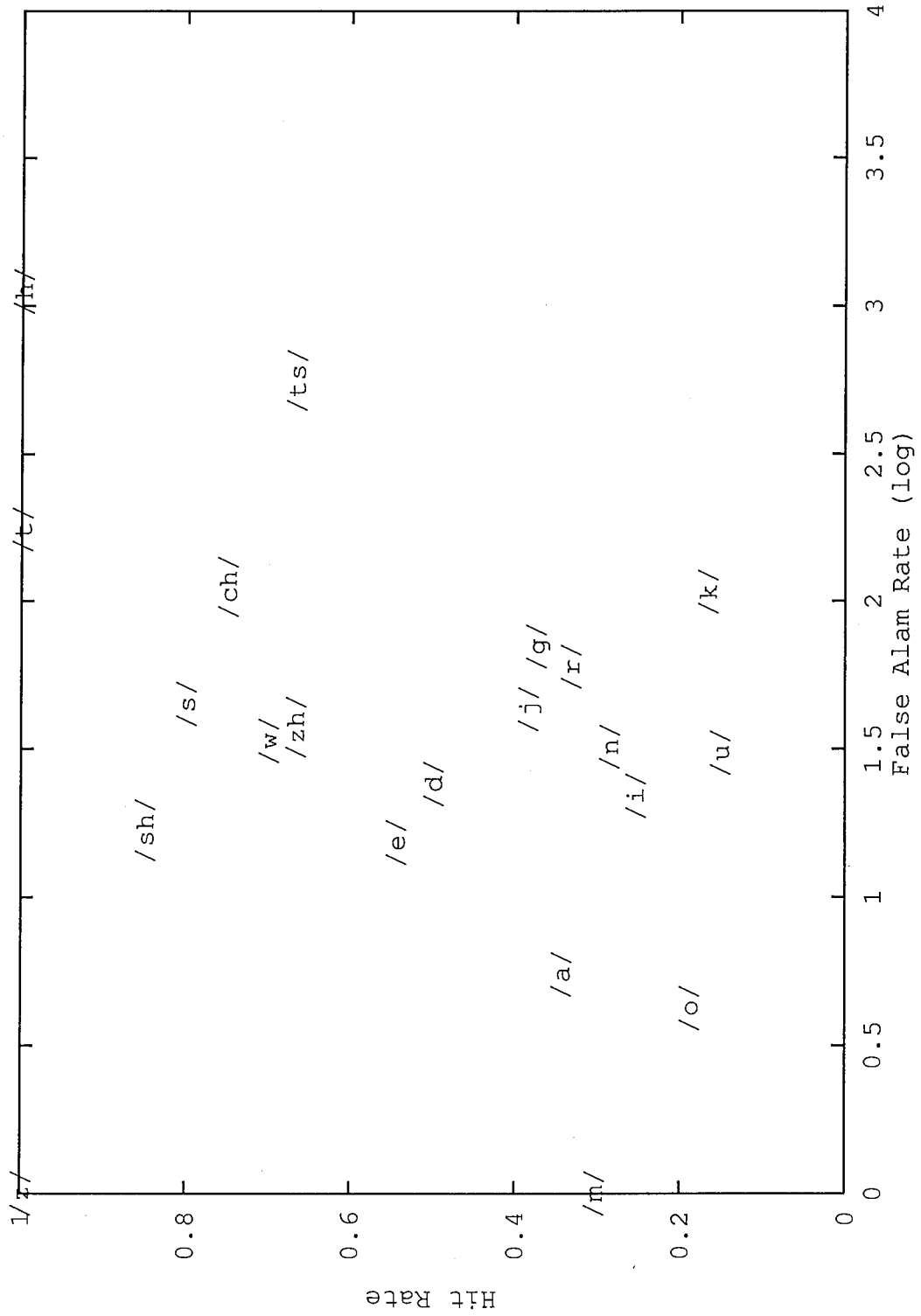

図 12: HR&FAR−ALTITUDE (±音素長＊15%)

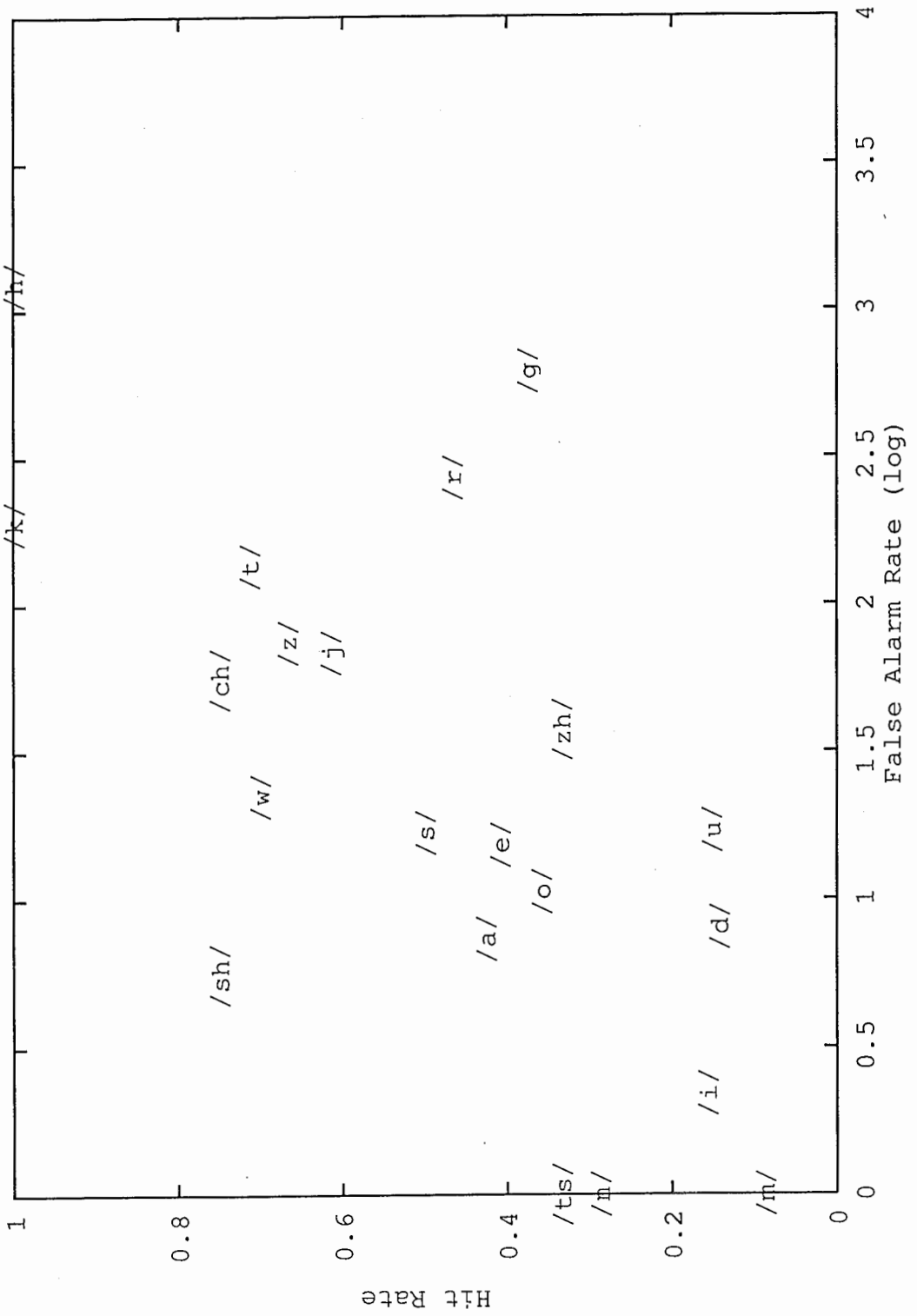

図 13: HR&FAR−BETA (±5Frame)

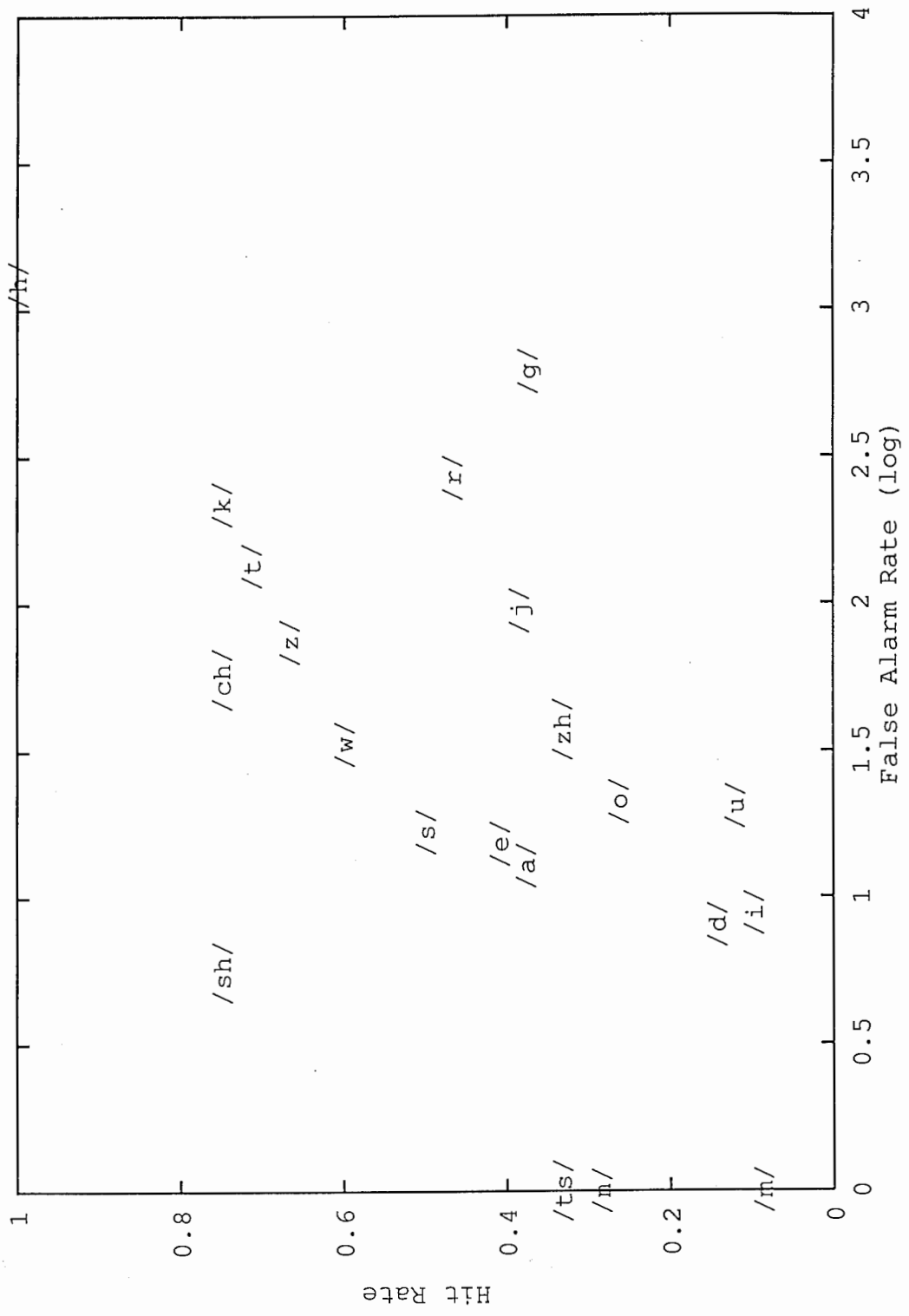

図 14: HR&FAR－BETA（±音素長＊5％）

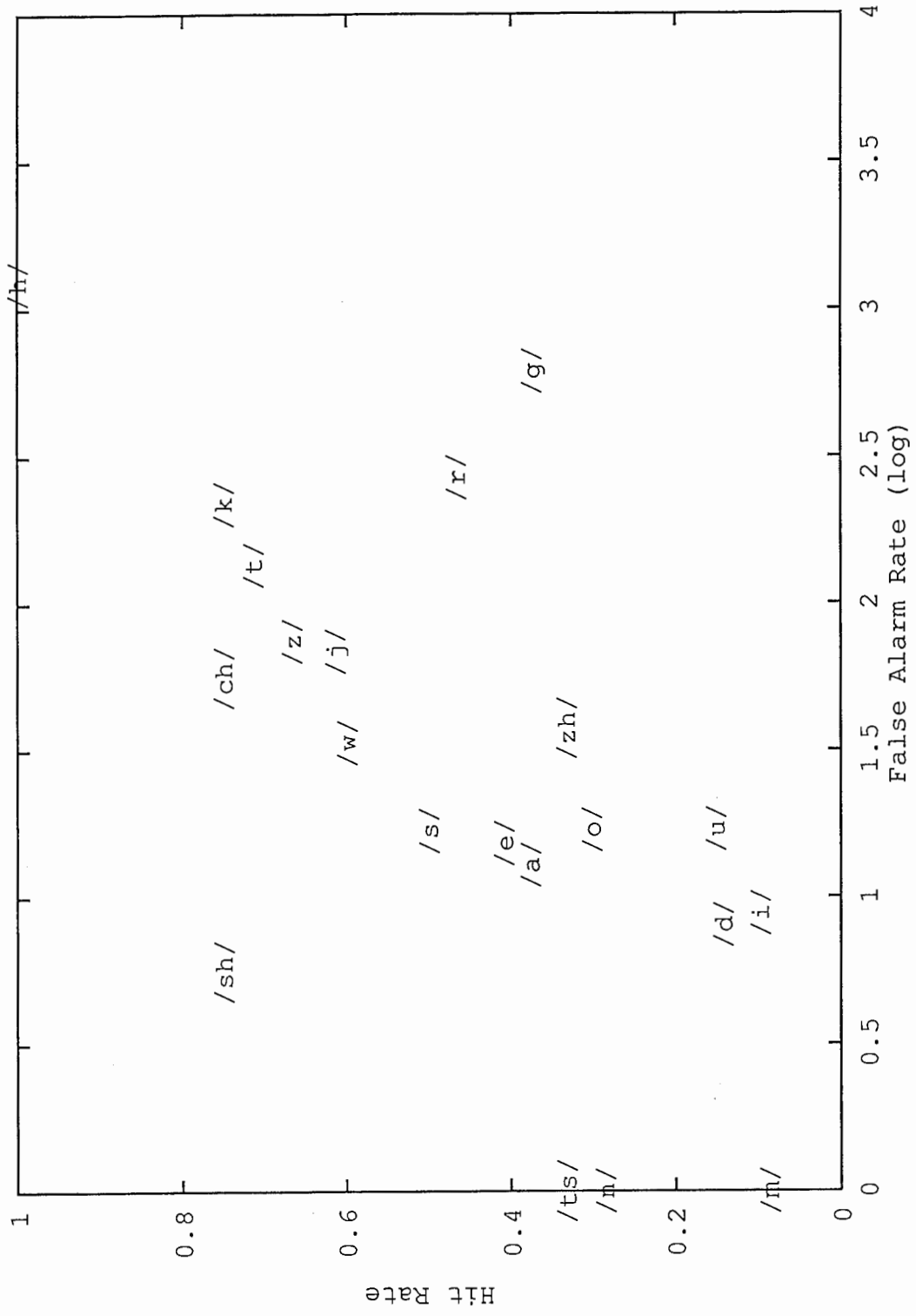

図 15: ＨＲ＆ＦＡＲ－ＢＥＴＡ（±音素長＊１０％）

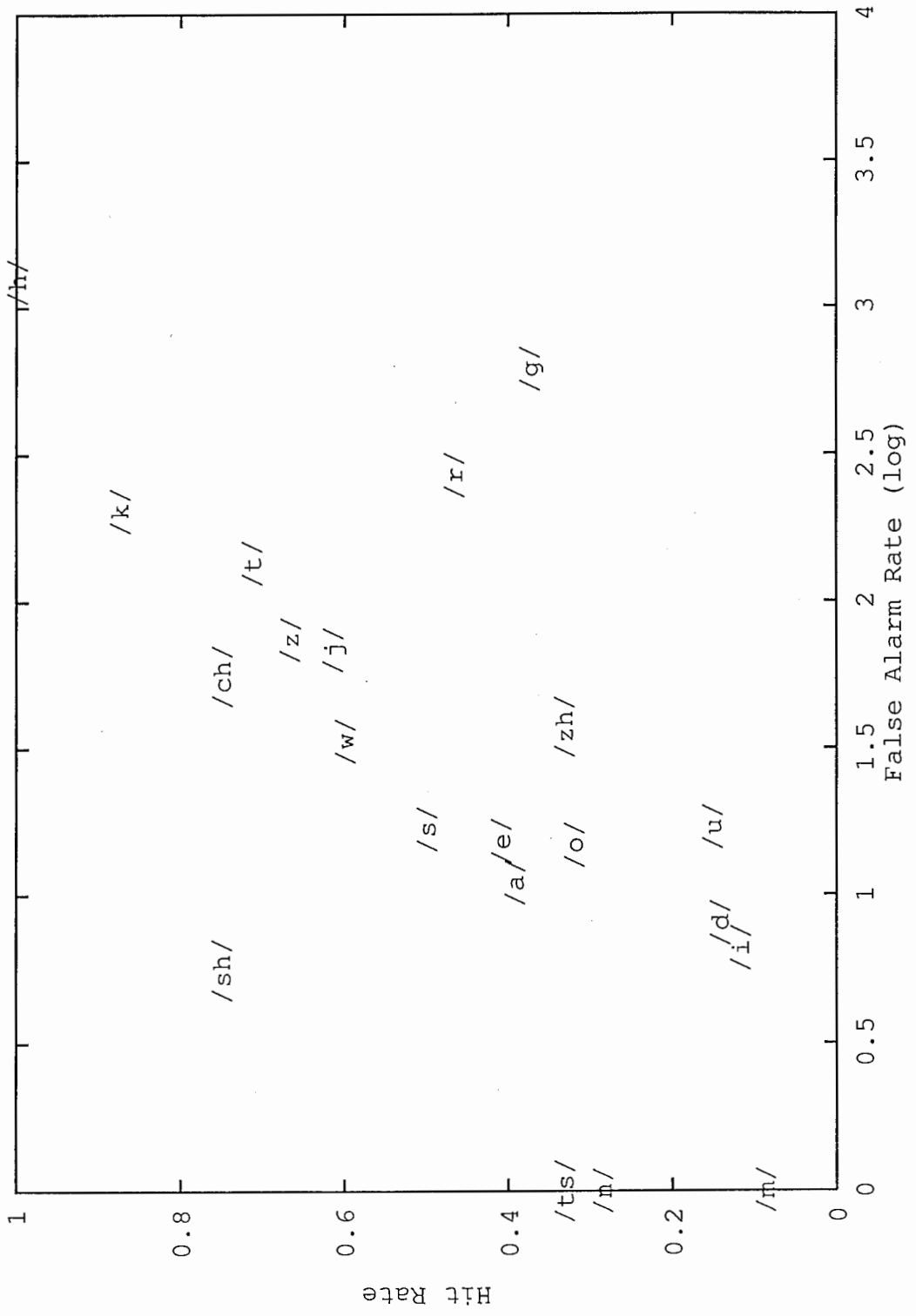

図 16: HR&FAR－BETA（±音素長＊15％）

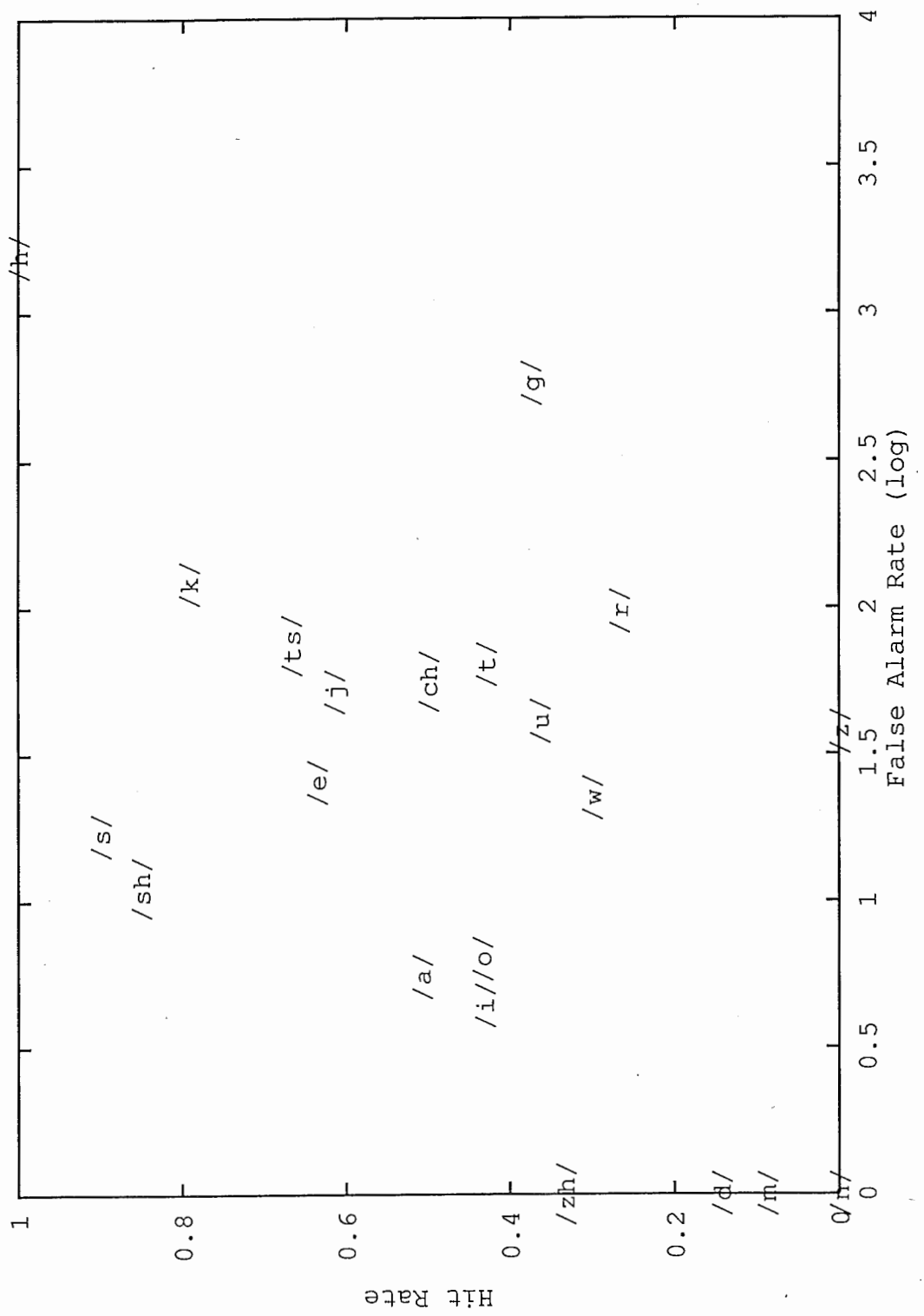

図 17: HR&FAR−AREA&VOLUME (±5Frame)

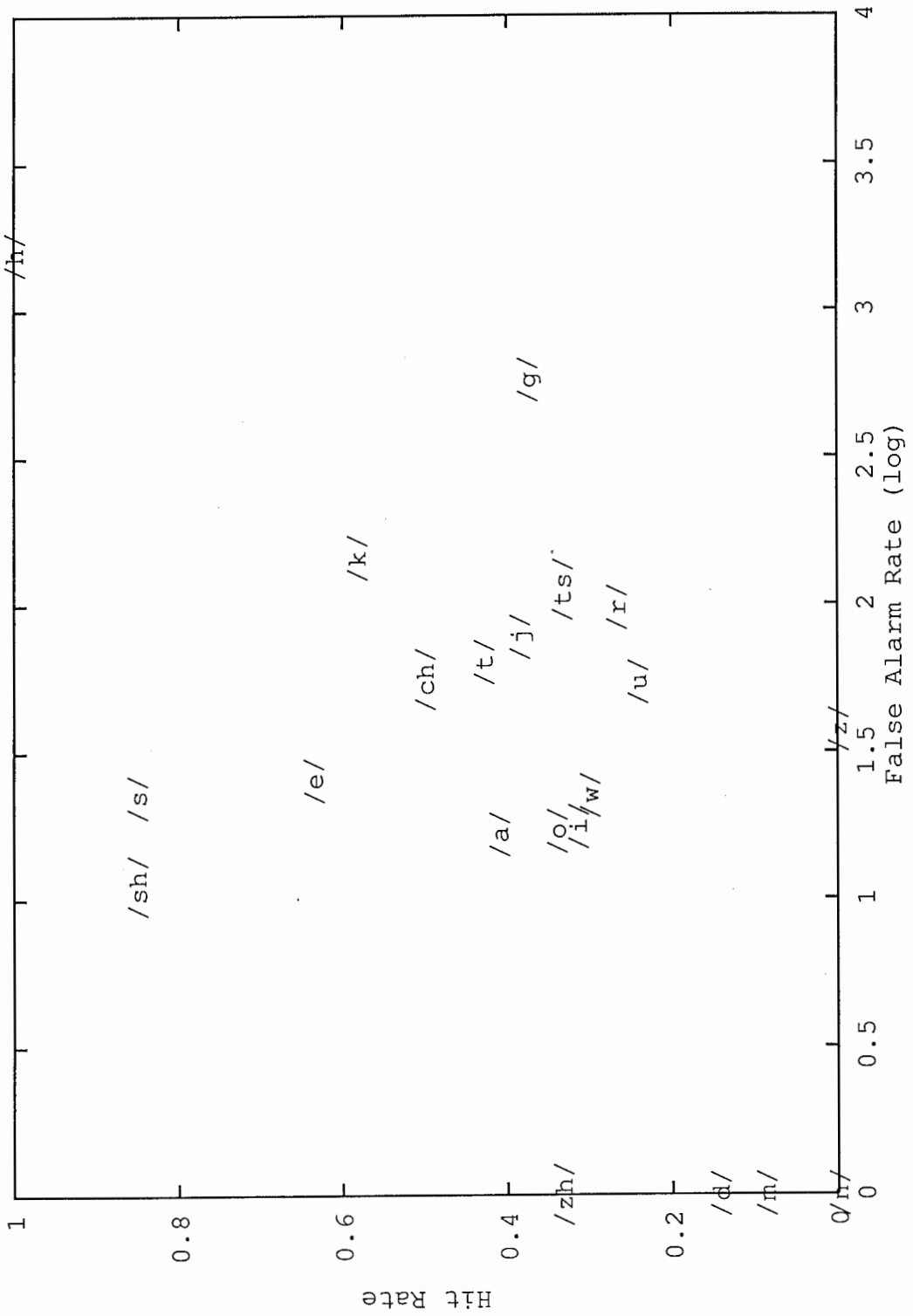

図 18: HR&FAR−AREA&VOLUME（±音素長＊5％）

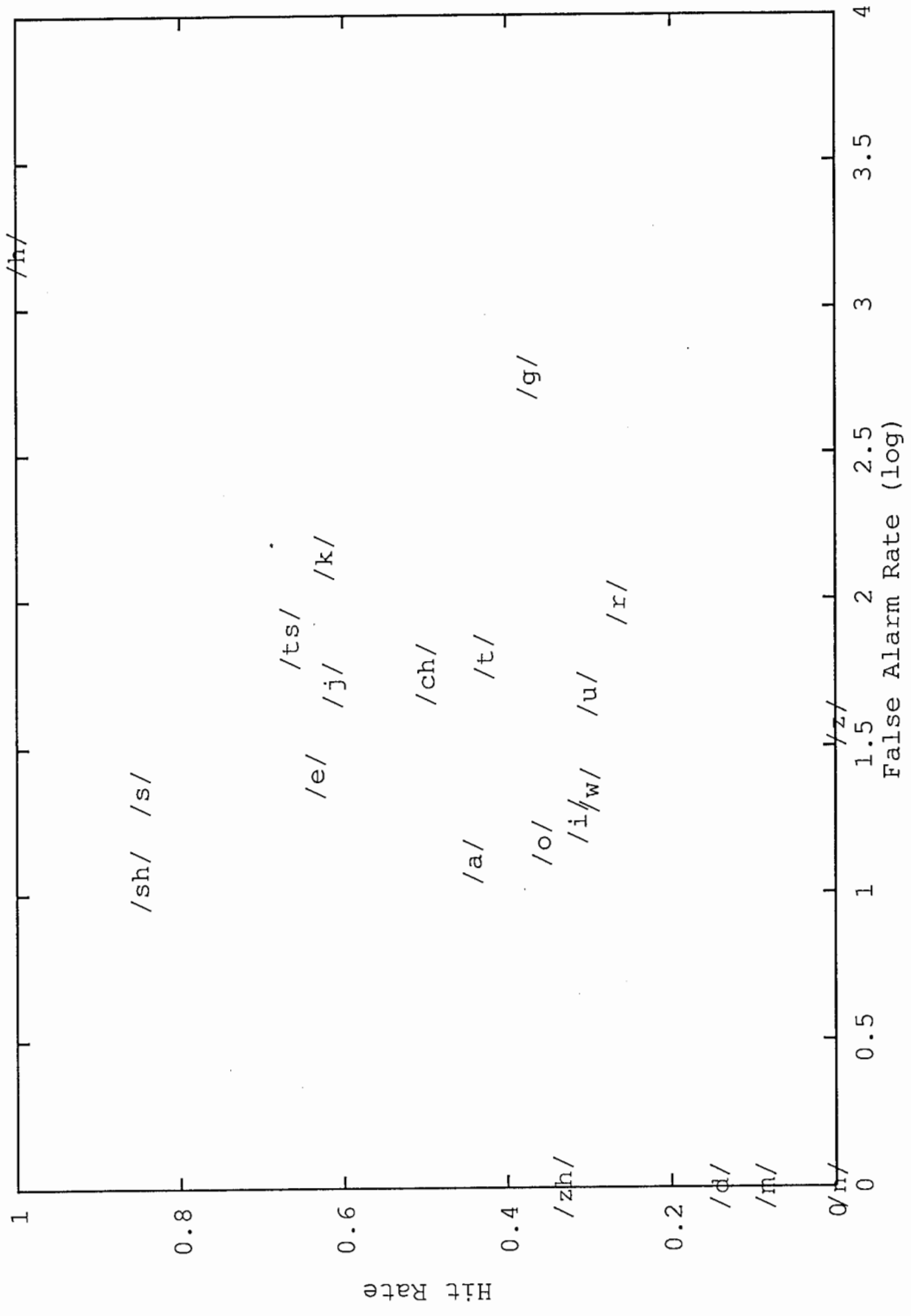

図 19: HR&FAR－AREA&VOLUME（±音素長＊１０％）

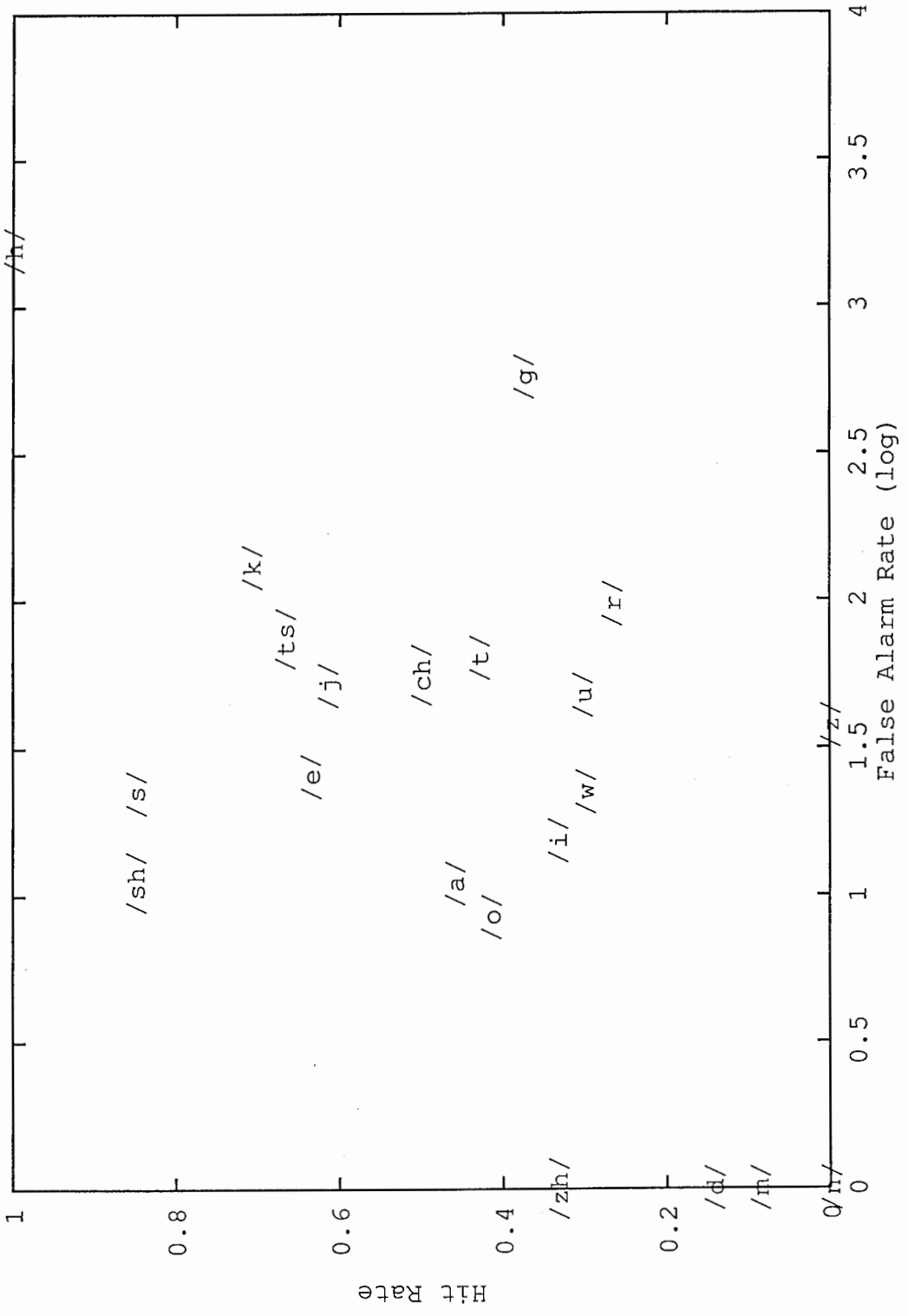

図 20: HR&FAR－AREA&VOLUME（±音素長＊15％）

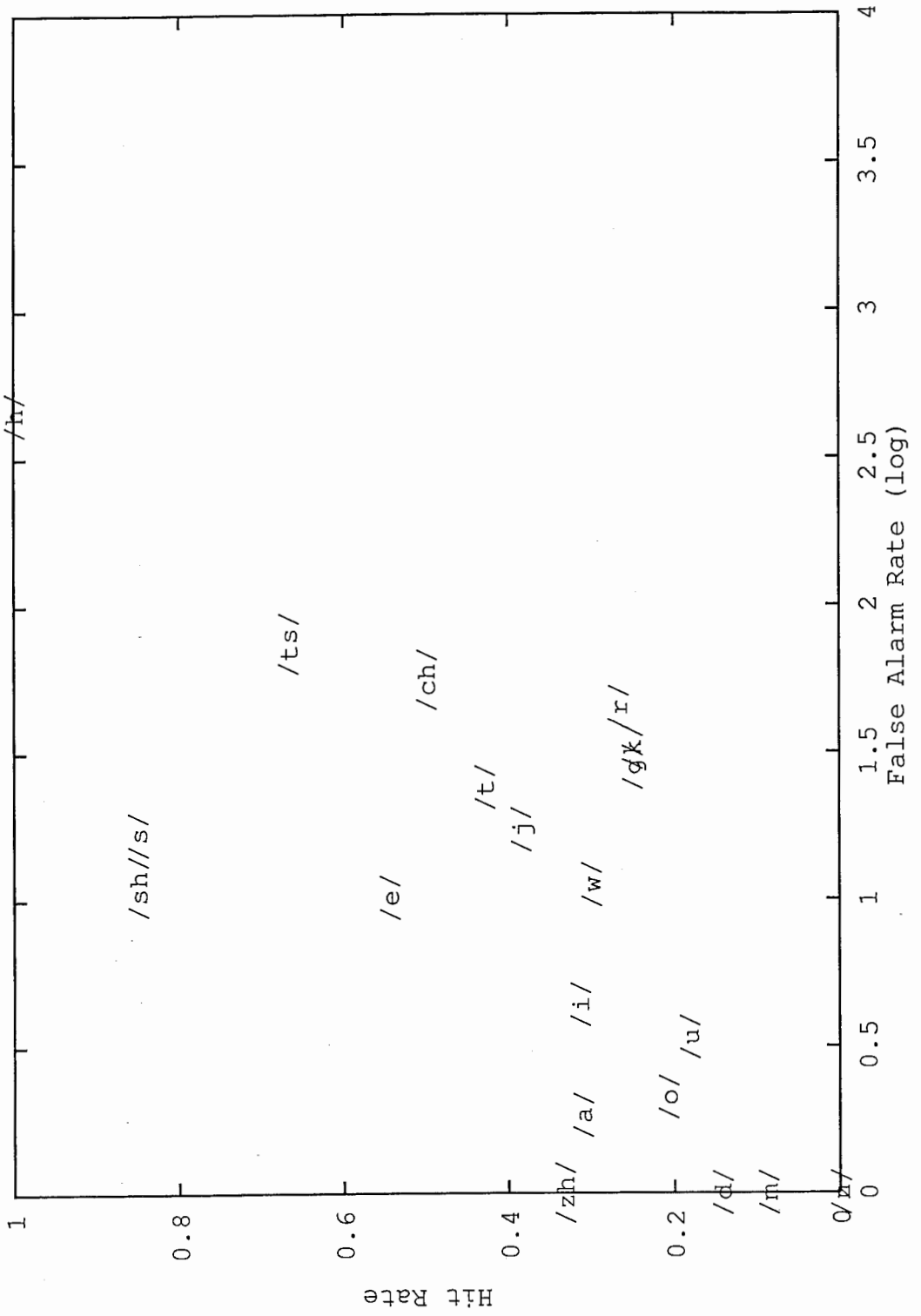

図 21: HR&FAR−AREA&ALTITUDE (±5Frame)

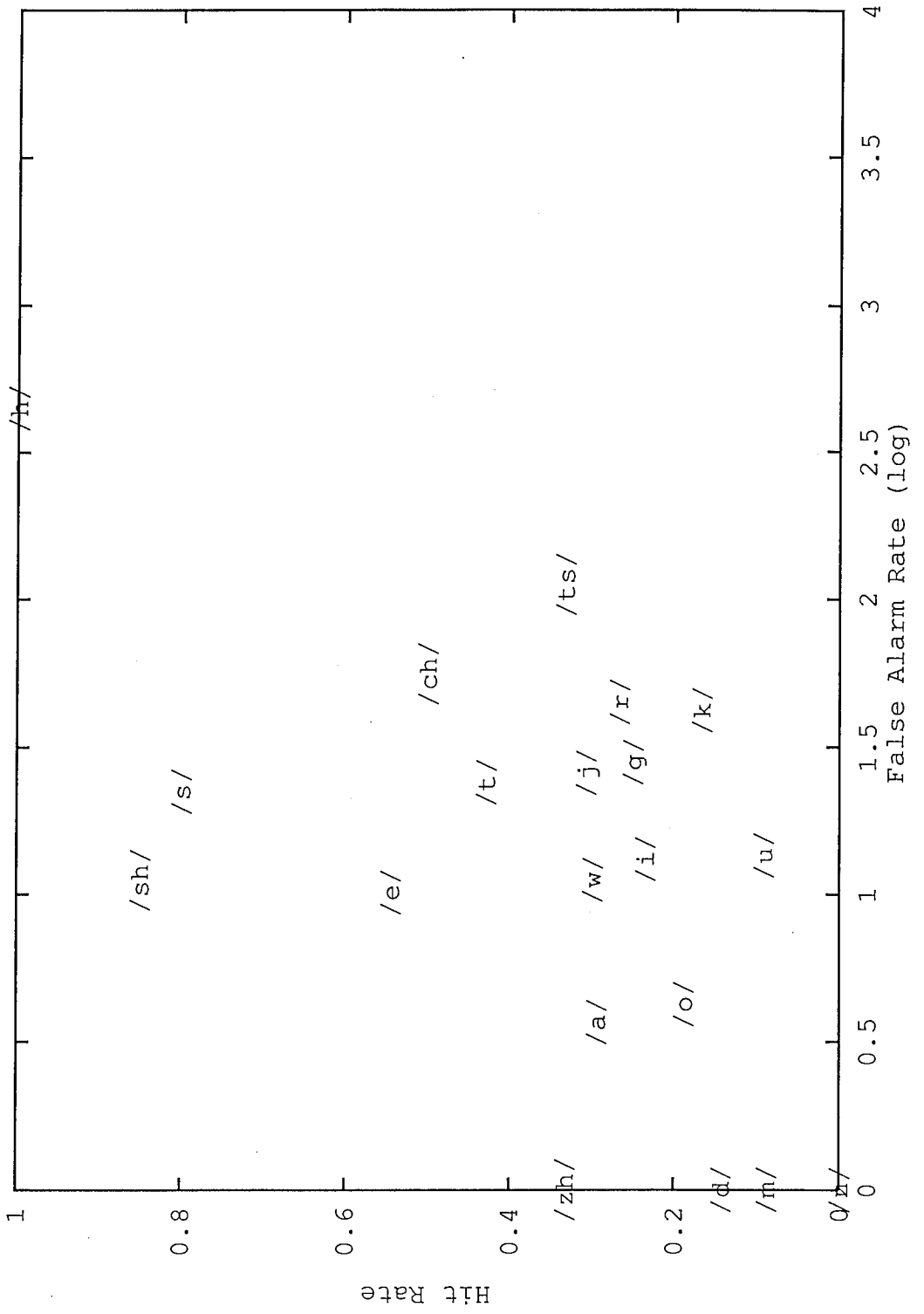

図 22: HR&FAR－AREA&ALTITUDE（±音素長＊5％）

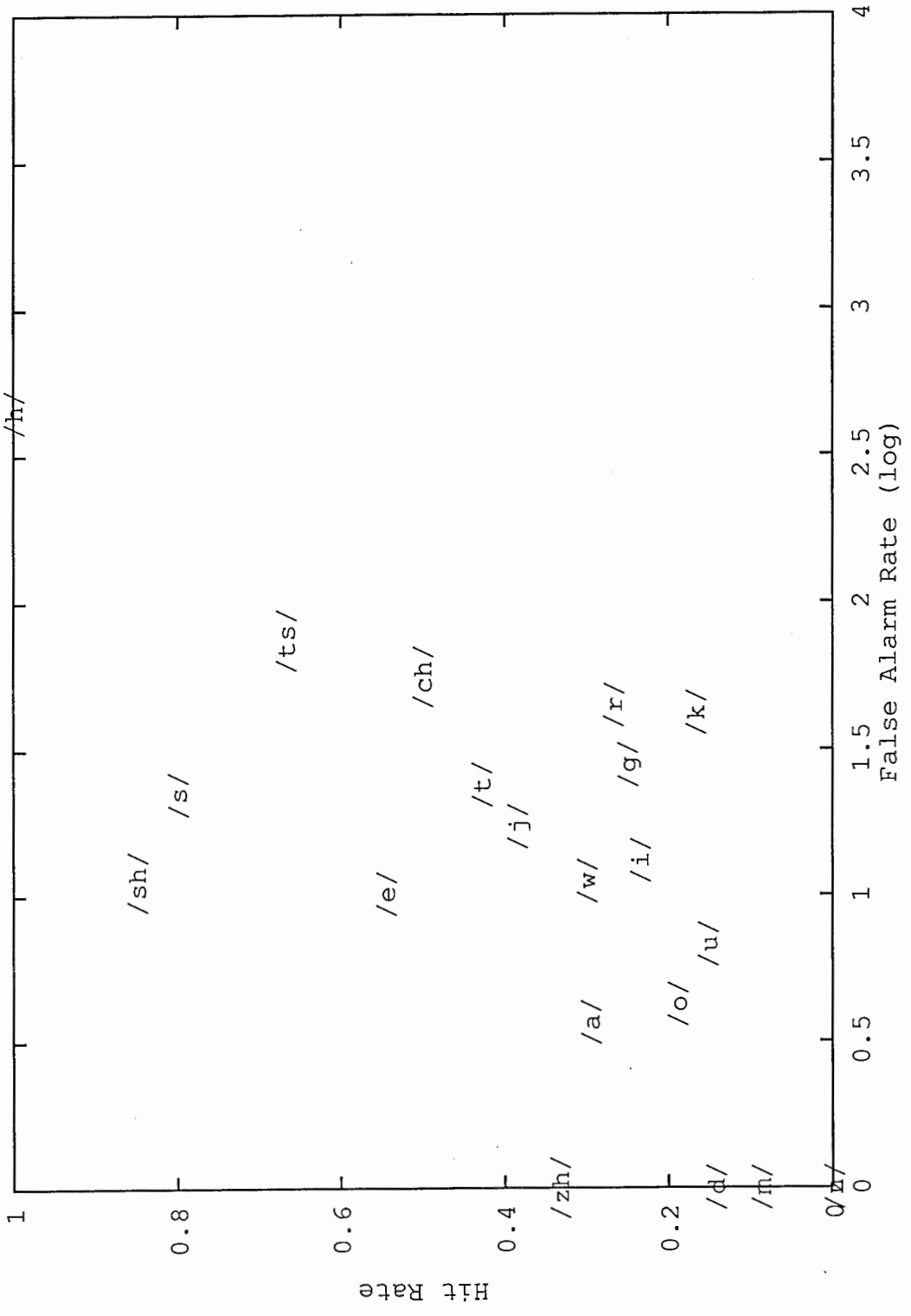

図 23: HR&FAR－AREA&ALTITUDE (±音素長＊10%)

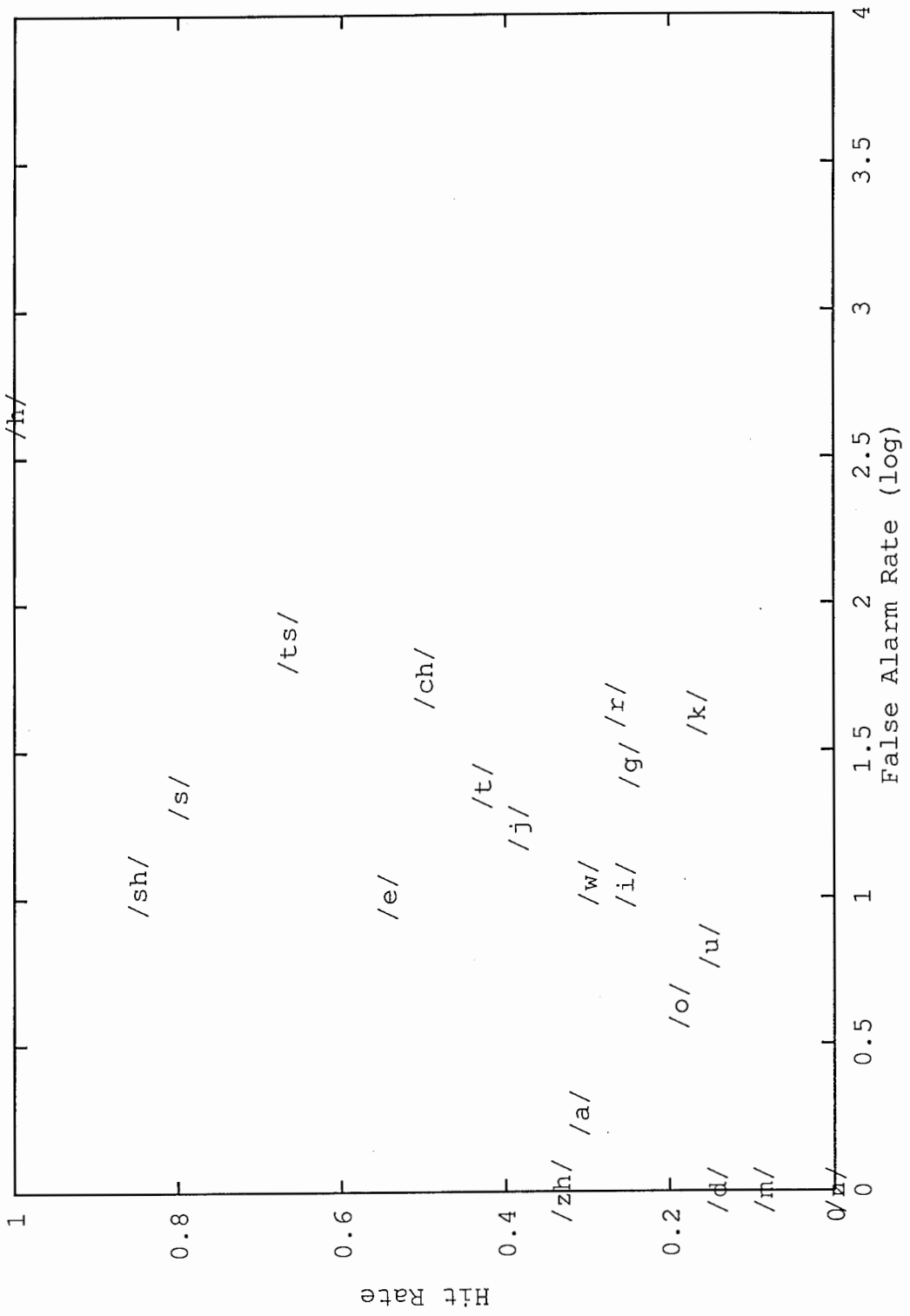

図 24: HR&FAR－AREA&ALTITUDE (±音素長＊15%)

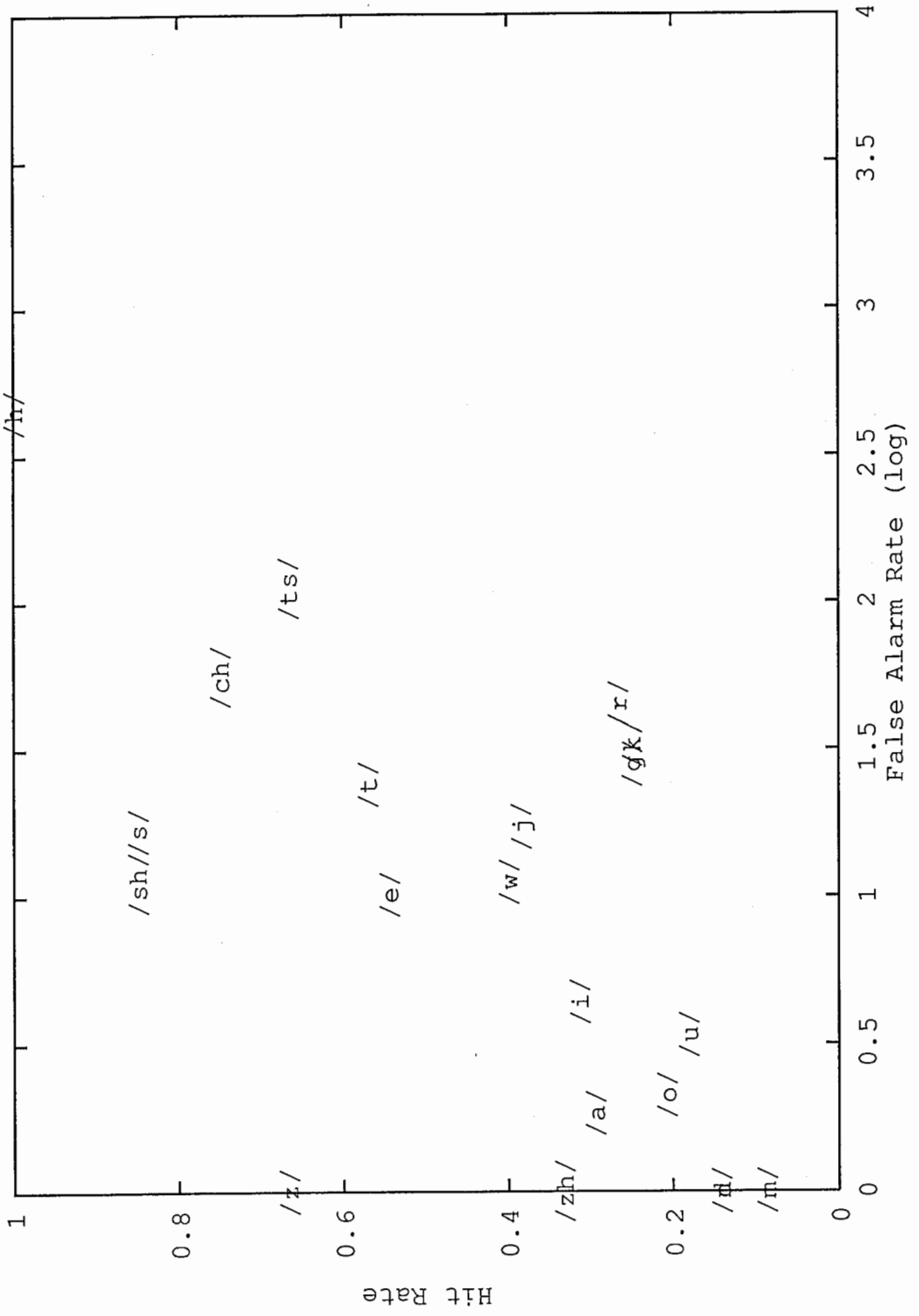

図 25: HR&FAR−VOLUME&ALTITUDE (±5Frame)

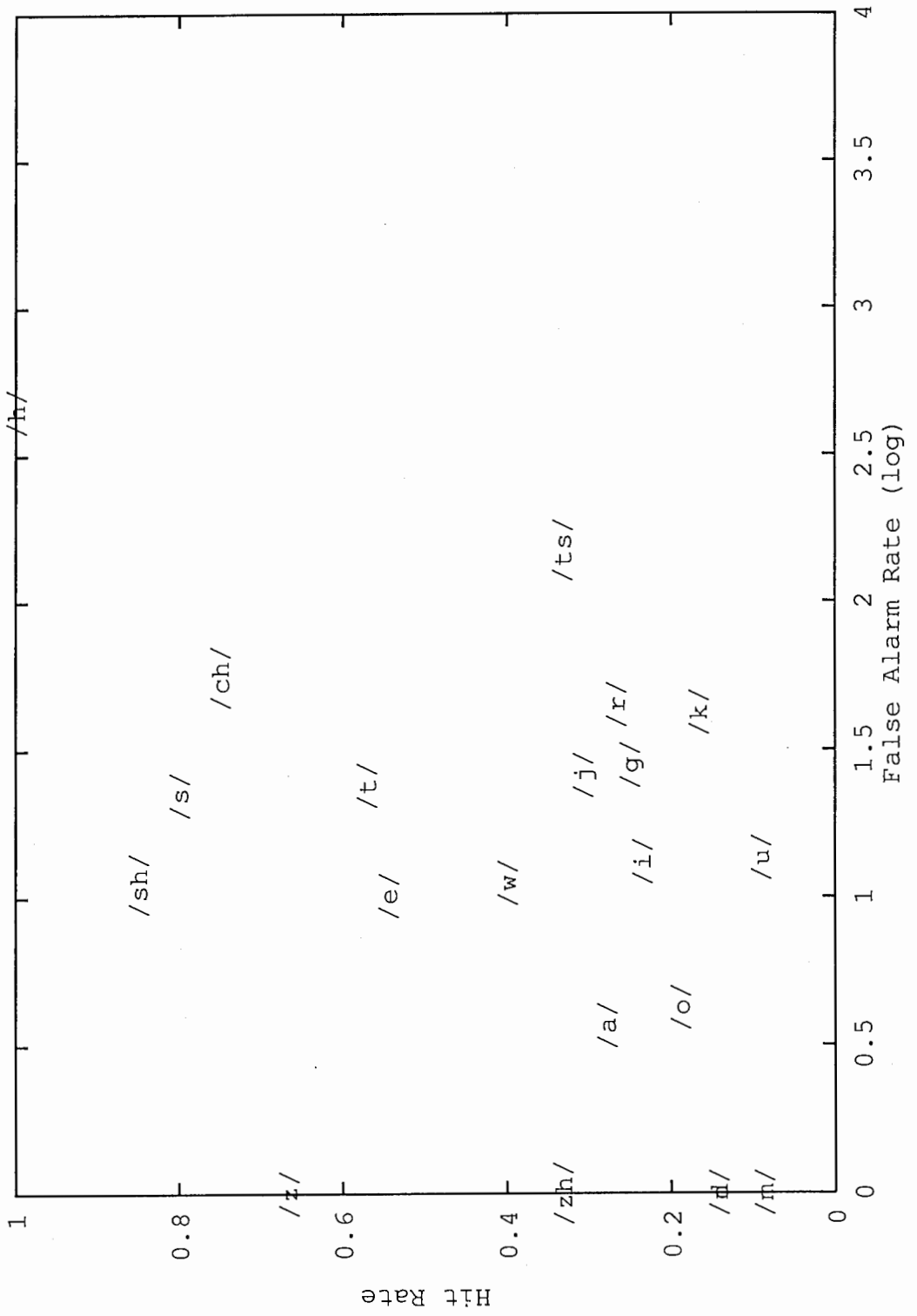

図 26: HR&FAR−VOLUME&ALTITUDE (±音素長＊5％)

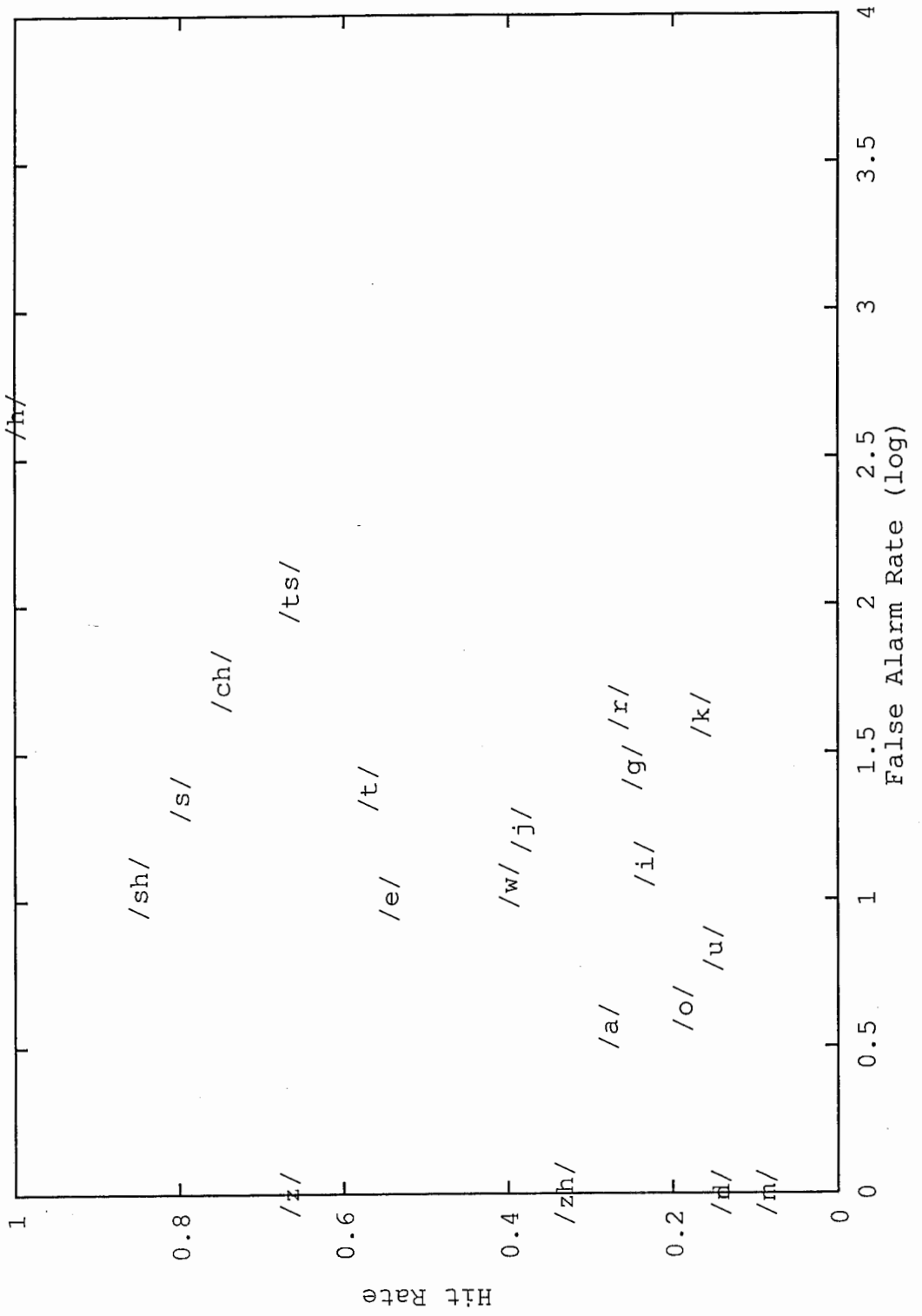

図 27: HR&FAR−VOLUME&ALTITUDE (±音素長＊10%)

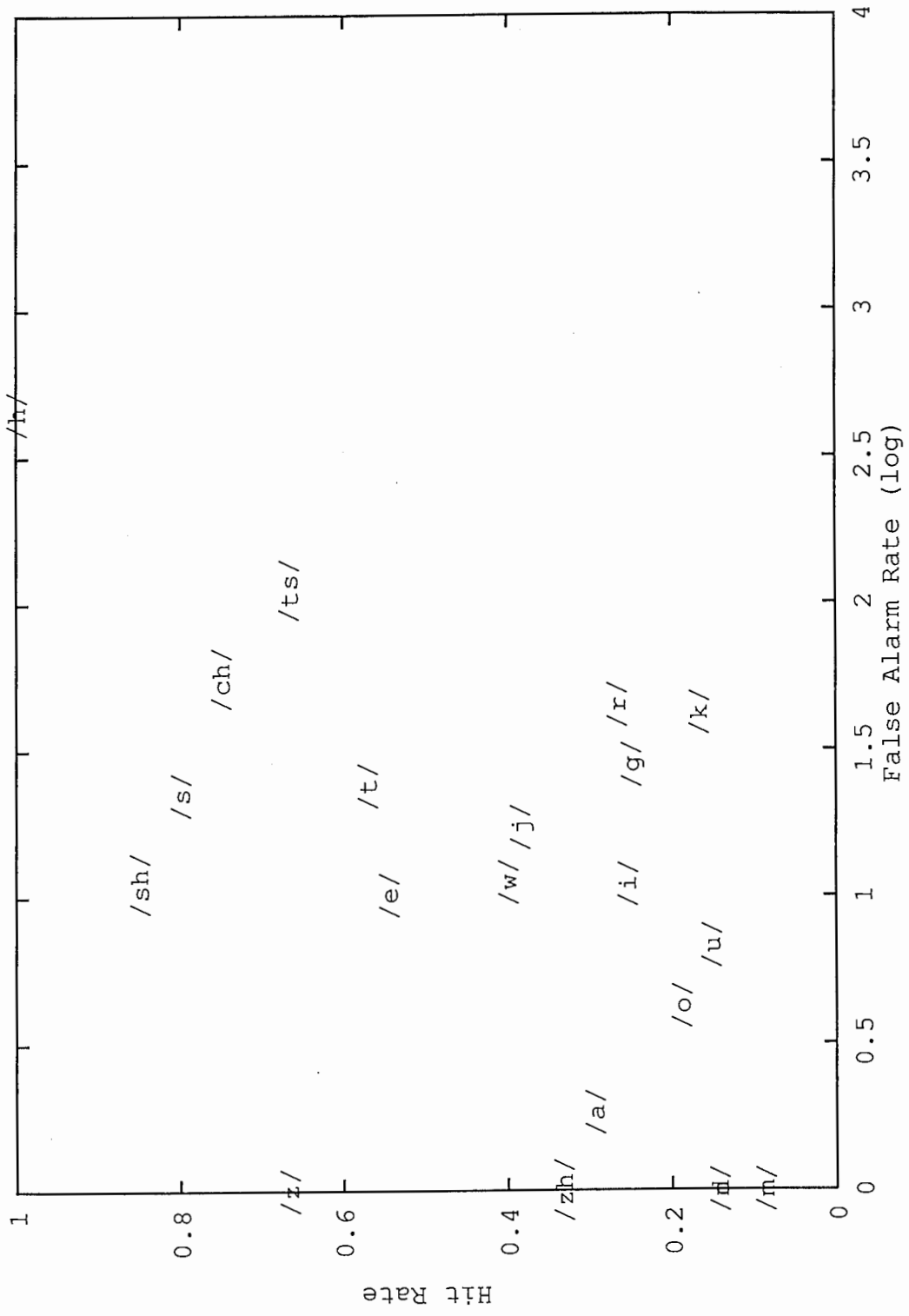

図 28: HR&FAR−VOLUME&ALTITUDE (±音素長＊15%)

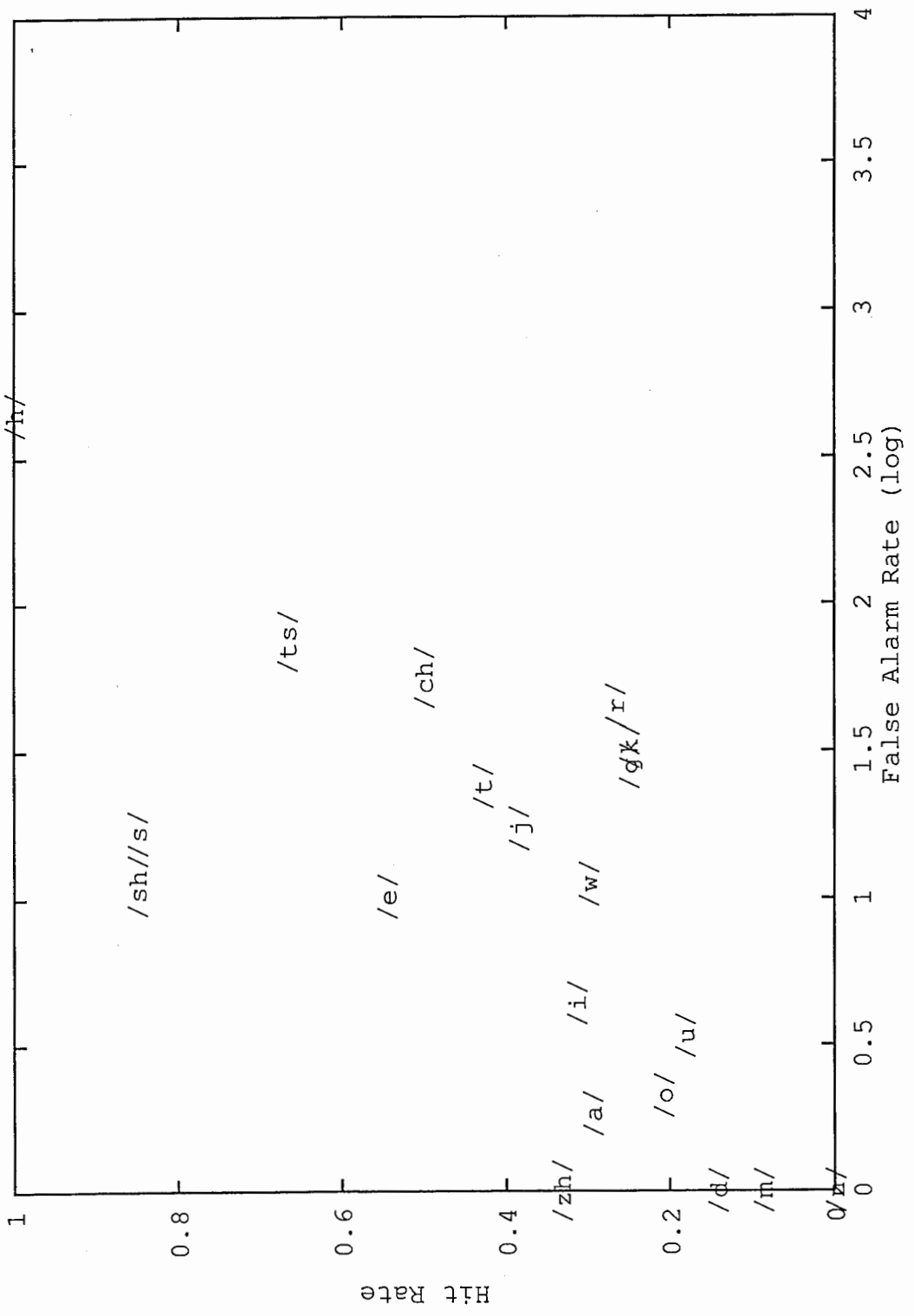

図 29: HR&FAR−AREA&VOLUME&ALTITUDE (±5Frame)

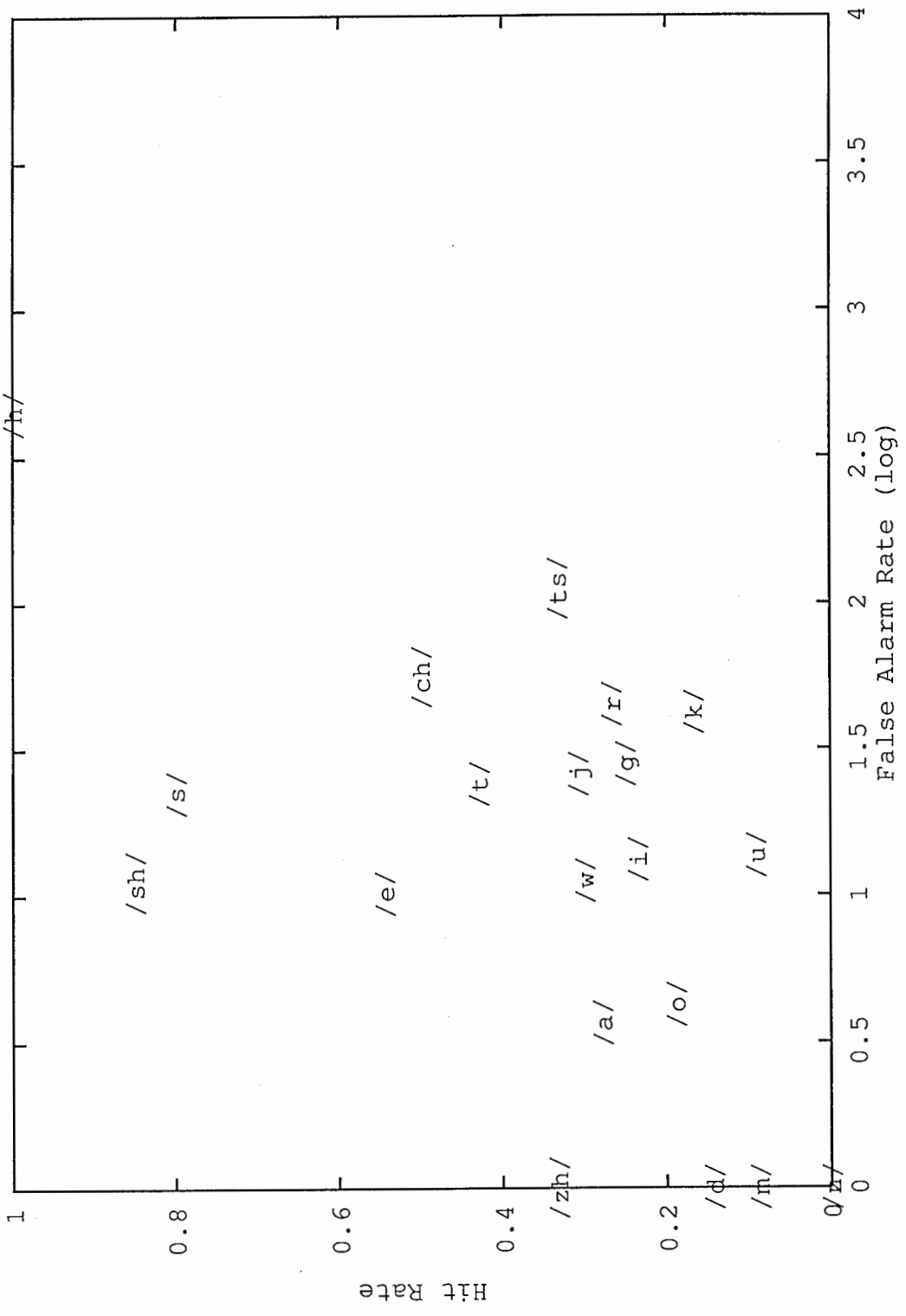

図 30: HR&FAR−AREA&VOLUME&ALTITUDE (±音素長＊5%)

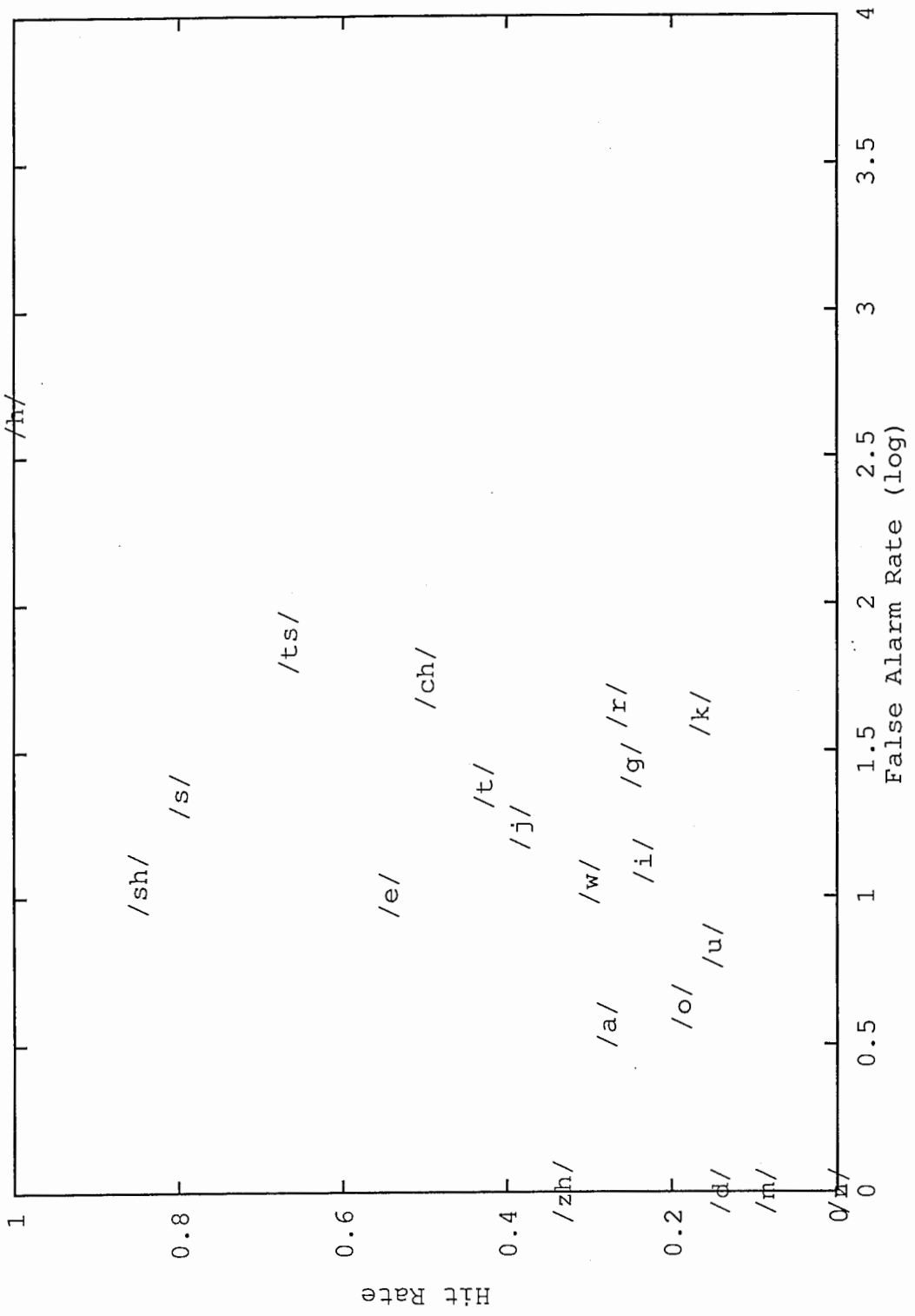

図 31: HR&FAR－AREA&VOLUME&ALTITUDE（±音素長＊10％）

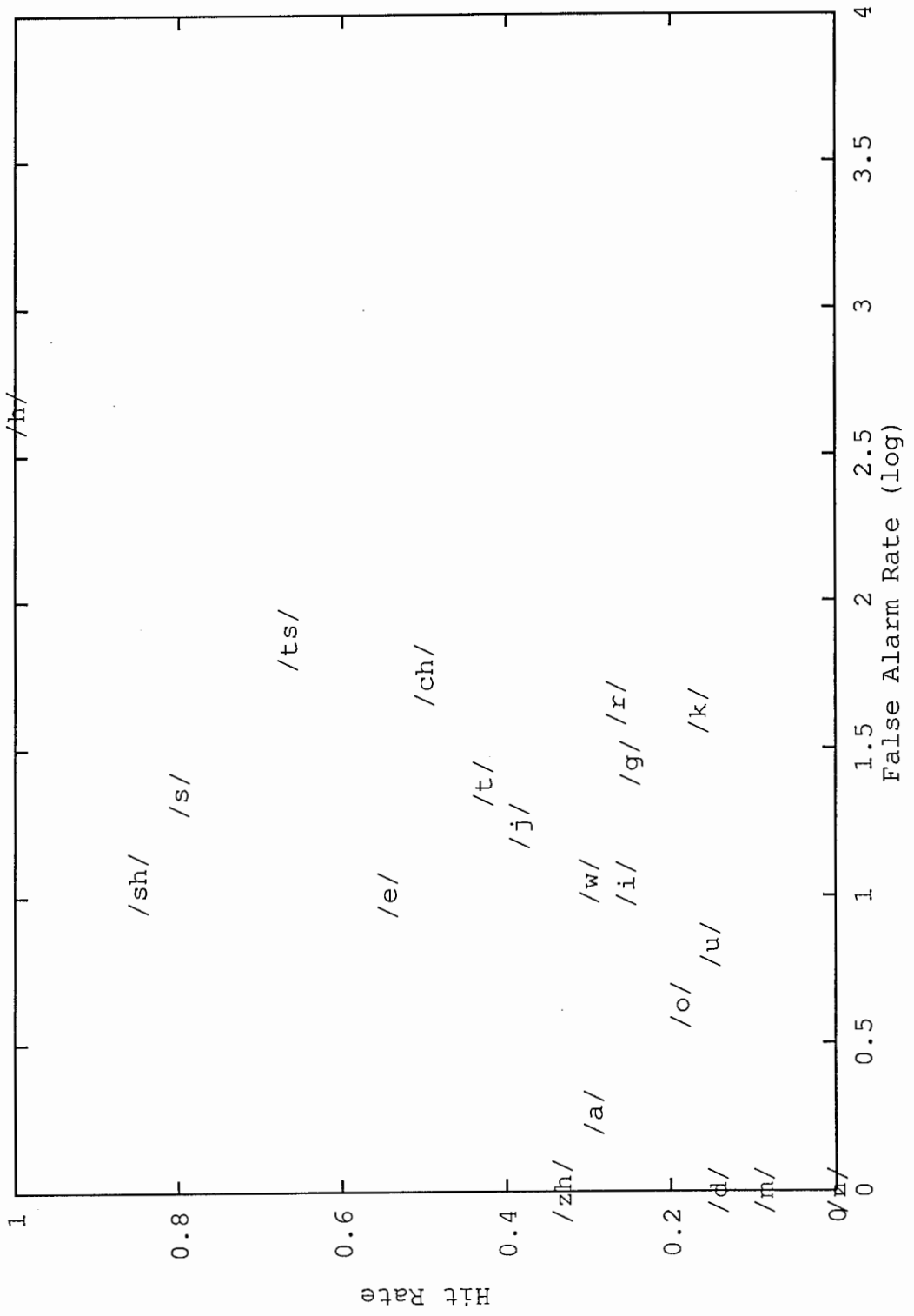

図 32: HR&FAR−AREA&VOLUME&ALTITUDE (±音素長*15%)

まず投影面積Ｔｏｐ１０、体積Ｔｏｐ１０、最大高さＴｏｐ１０、ＢｅｔａＴｏｐ１０の島データでこのグラフを作った場合、／ｓ／、／ｓｈ／（ＢｅｔａＴｏｐ１０のみ／ｓｈ／だけ）がすべて左上にきている。このことから／ｓ／、／ｓｈ／の音素は、ほとんどが認識され誤認識も少ないことがわかる。つまりもっとも信頼性の高い音素ということになる。また母音は投影面積、体積、最大高さ、Ｂｅｔａすべてにおいて左真中にきている。このことから母音は認識率はまあまあで誤認識は少ないことがわかる。日本語の場合、ほとんどの子音に母音がつくので、文節中に出現する母音の数は子音の数に比べはるかに多い。それなのに時間あたりの誤認識数が少ないということは、かなり信頼性があるということになる。投影面積、体積、最大高さＴｏｐ１０において／ｇ／、／ｈ／の音素は右に位置している。このことから／ｇ／、／ｈ／の信頼性はほとんどないことがわかる。その他の音素は大体中央に位置している。また音素の分類でこれらのグラフを見てみたが、母音が大体同じ位置に位置すること以外にこれといった特徴は見られなかった。

次に投影面積Ｔｏｐ１０＆体積Ｔｏｐ１０、投影面積Ｔｏｐ１０＆最大高さＴｏｐ１０、体積Ｔｏｐ１０＆最大高さＴｏｐ１０の島データでこのグラフを作った場合、投影面積&体積のグラフは投影面積のみのグラフとほとんど一緒であることがわかる。また、投影面積&最大高さ、体積&最大高さ、投影面積&体積&最大高さにおいては、投影面積Ｔｏｐ１０、体積Ｔｏｐ１０、最大高さＴｏｐ１０それぞれのみで、このグラフを書いた場合に比べ、まずすべての音素が左側によっていることがあげられる。つまりＦＡＲが小さくなってすべての音素の信頼性が向上したことを意味するが、その分ＨＲは下がっている。また音素の分類による区分がはっきりしてきた。まず、半母音である／ｊ／、／ｗ／が母音に近づいていることがあげられる。また同じ種類に分類される音素が、近い位置に集まっている。

7　考察

認識Ｓｃｏｒｅのみに注目した結果の場合、正解率は高いが音素正解率は大変低い。これは周辺の同じ音素ばかりがＴｏｐ１０内に出現したせいである。しかも／ｐ／，／ｑ／などの音素の出現数が大変多く、そのほとんどが誤認識である。これでは認識Ｓｃｏｒｅの高い音素が信頼性の高い音素であるとはいい難い。また日本語においてもっとも良く出現する母音がほとんどＴｏｐ１０内に出現していないことも問題である。しかしながらいくつかのある音素においては認識Ｓｃｏｒｅの高い音素が信頼性の高い音素とすることができる。／ｓ／、／ｓｈ／などの音素で認識Ｓｃｏｒｅが高ければかなり信頼性が高いといえる。しかしながら発声された音声で複数の信頼性のある音素が欲しい場合や／ｓ／、／ｓｈ／の出現していない音声で信頼性のある音素を探すのは認識Ｓｃｏｒｅのみからでは難しいと考えられる。

各音素ごとの音響マトリックスで投影面積、体積、最大高さ、ＢｅｔａのＴｏｐ１０を出した場合の結果では、全く良い結果は得られなかった。これは認識Ｓｃｏｒｅの低いものでも０より認識Ｓｃｏｒｅが高くなりやすい音素なら投影面積や体積が大きくなりやすいためであろう。そのため最大高さのＴｏｐ１０の場合は、ほんの少しだけ他の結果に比べ良い結果が出ている。

それに比べ最高認識Ｓｃｏｒｅの音素を集めた音響マトリックスでの投影面積、体積、最大高さＴｏｐ１０の結果では、正解率、音素正解率共にかなり良い結果を出している。またもっとも出現回数の多い母音の正解認識率が高く、信頼性が高いことは音声認識において非常に有用であると考えられる。また投影面積&最大高さ、体積&最大高さ、投影面積&体

積＆最大高さの様に複数のＴｏｐ１０に入っている音素を信頼性の高い音素とすることも有用であると考えられる。とくに投影面積＆体積＆最大高さでの信頼性の高い音素としては、母音、／ｊ／、／ｓ／、／ｓｈ／、／ｗ／などがあげられる。

このことから最大認識Ｓｃｏｒｅの音素を集めた音響マトリックスにおいて島という概念を用いることにより、認識Ｓｃｏｒｅのみから信頼性の高い音素を得る方法に比べ、信頼性を高い音素を得やすいことがわかる。

謝辞

本研究の機会を与えて下さった、ATR 音声翻訳通信研究所社長山崎泰弘氏に感謝いたします。また、適切な御助言、御激励を頂いた第 4 研究室室長森元 逞氏をはじめとするＡＴＲ音声翻訳通信研究所の皆様に、心から感謝いたします。

A　プログラムの作成と改良

今回、音響マトリックスを分析するためにいくつかのプログラムの作成と改良を行なった。

A.1　ラベル単位変換プログラムの作成

オートラベリング結果はｍｓｅｃ単位で表示されており、認識ＳｃｏｒｅはＦｒａｍｅ単位でつけられている。よって単位の変換が必要となる。今回はオートラベリングの単位をＦｒａｍｅ単位に変換した。

プログラムの位置
/home/atrp24/xitoh/Island/Display/labelchange.c
(使用方法)
labelchange 文節番号

プログラム内のＩｎｐｕｔＦｉｌｅＡが音声認識用ラベルファイル名、ＩｎｐｕｔＦｉｌｅＢがオートラベリング結果ファイル名、ＯｕｔｐｕｔＦｉｌｅが単位を変更したオートラベリング結果ファイル名である。

音声認識用ラベルファイルは、
/home/atrp24/xitoh/TEMP/FAK_MA2_01.LB ～ FAK_MA2_40.LB

オートラベリング結果ファイルは、
/home/atrp24/xitoh/TEMP2/FAK_MA2_01.LB ～ FAK_MA2_40.LB

Ｆｒａｍｅ単位オートラベリング結果ファイルは、
/home/atrp24/xitoh/Island/Display/temp/FAK_MA2_01.LB ～ FAK_MA2_40.LB
に実際に使用したファイルと結果がある

A.2　音素配置図表示と音素対応ファイル作成プログラム

最大認識Ｓｃｏｒｅの音素をマトリックス状に配置した音素配置図を標準出力に表示する。またオートラベリングによって確認された４０文節の各音素の位置を音素配置図に白抜きにすることにより示した。またＣｏｎｆｕｓｉｏｎＭａｔｒｉｘを作成するために、４０文節中に出てくる各音素と、音素配置図においてその音素のオートラベリングによって確認された位置（つまりＳｔａｒｔ－ＦｒａｍｅとＥｎｄ－Ｆｒａｍｅ）に出てくる音素をファイルに出力した。

プログラムの位置
/home/atrp24/xitoh/Island/Display/maxscore.c
(使用方法)
maxscore 文節番号 オプション
オプション：
　-L 数値（０～１０００）　　WaterLevel の設定

-P 数値	PlateauWidth の設定
-m	matrix 表示フラグ
-s	island slice フラグ

プログラム内のＩｎｐｕｔＦｉｌｅＡは先に述べたラベル単位変換プログラムによってＦｒａｍｅ単位に変換されたオートラベリング結果ファイル名、ＯｕｔｐｕｔＦｉｌａＡは最大認識Ｓｃｏｒｅにおける実際に発声された各音 素と音素配置図における音素の対応結果の出力ファイル名、ＯｕｔｐｕｔＦｉｌｅＢはＮｏ２認識Ｓｃｏｒｅにおける実際に発声された各音素と音素配置図における音素の対応結果の出力ファイル名である。関数 OutputMaxMatrix は上の音素配置図表示と音素対応ファイルの作成を行ない、関数 OutputMaxMatrix2 はＳｔａｒｔ－ＦｒａｍｅとＥｎｄ－ｆｒａｍｅの組合せにおける最高認識Ｓｃｏｒｅの音素配置から逆に音声認識結果ファイルを作成する。これはＩｓｌａｎｄ表示プログラムを変更せずに、最高認識ＳｃｏｒｅのＩｓｌａｎｄを表示するために行なったものである。

音素対応結果ファイルは
/home/atrp24/xitoh/Island/Display/result5/Conf_Matrix_(01～40).No1
に実際の結果ファイルがある。

A.3　文節の各音素数の計算プログラム

４０文節中に各音素がどれだけの数出現しているか計算するプログラム。

プログラムの位置
/home/atrp24/xitoh/Island/Display/CountOnso.c
（使用方法）
CountOnso

プログラム内のＩｎｐｕｔＦｉｌｅＡはＦｒａｍｅ単位か、ｍｓｅｃ単位のオートラベリング結果ファイル名、ＯｕｔｐｕｔＦｉｌｅＡは、４０文節中の各文節に各音素がどれだけ出現したかの結果ファイル名、ＯｕｔｐｕｔＦｉｌｅＢは４０文節中に各音素がどれだけ出現したかの結果ファイル名である。

各文節における各音素の出現回数結果ファイル
/home/atrp24/xitoh/Island/Display/result6/CountOnso-%02d

４０文節における各音素の出現回数結果ファイル
/home/atrp24/xitoh/Island/Display/result6/CountOnso-Total
に実際の結果ファイルがある。

A.4　Ｃｏｎｆｕｓｉｏｎ Ｍａｔｒｉｘ作成プログラム

音素対応ファイルからＣｏｎｆｕｓｉｏｎ Ｍａｔｒｉｘを作成するプログラム。Ｃｏｎｆｕｓｉｏｎ Ｍａｔｒｉｘは標準出力に出力される。

プログラムの位置
/home/atrp24/xitoh/Island/Display/CountOnso2.c
（使用方法）
CountOnso2

プログラム内のＩｎｐｕｔＦｉｌｅＡは前述の音素対応ファイル名である。

A.5　Ｉｓｌａｎｄ表示プログラムの改良

島のデータを自動的に計算し、表示するプログラムを、投影面積、体積、最大高さ、Ｂｅｔａ（Ｓｈａｐｅ　Ｐａｒａｍｅｔｅｒ）の４つのデータの範囲を指定することによってその範囲に入る島のデータだけを表示するように改良した。またその島のデータを投影面積、体積、最大高さ、ＢｅｔａでソートしＴｏｐｎまで出力できるようにした。また出力されたＴｏｐｎまでに出現した島の最大高さの位置が、オートラベリングされた音素の位置にと一致しているか判定している。さらに投影面積Ｔｏｐｎ＆体積Ｔｏｐｎ、投影面積Ｔｏｐｎ＆最大高さＴｏｐｎ、体積Ｔｏｐｎ＆最大高さＴｏｐｎ、投影面積Ｔｏｐｎ＆体積Ｔｏｐｎ＆最大高さＴｏｐｎなど両方、あるいはすべてに出現している音素でオートラベリングされた音素の位置と一致しているか判定している。

プログラムの位置
/home/atrp24/xitoh/Island/Display/island.c　：
　複合用（投影面積＆体積など）
/home/atrp24/xitoh/Island/Display/island5.c　：
　単体用（投影面積、体積など）
（使用方法）
island 文節番号 オプション…
オプション：

-L 数値（０～１０００）	WaterLevel の設定
-P 数値	PlateauWidth の設定
-m	matrix 表示フラグ
-s	island slice フラグ
-A 数値	投影面積が数値以上のみ表示
-a 数値	投影面積が数値以下のみ表示
-V 数値	体積が数値以上のみ表示
-v 数値	体積が数値以下のみ表示
-H 数値	最大高さが数値以上のみ表示
-h 数値	最大高さが数値以下のみ表示
-B 数値	Ｂｅｔａが数値以上のみ表示
-b 数値	Ｂｅｔａが数値以下のみ表示

A.6　認識ＳｃｏｒｅＴｏｐｎ表示プログラムの作成

各文節の各音素の認識Ｓｃｏｒｅのみに注目し、すべての音素の中で認識Ｓｃｏｒｅの多いものからＴｏｐｎまでを出力するプログラム。またＦｒａｍｅ単位に変換されたオート

ラベリング結果を使用し、そのＴｏｐｎに出現した音素が正しく認識された音素か判定している。

プログラムの位置
/home/atrp24/xitoh/Island/Display/S_top20.c
(使用方法)
S_top20.c 文節番号

プログラム内のＩｎｐｕｔＦｉｌｅＡは、音声認識結果ファイル名、ＩｎｐｕｔＦｉｌｅＢは、Ｆｒａｍｅ単位のオートラベリング結果ファイル名、ＯｕｔｐｕｔＦｉｌｅＡは、出力結果ファイル名である。

A.7　結果グラフ作成プログラムの作成

認識結果からグラフ作成用のデータを作り、ｇｎｕｐｌｏｔ用のグラフ表示プログラムを作る。

プログラムの位置
/home/atrp24/xitoh/Island/Display/gnuplot/plot2.csh
(使用方法)
plot2.csh 出力ファイル名 入力結果ファイルリスト グラフ表示番号
(例)
plot2.csh Area-1 Area 8

出力ファイル名とはｇｎｕｐｌｏｔ用のグラフ表示データ名、入力結果ファイルリストとは、グラフに出力したい認識結果（投影面積、体積など）のリストをファイルにした名前、グラフ表示番号は、認識結果の何番目をグラフに出すかの番号である。

グラフ表示番号：
2　　４０文節出現回数
3　　３８文節出現回数（オートラベリングに失敗した２文節を除いた）
4　　４０文節Ｔｏｐ１０出現回数
5　　３８文節Ｔｏｐ１０出現回数
6　　３８文節での正解認識数
7　　４０文節における正解認識率
8　　３８文節における正解認識率
9　　３８文節での正解音素数
１０　４０文節における音素認識率
１１　３８文節における音素認識率

このプログラムを実行すると、入力結果ファイルリストに入っている認識結果ファイルからグラフ作成用データである”～．ｄａｔａ”が自動的に作られる。この”～．ｄａｔａ”は後に述べるＨＲ＆ＦＡＲグラフにも使用する。

A.8　ＨＲ＆ＦＡＲグラフの作成

先に述べた認識結果から作られたグラフ作成用のデータ”～．ｄａｔａ”からＦａｌｓｅＡｌａｍ（ＦＡ）、Ｈｉｔ Ｒａｔｅ（ＨＲ）、ＦａｌｓｅＡｌａｍＲａｔｅ（ＦＡＲ）を計算し、結果を標準出力に出力する（ｐｌｏｔ３）。またこの出力結果からＨＲ＆ＦＡＲグラフを作成する（ｐｌｏｔ４）。

プログラムの位置
/home/atrp24/xitoh/Island/Display/gnuplot/plot3.csh
（使用方法）
plot3.csh グラフ作成用データファイル（～．ｄａｔａ）

出力結果は４０文節におけるＦＡ、３８文節におけるＦＡ２、４０文節におけるＨＲ、３８文節におけるＨＲ２、４０文節におけるＦＡＲ、３８文節におけるＦＡＲ２が出力される。またこの結果をファイルに落して、ｐｌｏｔ４のプログラムを動かし、ＨＲ＆ＦＡＲグラフ用のデータを作成する。

プログラムの位置
/home/atrp24/xitoh/Island/Display/gnuplot/plot4.csh
（使用方法）
polt4.csh ｐｌｏｔ３の結果ファイル グラフ選択番号
グラフ選択番号：
０　４０文節におけるＨＲ＆ＦＡＲのｇｎｕｐｌｏｔ用グラフデータ
１　３８文節におけるＨＲ＆ＦＡＲのｇｎｕｐｌｏｔ用グラフデータ