項目反応理論を用いた大語彙単語認識の期待正解率の推定

重田, 武弘 / SHIGETA, Takehiro

(発行年 / Year)
2008-03-24
(学位授与年月日 / Date of Granted)
2008-03-24
(学位名 / Degree Name)
修士(理学)
(学位授与機関 / Degree Grantor)
法政大学 (Hosei University)

項目反応理論を用いた大語彙単語認識の期待正解率の推定

Estimation of Expected Accuracy for large Vocabulary Word Recognizer using Item Response Theory

06T0011 重田　武弘
Takehiro Shigeta
法政大学情報科学研究科
E-mail: takehiro.shigeta.u3@gs-cis.k.hosei.ac.jp

目次

1 章　まえがき
1.1.　音声認識の現状と問題点
1.2.　本論文の構成

2 章　期待正解率を用いた個人の認識力の推定
2.1.　個人の認識力の推定
2.2.　期待正解率を用いた推定
2.2.1　期待正解率
2.2.2　多カテゴリー認識問題と期待正解率
2.3.　混合 2 項分布を用いた期待正解率の推定
2.3.1.　混合 2 項分布を用いた誤認識率推定のモデル化
2.3.1.　Rosenberg モデルの問題点
2.4.　項目反応理論を用いた期待正解率の推定
2.4.1.　項目反応理論
2.4.2.　項目反応理論と大語彙単語認識の対応関係
2.4.3.　IRT モデルを用いた期待正解率の推定

3 章　実験
3.1.　固有名詞単語発声コーパスの構築
3.2.　期待正解率の推定実験

4 章　考察
4.1.　最尤推定による推定結果と、項目反応理論による推定結果の比較
4.2.　Rosenberg モデルの大語彙での適用性について
4.2.1.　認識結果・平均順位・標準エラーに対する、Rosenberg（小語彙）の実験との比較
4.2.2.　混合モデルに対する、Rosenberg（小語彙）の実験との比較
4.3　誤認識に関する考察
4.3.1.　順位の出ない入力サンプルの原因について
4.3.2　一人当たりの必要サンプル数の推定

5 章　結論

謝辞

Abstract

This paper addresses one of the fundamental problems encountered in performance prediction for speech recognition. In particular we address problems related to the estimation of small-size development utterances that can give good error estimates and their confidence regarding very large vocabulary proper name recognition.

A purpose of this paper is estimation of expected accuracy to measure performance of large vocabulary proper name recognition. In this paper, large vocabulary proper name recognition is considered to multi-class pattern recognition problem. By using not each accuracy but expected accuracy, personal performance prediction with small samples is examined.

Generally, a word recognition rate (a word error rate) is dependent on variable factors. However in this paper, it is only dependent on vocabulary size, considering large vocabulary proper name recognition to multi-class pattern recognition problem. First, a probabilistic model of Rosenberg was used for this experiment, and next Item Response Theory (IRT) was used to calculate effectively a factor that is vaguely calculated by Rosenberg model.

Concluding, the IRT estimation method was effective comparing to maximum likelihood estimation.

1 章　まえがき

1.1　音声認識の現状と問題点

10 万以上の語彙を持つ、大語彙単語認識器はポータブルメディアプレイヤーの楽曲・動画検索、カーナビゲーションシステムの目的地検索、ディレクトリアシスタンス、その他様々なアプリケーションなど多くの選択リストを持つ音声インターフェースで要求されている。例えば Zenrin のカーナビゲーションシステムは 100～1000 万語彙を扱っているし、iTunes ミュージックストアでも数万語の曲を取り扱っている。そのほかの音声認識の実用システムへの応用は[1][2][3]などに詳しい。そういったアプリケーションはその選択肢の多さから、ボタンなどを用いて検索することが難しく、音声などを使って検索できると望ましいと思われる。しかし、現状では音声による認識はまだその認識率が十分でないことから、実用に当たってはなんらかの手段を講じなければならない。

例えば選択肢に重みをつけて、検索し易い選択肢の重みを重くする等の手段が考えられるが、この方法だと普段検索されない単語は認識されにくくなることになり、有効ではない。選択肢はあくまで同等にアクセスされることが望ましく、言語モデルのようなものも使うことが出来ない。そのような事前確率が、それの使用に適切でないことから、多くのパープレキシティを持つことになる。

最終手段として、システムの語彙サイズを減らすことは、一定以上の性能を要求する音声認識器の構造の最後の手段としては避けられない。アプリケーションの開発には、最小値、あるいは中間値など、適切な方法によって推定されたシステム辞書サイズを決定した方が良いだろうと考えられる。

語彙サイズの調整とは、音声認識システムにおいて誰もが一定以上の性能（認識率）を出すために、語彙サイズを各話者毎に調節することである。例えば下の図 1 を見てもらいたい。図 1 は語彙サイズと単語誤り確率（以下 WER）の関係をグラフ化した例である。図では 2 人の話者について描いてあるが、一般に、認識率は話者によって異なるのでグラフでも 2 つの曲線が描かれている。

図 1 からは、例えば語彙サイズが 50000 語のとき、A さんは誤認識率 20%、B さんは誤認識率 10%という結果が読み取れる。ここで今このシステムが 50000 語語彙で稼動しているとき、これを使う上で「誰もが誤認識率 10%の性能を出せる」ようにしたいと考えたとき、語彙サイズの調節を行えばそれが可能になる。この例では、B さんは語彙サイズの調整は必要ないが、A さんは語彙サイズを 50000 語から 40000 語に調整すれば良い。

このように各話者によって語彙サイズの調整が可能になれば大語彙システムの実用化に大きな貢献ができる。また、実用上の視点からは、その語彙サイズの決定はできるだけ少ないサンプルで決められると尚良い。今実験では語彙サイズの調整のための一人当たりのサンプル数を 100 語程度で行うことを目指した。しかしそのための問題は多々ある。このような性能を表す曲線は本当に描くことができるのか、それが出来たとしてどのようにその曲線を決めるのか。本論文ではそれらの問題を検討していく。

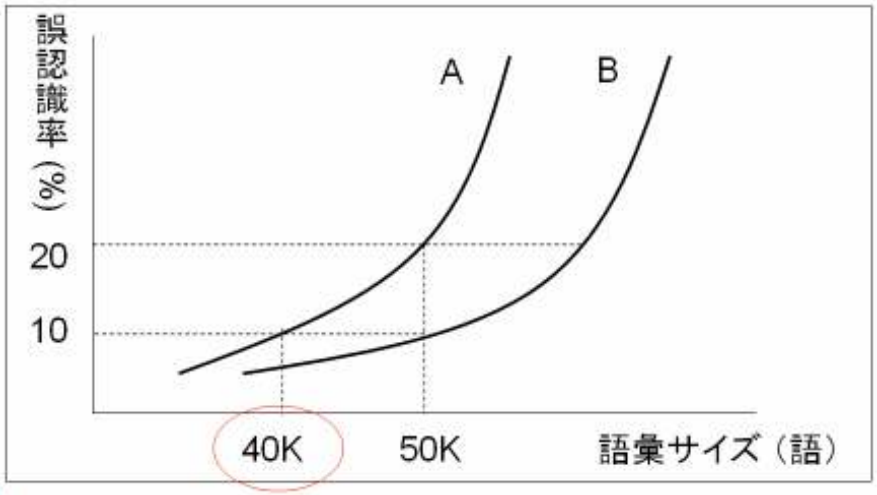

図 1　語彙サイズの調整

語彙サイズの決定には様々な要因が関わってくる。本論文では直感的に「システムの辞書サイズが増えれば認識率が低下する」という性質から語彙サイズと認識率の関係に着目した。大語彙単語認識のような多くの数を扱う認識問題では一般に個々の認識率の平均を取った「期待正解率」が用いられる。この期待正解率と多カテゴリー間の性質について議論したのが[4]である。ここでは先ず大語彙単語認識を多カテゴリー認識問題とみなし、語彙サイズと認識率に関する認識器の性能を考えている。

期待正解率の推定について、Rosenberg は 1983 年に誤認識率が辞書サイズの関数になるという確率モデルを提案している[2]。このモデルは、単語乱雑性の計算の為に全話者にシステム辞書の単語を全て発声させるものであり、大語彙システムに対しては適さないものであった。しかし大語彙に関してもその妥当性が示せれば、期待正解率の推定に有効な指標になると考えて、このモデルを利用した推定を行った。総合して、本実験の目的は、期待正解率を利用して、個人ごとに一定以上の性能が出せる語彙サイズの調整を行うことである。

1.2 本論文の構成

本論文の構成は、第 2 章で、期待正解率とそれを用いた個人の認識率の推定について述べる。ここで先行研究やその推定方法も説明する。第 3 章で実験の詳細なデータと期待正解率推定実験の結果を述べる。そして第 4 章で結果から考えられる考察を述べ、5 章で結論と今後の課題をあげる。

2章　期待正解率を用いた個人の認識力の推定

2.1　個人の認識力の推定

本実験の目的である、個人ごとの認識力の推定について説明する。実用上のシステムで使うイメージとしては例えば、50000 語の語彙のシステムで、100 語のサンプルを収録した場合、50000 語彙での認識率が出る。このように、まず使用においては一定の学習を必要とする。次に、この 50000 語彙、100 語サンプルでの認識率から、その話者における認識率の決定要因を抽出（本実験では期待正解率）、それを利用して任意の語彙サイズにおける認識率を推定できるようにするのが最終目標となる。

ここで注意すべき点は求めるべき期待正解率は、真の期待正解率とは限らないことである。例えば上記の例で 60000 語彙における認識率を推定しようとしても、実際に 60000 語で認識させたらどの語を正解してどの語を誤答するかは確率的にしかわからない。話者が語彙内の全単語を発話すれば、その認識率で近似することもできそうだが、大語彙において全単語の発話は難しい。今回の実験はそれらを考慮した小サンプルでの期待正解率・認識率の推定である。

2.2　期待正解率を用いた推定

2.2.1　期待正解率

期待正解率とは、個々の単語の正解率の平均を取ったものである。一般に特定話者の単語認識率といえばその話者が発声・認識した単語の（正答数/発声単語数×100）という式で表される（正答を 1、誤等を 0 とした場合）。これは、個々の単語の正解数を足したものを使用するので、この、個々の単語が正解する確率を全語彙にわたって平均したもの、すなわち期待正解率を認識率として考えることは不自然ではないと考えられる。

この期待正解率を理論的に扱った問題として、多カテゴリー認識問題というものがある。次節ではこの多カテゴリー認識問題と、それの単語認識への適用について説明する。

2.2.2　多カテゴリー認識問題と期待正解率

多カテゴリー認識問題とはパターン認識における問題の一種である。これは、例えば漢字カテゴリー全体のようなカテゴリー母集合を考え、その中の M 個のカテゴリーを対象とし、その正解率や一般的性質などについて問う問題である。この問題ではカテゴリー母集合の中から M の部分集合を選び出す組み合わせの数だけ考えられるが、その正解率を計算することは、組み合わせ的爆発により容易ではない。[5] では、個々の認識問題の正解率を直接求めるのではなく、統計的な取り扱いにより多カテゴリー認識問題一般に成り立つカテゴリー数と正解率の期待値（期待正解率）との関数関係を求めた。また、導出した式を用いて期待正解率を求めるための計算量と定義どおりに期待正解率を求める場合の計算量との比を評価し、カテゴリー母集合のカテゴリー数の増加と共にこれが指数関数のオーダで減少することを示している。

先ず、考察する問題を以下のように設定する。

多数のカテゴリーからなるカテゴリー集合 $\mathbb{C}$ と適当な分類方法 $\mathbb{S}$ が与えられている。このとき、任意の M について、$\mathbb{C}$ から大きさ M のカテゴリー部分集合を任意に取り出し、それを認識対象とする M カテゴリー認識問題を考える。この M カテゴリー認識問題の正解率の期待値 $EPc(M \parallel \mathbb{C}, \mathbb{S})$ を求めよ。

$\mathbb{C}$ をカテゴリー母集合と呼び、$EPc(M \parallel \mathbb{C}, \mathbb{S})$ を期待正解率と呼ぶ。期待正解率は全ての M カテゴリー認識問題の正解率の平均値であるが、その値を直接求めることは、カテゴリー数が大きくなると組み合わせ爆発により難しくなる。

・カテゴリーと分類方法

カテゴリー母集合を$\mathbb{C} = \{C_1, C_2, \cdots, C_{|\mathbb{C}|}\}$とし、カテゴリーCi に属するパターンの集合をΩ(Ci)と表す。$i \neq j$のとき、Ω(Ci)∩Ω(Cj)＝ϕとし、全パターンの集合$\bigcup_{i=1}^{|\mathbb{C}|}\Omega(C_i)$を$\Omega(\mathbb{C})$と表す。

ここでいうカテゴリー母集合の各カテゴリーに属するパターン数は有限とするが、カテゴリー数に関しては有限でも無限でも構わない。また、分類方法に関しては、入力パターンの観測値（特徴量ベクトル）を各カテゴリーの代表値（辞書ベクトル）と比較することによって認識を行う一般的パターン分類器を用いる。

パターンから抽出する特徴を定め、カテゴリー母集合 C の各カテゴリーCi に対し辞書ベクトルを用意する。全パターンの集合$\Omega(\mathbb{C})$の各パターンωと各カテゴリーCi に対し、ωから抽出される特徴量ベクトルと Ci の辞書ベクトルからパターン分類器が評価値を計算する。この評価値を得点と呼び S(ω,Ci)と表す。カテゴリー母集合から任意に取り出された M 個のカテゴリーのいずれかに属するパターンωが入力されるが、パターン分類器はωの観測値から M 個のカテゴリーのそれぞれが得る得点を計算し、最大の得点を得るカテゴリーをωが属するカテゴリーと分類し出力する。この分類法を$\mathbb{S}$で表す。

この分類法は一般的なものであり、得点は例えば距離法なら距離の逆数、類似度法では類似度、ベイズ決定法では尤度で与えればよい。次節に出てくる Rosenberg の確率モデルでは距離を用い、本論文における実験では尤度を用いている。

・全Mカテゴリー部分集合に対する期待正解率の計算手法

カテゴリー母集合$\mathbb{C}$から取り出された M 個のカテゴリーからなるカテゴリー部分集合のそれぞれが M カテゴリー部分集合を構成するが、このようなカテゴリー部分集合は$|\mathbb{C}|$（$\mathbb{C}$内の要素数）から M を取り出す組み合わせの数だけ存在する。M 個のカテゴリーからなるカテゴリー部分集合をC^Mで表し、すべての M カテゴリー部分集合の集合を$\mathbb{C}^M = \{C_1^M, C_2^M, \cdots, C_{|\mathbb{C}^M|}^M\}$で表す。各 M カテゴリー部分集合$C_j^M$が一つの M カテゴリー認識問題に対応しているので、今後C_j^Mを対応する M カテゴリー認識問題と同一視して扱う。期待正解率$EPc(M \parallel \mathbb{C}, \mathbb{S})$は$C_j^M$における正解率$Pc(C_j^M \parallel \mathbb{C}, \mathbb{S})$に$\mathbb{C}^M$における$C_j^M$の生起確率$P(C_j^M \mid \mathbb{C}^M)$を乗じて総和を取ったものとなる。すなわち、次式のようになる。

$$EPc(M \parallel \mathbb{C}, \mathbb{S}) \equiv \sum_{j=1}^{|\mathbb{C}^M|} P(C_j^M \mid \mathbb{C}^M) \cdot Pc(C_j^M \parallel \mathbb{C}, \mathbb{S})$$

この期待正解率に関しては以下のような性質が見られることがわかっている。

・減少関数である
期待正解率 EPc(M)はカテゴリー数 M の減少関数である。
・下に凸な関数である
期待正解率 EPc(M)は M に関して下に凸な関数である。
これらの性質は期待正解率の性質の良さを示している。

本論文においては、大語彙における単語認識をこの多カテゴリー認識問題とみなしている。各単語が誤認識する確率（正解単語よりスコアが高くなる確率）を pi とする。この pi は、単語の長さや音素構造などにより、本来単語ごとに異なった値を取る。それを本論文では M 種類の値であると設定、即ち M カテゴリーの問題に帰結させる。

本来、音声認識システムにおいて認識率は様々な要因によって変動する。話者による影響（発声方法、声道の長さ等）、認識用辞書による影響（語彙数、語彙単語の音素構造等）、入力音声の影響（入力サンプルの数、単語長、音素構造等）、収録環境による影響（収録する場所、マイクの違い等）などが例とし

て挙げられる。このような事情から、現在、音声認識においては、どの要因が決定的になるかなどが一概に決められないという問題があり、定量的な議論は難しい。これらの要因のどれが決定的かは一概に決められないが、全てではなく一部の要因を改善して、認識率を上げるという方法は現在もよく行われている[6]。

そのような要因はいくつかに分類できる。発声方法やマイクの違いなどの音響的な変動はそれこそ無数にあるのでこれを基に認識率を向上させようというのは現実的ではない。入力サンプル数の違いなどはより詳細な統計的検討が必要になるので現時点での扱いは難しい。これらの事実から、現時点で認識率決定・推定のための要因としては、語彙サイズや単語の音素構造辺りを用いることが考えられる。

今回は主に語彙サイズによる影響に着目した。直感的に、システムの辞書サイズが増えるにつれて減少する。そこから、一般に辞書サイズと認識性能には関係があることがわかる。今実験では単語認識に多カテゴリー認識問題の考え方を適用して、語彙サイズと認識率の問題を考えている。但しこれは認識率が語彙サイズ「のみ」に依存すると言っているのではなく、他の要因の影響が消えたわけではない。考え方として語彙サイズと関係があるといっているだけである。

しかし、辞書の規模や認識性能をどのような尺度で測ればよいかは明らかではない。この問題に対して、語彙サイズと認識率の関係を確率モデルで表したのが Rosenberg である。Rosenberg は上記の多カテゴリー認識問題において、単語認識を多カテゴリー認識問題とみなして、語彙サイズと正解率の関係を示す実験を行った。次に、この Rosenberg の確率モデルについて説明する。Rosenberg は、孤立単語認識において、平均順位の分布が混合 2 項分布に従うなら、エラー率が辞書サイズの関数になると主張した。

2.3 混合 2 項分布を用いた期待正解率の推定

Rosenberg は、単語認識において認識率と語彙サイズの関係を表す確率モデルを提唱した。そのモデルは孤立単語認識において、平均順位の分布が混合 2 項分布に従うなら、エラー率が辞書サイズの関数になるというもので、認識率の特定因子がはっきりしていない単語認識において性能評価の指標になると主張した。本節ではその確率モデルについて詳述する。

2.3.1 混合 2 分布を用いた誤認識率推定のモデル化

Rosenberg は、正解単語以外の単語のスコアが高くなる確率を定め、単語音声認識をシステム辞書に登録された単語数分のベルヌーイ試行とみなした。

実験では、認識器の性能評価に、2 種類の指標を定義する。個々のサンプルの認識において、入力サンプルとプロトタイプの距離 dij が入力サンプルと同じ単語のプロトタイプとの距離 dii よりも小さくなったとき（それを順位と定義し、その回数が 1 以上のときエラーが起こったとする）、それが起こる確率を pi、pi の確率でそのイベントが起きる回数の合計を si とする。するとこの認識は起こるか起こらないかの 2 項分布なので、今、全語彙数が N のとき si が定数 k になる確率は

$$\Pr ob\{s_I = k\} = \binom{N-1}{k} p_I^k (1-p_I)^{N-1-k} \quad (1)$$

になる。また、このときエラーの 1 つ目の指標として

$$e_I = 1 - \frac{1}{1+s_I} \quad (2)$$

を定め、その期待値を

$$E\{e_I\} = \sum_{k=0}^{N-1} P\{s_I = k\}\left(1 - \frac{1}{1+k}\right)$$

$$= 1 - \frac{1-(1-p_I)^N}{p_I N} \quad (3)$$

とする。このエラーは、Smith と Erman[7]によって紹介された認識器を性質付ける「効率性」に関係していることから「効率性エラー」と呼ぶ。「効率性」という言葉は、[7]において、システムが仮説単語生成機として使われるとき、この指標が全仮説単語と正解仮説単語の比率を表していることからきている。

また、2 つ目の指標として

$$E_I = \begin{cases} 1 \; if \; s_I > 0 \\ 0 \; if \; s_I = 0 \end{cases} \qquad (4)$$

を「標準認識エラー」と定め、その期待値を

$$E\{E_I\} = 1 - (1 - p_I)^{N-1} \qquad (5)$$

とする。これは s_I が 1 度でも起こったとき、即ち誤認識が起こったときに一律 1 と定めるもので、全体のエラーはこれを数え上げることで単純に計算できるという利点を持つ。しばしば誤認識率は「正解単語が認識器の出力する上位何位の中に含まれるか」を計算するために、(4)式において、si が 0 でなく c 以上なら 1、というように一般化して考えることも可能だが、本論文では 1 位のみを対象とする。

今、認識を個々のサンプルから語彙（辞書サイズ）全体に拡張することを考える。より一般的な仮説として、イベントの起こるベルヌーイ確率 pi はランダム変数である。辞書内の各単語よって異なるし、同じ単語でも試行毎に異なった値になり、複雑になる。そこで今 pi が M 種類の値しか取らないと仮定する。すると、確率変数 pv が、値 pm になる確率は

$$\Pr ob\{p_v = p_m\} = h_m, \quad m = 1,2,\cdots,M \qquad (6)$$

$$\overline{p}_v = \sum_{m=1}^{M} h_m p_m \qquad (7)$$

のように定義できる。(7)式は上述した期待正解率にあたるものであり、hm は pv が pm になる確率を表している。また、(1)式と(5)式から語彙全体に対して si が k になる確率は

$$\Pr ob\{s_v = k\} = \sum_{m=1}^{M} h_m \binom{N-1}{k} p_m^k (1 - p_m)^{N-1-k} \qquad (8)$$

このようにして、2 項分布の混合に単純化される。

また、その期待値と分散は

$$E\{s_v\} = (N-1)\overline{p}_v \qquad (9)$$

$$Var\{s_v\} = (N-1)\sum_{m=1}^{M} h_m p_m (1 - p_m) + (N-1)^2 \sum_{m=1}^{M} h_m (p_m - \overline{p}_v)^2 \qquad (10)$$

$$\overline{p}_v = \sum_{m=1}^{M} h_m p_m \qquad (11)$$

のように表される。

（9）式は一般に、期待値と等しい。

2 種のエラーについても、同様に拡張して、

$$E\{e_v\} = \sum_{m=1}^{M} h_m \left(1 - \frac{1-(1-p_m)^N}{p_m N}\right)$$

$$= 1 - \sum_{m=1}^{M} h_m \left(\frac{1-(1-p_m)^N}{p_m N}\right) \qquad (12)$$

$$E\{E_V\} = 1 - \sum_{m=1}^{M} h_m (1-p_m)^{N-1} \qquad (13)$$

のように定義する。(12)、(13)式の $\overline{p}_v$ の推定値の計算も(11)式と同様である。いずれも「平均順位から」「効率性エラーから」「標準エラーから」推定される期待正解率となっている。

この「pi が M 種類の確率しか取らない」ということは即ち、辞書内の全単語において、誤る確率が M 種類しかないと仮定することである。本来ならば、辞書内に含まれる単語はその長さや音素構造など、多様に異なっている。今回は先行研究やコスト等の問題から、後述するように M＝３程度で実験しているが、その妥当性は考慮していない。この問題は後に再掲する。

(11)、(12)、(13)式における $\overline{p}_v$ はイベントの起こる確率 pv の推定値（期待正解率）を表しており、即ち、この $\overline{p}_v$ を上手く推定できれば、精度の高い性能推定が行えることになる。この $\overline{p}_v$ を個人に対して推定することが、本論文における期待正解率の推定を意味する。

上述したとおり、Rosenberg のこのモデルでは期待正解率 pv の推定の元になる、個々の単語誤り確率 pi が、3 程度と、実際の収録単語数と比べて少ない値（代表値）になっている。これに対する問題点を次節で説明する。

2.3.2 Rosenberg モデルの問題点

Rosenberg の実験は被験者 6 人、最大辞書サイズ 1109 単語までしか扱っておらず、辞書サイズが大きくなったときにはどうなるかは定かではない。本実験では語彙サイズを最大 100000 語まで、被験者も 50 人まで扱っているため、大語彙に対してそのままこのモデルが適用できるかは定かではない。

Rosenberg のモデル適応には pi が全てわかっていることが前提になっている．大語彙においては語彙の多さから，小サンプルで性能推定しようとした場合，発声する単語数が少ないので，全単語分の pi を算出することは難しい．Rosenberg モデルではこれを回帰と最尤推定において推定するが，この方法では推定できる pi の個数に限界があり，いくつかの代表値を用いた混合モデル等に頼らざるを得なくなる．

また、Rosenberg は実験の指標として、単語間距離を用いている。これについては、語彙数が多くなると計算コストがかかりすぎることから、本実験では尤度を指標としている。

他、Rosenberg の実験では辞書サイズの単語 1109 語全てを各被験者に発声させているのに対して、本実験では一人 100 語程度しか発声させていない。Rosenberg モデルは全話者が全単語を発話させて認識させる必要があった。しかし本実験のような大語彙には全単語の発話は困難であり、一人あたりが苦痛なく取れるサンプル数として約 100 語とした。

Rosenberg の実験は小語彙（本実験と比較して）、多サンプルでの実験であり、本実験は大語彙、小サンプルでの実験となっている。いずれにせよ、本実験に Rosenberg のモデルをそのまま適応できるかどうかは検討する必要がある。このような上記の問題に対処するため pi の推定に項目反応理論（IRT）を用いた実験を行った。次節では、この項目反応理論について説明する。

2.4 項目反応理論を用いた期待正解率の推定

2.4.1 .項目反応理論

項目反応理論（Item Response Theory 略称 IRT）とは、評価項目群への応答に基づいて、被験者の特性（認識能力、物理的能力、技術、態度、人格特徴等）や、評価項目の難易度・識別力を測定するための試

験理論である。この理論の特徴は、個人の能力値、項目の難易度といったパラメータを、評価項目への正誤のような離散的な結果から確率的に求めようとする点である。

項目反応理論では、能力値や難易度のパラメータを推定し、データがモデルにどれくらい適合しているかを確かめ、評価項目の適切さを吟味することができる。従って、試験を開発・洗練させ、試験項目のストックを保守し、複数の試験の難易度を同等とみなす（例えば異なる時期に行われた試験の結果を比較する）ために、この理論は有効である。

より古典的なテスト理論（素点方式、偏差値方式）と比べると、この理論は、試験者が評価項目の信頼性の改善に役に立つ情報を提供し得る、標本（受験者）依存性・テスト依存性にとらわれずに不変的に受験者の能力値とテスト項目の難易度を求められる、という利点がある。

実際にこの理論は基本情報処理試験や、TOEIC，コンピュータテストなど、テストの作成に多く使われている[11]。

・IRT モデル

一般的なモデルでは、項目への離散的な応答（正誤など）の確率が、１つの人パラメータと１つ以上の項目パラメータによる関数であるという数学的な仮説に基づいている。用いられる変数は以下の通りである。

θ ：人パラメータ
　各受験者の特性の大きさを表す実数値。

a_i ：識別パラメータ
　項目 i が被験者の能力を識別する力を現す実数値

b_i ：難易度（困難度）パラメータ
　項目 i の難しさを現す数値。一般的には各項目に 50%の正答率を持つ被験者の能力値を基準として決められている

c_i ：当て推量パラメータ
　多肢選択形式のテストの場合に、項目 i に被験者が偶然に正答できる確率を表す実数値。

基本的な考え方としては、人パラメータと、項目の難易度パラメータの差を取り、ロジスティック曲線に当てはめて、正当する確率を求めるというものである。例えば能力試験において、ある項目が被験者にとって非常に簡単であった場合、その正答率は限りなく 1 に近づき、逆にある項目が被験者にとって非常に難しいものであった場合、その正答率は限りなく 0（パラメータ c を用いる場合にはc_i）に近づく。

もっとも簡単な 1 パラメータロジスティック（1PL）モデルでは、変数にθとb_iのみを用いる。しかし適用のための条件は厳しくなっている。このモデルでは、項目 i に正答する確率は次式のようになる。

$$p_i(\theta) = \frac{1}{1 + e^{-(\theta - b_i)}} \qquad (14)$$

2 パラメータロジスティック（2PL）モデルでは、更にa_iを用い、各項目が評価にとってどの程度適正な判断基準であるかを変数に組み込む。このモデルでは、項目 i に正答する確率は次の式で与えられる。

$$p_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}} \qquad (15)$$

ここで、定数 D は 1.701 という値で、ロジスティック関数を累積正規分布関数に近似するためのもので、確率が関数の定義域（一般に-3～3）内で 0.01 以上異ならないようになっている。なお、IRT モデルは当初は普通の累積正規分布関数が用いられたが、このように近似されたロジスティックモデルを使うことで、大きく計算を単純化することができた。

3 パラメータロジスティック（3PL）モデルでは多肢選択形式の場合において、適当に選択肢を選択しても偶然正答する確率c_iを考慮して、項目 i に正答する確率は次の式で与えられる。

$$p_i(\theta) = c_i + \frac{(1-c_i)}{1+e^{-Da_i(\theta-b_i)}} \qquad (16)$$

人パラメータは被験者の評価の対象となっている 1 次元的な特性の大きさを表す。この特性は因子分析の 1 つの因子に類似している。また、個々の項目や人は相互に独立であり、集合的に直交であると仮定されている。すなわち、ある項目の正誤は他の項目の正誤に影響せず、ある人の正誤は他の人の正誤に影響しないという仮定を置いている。

項目パラメータは、ある項目の性質を示す。項目パラメータが定まると、受験者がその項目に正答する確率 pi は各受験者の能力 θ の 1 変数のみを持つ関数になり、縦軸に正答率、横軸に能力値としたグラフが書ける。このグラフを項目特性曲線と呼ぶ。パラメータ b は項目の難しさであり、この値は人パラメータと同じスケール上にある。パラメータ a は項目特性曲線の傾きを決定し、その項目が個人の特性の水準を識別する程度を示す。曲線の傾きが大きいほど、項目の難しさと人の特性の大きさに差があるときに回答の正誤がくっきりわかれることを示す。最後のパラメータ c は、項目特性曲線の負の側の漸近線である。すなわち、これは非常に低い能力を持つ人がこの項目に偶然正答する確率を示す。

項目パラメータについてグラフを用いて説明する。

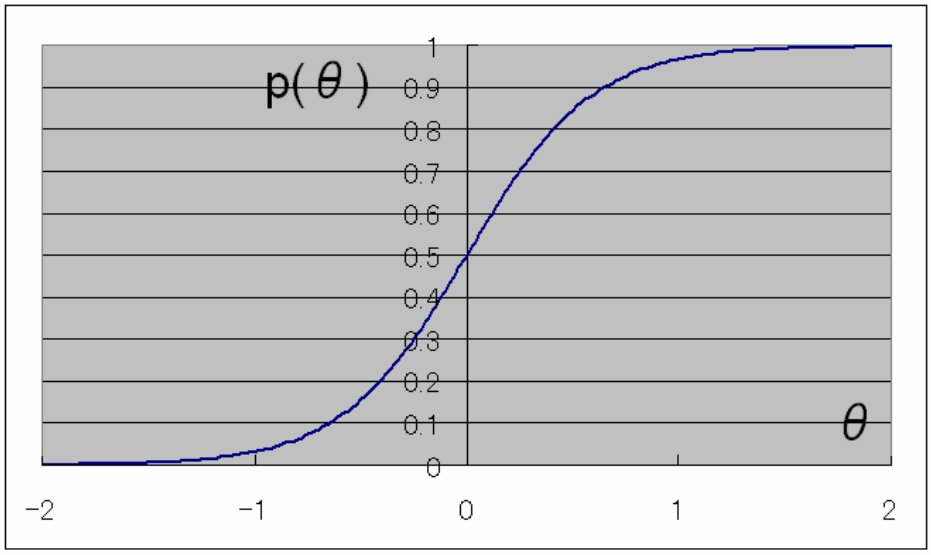

図 2　a=2、b=0、c=0 の IRT モデル

図 2 は a=2、b=0、c=0 の項目特性曲線である。横軸は能力値、縦軸は正答率であるグラフから、能力値が高い人ほど、正答する確率が高いことを示している。

先ず困難度パラメータについて。

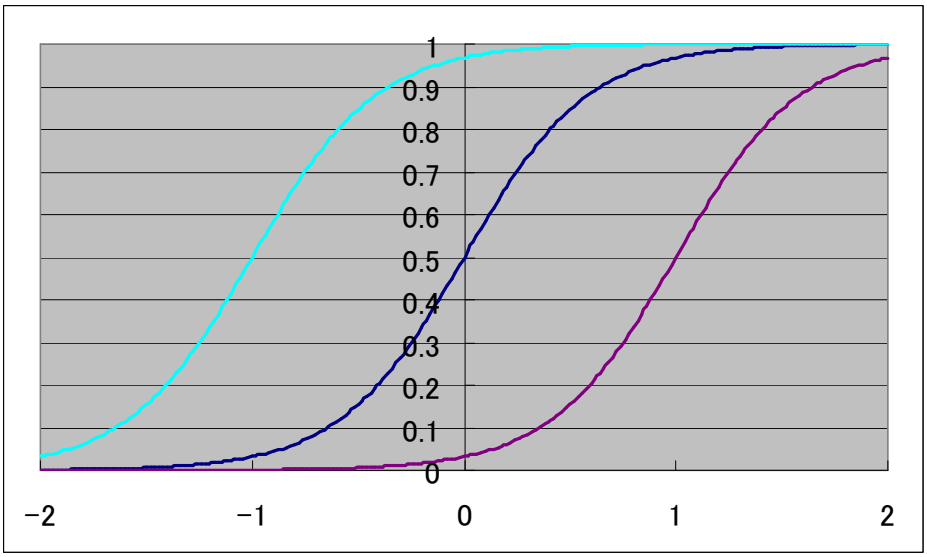

図 3　項目特性曲線（左から b=-1、0、1）

図 3 は左から b=-1、b=0、b=1 の項目特性曲線である。グラフから、困難度の高い単語ほど右よりになっていることがわかる。困難度の高い単語は能力値の高い人しか正答できず、困難度の低い単語は、比較的誰でも正答できることを示している。

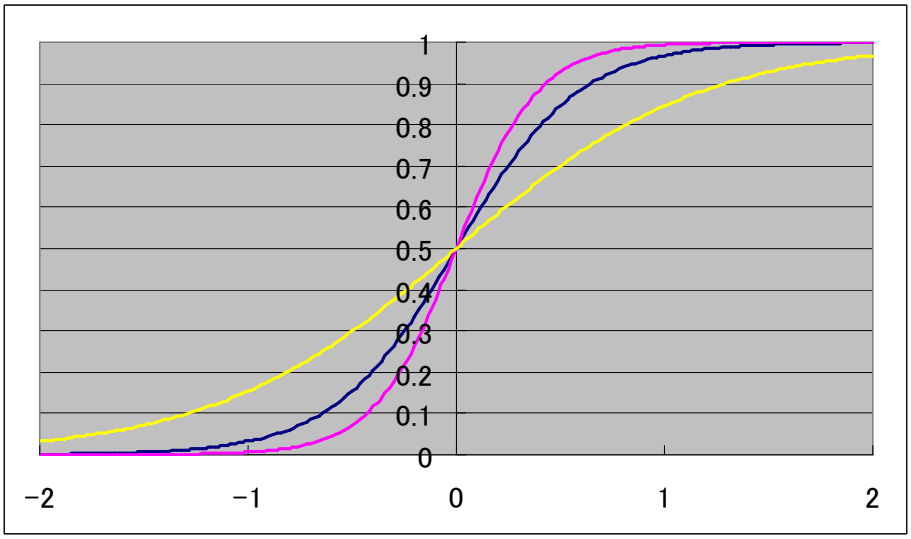

図 4　項目特性曲線（上から a=3、2、1）

図 4 は上から a=3、a=2、a=1 の項目特性曲線である。グラフから、識別性の高い単語ほど急峻になっていることがわかる。すなわち識別性の高い単語ほど、話者の能力値が正答できるか否かを鋭敏に反映していることを示している。

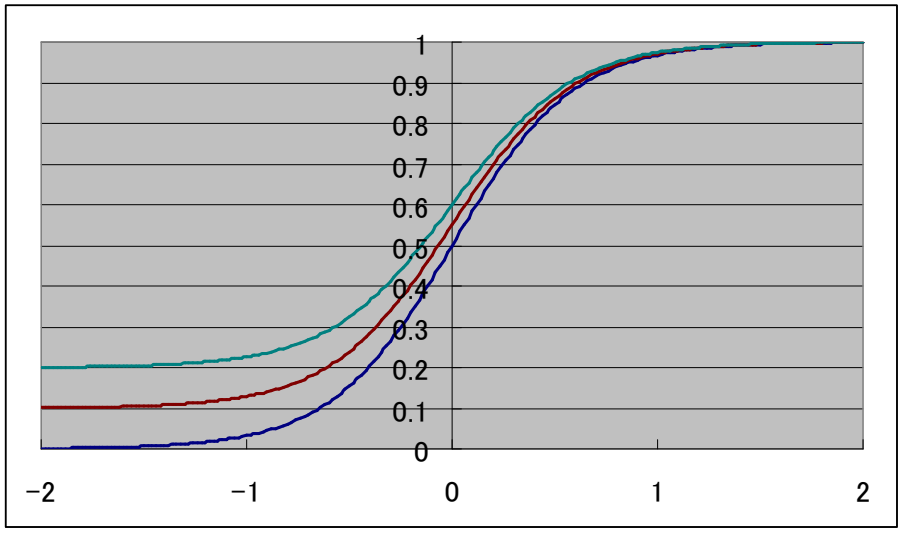

図 5　項目特性曲線（上から c=0.2、0.1、0）

図 5 は上から c=0.2、c=0.1、c=0 の項目特性曲線である。グラフから、当て推量パラメータの高い単語ほど上に寄っていることがわかる。当て推量パラメータの高い単語は能力値の低い人でもある程度正答できることが見て取れる。

各項目は互いに独立である。という前提を置いているので、項目特性曲線は加法的である。よって、全ての項目特性曲線を足したものが求められる。これはテスト特性曲線と呼ばれる。

$$T(\theta) = \sum_{i=1}^{N} p_i(\theta) \qquad (17)$$

試験のスコアはこのテスト特性曲線によって求められる。テスト特性曲線は θ の関数であり、$T(\theta)$ の値を受験者のスコアとする。よって、IRT によるスコアは従来の方法によるスコアと比べ、計算・解釈において非常に異なっている。しかし、殆どのテストにおいて、値 θ と従来のスコアとの（線形）相関関係は非常に高い（0.95 以上になることが多い）。したがって、従来のスコアに比べ、IRT のスコアのグラフは累積度数分布曲線の形に近くなる。

ここまで示したモデルでは、1 次元的な特性と、項目に対する正解・不正解のような 2 値のいずれかの応答を前提としていた。しかし、多値ラッシュモデルのように多値（例えば 0 : 全く誤り　1 : 殆ど誤り　2 : 概ね正しい　3 : 完全に正しい、の 4 値）をとるように拡張されたモデルや、多次元的な特性を仮定したモデルも存在する。

IRT の主な知見の 1 つは信頼性の概念を拡張したことである。伝統的に、信頼性とは測定の精度を示すものであり、真のスコアと観察されたスコアの誤差の比率など、様々な方法で定義される単一の指標で現される。古典的なテスト理論では、クロンバックの α 係数などがテスト全体としての信頼性の指標を表すものとして知られている。しかし IRT によると、評価の精度はテストの成績の全範囲にわたって均一ではないことが明らかになる。一般的に、試験点数の範囲の端のスコアは、中央に近いスコアより多くの誤差を含んでいる。

IRT では、項目・テストのそれぞれについて、信頼性の概念を置き換える情報関数という概念が用いられる。例えばフィッシャーの情報理論に従って、ラッシュモデルの場合には、項目情報関数は単純に正しい応答の確率と不正確な応答の確率の積で与えられる。すなわち、不正確な応答の確率を $q_i(\theta) = 1 - p_i(\theta)$ で表すと、以下の式で与えられる。

$I(\theta) = p_i(\theta) q_i(\theta)$

推定の標準誤差はテスト情報の逆数である。すなわち以下の式で表される。

$$\mathrm{SE}(\theta) = 1/\sqrt{I(\theta)}$$

従って、情報量が多いほど、測定の間違いがより少ない（被験者の能力の推定がより正確である ）ことを意味する。

2PL, 3PL モデルでもほぼ同様であるが、他のパラメータも考慮に入る。2PL, 3PL モデルのための項目情報関数はそれぞれ以下の式で表される。

$$I(\theta) = a_i^2 p_i(\theta) q_i(\theta)$$

$$I(\theta) = a_i^2 \frac{q_i(\theta)}{p_i(\theta)} \frac{(p_i(\theta) - c_i)^2}{(1 - c_i)^2}$$

各項目は互いに独立であるという前提を置いているので、項目情報関数は加法的である。テスト情報関数は単純にその試験における各項目の項目情報関数の和で求められる。テスト情報関数は、古典的なテスト理論における信頼性の概念を置き換えるものになる。

この性質を用いて、テスト項目の適切性に理論的根拠を与えることや、ある目的に特化したテストを作ることが可能になる。例えば、ある合格基準点を超えるか超えないかのみで合格・不合格が結果として与えられる（実際の合格点は重要でない）テストを作るのに有効なのは、合格基準点の近くで大きい情報が得られる項目だけを集めてテストを作ることである。また、コンピュータ適応型テストのように、ある時点での回答状況に応じて受験者の能力値を推定し、次にその受験者の能力値周辺で大きな情報が得られる問題を出題するということも可能になる。

・等化

等化とは、異なったテストの結果、異なった受験者に対してのテストの結果を、項目パラメータや被験者能力値に関係なく、共通の原点と単位をもつ尺度に変換することである。等化には、水平的等化、垂直的等化の 2 種類がある。

- 水平的等化(horizontal equiting)

同一の能力水準に対して複数のテストの難易度間に共通の尺度を設定すること

- 垂直的等化(vertical equiting)

異なった難易度のテスト間に異なった尺度を設定すること

古典的なテスト理論においては、テスト依存性や受験者依存性がつきまとうので等化を実現することは困難であった。しかし IRT による項目パラメータは不変的であり、理論的には等化の必要はない。しかし、実際には一定の定数によって、２つのテストの得点を同一尺度上に変換することがよく行われる。この手続きは以下の式で行われる。

$$\theta' = \alpha\ \theta\ +\ \beta$$

θ'は等化された能力値で、α，βは等化定数と呼ばれている。またこのとき、項目パラメータは以下のように調節される。

$$a_i' = \frac{a_i}{\alpha}$$

$$b_i' = \alpha\, b_i + \beta$$

等化定数α，βの推定には、共通の受験者または共通の項目が必要となる。そして、等化のための基準には回帰係数、平均値と標準偏差、項目特性曲線の特徴等が用いられる。

2.4.2　項目反応理論と大語彙単語認識の対応関係

項目反応理論と単語認識における対応関係について説明する。先ず、項目反応理論に必要なパラメータを、孤立短誤認識のパラメータに当てはめると以下のようになる。

θ：個人特性値　→　話者がその単語を正答する能力の高さを示す数値

a_i：項目の識別性　→　各単語が各話者をどれくらい鋭敏に識別するかを示す数値

b_i：項目の難しさ　→　単語の認識の困難さを示す数値

c_i：項目の当て推量特性　→　単語認識の当て推量（能力に関係なく当たる）の高さを示す数値

今回は 2PL モデルを使用して、各パラメータを推量した。
今実験における、各パラメータについて詳細に説明する。
先ず、個人特性値とは、単語認識における、話者の特性値を表す。単語認識において、これはその話者が単語を認識できる能力の高さを表している。すなわち、この値が高ければ高いほど単語の正答率が上がるということを示している。
次に、パラメータ a、各単語の識別性とは、各話者間の認識における識別のし易さを表す。すなわち、この値が高ければ高いほど、その単語は各話者の能力を鋭敏に区別することができる。
そして、パラメータ b、各単語の認識の困難さとは、そのまま各単語の認識のし難さを表す。すなわち、この値が高ければ高いほど、その単語は、話者特性値（能力値）の高い人しか正答できなくなり、逆に困難さの低い単語は、能力値の低い人でも正答できることを示す。

2.4.3 IRT モデルを用いた期待正解率の推定

以上ではθ、a_i、b_i、c_iの各パラメータが存在するものとして考えてきたが、それぞれの真の値は一般的に未知である。よって、離散的な回答から、それぞれの値を推定することも、この理論における重要な問題である。その推定法としては、最尤推定法、ベイズ推定法が知られている。
推定には主に以下の 3 種がある。
① 被験者の能力値が既知の母集団を用いて、テスト項目パラメータを推定
② テスト項目パラメータが既知のテストを用いて個々の被験者の能力値を推定
③ 能力値もテスト項目パラメータも未知のときに両方を推定（基本的に不可能 : 被験者全体の能力値の分布を仮定する）
今回はこの中の②を用いる。
これを利用した pi の推定は 4.1 の IRT モデルを用いることになるが、期待正解率の計算には、個々の単語の正解率 $p_i(\theta)$ を用いて、以下のように求める

$$Ep = \sum_{i=1}^{M} p_i(\theta) \cdot h_i$$

ここで M は単語数，hi は困難度ごとの補正係数を表している．
このモデルを使用して，期待正解率を推定する．基本的な流れは、共通語の単語の認識率から、各話者の特性値を推定し、等化を行って、個別語の期待正解率を推定するという形になる。
具体的な手順を以下に説明する．

＜1，共通語の標準エラー（誤認識率）を，悪い順に累積値を取り，GRG2 を用いて，単語毎の困難度パラメータと識別性パラメータを推定する．＞
50 人分の共通語の認識率データを X とする。（$X = x_1, x_2, \ldots x_{50}$）その X に対して

$$f(x_1 \ldots x_n \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

を取り、この尤度が最大になるような $\hat{\mu}, \hat{\sigma}^2$ を推定する。
そしてこのデータに対する正規累積分布関数値を取る。実際の計算には，Hastings の近似式を使用する。

$$p = F(x \mid \hat{\mu}, \hat{\sigma}) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt$$

を計算する

この累積値と、2PL モデルで非線形回帰(GRG2)を行い、困難度パラメータと、識別性パラメータを計算する例えば「ストレートアヘッド」の困難度は 0.874、識別性は 49.9 である。

＜2，困難度パラメータと共通語の認識率で線形回帰を行い，任意の音節数での困難度を出せるようにする（今回は個別語に適用）＞

困難度パラメータに対して、以下の式を用いて音節数で線形回帰を行う

$$b = \alpha o + \beta$$

ここで b は困難度パラメータ、o は音節数を示している。また α は-0.001、β は 0.711 と負の相関がある。これによって任意の音節数の単語の困難度の推定が可能になる。

＜3，個別語の正答パターンと，1，2 で算出した各単語パラメータを基に，IRT モデルによって個別語話者の話者特性値を推定する＞

話者 $i(i = 1,\cdots,N)$ の $j(j = 1,\cdots,n)$ 番目の単語に対する反応を u_{ij} と表すことにする。但し

$$u_{ij} = \begin{cases} 1 & \text{正解のとき} \\ 0 & \text{不正解のとき} \end{cases}$$

という 2 種類の形でデータ化されている（正答パターン）。項目反応モデルは 2PL モデルを用いる。先ず、ある特定の話者に対する単語への認識は、それぞれの単語で独立であるという局所独立の性質が成り立っていると仮定する。このとき、話者 i の反応が $u_i = [u_{i1}, u_{i2}, \cdots, u_{in}]$ に等しくなる確率は

$$P(u_i \mid \theta_i) = \prod_{j=1}^{n} p_j(\theta_i)^{u_{ij}} q_j(\theta_i)^{1-u_{ij}}$$

となる。ここで $q_j(\theta)$ は誤等確率であり、$q_j = 1 - p_j$ である。これが θ の関数であるので、$P(u_i \mid \theta_i) = L$ と書くことにする。最尤推定量を求めるためには、この式の対数を取り、それを θ i について微分したものを 0 とおいて θ i を求めればよい。すなわち

$$\log L = \sum_{j=1}^{n} [u_{ij} \log p_j(\theta_i) + (1 - u_{ij}) \log q_j(\theta_i)]$$

$$\frac{dL}{d\theta_i} = \sum_{j=1}^{n} \frac{a_j(u_{ij} - p_j(\theta_i))(p_j(\theta_i) - c_j)}{p_j(\theta_i)(1 - c_j)} = 0$$

である。

＜4，ここで個別語＋共通語、計 150 語を 75 語ずつの 2 グループに分ける。そしてその各々に対して期待正解率を計算し、両者の相関を検証する。期待正解率は単語の組み合わせによって大きく変わったりしないのが理想なので、この相関が近いほど、推定の精度が高いということができる。＞

共通語＋個別語、計 150 語を 75 語ずつの 2 グループに分ける。分け方はランダムである。この 75 語に対して、先ず個々の単語の正解率を IRT モデルを使って計算する。

単語のパラメータには困難度パラメータ、識別性パラメータ、（3PL なら当て推量パラメータも）があるが、その中で今実験では困難度パラメータを基準にした。このパラメータは認識における各単語の難しさを表すものであるがこれが認識率に一番関わってくると考えたことからである。また、今回任意の単語の困難度を推定するのに音節数を利用した。今回のこれは一例であり、音素構造など他にも基準になるものは考えられる。しかし、単語内の要素としては使える材料があまり無いことから、今回は音節を用いている。

3 章　実験

3.1　固有名詞単語発声コーパスの構築

今回の実験は、携帯音楽プレーヤーや、モバイル環境における楽曲ダウンロードサービスの音声フロントエンドとなるシステムを想定した語彙でおこなった。語彙は、音楽配信サイト mora(http://mora.jp)から、2006 年 6 月に抽出した情報を用いた。同サイトではアーティスト名、アルバム名、曲名合わせて 175962 語の名称が利用されている。それぞれの名前は全て単語ではあるが、中には文章やフレーズになっている長いものも存在する。

以下は、使用した単語の例である。

・ケレル
・クレージーケンバンド
・ボーチアルモニケボーカルアンサンブル
・ストップインザネームオブラブ
・アタシナンデダキシメタインダロー

評価用の音声サンプルは話者ごとに 150 語ずつ収録した。発声用のテキストとしては、全話者に共通の 50 語(アーティスト名 20 語、アルバム名 20 語、曲名 10 語)、話者ごとに異なる 100 語(アーティスト名 30 語、アルバム名 30 語、曲名 40 語)を選んだ。発話した話者は計 50 名で、内訳は男性 36 人、女性 14 人である。合計で、共通語が 2500 サンプル(50 語×50 名)、個別語が 5000 サンプル(100 語×50 名)である。

共通語は全員同じ単語を発話することによって、話者毎の違いを比較するために収録した。また、個別語は共通語によって推定された話者要因から、個別語の期待正解率を推定（比較）できるようにするために収録した。

収録は雑音の少ない静かな部屋で行った。収録の前に、５単語程度で発声と収録の練習を行った。
その他の音声データ、使用機器については以下の通りである。

・音声データ
サンプリング周波：16kHz、量子化ビット：16bit、チャンネル数：モノラル、PCM 音源
各音声データは正解単語と時間でアライメントを取り、音声部分前後 0.3 秒以外は切り取った。

・使用機器
音声収録：ヘッドセットマイク　ゼンハイザー
　　アンプ
　　音声キャプチャ
データ処理用ＰＣ：FMV-S8210
使用ソフトウェア：音声認識　julian-kit(version3.1)
　　音声収録　JuliusTest
　　波形処理　sp-wave
　　項目反応理論推定　EasyEstTheta

・使用した認識エンジンと認識方法

認識には音声認識エンジン julius（version3.5.3）[8]を用いた。
音響モデルに使用した HMM の詳細を表 1 にまとめた。

表 1　HMM の詳細

種類	mono-phone
特徴量	MFCC
パラメータ次元数	15

パラメータ混合数	256
共分散行列	対角行列

ここで tri-phone でなく mono-phone を用いる理由は、tri-phone の限定性による。tri-phone に含まれる連続要素は、特定の領域に限定されて構築されたものであるため、mono-phone よりも適応範囲が狭い。その分 tri-phone は高精度な認識が行えるが、今回は認識させる語彙の単語の多さもあって、mono-phone を使用している。
また、認識における設定値は表 2 にまとめた。

表 2　認識オプション

探索候補数	250
探索スタックサイズ	1000
スタックオーバーフロー	4000
第 1 パスの探索ビーム幅	3000
第 2 パスの探索ビーム幅	1000
CM アルファ係数	0.05

この認識手法ついて簡単に説明する。先ず、julius では認識を 2 段階に分けて行う。第 1 段階の認識（第 1 パス）で先ず荒い認識を行い、ある程度候補を絞ってから第 2 段階の認識（第 2 パス）で高精度の認識を行う。こうすることで高速で精度の良い認識が実現できる。それぞれの認識では枝狩りを行う「ビームサーチ」を用いる。枝狩りは一般的に「低い確率が与えられた状態については以降の計算を行わない」ことを意味するが、この認識でももっとも高い状態確率からビーム幅オプションで指定された数の状態のみ計算するような仕様になっている。
音響モデルの学習データには、評価用の話者は含まれない。

・認識用辞書

辞書サイズを増減させる実験では、上記の個別語 5000 語に、ランダムに抽出した単語を加えて認識用辞書を構築した。構築した辞書はサイズが、5000、10000、20000、30000、40000、50000、60000、70000、80000、90000、100000 の、計 11 個である。発声する単語は全て辞書内に含まれており、未知語は存在しない。
以下は、作成した辞書の例である。

・イトーヒデシ　i t o: h i d e sh i
・アンダーザムーンライト　a N d a: z a m u: N r a i t o
・ショーンポール　sh o: N p o: r u
・キョーリューノホネミタイナクモバンネン　ky o: ry u: n o h o n e m i t a i n a k u m o b a N n e N
・ジュエル　j u e r u

辞書内の単語は全てランダムに抽出して構築されているが、いくつかの単語は重複している。以下の表は、各辞書の単語の重なりを示すものである。

表 3　各辞書の単語の重なり

	5000	10000	20000	30000	40000	50000	60000	70000	80000	90000	100000
5000	5000	5000	5000	5000	5000	5000	5000	5000	5000	5000	5000
10000		10000	5501	5723	5958	6170	6369	6539	6807	7084	7220
20000			20000	11767	12684	13947	14718	15345	15958	16604	17000
30000				30000	17859	19361	20733	22051	23030	23990	24957
40000					40000	25116	27160	28733	30270	31671	32853
50000						50000	33356	35521	37301	39123	40756
60000							60000	42334	44735	46680	48802

70000								70000	51807	54372	56644
80000									80000	62114	64755
90000										90000	72662
100000											100000

また、以下に各辞書に含まれる音節数のヒストグラムを示す。

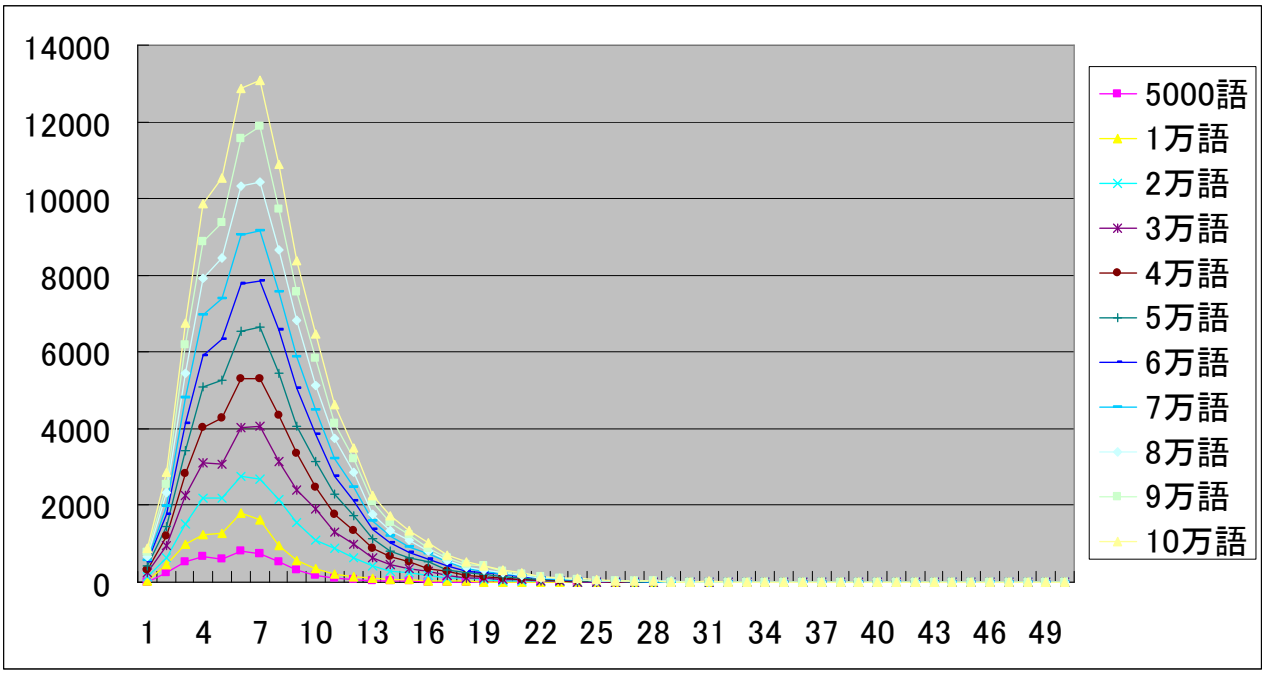

図 6　各辞書内の音節数

各辞書の音節の平均値はそれぞれ 6.22、6.36、7.29、7.36、7.44、7.5、7.52、7.54、7.55、7.56、7.56 であった。次に、各辞書に含まれる音素数のヒストグラムを示す。

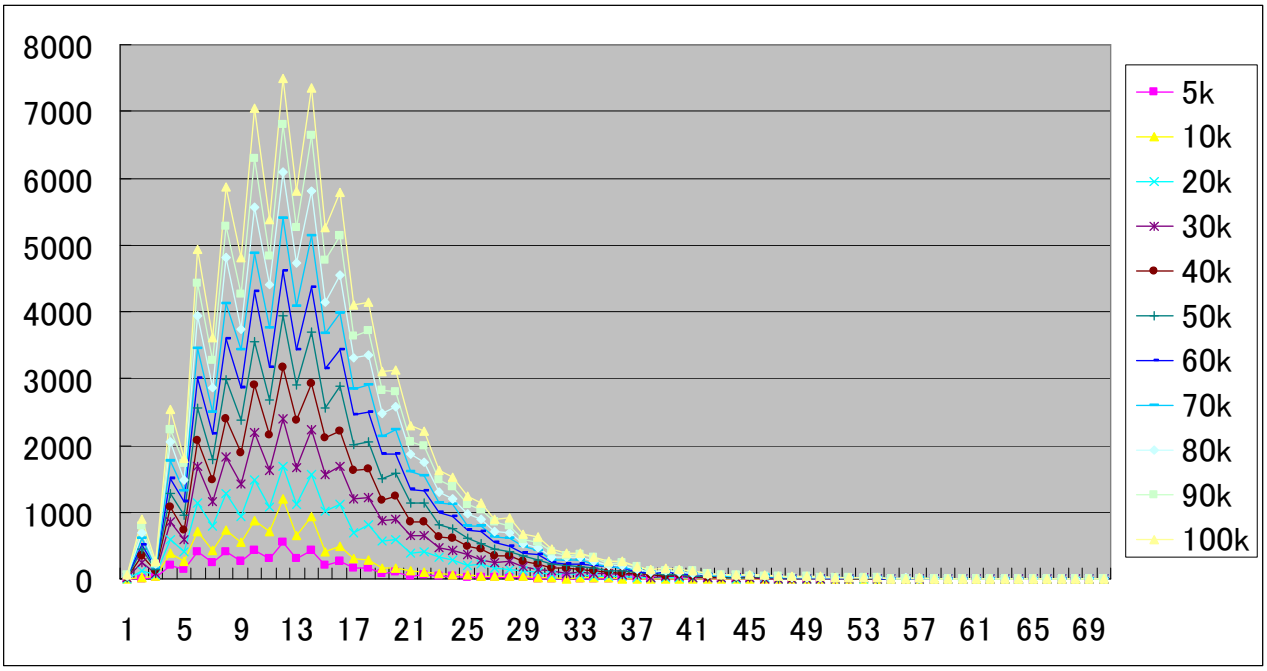

図 7　各辞書の音素数

各辞書の音節の平均値はそれぞれ 11.96、12.28、14、14.13、14.30、14.38、14.44、14.49、14.49、14.5、14.5 であった。

3.2　期待正解率の推定実験

ある話者がシステムを利用したとき，辞書サイズによって，どのような性能になるかを推定できるかどうかを検証するための実験を行った．

最初は混合 2 項分布モデルによる推定である。この実験は最尤推定によるものであり，ある単語を用いたときの相対的な値でしかない．つまり比較による実験である．システムを使う前に，システム語彙の中から，少数(100 語以下)の語を発話し，そのサンプルを用いて，この話者がシステムを利用したときの性能を推定するようなタスクを想定している．

5 名の話者を選択し、各話者それぞれに対し、共通語と、個別語それぞれを用いて各辞書に対して認識させ、辞書サイズ-エラー率曲線を描く。次に、そのエラー曲線から Rosenberg モデルの(6)式を用いて pv を推定し、理論曲線を描く。便宜上、前者を個人別エラー曲線、後者を基準エラー曲線と呼ぶ。個別語の基準エラー曲線と共通語の個人別エラー曲線、共通語の基準エラー曲線と個人別エラー曲線の相対誤差が、性能推定の精度を示すとみなせる。

先ず、この 5 人分の標準エラーを、混合数 3 で回帰したグラフを示す。表 4,5 にそれぞれの相関を示す。また、表 6 に、各人の正解確率の推定値を示す。

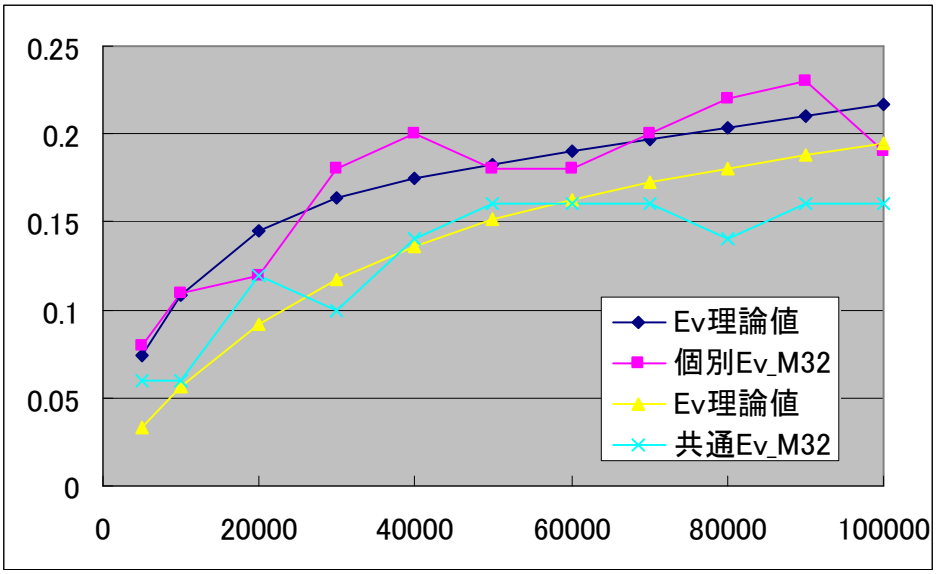

図 8a　男 32 の Ev 理論値と実測値

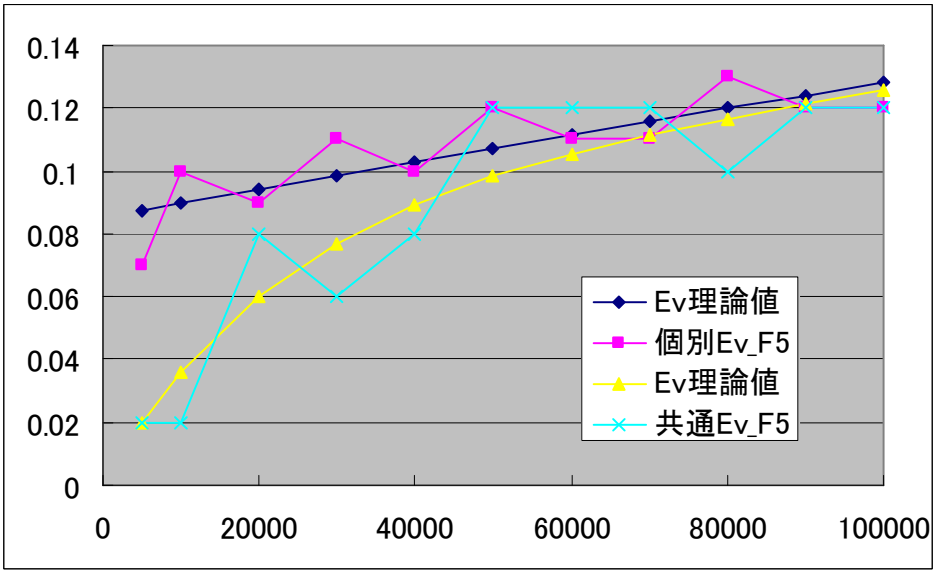

図 8b　女 5 の Ev 理論値と実測値

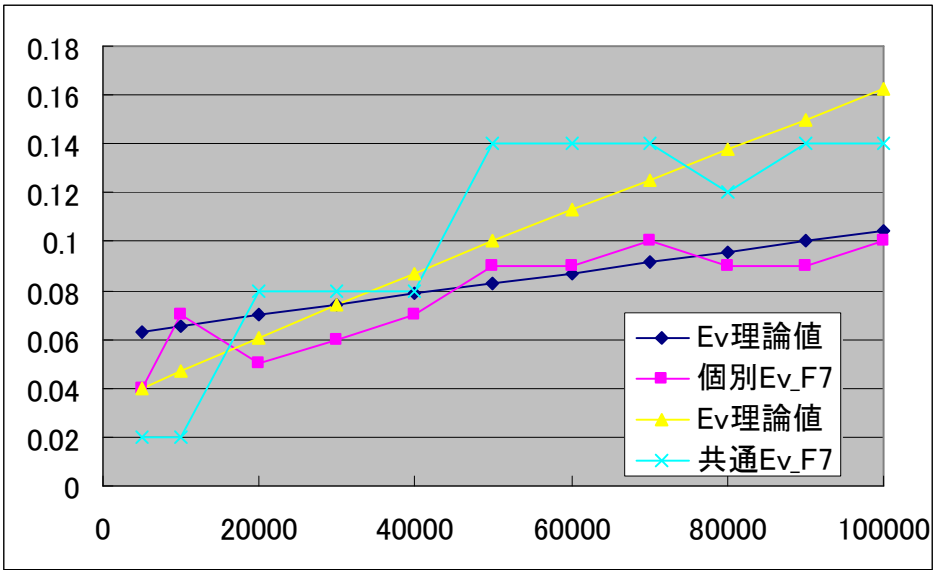

図 8c　女 7 の Ev 理論値と実測値

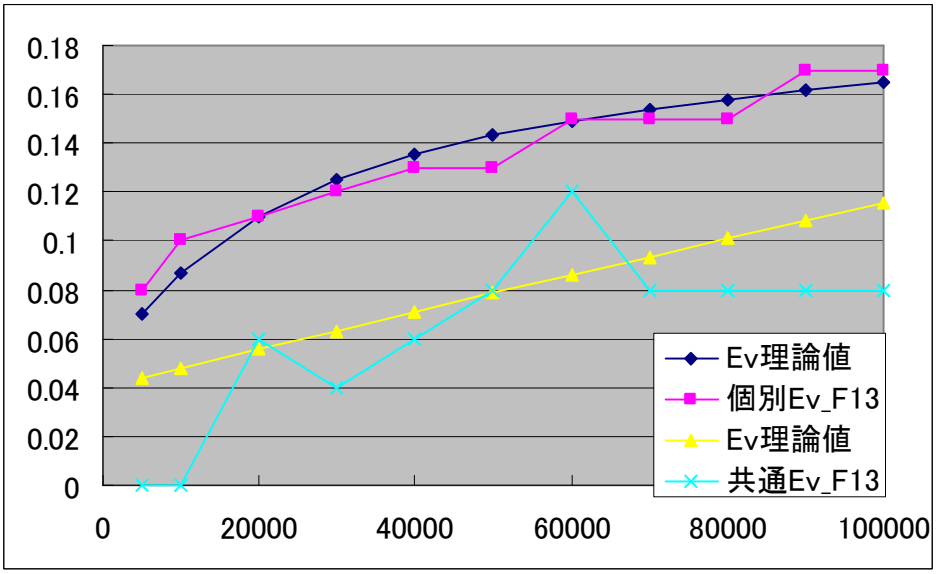

図 8d　女 13 の Ev 理論値と実測値

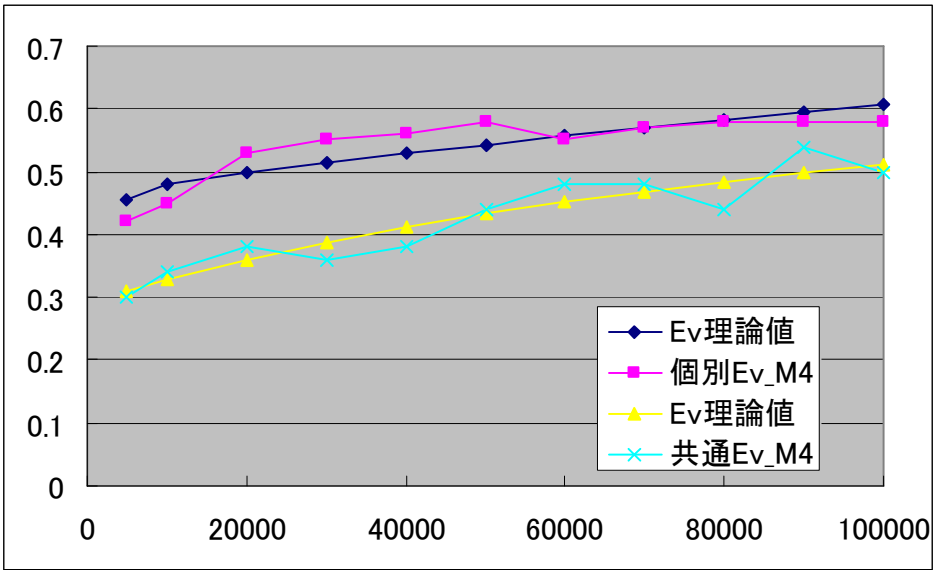

図 8e　男 4 の Ev 理論値と実測値

表 4　Ev 理論値と実測値の相関（個別同士・共通同士）

	個別	共通
M32	0.830633	0.857151
F5	0.888318	0.946846
F7	0.91912	0.891516
F13	0.92984	0.814138
M4	0.877835	0.883721

表 5　Ev 理論値と実測値の相関（個別実測値と共通推定値・共通実測値と個別推定値）

	個別実測・共通推定	共通実測・個別推定
M32	0.924719	0.683708
F5	0.942676	0.922106
F7	0.918934	0.88954
F13	0.905126	0.888165
M4	0.878845	0.883027

表 6　正解確率とその重み

	正解確率(m=3)	重み(m=3)
M32	0.9,0.9999,1	0.006,0.1,0.85
F5	0.9,0.9999,1	1E−10,0.08,0.91
F7	0.9,0.9999,1	0.03,1E−10,0.966
F13	0.9,0.9999,1	0.03,0.09,0.87
M4	0.9,0.9999,1	0.29,0.14,0.56

図 8 より、話者によって、発話サンプルが異なると、認識性能が大きく異なる(図 8c)ことがあることがわかる。また、話者ごとの個別語の Ev 実測値は、どれも辞書サイズに対して単調増加していないことがわかる。50 人の平均 Ev は図 10b（第 4 章）からも解るように単調に増加している。辞書サイズを変更した実験においては、発話サンプルはどのサイズにおいても同一のものを使用しているので、エラー率は辞書サイズの増加と共に増加、横ばいになることはあるにしろ、減少はしないと考えるのが普通である。

この違いは、100 語の分散と 5000 語の分散の違いから生じると考えられる。即ち 100 語ではまだサンプル数が少ないので分散が大きく、5000 語程度のサンプル数が無いと分散がある程度小さくならないということである。式(6)は単調増加を前提としているので、結果として 100 語程度では高精度なモデル化に必要なサンプル数が足りないと判断できる。

途中でエラー率が低下する理由として考えられる理由は大きく 2 つある。

1 つは辞書による影響。例えば 10000 語辞書を作るのに、5000 語辞書からプラス新しい語を 5000 語加えて作るのならば、エラーは横ばいになることはあるにしろ減少することはありえない。しかし今実験では辞書サイズの 5000 語と 10000 語では原データ 170000 語から全てランダムに選ばれて作られている。従ってたまたま 10000 語辞書の単語の中から、5000 語のときに誤認識された単語が消えて、認識エラーが減少する可能性はありうる。

2 つ目は探索エラーによる影響。今実験は Rosenberg のように単語間距離ではなく、枝刈を行う認識器の認識結果を使用している。従って、登録単語の変化によって、枝刈に影響を与えて最終的には上位となるようなスコアを得るような単語でも途中で枝刈されてしまうことが考えられる。

以下の図は、共通語 10 万語における、認識率別のヒストグラムである。

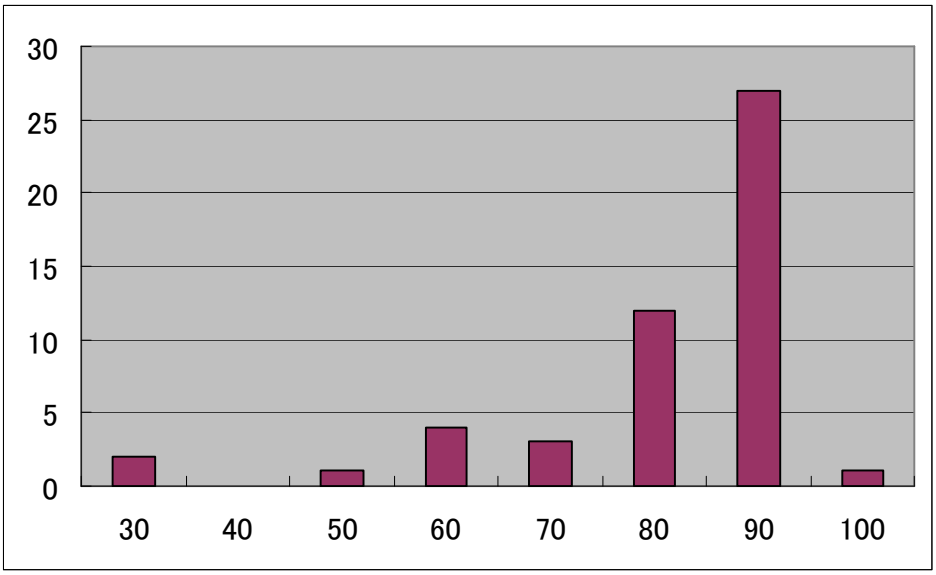

図 9　共通語 10 万語の認識率別ヒストグラム

この図から、共通語の各単語の認識率は 0.3～1 の間に収まることがわかる。この結果を表 6 の推定結果と比較すると、その値や比率は実態と乖離していることがわかる。また、図 8 から実際の個別語による認識率と比較しても、実際の認識率が変動していることに対して推定値は滑らかに推移している。これらのことから、最尤推定による期待正解率の推定は、精度があまりよくないということができる。

最尤推定による推定が上手くいかない理由は、単語の誤り確率の分布が、最尤推定で表せないくらいより複雑な形であることによると考えられる。

以下の図 10 は個別語 100 語を 50 語ずつにランダムに選んだときの、単語誤り率の散布図である。

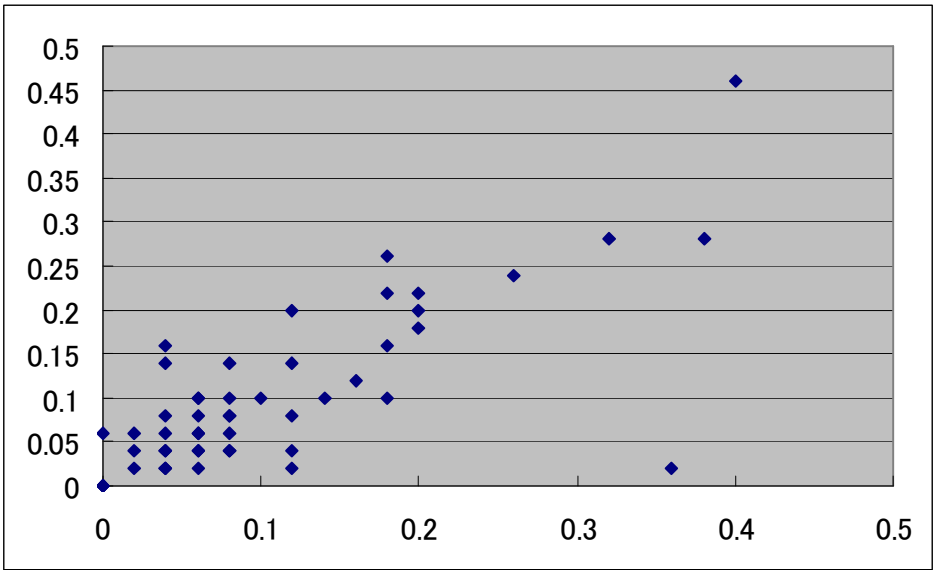

図 10　個別語 50 語ごとの相関

相関は 0.75 と余り大きくない値となった。このことから、同じ単語数でも誤り率は一概には決まらず、発声する単語によって期待正解率が大きく変動することがわかる。

以上をまとめると、小数サンプルで期待正解率を高精度に推定するためには、より単語の誤り率と個人の認識能力を結びつける推定方法を考える必要がある。

この問題に対処するため、項目反応理論を利用して，期待正解率の推定を行う．そこで先ず、任意の単語の誤り率を推定するために、全ての単語が持つ共通の属性でモデル化する方法を考える。以下の図 11 は個別語 10 万語における音節数ごとの認識率を表したものである。

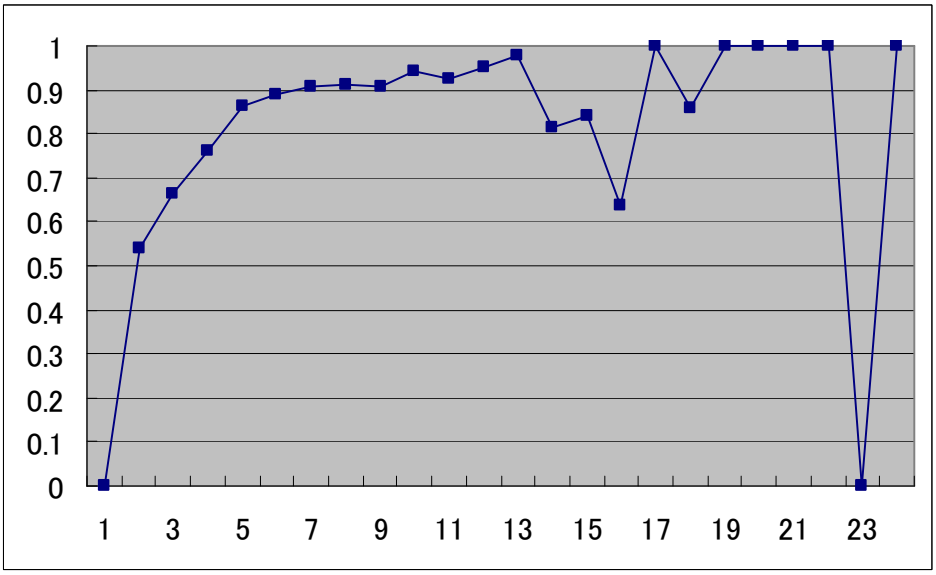

図 11　個別語 10 万語の音節数別認識率

音節数の大きい（13 以上）ものはそうでもないが、それより小さいものに対しては、音節数の小さいものほど、すなわち単語の短いものほどに認識率が悪くなっていることが見て取れる。このことから今実験では、音節数を全ての単語に共通する属性として、項目反応理論における認識の困難度を 5 段階に分類してモデル化した。

今回は 2PL モデルを用いるので pi の推定には，困難度パラメータ，識別性パラメータ，話者特性値の 3 つが推定できれば良いことになる．
推定の手順を以下に説明する．

1. 評価用の話者の発話を推定用 75 語と評価用 75 語に分割
2. 推定用サンプルで話者特性(1)を推定
3. 評価用サンプルを用いて同様に話者特性(2)を推定
4. 話者特性(1)と話者特性(2)を比較

期待正解率が高精度に推定できていれば、誤差は小さくなるはずである。比較用に単なる認識率の誤差を検証した。以下に、その結果を示す。

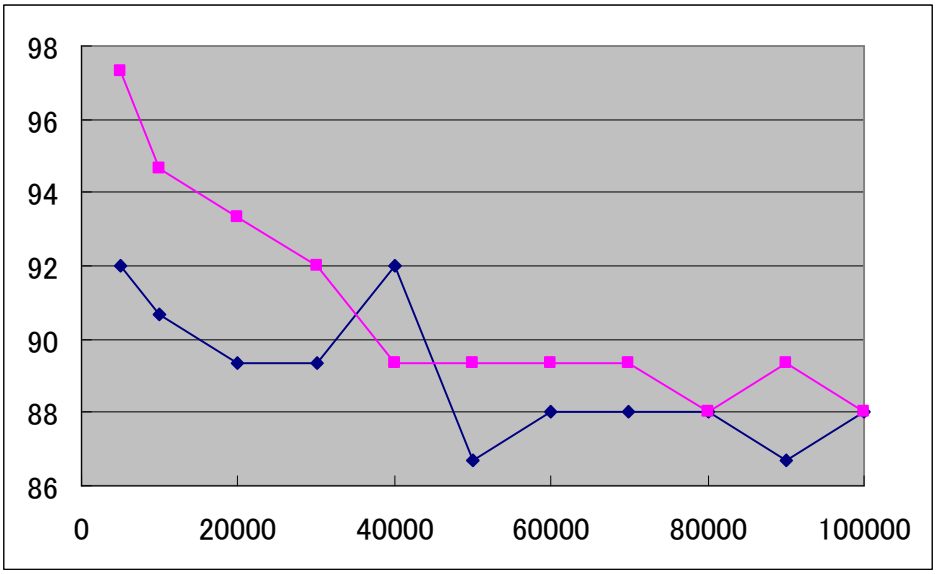

図 12a　女 5 の認識率（推定用 75 語と評価用 75 語）

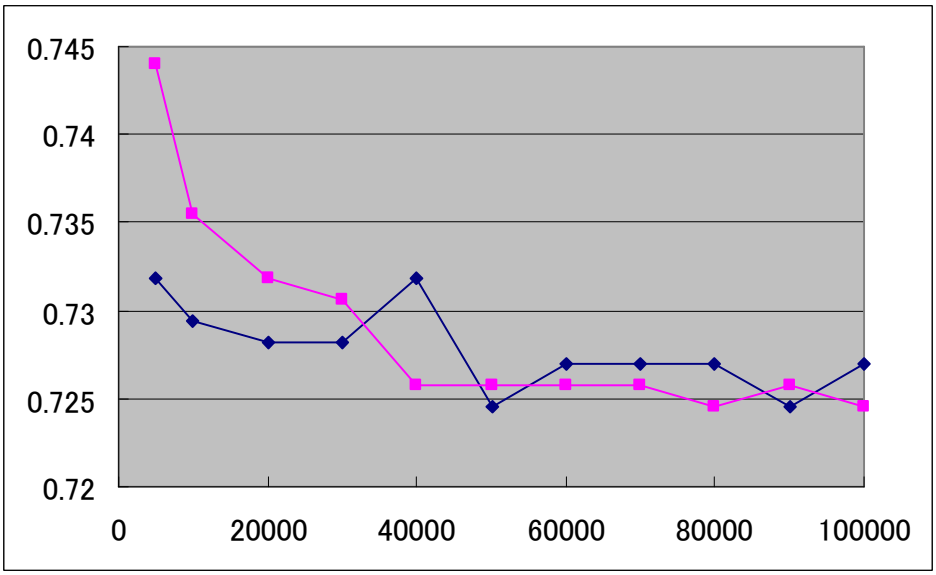

図 12b　女 5 の期待正解率（推定用 75 語と評価用 75 語）

表 7　F5 の認識率と期待正解率の 2 乗誤差合計

	2 乗誤差合計

認識率	0.009
期待正解率	0.0002

図 12 から、グラフの形はほとんど同じだが，2 乗誤差の合計は減少した。これは同数のテストセットが同じ期待値に近づいたということであり、項目反応理論を使用した推定方法は，大語彙・小サンプル推定に有用である可能性を示唆している.

4 章　考察

4.1　最尤推定による推定結果と、項目反応理論による推定結果の比較

最初に，項目反応理論による推定の結果を考察する．今実験のような、2 乗誤差による検証では、誤差が減り近い値にはなったものの、大きな改善とは行かなかった。これは IRT モデルによる差が小さいことに拠ると思われる。以下の図 13 は、今実験で 5 段階に分類した困難度の項目特性曲線である。

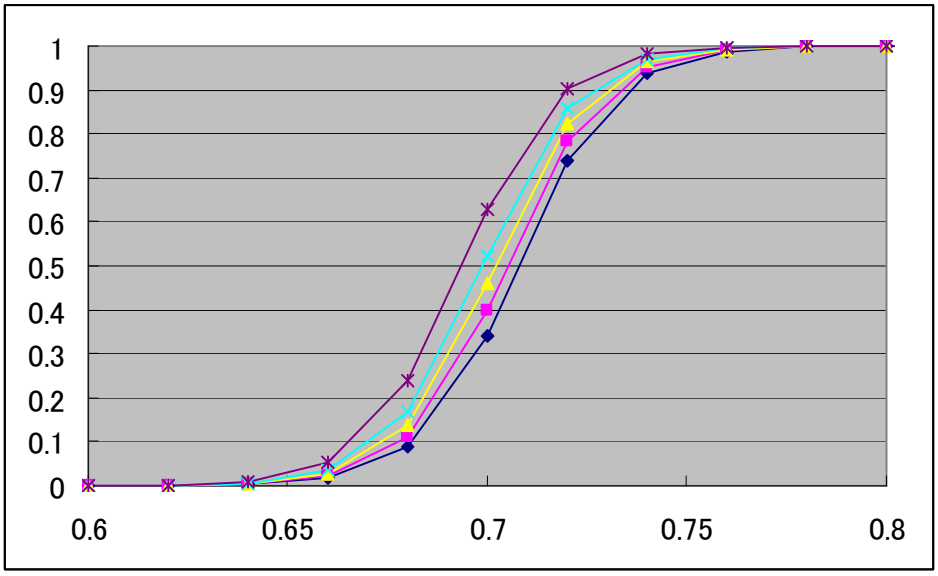

図 13　困難度 5 段階の項目特性曲線

図 13 より、困難度が近いことから、一人の話者が、書く困難度の単語をくっきりと識別しにくい状態となっている。このことから、大きな誤差の改善は見られなかったものと思われる。

今回項目反応理論を用いた推定が Rosenberg モデルのみの推定より制度が上がった理由は Rosenberg の期待正解率の推定は最尤と回帰によるもので、一般の統計モデルに多く見られるように画一的に推定される。しかし、この項目反応理論を用いた推定では音節数という、プラス要素を考慮して推定ことが良かったと考えられる。

4.2　Rosenberg モデルの大語彙での適用性について

本実験の本筋から少しそれるので、こちらに掲載する。Rosenberg モデルの問題の説についても述べたとおり、彼のモデルは、本実験と比較して、小語彙で有効とされたモデルであり、大語彙に対してそのまま適用できるかは定かではない。しかし、大語彙に関してもその有効性を検証することは、期待正解率から、個人ごとに最適な語彙サイズを決定させるステップにおいて必要である。本節では、Rosenberg の確率モデルを大語彙に対して追試する。平均順位、標準エラーの実測値と推定値を比較し、検証する。また、混合モデルによる m と、推定精度の関係についても確認する。

実験において、順位、標準認識エラーの各値は、個々の単語、話者の平均値を取る。従って各変数は以下のように書き直される。今、単語 vi が $V(N)$（サイズ N の辞書）に含まれるとする。与えられたランクを $r_{I,V(N),s}$ とすると、平均順位、効率性エラー、標準認識エラーはそれぞれ

$$\bar{r}_{s,v}(N) = \frac{1}{N}\sum_{I \in V(N)} r_{I,V(N),s}$$

$$\bar{e}_{r,s,V}(N) = 1 - \frac{1}{N}\sum_{I \in V(N)} \frac{1}{1 + r_{I,V(N),s}}$$

$$\overline{E}_{r,s,V}(N) = \frac{1}{N} | \{ I : I \in V(N), r_{I,V(N),s} \geq 0 \} |$$

と表される。|{}|は、カッコ内のイベントが起こる回数を表す。

今実験はこの３つの指標の中で平均順位と標準認識エラーのみを考察材料にしている。

4.2.1　認識結果・平均順位・標準エラーに対する、Rosenberg（小語彙）の実験との比較

共通語 2500 語、個別語 5000 語分の認識結果を以下に示す。

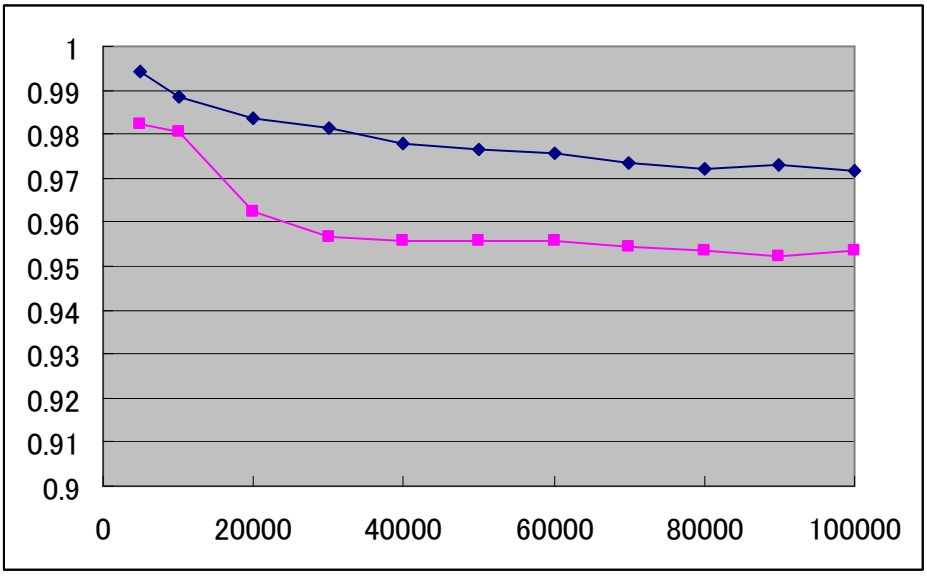

図 14　共通語と個別語の認識率

今回は 250 位まで出力して、順位の出なかったものに対しては一律で 1001 位と定めてあるが、辞書内に正解単語が必ず含まれていることから、このような単語があるのは適切とは言えない。1001 位になってしまった単語の数は全体の 5%程度ではあるものの、無視して良い現象とは言い難い。これについては後に考察対象とする。

認識率は、共通語・個別語どちらも 10 万語で 95%を超えていうように十分な値を出している。Julius の平均的な認識率が 20000 語辞書で 93～95%であったことから比較しても、遜色ない値であると言える。

以下の図 15 は、辞書サイズの関数として、50 人全員の平均順位と標準認識エラーをグラフ化したものである。また、標準エラーのグラフの横軸は対数表記してある。

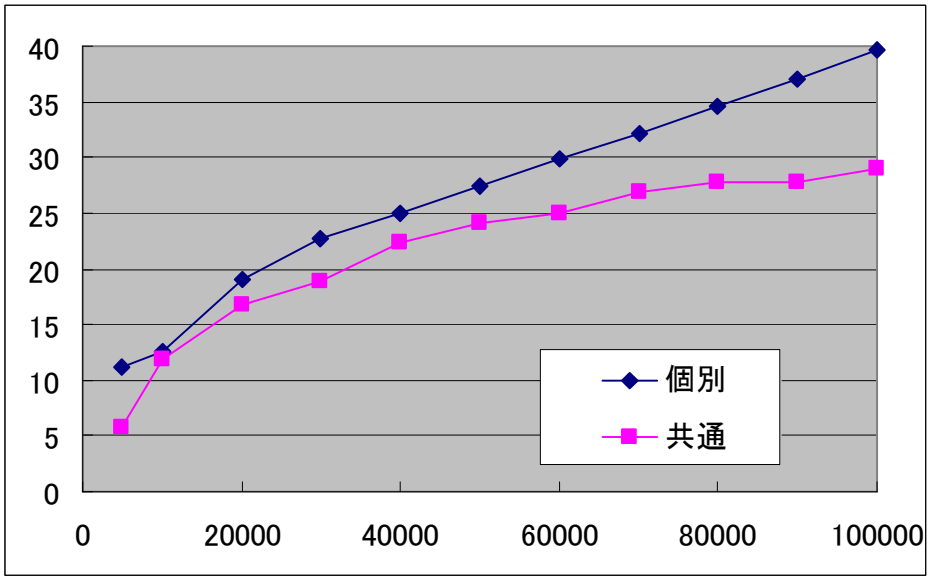

図 15a　平均順位

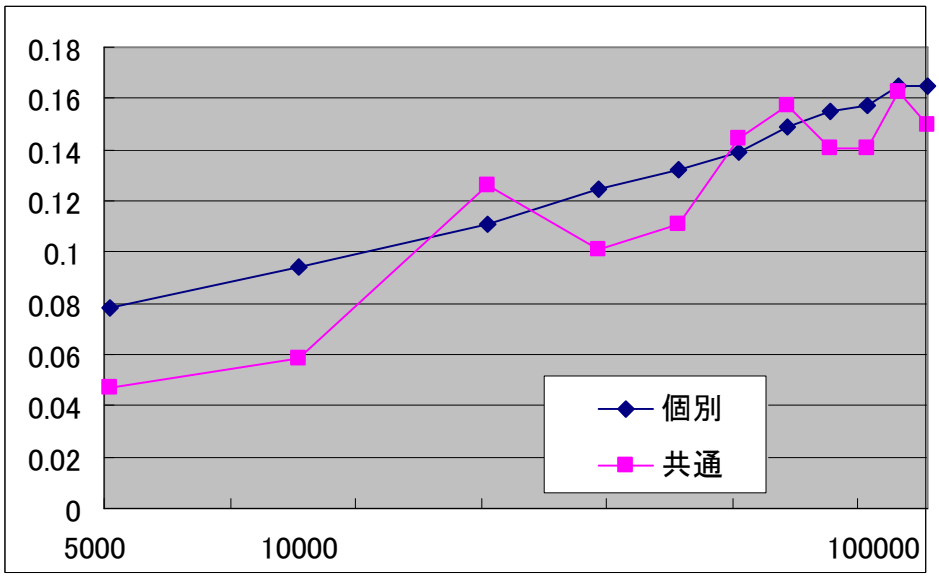

図 15b 標準エラー

表 8　本実験から推定された pv と、Rosenberg の実験で推定された pv

	本実験（個別）	本実験（共通）	Rosenberg
平均順位からの pv 推定値	0.000288	0.000213	0.00667
標準エラー(m=3)からの pv 推定値	0.004405	0.004403	0.00835

平均順位は（4）式のモデルに従うなら、辞書サイズに線形になる筈である。図 15a を見ると、個別語はほぼ直線になったが、共通語はそうならなかった。切片をバイアスだとみなして、（4）式のモデルを使用する。すると、この関数の傾きが、（4）式で表現される pv の推定値になる。

図 15b から、個別語の標準エラーは大体辞書サイズの対数に線形に増加していることがわかる。また、共通語の標準エラーを示すグラフは個別語のそれに比べて、上下の変動が大きくなっている。

ここで、表 8 を見て欲しい。表 8 は、平均順位から推定された pv と、標準エラーから推定された pv を表したものであるが、その値は大きく異なっている。実は Rosenberg の実験でも(4)式の傾きで推定される pv と（6）式から推定される pv には食い違いが生じている。

Rosenberg はこの原因について、混合数が 3 つしかないことに統計的な問題を提示している。本実験ではこれについて野値の項目反応理論を用いることで対処しているが、Rosenberg 自身もこのモデルが完全でないことを示していたのは注目すべきである。この事実と、（4）式では、複数の pi を推定できないことから、今後 pv の推定は（6）式を用いる。

本実験の pv 推定値と、Rosenberg の推定値について比較する。本実験の個別語と共通語から推定される pv はどちらも似た値になった。また、Rosenberg モデルの推定値との比較に関しては、標準エラーからの推定値は比較的近い値ではあるものの、平均順位からの推定値は一桁ほど異なっている。これは、Rosenberg モデルでの推定値と実測値の相関が 0.99 以上であるのに対し、本実験の相関値は個別が 0.95、共通語が 0.92 と、それほど大きくない値になっていることから、まだ Rosenberg モデルと比べて、サンプルが近い値になっていないことを示すものだと考えられる。これと比べると、標準エラーからの推定値は、本実験も Rosenberg モデルも、相関値が 0.99 を超えているので、ほぼ近い値になっていると思われる。

この実験において、個別語に対してはおおよそ Rosenberg モデル通りになったが、共通語はそうならなかったことがわかる。原因は共通語の方に、カクカクとした変動が起きていることである。

これらの違いは発声テキストのバリエーションの違いによって起こると考えられる。共通語のサンプル数が合計 2500 語（各 50 語×50 人）、個別語のサンプル数が合計 5000 語（各 100 語×50 人）で、サンプル数自体は 2 倍の違いしかない。しかし、共通語の 50 語が全員に共通して発声されているのに対して、個別語の 100 語は個人によって全く違うものが発声されている。従って、共通語は全体でもバリエーションが 50 語なのに対して、個別語は 5000 語のバリエーションを持っている。つまり発声テキストのバリエーションは 100 倍異なる。したがって、発声テキストのばらつきが不十分だと性能推定にばらつきが生じうると考えられる。

4.2.2　混合モデルに対する、Rosenberg（小語彙）の実験との比較

次に、標準エラーのグラフから、hm、pm を推定する。推定の式は(6)式を用い、非線形回帰をおこなった。回帰には GRG2 を用いた[9][10]。初期値はいくつかを適当に選び、得られた回帰曲線の相関が高いものを採用している。混合するパラメータの数 m は、数が大きいほどよく近似することが知られている。しかし同時に、他との峻別が十分であること、単純である等の理由から、Rosenberg の論文では混合数 2 が採用されている。以下に標準エラーの実測値と m=1~4 に対する回帰線、パラメータ、モデルとの相関を示す。

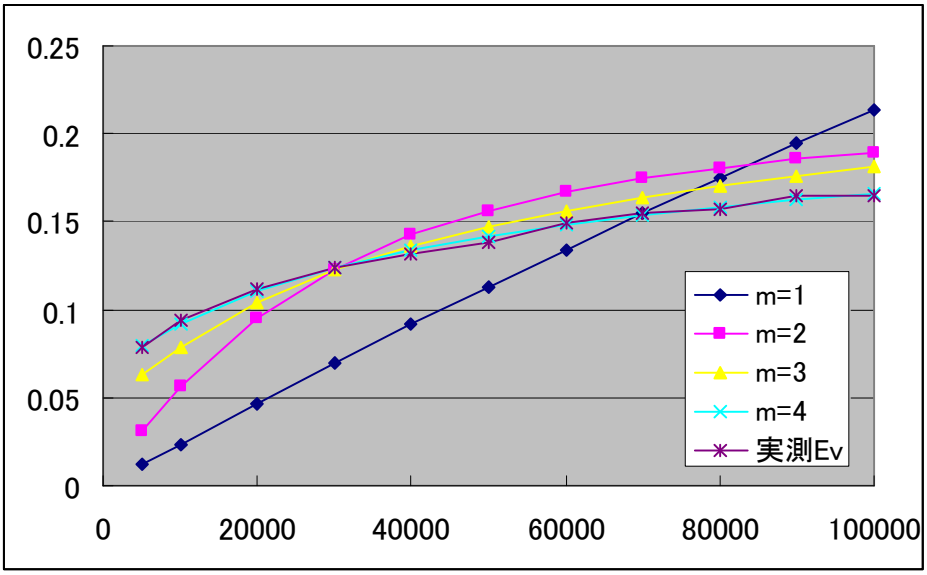

図 16　標準エラーと m=1~4 の混合モデル回帰線

表 9　標準エラーから推定されたパラメータ

	H1,h2,…,hm-1	p1,p2,…,pm	相関 r
M=1	-	2.759E-06	0.9620
M=2	0.1662	0.0001 6.7583E-07	0.9753
M=3	0.0434 0.0898	0.1 0.0001 5.8718E-07	0.9905
M=4	0.0286 0.0995 0.1873 0.6844	0.1 0.0002 3.18E-06 1E-07	0.9921

相関 r は実測値と推定値に対して

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

を取ったものである。σ_x は x の標準偏差、Cov(X,Y)は X と Y の共分散を表す。

表 9 からも解るとおり、m が大きくなるにつれて相関が高くなっていくことがわかる。

4.3　誤認識に関する考察

4.3.1.　順位の出ない入力サンプルの原因について

今実験においては、順位が 250 位までを出力して計算したが、それを超えるもの（250 位以内にスコアが入らなかったもの）に関しては、一律 1001 位として計算している。しかし正解単語が辞書内に必ず含まれていることから、250 位以内に含まれないというのは適当とは言えない。この説では先ず認識の結果が 1001 位になってしまった入力サンプルの原因を考察する。

今回は、その中でも、一番小さい 5000 語から 100000 語辞書までの認識結果全てで 1001 位になってしまった入力サンプルに限定する。そのような単語は個別語 38 個、共通語 11 個で合わせて 49 語あった。これらのサンプルを実際に聞いてみた感じ、音声波形、出力された解候補等から 1001 位になってしまった原因を調べる。今後実験の収録の参考になれば幸いである。

考察した結果、以下のような原因が考えられた。

・英語特有の表記「トゥ」「ファ」「デュ」「ウィ」「ウェ」「ウォ」等の間違いが目立つ
今回の実験は携帯音楽プレーヤーや、モバイル環境における楽曲ダウンロードサービスの音声フロントエンドとなるシステムを想定した語彙で行っているので、や外国のアーティスト名や楽曲名が多く含まれる。そのような語彙に対し、認識エンジンは完全日本語対応の音素で行われるため、カバーしきれないという印象があった。単語とエンジンにはそれぞれ日本語表記がされているので、その単語の通りに読めば問題は無いのだが、発声者の先入観と英語の経験から、書かれている単語が英語だとわかると、そのような発音になってしまうことが少なからずあった。

・「ア」と「オ」の間違いが目立つ。
これはフォルマント的にも近しいので無理も無いが、出力された解候補では、この間違いが多く見られた。発声の際にはここの区別に注意して発声するように注意したい。以下は単語「アーワイユーレディ」の間違いの例である。

1 位　ソーユーセイ
2 位　モユルヒ

3 位　イトーヒデシ
4 位　フォーユー
最初の音素「ア」が、子音は除くが 1 位、2 位、4 位で「オ」と認識されている。勿論後ろの音素の数や構造も含めての認識結果だが、見た感じでは少し目立った。

・誤認識のされかたは話者に大きく依存する
単語の誤認識のされかたには様々なものがある。音素的に近いと思われる間違いもあれば、どうしてこの単語に間違えるのかわからないほど外れたようなものも見たところではあった。そのような「間違いの近さ」は話者に依存するような傾向が見られた。例えば A さんがある単語を一つだけ間違ったとしても、降雨目反応理論でいう能力値の高い人ならば、たいてい間違えた単語は候補を見てどういう現象（置換、挿入、脱落など）で間違えたかなどが納得できるような間違い方をする。しかし、例えば B さんが単語をいくつも間違えているような能力値の低い人だとするとたいていその人の単語の間違いは外れた（ように感じる）ものになる。原因としては、たいてい発声に問題がある（「不明瞭発音」「発声が速い」）ように感じるものが多い。

・単語間の切れは「ッ」と解釈される
英語の名前など、日本人にとって意味の無い言葉の羅列では仕方ないかもしれないが、不必要に単語を切って発声する人が多かった。このような間は、促音に解釈されることがある。被験者は読む前に一度は目を通して声に出してみて、読み方を理解しておく必要がある。

その他には、音量の小ささに関しては特に影響は見られなかった。大体音量のレベルは 10000 かそれ以上に最大値がくるのが良いとされるが、今回は非常に声の小さい人で 2000～4000 レベルの人も少なくなかった。しかし、そのような声の人も 1 位になるものはなるし、そうでないものは大抵他の原因（発音が悪い、発声が速い等）が考えられるものが多かった。このことから、音量に関してはさほど厳密に考慮する必要は無いのではないかと思われる。

4.3.2　一人当たりの必要サンプル数の推定

上述したように、二項分布モデルを用いた場合、100 語程度では、高精度の推定が難しい。それでは、どのくらいデータがあれば安定した性能推定ができるのか考察してみた。一人当たり 100 語から 1000 語のサンプルを取り直し、100 語から 1000 語までランダムに取り出したときの標準エラーを再計算した（図 13）。同時に、多サンプルの場合、平均を推定するのに必要なサンプル数も考察した、（図 14）

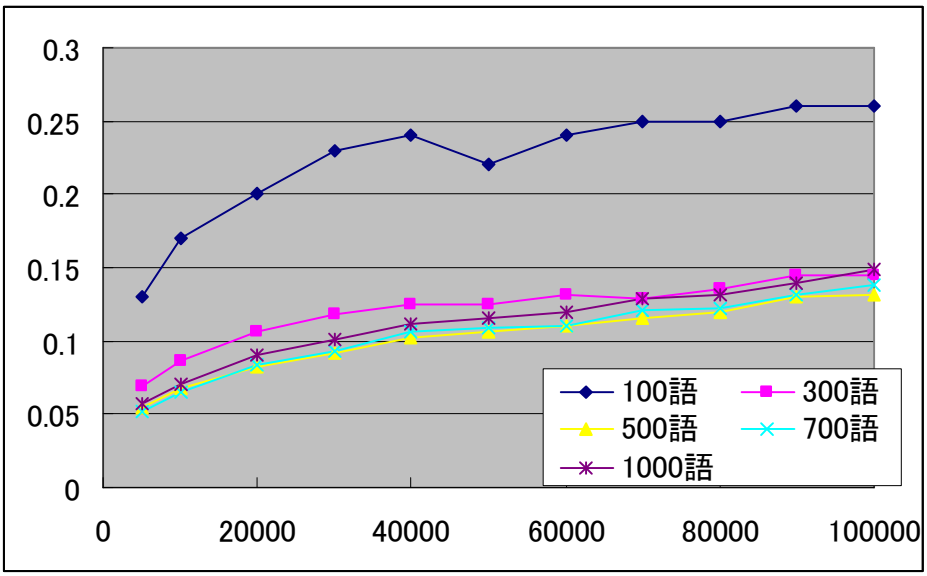

図 17　単一話者の 100～1000 語のサンプル数による標準エラー

図 17 より、一人当たり約 500～750 サンプル数あれば、途中で減少が起きないことがわかる。また、

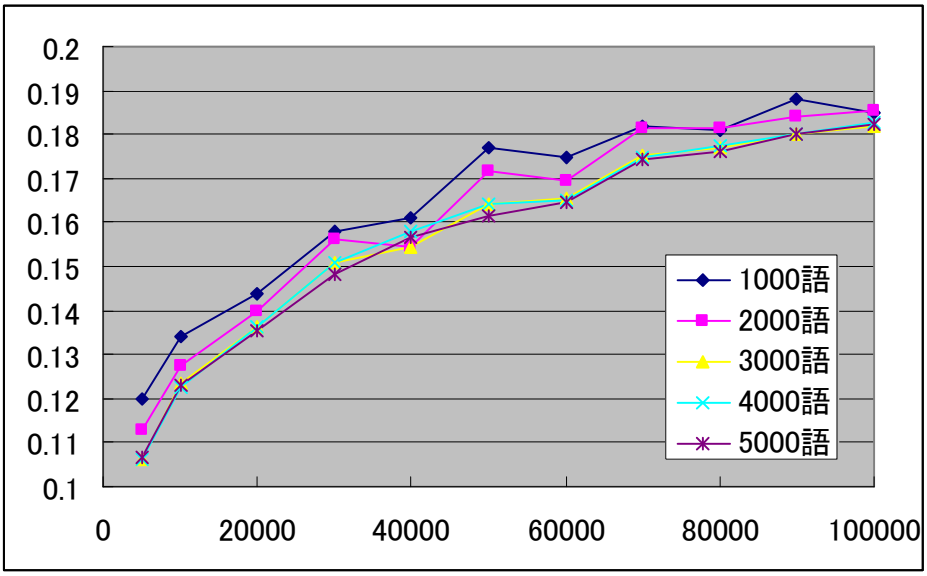

図 18　全話者・全単語からランダムに 1000～5000 語選んで計算したときの標準エラー

図 18 より、合計約 2500～3000 サンプル数あれば、途中で変動がおきないことがわかる。

5 章　結論

本論文では、項目反応理論による期待正解率の推定法を提案した。同数のテストセットによる精度検証で、誤差が小さくなったことから、安定した推定ができる可能性を示唆した。

今後の課題としては先ず，この推定結果を用いて、各個人に最適な語彙サイズの推定実験を行うことがあげられる。

また、今回は音節数をすべての単語の共通の属性として基準にとり、困難度を分類したが、音節数と困難度の相関の絶対値は実は 0.4 程度であった。もともと単語が共通に持つ属性はそれほど多くないが、その他に基準として考えられる属性として、含まれる音節の種類や音節の並び方などで検証する必要もあると考えられる。

それから、推定用サンプル数と推定精度の関係をはっきりさせることも考察材料となる。今回は一人当たり 150 語を収録したので、75 語ずつの比較となったが、これを増やすことにより、より精度が高まる可能性もある。

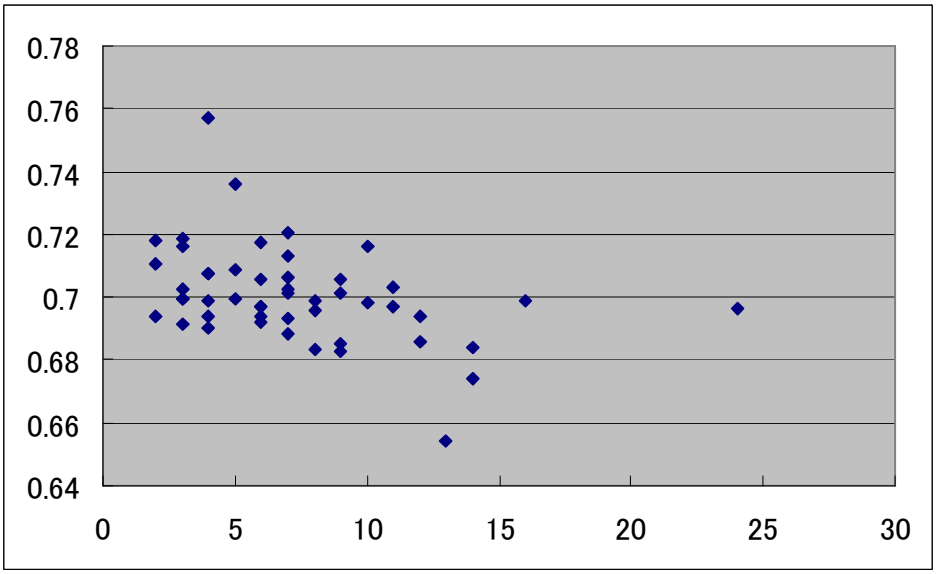

図 19　音節数と困難度の相関

他、今回の実験では、認識用辞書は、各語異数で 1 種類しか作成しておらず、語彙サンプルの違いによる、認識率の違いを考慮していない。認識する対象に伴って、認識され易い単語で構成された辞書など、サンプルの違いによる認識率の変動を考えることは意義のあることと思う。これは、更に語彙サンプルの音素数、音素構造等より深く掘り下げて考えることも可能だが実用的な汎用性を考えると統計的な検証が重要になると思われる。

他に、今回は各辞書を引用元の 170000 語のデータからランダムに選択して作成したが、そうではなく辞書が大きくなるごとに前の辞書に単語を上乗せして辞書を作成するやり方で認識させることも、大いに意味のあることと思われる。少なくとも今十実験の結果のように、変動が起きることはなく、推定もし易くなるであろう。しかし同時に、そのようにして作られるシステムの有用性についても考える必要がある。

また, Rosenberg モデルを大語彙に対して改変するという手段もある. 現在、Rosenberg モデルにおいて、標準エラーの計算には (6) 式を用いている。この式では、辞書内の単語の数が増えるに従って、本来右辺は指数的に増えていく。しかし、左辺の標準エラーは 0 から 1 までの値しか取りえないことから、右辺は単語数 N が増えるに従って、確率 pm が非常に小さくなってしまう。

これは辞書サイズが大きくなるに従って避けられない問題である、この問題を解決するために、(15)式に N を変動させ、pm を固定させる試みに取り組んでいる。辞書内単語数 N を全て扱うのではなく、誤認識された単語数 Nm だけを計算に使用する。欲しいのは誤認識率なので、ここで Nm を用いるのは不自然

ではない。こうすることで、辞書サイズの増加に伴う、確率 pm の極端な減少を防ぐことができ、大語彙化に伴って適切な推定をもたらすと考えられる.

$$WER = 1 - \sum_{m=1}^{M} h_m (1-p)^{N_m - 1}$$

他に考えるべき課題としては、語彙サイズの変更に伴う抽出単語の選出である。認識率を上げるためには一般に語彙サイズを減らさなければならないが、例えば 5 万語から 4 万語に減らすときに、5 万語のうちどの 4 万語を使用させるかという問題がある。現在の使用イメージでは例えばカーナビゲーションシステムを用いるとき、最初に 5 万語の検索地候補があるとする。そして推定の結果、例えば 90%以上の認識率を得るためには 1 万語が最適だったとする。この場合、荒い検索で先ず 1 万語以内の語彙数で認識させ、対象範囲を絞ってまた認識させる。このような対象を限定していく方法が考えられる。あるいは始めから 5 万語の対象を 1 万語ずつ 5 グループにわけ、5 回検索を行わせるような、並列的な検索方法も考えられる。このように抽出をしなくても、実用上に使えると考えられる。

また、単語をどうしても抽出しなくてはならない場合は、単純に頻度順に抜き出す等の、統計的な考えが必要になってくると考えられるので、こちらも合わせて課題としたい。

謝辞

今実験で、音声の収録に協力してくださった佐久間康裕君に感謝致します。

参考文献

［1］Masahiko Matsushita, Hiromitsu Nishizaki Takehito Utsuro and Seiichi Nakagawa “Keyword recognition and extraction for speech –driven Web retrieval task” IPSJ SIG Notes Vol.2003, No.104(20031017) pp. 21-28

［2］Sao Hwa Jeong, Sakabe Nagamasa, Sekiya Tomio, Okuyama Fumio, Gu Yeun Hwa, Koga Tatuhiko, Kawai Masakazu, Kanagawa Katumi, Lee Ki Han ”The Study of Electronic Medical Record using Speech Recognition Pocessing” ITE Technical Report Vol.25 No.29(20010321) pp.133-137

［3］NISHIZAKI Hiromitsu, NAKAGAWA Seiichi “Modeling for Performance Estimation in Spoken Document Retrieval and Simulation Experiments” IEICE technical report. Natural language understanding and model of communination Vol.102. No.527(20021212) pp.159-164

［4］A.E.Rosenberg,“A Probabilistic Model for the Performance of Word Recognizers”,AT&T BELL LABORATORIES TECHNICAL JOURNAL Vol.63,No.1,January 1984

［5］Mitsuru SAKAI, Masaki YONEDA and Hiroyuki HASE “A Theoretical Consideration in Muili-Class Pattern Recognition Problem –Derivation of an Effective Formula for the Expected Accuracy-” The Transactions of the Institute of Electronics, Information and Communication Engineers. A Vol.J79-A No.11 pp. 1907-1918

［6］Sumio OHNO, Keikichi and Hiroya FUJISAKI “Utterance Normalization Using Vowel Features for the Recognition of Spoken Words by Multiple Speakers” The Transactions of the Institute of Electronics, Information and Communication Engineers. A Vol.J76-A, No.12(19931225) pp. 1661-1667

［7］A. R. Smith and L. D. Erman, “Noah—A Bottom Up Word Hypothesizer for Large Vocabulary Speech Understanding System,” IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-3(January 1981), pp. 41-51.

［8］http://julius.sourceforge.jp

［9］Lasdon, L.S., Waren, A.D., Jain, A., and Ratner, M., "Design and Testing of a Generalized Reduced Gradient Code for Nonlinear Programming", ACM Transactions on Mathematical Software, Vol. 4, No. 1, March 1978, pp. 34-50.

［10］W.R.Blischke “Estimating the Parameter of Mixtures Of Binomial Distributions” Physical and Engineering Sciences October 1962

［11］Hiroshi Furutani, et.al, “Analysis of Programming Test by Item Response Theory”, Memories of the Faculty of Engineering, Miyazaki University, Vol.36. pp.333-338

[12] Seiichi NAKAGAWA and Yoshihisa OHGURO “Relationship between Phoneme Recognition Accuracy and Sentence Recognition Accuracy for Continuous Speech Recognition” The Transactions of the Institute of Electronics, Information and Communication Engineers. D-Ⅱ Vol.J72-D-Ⅱ No.2 pp.207-217

[13]KEINOSUKE FUKUNAGA and THOMAS E. FLICK “Classification Error for a Very Large Number of Classes” IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-6 No.6, pp. 779-788.

[14]R.Y.Kain “The Mean Accuracy of Pattern Recognizers with Many Pattern Class” IEEE Trans. on Information Theory, pp. 424-425.

付録

使用データ CD 内、ファイル一覧