認識タスクを考慮した雑音下音声認識の性能推定の研究

山　田　武　志　　筑波大学大学院システム情報工学研究科准教授

1 はじめに

現在の音声認識技術では，雑音が混入した音声を高精度に認識することは困難であり，雑音の特性や大きさ，前処理として用いる雑音抑圧アルゴリズムなどによって認識性能は大きく変動する．よって，音声認識サービスを提供する際には，サービス品質（認識性能）の保証という観点から，対象とする環境でどの程度の認識性能が得られるのかを事前に調査する必要がある．現時点で最も確実な方法は，サービスを運用する現場で認識実験を行うことである．しかし，人的，時間的コストが極めて大きく，また専門的な知識や技術を要するという問題があり，音声認識サービスの普及を妨げる一因となっている．現状の技術レベルであっても実用的な認識性能を得られる環境は数多く存在することから，認識性能を簡便に推定する技術を確立することが急務である．

従来，音声のひずみの大きさから認識性能を推定するというアプローチが提案されている[1-2]．これは，音声のひずみの大きさと認識性能の関係式（以下では推定式と呼ぶ）をあらかじめ実験的に求めておき，調査対象の雑音環境で求めた音声のひずみの大きさをその推定式に代入することにより認識性能を推定するものである．このアプローチにより，認識実験を行う場合と比べて大幅なコスト削減が実現できる．

これまでに我々は，ITU-T 勧告 P.862 [3]の PESQ を用いて認識性能を推定する手法を開発した[4,5]．本手法により，雑音や雑音抑圧アルゴリズムの種類によらず高い精度で認識性能を推定できるものの，それは認識タスク毎に最適化した推定式を用意する場合に限られていた．一般に，雑音環境や前処理が同じでも，認識タスクの難しさ，すなわち認識対象語彙数や文法的複雑さ，文の長さなどによって認識性能は変動する．このことは，認識タスクが変わった場合には，それに最適化した推定式をあらためて求める必要があることを意味する．しかし，実用上は一つの推定式で様々な認識タスクに適用できることが望まれる．

この問題を解決する方法としては，認識タスクの難しさを表すパラメータを推定式に導入することが考えられる．このような推定式を一度求めておけば，以降は認識タスクの難しさを指定することにより，任意の認識タスクに対する推定式が容易に得られることになる．本研究では，認識対象語彙数をパラメータに持つ推定式，及び文法的複雑さと文の長さをパラメータに持つ推定式を提案する．種々の認識タスク（小～中語彙の孤立単語認識や記述文法認識，大語彙の連続音声認識）の認識性能を推定する実験を行い，その有効性を示す．

2 ひずみ尺度を用いた認識性能の推定

ひずみ尺度を用いた認識性能の推定の流れを図 1 に示す．

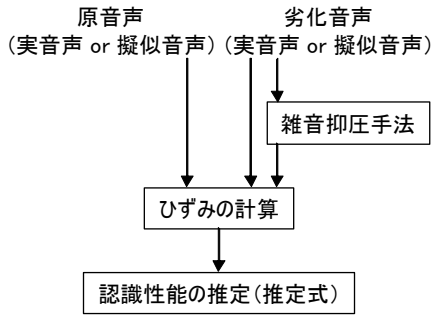


図 1. 認識性能の推定の流れ

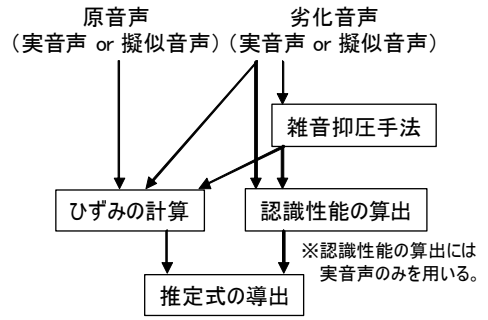


図 2. 推定式の導出の流れ

まず，原音声（雑音が重畳していない音声）と劣化音声（雑音が重畳している音声，あるいは雑音抑圧後の

音声）を入力とし，劣化音声のひずみの大きさを計算する．そして，そのひずみの大きさを推定式に代入することにより認識性能を推定する．なお，音声認識の前処理として雑音抑圧手法を用いることや，ひずみの計算の際に大量の実音声の代わりに数秒程度の擬似音声を用いることも可能である．

推定式の導出の流れを図 2 に示す．まず，劣化音声のひずみの大きさと，劣化音声に対する認識性能を求める．そして，両者の関係を最適近似する式を求め，推定式とする．これまでに我々は，推定式として次式が有効であることを明らかにした[4]．

$$y = f(x) = \frac{a}{1 + e^{-b(x-c)}} \tag{1}$$

ここで，y は推定認識性能，x はひずみの大きさである．a，b，c は定数であり，a はクリーン音声に対する認識性能，b は認識性能の低下の急峻さ，c はひずみに対する頑健性に相当する．各定数の値は，劣化音声のひずみの大きさと認識性能を実験的に求め，両者の関係を最適近似することにより決定する．

適切なひずみ尺度を用いることにより，雑音や雑音抑圧アルゴリズムの種類によらず高い精度で認識性能を推定できるものの，それは認識タスク毎に最適化した推定式を用意する場合に限られていた．実用上は一つの推定式で様々な認識タスクに適用できることが望まれることから，以下では認識タスクの難しさを表すパラメータを推定式に導入する．

3 認識対象語彙数を考慮した認識性能の推定

3-1 提案法

前述の通り，認識タスクの難しさは，認識対象語彙数や文法的複雑さ，文の長さなどによって表される．本章では，まず認識対象語彙数に着目し，孤立単語認識を対象とする推定式について述べる．雑音下孤立単語認識の性能は，認識対象語彙数が増加するにつれて低下すると考えられることから，次式に示すように，式(1)の定数を認識対象語彙数 n によって表現するように変更する．

$$y = f(x,n) = \frac{p_1 n^{q_1}}{1 + e^{-p_2 n^{q_2}\left(x - p_3 n^{q_3}\right)}} \tag{2}$$

ここで，p_1～p_3，q_1～q_3 は定数であり，様々な語彙数の孤立単語認識を対象として決定される．この推定式を一度求めておけば，以降は n を指定することにより，任意の語彙数の孤立単語認識に対する推定式が容易に得られることになる．

3-2 実験条件

音声データは，東北大ー松下単語音声データベース[6]の鉄道駅名 3285 語である．本実験では，認識対象語彙数を 50，100，200，400，800，1600，2400，3285 と変化させ，孤立単語認識を行った．なお，2400 は未知の語彙数として扱うこととし，それ以外の語彙数を対象として推定式の定数を決定する．

音響モデルとしては，IPA の「日本語ディクテーション基本ソフトウェア 1999 年度版」に収録されているモノフォン性別非依存モデル（16 混合分布）[7]を用いた．また，雑音データは，電子協騒音データベース[8]の car1，hall1，train2，lift2（以下ではテストセット A と呼ぶ），及び factory1，road2，crowd，lift1（テストセット B）である．クリーンな音声データに雑音データを計算機上で加算することにより，雑音重畳音声データを作成した．ここで，SNR は 20，15，10，5，0，-5 dB である．なお，本実験では雑音抑圧手法を用いていない．

ひずみ尺度としては ITU-T 勧告 P.862 の PESQ[3]を用いた．PESQ は人間の知覚・認知過程を考慮したひずみ尺度であり，ひずみの大きさを品質（5 が最高，1 が最低）により表すことに注意されたい．また，テストセット A を用いて推定式の係数を決定し，テストセット A，B の認識性能を各々推定した．テストセット A は雑音既知，テストセット B は雑音未知という位置付けである．

3-3 実験結果

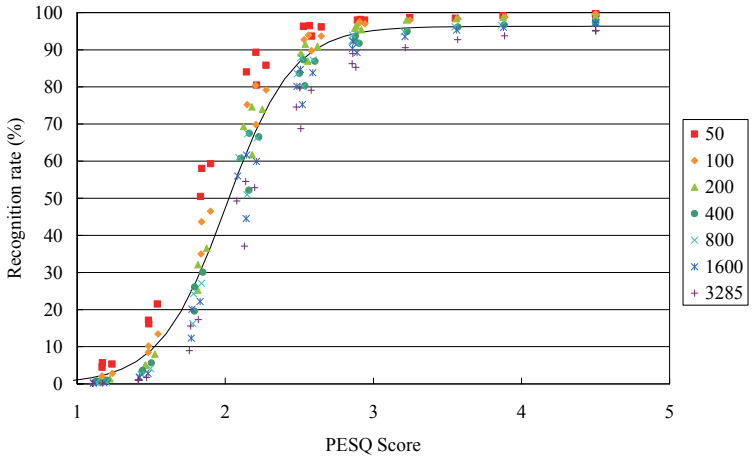
本実験では，次の 3 通りの方法で推定式を求め，各々の推定精度を比較する．

(a) 全ての認識タスクを対象とする推定式を式(1)により求める．推定式は 1 個である．

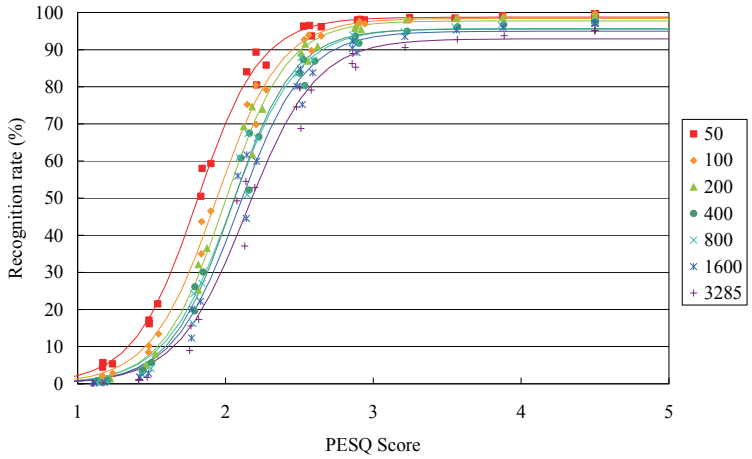
(b) 認識タスク毎の推定式を式(1)により求める．推定式は 7 個である（タスク毎に 1 個）．

(c) 認識タスク毎の推定式を式(2)により求める．推定式は 7 個である（タスク毎に 1 個）．

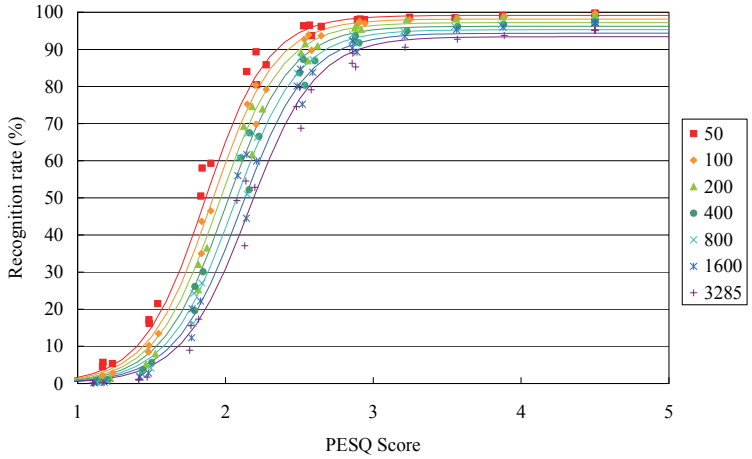
まず，単語認識率と PESQ スコアの関係を最適近似することにより求めた推定式を図 3 に示す．



(a) 式(1)で求めた全ての認識タスクを対象とする推定式



(b) 式(1)で求めた認識タスク毎の推定式



(c) 式(2)で求めた認識タスク毎の推定式

図 3. 推定式の比較

図 3(a)～(c)は，各々上記の(a)～(c)の推定式に相当する．ここで，図中の曲線は推定式であり，マーカー

はテストセットAの28種類の雑音環境の一つから得られたPESQスコアと単語認識率を表している．なお，図3(c)の推定式は，具体的には次式に語彙数 n を代入することにより得られた．

$$y = f(x,n) = \frac{104.86n^{-0.0143}}{1+e^{-5.0396n^{-0.0157}(x-1.6234n^{0.0352})}} \tag{3}$$

ここで，この推定式の係数はテストセットAを用いて最適化された（語彙数2400を除く）．図3より，(a)の推定式よりも(b)の推定式の方が近似精度が高いことが分かる．このことから，従来の式(1)の推定式を用いる場合は，認識タスク毎に最適化した推定式を用意すべきであると言える．一方，(b)の推定式と(c)の推定式を比べると大きな違いが見られない．このことは，認識対象語彙数をパラメータとする式(2)により，適切な推定式が得られていることを意味する．

次に，図3(a)～(c)の推定式を用いてテストセットA，テストセットBの単語認識率を推定した結果を図4～5に示す．ここで，図4はテストセットA（雑音既知），図5はテストセットB（雑音未知）に対する結果である．また，このときの決定係数 R^2 とRMSEを表1に示す．

表1．決定係数とRMSE

	テストセットA		テストセットB	
推定式	R^2	RMSE	R^2	RMSE
(a)	0.97	6.6	0.98	5.1
(b)	0.99	3.0	0.99	3.2
(c)	0.99	3.5	0.99	3.5

ここで，(b)と(c)の R^2 とRMSEは，語彙数毎に求めたものの平均である．なお，R^2 とRMSEは次式で定義される．

$$R^2 = 1 - \frac{\overline{(\text{真の単語認識率} - \text{推定単語認識率})^2}}{\overline{(\text{真の単語認識率} - \overline{\text{真の単語認識率}})^2}} \tag{4}$$

$$RMSE = \sqrt{\overline{(\text{真の単語認識率} - \text{推定単語認識率})^2}} \tag{5}$$

図4～5と表1から，(a)の推定式を用いた場合は他と比べて推定誤差が大きいことが分かる．RMSEで見るとその差は比較的小さいものの，大きな推定誤りを起こしている箇所が見受けられる．一方，(c)の推定式を用いた場合は，(b)の推定式を用いた場合と同等の推定精度が得られている．また，このことは雑音が未知の場合にも言える．

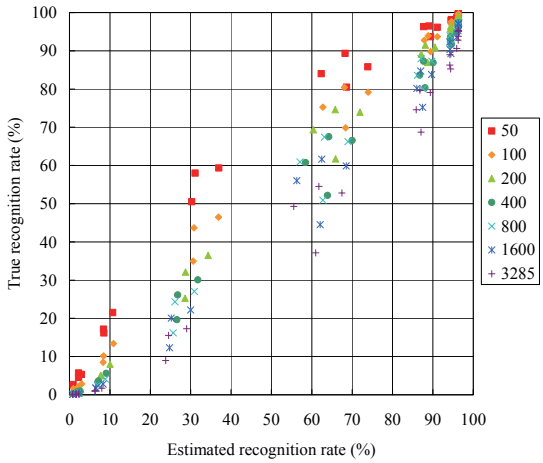
以上の実験では，推定式を求める際の語彙数と単語認識率を推定する際の語彙数は同じであった．最後に，未知の語彙数（推定式を求める際に対象としていない語彙数）に対する単語認識率を提案法により推定する．なお，(b)の推定式を求めるためには追加の認識実験などが必要となる一方，提案法では式(3)に語彙数（ここでは $n = 2400$）を代入することにより容易に推定式を導出することができる．提案法により推定したテストセットBの単語認識率を図6に示す．R^2 は0.99，RMSEは3.3であり，語彙数が未知の場合でも，既知の場合と同等の精度で単語認識率を推定できることが分かった．

以上のことから，提案法は，認識対象語彙数の違いによる孤立単語認識の性能の変動を適切に吸収できていると考えられる．

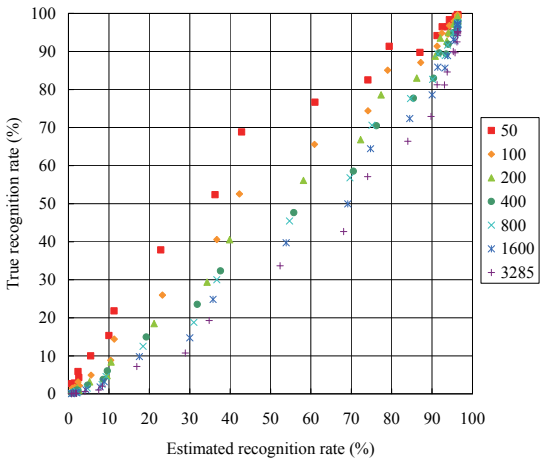
4 文法的複雑さと文の長さを考慮した認識性能の推定

4-1 提案法

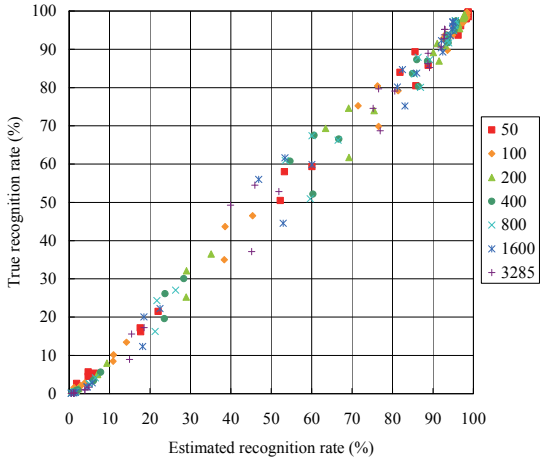
一般に，雑音下音声認識の性能は，認識時の探索空間が大きいほど低下すると考えられる．実際，3章で推定式に導入した認識対象語彙数は，孤立単語認識における探索空間の大きさに相当すると考えられる．本章では，孤立単語認識から記述文法認識や大語彙連続音声認識までの幅広い認識タスクを対象とするために，探索空間の大きさを文法的複雑さと文の長さにより表現することを考える．具体的には，文法的複雑さを平均接続可能単語数 p，文の長さを一文あたりの平均単語数 l によって表すこととし，次式のような推定式を提案する．



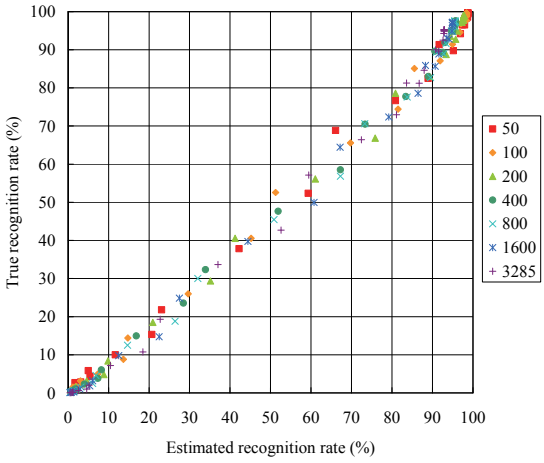
(a) 式(1)で求めた全ての認識タスクを対象とする推定式



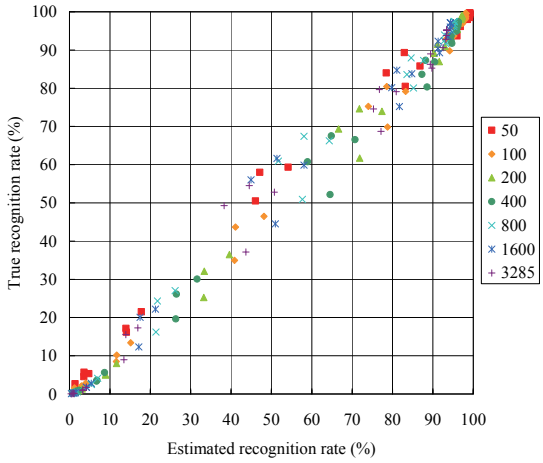
(a) 式(1)で求めた全ての認識タスクを対象とする推定式



(b) 式(1)で求めた認識タスク毎の推定式

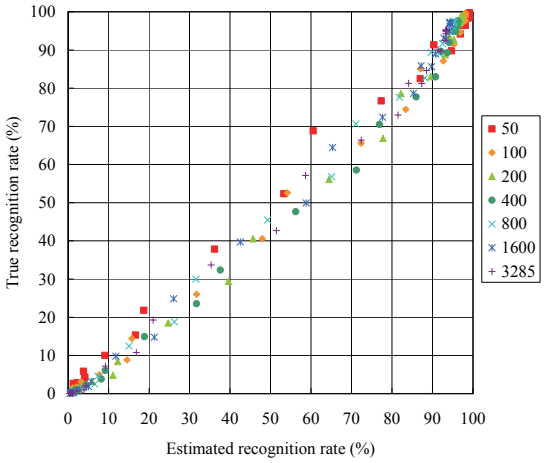


(b) 式(1)で求めた認識タスク毎の推定式



(c) 式(2)で求めた認識タスク毎の推定式

図 4. 単語認識率の推定結果（テストセット A）



(c) 式(2)で求めた認識タスク毎の推定式

図 5. 単語認識率の推定結果（テストセット B）

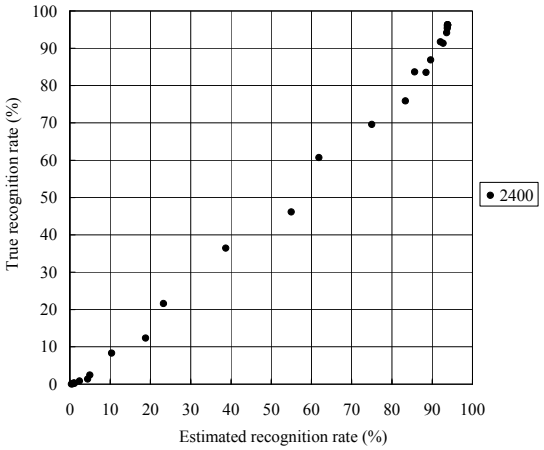


図 6. 提案法による単語認識率の推定結果（語彙数 2400，テストセット B）

$$y = f(x,n) = \frac{p_1\left(p^l\right)^{q_1}}{1+e^{-p_2\left(p^l\right)^{q_2}\left(x-p_3\left(p^l\right)^{q_3}\right)}} \tag{6}$$

これは，式(2)の n を p^l に置き換えたものである．$p = n$，$l = 1$ のときは式(2)と等価になるので，式(6)は式(2)の自然な拡張であると言える．

4-2 実験条件

本実験では，孤立単語認識，記述文法認識，大語彙連続音声認識の認識性能を推定する．各々の認識タスクの詳細は以下の通りである．

- 孤立単語認識：3.2 節と同じである．ただし，語彙数 2400 を除く．
- 記述文法認識：連続数字認識を認識タスクとする AURORA-2J[9]を用いた．発話内容は 1〜7 桁の数字列であり，これを記述文法に基づいて認識する．なお，単語（数字）の数は読みの違いを含めて 11 である．雑音は 8 種類（テストセット A とテストセット B の各々について 4 種類），SNR は 20 dB から-5 dB の 7 通りである．音響モデルは，AURORA-2J の学習用クリーン音声データで学習したものを用いた．
- 大語彙連続音声認識：音声データは，JNAS[10]のテストセット 100 文（男性話者）であり，語彙数 5000（MID）と語彙数 20000（LARGE）の 2 種類を用いた．この音声データに 3.2 節と同じ条件で雑音を重畳した．音響モデルと言語モデルとしては，IPA の「日本語ディクテーション基本ソフトウェア 1999 年度版」[7]に収録されているものを用いた．ここで，音響モデルはモノフォン性別依存モデル（16 混合分布）である．また，言語モデルは 3-gram モデルであり，語彙数 5000（5k），20000（20k），60000（60k）の 3 種類である．テストセットと言語モデルを組合せることにより 6 種類の認識タスクを設定した．

各認識タスクの平均接続可能単語数と一文あたりの平均単語数を表 2 に示しておく．なお，本実験でも雑音抑圧手法を用いておらず，ひずみ尺度として PESQ を採用した．また，テストセット A を用いて推定式の係数を決定し，テストセット A の認識性能を推定した．

4-3 実験結果

本実験では，次の 2 通りの方法で推定式を求め，各々の推定精度を比較する．

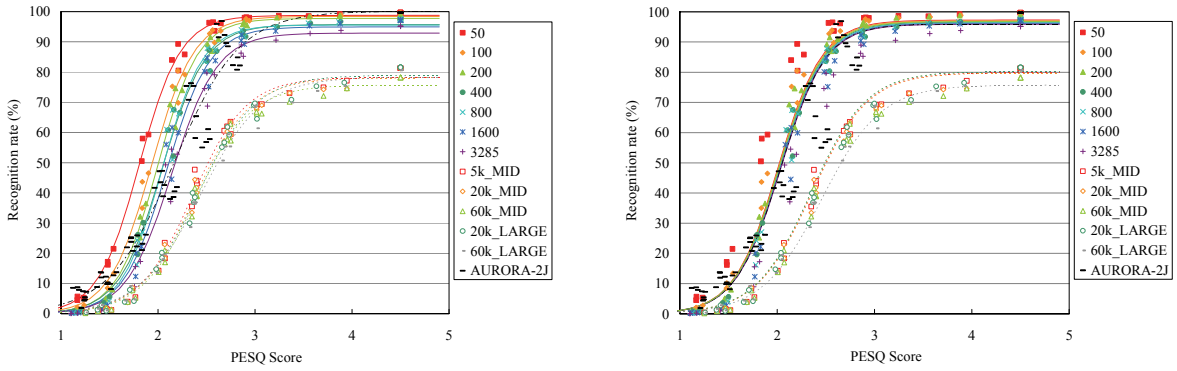
(b) 認識タスク毎の推定式を式(1)により求める．

(c) 認識タスク毎の推定式を式(6)により求める．

まず，単語認識率と PESQ スコアの関係を最適近似することにより求めた推定式を図 7 に示す．

表 2. 平均接続可能単語数と平均単語数

認識タスク		平均接続可能単語数	一文あたりの平均単語数
孤立単語認識	50 単語	50.00	1.00
孤立単語認識	100 単語	100.00	1.00
孤立単語認識	200 単語	200.00	1.00
孤立単語認識	400 単語	400.00	1.00
孤立単語認識	800 単語	800.00	1.00
孤立単語認識	1600 単語	1600.00	1.00
孤立単語認識	3285 単語	3285.00	1.00
記述文法認識	AURORA-2J	11.00	3.29
大語彙連続音声認識	5k_MID	120.25	12.59
大語彙連続音声認識	20k_MID	125.12	12.59
大語彙連続音声認識	60k_MID	109.44	12.59
大語彙連続音声認識	20k_LARGE	116.25	16.03
大語彙連続音声認識	60k_LARGE	110.87	16.03



(b) 式(1)で求めた認識タスク毎の推定式　　(c) 式(6)で求めた認識タスク毎の推定式

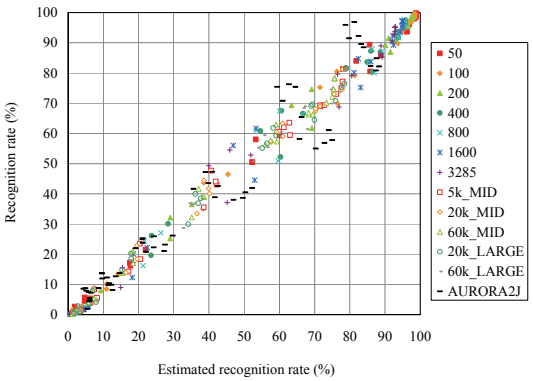
図 7. 推定式の比較

図 7(b)～(c)は，各々上記の(b)～(c)の推定式に相当する．ここで，図中の曲線は推定式であり，マーカーはテストセット A の 28 種類の雑音環境の一つから得られた PESQ スコアと単語認識率を表している．なお，図 7(c)の推定式は，具体的には次式に表 2 の p と l を代入することにより得られた．

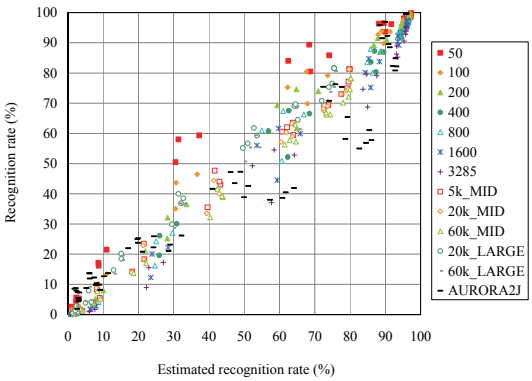
$$y = f(x,n) = \frac{98.60\left(p^{l}\right)^{-3.49\times10^{-3}}}{1+e^{-4.468\left(p^{l}\right)^{-4.04\times10^{-3}}\left(x-1.989\left(p^{l}\right)^{2.799\times10^{-3}}\right)}} \tag{7}$$

ここで，この推定式の係数は全ての認識タスクのテストセット A を用いて最適化された．図 7 からは，(c)の推定式は認識タスクの違いによる認識性能の変動を大局的には捉えているものの，局所的には近似精度があまり良くないことが分かる．特に孤立単語認識と記述文法認識に対する推定式は，区別が付き難くなっていることが見て取れる．

次に，図 7(b)～(c)の推定式を用いてテストセット A の単語認識率を推定した結果を図 8(b)～(c)に示す．また，そのときの決定係数 R^2 と RMSE を表 3 に示す．図 8 と表 3 から，(c)の推定式を用いた場合は，(b)の推定式を用いた場合と比べて推定精度が低いことが分かる．RMSE で見るとその差は比較的小さいものの，孤立単語認識と記述文法認識に対しては大きな推定誤りを起こしていることが見て取れる．本研究では，認識タスクの違いのみによって認識性能が変動することを前提としている．しかし，本実験では，各認識タスクで使用している音響モデルが違っており，その違いが予想以上に認識性能の変動に影響を及ぼしている恐れがある．提案した推定式では音響モデルの違いを吸収できないため，共通の音響モデルを用いて再度検討する必要があると考えられる．



(b) 式(1)で求めた認識タスク毎の推定式



(c) 式(6)で求めた認識タスク毎の推定式

図 8. 認識性能の推定結果

表 3. 決定係数と RMSE

推定式	R^2	RMSE
(b)	0.99	3.1
(c)	0.95	5.2

5 おわりに

これまでに我々は，ひずみ尺度を用いて認識性能を推定する手法を開発した．適切なひずみ尺度を用いることにより，雑音や雑音抑圧アルゴリズムの種類によらず高い精度で認識性能を推定できるものの，それは認識タスク毎に最適化した推定式を用意する場合に限られていた．一般に，雑音環境や前処理が同じでも，認識タスクの難しさ,すなわち認識対象語彙数や文法的複雑さ,文の長さなどによって認識性能は変動する．このことは，認識タスクが変わった場合には，それに最適化した推定式をあらためて求める必要があることを意味する．しかし，実用上は一つの推定式で様々な認識タスクに適用できることが望まれる．

本研究では，認識タスクの難しさを表すパラメータを推定式に導入することによりこの問題の解決を図った．まず，認識対象語彙数をパラメータに持つ推定式を提案し，実験により認識対象語彙数の違いによる孤立単語認識の性能の変動を適切に吸収できることを示した．次に，この結果を踏まえて，文法的複雑さと文の長さをパラメータに持つ推定式を提案した．種々の認識タスクの認識性能を推定する実験を行った結果，局所的には近似精度があまり良くないものの，認識タスクの違いによる認識性能の変動を大局的には捉えていることが分かった．今後，多種多様な認識タスクを対象として実験データを積み重ねることにより，この問題の解決を図る予定である．また，実環境において認識性能に影響を及ぼす要因としては，雑音の他にも残響や入力デバイスの音響特性，認識システムの構成などが考えられる．これらの要因を考慮するように推定式を拡張していきたい．

【参考文献】

[1] M. Kondo, K. Takeda, F. Itakura, "Predicting the degradation of speech recognition performance from sub-band dynamic ranges," 情報処理学会論文誌, Vol. 43, No. 7, pp. 2242-2248, July 2002.

[2] H. Sun, L. Shue, J. Chen, "Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP2004, Vol. 1, pp. 865-868, May 2004.

[3] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb. 2001.

[4] T. Yamada, M. Kumakura, N. Kitawaki, ″Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice,″ IEEE Transactions on Audio, Speech and Language Processing, Vol. 14, No. 6, pp. 2006-2013, Nov. 2006.

[5] 橋本倫和, 山田武志, 北脇信彦, ″雑音下音声認識の性能推定のためのひずみ尺度の検討,″ 情報処理学会研究報告, 2007-SLP-69-4, pp. 19-24, Dec. 2007.

[6] 牧野正三, 二矢田勝行, 真船裕雄, 城戸健一, ″東北大ー松下単語音声データベース,″ 日本音響学会誌, Vol. 48, No. 12, pp. 899-905, Nov. 1992.

[7] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峰松信明, 嵯峨山茂樹, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏, ″日本語ディクテーション基本ソフトウェア（99 年度版）,″ 日本音響学会誌, Vol. 57, No. 3, pp. 210-214, March 2001.

[8] 板橋秀一, ″騒音データベースと日本語共通音声データ DAT 版,″ 日本音響学会誌, Vol. 47, No. 2, pp. 951-953, Feb. 1991.

[9] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, T. Endo, ″AURORA-2J: An evaluation framework for Japanese noisy speech recognition,″ IEICE Transactions on Information and Systems, Vol. E88-D, No. 3, pp. 535-544, Mar. 2005.

[10] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi, ″JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research,″ The Journal of the Acoustical Society of Japan (E), Vol. 20, No. 3, pp. 199-206, May 1999.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
認識タスクを考慮した雑音下音声認識の性能推定の検討	日本音響学会 2008 年春季研究発表会	2008 年 3 月
認識対象語彙数を考慮した雑音下孤立単語認識の性能推定	情報処理学会研究報告 (2008-SLP-72-12)	2008 年 7 月 (発表予定)
文法的複雑さを考慮した雑音下音声認識の性能推定の検討	日本音響学会 2008 年秋季研究発表会	2008 年 9 月 (発表予定)