社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－SLP－72 (12)
2008／7／19

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

中島智弘　山田武志　北脇信彦

筑波大学大学院システム情報工学研究科
〒305–8573 茨城県つくば市天王台 1-1-1
E-mail: nakajima@mmlab.cs.tsukuba.ac.jp

概要　これまでに我々は，ITU-T 勧告 P.862 の PESQ を用いて認識性能を推定する手法を開発した．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，すなわち認識対象語彙数や文法的複雑さ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．本稿では，まず認識対象語彙数をパラメータに持つ推定式を提案する．様々な語彙数の孤立単語認識の性能を推定した結果，提案法により高い精度で単語認識率を推定できることが分かった．

Performance Estimation of Noisy Isolated Word Recognition Considering the Number of Vocabulary Words

Tomohiro Nakajima, Takeshi Yamada, Nobuhiko Kitawaki

Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305–8573 Japan
E-mail: nakajima@mmlab.cs.tsukuba.ac.jp

Abstract　It is essential to ensure QoS (Quality of Service) when offering a speech recognition service for use in noisy environments. This means that the recognition performance in the target noise environment must be investigated. Previously, we proposed a method for estimating the recognition performance from a distortion value, which represents the difference between the noisy speech in the target noise environment and its original clean version. However, there is the problem that when the recognition task is changed, an additional recognition experiment must be performed in order to obtain the estimator corresponding to it. To reduce this cost, we propose a new method considering the number of vocabulary words. Experimental results confirmed that the proposed method can give an accurate estimate for isolated word recognition with the arbitrary number of vocabulary words.

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[1] 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，すなわち認識対象語彙数や文法的複雑さ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．様々な認識タスクを対象として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．本稿では，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，様々な語彙数の孤立単語認識の性能を推定することにより，その有効性を示す．

2 提案法

認識性能の推定の流れを図 1 に示す．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以下に示す推定式に代入することにより認識性能を推定する．

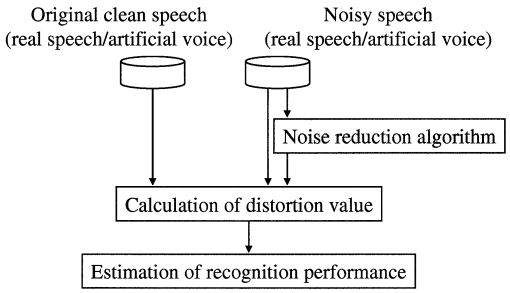


図 1: 認識性能推定の流れ

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

ここで，y は推定認識性能，x はひずみの大きさである．また，a，b，c は定数であり，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．

前述の通り，認識タスクが異なると雑音環境や雑音抑圧アルゴリズムが同じであっても認識性能は変動する．よって，認識タスク毎に最適化した推定式を求める必要がある．本稿では，この問題を解決するために，まず認識対象語彙数をパラメータとして推定式に導入する．具体的には，次式に示すように，式 (1) の定数を認識対象語彙数を用いて表すように変更する．

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

ここで，n は認識対象語彙数である．また，p_1〜p_3，q_1〜q_3 は定数であり，様々な認識タスクを対象として求めた劣化音声のひずみの大きさと認識性能の関係から決定される．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．

3 認識性能の推定実験

3.1 実験条件

音声データは，東北大－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

[1] 人間の知覚・認知過程を考慮したひずみ尺度．PESQ では，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．

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を求める．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．

本実験では，ひずみ尺度として ITU-T 勧告 P.862 の PESQ[3] を用いた．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，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．

3.2 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．

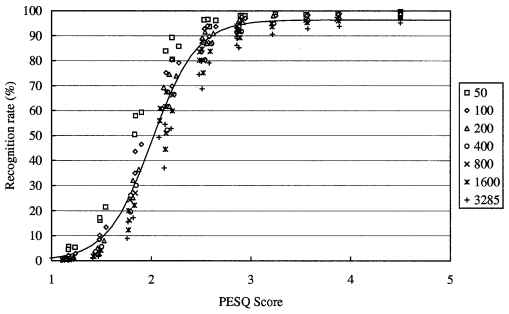
(b) 認識タスク毎の推定式を式 (1) により求める．推定式は 7 個である（タスク毎に 1 個）．

(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）．

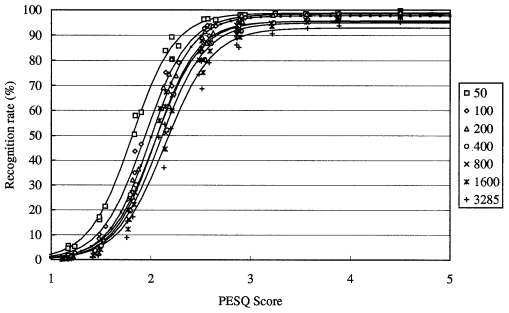
まず，単語認識率と PESQ スコアの関係を最適近似することにより求めた推定式を図 2 に示す．図 2(a)〜(c) は，各々上記の (a)〜(c) の推定式に相当する．ここで，図中の実線は推定式であり，図 2(b)(c) においては上から順に語彙数が少ないタスクに対応する．また，マーカーはテストセット A の 28 種類の雑音環境の一つから得られた PESQ スコアと単語認識率を表している．なお，図 2(c) の推定式は，具体的には次式に語彙数 n を代入することにより得られた．

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

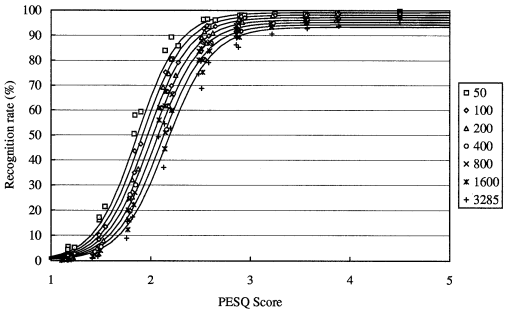
ここで，この推定式の係数はテストセット A を用いて最適化された．(a) の推定式と (b) の推定式を比べると，当然ながら後者の方が近似精度が高い．このことから，従来の式 (1) の推定式を用いる場合は，認識タスク毎に最適化した推定式を用意すべきであると言える．(b) の推定式と (c) の推定式を比べるとあまり違いが見られない．このことは，認識対象語彙数をパラメータとする式 (2) により，適切な推定式が得られていることを意味する．



(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 2: 推定式の比較

次に，図 2(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

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

図 3〜4 に示す．ここで，図 3 はテストセット A（雑音既知），図 4 はテストセット B（雑音未知）に対する結果である．また，図 2(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ときの決定係数 R^2 と RMSE を表 1 に示す．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．

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

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

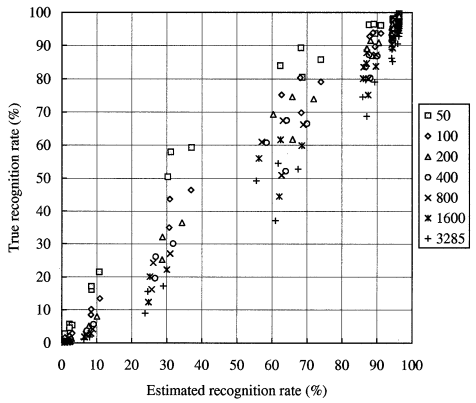
図 3〜4 と表 1 から，(a) の推定式を用いた場合は他と比べて推定誤差が大きいことが分かる．一方，(c) の推定式を用いた場合は，(b) の推定式を用いた場合と同等の推定精度が得られている．また，このことは雑音が未知の場合にも言える．

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 5 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．

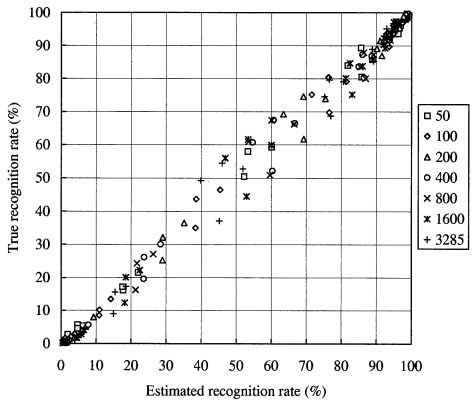
以上のことから，提案法は，認識対象語彙数の違いによる認識性能の変動を適切に吸収できていると考えられる．

4 おわりに

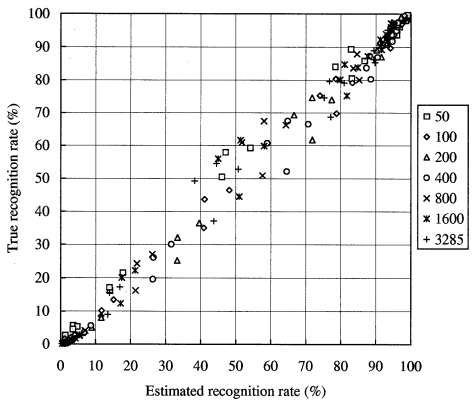
本稿では，音声のひずみの大きさと認識対象語彙数から，雑音下孤立単語認識の性能を語彙数によら



(a) 式 (1) で求めた全ての認識タスクを対象とする推定式

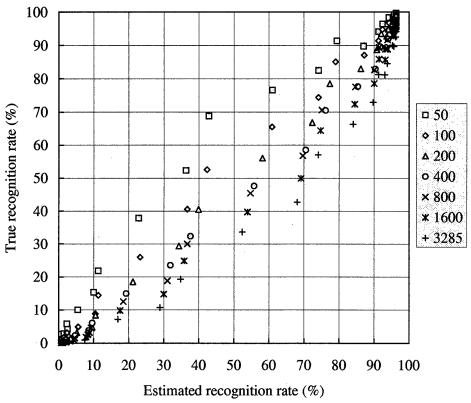


(b) 式 (1) で求めた認識タスク毎の推定式

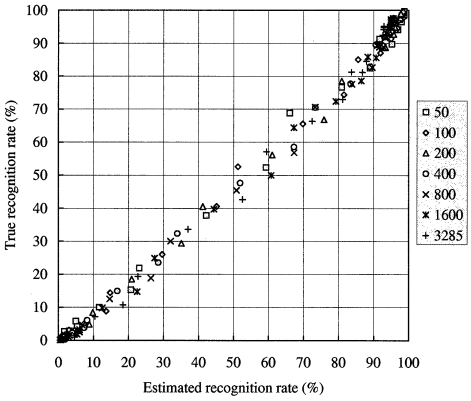


(c) 式 (2) で求めた認識タスク毎の推定式

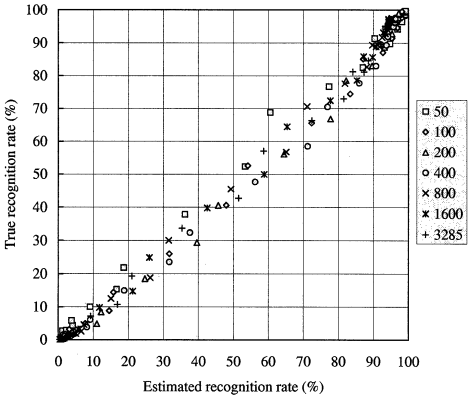
図 3: 単語認識率の推定結果（テストセット A）



(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 4: 単語認識率の推定結果（テストセット B）

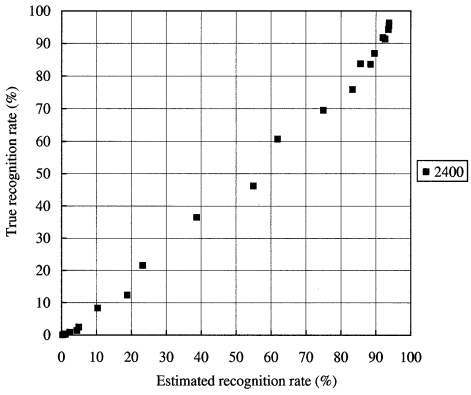


図 5: 提案法による単語認識率の推定結果（語彙数 2400，テストセット B）

ず高い精度で推定できることを示した．今後は，認識対象語彙数に加えて，文法的複雑さをパラメータとして推定式に導入する予定である．また，認識性能を変動させる他の要因（乗法性雑音や認識システムの構成など）についても検討する．

謝辞

本研究の一部は，財団法人電気通信普及財団の研究助成による．

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－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.