雑音下音声認識の性能推定法の実環境における評価*

☆中島智弘，山田武志，北脇信彦，牧野昭二（筑波大）

1　はじめに

一般に音声認識の性能は雑音の混入によって大きく変動するため，音声認識サービスを提供する際には，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ったことが問題となっている．現状の音声認識技術であってもサービスの運用に耐え得る雑音環境は少なからず存在することから，この問題を解決することにより音声認識サービスのさらなる普及が期待できる．よって，雑音下音声認識の性能を簡便に推定する手法が必要不可欠である．

我々はこれまでに，音声のひずみの大きさから認識性能を推定する手法を開発し，その有効性を示した [1]．本手法では，ひずみの大きさと認識性能の関係を表す推定式を実験的に求める．しかし，認識タスク（例えば語彙サイズや文法的複雑さ）が異なると同じ雑音環境であっても認識性能は変動するため，個々の認識タスクに最適化した推定式を用意しなければならなかった．我々はこの問題に対処するために，認識タスクの複雑さをパラメータとして持つ推定式を提案した [2, 3, 4]．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複雑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を容易に得ることができる．本稿では，提案法の有効性を検証するために，雑音や残響が存在する実環境において認識性能の推定実験を行う．

2　雑音下音声認識の性能推定法

認識性能推定の基本的な流れを Fig. 1 を用いて説明する．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 [1]．

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

ここで，y は認識性能の推定値，x はひずみの大きさである．また，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 はひずみが大きくなったときの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．これらの値は，劣化音声に対する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．ただし，上述したように認識タスクが異なると同じ雑音環境であっても認識性能が変動するため，個々の認識タスクに最適化した推定式を求める必要がある．

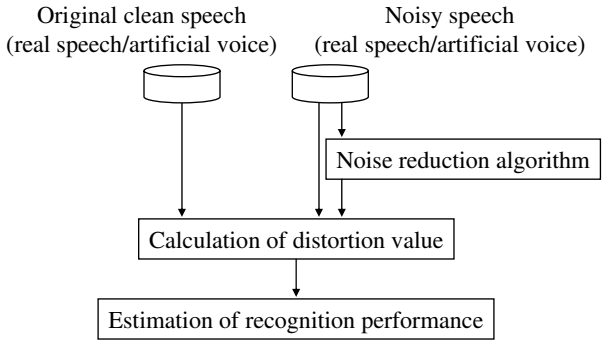

Fig. 1　Estimation of the recognition performance from the distortion value.

我々はこの問題に対処するために，認識タスクの複雑さをパラメータとして持つ推定式を提案した [4]．式 (1) の a，b，c は認識タスクの違いを反映するので，タスクの複雑さをパラメータとする関数によって各定数を表すことができると考えられる．よって，次式のように，式 (1) の定数をタスクの複雑さ α によって表すように変更した．

$$\begin{aligned} y &= f(x, \alpha) \\ &= \frac{a(\alpha)}{1 + e^{-b(\alpha)(x - c(\alpha))}} \end{aligned} \tag{2}$$

タスクの複雑さを表す尺度としては様々なものが考えられるが，これまでの検討により SMR パープレキシティ [5]（各単語の出現確率の逆数を単語パープレキシティとしたときの相加平均）を次式の推定式と共に用いる場合に優れた推定性能を得ることができると分かったので [4]，本稿でもこれを採用することにする．

$$\begin{aligned} y &= f(x, \alpha) \\ &= \frac{p_1 \alpha^{q_1} + r_1}{1 + e^{-(p_2 \alpha^{q_2} + r_2)(x - (p_3 \alpha^{q_3} + r_3))}} \end{aligned} \tag{3}$$

ここで，α は SMR パープレキシティ（以下では p_{SMR} と称す）である．p_1〜p_3，q_1〜q_3，r_1〜r_3 は定数であり，様々な認識タスクを対象として劣化音声に対する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．

*Evaluation of performance estimation of noisy speech recognition in a real environment, by Tomohiro NAKAJIMA, Takeshi YAMADA, Nobuhiko KITAWAKI, and Shoji MAKINO (University of Tsukuba).

3　実環境における有効性の検証

本章では，様々な認識タスクを対象とし，提案法による認識性能の推定精度を検証する．ここで，推定式の最適化には，計算機上で加法性雑音のみを重畳した音声データを用いる．一方，認識性能の推定には，実環境において収録した音声データを用いる．推定性能を高めるためには，推定式の最適化のために実環境において収録した音声データを用いることが望ましいと考えられるが，収録には膨大なコストがかかることを考慮し，本実験では現実的な状況を設定することにした．なお，本実験ではひずみ尺度として PESQ[6] を用いており，雑音抑圧手法は適用していない．

以下，3.1 節では推定式の最適化について述べ，3.2 節では実環境における音声データの収録について説明する．最後に 3.3 節では実験結果について述べる．

3.1　推定式の最適化

まず，推定式の最適化のために用いた認識タスクと音声データについて述べる．

- 孤立単語認識タスク：音声データとして，東北大ー松下単語音声データベース [7] に収録されているものを用いた．これは鉄道駅名の 3,285 語を読み上げたものである．本実験では，語彙サイズを 50，100，200，400，800，1600，3285 とすることにより，7 種類の認識タスクを設定した．
- 記述文法認識タスク：音声データとして，AURORA-2J [10] と全く同じ 1〜7 桁の数字列を読み上げたものを用いた．ただし，AURORA-2J とは異なり，サンプリング周波数は 16kHz である．単語（数字）の数は読みの違いを含めて 11 であり，これらを任意回数繰り返すという記述文法を作成した．認識タスクは 1 種類である．
- 大語彙連続認識タスク：音声データは，新聞記事読み上げ音声コーパス（JNAS）のテストセット 100 文（男性話者）である．語彙サイズ 5k（MID）と語彙サイズ 20k（LARGE）の 2 種類を用いた．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」[8] に含まれている 3-gram モデルのうち，語彙サイズ 5k，20k，60k の 3 種類を用いた．テストセットと言語モデルを組合せることにより 5 種類の認識タスクを設定した．

各認識タスクの SMR パープレキシティ p_{SMR} を Table 1 に示しておく．

各認識タスクの音声データには，電子協騒音データベース [9] の雑音（car1，hall1，train2，lift2）を 20，15，10，5，0，−5dB の SNR で重畳した．これらの雑音重畳音声データを認識するための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」に含まれているモノフォン性別非依存モデル（16 混合分布）を用いた．

Table 1　p_{SMR} for each recognition task.

認識タスク		p_{SMR}
孤立単語	50 words	50
	100 words	100
	200 words	200
	400 words	400
	800 words	800
	1,600 words	1,600
	3,285 words	3,285
記述文法	Connected digits	11
大語彙連続	5k_MID	40,588
	20k_MID	44,073
	60k_MID	33,381
	20k_LARGE	33,976
	60k_LARGE	57,424

上記の 13 種類の認識タスクに対し，SMR パープレキシティ，単語正解精度，ひずみの大きさを求め，これらの関係を最適近似するように式 (3) の係数を求めた結果，次式が得られた．

$$\begin{aligned} y &= f(x, p_{SMR}) \\ &= \frac{a(p_{SMR})}{1 + e^{-b(p_{SMR})(x - c(p_{SMR}))}} \end{aligned} \tag{4}$$

ただし，

$$\begin{aligned} a(p_{SMR}) &= -0.335(p_{SMR})^{0.401} + 99.31 \\ b(p_{SMR}) &= -1.66 \times 10^{-8}\ (p_{SMR})^{1.67} + 4.44 \\ c(p_{SMR}) &= -13.42\ (p_{SMR})^{-0.00518} + 15.15 \end{aligned}$$

式 (4) の x と p_{SMR} に値を代入することにより，単語正解精度の推定値を得ることができる．また，式 (4) の p_{SMR} にのみ値を代入することにより，対象とする認識タスクの推定式（式 (1) に相当）を得ることもできる．

3.2　実環境における音声データの収録

Fig. 2 の設定のもとで音声データを収録した．実験室のサイズは 7.2m × 7.7m であり，内部には防音室，机，棚などが設置されている．なお，壁面はコンクリートである．スピーカとしては，クリーン音声再生用と雑音再生用の 2 台を用いた．音声用のスピーカはマイクロホンの方に向け，雑音用のスピーカは壁面に向けている．マイクロホンとしては，近接マイク（スピーカからの距離 10cm）と遠隔マイク（スピーカからの距離 110cm）の 2 本を用いた．以上の設定のもと，音声と雑音をスピーカから同時に再生することにより，雑音重畳音声データを収録した．なお，

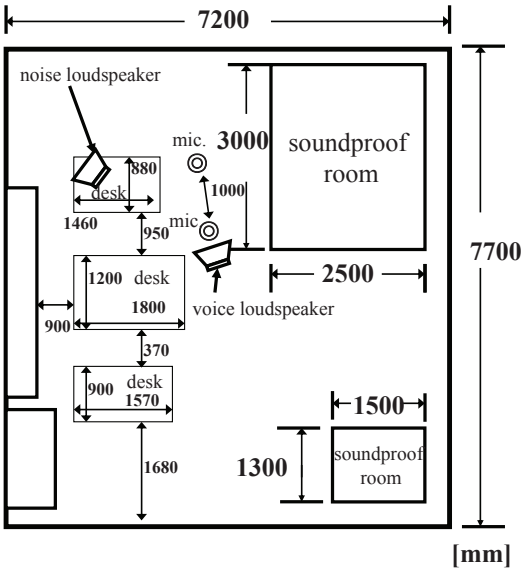

Fig. 2　The recording setup.

雑音は hall1 と factory1 の 2 種類であり，雑音の音量は 3 段階に変化させた．

収録の対象とした認識タスクと音声データは次の通りである．

- 孤立単語認識タスク：3.1 節の認識タスクのうち，語彙サイズ 800 と語彙サイズ 1,600 に対応する音声データを収録した．収録データの SNR の推定値は，近接マイクの場合は 20，15，10dB，遠隔マイクの場合は 10，5，0dB であった．
- 記述文法認識タスク：3.1 節の全ての認識タスクに対応する音声データを収録した．収録データの SNR の推定値は，近接マイクの場合は 20，15，10dB，遠隔マイクの場合は 10，5，0dB であった．
- 大語彙連続認識タスク：3.1 節の全ての認識タスクに対応する音声データを収録した．収録データの SNR の推定値は，近接マイクの場合は 15，13，10dB，遠隔マイクの場合は 5，0，−5dB であった．

以上の収録データに対して，3.1 節と全く同じ条件で単語正解精度とひずみの大きさを求めた．

3.3　実験結果

3.1 節で述べた 13 種類の認識タスクの各々に対する推定式を Fig. 3 に示す．ここで，横軸は PESQ の値であり，値が小さいほど音声が劣化していることを表す（値が 5 のときはクリーン音声に一致する）．また，図中の曲線は各認識タスクに対する推定式を表しており，式 (4) の p_{SMR} に Table 1 の値を代入することにより求めた．Fig. 3 と Table 1 より，各推定式の上下の位置関係は概ね p_{SMR} の値によって定まっていることが分かる．

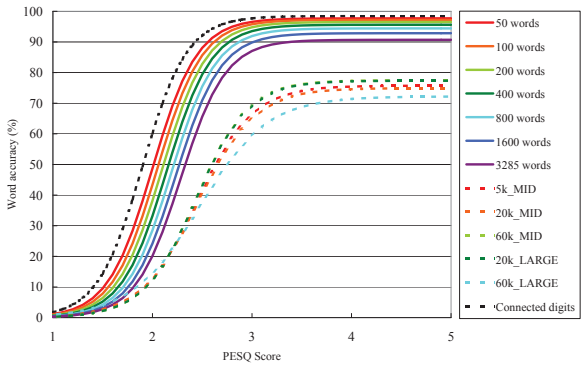

Fig. 3　The estimators for each recognition task.

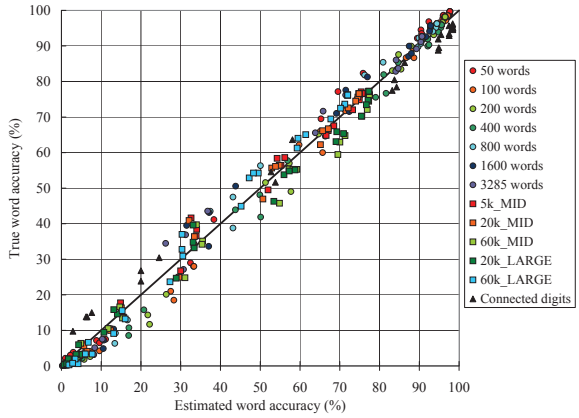

Fig. 4　The relationship between true word accuracy and word accuracy estimated by using simulation data.

次に，3.1 節で述べた 13 種類の認識タスクの各々に対して単語正解精度の推定を行った．推定のために用いた雑音重畳音声データは推定式の最適化のために用いたものと同じであり，残響などの影響を全く受けていないため，このときの推定精度は提案法の上限とみなすことができる．Fig. 4 に真の単語正解精度と推定した単語正解精度の関係を示す．真の単語正解精度と推定した単語正解精度の決定係数は 0.99，RMSE は 3.4 であり，高い精度で推定できていることが分かる．

最後に，3.2 節で述べた 8 種類の認識タスクの各々に対して単語正解精度の推定を行った．推定のために用いた雑音重畳音声データは実環境において収録したものであり，推定式の最適化のために用いたものとは残響などの影響を受けているという点で異なる．Fig. 5 に真の単語正解精度と推定した単語正解精度の関係を示す．真の単語正解精度と推定した単語正解精度の決定係数は 0.98，RMSE は 3.2 であり，残響などの影響を全く受けていないデータを用いて推定したときと同等の精度で推定できていることが確認できる．今回は特定の音響環境でしか評価していない

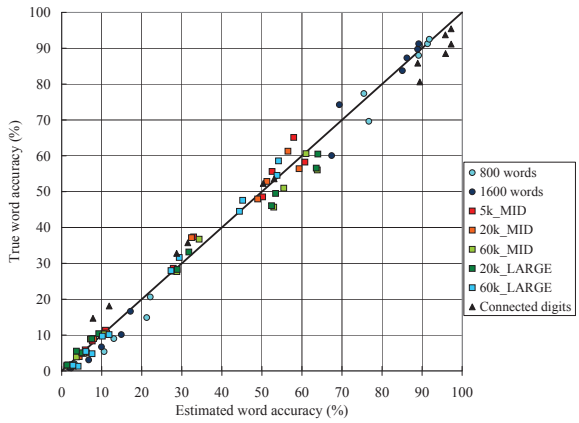

Fig. 5 The relationship between true word accuracy and word accuracy estimated by using real data.

ものの，推定式の最適化のために用いる音声データは計算機上で作成したもので十分であるとの見込みを得た．

4 おわりに

本稿では，提案法の有効性を検証するために，雑音や残響が存在する実環境において認識性能の推定実験を行った．その結果，実環境で収録した雑音重畳音声データを用いても高い精度で認識性能を推定できることが分かった．また，推定式の最適化のために用いる音声データは計算機上で作成したもので十分であるとの見込みを得た．今後は，多様な実環境や未知の認識タスクに対する提案法の有効性を検証していきたい．

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] 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.

[2] 中島智弘, 山田武志, 北脇信彦, “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, pp. 63–68, July 2008.

[3] 中島智弘，山田武志，北脇信彦，”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,” 日本音響学会秋季研究発表会, pp. 167–168, Sept. 2008.

[4] 中島智弘, 山田武志, 北脇信彦, ”雑音下音声認識の性能推定に用いるタスクの複雑さを表す尺度の検討,” 日本音響学会春季研究発表会, pp. 147–150, Mar. 2009.

[5] 中川聖一，伊田 政樹，“連続音声認識のタスクの複雑さを表す新しい尺度,” 電子情報通信学会論文誌, Vol. J81-D-2, No. 7,pp. 1491–1500, July 1998.

[6] 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.

[7] S. Makino, N. Niyada, Y. Mafune, K. Kido, “Tohoku University and Matsushita isolated spoken word database,” Journal of the Acoustical Society of Japan, Vol. 48, No. 12, pp. 899–905, 1992.

[8] 河原達也 *et al.*, “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, Mar. 2001.

[9] 電子協騒音データベース，http://research.nii.ac.jp/src/list/detail.html#JEIDA-NOISE.

[10] S. Nakamura *et al.*, “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.