複数のひずみ尺度を用いた雑音下音声認識の性能推定の検討*

☆ Ling GUO，山田武志，牧野昭二，北脇信彦（筑波大）

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正しく認識することは困難である．音声認識の前処理として雑音抑圧を行うことにより，認識性能をある程度改善することができるが，雑音の特性や雑音抑圧アルゴリズムの種類によって性能改善の度合いは異なる．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，対象とする雑音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，音声認識サービスを運用する現場で，あるいはそこで収録した音声データを用いて，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．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．

この問題を解決するためには，音声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有効であると考えられる [1, 2, 3]．これは，音声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，対象とする雑音環境で求めた音声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．これまでに我々は，ひずみ尺度として ITU-T 勧告 P.862 の PESQ（Perceptual Evaluation of Speech Quality）[4] を用いることにより，認識性能を高い精度で推定できることを示した [3]．しかし，個々の雑音抑圧アルゴリズムに最適化した推定式を用いる必要があり，そのためのコストが問題となっている．これは，雑音抑圧アルゴリズムの内部パラメータを変更した場合や，新しい雑音抑圧アルゴリズムを開発した場合には，それに最適化した推定式をその都度求める必要があることを意味している．

本稿では，この問題を解決するために，音声ひずみと出力 SNR（Signal to Noise Ratio）を併用した認識性能推定法を提案する．ここで，出力 SNR とは，雑音抑圧アルゴリズムの出力音声（雑音抑圧後の音声）の SNR である．提案法における推定式は，変更を加えることなく様々な雑音抑圧アルゴリズムに適用可能である．認識性能推定の実験を行うことにより，提案法の有効性を検証する．

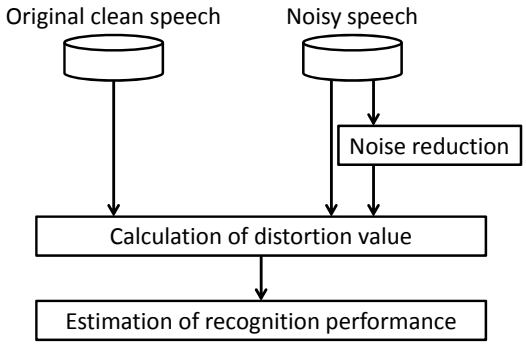


Fig. 1: Estimation of the recognition performance from the distortion value.

2 提案法

音声ひずみの大きさから認識性能を推定する処理の流れを Fig. 1 に示す．まず，原音声（雑音が重畳していない音声）と認識対象の劣化音声（雑音が重畳している音声や雑音抑圧後の音声）から，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から認識性能を推定する．従来法ではひずみ尺度として PESQ を用いており，その場合の推定式は次式で表される [3]．

$$y = \frac{a}{1 + e^{-bx+c}} \tag{1}$$

ここで，y は単語正解精度，x は PESQ スコアである．PESQ スコアは主観品質スコアである MOS（Mean Opinion Score）に対応しており，-0.5（最も悪い）から 4.5（最も良い）までの実数値をとる．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，様々な雑音条件のもとで推定誤差を最小にするように最適化される．

単語正解精度と PESQ スコアの関係の例を Fig. 2 に示す．ここで，マーカの種類は雑音抑圧アルゴリズムの違いを表す．個々の点は，雑音（4 種類）と入力 SNR（7 通り）の組に対応する．実験条件の詳細は 3.1 節で述べる．Fig. 2 より，雑音抑圧アルゴリズム (G) が他とは顕著に異なる傾向を示しており，認

*Performance estimation of noisy speech recognition using spectral distortion and SNR of Noise-reduced speech, by Ling Guo, Takeshi Yamada, Shoji Makino, Nobuhiko Kitawaki (University of Tsukuba).

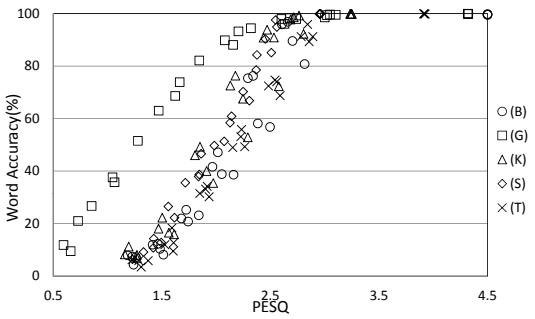


Fig. 2: Relationship between the word accuracy and the PESQ score.

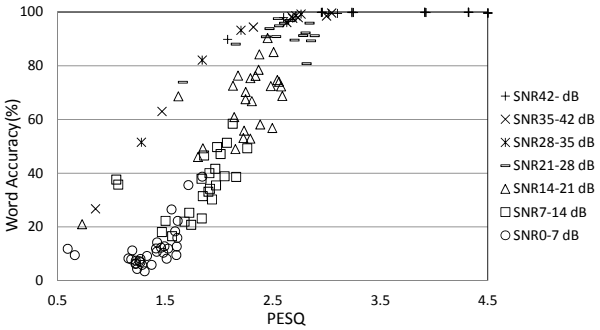


Fig. 3: Relationship between the word accuracy, the PESQ score, and the output SNR.

識性能を高精度に推定するためには，個々の雑音抑圧アルゴリズムに最適化した推定式が必要となることが確認できる．

本稿では，この問題を解決するために，音声ひずみと出力 SNR を併用した認識性能推定法を提案する．単語正解精度，PESQ スコア，出力 SNR の関係を Fig. 3 に示す．ここで，本稿では出力 SNR を次式により近似計算している．

$$\text{Output SNR} \quad = \quad 10\log_{10}\frac{P_x}{P_n} \qquad (2)$$

P_x は音声区間の時間平均パワー，P_n は非音声区間の時間平均パワーである．音声区間検出はクリーン音声を用いて行う．Fig. 3 は Fig. 2 におけるマーカの種類を出力 SNR の違いに変更したものであり，PESQ スコアが同じときには SNR が高いほど単語正解精度が高いことが分かる．これは，音声ひずみと出力 SNR を併用することにより，単語正解精度の変動を説明できることを示唆している．

以上より，提案法では推定式として次式を用いることとする．

$$y \quad = \quad \frac{a}{1+e^{-b_1x_1-b_2x_2+c}} \qquad (3)$$

ここで，y は単語正解精度，x_1 は PESQ スコア，x_2 は出力 SNR である．また，a，b_1，b_2，c は定数であり，様々な雑音条件・雑音抑圧アルゴリズムのもとで推定誤差を最小にするように最適化される．

3 提案法の有効性の検証

3.1 実験条件

認識実験には，雑音環境下連続数字認識タスク AURORA-2J[8] を用いる．本実験で用いた音声データを Table 1 に示す．音声認識に用いる音響モデルの学習データには，雑音が重畳されていない学習データである Clean Training を用いる．Clean Training は，成人男性 55 名，成人女性 55 名による計 8,440 発話を，伝送路を模擬したフィルタである G.712 に通したものである．なお，学習データに対しても認識時と同様の雑音抑圧を行い，各雑音抑圧アルゴリズムに専用の音響モデルを作成する．認識対象のテストデータには，Test set A と Test set B の 2 種類を用いる．雑音は，Subway（地下鉄），Babble（人混み），Car（自動車），Exhibition（展示場），Restaurant（飲食店），Street（街頭），Airport（空港），Station（駅構内）の 8 種類，SNR は Clean から −5dB までの 7 通りである．本実験では，以下に示すように，雑音抑圧を行わない場合と 4 種類の雑音抑圧アルゴリズムを用いる場合を考える．

(B) ベースライン（雑音抑圧を行わない場合）

(G) GMM に基づく音声信号推定法 [5]

(K) ピッチ同期 KLT 法 [6]

(S) SS-SMT 法（スペクトルサブトラクション法）[7]

(T) 時間領域 SVD に基づく音声強調法 [5]

これら 4 種類の雑音抑圧アルゴリズムは，非音声区間で推定した雑音を雑音重畳音声から減算するという点では共通であるものの，雑音の推定方法や減算方法が異なっている．

認識性能推定の実験条件は以下の通りである．

(C1) (B) を含む計 5 種類の雑音抑圧アルゴリズムを Test set A に適用した後，単語正解精度，PESQ スコア，出力 SNR を算出し，式 (1) と式 (3) の推定式の係数を最適化する．そして，5 種類の雑音抑圧アルゴリズムを Test set A に適用したときの単語正解精度を推定する．

(C2) 推定式の係数は (C1) と同じとし，5 種類の雑音抑圧アルゴリズムを Test set B に適用したときの単語正解精度を推定する．これは雑音に対してオープンなテストである．

Table 1: Speech data used in the experiment.

Training and Test sets	Speech	Noise	Channel	SNR (dB)
Clean training	8,440 utterances of 110 people	None	G.712	Clean
Test set A	4,004 utterances of 104 people	Subway, Babble, Car, Exhibition		Clean, 20, 15, 10, 5, 0, −5
Test set B		Restaurant, Street, Airport, Station		

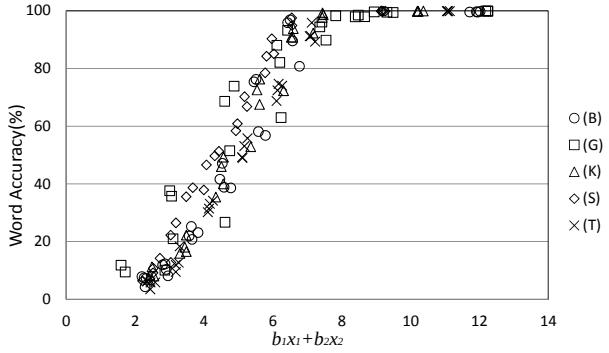


Fig. 4: Relationship between the word accuracy and the distortion calculated by $b_1x_1 + b_2x_2$ in the proposed method.

条件 (C1) のもとで最適化した従来法の推定式を式 (4)，提案法の推定式を式 (5) にそれぞれ示しておく．

$$y = \frac{102.7779}{1 + e^{-2.807325x + 5.618707}} \quad (4)$$

$$y = \frac{100.8289}{1 + e^{-0.088535x_1 - 1.738907x_2 + 4.737119}} \quad (5)$$

3.2 実験結果

まず，提案法におけるひずみ相当量（$b_1x_1 + b_2x_2$）と単語正解精度の関係を Fig. 4 に示す．Fig. 4（提案法）と Fig. 2（従来法）を比較すると，提案法の方が雑音抑圧アルゴリズムの違いに頑健であることが分かる．次に，提案法と従来法により認識性能を推定した結果を Fig. 5 と Fig. 6 に示す．ここで，Fig. 5 は条件 (C1) の場合，Fig. 6 は条件 (C2) の場合である．各図の横軸は真の単語正解精度，縦軸は推定した単語正解精度である．また，真の単語正解精度と推定した単語正解精度の決定係数 R^2 と RMSE（Root Mean Square Error）を Table 2 に示す．Fig. 5，Fig. 6，Table 2 より，提案法の方が従来法よりも高精度に単語正解精度を推定できていることが確認できる．特に雑音抑圧アルゴリズム (G) に対する推定精度には明確な差が見て取れる．また，条件 (C1) と条件 (C2) を比較することにより，雑音の種類にさほど影響を受けずに推定できていることが分かる．以上より提案法の有効性が示された．

Table 2: R^2 and RMSE.

	(C1)		(C2)	
	R^2	RMSE	R^2	RMSE
Conventional	0.86	13.2	0.85	14.0
Proposed	0.96	7.0	0.96	7.37

4 おわりに

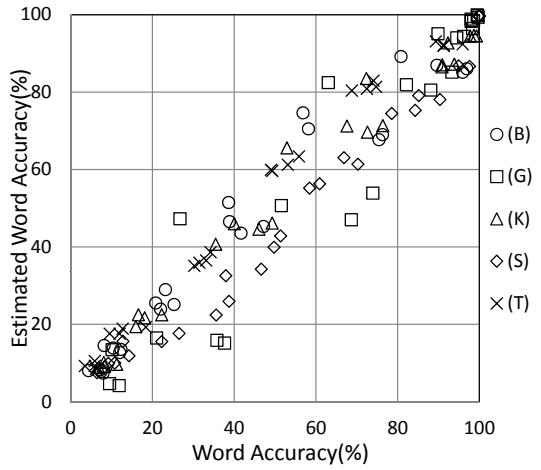
本稿では，音声ひずみと出力 SNR を併用した認識性能推定法を提案した．そして，提案法は従来法よりも高精度に認識性能を推定できること，及び提案法における推定式は変更を加えることなく様々な雑音抑圧アルゴリズムに適用可能であることを示した．今後は，推定式の係数の最適化に用いていない雑音抑圧アルゴリズムに対する有効性を検証する予定である．

参考文献

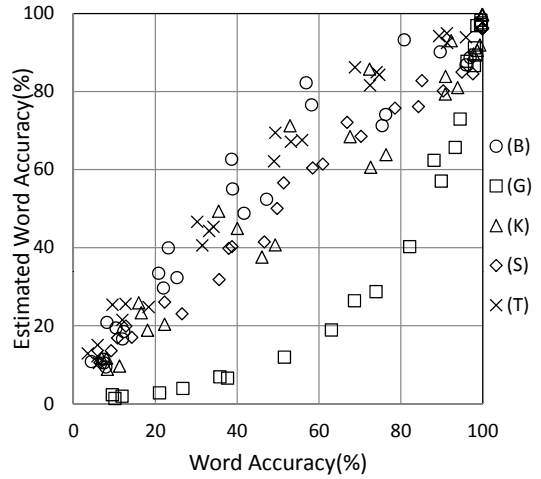
[1] H. Sun, L. Shue, J. Chen, “Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech,” Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP2004, Vol. 1, pp. 865–868, May 2004.

[2] M. Kondo, K. Takeda, F. Itakura, “Predicting the degradation of speech recognition performance from sub-band dynamic ranges,” IPSJ Journal, Vol. 43, No. 7, pp. 2242–2248, July 2002.

[3] T. Yamada, M. Kumakura, N. Kitawaki, “Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice,” IEEE Trans-

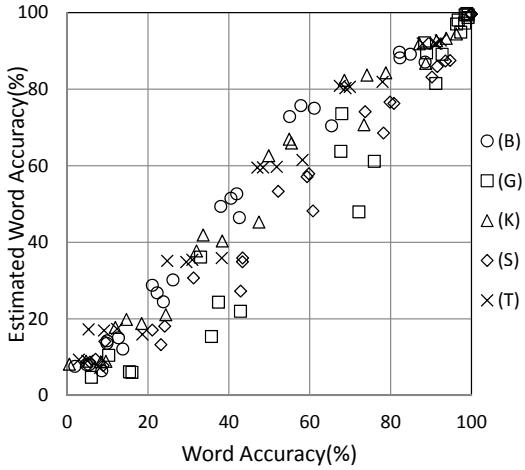


(a) Proposed method

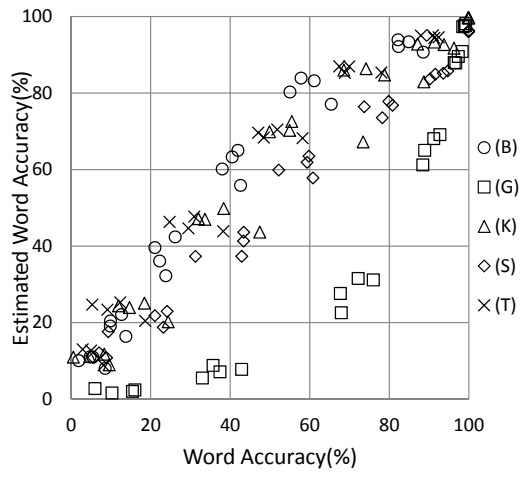


(b) Conventional method

Fig. 5: Relationship between the true word accuracy and the estimated word accuracy in the condition (C1).



(a) Proposed method



(b) Conventional method

Fig. 6: Relationship between the true word accuracy and the estimated word accuracy in the condition (C2).

actions on Audio, Speech and Language Processing, Vol. 14, No. 6, pp. 2006–2013, Nov. 2006.

[4] 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.

[5] M. Fujimoto, Y. Ariki, "Combination of temporal domain SVD based speech enhancement and GMM based speech estimation for ASR in noise – evaluation on the AURORA2 task –," Proc. European Conference on Speech Communication and Technology, EUROSPEECH2003, pp. 1781–1784, 2003.

[6] S.-J. Park, M. Ikeda, K. Takeda, F. Itakura, "Improvement of the ASR robustness using combinations of spectral subtraction and KLT based adaptive comb-filtering," IPSJ SIGNotes, SLP-44-3, pp. 13–18, 2002.

[7] N. Kitaoka, S. Nakagawa, "Evaluation of spectral subtraction with smoothing of time direction on the AURORA 2 task," Proc. International Conference on Spoken Language Processing, ICSLP2002, pp. 465–468, 2002.

[8] 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.