種々の雑音抑圧手法と認識タスクに適用可能な音声認識性能推定法の検討*

☆ Ling GUO，山田武志，宮部滋樹，牧野昭二，北脇信彦（筑波大）

1 はじめに

雑音環境における音声認識性能は，雑音の特性や雑音抑圧アルゴリズムの種類，音声認識システムの構成によって大きく変動する．しかし，その変動を容易に推定する手段は存在しておらず，音声認識を実利用する上での足かせとなっている．よって，雑音環境における認識性能を簡単かつ短時間に推定する手法が必要不可欠である．

我々はこれまでに，音声ひずみの大きさから認識性能を推定する手法を提案し，その有効性を示した [1]．さらに，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という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]．しかし，この推定式は個々の雑音抑圧アルゴリズムに最適化する必要があった．

本稿では，種々の雑音抑圧手法と認識タスクに適用可能な認識性能推定法を提案する．提案手法では，音声ひずみと出力 SNR（雑音抑圧後の音声の SNR）から認識性能を推定する推定式 [3] に，認識タスクの複雑さをパラメータとして導入する．認識性能を推定する実験を行うことにより，提案手法の有効性を検証する．

2 提案手法

我々は，雑音抑圧アルゴリズムの違いに頑健な認識性能推定を行うために，音声ひずみと出力 SNR（雑音抑圧後の音声の SNR）を併用した推定式を提案した [3]．

$$y = f(x_1, x_2) = \frac{a}{1 + e^{-b_1 x_1 - b_2 x_2 + c}} \tag{1}$$

ここで，y は単語正解精度，x_1 は PESQ[4] スコア，x_2 は出力 SNR である（PESQ は音声ひずみの尺度であり，PESQ スコアが大きいほど音声ひずみは小さい）．また，a，b_1，b_2，c は定数であり，PESQ スコアと単語正解精度を実験的に求め，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定数を求める必要がある．

そこで、文献 [2] と同様に，認識タスクの複雑さ α をパラメータとして持つ，すなわち式 (1) の定数を α の関数で置き換えた推定式を提案する．

$$\begin{aligned} y &= f(x_1, x_2, \alpha) = \frac{a(\alpha)}{1 + e^{-b_1(\alpha) x_1 - b_2(\alpha) x_2 + c(\alpha)}} \\ &= \frac{p_1 \cdot \alpha^{q_1}}{1 + e^{-p_2 \cdot \alpha^{q_2} x_1 - p_3 \cdot \alpha^{q_3} x_2 + p_4 \cdot \alpha^{q_4}}} \end{aligned} \tag{2}$$

ここで，認識タスクの複雑さの尺度として SMR-パープレキシティ [5] を用いている．また，p_1〜p_4，q_1〜q_4 は定数であり，様々な雑音抑圧アルゴリズムと認識タスクを対象として PESQ スコア，出力 SNR，SMR-パープレキシティ，単語正解精度を求め，これらの関係を最適近似することにより決定する．

3 提案手法の有効性の検証

3.1 実験条件

提案手法の有効性を検証するために，様々な語彙サイズの孤立単語認識の単語正解精度を推定する．

音声データは，東北大−松下単語音声データベース [6] の鉄道駅名の 3,285 語である．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（各認識タスクの SMR-パープレキシティは語彙サイズと等しい）．この音声データに電子協騒音データベース [7] の雑音（car1，hall1，train2，lift2）を 20，15，10，5，0，5dB の SNR で重畳した．雑音抑圧アルゴリズムには，(B) 雑音抑圧を行わない場合，及び (G)GMM に基づく音声信号推定 [8]，(K) ピッチ同期 KLT[9]，(S)SS-SMT 法 (スペクトルサブトラクション法)[10]，(T) 時間領域 SVD に基づく音声強調 [8] を用いた．孤立単語認識のための音響モデルは，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）[11] である．

認識性能推定の条件は以下の通りである．

(C1) 7 種類の認識タスクに対して， 5 種類の雑音抑圧アルゴリズムを音声データに適用した後，単語正解精度，PESQ スコア，出力 SNR，SMR-パープレキシティを算出し，従来手法 [2] と提案手法の推定式 (2) の定数をそれぞれ最適化する．そして，5 種類の雑音抑圧アルゴリズムを同じ音声データに適用したときの単語正解精度を推定する．

(C2) 4 種類の認識タスク（語彙サイズ 50， 200， 800， 3285）に対して，5 種類の雑音抑圧アルゴリズムをその 4 種類の認識タスクの音声データに適用した後，(C1) と同じように 2 つの推定式の定数をそれぞれ最適化する．そして，5 種類の雑音抑圧アルゴリズムを残り 3 種類の認識タスクの音声データに適用したときの単語正解精度を推定する．これは認識タスクに対してオープンなテストである

3.2 実験結果

まず，条件 (C1) の場合の，従来手法 [2] における PESQ スコアと単語正解精度の関係，及び提案手法における式 (1) のひずみ相当量 $(b_1 x_1 + b_2 x_2)$ と単語正解精度の関係をそれぞれ図 1 と図 2 に示す．図中のマーカの種類は認識タスクの違いを表す．これらの図から，提案手法の方がタスク毎のばらつき（これは雑音抑圧アルゴリズムの違いに起因する）が小さいことが分かる．また，語彙サイズ 50 の認識タスクの推定式を図 3 に示す．ここで，実線は提案手法の推定式 (2) に $\alpha = 50$ を代入して得られた推定式，点線は

* Performance estimation of noisy speech recognition applicable to different noise reduction algorithms and recognition tasks, by Ling GUO, Takeshi YAMADA, Shigeki MIYABE, Shoji MAKINO, and Nobuhiko KITAWAKI (University of Tsukuba).

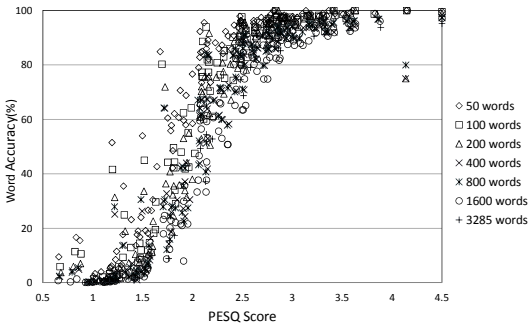


Fig. 1 Relationship between the word accuracy and the PESQ score in different recognition tasks.

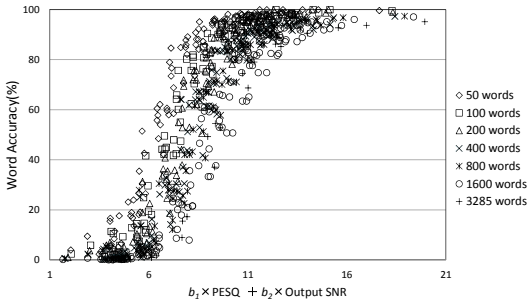


Fig. 2 Relationship between the word accuracy and the total distortion calculated by $b_1x_1 + b_2x_2$ in different recognition tasks.

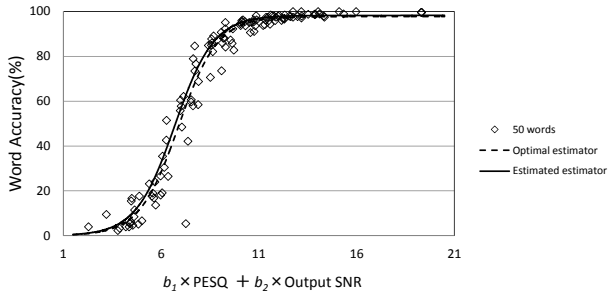


Fig. 3 The estimators for the 50 words task.

語彙サイズ 50 の認識タスクのみを用いて提案手法の推定式 (1) の定数を最適化したときの推定式である．提案手法により得られた推定式は，最適化した推定式に近いことを確認できる．

次に，条件 (C2) の場合の，推定した単語正解精度と真の単語正解精度の関係を図 4 に示す．図 4(a) における従来手法 [2] の決定係数は 0.95，RMSE は 8.6 であるのに対して、図 4(b) における提案手法の決定係数は 0.96，RMSE は 7.2 であり，提案手法の方が単語正解精度を高精度に推定できることが分かる．

4 まとめ

本稿では，音声ひずみ，出力 SNR，SMR-パープレキシティを用いた，種々の雑音抑圧手法と認識タスクに適用可能な認識性能推定法を提案した．また，実験により従来手法よりも高精度な推定が可能であることを確認した．

参考文献

[1] T. Yamada *et al.*, “Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice,” IEEE Trans. ASLP, Vol. 14, No. 6, pp. 2006–2013, Nov. 2006.

[2] T. Yamada *et al.*, “Performance estimation of noisy speech recognition considering recognition task complexity,” Proc. INTERSPEECH 2010, pp. 2042–2045, Sep. 2010.

[3] L. Guo *et al.*, “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,” 日本音響学会秋季研究発表会, pp. 145–148, Sep. 2013.

[4] ITU-T Rec. P.862, “Perceptual evalation of speech quality(PESQ):An objective method for end-to-end speech quality assesment of narrow-band telephone networks and speech codecs,” Feb. 2001.

[5] 中川聖一ら，“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] 牧野正三ら，“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, 1992.

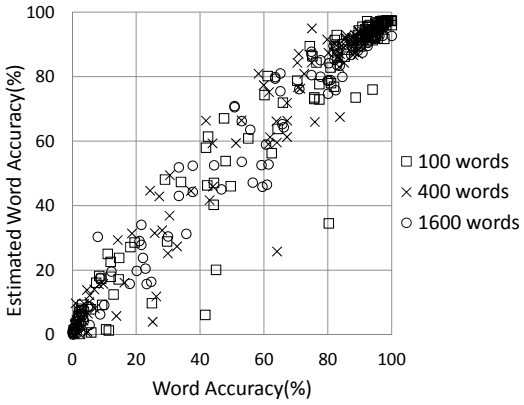
[7] 電子協騒音データベース, http://research.nii.ac.jp/src/list/detail.html#JEIDA-NOISE.

[8] M. Fujimoto *et al.*, “Combination of temporal domain SVD based speech enhancement and GMM based speech estimation for ASR in noise – evaluation on the AURORA2 task –,” Proc. EUROSPEECH 2003, pp. 1781–1784, 2003.

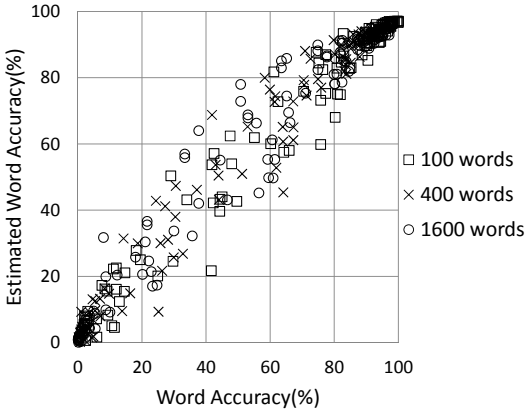
[9] S. -J. Park *et al.*, “Improvement of the ASR robustness using combinations of spectral subtraction and KLT based adaptive comb-filtering,” IPSJ SIGNotes, SLP-44-3, pp. 13–18, 2002.

[10] N. Kitaoka *et al.*, “Evaluation of spectral subtraction with smoothing of time direction on the AURORA 2 task,” Proc. ICSLP 2002, pp. 465–468, 2002.

[11] 河原達也ら, “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.



(a) Conventional method



(b) Proposed method

Fig. 4 Relationship between the true word accuracy and the estimated word accuracy in the condition (C2).