ノンリファレンスひずみ特徴量を用いた雑音下音声認識性能推定の検討*

☆郭翎，山田武志，宮部滋樹，牧野昭二，北脇信彦（筑波大）

1 はじめに

一般に雑音環境において発話された音声を正しく認識することは困難であり，雑音の特性や認識システムの構成によって認識性能は大きく変動する．よって，サービス品質（認識性能）の保証という観点から，音声認識サービスの提供を開始する前に最適なシステム構成やそれにより得られる認識性能を調査する必要がある．また，音声認識サービスの提供を開始した後には，保証した認識性能が得られているのかを継続的にモニタリングする必要がある．

サービス提供前の性能調査を効率的に行うためには，各種の認識性能推定手法 [1]〜[3] の利用が有効である．これは，対象とする雑音環境における音声のひずみの大きさから認識性能を推定するものである．これまでに我々は，雑音の性質や雑音抑圧手法の種類，及び認識タスクの違いに依存しない認識性能推定手法を提案した [4]〜[5]．しかし，これらの手法は，ひずみの大きさを正確に求めるために原音声（雑音が重畳する前のクリーン音声）を必要とするため，サービス提供後の性能モニタリングに適用することはできない．

そこで本稿では，ノンリファレンス（原音声を必要としない）ひずみ特徴量を用いた認識性能推定手法を提案する．提案手法では，まずノンリファレンス型客観品質評価法である ITU-T 勧告 P.563[6] の内部で使用されている，43 種類のノンリファレンスひずみ特徴量を抽出する．さらに，認識性能のタスク依存性を解消するために，認識タスクの複雑さを表す SMR パープレキシティ [7] の値を特徴量に追加する．そして，SVR（Support Vector Regression）を用いて認識性能を推定する．提案手法は，性能モニタリングのみではなく，例えば音声対話システムにおける信頼度尺度として，また十分な認識性能が得られない音声データの効率的な収集のために用いることができると考えられる．

2 提案手法

提案手法の処理フローを Fig. 1 に示す．

提案手法では，まずノンリファレンス型客観品質評価法である ITU-T 勧告 P.563[6] の内部で使用されている，43 種類のノンリファレンスひずみ特徴量を抽出する．ここで，抽出される特徴量には，発話者の声道特性，背景 SNR，無音長などがある．さらに，認識タスクの複雑さを表す SMR パープレキシティ [7] の値を特徴量に追加する．SMR パープレキシティを用いることにより，認識性能のタスク依存性を解消することが可能になると考えられる [8]．

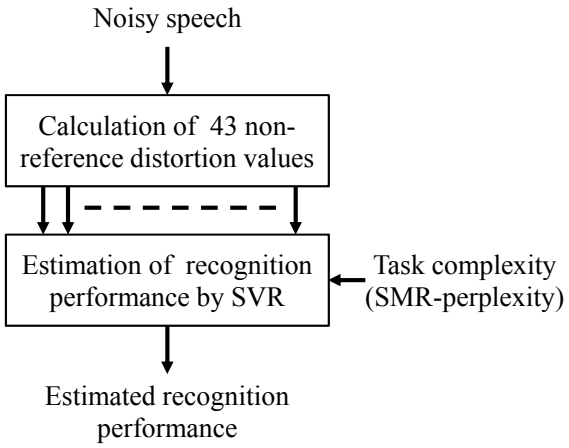


Fig. 1 Estimation of the recognition performance from the non-reference distortion values and task complexity.

最後に SVR を用いて認識性能を推定する．SVR の学習は，雑音，雑音抑圧手法，及び認識タスクの様々な組合せに対して求めた特徴量と認識率のペアデータを用いて行う．一度学習すれば，以降は変更を加えることなく，未知の条件に対して適用可能であると期待できる．

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，様々な語彙サイズを持つ孤立単語認識の単語正解精度を推定する実験を行う．

まず，実験に用いる音声データと認識システムの構成について述べる．音声データとして，東北大−松下単語音声データベース [9] に収録されている鉄道駅名の 3200 語を用いる．本実験では，語彙サイズを 50，100，200，400，800，1600，3200 とすることにより，7 種類の認識タスクを設定する．ここで，各認識タスクの SMR パープレキシティは語彙サイズと同じである．また，雑音データとして，電子協騒音デー

*Performance estimation of noisy speech recognition using non-reference distortion features, by Ling GUO, Takeshi YAMADA, Shigeki MIYABE, Shoji MAKINO, and Nobuhiko KITAWAKI (University of Tsukuba).

タベース [10] に収録されている 4 種類の雑音（Car1，Hall1，Train2，Lift2）を用い，各音声データに 20，15，10，5，0，−5 dB の SNR で重畳する．雑音重畳音声データのサンプリング周波数は 16kHz，量子化ビット数は 16 である．本実験では，音声認識のフロントエンドとして次の 3 種類を用いる．

- NON（雑音抑圧を行わない場合）
- MMSE（Minimum Mean Square Error Short-Time Spectral Amplitude Estimator ）[11]
- WF（Wiener Filtering ）[12]

ここで，MMSE と WF は音声波形を出力する雑音抑圧手法である．以上の雑音重畳音声データ，及び雑音抑圧音声データをモノフォン性別非依存モデル（16 混合分布）[13] を用いて認識する．

次に，ノンリファレンスひずみ特徴量の算出について述べる．P.563 は電話帯域の音声を対象としているため，まず上記の雑音重畳音声データ，及び雑音抑圧音声データを 8kHz にダウンサンプリングする．そして，同一話者の 5 単語を一つに連結して特徴量を算出する．最後に各連結単語から求めた特徴量を連結単語の総数で平均する．

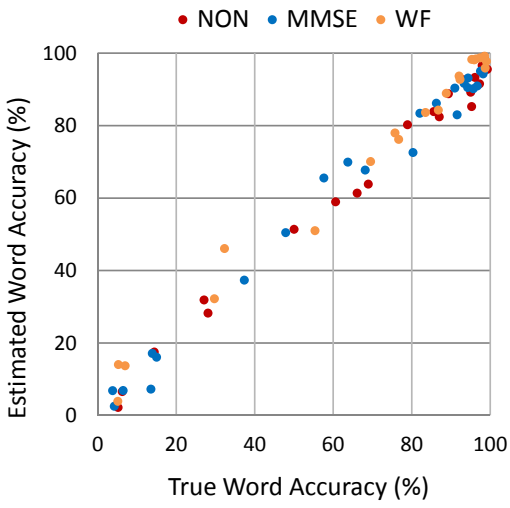
最後に SVR について述べる．SVR の学習と認識性能の推定には LIBSVM[14] を用いる（epsilon-SVR，radial basis function を使用）．コストパラメータは評価データに対する推定誤差が小さくなるように設定する．

3.2　ノンリファレンスひずみ特徴量の有効性の検証

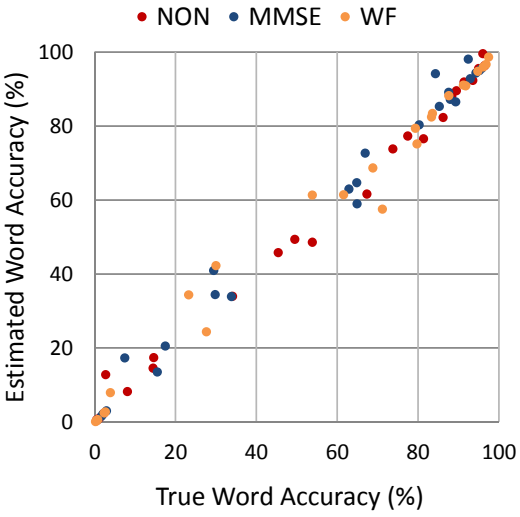
本節では，特定の語彙サイズを持つ孤立単語認識の単語正解精度を推定し，特徴量に P.563 のノンリファレンスひずみ特徴量を用いることの有効性を検証する．

本実験では，語彙サイズが 100，400，1600 の認識タスクを用い，認識タスク毎に学習と評価を行う．各認識タスクからは，学習用と評価用に異なる 50 個の語彙をそれぞれ選択し，SVR の学習には学習用語彙に対応する雑音重畳音声データと雑音抑圧音声データを用いる．同様に，単語正解精度の推定には評価用語彙に対応するデータを用いる．特徴量としては，43 種類のノンリファレンス特徴量のみを用いる（SMR パープレキシティは用いない）．

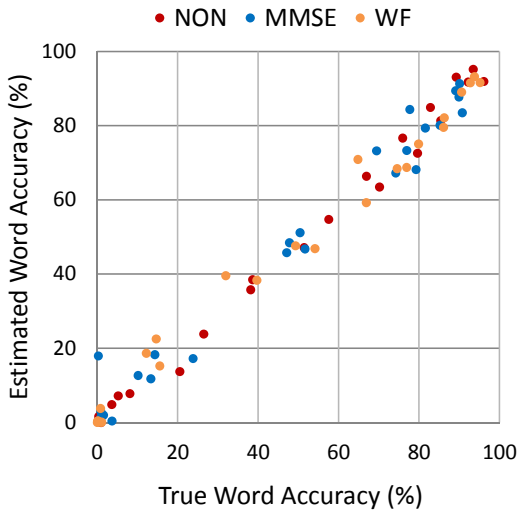
Fig. 2 に，語彙サイズが 100，400，1600 のときの，真の単語正解精度と提案手法により推定した単語正解精度の関係をそれぞれ示す．図中の個々の点は，75 種類の雑音条件（4 雑音× 6 SNR × 3 手法＋ 1 クリー



(a) 100 words



(b) 400 words



(c) 1600 words

Fig. 2　Relationship between the three word accuracy and the estimated word accuracy for each task.

Table 1　Correlation coefficient R and RMSE.

Recognition task	R	RMSE
100 words	0.99	4.3
400 words	0.99	4.1
1600 words	0.99	4.5

ン× 3 手法）の一つにおいて得られた推定単語正解精度と真の単語正解精度を表している．また，真の単語正解精度と推定単語正解精度の相関係数を Table 1 にそれぞれ示す．

Fig. 2 と Table 1 から，提案手法により雑音抑圧手法の違いに依存せずに高い精度で推定できていることが分かる．よって，P.563 のノンリファレンスひずみ特徴量は有効であると言える．

3.3　SMR パープレキシティの有効性の検証

本節では，様々な語彙サイズを持つ孤立単語認識の単語正解精度を推定し，特徴量に SMR パープレキシティを追加することの有効性を検証する．

本実験では，語彙サイズが 50，200，800，3200 の 4 種類の認識タスクを学習に用い，残りの 3 種類の認識タスクを評価に用いる．各認識タスクからは，学習用と評価用にそれぞれ 50 個の語彙を選択する．ここで，選択した語彙は認識タスク間で重複していない．3.2 節と同様に，SVR の学習には学習用語彙に対応する雑音重畳音声データと雑音抑圧音声データを用いる．また，単語正解精度の推定には評価用語彙に対応するデータを用いる．特徴量としては，43 種類のノンリファレンス特徴量のみを用いる場合，及び 43 種類のノンリファレンス特徴量と SMR パープレキシティを用いる場合の 2 通りを比較する．

Fig. 3 と Fig. 4 に真の単語正解精度と提案手法により推定した単語正解精度の関係を示す．Fig. 3 は 43 種類のノンリファレンス特徴量のみを用いる場合であり，相関係数と RMSE は 0.97，7.6 であった．Fig. 4 は 43 種類のノンリファレンス特徴量と SMR パープレキシティを用いる場合であり，相関係数と RMSE は 0.98，6.0 であった．

Fig. 3 を見ると，特に単語正解精度が 20〜80%の範囲において推定誤差が大きいことが分かる．一方，Fig. 4 からはこの範囲の誤差が小さくなっていることが分かる．これは，特徴量に SMR パープレキシティを追加したことの効果であると考えられる．

4　まとめ

本稿では，ノンリファレンスひずみ特徴量を用いた認識性能推定手法を提案した．様々な語彙サイズを持つ孤立単語認識の単語正解精度を推定する実験を

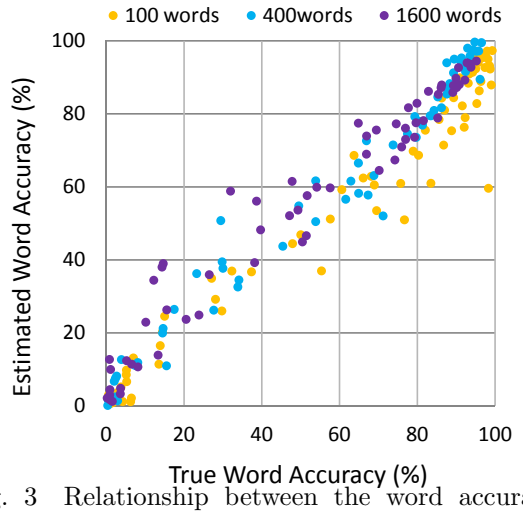


Fig. 3　Relationship between the word accuracy and the estimated word accuracy (Use of 43 non-reference distortion values).

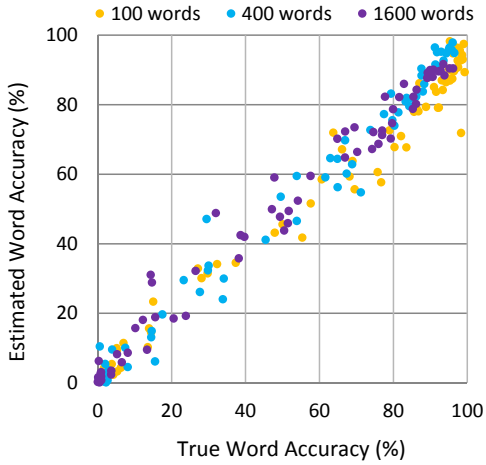


Fig. 4　Relationship between the word accuracy and the estimated word accuracy (Use of 43 non-reference distortion values and SMR-perplexity).

行った結果，ノンリファレンスひずみ特徴量を用いることにより高精度な推定ができること，及び SMR パープレキシティを特徴量に追加することによりタスク依存性を解消できることを確認した．

今後は，孤立単語認識以外の認識タスクに対する有効性を検証し，また他のノンリファレンスひずみ特徴量（例えば [15]）との比較等を行う予定である．

参考文献

[1] H. Sun *et al.*, “Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech,” Proc. ICASSP2004, pp. 865–868, May 2004.

[2] T. Yamada *et al.*, “Performance estimation of speech recognition system under noise conditions using objective quality measures and arti-

ficial voice," IEEE Trans. ASLP, Vol. 14, No. 6, pp. 2006–2013, Nov. 2006.

[3] 福森隆寛 他，"PESQ と室内音響指標を用いた雑音・残響指標 NRSR-PA に基づく雑音・残響下音声認識性能の予測," 電子情報通信学会論文誌, Vol. J94-D-2, No. 4, pp. 1–10, Sep. 2014.

[4] L. Guo *et al.*, "Performance estimation of noisy speech recognition using spectral distortion and SNR of noise-reduced speech," Proc. TENCON 2013, PAPER ID 540, Oct. 2013.

[5] 郭翎 他, "ケプストラム距離と SMR-パープレキシティを用いた雑音下音声認識の性能推定の検討," 日本音響学会春季研究発表会, pp. 145–148, Mar. 2015.

[6] ITU-T Rec. P.563, "Single ended method for objective speech quality assessment in narrow-band telephony applications," May. 2004.

[7] 中川聖一 他，"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.

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

[9] 牧野正三 他，"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.

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

[11] Y. Ephraim *et al.*, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Transactions on Acoustics, Speech, and SignalProcessing, Vol. ASSP-32, No. 6, pp. 1109-1121, Dec. 1984.

[12] S.V. Vaseghi, "Advanced Digital Signal Processing and Noise Reduction," Second Edition, Wiley, 2000.

[13] 河原達也 他，"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.

[14] C.-C. Chang, C.-J. Lin, "LIBSVM: a library for support vector machines," ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.

[15] M. Kondo *et al.*, "Predicting the degradation of speech recognition performance from sub-band dynamic ranges," IPSJ Journal, Vol. 43, No. 7, pp. 2242–2248, July 2002.