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

2007－ＳＬＰ－69（４）
2007／12／20

雑音下音声認識の性能推定のためのひずみ尺度の検討

橋本 倫和† 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: †hashimoto@mmlab.cs.tsukuba.ac.jp, {takeshi,kitawaki}@cs.tsukuba.ac.jp

あらまし これまでに我々は，ひずみ尺度として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．本手法により高い精度で認識性能を推定できるものの，それは個々の雑音抑圧アルゴリズムに最適化した推定式を用意する場合に限られていた．しかし，実用上は一つの推定式で様々な雑音抑圧アルゴリズムに適用できることが望まれる．この問題は，ひずみの大きさと認識性能の関係が雑音抑圧アルゴリズムによって異なることに起因するので，本稿ではひずみ尺度を修正することによりその解決を図った．認識性能の推定実験を行った結果，修正したひずみ尺度の有効性が明らかとなった．また，大量の実音声データの代わりに用いるために開発したタスク依存擬似音声は，修正したひずみ尺度においても有効であることを確認した．

キーワード 雑音下音声認識，性能推定，ひずみ尺度，擬似音声

Distortion measure used for performance estimation of speech recognition system under noise conditions

Tomokazu HASHIMOTO†, Takeshi YAMADA†, and Nobuhiko KITAWAKI†

† Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573 Japan
E-mail: †hashimoto@mmlab.cs.tsukuba.ac.jp, {takeshi,kitawaki}@cs.tsukuba.ac.jp

Abstract Recently, we have developed a method for estimating the recognition performance using the PESQ that is the distortion measure standardized by the ITU-T as Rec. P.862. This method can give an accurate estimate only when the estimator optimized for the noise reduction algorithm used is available. However, practically the single estimator that can be used for various noise reduction algorithms is required. This problem is caused by an inconsistency between the distortion value and the recognition performance. This paper describes the effectiveness of the modified PESQ developed to solve this problem.

Key words noisy speech recognition, performance estimation, distortion measure, artificial voice

1. まえがき

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．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．

これまでに我々は，ひずみ尺度としてITU-T 勧告 P.862 の PESQ [1] を用いる手法を開発した [2]．本手法により高い精度で認識性能を推定できるものの，それは個々の雑音抑圧アルゴリズムに最適化した推定式を用意する場合に限られていた．このことは，雑音抑圧アルゴリズムの内部パラメータを変更した

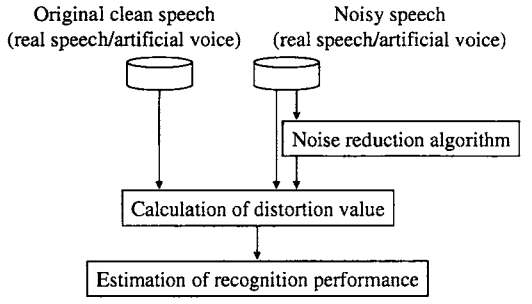


図 1　認識性能の推定の流れ

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

場合や，新しいタイプの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する．しかし，実用上は一つの推定式で様々な雑音抑圧アルゴリズムに適用できることが望まれる．この問題は，ひずみの大きさと認識性能の関係が雑音抑圧アルゴリズムによって異なることに起因するので，本稿ではひずみ尺度を修正することによりその解決を図る．

また，本手法で認識性能を推定する際には，調査対象の雑音環境で音声データを収録する必要がある．我々はこの収録コストを削減するために，大量の実音声データの代わりに数秒程度の擬似音声（ITU-T 勧告 P.50 [3]）を用いる手法を開発した [2]．また，推定精度を改善するために，認識対象語彙を考慮したタスク依存擬似音声を開発し，その有効性を示した [4]．本稿では，上記の修正したひずみ尺度においてもタスク依存擬似音声が有効であるかを検証する．

2. ひずみ尺度と擬似音声を用いた認識性能の推定

2.1 認識性能の推定

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，劣化音声のひずみ値を計算する．そして，そのひずみ値を推定式に代入することにより認識性能を推定する．推定式は次式で表される．

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

ここで，y は認識性能の推定値（本稿では認識性能として単語正解精度を用いる），x はひずみ値である．また，a，b，c は定数であり，様々な雑音環境において劣化音声のひずみ値と認識性能を求め，両者の関係を最適近似するように決定される．

2.2 ひずみ尺度の修正

ITU-T 勧告 P.862 の PESQ [1] は，CODEC などにより劣化した音声の主観品質（主観 MOS）を人間の聴覚心理特性を考慮して求める客観品質評価法である．PESQ 値の算出過程を図 2 に示す．まず，知覚モデルを用いて，原信号と劣化信号をセルと呼ばれる時間，バークスペクトル領域の区画にマッピングする．そして，セル間のひずみをバークスペクトルひずみのラウドネスとして算出し，認知モデルを用いて主観 MOS の推定値（PESQ 値）を得る．

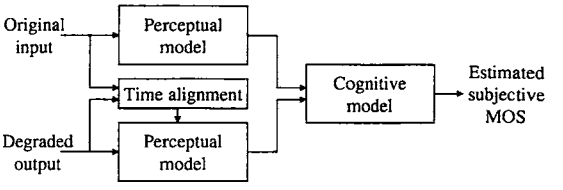


図 2　PESQ 値の算出過程

Fig. 2　Calculation process of the PESQ score.

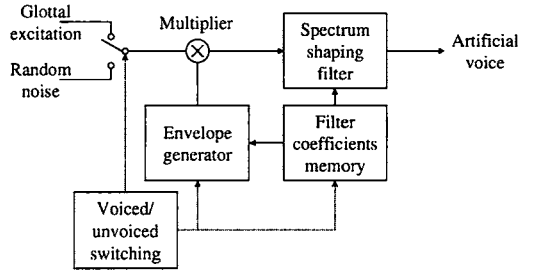


図 3　擬似音声の生成過程

Fig. 3　Generation process of the artificial voice.

PESQ 値は最終的に次式により決定される．

$$\text{PESQ 値} = 4.5 - (\alpha \times D + \beta \times A) \tag{2}$$

以下では，第 1 項は主観 MOS に対応付けるためのものなので無視し，ひずみ値を表す第 2 項のみを扱う．

$$\text{ひずみ値} = \alpha \times D + \beta \times A \tag{3}$$

ここで，D は減算ひずみ（音声成分が減ったことによるひずみ）と加算ひずみ（音声成分が増えたことによるひずみ）を同等に評価した劣化量，A は加算ひずみをより厳しく評価した劣化量である．α と β は，減算ひずみと加算ひずみのどちらを厳しく評価するかを調節する重み係数として機能する．PESQ では，主観品質との対応を良くするように重み係数を設定している．しかし，認識性能との対応は必ずしも良くなく，これまでの実験結果から減算ひずみを厳しく評価し過ぎる傾向があることが分かった．3.2 節では，認識性能との対応が良くなるように重み係数を設定し，その有効性を検証する．

2.3 タスク依存擬似音声

ITU-T 勧告 P.50 の擬似音声 [3] は，音声の平均的特性を有する合成信号である．擬似音声の生成過程を図 3 に示す．有声音と無声音の音源に相当する 2 種類の三角波を組合せた励起信号を，PARCOR 係数をパラメータとする時変係数のスペクトル整形フィルタに通すことにより，擬似音声を生成する．ここで，スペクトルの変化パターンは，ベクトル量子化された 16 個の短時間スペクトルパターンをランダムに選択して与える．また，パワーの変化特性は，無声音と有声音の組合せによる 4 種類のパターンをランダムに選択して与える．さらに，有声音の場合には，人間の声の高さの変化に対応して，ピッチ周波数

表 1　学習セットとテストセット
Table 1　Training and test sets.

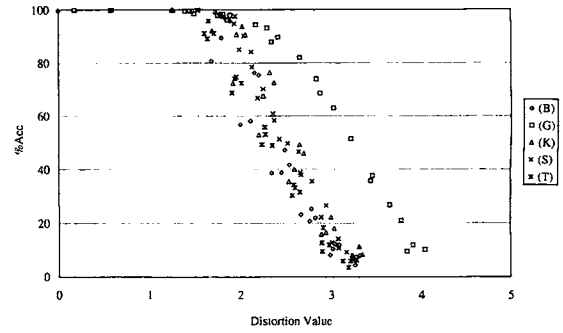
	音声	雑音	チャネル	SNR
Clean training	110 名 8,440 発話	なし	G.712	Clean
テストセット A	104 名 4,004 発話	Subway Babble Car Exhibition	G.712	Clean, 20, 15, 10, 5, 0, −5 [dB]
テストセット B		Restaurant Street Airport Station		

表 2　認識実験の条件
Table 2　Conditions of the recognition experiments.

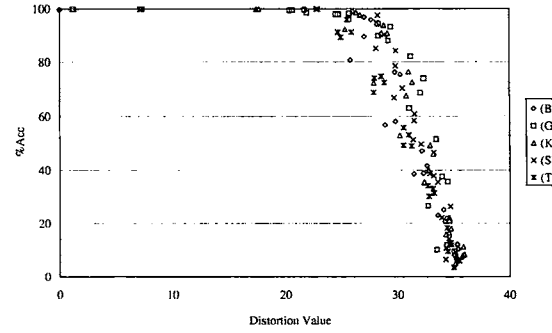
窓関数	ハミング窓
フレーム長	25msec
フレーム周期	10msec
高域強調	$1-0.97z^{-1}$
特徴量	メルケプストラム係数（12 次元）
	＋対数パワー（1 次元）
	＋Δ係数（13 次元）
	＋ΔΔ係数（13 次元）
HMM（数字）	16 状態，混合分布数 20
HMM（sil）	3 状態，混合分布数 36
HMM（sp）	1 状態（sil の第 2 状態と共有）

を変化させる．生成された擬似音声は，長時間平均スペクトル特性，短時間スペクトル変化特性，瞬時振幅レベル分布特性，長時間パワー累積分布特性などが，実音声の特性に最適近似されている．

認識性能の推定の際，大量の実音声データの代わりに P.50 の擬似音声（以下では標準擬似音声と呼ぶ）を用いても，比較的高い精度で認識性能を推定できる．しかし，実音声と比べると推定精度が低下してしまう．これは，標準擬似音声は元来符号化音声の客観品質評価のために開発されたテスト信号であり，発話内容や言語に依存しない特性を持つように設計されているからだと考えられる．一方，音声認識の場合，言語は通常 1 つであり，また小語彙の場合には発話内容が制限されるという特徴がある．そこで，以下のような方法で，認識対象語彙の特性を持つタスク依存擬似音声を生成した [4]．上述したスペクトル整形フィルタのフィルタ群は，ベクトル量子化された 16 個（有声音 13 個，無声音 3 個）の短時間スペクトル特性に相当する．タスク依存擬似音声では，このフィルタ群を認識対象語彙に対応する音声データを用いて再設計する．まず，本稿で対象とする雑音下連続数字認識タスク AURORA-2J [5] の学習データ（クリーン音声）から求めた自己相関係数をクラスタリングし，有声音に対して 13 個，無声音に対して 3 個の代表パターンを求める．そして，各代表パターンの自己相関係数を PARCOR 係数に変換し，擬似音声の生成に用いる．代表パターンの詳細な求め方については文献 [6] を参照されたい．



(a) 修正前のひずみ尺度を用いた場合



(b) 修正後のひずみ尺度を用いた場合

図 4　テストセット A におけるひずみ値と単語正解精度の関係
Fig. 4　Relationship between the distortion value and the word accuracy for the test set A.

3. 認識性能の推定実験

3.1 実験条件

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雑音抑圧アルゴリズムを用いる場合（4 種類）を考えた．

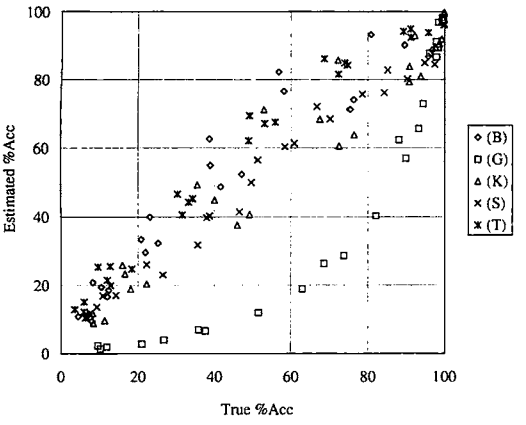
(B)　ベースライン（雑音抑圧を行わない場合）
(G)　GMM に基づく音声信号推定 [7]
(K)　ピッチ同期 KLT [8]
(S)　SS-SMT 法（スペクトルサブトラクション法）[9]
(T)　時間領域 SVD に基づく音声強調 [7]

ここで，各アルゴリズムは時間信号を入出力とし，雑音の推定方法や減算方法に特徴がある．

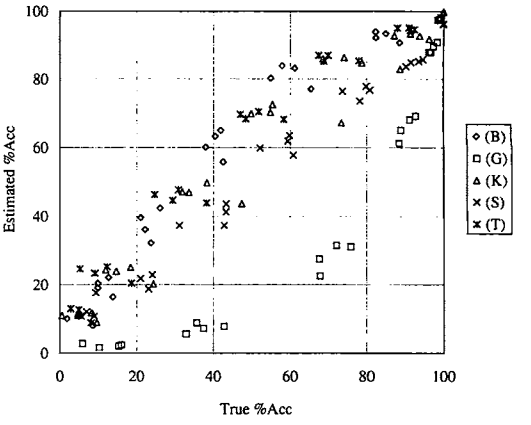
認識実験には，雑音下連続数字認識タスク AURORA-2J [5] を用いた．本実験で用いる学習データ（Clean training）とテストデータ（テストセット A とテストセット B）を表 1 に，認識実験の条件を表 2 に示しておく．学習と認識には，AURORA-2J に添付されている標準スクリプトを用いた．本実験では，学習データに対しても認識時と同様の雑音抑圧処理を行っており，雑音抑圧アルゴリズム毎に専用の音響モデルを作成した．

3.2 修正したひずみ尺度の有効性の検証

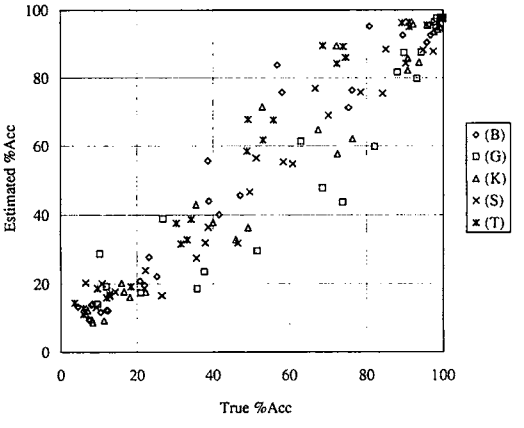
テストセット A におけるひずみ値と単語正解精度の関係を図 4(a)(b) に示す．図中の各点は，雑音抑圧アルゴリズム（5 種類），雑音（4 種類），SNR（7 種類）の組として区別されてい



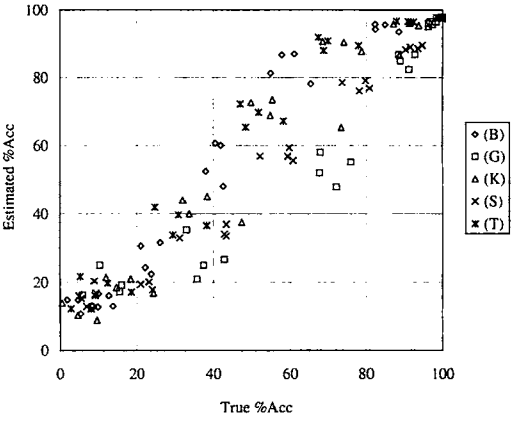
(a) テストセット A，修正前のひずみ尺度を用いた場合
$R^2 = 0.861$，RMSE = 13.19



(c) テストセット B，修正前のひずみ尺度を用いた場合
$R^2 = 0.848$，RMSE = 14.00



(b) テストセット A，修正後のひずみ尺度を用いた場合
$R^2 = 0.937$，RMSE = 8.89



(d) テストセット B，修正後のひずみ尺度を用いた場合
$R^2 = 0.923$，RMSE = 10.42

図 5　テストセット A，テストセット B における真の単語正解精度と推定した単語正解精度の関係

Fig. 5　Relationship between the true word accuracy and the estimated word accuracy for the test sets A and B.

る．また，ひずみ値は，その組に含まれる音声データ（1,001個）から各々算出したものの平均である．

まず，図 4(a) は修正前のひずみ尺度を用いた場合であり，アルゴリズム (G) が他とは異なる傾向を示している．このことから，認識性能を推定するためには，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であることが分かる．次に，図 4(b) は修正後のひずみ尺度を用いた場合である．図 4(a) と比べて雑音抑圧アルゴリズムの違いによるばらつきが軽減していることが分かる．ここで，式 (3) の重み係数は，

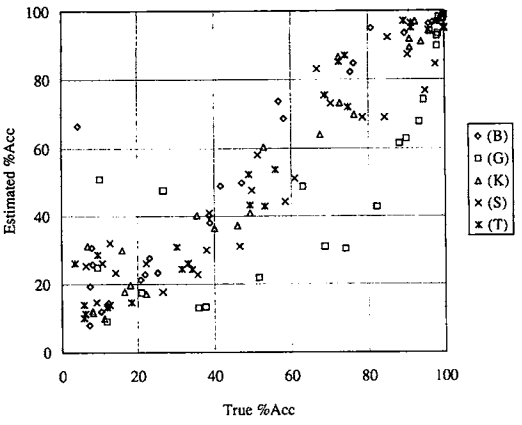
(a)　修正前：$\alpha = 0.10$，$\beta = 0.0309$

(b)　修正後：$\alpha = 0.28$，$\beta = 0.72$

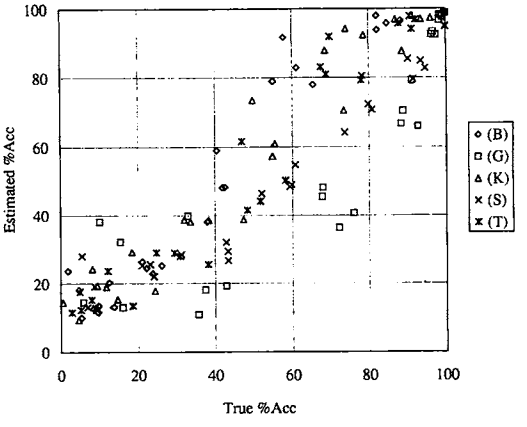
である．修正後の重み係数は，後述する認識性能の推定誤差 (RMSE) が最小となるように設定した．修正前後の重み係数を比べると，修正前の方が減算ひずみを厳しく評価していることが分かる．その結果，他と比べて減算ひずみが大きいアルゴリズム (G) のひずみ値を過大に評価することになっていたと考えられる．

テストセット A，テストセット B における真の単語正解精度と推定した単語正解精度の関係を図 5(a)〜(d) に示す．ここで，単語正解精度の推定には，個々の雑音抑圧アルゴリズムに最適化した推定式ではなく，雑音抑圧アルゴリズムを区別せずに最適化した推定式を一つだけ用いた．例えば，図 5(a) の推定単語正解精度は，図 4(a) の全ての点を最適近似する推定式を用いて推定した．また，図 5(a)(b) では，テストセット A を用いて推定式を求め，テストセット A の単語正解精度を推定した．図 5(c)(d) では，テストセット A を用いて推定式を求め，テストセット B の単語正解精度を推定した．前者は雑音既知，後者は雑音未知という位置付けである．

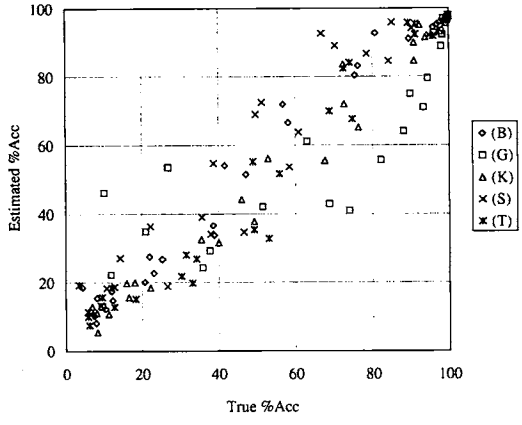
まず，図 5(a) と (b)，及び図 5(c) と (d) を比較することにより，修正後のひずみ尺度を用いた場合の方が，アルゴリズム



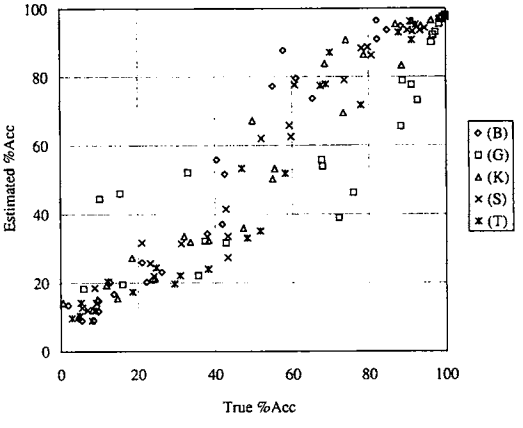
(a) テストセット A，標準擬似音声
$R^2 = 0.857$, $RMSE = 13.37$



(c) テストセット B，標準擬似音声
$R^2 = 0.881$, $RMSE = 12.17$



(b) テストセット A，タスク依存擬似音声
$R^2 = 0.921$, $RMSE = 9.94$



(d) テストセット B，タスク依存擬似音声
$R^2 = 0.915$, $RMSE = 10.37$

図 6　テストセット A，テストセット B における真の単語正解精度と推定した単語正解精度の関係

Fig. 6　Relationship between the true word accuracy and the estimated word accuracy for the test sets A and B.

(G) に対する推定精度が改善されていることが分かる．実際，決定係数 R^2，RMSE 共に改善が見られた．次に，図 5(b) と (d) 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，修正したひずみ尺度の有効性が示された．

3.3　タスク依存擬似音声の有効性の検証

標準擬似音声，タスク依存擬似音声共に各々10 秒程度の男声データと女声データを用意し，AURORA-2J と全く同じ方法で雑音重畳や送話特性の付与を行うことにより，実音声のテストセット A とテストセット B に相当するものを作成した．本実験では，ひずみ値は擬似音声のテストセットを用いて求め，単語正解精度は実音声のテストセットを用いて求めた．また，本実験では，修正後のひずみ尺度のみを用いた．式 (3) のひずみ尺度の重み係数を前節と同様の方法で求めた結果，

- 標準擬似音声：$\alpha = 0.35$, $\beta = 0.65$
- タスク依存擬似音声：$\alpha = 0.26$, $\beta = 0.74$

となった．いずれも前節で述べた実音声の場合の重み係数に近い値であることが分かる．

テストセット A，テストセット B における真の単語正解精度と推定した単語正解精度の関係を図 6(a)〜(d) に示す．ここで，単語正解精度の推定には，前節と同様に，雑音抑圧アルゴリズムを区別せずに最適化した推定式を一つだけ用いた．また，図 6(a)(b) では，テストセット A を用いて推定式を求め，テストセット A の単語正解精度を推定した．図 6(c)(d) では，テストセット A を用いて推定式を求め，テストセット B の単語正解精度を推定した．

図から，標準擬似音声を用いた場合よりもタスク依存擬似音声を用いた場合の方が推定精度が高いことが分かる．また，前節の結果と比較することにより，タスク依存擬似音声を用いた場合と実音声を用いた場合の推定精度は同程度であることが分

かる．以上より，修正後のひずみ尺度に対するタスク依存擬似音声の有効性が確認された．

4. むすび

これまでに我々は，ひずみ尺度として 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．本手法により高い精度で認識性能を推定できるものの，それは個々の雑音抑圧アルゴリズムに最適化した推定式を用意する場合に限られていた．しかし，実用上は一つの推定式で様々な雑音抑圧アルゴリズムに適用できることが望まれる．この問題は，ひずみの大きさと認識性能の関係が雑音抑圧アルゴリズムによって異なることに起因するので，本稿ではひずみ尺度を修正することによりその解決を図った．認識性能の推定実験を行った結果，修正したひずみ尺度の有効性が明らかとなった．また，大量の実音声データの代わりに用いるために開発したタスク依存擬似音声は，修正したひずみ尺度においても有効であることを確認した．今後は，残響などの乗法性雑音や認識システムの構成（認識タスクの複雑さ，バックエンドのロバスト処理など）を考慮した認識性能の推定法を開発する予定である．

謝辞 雑音抑圧アルゴリズムのプログラムをご提供頂いた，武田一哉氏，北岡教英氏，藤本雅清氏に感謝する．本研究の一部は，財団法人電気通信普及財団，財団法人立石科学技術振興財団の研究助成による．本研究では，IPSJ SIG-SLP 雑音下音声認識評価 WG の雑音下音声認識評価環境（AURORA-2J）を利用した．

文　献

[1] 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.

[2] 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.

[3] ITU-T Rec. P.50, "Artificial voices," Sep. 1999.

[4] 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擬似音声信号の検討，" 日本音響学会秋季研究発表会，pp. 119–120, Sep. 2006.

[5] 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.

[6] K. Itoh, N. Kitawaki, H. Nagabuchi, H. Irii, "A new artificial speech signal for objective quality evaluation of speech coding systems," IEEE Transactions on Communications, Vol. 42, No. 2/3/4, pp. 664–672, Feb./Mar./Apr. 1994.

[7] 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.

[8] 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.

[9] 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.