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*

☆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

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手法の種類，及び認識タスクなど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．よって，雑音環境において得られる認識性能を推定する手法が必要不可欠である．

これまでに我々は，ひずみ尺度にケプストラム距離を用いて認識性能を推定する手法を提案した [1]．この手法では，音声区間と非音声区間に分けて算出したケプストラム距離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ケプストラム距離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．この手法は雑音抑圧手法の違いにロバストであり，音声波形を出力するタイプの雑音抑圧手法と特徴量を出力するタイプの雑音抑圧手法の両方に適用できるという特徴がある．しかし，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が，この問題には未対応であった．

そこで本稿では，ケプストラム距離とSMR-パープレキシティを用いた認識性能推定法を提案する．提案手法では，音声区間と非音声区間に分けて算出したケプストラム距離から認識性能を推定する推定式に，認識タスクの複雑さをパラメータとして導入する．これは文献 [2] の手法と同じアプローチである．この手法では単一の雑音抑圧手法のみを対象としていたが，提案手法における推定式は，一度学習すれば以降は変更を加えることなく，様々な雑音抑圧手法と認識タスクに適用可能である．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

提案手法の処理の流れを 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，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と認識タスクの複雑さから認識性能を推定する．

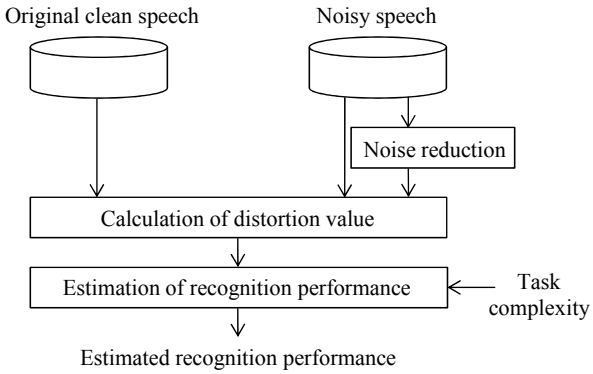

Fig. 1 Estimation of the recognition performance from the distortion value and task complexity.

これまでに我々は，雑音抑圧手法の違いに頑健な認識性能推定を行うために，次式で表される推定式を提案した [1]．

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

ここで，y は単語正解精度，x_1 は音声区間のケプストラム距離，x_2 は非音声区間のケプストラム距離である．また，a，b_1，b_2，c は定数であり，様々な雑音抑圧手法を対象として音声区間・非音声区間のケプストラム距離と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．ただし，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] の手法と同様に，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用いる．

$$\begin{aligned} y &= f(x_1, x_2, \alpha) \\ &= \frac{a(\alpha)}{1 + e^{-b_1(\alpha)x_1 - b_2(\alpha)x_2 + c(\alpha)}} \end{aligned} \tag{2}$$

これは，式 (1) の定数が α に応じて変動することに基づいている．3 章で述べる実験条件で，7 種類の孤

*Performance estimation of noisy speech recognition using cepstral distance and SMR-perplexity, by Ling GUO, Takeshi YAMADA, Shigeki MIYABE, Shoji MAKINO, and Nobuhiko KITAWAKI (University of Tsukuba).

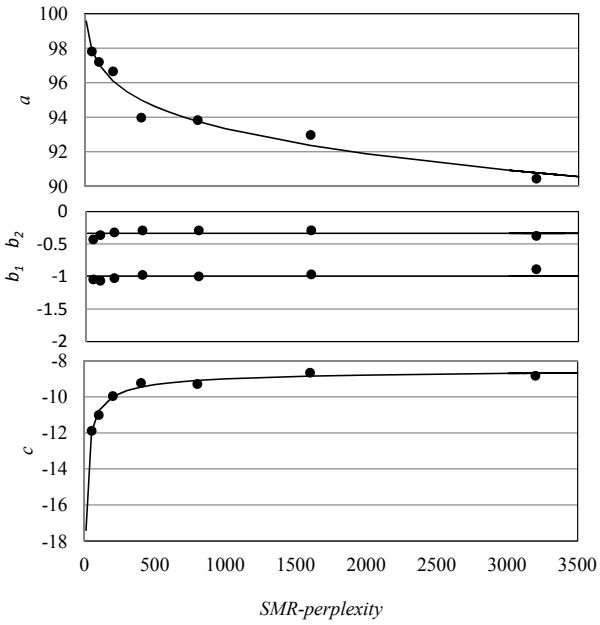

Fig. 2　Relationship between constants a, b_1, b_2, c and α.

立単語認識タスクに対してそれぞれ求めた推定式 (1) の定数 a，b_1，b_2，c と α の関係を Fig. 2 に示す．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 [8] を用いている．この関係から各定数は認識タスクの複雑さにより決まることが確認できる．また，図の中の実線は各定数と α の近似線である．これを参考にし，以下の推定式を提案する．

$$y = \frac{p_1 \cdot \alpha^{q_1} + r_1}{1 + e^{-b_1 x_1 - b_2 x_2 + -p_2 \cdot \alpha^{q_2} + r_2}} \quad (3)$$

ここで，p_1，p_2，q_1，q_2，r_1，r_2 は定数であり，様々な雑音抑圧手法と認識タスクを対象として音声区間と非音声区間のケプストラム距離，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．この推定式は，一度学習すれば，以降は変更を加えることなく様々な雑音抑圧手法と認識タスクに適用できる．

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，孤立単語認識，記述文法認識，大語彙連続音声認識の単語正解精度を推定する．各々の認識タスクの詳細は以下の通りである．

孤立単語認識タスク　音声データとして，東北大−松下単語音声データベース [3] に収録されているものを用いた．これは鉄道駅名の 3,200 語を読み上げたものである．本実験では，語彙サイズを 50，100，200，400，800，1,600，3,200 とすることにより，7 種類の認識タスクを設定した．

Table 1　SMR-perplexity(α) of each task.

Task		α
TMW	50 words	50
	100 words	100
	200 words	200
	400 words	400
	800 words	800
	1600 words	1,600
	3200 words	3,200
AURORA-2J	Connected digits	11
JNAS	5k_MID	40,588
	20k_LARGE	33,975

記述文法認識タスク　音声データとして，AURORA-2J[5] 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．

大語彙連続音声認識タスク　音声データは，新聞記事読み上げ音声コーパス（JNAS）[6] のテストセット 100 文（男性話者）であり，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」[7] に含まれている 3-gram モデルのうち，語彙サイズ 5k，20k の 2 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 2 種類の認識タスクを設定した．

各認識タスクの SMR-パープレキシティを Table 1 に示しておく．雑音データは，電子協騒音データベース [9] の雑音を Test A（car1，hall1，train2，lift2）と Test B（factory1，road2，crowd，lift1）に分けて，それぞれの音声データに 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）である．音声データはすべて 16kHz，16bit である．特徴量として，12 次元のメル周波数ケプストラム係数（MFCC），その一次差分（Δ MFCC）．及びパワーの一次差分（Δ LogPow）を用いる．つまり，各フレー

Table 2　Conditions of the recognition performance estimation.

Condition		Noise	Task	Algorithm
C1	Train	Test A	TMW	NON，MMSE，AFE，SPLICE
	Test	Test B	TMW	NON，MMSE，AFE，SPLICE
C2	Train	Test A	TMW	NON，MMSE，AFE，SPLICE
	Test	Test B	AURORA-2J，JNAS	NON，MMSE，AFE，SPLICE
C3	Train	Test A	TMW	NON，MMSE，AFE，SPLICE
	Test	Test B	AURORA-2J，JNAS	WF

Table 3　Correlation coefficient R and RMSE.

	R	RMSE
C1	0.97	12.2
C2	0.98	8.5
C3	0.98	6.2

ムの特徴量ベクトルは 25（=12+12+1）次元である．

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 種類の雑音抑圧手法を用いる場合を考える．

- NON（雑音抑圧を行わない場合)
- MMSE（Minimum Mean Square Error Short-Time Spectral Amplitude Estimator 法）[10]
- AFE（Advanced Front-End，ETSI ES 202 050 v1.1.5）[11]
- SPLICE（Stereo based Piecewise Linear Compensation for Environments 法）[12]
- WF（Wiener Filtering 法）[13]

ここで，MMSE 法と WF 法は音声波形を出力する雑音抑圧手法であり，AFE と SPLICE 法は音声認識の特徴量であるメルケプストラム係数を直接出力する雑音抑圧手法である．

認識性能推定の条件を Table 2 に示す．条件 C1 では，雑音データ Test A，孤立単語認識タスク，及び雑音抑圧手法 NON，MMSE，AFE，SPLICE を用いて，提案手法の式 (3) の定数をそれぞれ最適化する．そして，雑音データ Test B に対して単語正解精度を推定する．ここで，タスクと雑音抑圧手法は最適化に用いたものと同じである．雑音の種類に対してオープンなテストである．

条件 C2 では、条件 C1 で求めた推定式を用いて，雑音データ Test B，記述文法認識タスクと大語彙連続音声認識タスクに対して単語正解精度を推定する．ここで、雑音抑圧手法は最適化に用いたものと同じである。これは，雑音の種類と認識タスクに対してオープンなテストである．

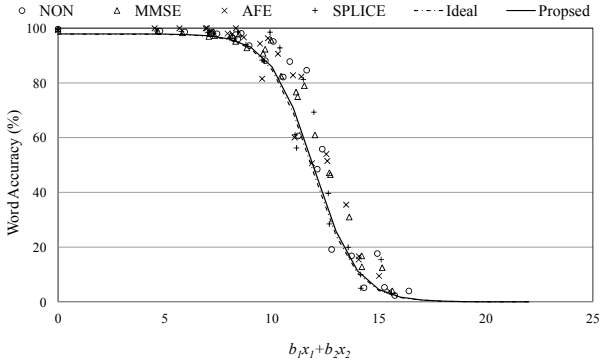

Fig. 3　Estimator obtained by the proposed method for the 50 words isolated word recognition task.

条件 C3 では、条件 C1 で求めた推定式を用いて，雑音データ TestB，記述文法認識タスクと大語彙連続音声認識タスク，雑音抑圧手法 WF に対して単語正解精度を推定する．これは，雑音の種類，認識タスク及び雑音抑圧手法に対してオープンなテストである．

3.2　実験結果

提案手法により得られる推定式の例を Fig. 3 に示す．ここで、認識タスクは語彙サイズ 50 の孤立単語認識タスクである。横軸は音声区間と非音声区間から求めた総合ひずみ（$b_1x_1+b_2x_2$），縦軸は単語正解精度であり、マーカの種類は雑音抑圧手法の違いを表す。また、点線は語彙サイズ 50 の孤立単語認識タスクのみを用いて従来手法の推定式 (1) の定数を最適化したときの理想的な推定式，実線は提案手法の推定式 (3) に $\alpha = 50$ を代入して得られた推定式である．図から，提案手法により得られた推定式と理想的な推定式はほぼ一致していることが確認できる．

条件 C1，C2，C3 における真の単語正解精度と推定した単語正解精度の相関係数と RMSE(Root Mean Square Error) を Table 3 に示す．相関係数は 0.97〜0.98，RMSE は 6.2〜12.2 であり，いずれの条件でも良好な精度で認識性能を推定できることを確認した．

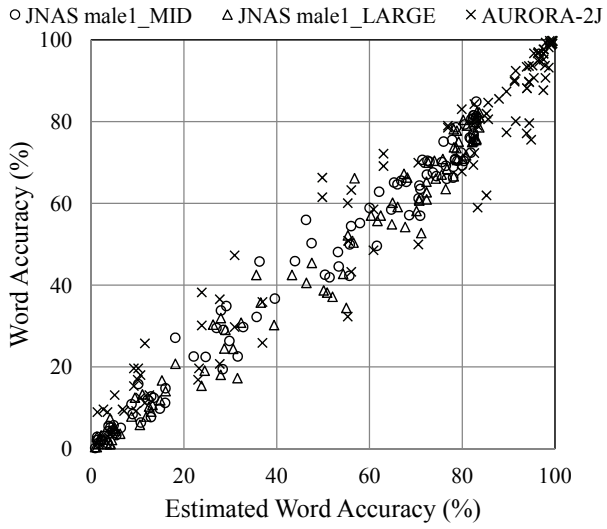

Fig. 4 Relationship between the word accuracy

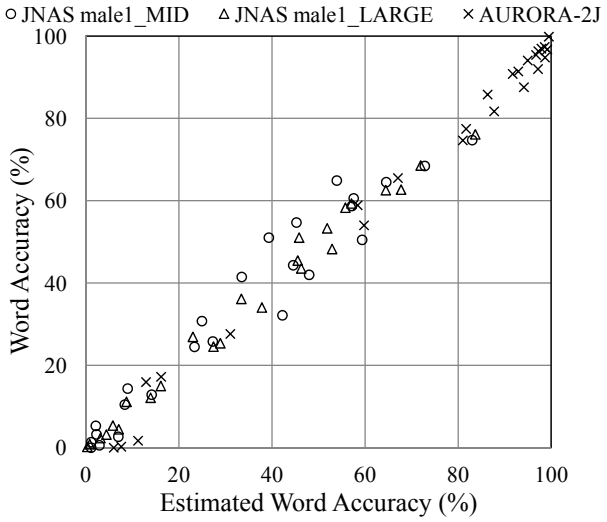

Fig. 5 Relationship between the word accuracy and the estimated word accuracy in the C3.

なお，条件 C1 の RMSE が他の条件より大きい理由を述べる．提案手法の式 (3) において，b_1 と b_2 を α に依存しないとして近似したが、Fig. 2 からは SMR-パープレキシティが小さいときに，近似精度が低下しているように見える．これがその理由であると考えられる．

次に条件 C2 と条件 C3 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 Fig. 5 と Fig. 6 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，雑音の種類，雑音抑圧手法及び認識タスクに対してオープンなテストでも高精度で認識性能を推定できていることが分かる．

4 まとめ

本稿では，音声区間と非音声区間に分けて算出したケプストラム距離と SMR-パープレキシティを併用した認識性能推定法を提案した．提案手法の推定式は，一度学習すれば以降は変更を加えることなく，様々な雑音抑圧手法と認識タスクに適用可能である．認識性能を推定する実験により，良好な精度で認識性能を推定できることを示した．

参考文献

[1] L. Guo ら, “ケプストラム距離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. 2014.

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

[3] 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.

[4] S. Young *et al.*, “The HTK Book(for HTK Version 3.1),” Cambridge Univ., Dec. 2001.

[5] 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.

[6] 新聞記事読み上げ音声コーパス,JNAS: Japanese Newspaper Article Sentences, http://www.mibel.cs.tsukuba.ac.jp/_090624/jnas.

[7] 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.

[8] 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.

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

[10] 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.

[11] ETSI ES 202 050 v1.1.5, “Transmission and Quality Aspects（STQ）, Distributed speech recognition, Advanced front-end feature extraction algorithm, Compression algorithms,” 2007.

[12] J. Droppo *et al.*, “Evaluation of the SPLICE on the AURORA2 and 3 Tasks,” in Proc. ICSLP, pp. 29-32, 2002.

[13] S.F. Boll,“Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” IEEE Trans. Acoustics,Speech and Signal Processing, Vol.ASSP-27, No.7, pp.113-120, 1979.