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*

☆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

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手法の種類，音声認識システムの構成など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法は存在しておらず，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，雑音環境において得られ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．

これまでに我々は，ITU-T 勧告 P.862 の PESQ[1] と雑音抑圧後の音声の SNR を併用して認識性能を推定する手法を提案し，雑音抑圧手法の違いによらず頑健な認識性能推定ができることを示した [2]．しかし，この手法が適用可能なのは，音声波形を出力するタイプの雑音抑圧手法である．音声認識の特徴量を直接出力するタイプの雑音抑圧手法に適用するためには，PESQ スコアを算出するために特徴量を音声波形，あるいはバークスペクトルに変換する必要があり，変換精度といった問題が生じることになる．

現在の音声認識では特徴量としてメルケプストラム係数が標準的に用いられており，ETSI ES202[3] のようにこれを直接出力する雑音抑圧手法が数多く提案されている．よって，本稿では，ケプストラム距離を用いた認識性能推定法を提案する．提案手法の特徴は，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を併用することである．これにより，我々がこれまでに提案してきた PESQ に基づく手法 [2] と同様に，雑音抑圧手法の違いによらない頑健な認識性能推定が可能となる．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

これまでに，ひずみ尺度としてケプストラム距離を用いた認識性能推定の検討がなされている [4]．またこの手法におけるケプストラム距離は，音声区間と非音声区間を含む全フレームで求めたケプストラム距離の平均として定義されている．ここで，単語正解精度とケプストラム距離の関係の例を図 1 に示す．マーカの種類は雑音抑圧手法の違いを表す．個々の点は，雑音（4 種類）と入力 SNR（7 通り）の組に対応する．実線は雑音抑圧手法毎に求めた近似式（推定式）を示す．なお，実験条件の詳細は 3.1 節で述べる．図 1 より，雑音抑圧手法によって推定式が大きく異なることが分かる．これは，認識性能を推定するためには，個々の雑音抑圧手法に最適化した推定式が必要で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，音声区間と非音声区間に分けて算出したケプストラム距離を併用して認識性能を推定する手法を提案する．提案手法では，まず雑音が重畳していないリファレンス音声に対して音声区間検出を行い，音声区間と非音声区間を決定する．そして音声区間と非音声区間のそれぞれに対して次式に示すケプストラム距離を計算する．

$$x = \frac{1}{M}\sum_{m=0}^{M-1}\left\{\frac{1}{K+1}\sum_{k=0}^{K}|c_d(k;m)-c_r(k;m)|^2\right\} \quad (1)$$

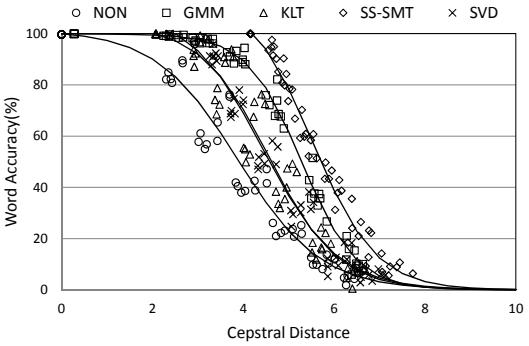

Fig. 1 Relationship between the word accuracy and the cepstral distance.

ここで，M はフレーム数，K はケプストラムの次元，$c_d(k;m)$ と $c_r(k;m)$ は雑音抑圧音声とリファレンス音声のケプストラム係数である．ここで，ケプストラム係数は 0 次を含み，またリファレンス音声に対して対象とする雑音抑圧処理を行っている．最後に，次式を用いて認識性能を推定する．

$$y = f(x_1, x_2) = \frac{a}{1+e^{-b_1x_1-b_2x_2+c}} \quad (2)$$

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と非音声区間のケプストラム距離，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，雑音環境下連続数字認識タスク AURORA-2J[5] を用いて単語正解精度を推定する実験を行った．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 発話である．なお，学習データに対しても認識時と同様に雑音抑圧を行い，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，Babble，Car，Exhibition，Restaurant，Street，Airport，Station の 8 種類，SNR は Clean から −5dB までの 7 通りである．式（2）の推定式の係数を決定するために，雑音抑圧を行わない場合 (NON)，GMM に基づく音声信号推定 (GMM)[6]，ピッチ同期 KLT(KLT)[7]，時間方向スムージングを併用したスペクトルサブトラクション法 (SS-SMT)[8]，時間領域 SVD に基づく音声強調 (SVD)[6] を用いる．これらは，音声波形を出力するタイプの雑音抑圧手法である．また，推定精度を評価するために，Minimum Mean Square Error 法

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

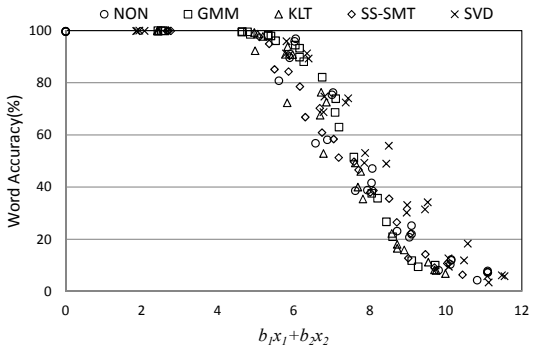

Fig. 2 Relationship between the word accuracy and the distortion calculated by $b_1x_1+b_2x_2$.

Table 1 Correlation coefficient and RMSE（C1）.

	Algorithm	Correlation coefficient	RMSE
Conventional	NON	0.98	7.4
	GMM	0.97	8.8
	KLT	0.99	6.0
	SS-SMT	0.98	6.5
	SVD	0.99	7.1
	Average	0.98	7.2
Proposed	NON	0.98	7.7
	GMM	0.99	4.2
	KLT	0.97	11.8
	SS-SMT	0.99	8.2
	SVD	0.97	7.4
	Average	0.98	7.9

(MMSE)[9], Stereo based Piecewise Linear Compensation for Environments 法 (SPLICE)[10] と ETSI ES202 の Advanced Front-End(AFE)[3] を用いる．これらは，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．

認識性能推定の条件は以下の通りである．

(C1) 音声波形を出力する 5 種類の雑音抑圧手法を Test set A に適用した後，単語正解精度，音声区間と非音声区間のケプストラム距離をそれぞれ算出し，式 (2) の推定式の係数を最適化する．そして， 同じ 5 種類の雑音抑圧手法を Test set B に適用したときの単語正解精度を推定する．これは雑音の種類に対してオープン，雑音抑圧手法に対してクローズドなテストであり，従来手法 [2] についても同様の実験を行う．

(C2) (C1) で求めた推定式を用いて，音声波形を出力しない 3 種類の雑音抑圧手法を Test set B に適用したときの単語正解精度を推定する．これは雑音の種類と雑音抑圧手法の両方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) の場合について，提案手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図 2 に示す．ここで，係数 b_1 と b_2 は最適化後の係数を用いている．図 2 の提案手法と図 1 の従来手法を比較すると，提案手法の方が雑音抑圧手法の違いに頑健であることが分かる．また，各雑音抑圧手法に対して求めた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）を表 1 に示す．表 1 より，提案手法は従来手法と同等の精度で推定できていることが分かる．

Table 2 Correlation coefficient and RMSE（C2）.

	Algorithm	Correlation coefficient	RMSE
Proposed	MMSE	0.96	14.2
	SPLICE	0.94	16.7
	AFE	0.93	14.1
	Average	0.94	15.0

次に条件 (C2) の場合について，MMSE，SPLICE，AFE に対して求めた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 を表 2 に示す．雑音抑圧手法に対してオープンなテストであるために，表 1 と比べて若干推定精度が低下しているものの，特徴量を出力するタイプの雑音抑圧手法についても良好に推定できることが分かる．

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を併用して認識性能を推定する手法を提案した．認識性能を推定する実験により，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双方に適用できることを示した．

参考文献

[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] L. Guo ら, “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.

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

[4] 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.

[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] 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.

[7] 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.

[8] 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.

[9] 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.

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