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

2004－SLP－54（39）
2004／12／21

雑音抑圧手法の主観・客観品質と音声認識性能の関係

山田　武志†　　熊倉　正和††　　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
†† 筑波大学大学院理工学研究科　〒305-8573 茨城県つくば市天王台 1-1-1
E-mail: †{takeshi,kitawaki}@cs.tsukuba.ac.jp, ††kumakura@mmlab.cs.tsukuba.ac.jp

あらまし　雑音環境下での認識率を推定する一つの方法は，雑音が重畳した音声や雑音抑圧後の音声のひずみ値を算出し，その大きさから推定することである．本稿では，雑音抑圧手法の主観・客観品質と音声認識性能の関係について述べる．まず，雑音が重畳した音声と雑音抑圧後の音声の主観品質評価を行った．その結果，音声のひずみ感が大きいために雑音抑圧手法を適用しても主観品質に改善は見られなかったものの，雑音抑圧手法毎に見ると残留雑音が小さくなるにつれて主観品質も高くなっており，認識性能との対応が良いことを確認した．次に，主観品質と認識性能の関係を調べたところ，主観品質から認識性能を高い精度で推定できることが分かった．また，PESQ による客観品質は主観品質との対応が良いこと，主観品質の場合よりも高い精度で認識性能を推定できることを示した．
キーワード　雑音抑圧手法，主観・客観品質，PESQ，音声認識性能

Relationship Between Subjective/Objective Quality of Noise Reduction Algorithms and Speech Recognition Performance

Takeshi YAMADA†, Masakazu KUMAKURA††, and Nobuhiko KITAWAKI†

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

Abstract　One way for estimating speech recognition performance in noisy environments is to use a distortion value calculated from the input speech signal. This paper focuses on ITU-T Recommendation P.862 as a distortion measure and describes the relationship between the subjective/objective quality of noise reduction algorithms and the recognition performance. First, the subjective quality assessment of the noisy speech signals and the output speech signals of noise reduction algorithms is performed. Second, the relationship between the subjective/objective quality and the recognition performance is investigated. These results confirmed that the recognition performance for each noise reduction algorithm can be estimated accurately from the subjective/objective quality.
Key words　Noise reduction algorithms, subjective/objective quality, PESQ, speech recognition performance

1. まえがき

現在の音声認識技術では，雑音が重畳した音声を正しく認識することは困難である [1]．よって，音声認識サービスを提供する際には，サービス品質を保証するために，対象とする雑音環境でどの程度の認識率が得られるのかを事前に知る必要がある．しかし，認識率に影響を与える要因は多岐に渡るため，実際にその場で認識してみないと分からないことが多い．例えば，同じ部屋の中であっても，場所が違うと認識率に数 10%の差が生じることもある．また，雑音の特性や音声認識と併用される雑音抑圧アルゴリズムの性質によっても認識率は大きく変動する．このような現状は，音声認識サービスを広く一般に提供することの妨げとなっている．よって，雑音環境下での認識率を推定する技術が必要不可欠である．

このような技術を利用する人は必ずしも専門的知識を有しているわけではなく，さらには対象となる雑音環境も多岐に渡ることから，複雑な手順を要さない，時間を要さないという点が重要となる．これらの条件を満たす一つの方法は，雑音が重

表 1　主観品質評価の実験条件

Table 1　Experimental conditions of subjective quality assessment.

実験条件	ITU-T 勧告 P.800 [4]
評価法	ACR 法（5 段階絶対品質尺度）
被験者	成人男性 11 名，成人女性 9 名

表 2　受聴実験に用いた音声サンプル

Table 2　Speech samples used for subjective quality assessment.

データベース	AURORA-2J [6] のテストセット A
発話者	男性 2 名，女性 2 名
発話数	雑音 1 種類につき 1 発話
発話内容	7 桁の数字
雑音	Subway，Babble，Car，Exhibition
チャネル	G.712
SNR (dB)	Clean，20，15，10，5，0，-5
雑音抑圧手法	(B) ベースライン（雑音抑圧を行わない場合）
	(S) SS-SMT 法（スペクトルサブトラクション）[7]
	(T) 時間領域 SVD に基づく音声強調 [8]
	(G) GMM に基づく音声信号推定 [8]
	(K) ピッチ同期 KLT [9]
総サンプル数	560

畳した音声や雑音抑圧後の音声のひずみ値を算出し，その大きさから認識性能を推定することである．これまでに著者らは，ひずみ尺度として ITU-T 勧告 P.862 の PESQ (Perceptual Evaluation of Speech Quality) [2]，ケプストラム距離，セグメンタル SNR に着目し，比較実験により PESQ と認識性能の対応が最も良いことを示した [3]．PESQ は人間の主観品質との対応が最も良いとされている客観品質評価法であることから，本稿では人間の主観品質と認識性能の関係を調査する．また，主観品質と客観品質から実際に認識性能を推定し，その性能を評価する．

2.　主観品質と認識性能の関係

2.1　主観品質評価の実験条件

主観品質評価の実験条件を表 1 に示す．ITU-T 勧告 P.800 [4] に準拠する環境条件で受聴実験を実施した．被験者は，成人男性 11 名，成人女性 9 名の計 20 名であり，防音室内でヘッドホンにより音声サンプルを受聴した．実験時間は約 2 時間（10 分程度の受聴と 10 分の休憩を繰返す形態）である．また，評価法は ACR 法，評価尺度は 5 段階絶対品質尺度である．なお，評点の偏りを防ぐために，被験者は最初に MNRU 信号 [5] を含む音声サンプルを受聴した．

受聴実験に用いた音声サンプルを表 2 に示す．受聴実験には，AURORA-2J [6] のテストセット A に含まれる男性 2 名，女性 2 名の音声サンプルから，ランダムに選択したものを使用した．発話内容は 7 桁の数字であり，発話者と雑音の種類が同じならば SNR によらず同一である．以上の音声サンプル，及びそれらに雑音抑圧を行ったものを受聴実験に用いた．ここで，雑音抑圧手法としては，音声に生じるひずみの大きさ・特性，残留雑音の大きさ・特性が異なる 4 種類のものを用いた．被験者は 4（発話者）× 4（雑音）× 7（SNR）× 5（手法），すなわち 560 個の音声サンプルを評価することになる．

表 3　受聴実験に用いた音声サンプルの番号

Table 3　Speech sample numbers used for subjective quality assessment.

Subway	Babble	Car	Exhibition
MGA1ZZ5837	MGA0199113	MGA2810240	MGA5314803
MCE3943ZZ8	MCE18ZZ985	MCE2028691	MCE591Z5Z7
FMJZ5797Z2	FMJZ931ZZ3	FMJ1817945	FMJ4273297
FBA98Z7437	FBA7976050	FBA6924Z63	FBA3468927

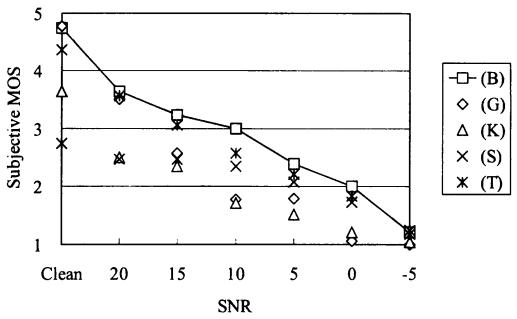


図 1　雑音が Subway のときの主観 MOS

Fig. 1　Subjective MOS, where the noise is Subway.

受聴実験に使用した音声サンプルの番号を表 3 に示しておく．表中の MGA や FMJ は話者コードであり，1 文字目は男性，女性の別を表す．そして，1ZZ5837 や 0199113 は発話内容である．ここで，Z はぜろ，0 はまるを意味する．

2.2　主観品質評価の結果

雑音が Subway のときの主観 MOS を図 1 に示す．図より，(B)，すなわち雑音抑圧を行わない場合の主観 MOS が最も高く，雑音抑圧手法を適用しても主観 MOS は改善していないことが分かる．他の雑音についても同様の傾向が見られた．

雑音抑圧手法が主観品質に及ぼす主な影響としては，(1) 音声に生じるひずみの大きさ・特性，(2) 残留雑音の大きさ・特性がある．本実験では，被験者は (1) を厳しく評価したものと考えられる．特に，本実験で用いた雑音抑圧手法は音声認識性能の改善，すなわち雑音の抑圧性能を重視していることから，(1) がなおさら目立つ結果になったと考えられる．一方，(1) の傾向が同様，すなわち雑音抑圧手法が同じであれば，残留雑音が小さくなるにつれて主観品質も高くなっていることが分かる．このことから，雑音抑圧手法毎に見ると，主観品質と認識性能は良好な関係を示すと期待できる．

2.3　主観 MOS と単語正解精度の関係

本実験では，Clean training を対象として認識実験を行った．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．ベースラインと唯一異なるのは，特徴量の計算の際に CMN を適用していることである．なお，本実験では，

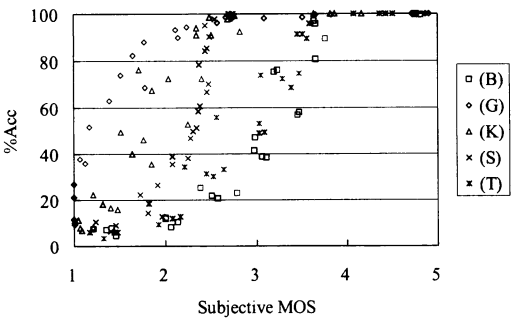


図 2　主観 MOS と単語正解精度の関係

Fig. 2　Relationship between the subjective MOS and the word accuracy.

学習データに対しても認識時と同じ雑音抑圧を行っている(注1).

主観 MOS と単語正解精度の関係を図 2 に示す．ここで，雑音抑圧手法の違いは記号の種類により表される．また，記号が同じとき，個々の点は雑音条件（雑音の種類，SNR）により区別される．図より，全体的に見るとばらつきが大きいものの，雑音抑圧手法毎に見ると主観 MOS と単語正解精度には強い相関が認められる．これは，上述したように，雑音抑圧手法が同じであれば，残留雑音が小さくなるにつれて主観品質，認識性能共に高くなるからであると考えられる．

次に，単語正解精度と主観 MOS から推定した単語正解精度の関係を図 3 に示す．ここで，個々の点は雑音条件（雑音の種類，SNR）により区別される．全手法の場合は，雑音抑圧手法によっても区別される．また，単語正解精度の推定値は，主観 MOS と単語正解精度の関係を雑音抑圧手法毎に 3 次多項式で近似し，その式から求めたものである．図より，雑音抑圧手法毎に見ると，単語正解精度を高い精度で推定できていることが分かる．決定係数 R^2 は 0.90～0.96，RMSE は 5.22～8.86 であった．

3.　客観品質と認識性能の関係

3.1　実験条件

客観品質評価には，ITU-T 勧告 P.862 の PESQ [2] を用いた．PESQ は，人間の主観品質との対応が最も良いとされている客観品質評価法である．客観 MOS（主観 MOS の推定値）の算出過程を図 4 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客観 MOS を得る．

本実験では，AURORA-2J のテストデータの元になっているクリーンな音声データに，各テストセットと同じ送話特性を付与したものを原信号として用いた．また，劣化信号としては，

(注1)：学習データに雑音抑圧を行わない場合の実験も行ったが，Clean training の場合には認識性能に大きな差は生じなかった．

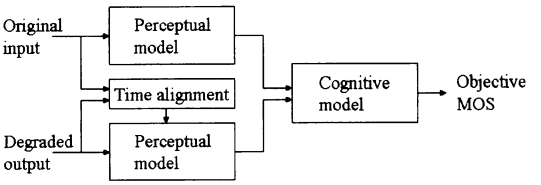


図 4　客観 MOS の算出過程

Fig. 4　Calculation process of the objective MOS.

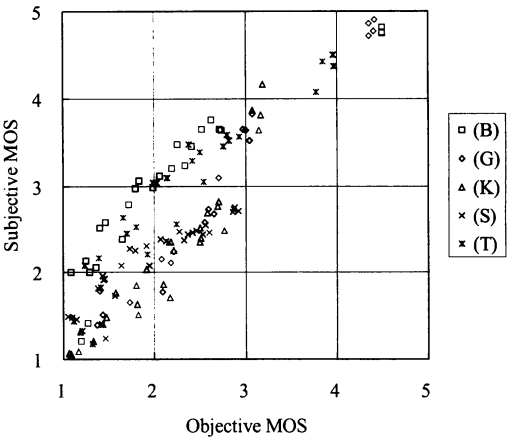


図 5　主観 MOS と客観 MOS の関係

Fig. 5　Relationship between the subjective MOS and the objective MOS.

AURORA-2J のテストデータ，及びそれに雑音抑圧を行ったものを用いた．これらのデータは，受聴実験に用いたものと完全に同じものである．

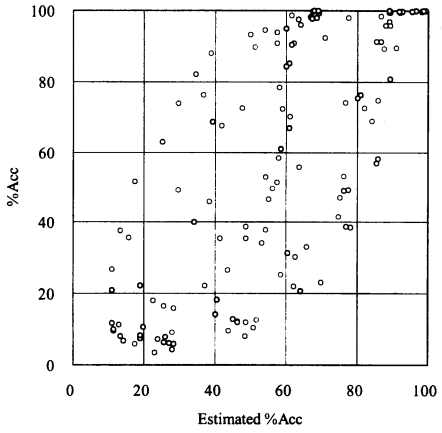
3.2　主観 MOS と客観 MOS の関係

主観 MOS と客観 MOS の関係を図 5 に示す．ここで，個々の点の意味は図 2 と同様である．図より，主観 MOS と客観 MOS には強い相関が認められる．決定係数 R^2 は 0.84，RMSE は 0.44 であった．一方，(B) と (T) については客観 MOS の方が低くなっていることが分かる．その原因については今後調査する必要がある．

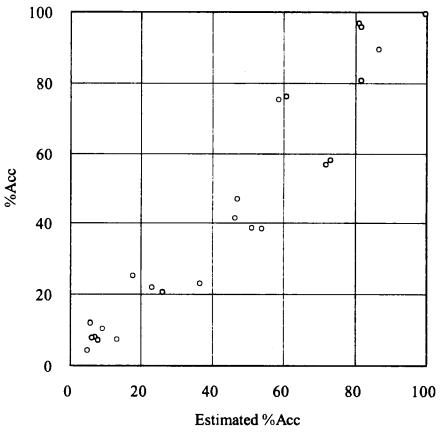
3.3　客観 MOS と単語正解精度の関係

客観 MOS と単語正解精度の関係を図 6 に示す．ここで，認識実験の条件は 2.3 節と同様である．図より，客観 MOS と単語正解精度の間にはやはりばらつきが見られるものの，図 2 と比較するとかなり小さくなっていることが分かる．これは，上述したように，(B) と (T) については客観 MOS の方が低い値を示したことによる．また，雑音抑圧手法毎には強い相関が認められる．

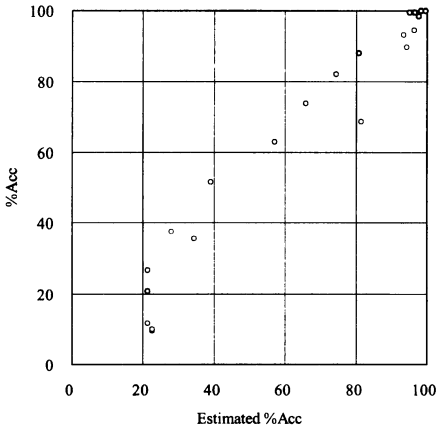
次に，単語正解精度と客観 MOS から推定した単語正解精度の関係を図 7 に示す．ここで，単語正解精度の推定値は，2.3 節と同様の方法で求めた．図から，雑音抑圧手法毎に見ると，単語正解精度を非常に高い精度で推定できていることが分かる．



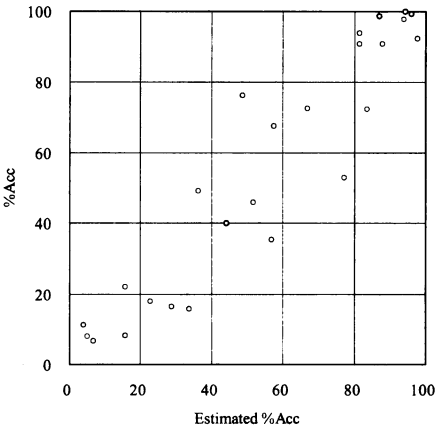
全手法：$R^2 = 0.56$, RMSE = 19.06



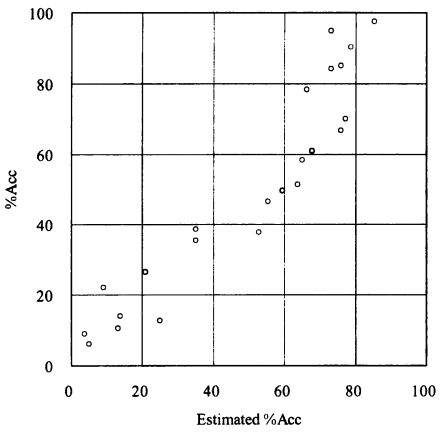
(B)：$R^2 = 0.94$, RMSE = 6.55



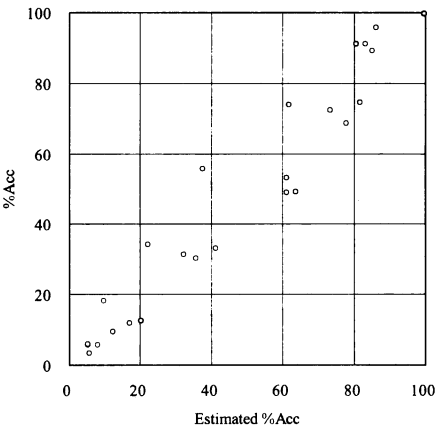
(G)：$R^2 = 0.96$, RMSE = 5.22



(K)：$R^2 = 0.90$, RMSE = 8.86



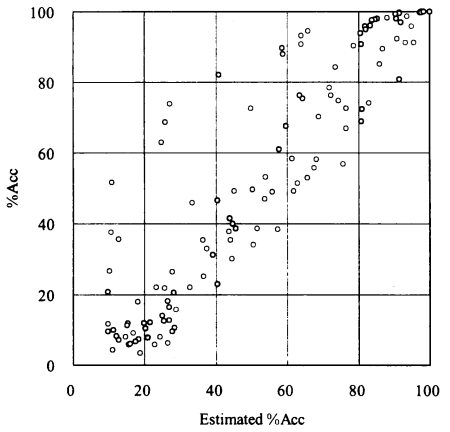
(S)：$R^2 = 0.92$, RMSE = 7.85



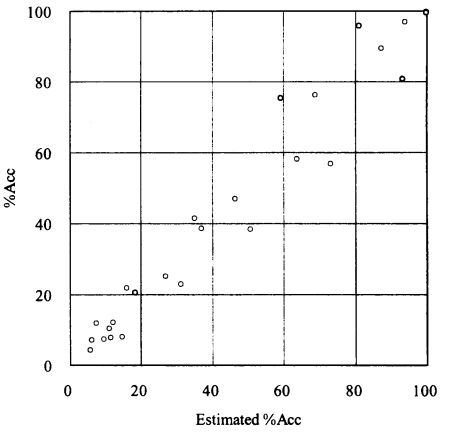
(T)：$R^2 = 0.95$, RMSE = 6.23

図 3　単語正解精度と主観 MOS から推定した単語正解精度の関係

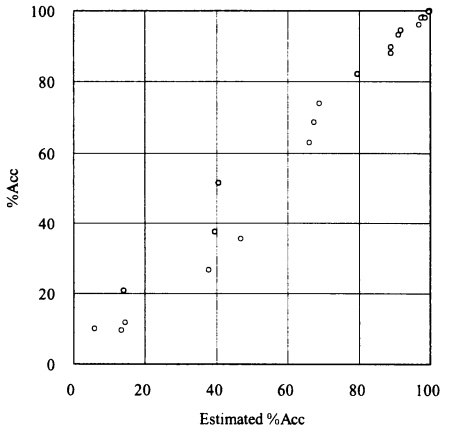
Fig. 3　Relationship between the word accuracy and the word accuracy estimated from the subjective MOS.



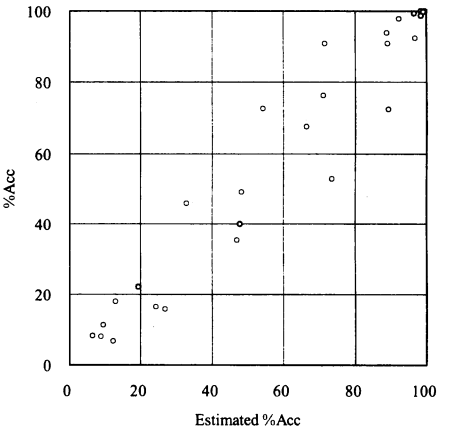
全手法：$R^2 = 0.84$, RMSE = 10.63



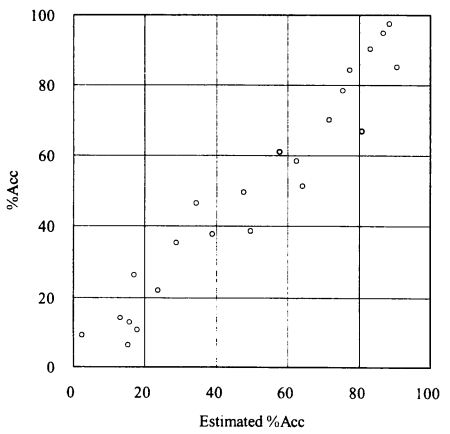
(B)：$R^2 = 0.96$, RMSE = 4.93



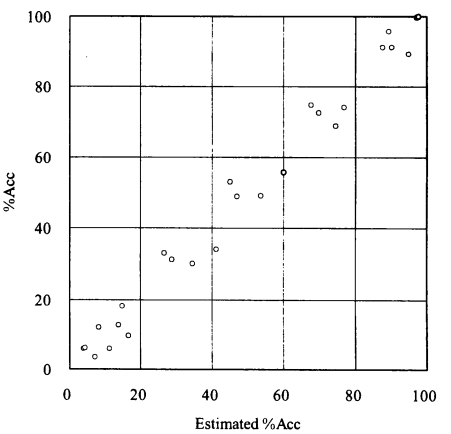
(G)：$R^2 = 0.98$, RMSE = 2.86



(K)：$R^2 = 0.94$, RMSE = 6.24



(S)：$R^2 = 0.96$, RMSE = 5.50



(T)：$R^2 = 0.98$, RMSE = 4.07

図 7　単語正解精度と客観 MOS から推定した単語正解精度の関係

Fig. 7　Relationship between the word accuracy and the word accuracy estimated from the objective MOS.

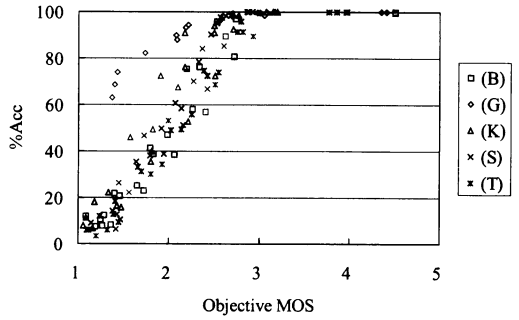


図 6　客観 MOS と単語正解精度の関係

Fig. 6　Relationship between the objective MOS and the word accuracy.

決定係数 R^2 は 0.94〜0.98，RMSE は 2.86〜6.24 であった．

4. むすび

本稿では，雑音抑圧手法の主観・客観品質と音声認識性能の関係について述べた．まず，雑音が重畳した音声と雑音抑圧後の音声の主観品質評価を行った．その結果，音声のひずみ感が大きいために雑音抑圧手法を適用しても主観品質に改善は見られなかったものの，雑音抑圧手法毎に見ると残留雑音が小さくなるにつれて主観品質も高くなっており，認識性能との対応が良いことを確認した．次に，主観品質と認識性能の関係を調べたところ，主観 MOS から単語正解精度を高い精度で推定できることが分かった．また，PESQ による客観品質は主観品質との対応が良いこと，主観品質の場合よりも高い精度で単語正解精度を推定できることを示した．今後は，PESQ のひずみ算出アルゴリズムを音声認識の特性に特化させることにより，雑音抑圧手法の違いにロバストな推定を実現したいと考えている．

謝辞

音声データや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，総務省戦略的情報通信研究開発推進制度，及び NTT サービスインテグレーション基盤研究所の研究委託による．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] 中村哲，“外来に強い音声認識を目指して，” 日本音響学会誌，Vol. 57, No. 10, pp. 662–667, 2001.

[2] ITU-T Recommendation 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.

[3] 山田武志，北脇信彦，“PESQ と擬似音声を用いた雑音下音声認識の性能予測の検討，” 情報処理学会研究報告，SLP-49-7, pp. 37–42, Dec. 2003.

[4] ITU-T Recommendation P.800, “Methods for subjective determination of transmission quality,” Aug. 1996.

[5] ITU-T Recommendation P.810, “Modulated noise reference unit (MNRU),” Feb. 1996.

[6] 山本一公，中村哲，武田一哉，黒岩眞吾，北岡教英，山田武志，水町光徳，西浦敬信，藤本雅清，“AURORA-2J/AURORA-3J データベースとその評価ベースライン，” 情報処理学会研究報告，SLP-47-19, pp. 101–106, July 2003.

[7] 北岡教英，赤堀一郎，中川聖一，“スペクトルサブトラクションと時間方向スムージングを用いた雑音環境下音声認識，” 電子情報通信学会論文誌，Vol. J83-D-II, No. 2, pp. 500–509, 2000.

[8] 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.

[9] S.-J. Park, M. Ikeda, K. Takeda, F. Itakura, “Improvement of the ASR robustness using combinations of spectral subtraciton and KLT based adaptive comb-filtering,” IPSJ SIGNotes, SLP-44-3, pp. 13–18, 2002.