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

2003－ＳＬＰ－49　（７）
2003／12／18

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

山田　武志†　　北脇　信彦†

† 筑波大学電子・情報工学系
〒305-8573 茨城県つくば市天王台1-1-1
E-mail: †{takeshi,kitawaki}@is.tsukuba.ac.jp

あらまし　雑音抑圧手法を音声認識の前処理として用いたときの認識性能を予測する方法としては，雑音抑圧後の音声信号から算出したひずみ値を用いることが考えられる．この方法では認識実験を行わないので，数秒程度の擬似音声を用いることが可能となり，音声認識の運用時の手法選択や新しい手法の研究開発の効率を大幅に高めることができると考えられる．本稿では，音声の客観品質評価のためのひずみ尺度であるITU-T勧告P.862のPESQ，ケプストラム距離，セグメンタルSNRに着目し，ひずみ値と認識性能の関係を雑音下連続数字認識タスクであるAURORA-2Jを用いて調べた．その結果，単語正解精度との相関が最も強いのはPESQ値であることが明らかとなった．また，ITU-T勧告P.50の擬似音声を用いて，擬似音声の適用可能性を調査した．その結果，実音声から算出したPESQ値と擬似音声から算出したPESQ値は概ね線形関係にあり，実音声の代わりに擬似音声を用いても，単語正解精度との強い相関が保たれていることが明らかとなった．

キーワード　雑音下音声認識，性能予測，擬似音声，PESQ，ケプストラム距離，セグメンタルSNR

Performance Prediction of Noisy Speech Recognition Using PESQ and Artificial Voice

Takeshi YAMADA† and Nobuhiko KITAWAKI†

† Institute of Information Sciences and Electronics, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573 Japan
E-mail: †{takeshi,kitawaki}@is.tsukuba.ac.jp

Abstract　One approach for predicting the performance of speech recognizers using noise reduction algorithms is to use distortion values calculated from the output speech signals of the noise reduction algorithm. This paper focuses on the PESQ (ITU-T Recommendation P.862), the cepstrum distance and the segmental SNR as the distortion measure, and investigates the relationship between the recognition performance and the distortion value. Recognition experiments using four noise reduction algorithms were performed on the AURORA-2J connected digit recognition task. Also, the distortion values were calculated from the real speech and the artificial voice (ITU-T P.50). These results confirmed that there is the strong linear relationship between the word accuracy and the PESQ score for both the real speech and the artificial voice.

Key words　noisy speech recognition, performance prediction, artificial voice, PESQ, cepstrum distance, segmental SNR

1. まえがき

従来，ロバストな雑音下音声認識を実現するために，様々な雑音抑圧手法が提案されている[1]．しかし，これらの手法の有効性は，雑音の種類やパワーなどの雑音条件に左右されることが多い．これは，ある環境において最高の認識性能を得るためには，数多くの手法の中から最適な手法を選択する必要があることを意味している．その際には，各手法の動作を制御するパラメータの最適化も行う必要がある．

雑音抑圧手法を選択する一つの方法は，その環境における各手法の認識性能の予測値を用いることである．現状では，予測値を得るために，あらかじめその環境で音声データや雑音データを収録し，実際に認識実験を行うことが多い．しかし，音声認識を運用する現場で大量のデータを収録することは一般に困

難であり，携帯端末のように対象とする環境を絞れない場合や，自動車内のように収録に危険を伴う場合もある．また，データを収録できる場合でも，認識実験に要するコストの大きさが問題となる．これは，音声認識の運用時だけではなく，雑音抑圧手法の研究開発時にもあてはまることである．

このような問題に対処する方法としては，雑音抑圧後の音声信号のひずみ値を算出し，その大きさから認識性能を予測することが考えられる．この方法では認識実験を行わないので，大量の音声データではなく，数秒程度の擬似音声を用いることが可能となる．よって，事前のデータ収録を要するものの，その量は極めて少ないので，収録や実験に要するコストを大幅に削減できると考えられる．

以上より，本稿では，雑音下連続数字認識タスクであるAURORA-2J [2] を用いて，ひずみ値と認識性能の関係を調べる．また，ITU-T勧告P.50の擬似音声 [3] を用いて，擬似音声の適用可能性を調査する．ここで，ひずみ尺度としては，

- ITU-T勧告P.862のPESQ [4]
- ケプストラム距離
- セグメンタルSNR

を用いる．これらは，音声の客観品質評価に採用されているものである [5]．

2. ひずみ尺度と擬似音声

2.1 セグメンタルSNR

セグメンタルSNR（SNRseg）は，時間領域における波形のひずみの大きさを表す尺度であり，次式で定義される．

$$\mathrm{SNRseg} = \frac{1}{M}\sum_{m=1}^{M}\left[10\log\left\{\frac{\sum_{n=0}^{N-1}\{x(n;m)\}^2}{\sum_{n=0}^{N-1}\{x(n;m)-y(n;m)\}^2}\right\}\right]$$

ここで，Mはフレーム数，Nはフレーム内のサンプル数，$x(n;m)$は原信号，$y(n;m)$は劣化信号である．なお，本稿では，フレーム長を25msec，フレーム周期を10msecとし，SNRseg値の上限を30dBに設定している．

2.2 ケプストラム距離

ケプストラム距離（CD）は，周波数領域におけるスペクトルのひずみの大きさを表す尺度であり，次式で定義される．

$$\mathrm{CD} = \frac{1}{M}\sum_{m=1}^{M}\left\{\frac{1}{K}\sum_{k=1}^{K}|c_x(k;m)-c_y(k;m)|\right\}$$

ここで，Kはケプストラムの次元数，$c_x(k;m)$は原信号のケプストラム係数，$c_y(k;m)$は劣化信号のケプストラム係数である．なお，本稿では，メルケプストラム係数（1〜12次）を用いている．

一般に，時間領域のひずみ尺度よりも，周波数領域のひずみ尺度の方が，人間の主観品質評価との対応が良いとされている．音声認識では周波数領域の特徴量を用いていることから，同様に周波数領域のひずみ尺度の方が認識性能との対応が良いと考えられる．

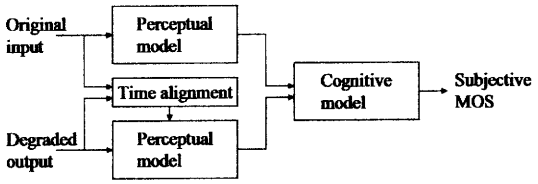


図1 PESQ値の算出過程

Fig. 1 Calculation process of the PESQ score.

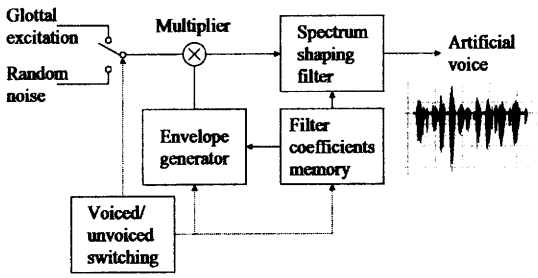


図2 擬似音声の生成過程

Fig. 2 Generation process of the artificial voice.

2.3 PESQ

ITU-T P.862として勧告されているPESQ（Perceptual Evaluation of Speech Quality）[4] は，人間の主観品質評価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．PESQの主な特徴は，人間の知覚過程と認知過程をモデル化していること，パケット損などの時間軸上で離散的に発生するひずみを扱えることである．

PESQ値の算出過程を図1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．

2.4 擬似音声

ITU-T P.50として勧告されている擬似音声 [3] は，音声の平均的特性を有する合成信号である．これは，符号化音声の客観品質評価のためのテスト信号として提案されたものであるが，エコーキャンセラの客観品質評価においても有効であることが示されている [6]．

擬似音声の生成過程を図2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．

表1　AURORA-2Jの学習セットとテストセット
Table 1　Training and test sets of the AURORA-2J.

学習・テストセット	音声	雑音	チャネル	SNR
Clean training	110名，8,440発話	なし	G.712	Clean
Multicondition training	同上	Subway, Babble, Car, Exhibition	G.712	Clean, 20, 15, 10, 5
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
テストセットB	同上	Restaurant, Street, Airport, Station	G.712	同上
テストセットC	104名，2,002発話	Subway, Street	MIRS	同上

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）

3. 評価実験

3.1 実験条件

本実験では，ひずみ値と認識性能の関係が，雑音抑圧手法の違い（ひずみの性質の違い）に左右されるかどうかを調べるために，次の5つの雑音抑圧手法を用いる．

(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法）[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]

ここで，各手法は時間信号を入出力とする．

認識実験には，雑音下連続数字認識タスクであるAURORA-2J[2]を用いる．AURORA-2Jの学習セットとテストセットを表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ている．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．よって，評価カテゴリ[2]は0（バックエンドの変更なし）である．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いる．

ひずみ値の算出について述べる．原信号としては，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と擬似音声（各々10秒程度の男声データと女声データ）を用いる．ここで，原信号には，各テストセットと同じ送話特性を付与している．また，劣化信号としては，雑音抑圧後の音声データと擬似音声を用いる．ここで，擬似音声への雑音重畳や送話特性の付与は，AURORA-2Jのテストデータと全く同じ方法で行っている．

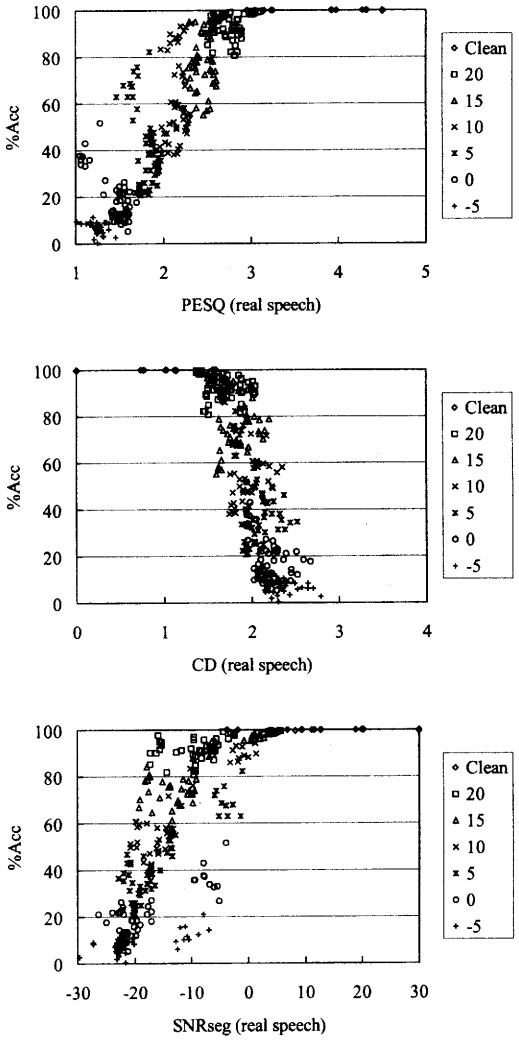


図3　単語正解精度と実音声から算出したひずみ値の関係（Clean training）
Fig. 3　Relationship between the word accuracy and the distortion value calculated from the real speech in the clean training.

3.2 単語正解精度と実音声から算出したひずみ値の関係

単語正解精度（%Acc）と実音声から算出したひずみ値の関係を図3と図4に示す．ここで，図3はClean trainingの場合，

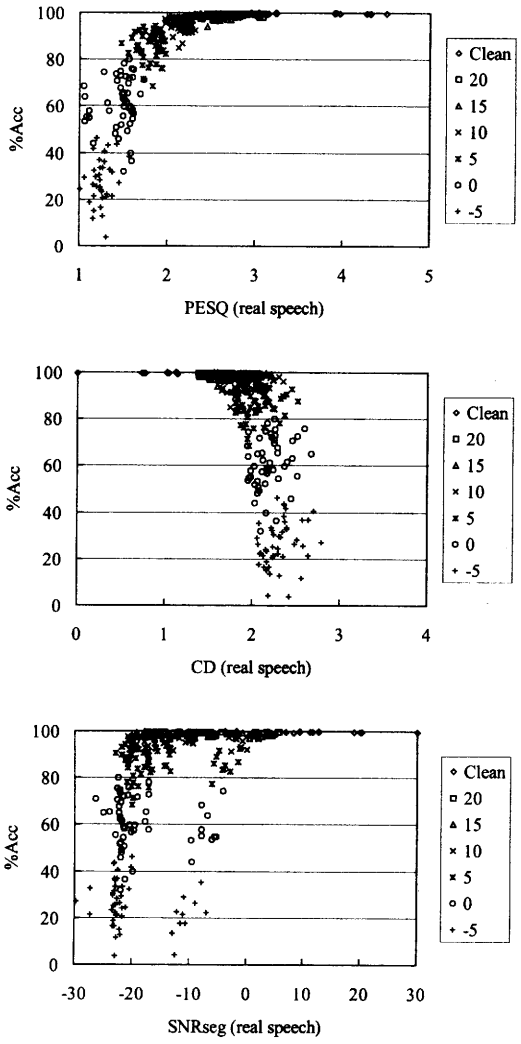


図4 単語正解精度と実音声から算出したひずみ値の関係 (Multicondition training)

Fig. 4 Relationship between the word accuracy and the distortion value calculated from the real speech in the multicondition training.

図4はMulticondition trainingの場合である．図中の個々の点は，雑音の種類（10通り），SNR（7通り），雑音抑圧手法（5通り）の組として区別されている．また，ひずみ値は，その組の全ての音声データから算出した値の平均である．

実験結果を以下にまとめる．

- どのひずみ尺度を用いても，ひずみがある程度小さくなると，単語正解精度は上限に近くなっている．また，Multicondition trainingの場合は，多少ひずみが大きくても，高い単語正解精度が得られることが分かる．ここで，90%以上の単語正解精度を得るために必要なひずみ値を表3に示す．例えばPESQ値に着目すると，Clean trainingの場合は2.90以上，Multicondition trainingの場合は2.17以上であれば，90%以上の単語正解精度が得られることが分かる．

表3 90%以上の単語正解精度を得るために必要なひずみ値

Table 3 Distortion value required for obtaining the word accuracy more than 90%.

	PESQ	CD	SNRseg
Clean training	2.90	1.44	1.59
Multicondition training	2.17	1.74	−1.11

表4 単語正解精度と実音声から算出したひずみ値の相関係数

Table 4 Correlation coefficient between the word accuracy and the distortion value calculated from the real speech.

	PESQ	CD	SNRseg
Clean training	0.83	−0.74	0.78
Multicondition training	0.81	−0.54	0.43

- Clean trainingの場合はPESQ値とCD値に，Multicondition trainingの場合はPESQ値に，単語正解精度との強い相関が認められる．ここで，SNR20〜0dBにおける，単語正解精度と実音声から算出したひずみ値の相関係数を表4に示す．表4から，単語正解精度と最も相関が強いのはPESQ値であることが分かる．特にMulticondition trainingの場合は，PESQ値の優位性が目立っている．

次に，単語正解精度と実音声から求めたPESQ値の手法毎の関係を図5と図6に示す．ここで，図5はClean trainingの場合，図6はMulticondition trainingの場合である．

図より，ひずみ値が同じであれば，手法(G)の単語正解精度が一番高いことが分かる．これは，手法(G)によるひずみの大きさを，過剰に評価していることを意味している．この問題に対処するためには，PESQ値の算出に用いる知覚・認知モデルを，音声認識の特性を考慮して改良する必要があると考えられる．なお，手法(G)に対応する点を除いて，表4に示した相関係数を算出すると，Clean trainingの場合には0.92となる．

3.3 実音声から算出したPESQ値と擬似音声から算出したPESQ値の関係

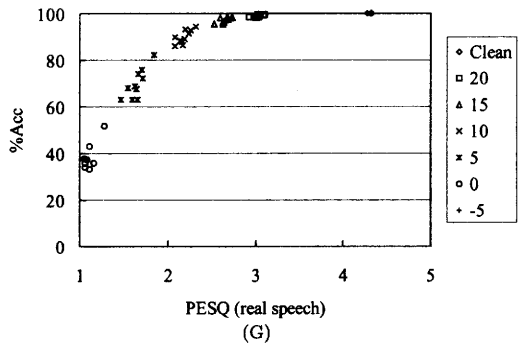
実音声から算出したPESQ値と擬似音声から算出したPESQ値の関係を図7に示す．

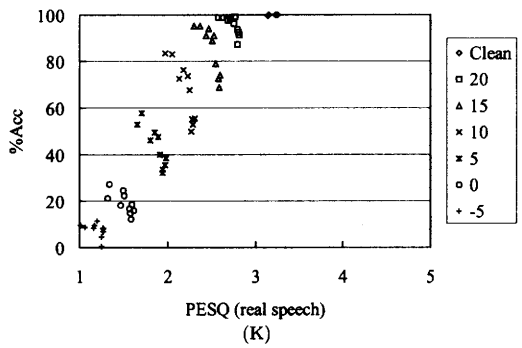
図より，擬似音声から算出したPESQ値の方が高い値を示す傾向がみられるものの，両者は概ね線形関係にあることが分かる．ここで，SNR20〜0dBにおける相関係数は0.90であり，CD値の場合の0.79やSNRseg値の場合の0.78よりも高い値を示している．さらに相関を強くするためには，音声の平均的な特性ではなく，音声認識タスク（語彙セット）の特性を持つように，擬似音声を改良する必要があると考えられる．

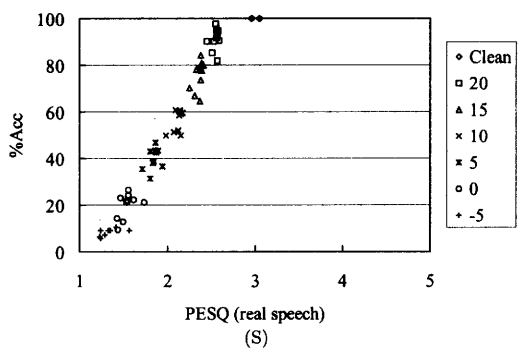
3.4 単語正解精度と擬似音声から算出したPESQ値の関係

単語正解精度と擬似音声から算出したPESQ値の関係を図8に示す．ここで，手法(G)に対応する点は除いている．

図より，実音声の代わりに擬似音声を用いても，単語正解精度との強い相関が保たれていることが分かる．ここで，SNR20〜0dBにおける相関係数は，Clean trainingの場合は0.89，Mul-







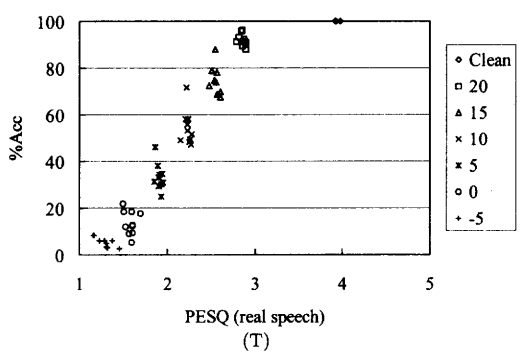
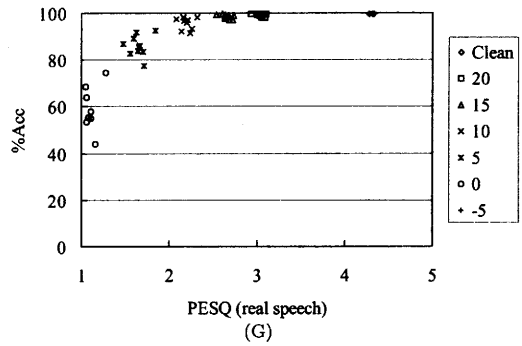
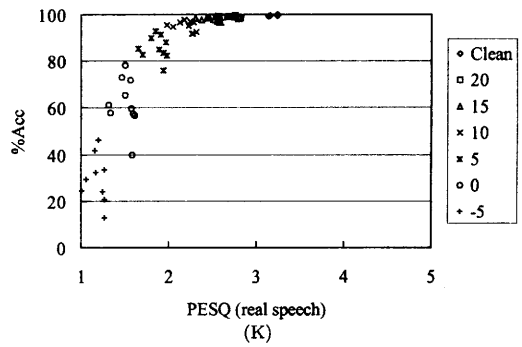
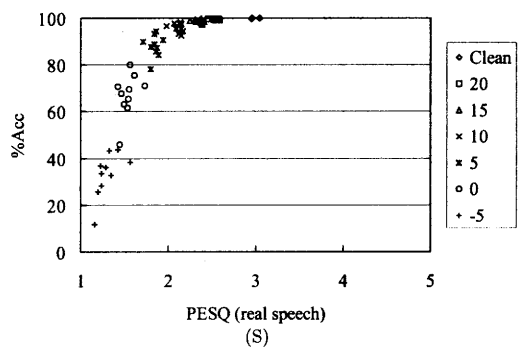


図5 単語正解精度と実音声から求めたPESQ値の手法毎の関係 (Clean training)

Fig. 5 Relationship for each algorithm between the word accuracy and the PESQ score calculated from the real speech in the clean training.







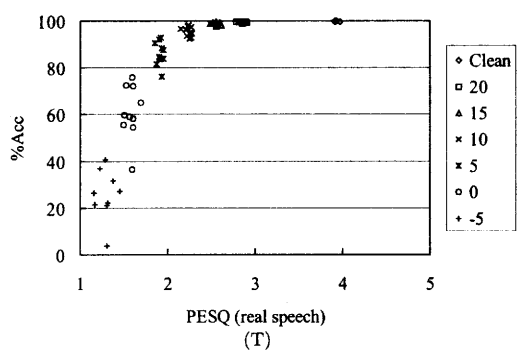


図6 単語正解精度と実音声から求めたPESQ値の手法毎の関係 (Multicondition training)

Fig. 6 Relationship for each algorithm between the word accuracy and the PESQ score calculated from the real speech in the multicondition training.

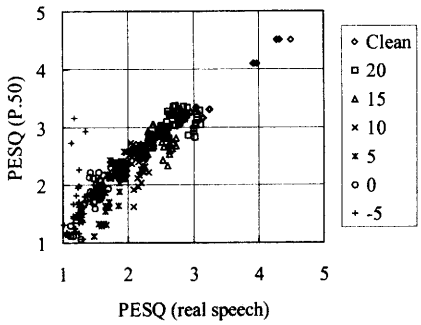


図7 実音声から算出したPESQ値と擬似音声から算出したPESQ値の関係

Fig. 7 Relationship between the PESQ score calculated from the real speech and the PESQ score calculated from the artificial voice.

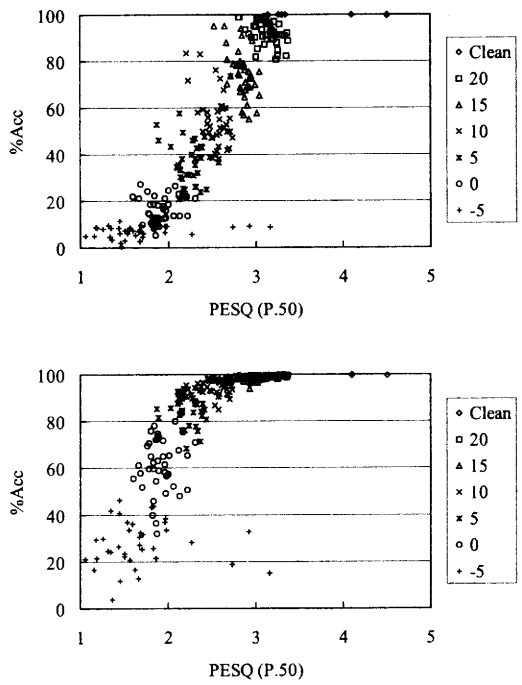


図8 単語正解精度と擬似音声から算出したPESQ値の関係（上段：Clean training，下段：Multicondition training）

Fig. 8 Relationship between the word accuracy and the PESQ score calculated from the artificial voice in the clean training (upper) and the multicondition training (lower).

ticondition trainingの場合は0.79である．

4. むすび

本稿では，音声の客観品質評価のためのひずみ尺度であるPESQ，ケプストラム距離，セグメンタルSNRに着目し，ひずみ値と認識性能の関係をAURORA-2Jを用いて調べた．その結果，単語正解精度との相関が最も強いのはPESQ値であることが明らかとなった．その一方で，PESQ値は，手法(G)によるひずみを適切に表していないことが分かった．今後は，この問題に対処するために，PESQ値の算出に用いる知覚・認知モデルを，音声認識の特性を考慮して改良する予定である．

また，ITU-T勧告P.50の擬似音声を用いて，擬似音声の適用可能性を調査した．その結果，実音声から算出したPESQ値と擬似音声から算出したPESQ値は概ね線形関係にあり，実音声の代わりに擬似音声を用いても，単語正解精度との強い相関が保たれていることが明らかとなった．今後は，音声の平均的な特性ではなく，音声認識タスク（語彙セット）の特性を持つように，擬似音声を改良する予定である．

謝辞

音声データや各手法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，総務省戦略的情報通信研究開発推進制度の研究委託による．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] 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, 2003.

[3] ITU-T Recommendation P.50, “Artificial voices,” Sep. 1999.

[4] 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.

[5] 北脇信彦編，“音のコミュニケーション工学，” コロナ社，1996.

[6] N. Kitawaki, T. Yamada, F. Asano, “Comparative assessment of test signals used for measuring residual echo characteristics,” IEICE Transactions on Communications, Vol. E86-B, No. 3, pp. 1102–1108, 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. Eurospeech2003, 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 SIG-Notes, SLP-44-3, pp. 13–18, 2002.