日本音響学会誌 67 巻 10 号（2011），pp. 476–481

解説 雑音抑圧音声の主観・客観品質評価法*

山田武志，牧野昭二，北脇信彦（筑波大学）**

43.72.+q

1. はじめに

駅の構内のような騒々しい環境で音声通話を行うと，音声が周囲の雑音に埋もれてしまい，正常に会話を行うことが困難となる。この問題を解決するためには，音声に重畳している雑音を抑圧することが有効である。しかし，雑音の音量感が低減する一方で，音声成分には歪が生じ，また抑圧しきれなかった雑音成分が聞き苦しい音に変形するという副作用が起こる。その程度は，抑圧対象となる雑音の特性や雑音抑圧の原理によって異なり，ユーザ体感品質に大きな影響を及ぼす。従って，雑音抑圧アルゴリズムの性能評価・性能比較はもちろんのこと，音声通話サービスの品質設計・品質管理を行うにあたって，雑音抑圧音声の品質を適切に評価する手法の確立が必要不可欠である。

音声の品質評価は，人間が実際に被評価信号を受聴し，その品質を主観的に判断することを基本とする。これを主観品質評価という。音声の総合的な品質の評価には，平均オピニオン評点（MOS：Mean Opinion Score）がよく用いられる。これは，多数の被験者が被評価信号に付けた評点の平均値として定義される。雑音抑圧音声のオピニオン評価試験法は，ITU-T 勧告 P.835 [1] により定められている。音声に雑音が重畳した結果，発話内容を聴き取ることさえ困難となる場合があるので，明瞭性という観点からの評価も重要である。明瞭性の評価には単語了解度を用いることが多い。単語了解度試験法は，聴き取った内容を記述する方式 [2]，聴き取った内容を複数の候補の中から選択する方式 [3] に大別できる。このように，雑音抑圧音声を対象とする主観品質評価法については既に確立されていると言えるであろう。

主観品質評価を実施するには，専用の設備・機器や多大な時間・労力を要するという問題が伴う。従って，被評価信号から品質に対応する特徴量を抽出し，その特徴量から主観品質を推定する手法，すなわち客観品質評価法が求められる。雑音抑圧音声を対象とする客観品質評価法については，ここ数年の間に大きな進展があり，主観品質を高い精度で推定できるようになってきている [4–10]。本解説では，雑音抑圧音声のオピニオン評価試験と単語了解度試験の実施例，及び我々がこれまでに開発してきた客観品質評価法について述べる [5, 6, 9]。

2. 雑音抑圧音声の総合品質の評価

2.1 オピニオン評価試験

雑音抑圧音声のオピニオン評価試験法は，ITU-T 勧告 P.835 [1] により定められている。被験者は一つの被評価信号（すなわち雑音抑圧音声）を 3 回受聴する。1 回目と 2 回目の受聴時には，音声成分のみに注目したときの音声品質，雑音成分のみに注目したときの雑音品質をそれぞれ評価する。そして，3 回目の受聴時には雑音抑圧音声全体の総合品質を評価する。このように，総合品質を評価する前に音声成分と雑音成分の双方に注目させることにより，どちらか一方の影響が強くなりすぎるのを防いでいる。

各品質は，表–1 の 5 段階絶対品質評価尺度を用いて評価される。P.835 ではカテゴリ（評定語）を英語により定めているため，英語以外の言語に適用する際にはカテゴリを適切に翻訳する必要がある。しかし，例えば「Good」と「Fair」の間の感覚的な距離と，それらの訳語の間の感覚的な距離が必ずしも同じではないという問題がある。また，主観的な評価であるが故に，言語のみならず，国や地域の文化的・経済的背景の影響を受けることが知られている（例えば，日本人は比較的厳しい

* Subjective and objective quality evaluation of noise-reduced speech.

** Takeshi Yamada, Shoji Makino and Nobuhiko Kitawaki (University of Tsukuba, Tsukuba, 305–8573)

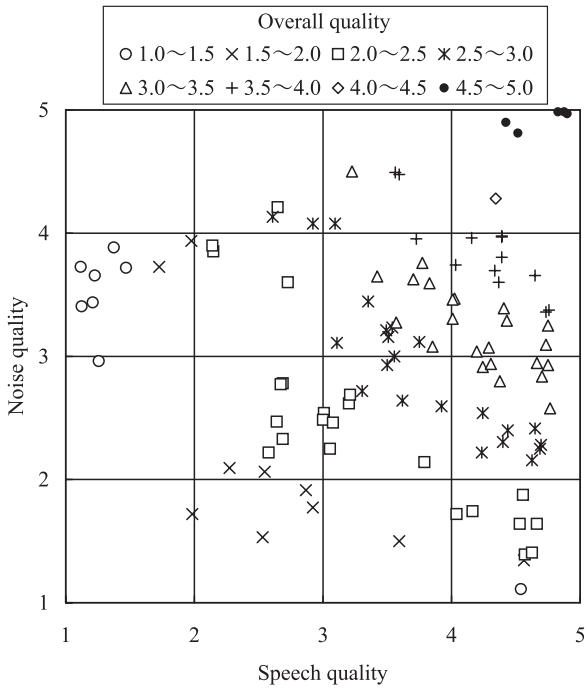
表–1　5 段階絶対品質評価尺度

	Speech quality	Noise quality	Overall quality
Score	Category	Category	Category
5	Not distorted	Not noticeable	Excellent
4	Slightly distorted	Slightly noticeable	Good
3	Somewhat distorted	Noticeable but not intrusive	Fair
2	Fairly distorted	Somewhat intrusive	Poor
1	Very distorted	Very intrusive	Bad

評価を行う傾向があると言われている）[11]。従って，複数の言語や国を対象としてオピニオン評価試験を実施する際には注意が必要である。

P.835 により定められるオピニオン評価試験の実施例について述べる。32 名の被験者が防音室において音声サンプルをヘッドホン受聴し，音声品質，雑音品質，総合品質を評価した。音声サンプルには男女各 2 名の計 4 発話を用いた。ここで，各発話は連続する二つの日本語文からなる。これらの音声サンプルに，電子協騒音データベース [12] の走行自動車内雑音，展示会場雑音，列車走行音，及び別途生成した白色雑音を計算機上で加算することにより，雑音重畳音声を生成した。SNR は Clean，20，15，10，5，0 dB の 6 種類である。雑音抑圧アルゴリズムには，EVRC（Enhanced Variable Rate Codec）の雑音抑圧 [13]，スペクトル減算と振幅抑圧の相互制御に基づく雑音抑圧 [14]，時間領域 SVD に基づく雑音抑圧 [15]，GMM に基づく雑音抑圧 [15]，雑音抑圧を行わない場合の 5 種類を用いた。なお，サンプリング周波数は 8 kHz である。

オピニオン評価試験の結果を図–1 に示す。ここで，横軸は音声品質，縦軸は雑音品質である。個々のマーカは，雑音抑圧アルゴリズム，雑音，SNR の組合せの一つに対する MOS（32 名の被験者が 4 個の音声サンプルに付けた評点の平均）を表しており，マーカの位置から音声品質と雑音品質，マーカの種類から総合品質を読み取ることができる。図–1 より，例えば音声品質と雑音品質の一方が顕著に低ければ総合品質を低く評価するなど，被験者は音声品質と雑音品質のバランスを考慮して総合品質を評価していることが分かる。これは，音声品質と雑音品質から総合品質を推定できることを示唆している。次節以降では，この結果に基づく客観品質評価法について述べる。



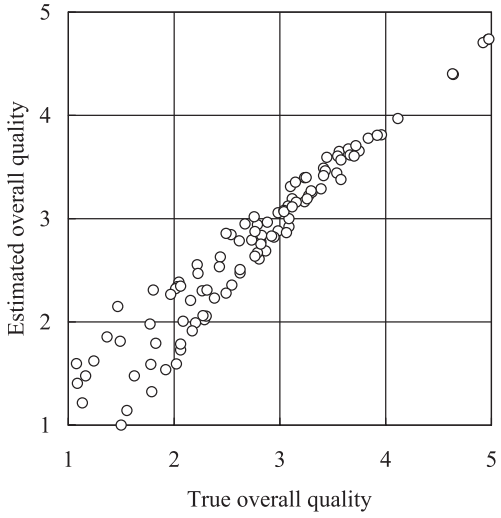
図–1　オピニオン評価試験の結果

2.2 総合品質推定モデル

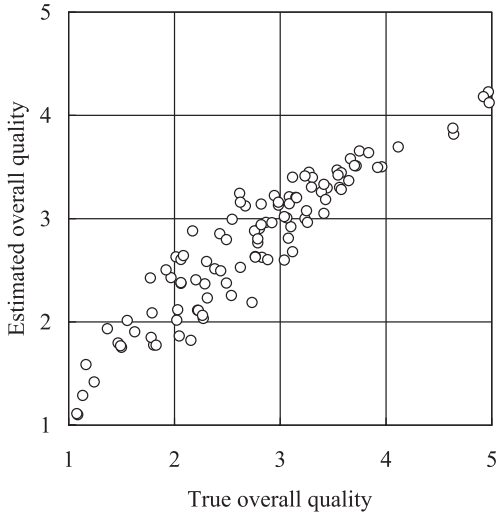
前節のオピニオン評価試験の結果に基づき，音声品質と雑音品質から総合品質を推定する総合品質推定モデルを次式により定めた。

$$\text{Overall quality} = 0.6303 \times \text{Speech quality} + 0.6125 \times \text{Noise quality} - 1.3917 \quad (1)$$

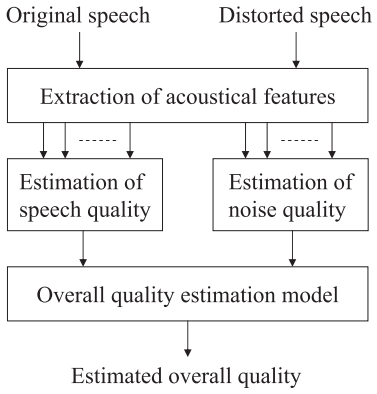
ここで，式中の数値は重回帰分析により決定している。図–1 の総合品質を，同じく図–1 の音声品質と雑音品質から推定した結果を図–2 に示す。ここで，横軸は真の総合品質，縦軸は推定した総合品質である。総合品質の低い領域においてばらつきがやや大きいものの，推定誤差を表す RMSE（Root Mean Square Error）は 0.26 であり，総じて高い精度で推定できていることが分かる（別に実施したオピニオン評価試験の結果に対しても同様であった [6]）。式 (1) を音声品質と雑音品質の 2 次



図–2　総合品質推定モデルによる総合品質の推定結果



図–4　FR 型客観品質評価法による総合品質の推定結果



図–3　総合品質推定モデルを用いた客観品質評価法

関数により表現することにより，推定誤差を更に低減することが可能となる [7]。

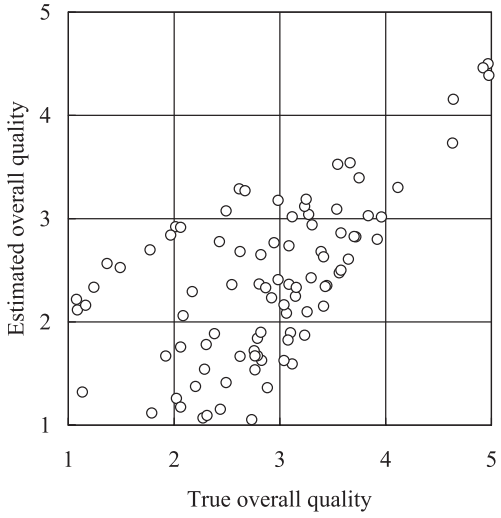
総合品質推定モデルを用いた客観品質評価法の基本構成を図–3 に示す。まず音声品質と雑音品質を個別に推定し，次に推定した音声品質と雑音品質から総合品質を推定する。このように P.835 における品質評価過程を模擬することにより，以下のような利点が生じる。

- 総合品質と特徴量の複雑な関係を直接モデル化する必要がない。音声品質と特徴量の関係，及び雑音品質と特徴量の関係をモデル化する必要があるが，これらは比較的容易であると期待できる。
- 特徴量の種類や数，求め方を定めていないため，特徴量の抽出に被評価信号とその原信号（雑音が重畳しておらず，かつ雑音抑圧の処理を行っていない元の音声信号）を用いるフルリファレンス（FR：Full-Reference）型客観品質評価法，被評価信号のみを用いるノンリファレンス（NR：Non-Reference）型客観品質評価法の双方を構築することができる。

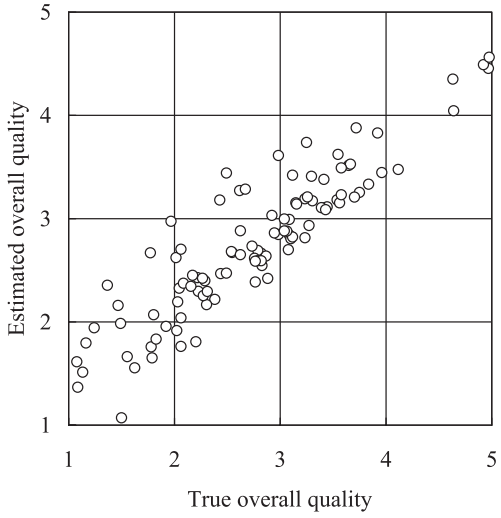
2.3 総合品質推定モデルを用いた FR 型客観品質評価法

総合品質推定モデルを用いた FR 型客観品質評価法について述べる。本手法では，まず音声区間と非音声区間のおのおのから信号の成分が増えるタイプの歪と信号の成分が減るタイプの歪を求める。ここで，歪尺度は耳内音圧スペクトル歪尺度である。これは，主に符号化音声を対象とする FR 型客観品質評価法である，ITU-T 勧告 P.862 [16]（以下では PESQ と呼ぶ）に採用されている。また，雑音抑圧音声の非音声区間から残留雑音の平均対数パワーを求める。次に，これら 5 種類の特徴量から音声品質と雑音品質をおのおの推定する。音声品質と雑音品質の各推定式は，上述した特徴量の 1 次結合として定義している。最後に，推定した音声品質と雑音品質を式 (1) に代入することにより，総合品質を推定する。

本手法により総合品質を推定した結果を図–4 に示す。ここで，横軸は真の総合品質，縦軸は推定した総合品質である。個々のマーカは，雑音抑圧アルゴリズム，雑音，SNR の組み合わせの一つに対応している。なお，音声品質と雑音品質の推定に用いた音声サンプルは，2.1 節のオピニオン評価試験に用いたものと同じである。RMSE は 0.33 であり，総合品質を良好な精度で推定できていることが分かる。参考までに，PESQ により総合品



図–5　PESQ による総合品質の推定結果



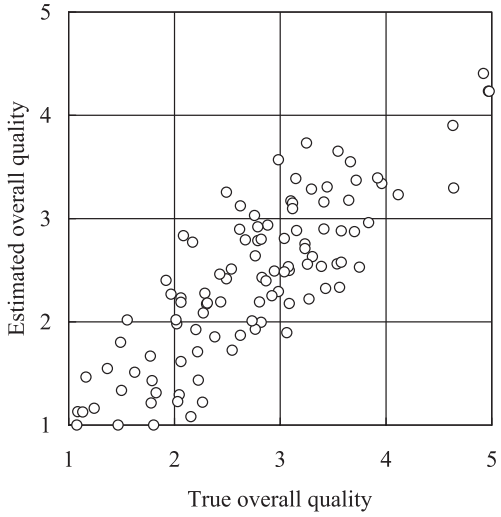
図–6　NR 型客観品質評価法による総合品質の推定結果

質を推定した結果を図–5 に示す。RMSE は 0.94 であり，PESQ は雑音抑圧音声の総合品質の推定には適していないことを確認できる。

2.4 総合品質推定モデルを用いた NR 型客観品質評価法

総合品質推定モデルを用いた NR 型客観品質評価法について述べる。本手法では，まず ITU-T 勧告 P.563 [17] において採用されている特徴量を抽出する。ここで，P.563 は，主に符号化音声を対象とする NR 型客観品質評価法である。次に，これらの特徴量のうち，Basic speech descriptors，及び Unnatural speech というクラスに属する 27 種類の特徴量から音声品質を，Noise analysis，及び Interruptions/Mutes というクラスに属する 24 種類の特徴量から雑音品質をおのおの推定する。音声品質と雑音品質の各推定式は，上述した特徴量の 1 次結合として定義している。最後に，推定した音声品質と雑音品質を式 (1) に代入することにより，総合品質を推定する。

2.3 節と同じ条件のもと，本手法により総合品質を推定した結果を図–6 に示す。RMSE は 0.37 であり，FR 型客観品質評価法と同程度の精度で推定できていることが分かる。参考までに，P.563 により総合品質を推定した結果を図–7 に示す。RMSE は 0.58 であり，本手法よりも推定誤差が大きいことが読み取れる。本手法と P.563 は全く同じ特徴量を用いていることから，これは総合品質推定モデルの有効性を示していると考えられる。



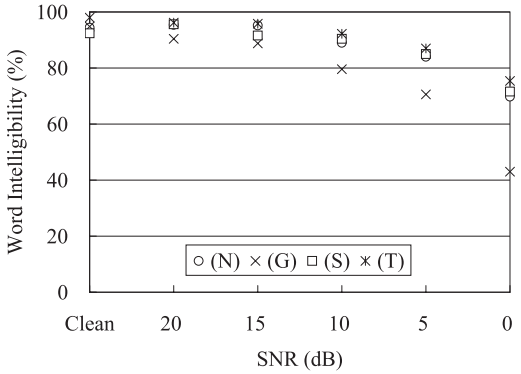
図–7　P.563 による総合品質の推定結果

3. 雑音抑圧音声の明瞭性の評価

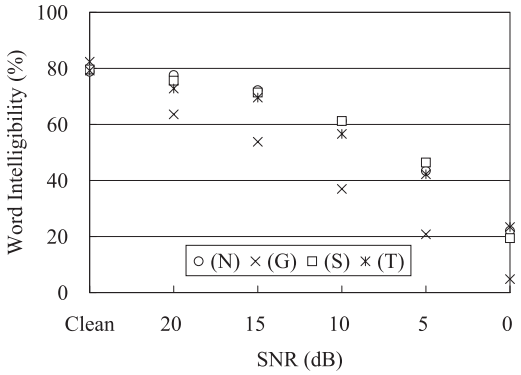
3.1 単語了解度試験

単語親密度と音韻バランスを考慮した単語リスト [2] を用いて実施した単語了解度試験について述べる。ここで，単語親密度とは単語に対する馴染みの度合いを指し，7.0（親密度高）から 1.0（親密度低）の数値により表される。単語リストは，単語親密度の四つのランク（F4：7.0～5.5，F3：5.5～4.0，F2：4.0～2.5，F1：2.5～1.0）ごとに構築されている。

20 名の被験者が防音室において音声サンプルをヘッドホン受聴し，聴き取った発話内容を仮名表記により記述した。音声サンプルには NTT・東北大親密度別単語了解度試験用音声データベース [18] を用いた。ここで，発話者は男性 1 名であり，発話



図–8　単語了解度試験の結果（F4）



図–9　単語了解度試験の結果（F1）

内容は 4 モーラの単語である。これらの音声サンプルに，AURORA-2J [19] の走行自動車内雑音と列車走行音を計算機上で加算することにより，雑音重畳音声を生成した。SNR は Clean，20，15，10，5，0 dB の 6 種類である。雑音抑圧アルゴリズムには，(S) SS-SMT 法 [20]，(T) 時間領域 SVD に基づく雑音抑圧 [15]，(G) GMM に基づく雑音抑圧 [15]，(N) 雑音抑圧を行わない場合の 4 種類を用いた。なお，サンプリング周波数は 8 kHz である。

走行自動車内雑音の場合の F4 と F1 に対する単語了解度試験の結果をそれぞれ図–8 と図–9 に示す。ここで，横軸は雑音抑圧前の SNR，縦軸は単語了解度（正答した単語の割合）である。SNR が下がるにつれて単語了解度が大きく低下すること，及び F1 における単語了解度は Clean のときでも 80%程度しかなく，単語親密度の影響はかなり大きいことが分かる。雑音抑圧を施すことにより単語了解度をむしろ低下させているケースがあるが，これは雑音の音量感を大幅に低減している一方で，音声成分に顕著な歪を与えていることによる。

3.2 単語了解度の客観推定法

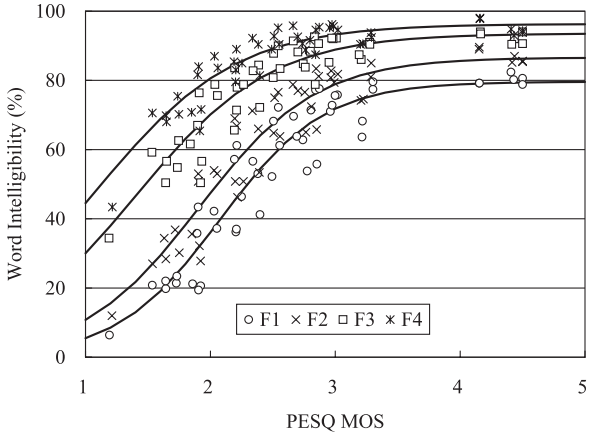
PESQ により推定した MOS（以下では PESQ MOS と呼ぶ）から単語了解度を客観推定する手法について述べる。本手法では単語了解度を次式により推定する

$$y = \frac{a}{1 + e^{-b(x-c)}} \tag{2}$$

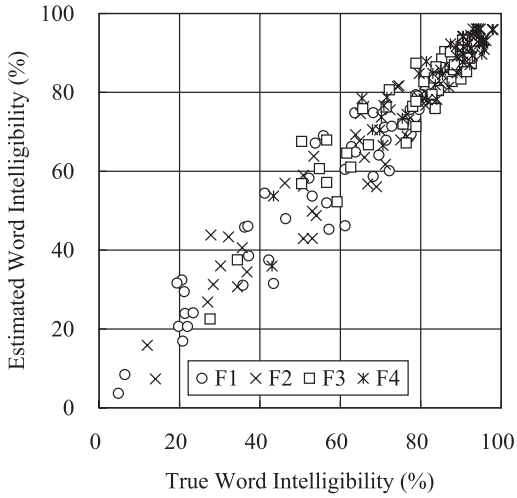
ここで，y は単語了解度，x は PESQ MOS である。a，b，c は単語了解度の推定誤差を最小にするように決定される。なお，音声認識の性能も PESQ MOS と式 (2) を用いて推定可能である [21]。

まず，前節の単語了解度試験の結果に基づき，F4～F1 のおのおのに対して求めた推定式を図–10 に示す。ここで，横軸は PESQ MOS，縦軸は単語了解度である。また，個々のマーカは，単語親密度のランク，雑音抑圧アルゴリズム，雑音，SNR の組合せの一つに対応している。図–10 から，単語了解度と PESQ MOS の関係を式 (2) により表すことが妥当であること，単語親密度のランクに応じた推定式が必要であることが分かる。

本手法により単語了解度を推定した結果を図–11



図–10　F4～F1 のおのおのに対する推定式



図–11　単語了解度の推定結果

に示す。ここで，横軸は真の単語了解度，縦軸は推定した単語了解度である。なお，単語了解度の推定に用いた音声サンプルは，3.1 節の単語了解度試験に用いたものと同じである。図–11 から，単語了解度を高精度に推定できていることが分かる。単語親密度のランクごとに求めた RMSE は 4.2〜7.0 であった。上述したように，PESQ は雑音抑圧音声の総合品質の推定には適していないにも関わらず，明瞭性を表す単語了解度の推定には適しているという事実は興味深く，今後詳しく調査する必要がある。

4. おわりに

本解説では，雑音抑圧音声のオピニオン評価試験と単語了解度試験の実施例，及び我々がこれまでに開発してきた雑音抑圧音声の客観品質評価法について述べた。今後の課題としては，推定精度を実用に耐え得るレベルにまで高めることが挙げられる。また，雑音抑圧音声の品質評価全般において，どのような雑音を選定すべきなのかが明らかではないことから，雑音選定基準を明確化することが必要不可欠である。

謝　辞

雑音抑圧アルゴルリズムのプログラムをご提供いただいた，北岡教英博士，藤本雅清博士に感謝する。

文　献

[1] ITU-T Rec. P.835, “Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm” (2003).
[2] 坂本修一, 鈴木陽一, 天野成昭, 小澤賢司, 近藤公久, 曽根敏夫, “親密度と音韻バランスを考慮した単語了解度試験用リストの構築,” 音響学会誌, 54, 842–849 (1998).
[3] 近藤和弘, 泉　良, 藤森雅也, 加賀　類, 中川清司, “二者択一型日本語音声了解度試験方法の検討,” 音響学会誌, 63, 196–205 (2007).
[4] N. Egi, H. Aoki and A. Takahashi, “Objective quality evaluation method for noise-reduced speech,” *IEICE Trans. Commun.*, **E91-B**, 1279–1286 (2008).
[5] 篠原佑基, 山田武志, 北脇信彦, 牧野昭二, “雑音抑圧音声の総合品質モデルを用いたフルリファレンス客観品質評価法の検討,” 第 7 回 QoS ワークショップ, pp. 40–41 (2009).
[6] T. Yamada, Y. Kasuya, Y. Shinohara and N. Kitawaki, “Non-reference objective quality evaluation for noise-reduced speech using overall quality estimation model,” *IEICE Trans. Commun.*, **E93-B**, 1367–1372 (2010).
[7] 藤田悠希, 山田武志, 牧野昭二, “雑音抑圧音声の客観品質評価に用いる総合品質推定モデルの改良,” 信学会総合大会, B-11-18, p. 447 (2011.3).
[8] ETSI EG 202 396-3 V1.3.1, “Speech and multimedia Transmission Quality (STQ); Speech quality performance in the presence of background noise Part 3: Background noise transmission — Objective test methods” (2011).
[9] T. Yamada, M. Kumakura and N. Kitawaki, “Objective estimation of word intelligibility for noise-reduced speech,” *IEICE Trans. Commun.*, **E91-B**, 4075–4077 (2008).
[10] K. Kondo and Y. Takano, “Estimation of two-to-one forced selection intelligibility scores by speech recognizers using noise-adapted models,” *Proc. Interspeech 2010*, pp. 302–305 (2010).
[11] Z. Cai, N. Kitawaki, T. Yamada and S. Makino, “Comparison of MOS evaluation characteristics for Chinese, Japanese, and English in IP telephony,” *Proc. Int. Universal Communication Symp.*, *IUCS2010*, pp. 111–114 (2010).
[12] 電子協騒音データベース, http://research.nii.ac.jp/src/list/detail.html#JEIDA-NOISE.
[13] 3GPP2 C.S0014-A Version 1.0, “Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems” (2004).
[14] 古田　訓, 高橋真哉, 中島邦男, “スペクトル減算と振幅抑圧の相互制御に基づく雑音抑圧法の検討,” 信学論, **J87-D-II**, 464–474 (2004).
[15] M. Fujimoto and 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. Eurospeech 2003*, pp. 1781–1784 (2003).
[16] 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” (2001).
[17] ITU-T Rec. P.563, “Single ended method for objective speech quality assessment in narrow-band telephony applications” (2004).
[18] NTT・東北大親密度別単語了解度試験用音声データベース, http://research.nii.ac.jp/src/list/detail.html#FW03.
[19] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto and T. Endo, “AURORA-2J: An evaluation framework for Japanese noisy speech recognition,” *IEICE Trans. Inf. Syst.*, **E88-D**, 535–544 (2005).
[20] 北岡教英, 赤堀一郎, 中川聖一, “スペクトルサブと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,” 信学論, **J83-D-II**, 500–509 (2000).
[21] T. Yamada, M. Kumakura and N. Kitawaki, “Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice,” *IEEE Trans. Audio Speech Lang. Process.*, 14, 2006–2013 (2006).