総合品質と明瞭性の客観推定に基づく スペクトルサブトラクションの減算係数の最適化*

☆中里徹, 山田武志, 宮部滋樹, 牧野昭二, 北脇信彦（筑波大）

1 はじめに

近年，音声通信技術の発達により，携帯電話に代表されるモバイル音声通話や，テレビ会議・IP 電話のようなハンズフリー通話によるコミュニケーションが普及している．しかし，これらの通話形態においては，雑音が重畳しやすいために通話品質が著しく低下してしまうという欠点がある．この問題を解決する方法の一つとして，音声に重畳している雑音成分を取り除く雑音抑圧処理がある．雑音抑圧処理の原理は様々であるものの，雑音感の低減と音声の品質保持が基本方針となっている [1].

雑音抑圧を行った音声の品質を人間が評価する際には，総合品質と明瞭性がよく取り上げられる．総合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．これは多数の評定者が被評価音声に付けた評点の平均値として定義される．また，明瞭性の評価には了解度がよく用いられる．これは発話された内容が受聴者に正しく伝わった割合で定義される．中でも，単語了解度（WI：Word Intelligibility）が広く使われている．

Fig. 1 はある条件で雑音抑圧処理を行った結果得られた MOS と WI それぞれの改善を示している．横軸は総合品質の尺度である MOS，縦軸は明瞭性の尺度である WI を表す．図中の○は雑音抑圧処理を行う前の雑音重畳音声の MOS と WI，□は 3 つの異なる雑音抑圧手法をこの音声に適用したときの MOS と WI をそれぞれ示している．Fig. 1 から，雑音抑圧手法毎に改善傾向（矢印の向きと長さ）が異なっていることが分かる．ここで注目すべき点は，それぞれの改善傾向は明確に意図して得られたものではないことである．

そこで本稿では，MOS と WI の改善傾向を制御可能な雑音抑圧手法を提案する．提案手法により，対象とする環境やシーンに適した雑音抑圧音声を得ることが可能になる．例えば，オフィス内などの雑音がそれほど大きくない環境での通話では特に MOS を向上させること，また工事現場などの雑音が非常に大きい環境での通話では特に WI を向上させることが可能となる．このように，ユーザは要望を指定することでそれに合った最適な音声を得ることができる．

提案手法では，まず雑音重畳音声を入力し，複数の雑音抑圧手法で処理する．そして，各雑音抑圧音声の MOS と WI を推定し，ユーザの要望に合った最適な雑音抑圧音声を出力する．各雑音抑圧音声の MOS と WI の推定にはノンリファレンス型客観品質評価法 [2] を適用する．最後に，実験により提案手法の有効性を示す.

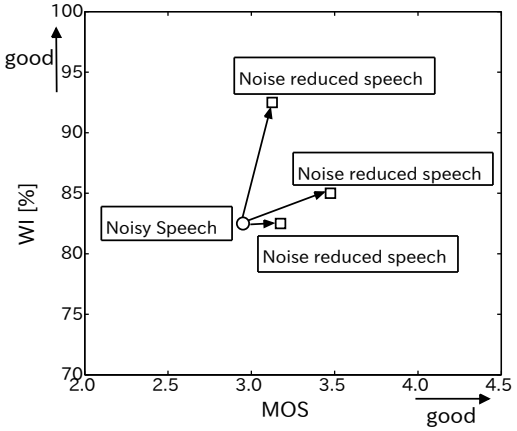

Fig. 1 Tendency of MOS and WI improvement by three different noise reduction algorithms.

2 提案手法

提案手法の処理フローを Fig. 2 に示す．まず，雑音重畳音声を入力し，複数の雑音抑圧手法で処理する．この複数の雑音抑圧手法は，原理の異なる雑音抑圧手法，内部パラメータを様々に設定した同じ雑音抑圧手法，またはその組み合わせのいずれでも良い．本稿では，時間方向スムージングを用いたスペクトルサブトラクション法 [3]（以下，SS-SMT 法と呼ぶ）の減算係数 α を様々に設定したものを雑音抑圧手法として用いる．SS-SMT 法では，次式に示すように推定した雑音のパワースペクトルを入力信号のパワースペクトルから減算する．

$$|\hat{H}(\omega,t)| = \overline{|X(\omega,t)|^2} - \alpha|\hat{W}(\omega,t)|^2 \quad (1)$$

$$\hat{S}(\omega,t) = \begin{cases} |\hat{H}(\omega,t)|e^{j\arg X(\omega,t)} & (|\hat{H}(\omega,t)| > \epsilon) \\ \epsilon & (otherwise) \end{cases} \quad (2)$$

$$\overline{|X(\omega,t)|^2} = \sum_{i=0}^{2} |X(\omega,t-i)|^2 \quad (3)$$

$\hat{S}(\omega,t)$ は出力音声，$X(\omega,t)$ は入力音声，$\hat{W}(\omega,t)$ は推定雑音，ω は周波数インデックス，t は時間フレームインデックス，ϵ は閾値である．ここで，式 (1) 中の減算係数 α により減算の強さを調整するため，α の値により MOS と WI の改善傾向が変動する．そこ

*Subtraction coefficient optimization of spectral subtraction based on objective estimation of overall quality and intelligibility, by Toru Nakazato, Takeshi Yamada, Shigeki Miyabe, Shoji Makino, Nobuhiko Kitawaki (University of Tsukuba).

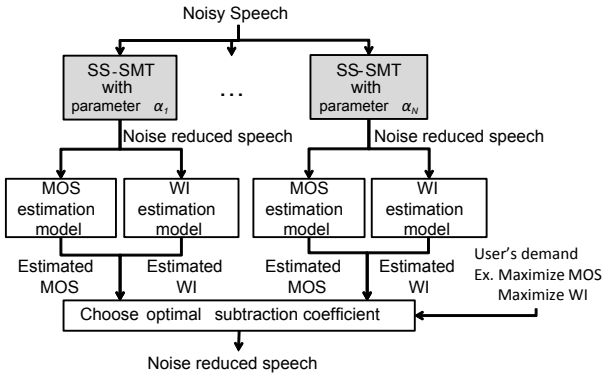

Fig. 2　Process flow of the proposed method.

で，提案手法では，$\alpha = \alpha_1, \alpha_2, \cdots, \alpha_N$ の N 通りの SS-SMT 法を雑音抑圧手法として用いる．

次に，各雑音抑圧手法から出力される雑音抑圧音声それぞれの MOS，WI を推定する．提案手法では MOS と WI の推定にノンリファレンス型客観品質評価法 [2] を用いる．これは，ITU-T 勧告 P.563[4] の内部で使用されている 44 種類の物理的特徴量を雑音抑圧音声のみから抽出し，線形重回帰で MOS や WI を推定する手法である．抽出される特徴量には，発話者の声道特性，背景 SNR，無音長などがある．本稿では，より頑健な推定をするため，線形重回帰から SVR（Support Vector Regression）[5] に推定モデルを変更する．

最後に，推定モデルにより得られた推定 MOS，推定 WI に基づいてユーザの要望に合った最適な減算係数の選択を行い，雑音抑圧音声を出力する．本稿では，減算係数の選択を次式のように行う．

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}\, \mathrm{MOS}_\alpha \tag{4}$$

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}\, \mathrm{WI}_\alpha \tag{5}$$

ここで，$\hat{\alpha}$ は選択する減算係数，MOS_α は減算係数 α のときの推定 MOS，WI_α は減算係数 α のときの推定 WI を表す．式（4）の $\hat{\alpha}$ を用いることで，推定 MOS が最大となる減算係数の選択ができる．また，式（5）の $\hat{\alpha}$ を用いることで，推定 WI が最大となる減算係数の選択ができる．

3　MOS・WI 推定モデルの学習

3.1　主観評価試験

3.1.1　試験条件

被評価音声の条件を Table. 1 に示す．サンプルソースとして，親密度別単語了解度試験用音声データベース [6][7] に収録されている 4 モーラの単語群を用いた．発話者は男性 2 名，女性 2 名の計 4 名である．収録されている各単語は，単語の馴染みの程度を表す親密度別に 4 グループ (1.0〜2.5，2.5〜4.0，4.0〜5.5，5.5〜7.0) に分類されている．本稿では親密度が最も高い 5.5〜7.0 の単語群 1000 語を用いた．この音声サンプルに AURORA-2J[8] に収録されている 4 種類の雑音を，15, 10, 5, 0[dB] の 4 種類の SNR で音声に重畳する．音声サンプルに雑音を重畳する際，雑音データから音声サンプルの長さの雑音を切り出し重畳した．ここで，切り出す時間区間はランダムに設定した．雑音抑圧手法として，減算係数の異なる SS-SMT 法 [3] を 6 種と雑音抑圧を行わない場合 1 種の計 7 種を用いた．SS-SMT 法の減算係数 α は，予備実験により決定した．予備実験では，減算係数 α が 0.1, 0.2, …, 1.5, 2.0, 3.0, …,10.0 のときの雑音抑圧音声を受聴し，MOS と WI の測定を行った．それにより得られた各評価値の差がほぼない冗長な減算係数 α を除外し，$\alpha = 0.5, 1.0, 1.5, 2.0, 4.0, 8.0$ とした．被験者は，重畳雑音，SNR，減算係数の組毎に，4 単語を受聴する．ここで，各組で発話内容は異なるようにした．また，被験者は同じ単語は 1 度しか受聴しないようにした．

Table 1　Speech samples used in the experiment.

データソース	親密度別単語了解度試験用 音声データベース [6][7]
発話者	男女各 2 名の計 4 名
発話内容	4 モーラの単語
雑音	Subway, Car, Babble, Train[8]
チャネル	G.712
標本化周波数	8kHz
SNR(dB)	Clean，15，10，5，0
雑音抑圧手法	None(適用なし) SS-SMT 法 [3]：$\alpha = 0.5, 1, 1.5, 2, 4, 8$
一人あたりの サンプル数	452 音声 (4 音声 ×7 手法 ×4 雑音 ×4SNR+4Clean)

Table 2　Rating scale used in the experiment.

評点	音声品質	雑音品質	総合品質
5	歪んでいない	気にならない	非常に良い
4	わずかに歪んでいる	わずかに気になる	良い
3	多少歪んでいる	多少気になる	まあ良い
2	かなり歪んでいる	かなり気になる	悪い
1	非常に歪んでいる	非常に気になる	非常に悪い

次に，MOS 試験について述べる．被験者は 14 名で，防音室でヘッドホンにより音声の受聴を行った．MOS 試験は，ITU-T 勧告 P.835[9] によって勧告されているように 2 文章を約 1 秒の無音区間で連結した音声サンプルで行う必要がある．しかし，本稿では MOS 試験と WI 試験で同じ音声データセットを使用するため，4 モーラの単語を用いて MOS 試験を行う．被験者一人が評価する音声は Table. 1 で示した 452 音声である．被験者は同じ音声を 3 回受聴し，音声品質，雑音品質，総合品質の 3 項目の評価を行う．評価尺度として Table. 2 に示す 5 段階絶対品質評価尺度を用いる．重畳雑音，SNR，減算係数の組毎に，14 名が異なる 4 単語を評価しているため，56 個の評価値が得られる．これらを平均することにより各組の

MOS を算出する．

最後に，WI 試験について述べる．WI 試験では，被験者は聞き取った単語の読みをキーボードで入力した．被験者は MOS 試験と同じ 14 名である．被験者一人が評価する音声は Table. 1 で示した 452 音声である．MOS 試験と同様に，重畳雑音，SNR，減算係数の組毎に，14 名が異なる 4 単語を評価しているため，56 個の解答単語が得られる．これらを用いて各組の WI を算出する．

3.1.2　主観評価試験の結果

Car 雑音に対する主観評価試験の結果を Fig. 3 に示す．縦軸は主観 WI，横軸は主観 MOS を示している．マーカーの違いは SS-SMT 法の減算係数の違いを表し，点線の領域は雑音抑圧処理前の SNR が同じであることを示している．

Fig. 3 から，各 SNR において MOS が最大となる減算係数，WI が最大となる減算係数が異なることが分かる．他の 3 種類の雑音についても Car 雑音と同様の結果が得られた．このことから，雑音環境により減算係数の最適化を行う必要があると言える．

3.2　推定モデルの学習方法

提案手法では，MOS と WI の推定に SVR を用いる．推定モデルの学習には，教師データとして 3.1.2 節で述べた主観評価試験の MOS と WI，特徴量には 2 章で述べた 44 種類の物理的特徴量を用いる．なお，SVR ツールとして LIBSVM[10] を用いた．

本来 MOS と WI は一つの被評価音声サンプルに対する多人数の評価値・解答単語を用いて算出されるが，先述の主観評価試験では 1 音声サンプルに対して一人の評価値・解答単語のみが得られている．そこで本稿では，一人の評価値をその音声サンプルの MOS，WI として用いることにする．ここで，被験者一人の WI を使用した場合，WI が 0%か 100%という荒い値になってしまう．そこで本稿では，音節了解度を算出することにした．このように WI 推定モデルでは実際には音節了解度を推定することになるが，便宜上 WI 推定モデルと呼ぶことにする．

4　提案手法の有効性の評価

4.1　実験条件

本実験では，雑音クローズドテストと雑音オープンテストを行う．雑音クローズドテストでは，推定モデルの学習と提案手法の評価に Table. 1 で示した 6328 音声サンプル（452 音声サンプル ×14 名）全てを用いる．雑音オープンテストでは，Table. 1 で示した 4 雑音で音声サンプルを分割する 4-fold cross-varidation を行った．これは，1 雑音が重畳した音声サンプル群 1568 音声をテストデータとし，残りの 3 雑音が重畳した音声サンプル群と Clean 音声サンプル群の計 4760 音声を学習データとするものである．これを 4 雑音それぞれがテストデータとなるよう 4 回検証を行った．

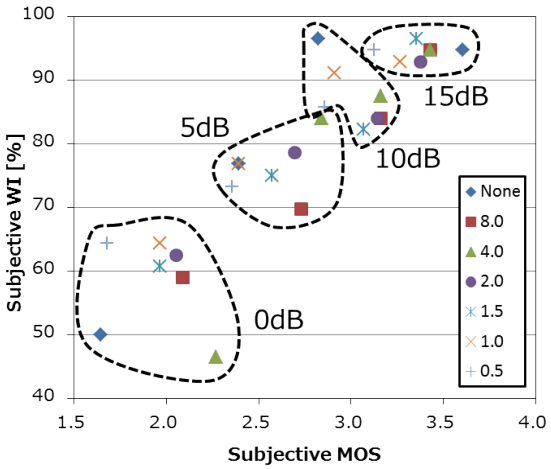

Fig. 3　Relationship between Subjective MOS and Subjective WI for the Car noise.

提案手法の最適化基準として，推定 MOS が最大となる減算係数の選択（以下，提案手法（MOS 最大）と呼ぶ）と，推定 WI が最大となる減算係数の選択（以下，提案手法（WI 最大）と呼ぶ）を用いる．

提案手法の MOS と WI の算出方法について述べる．重畳雑音，SNR の組毎に 56 個の雑音重畳音声サンプルがある．この音声サンプル一つ一つを提案手法に入力として与え，減算係数 α を選択して雑音抑圧音声を出力する．3.2 節で述べたように，出力された音声それぞれについて 1 被験者の評価値・解答単語がある．56 音声サンプルの評価値を平均することで，MOS を算出する．また，56 音声サンプルの解答単語を用いて WI を算出する．

4.2　実験結果

4.2.1　雑音クローズドテスト

Car 雑音に対する主観 MOS と主観 WI の関係を Fig. 4 に示す．Conventional（α）は減算係数を α に固定したときの従来手法を表し，Proposed（maxMOS），Proposed（maxWI）は提案手法（MOS 最大）と提案手法（WI 最大）を表している．

まず提案手法（MOS 最大）について考察する．提案手法（MOS 最大）では，SNR 5dB，10dB，15dB のときに従来手法よりも高い主観 MOS が得られた．他の重畳雑音も含めると，重畳雑音，SNR の 16 個の組み合わせのうち，14 組で従来手法よりも高い主観 MOS が得られた．残りの 2 組は，Car 雑音と Babble 雑音の SNR 0dB のときである．これは，SNR が非常に低いときは MOS の推定精度が十分ではないことを示唆している．なお，提案手法の主観 MOS が従来手法よりも高くなっているのは，従来手法では 56 個の雑音重畳音声に対して共通の減算係数を用いているのに対して，提案手法では雑音重畳音声毎に最適な減算係数を設定していることによる．56 音声のそれぞれに重畳している雑音は異なるため（雑音の種類は同じであるが時間区間は異なる），最適な減算係

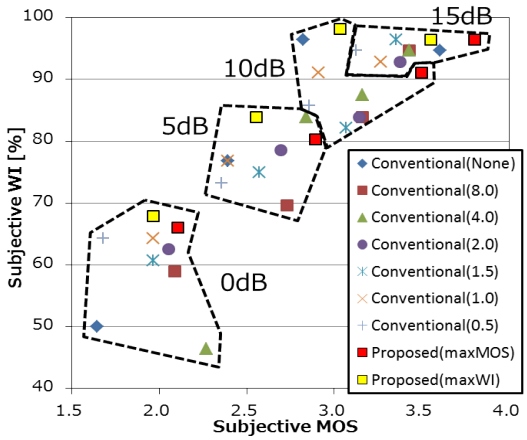

Fig. 4　Relationship between Subjective MOS and Subjective WI by the proposed method (Noise-closed test, Car noise).

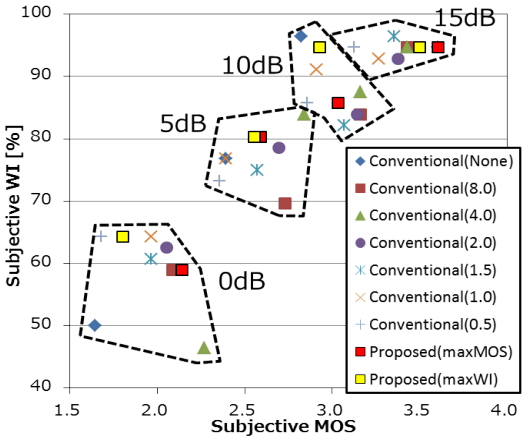

Fig. 5　Relationship between Subjective MOS and Subjective WI by the proposed method (Noise-open test, Car noise).

数は雑音重畳音声毎に異なる．

次に提案手法（WI 最大）について考察する．提案手法（WI 最大）では，全ての SNR の組において従来手法よりも高い主観 WI が得られた．他の重畳雑音も含めると，重畳雑音，SNR の 16 個の組み合わせのうち，14 組で従来手法よりも高い主観 WI が得られた．残りの 2 組は，Subway 雑音の SNR 0dB のときと Train 雑音の SNR 5dB のときである．

4.2.2　雑音オープンテスト

Car 雑音に対する主観 MOS と主観 WI の関係を Fig. 5 に示す．提案手法（MOS 最大）では，SNR 15dB のときに従来手法よりも高い主観 MOS が得られた．他の重畳雑音も含めると，重畳雑音，SNR の 16 個の組み合わせのうち，2 組で従来手法よりも高い主観 MOS が得られた．また，提案手法（WI 最大）では，SNR 0dB のときに従来手法よりも高い主観 WI が得られた．他の重畳雑音も含めると，重畳雑音，SNR の 16 個の組み合わせのうち，3 組で従来手法よりも高い主観 WI が得られた．

雑音クローズドテストと比較すると，提案手法の主観 MOS・主観 WI が従来手法の主観 MOS・主観 WI より高くなる組が少なくなっている．これは，過学習が起きていることが原因と考えられる．ここで，従来手法の中で最大となる主観 MOS・主観 WI と，提案手法の主観 MOS・主観 WI の差の平均を求めて比較する．その結果，提案手法（MOS 最大）では約–0.12，提案手法（WI 最大）では約–3.58%であった．これにより各提案手法は，減算係数固定の従来手法のうち，最も主観 MOS・主観 WI が高いものに近い性能を得られていることが確認された．

5　おわりに

本稿では，総合品質と明瞭性の改善傾向を制御可能な雑音抑圧手法の実現を目的とし，SS-SMT 法の減算係数の最適化を行う手法を提案した．実験により，提案手法が有効であることを確認した．今後は，未知の雑音に対する頑健性を高める必要がある．

謝辞　雑音抑圧手法のプログラムをご提供頂いた，北岡教英氏に感謝する．

参考文献

[1] 吉岡拓也, 中谷智広, “確率モデルを用いた音声強調:雑音抑圧, 音源分離, 残響除去, 統合技術及びその応用,” 日本音響学会誌, Vol. 68, No. 11, pp. 572–577, 2012.

[2] T.Yamada, Y.Kasuya, Y.Shinohara, N.Kitawaki, “Non-reference objective quality evaluation for noise-reduced speech using overall quality estimation model,” IEICE Trans. Communications, Vol. E93-B, No. 6, pp. 1367–1372, 2010.

[3] 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 D-II, Vol. J83-D-II, No. 2, pp. 500–508, 2000.

[4] ITU-T Rec. P.563, “Single ended method for objective speech quality assessment in narrowband telephony applications,” 2004.

[5] Harris Drucker, Chris J. C. Burges, Linda Kaufman, Alex Smola, Vladimir Vapnik, “Support vector regression machines,” Advances in neural information processing systems, Vol. 9, pp. 155–161, 1997.

[6] 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単語リストの構築,” 日本音響学会誌, Vol. 54, No. 12, pp. 842–849, 1998.

[7] NTT-AT, “親密度別単語了解度試験用音声データベース,” http://www.ntt-at.co.jp/product/wordtest/.

[8] S.Nakamura, et al., “AURORA-2J: An evaluation framework for Japanese noisy speech recognition,” IEICE Trans. Information and Systems, Vol. E88-D, No. 3, pp. 535–544, 2005.

[9] ITU-T Rec. P.835, “Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm,” 2003.

[10] Chih-Chung Chang, Chih-Jen Lin, “LIBSVM: a library for support vector machines,” ACM Trans. Intelligent Systems and Technology (TIST), Vol. 2, No. 3, pp. 1–27, 2011.