社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

カートシスマッチングに基づく 低ミュージカルノイズDNN音声強調の評価

溝口　聡† 齋藤　佑樹† 高道慎之介† 猿渡　洋†

† 東京大学大学院情報理工学系研究科　〒113–8656 東京都文京区本郷 7–3–1
E-mail: †{satoshi_mizoguchi,yuki_saito,shinnosuke_takamichi,hiroshi_saruwatari}@ipc.i.u-tokyo.ac.jp

あらまし　本稿では，DNN 音声強調にカートシスマッチングを適用し，ミュージカルノイズの発生を低減させる方法を提案する．非線形信号処理によって発生する人工的な歪みをミュージカルノイズと呼び，これは聴覚的不愉快さをもたらすことが知られている．また，ミュージカルノイズの発生量は強調前後のカートシスの上昇と大きな相関があることが知られている．DNN 音声強調は，DNN の豊かな表現力によって強力な雑音抑圧性能を誇るが，ミュージカルノイズの発生について考慮していない．本稿では，DNN 音声強調にカートシスの上昇を抑えるような正則化，すなわちカートシスマッチングを行うことによって，雑音抑圧性能や音声歪み発生量を維持したまま，低ミュージカルノイズな音声強調を実現する手法を提案する．また，音声強調実験の結果に対して客観評価を行い，提案手法の有効性を示す．

キーワード　音声強調，ミュージカルノイズ，カートシスマッチング，深層学習

Evaluation of DNN-based Low-Musical-Noise Speech Enhancement Using Kurtosis Matching

Satoshi MIZOGUCHI†, Yuki SAITO†, Shinnosuke TAKAMICHI†, and Hiroshi SARUWATARI†

† Graduate School of Information Science and Technology, The University of Tokyo　7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–8656 Japan
E-mail: †{satoshi_mizoguchi,yuki_saito,shinnosuke_takamichi,hiroshi_saruwatari}@ipc.i.u-tokyo.ac.jp

Abstract　This paper proposes DNN-based speech enhancement with low musical noise by kurtosis matching. Musical noise, artifacts generated by nonlinear signal processing, causes a negative effect on the auditory impression. Quantity of the generated musical noise is significantly correlated with increase in kurtosis from observed signal to enhanced signal. Although soft-mask-based DNN speech enhancement has a high performance on noise reduction thanks to rich power of expression of DNN, it does not consider generation of musical noise. This paper proposes low-musical-noise speech enhancement without degrading noise-reduction-rate and generating significant speech distortion by applying kurtosis matching, which is regularization to prevent kurtosis from increasing, to DNN-based speech enhancement. We give objective evaluation of the enhanced speech signal to demonstrate the efficiency of the proposed method.

Key words　speech enhancement, musical noise, kurtosis matching, deep learning

1. はじめに

音声通信において，音声信号に重畳される環境雑音は，話者間のコミュニケーションを阻害する要因として望ましくないものである．特に，単一のマイクロフォンしか用いることができない状況でのハンズフリーな音声通信では，話者とマイクの位置関係の特定が難しく，単一チャネル信号処理による音声強調技術が必須である．スペクトル減算法（Spectral Subtraction: SS）[1] やウィーナフィルタ（Wiener Filtering: WF）に代表される従来の単一チャネルの音声強調技術では，非線形な信号処理に由来する人工的な歪みが生じ，聴覚的な印象を大きく損なうことが知られている．この人工的な歪みをミュージカルノイズ（musical noise）と呼ぶ [2], [3]．

近年，ミュージカルノイズを考慮した，事前学習不要の音声強調技術が盛んに研究されている [2]〜[6]．ミュージカルノイズの知覚の度合いと大きな相関を持つ数値としては，音声強調

前後での非音声区間のカートシス比 [6] がよく知られ，これをミュージカルノイズの発生量の指標とすることが多い．SS や WF において，雑音のパワーがガンマ分布に従うと仮定したときに，幾つかの近似のもとで音声強調前後の雑音のパワーのカートシス比が不変となるパラメータが発見されている [4], [5]．このようなパラメータが与えるミュージカルノイズが発生しない状態をミュージカルノイズフリーと呼ぶ．ミュージカルノイズフリーな SS や WF は雑音抑圧としての性能が低いため，反復して用いることで所望の雑音抑圧率（Noise Reduction Rate: NRR）を達成する．

一方，近年は，事前学習を必要とするが強力な手段として，ディープニューラルネットワーク（Deep Neural Network: DNN）による音声強調技術も多数提案されている（e.g., [7]〜[10]）．特に，[10] はソフトマスクベースの DNN 音声強調であり，これらは DNN の高い表現能力を利用した強力な雑音抑圧性能を誇る有力な手法であるが，強調処理後の信号にミュージカルノイズを発生させないという保証はない．

これに対し本稿では，カートシスの乖離度（Kurtosis Discrepancy: KD）による正則化（カートシスマッチング; kurtosis matching）を導入することで，ミュージカルノイズ発生量の小さい DNN 音声強調法を提案する．提案手法における DNN は，観測された音声信号の振幅スペクトログラムを入力とし，ソフトマスクを出力とする．この DNN は，通常用いられる，クリーンな音声のスペクトログラムとの誤差の最小化に加え，マスクにより得られた非音声区間における音声強調後のカートシスが，音声強調前の同区間のカートシスと一致するように学習される．提案手法では，カートシスマッチングによりミュージカルノイズの発生を抑えるため，主観的音質の高い音声強調が可能になると期待される．実験的評価により，提案手法が雑音抑圧性能を保持しつつ，カートシスの上昇を避けられることを示す．また，主観的評価によって，提案手法による音声強調における聴覚品質の向上を確認する．

2. ソフトマスクによる DNN 音声強調 [10]

観測信号の短時間フーリエ変換によって得られた振幅スペクトログラムを X とする．これを入力とする DNN のパラメータを θ とし，その出力を $S = f(X;\theta)$ とおく．また，ターゲットであるクリーンな音声信号の振幅スペクトログラムを Y とする．このとき，損失関数を

$$L_0(X,Y;\theta) := \| S \circ X - Y \|_{1,1} \tag{1}$$

によって定義する．ただし，$\|\cdot\|_{1,1}$ は $L_{1,1}$ ノルムであり，行列の各成分の絶対値を表すものである．また，$\circ$ は行列のアダマール積であり，要素ごとに積をとるものである．この損失関数の訓練データに関する標本期待値について最小化を行う．すなわち，

$$\hat{\theta} = \arg\max_{\theta} \mathrm{E}[L_0(X,Y;\theta)] \tag{2}$$

とする．このようにして得られる θ は，観測信号 X の雑音を抑圧し，音声信号を抽出するマスクを生成するように学習される．最後に，DNN より出力されるソフトマスクをかけた観測信号の振幅スペクトログラム $S \circ X$ に，観測信号の位相スペクトログラムをかけて短時間逆フーリエ変換を行うことで，所望の強調音声を推定する．

3. 提 案 手 法

3.1 動　　機

従来の DNN 音声強調においては，非線形処理によるミュージカルノイズの発生が考慮されていない．そこで，ソフトマスク推定に基づく DNN 音声強調の強力な雑音抑圧を達成し，尚且つミュージカルノイズの発生が少ないような音声強調法を提案する．具体的には，先述のソフトマスクによる DNN 雑音抑圧において，損失関数にカートシスマッチングを実現するような正則化項を加えることでミュージカルノイズの発生を低減させる．提案手法の概要は図 1 である．

3.2 カートシスマッチングを考慮した DNN 学習

3.2.1 カートシス（kurtosis）

一変数確率変数 W は正実数値をとり，確率分布 $p(w)$ に従うとする．その n 次モーメントを

$$\mu_n = \int_0^\infty w^n p(w)\mathrm{d}w \tag{3}$$

で定義する．また，確率変数 W のカートシスとは，

$$K_W := \frac{\mu_4}{\mu_2^2} \tag{4}$$

によって定義する．この定義は一般的な統計学における平均周りのカートシスとは異なることに注意されたい．このカートシスは確率分布の裾の重さ，すなわち，外れ値の多さを表している．また，標本カートシスは式 (4) のモンテカルロ積分より

$$\kappa_W = \frac{1}{T} \frac{\sum_{t=1}^{T} W_t^4}{\left(\sum_{t=1}^{T} W_t^2\right)^2} \tag{5}$$

によって計算できる．

音声強調前後におけるパワースペクトログラムのカートシスの上昇は，ミュージカルノイズの発生と強い相関があることが知られている [6]．本稿では，振幅スペクトログラムのカートシスの上昇について考えるが，カートシスが外れ値の多さを反映する統計量であることから，ミュージカルノイズ発生量と相関があると考えて議論する．

3.2.2 カートシスの乖離度（Kurtosis Discrepancy: KD

本稿では，DNN の損失関数にカートシスの変化が発生しないような項を組み込むために，KD を定義する．本定義は kernelized discrepancy を損失とする Generative Moment Matching Networks（GMMN）[11] に着想を得たものであるが，本定義における discrepancy はカーネル化されておらず，その点で意味が異なる．本稿では，以降に述べるように，非音声区間の振幅スペクトログラムにおける周波数サブバンド毎の KD を用いる．

観測信号の振幅スペクトログラム X の行列成分を $X_{k,t}$，強調後の音声信号の振幅スペクトログラム Z の行列成分を $Z_{k,t} = S_{k,t} X_{k,t}$（ただし，$S_{k,t}$ は S の行列成分）とする．ここで，$k \quad K := \{0, \cdots, K\}$ は周波数サブバンドのインデックス，$t \quad T := \{1, \cdots, T\}$ は時間フレームのインデックスである．また，周波数サブバンドのインデックス集合の分割を

$K_i := \{k_i, \cdots, k_{i+1} - 1\}$（ただし，$i = 1, \cdots, N-1$，$k_1 = 0$，$k_N = K + 1$）とし，非音声区間の時間フレームインデックスの集合を T　　T とする．と書くことにする．このとき，非音声区間の KD を

$$\mathrm{KD}(X, Z) := \sum_{i=1}^{N} \alpha_i \; K_k^{t\ T'}{}_{K_i}(X_{k,t}) - K_k^{t\ T'}{}_{K_i}(Z_{k,t}) \qquad (6)$$

で定義する．ここで，$K_k^{t\ T'}{}_{K_i}(X_{k,t})$ は，行列 X の成分のうち，添え字が集合 $K_i \times T$ の元であるものについての全要素での標本カートシス（すなわち，非音声区間における当該サブバンドの標本カートシス）であり，式 (5) によって計算する．ただし，$\alpha = [\alpha_1, \cdots, \alpha_N]$ は周波数サブバンドの分割ごとの KD の重みを決めるパラメータである．

式 (6) によって定義される KD は，学習におけるカートシスの上昇度の評価がその絶対値に依存するという問題がある [12]．その場合，異なるカートシスを持った雑音を混在させて学習すると，カートシスの絶対値が低い雑音の抑圧時にカートシスの上昇を抑制できない．実際，予備実験を行った結果，そのような事象が観測された．そこで，乖離度を適切に評価するために，Scaled Kurtosis Discrepancy（SKD）を

$$\mathrm{SKD}(X, Z) := \sum_{i=1}^{N} \alpha_i \; \frac{K_k^{t\ T'}{}_{K_i}(X_{k,t}) - K_k^{t\ T'}{}_{K_i}(Z_{k,t})}{K_k^{t\ T'}{}_{K_i}(X_{k,t})} \qquad (7)$$

によって定義する．これは，[6] におけるカートシス比と 1 の距離に対応している．

3.2.3　DNN 学習

ソフトマスクを出力とするような雑音抑圧の DNN を考える．このとき，損失関数に KD を正則化項として加えることで，ミュージカルノイズの発生を回避することを期待する．すなわち，損失関数を，

$$L(X, Y; \theta) := L_0(X, Y; \theta) + \lambda \mathrm{SKD}(X, f(X; \theta) \circ X) \qquad (8)$$

とし，DNN のパラメータを

$$\tilde{\theta} = \arg\max_{\theta} \mathrm{E}[L(X, Y; \theta)] \qquad (9)$$

として推定する．ただし，λ はカートシスマッチングの重みを表すハイパーパラメータである．

学習にあたって，非音声区間はターゲットの音声より決定する．また，K の分割 K_i と α も任意に固定する．

最終的に得られた強調後の振幅スペクトログラム $S \circ X$ に観測信号の位相スペクトログラムを乗じ，短時間逆フーリエ変換をして所望の強調音声を得る．

4.　実験的評価

提案手法の有効性の検証のために，音声強調実験を行った．

4.1　実 験 条 件

訓練データには JNAS [13] より任意に選んだ新聞読み上げ音声 31896 文の前後に非音声区間を付与したデータを 24 個の集合に分け，それぞれ入力 Signal-to-Noise Ratio（SNR）が −5 dB，0 dB，5 dB，10 dB となるような 6 種類の雑音を加えたパラレルデータを作成した．6 種類の雑音の内約は，ガウス性雑音 (GAUSS) と DEMAND [14] よりカートシスの大きく異なる 5 種類の雑音 PSTATION，NFIELD，PRESTO，TBUS，SPSQUARE とした．そのカートシスの一覧を表 1 に示す．音声のサンプルレートは 16 kHz であった．また，短時間フーリエ変換の窓関数には窓長 1024 の Hanning 窓を用い，ホップサイズは 80 とした．また，テストデータには JSUT [15] より任意に選んだ発話音声の前に 1.25 秒の非音声区間を付与した 200 文に対し，SNR が −5 dB，0 dB，5 dB，10 dB となるような先述の 6 種類の雑音を加えたものを用意した．

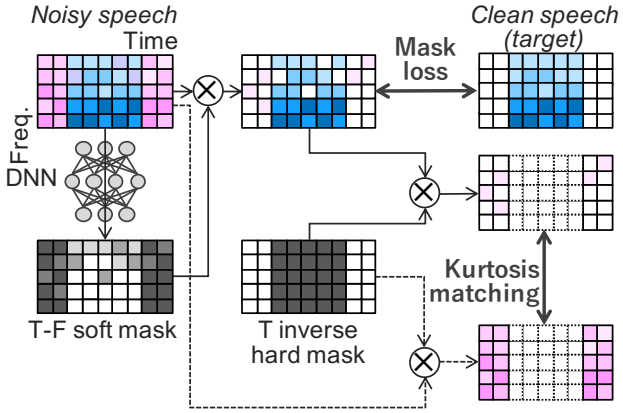

図 1　提案手法の概要．Hard mask は Clean speech より直接決定する非音声区間判定マスクである．この Hard mask を用いて，非音声区間のみの SKD を雑音抑圧の損失関数に加えて学習を実行する．Mask loss は式 (1)，kurtosis matching は式 (7) にそれぞれ対応する．

Fig. 1　Overview of the proposed method. The hard mask for non-speech regions is determined from clean speech. The scaled kurtosis discrepancy of the non-speech frames is added to a loss function for training. Mask loss and kurtosis matching correspond to Eqs. (1) and (7), respectively.

表 1　実験に用いた雑音の種類とカートシスの一覧．

Table 1　List of noise used in evaluation and its kurtosis.

雑音の種類	カートシス
GAUSS	3.00
PSTATION	5.56
PRESTO	12.1
NFIELD	13.3
SPSQUARE	29.8
TBUS	35.8

DNN のアーキテクチャには，中間層 12 層の U-Net [16] を用いた．U-Net の構造は [17] と同様とした．学習にはミニバッチ法を適用し，バッチサイズは 32 とした．また，パッチ長は 256 とした．本稿で示したハイパーパラメータは，$N = 4$，$K_1 = \{0, \cdots, 127\}$，$K_2 = \{128, \cdots, 255\}$，$K_3 = \{256, \cdots, 383\}$，$K_4 = \{384, \cdots, 512\}$，$\alpha = [0.01, 1, 1, 1]$，$\lambda = 1 \times 10^{-4}$ とした．勾配には Adam [18] を用い，ステップサイズは 0.01 とした．そして，エポック回数を 30 として学習を行った．

4.2　雑音抑圧性能と音声歪みの客観評価

従来手法と提案手法について，テストデータを入力として得られた強調音声の Signal-to-Distortion Ratio（SDR）改善量，

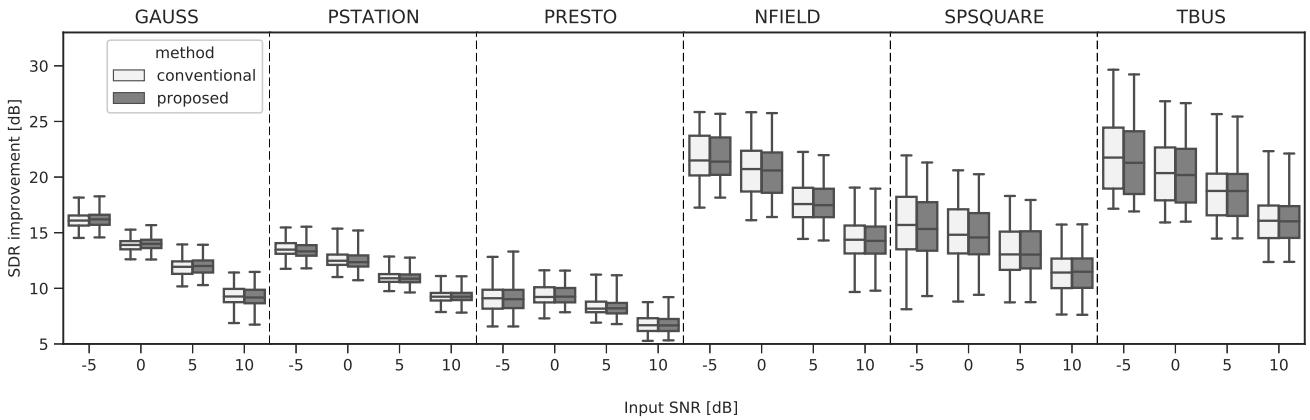

図 2　従来手法と提案手法における SDR 改善量の箱ひげ図．この値が大きいほど雑音抑圧性能が良いことを表している．

Fig. 2　Boxplot of SDR improvements on conventional and proposed methods. The larger value indicates the better performance of noise reduction.

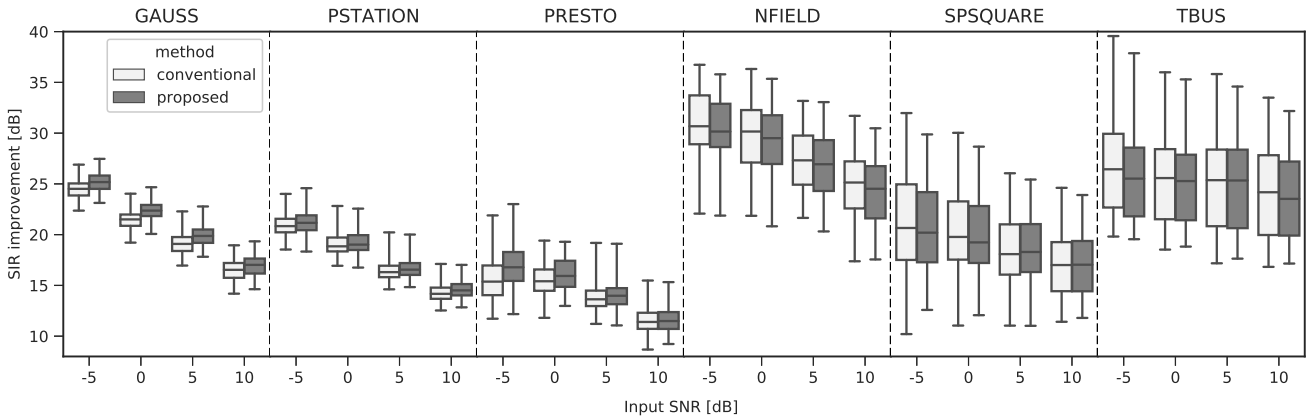

図 3　従来手法と提案手法における SIR 改善量の箱ひげ図．この値が大きいほど雑音抑圧性能が良いことを表している．

Fig. 3　Boxplot of SIR improvements on conventional and proposed methods. The larger value indicates the better performance of noise reduction.

Signal-to-Interferance Ratio（SIR）改善量，Signal-to-Artifact Ratio（SAR），ケプストラム歪み（Cepstral Distortion; CD）を図 2–図 5 に示す．

まず，雑音抑圧性能（図 2–図 4）に関して述べる．いずれの種類の雑音，いずれの入力 SN 比についても，SDR 改善量，SIR 改善量，SAR の差は従来手法と提案手法の間で見られない．したがって，提案手法が雑音抑圧性能を低下させることことは少ないと考えられる．

次に，音声歪み発生量（図 5）について述べる．いずれの種類の雑音，いずれの入力 SN 比についても，従来手法に比べて提案手法では，CD の有意な差は見られないか，あるいは減少していることがわかる．したがって，提案手法が音声歪み発生量を増大させることは少ないと考えられる．

4.3　ミュージカルノイズ発生量の客観評価

強調音声の非音声区間の振幅スペクトログラムのカートシス比を図 6 に示す．この値が 1 になるときに，強調前後でカートシスが不変であることを表す．また，入力 SN 比が 0 dB、雑音が GAUSS，PSTATION のときの強調音声の一つについて，両手法の対数振幅スペクトログラムをそれぞれ図 7，図 8 に示す．

まず，周波数領域でのカートシスの変化（図 6）について述べる．振幅スペクトログラムの非音声区間のカートシス比は，いずれの雑音・入力 SN 比の場合も，従来手法に比べて提案手法では有意に小さくなっている．これは，提案手法がミュージカルノイズの発生の抑圧を達成していることを示唆している．

最後に，スペクトログラムの変化を定性的に評価する（図 7，図 8）．従来手法においては強調音声のスペクトログラム上の中高域の部分に斑状に見える雑音が見られるが，提案手法においては比較的に目立たない．この縞状の雑音は，DNN によって音声として誤特定されたことで残留したミュージカルノイズであると考えられる．しかしながら，ガウス雑音の場合には残留雑音が見受けられる上に，音声区間における局所的なノイズが目立つ．

5.　結　　論

ミュージカルノイズの発生量が低く雑音抑圧性能が高い音声強調を，カートシスマッチングを反映した DNN によるソフト

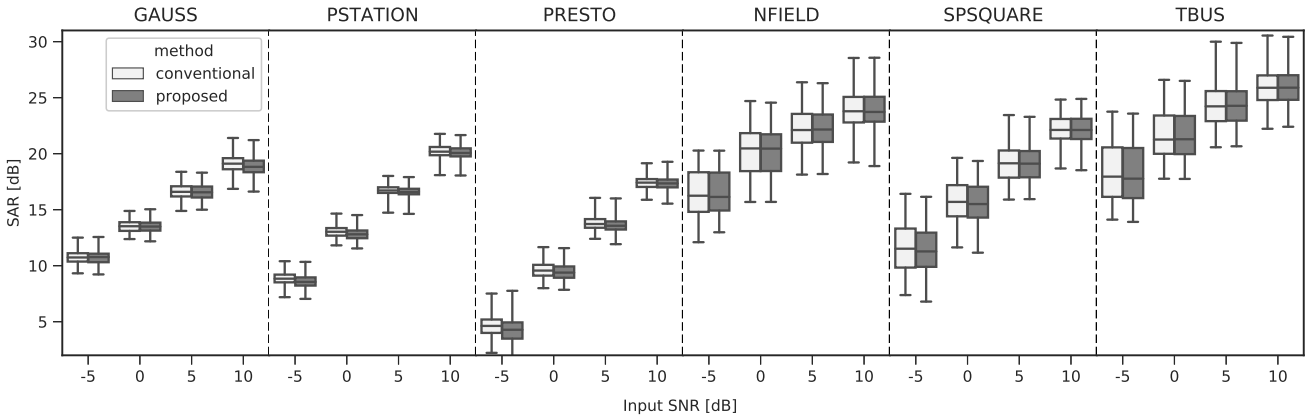

図 4　従来手法と提案手法における SAR の箱ひげ図．この値が大きいほど強調音声におけるアーチファクトが少なく，品質が良いことを表している．

Fig. 4　Boxplot of SARs on conventional and proposed methods. The larger value indicates smaller artifacts in emphasized speech.

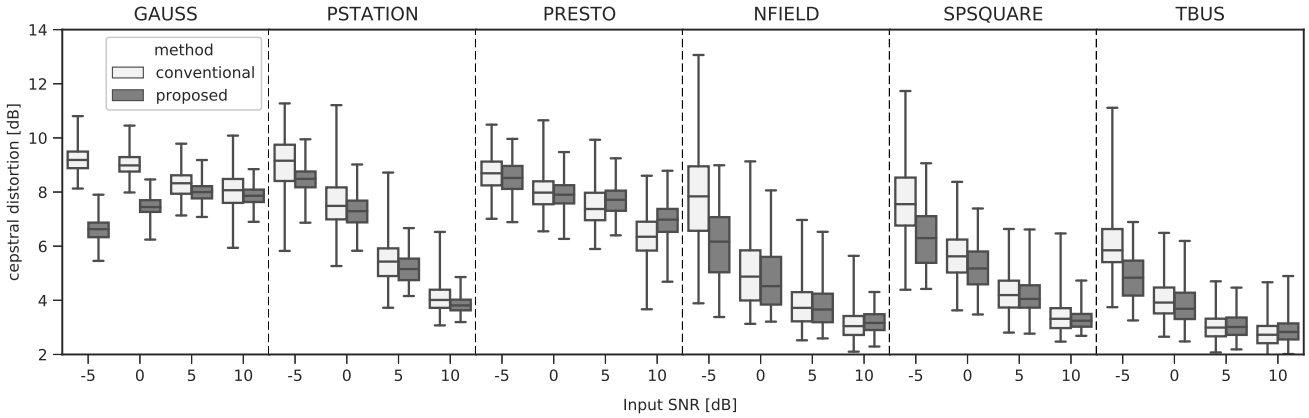

図 5　従来手法と提案手法におけるケプストラム歪みの箱ひげ図．この値が小さいほど強調音声における音声歪み発生量が小さく，品質が良いことを表している．

Fig. 5　Boxplot of cepstrum distortion on conventional and proposed methods. The smaller value indicates smaller distortion of emphasized speech.

マスク雑音抑圧によって定式化した．また，実験的評価によって提案手法が雑音抑圧性能と音声の歪みの発生量を維持したままカートシスの上昇率を低減させることを確認し，その有効性を示した．今後の課題として，音声区間の残留雑音を考慮した損失関数の検討や，より直接的にミュージカルノイズの発生量を定量化できる手法の探求が挙げられる．

謝辞: 本研究の一部は，セコム科学技術支援財団の助成を受け実施した．

文　献

[1] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[2] O. Cappé, “Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[3] Z. Goh, K.-C. Tan, and B. Tan, “Postprocessing method for suppressing musical noise generated by spectral subtraction,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 287–292, Mar. 1998.

[4] R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, “Musical-noise-free speech enhancement based on optimized iterative spectral subtraction,” *IEEE Transactions on Audio Speech And Langage Proceeding*, vol. 20, no. 7, pp. 2080–2094, Sep. 2012.

[5] R. Miyazaki, H. Saruwatari, K. Shikano, and K. Kondo, “Musical-noise-free speech enhancement based on iterative Wiener filtering,” in *Proceedings of The 12th IEEE International Symposium on Signal Processing and Information Technology*, Ho Chi Minh City, Vietnam, Dec. 2012.

[6] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, “Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics,” in *Proceedings of International Workshop for Acoustic Echo and Noise Control 2008*, Seattle, W.A., U.S.A., Sep. 2008.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[8] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in *Proceedings of INTERSPEECH 2013*, Lyon, France, Aug 2013, pp. 436–