カートシスマッチングと深層学習に基づく低ミュージカルノイズ音声強調*

☆溝口 聡，齋藤 佑樹，高道 慎之介，猿渡 洋（東大院・情報理工）

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, 3, 4, 5, 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尖度比（kurtosis ratio）[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尖度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, 8, 9, 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，尖度の乖離度（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，尖度の上昇を避けられることを示す．

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 $\boldsymbol{X}$ とする．これを入力とする DNN のパラメータを $\boldsymbol{\Theta}$ とし，その出力を $\boldsymbol{S} = f(\boldsymbol{X};\boldsymbol{\Theta})$ 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 $\boldsymbol{Y}$ とする．このとき，損失関数を

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

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，

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}\,\mathrm{E}[L_0(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\Theta})] \tag{2}$$

とする．このようにして得られる $\boldsymbol{\Theta}$ は，観測信号 $\boldsymbol{X}$ 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 $\boldsymbol{S}\circ\boldsymbol{X}$ 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．提案法の概要は Fig. 1 である．

---

*Low-musical-noise speech enhancement based on DNNs and kurtosis matching, by MIZOGUCHI, Satoshi, SAITO, Yuki, TAKAMICHI, Shinnosuke and SARUWATARI, Hiroshi (The University of Tokyo)

3.2 カートシスマッチングを考慮した DNN 学習

3.2.1 尖度（kurtosis）

一変数確率変数 W は実数値をとり，確率分布 $p(w)$ に従うとする．また，その平均を μ とする．このとき，確率変数 w の尖度とは，

$$K_W := \frac{\int_{w\in\mathbb{R}}(w-\mu)^4 p(w)\mathrm{d}w}{\left(\int_{w\in\mathbb{R}}(w-\mu)^2 p(w)\mathrm{d}w\right)^2} \tag{3}$$

で定義される統計量であり，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，標本尖度は式 (3) のモンテカルロ積分より

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

によって計算できる．正規分布における尖度は 3 であり，尖度が 3 より大きな分布を優ガウス的であるといい，小さな分布を劣ガウス的であるという．

一般に，自然界に発生する雑音をサンプリングして得られる経験分布は，正規分布に近い形状をしているとされる．対して，ミュージカルノイズのような人工的な歪みは正規分布から乖離しており，優ガウス的である．

3.2.2 尖度の乖離度（Kurtosis Discrepancy: KD）

本稿では，DNN の損失関数に尖度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 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 $\boldsymbol{X}$ 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 $\boldsymbol{Z}$ の行列成分を $Z_{k,t}=S_{k,t}X_{k,t}$（ただし，$S_{k,t}$ は $\boldsymbol{S}$ の行列成分）とする．ここで，$k\in\mathcal{K}:=\{0,\cdots,K\}$ は周波数サブバンドのインデックス，$t\in\mathcal{T}:=\{1,\cdots,T\}$ 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 $\mathcal{K}_i := \{k_i,\cdots,k_{i+1}-1\}$（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 $\mathcal{T}'\subset\mathcal{T}$ とする．と書くことにする．このとき，非音声区間の KD を

$$\mathrm{KD}(\boldsymbol{X},\boldsymbol{Z}) := \sum_{i=1}^{N}\alpha_i\left|\mathcal{K}_{k\in\mathcal{K}_i}^{t\in\mathcal{T}'}(X_{k,t}) - \mathcal{K}_{k\in\mathcal{K}_i}^{t\in\mathcal{T}'}(Z_{k,t})\right| \tag{5}$$

で定義する．ここで，$\mathcal{K}_{k\in\mathcal{K}_i}^{t\in\mathcal{T}'}(X_{k,t})$ は，行列 $\boldsymbol{X}$ の成分のうち，添え字が集合 $\mathcal{K}_i\times\mathcal{T}'$ の元であるものについての全要素での標本尖度（すなわち，非音声区間における当該サブバンドの標本尖度）であり，式 (4) によって計算する．ただし，$\boldsymbol{\alpha}=[\alpha_1,\cdots,\alpha_N]$ 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．

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，

$$\begin{aligned}L(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\Theta}) :=& L_0(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\Theta})\\ &+\lambda\mathrm{KD}(\boldsymbol{X},f(\boldsymbol{X};\boldsymbol{\Theta})\circ\boldsymbol{X})\end{aligned} \tag{6}$$

とし，DNN のパラメータを

$$\tilde{\boldsymbol{\Theta}} = \operatorname*{argmax}_{\boldsymbol{\Theta}}\mathrm{E}[L(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\Theta})] \tag{7}$$

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，観測信号の振幅スペクトログラムから音声信号の振幅スペクトログラムのみを抽出するような DNN を式 (2) によって事前学習する．事前学習した DNN のパラメータを初期値として用いることで，学習の効率化が期待できる．また，事前学習によって得られた教師データ $X_{k,t}$ を入力とする DNN の出力，すなわちソフトマスク $S_{k,t}$ から，教師データの非音声区間のみを判定する．すなわち，

$$\frac{1}{k_+-k_-+1}\sum_{k=k_-}^{k_+}S_{k,t}<\beta \tag{8}$$

なる t の集合を $\mathcal{T}'$ として固定する．ここで，$\beta\in(0,1]$ は音声区間判定のための閾値，$\{k_-,k_+\}$ は音声成分が集中する周波数サブバンドを指定する閾値の集合であり，任意に設定する．また，$\mathcal{K}$ の分割 $\mathcal{K}_i$ と $\boldsymbol{\alpha}$ 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 $\boldsymbol{S}\circ\boldsymbol{X}$ 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 [12] より任意に選んだ女性と男性の発話音声それぞれ 25 文の前後に非音声区間を付与した 50 文に対し，入力 Signal-to-Noise Ratio（SNR）が 0 dB，5 dB，10 dB となるようなガウス性雑音を加えたパラレルデータを作成し，計 150 組の音声を用いた．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 [13] より任意に選んだ発話音声の前後に 1.25 秒の非音声区間を付与した 100 文に対し，SNR が 0 dB，5 dB，10 dB となるようなガウス性雑音を加えたものを用意した．

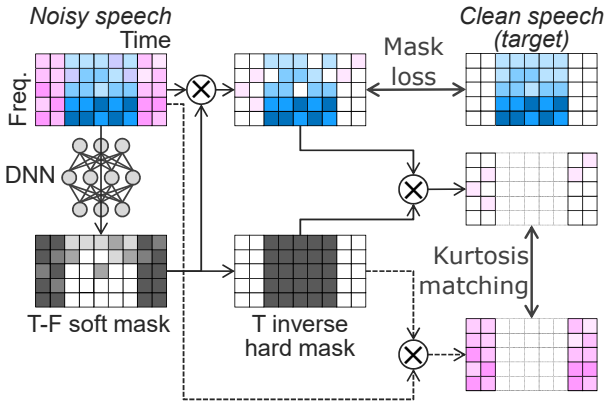

Fig. 1　提案法の概要．事前学習によって得られた DNN に雑音を含む観測データを入力し，出力である時間周波数ソフトマスクから雑音区間のみを抽出するハードマスクを生成．このハードマスクを用いて，雑音区間のみの KD を損失関数に加えて学習を実行する．Mask loss は式 (1)，T inverse hard mask は式 (8)，kurtosis matching は式 (5) にそれぞれ対応する．

DNN のアーキテクチャには，中間層 12 層の U-Net [14] を用いた．U-Net の構造は [15] 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，パッチサイズは 256 とした．本稿で示したハイパーパラメータは，$N = 6$，$\mathcal{K}_1 = \{0, \cdots, 7\}$，$\mathcal{K}_2 = \{8, \cdots, 127\}$，$\mathcal{K}_3 = \{128, \cdots, 255\}$，$\mathcal{K}_4 = \{256, \cdots, 383\}$，$\mathcal{K}_5 = \{384, \cdots, 495\}$，$\mathcal{K}_6 = \{496, \cdots, 512\}$，$\boldsymbol{\alpha} = [0, 0.01, 1, 1, 1, 0]$，$\beta = 0.2$，$k_- = 10$，$k_+ = 74$，$\lambda = 3 \times 10^{-5}$ とした．勾配には Adam [16] を用い，ステップサイズは 0.01 とした．

まず，損失を式 (2) に設定してエポック回数を 100 として事前学習した後，損失を式 (2) に固定したものと，提案手法である式 (6) に変更したものでそれぞれさらにエポック回数を 100 として学習を行った．

4.2　雑音抑圧性能と音声歪みの評価

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）を Fig. 2 に示す．

いずれの入力 SNR についても両手法の SDR が変化がほとんど変化していないか，あるいは上昇していることから，提案手法が雑音抑圧性能を低下させることやと音声の歪みを増大させることは少ないと考えられる．

4.3　KD 及び尖度の評価

強調音声の非音声区間の振幅スペクトログラムの KD を Fig. 3，強調音声の非音声区間の時間領域での尖度を Fig. 4 に示す．また，入力 SNR が 5 dB 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 Fig. 5 に示す．

まず，周波数領域での尖度の変化について述べる．

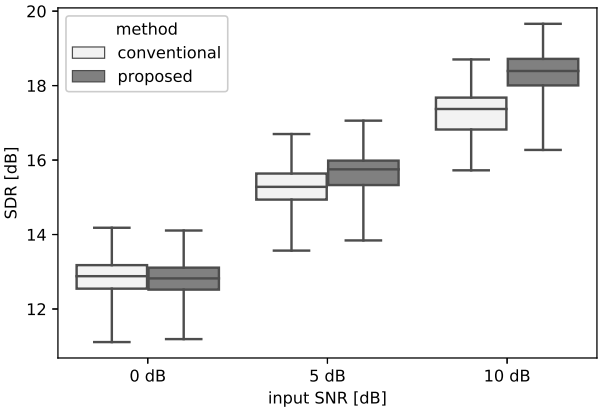

Fig. 2　従来手法と提案手法における SDR の箱ひげ図．

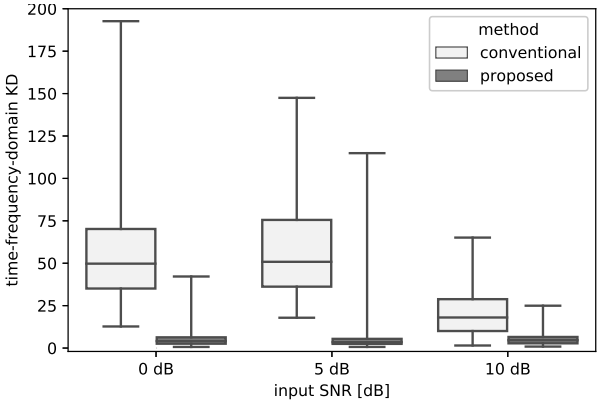

Fig. 3　従来手法と提案手法における振幅スペクトログラムの非音声区間における KD の箱ひげ図．

振幅スペクトログラムの非音声区間の KD は，いずれの入力 SNR 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している．

次に，時間領域における尖度の変化について述べる．時間領域における非音声区間の尖度は，いずれの入力 SNR の場合でも提案手法の方が低くなり，正規乱数の理論値である 3 に極めて近い値をとっている．これは，時間領域で見たときの雑音の従う分布が，四次統計量の意味で正規分布に近くなっていることを示唆している．

最後に，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．特に興味深いのは，損失関数において考慮されていなかった音声区間の中高域においても，この雑音が抑圧されていることである．

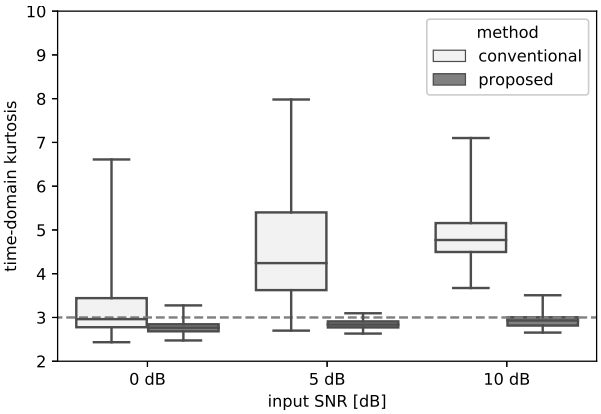

Fig. 4　従来手法と提案手法における時間領域での尖度の箱ひげ図．尖度は非音声区間について計算した．点線で表した値は正規乱数の理論的な尖度の値である．

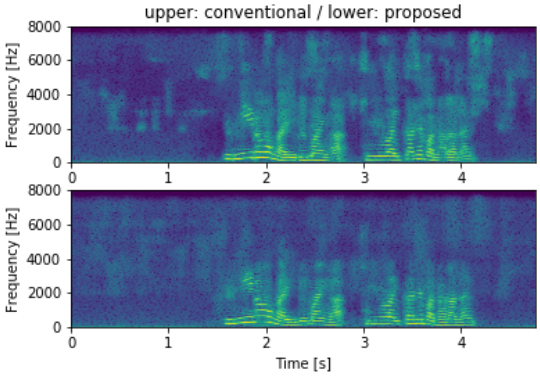

Fig. 5　両手法における入力 SNR が 5 dB のときの強調音声の対数振幅スペクトログラムの一つ．従来手法では高域に縞状の雑音が散見されるが，提案手法では目立たない．

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尖度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，雑音の種類を増やした実験や，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．

参考文献

[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. 436–440.

[9] S.-W. Fu, Y. Tsao, and X. Lu, “SNR-aware convolutional neural network modeling for speech enhancement,” in *Proceedings of INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 3678–3772.

[10] Z. Chen, Y. Huang, J. Li, and Y. Gong, “Speech enhancement based on deep denoising autoencoder,” in *Proceedings of INTERSPEECH 2017*, Stockholm, Sweden, Aug 2017, pp. 3632–3636.

[11] Y. Li, K. Swersky, and R. zemel, “Generative moment matching networks,” in *Proceedings of The 32th International Conference on Machine Learning*, vol. 37, Lille, France, Jul. 2015, pp. 1718–1727.

[12] “新聞記事読み上げ音声コーパス（JNAS）,” http://www.mibel.cs.tsukuba.ac.jp/_090624/jnas/.

[13] R. Sonobe, S. Takamichi, and H. Saruwatari, “JSUT corpus: free large-scale japanese speech corpus for end-to-end speech synthesis,” arXiv preprint, 1711.00354, 2017.

[14] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Proceedings of The 18th International Conference on Medical Image Computing and Computer Assisted Intervention*, Munich, Germany, Oct. 2015, pp. 234–241.

[15] N. Jansson, E. J. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, “Singing voice separation with deep u-net convolutional networks,” in *Proceedings of The 18th International Society for Music Information Retrieval Conference*, Suzhou, China, Oct. 2017, pp. 745–751.

[16] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proceedings of International Conference on Learning Representations*, Banff, Canada, Dec. 2014.