多様なカートシスを持つ雑音に対応した低ミュージカルノイズDNN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 [1] やウィーナフィルタ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と呼ぶ [2, 3]．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（kurtosis ratio）[4] 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．

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（例えば，[5, 6, 7, 8]）．特に，[8] 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．著者らは，単一種類の雑音に対して低ミュージカルノイズなDNN雑音抑圧の手法を提案している [9]．しかし，この手法は雑音が多様なカートシスを持つ場合には対応していない．

これに対し本稿では，Scaled Kurtosis Discrepancy（SKD）による正則化，カートシスマッチングを導入することで，多様なカートシスを持つ雑音を含む音声に対してミュージカルノイズ発生量が小さいDNN音声強調の手法を提案する．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前後のカートシス比が1にな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，提案手法が多様な雑音を含む音声に対して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．

2 ソフトマスクによるDNN音声強調 [8]

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．提案手法の概要はFig. 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

一変数確率変数 W は正実数値の標本カートシスを

$$\kappa_W = \frac{1}{T}\frac{\sum_{t=1}^{T} W_t^4}{\left(\sum_{t=1}^{T} W_t^2\right)^2} \tag{3}$$

によって定義する．

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 [4]．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 SKD

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，SKDを定義する．

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 $\boldsymbol{Z}$

*Low-Musical-Noise DNN-Based Speech Enhancement Applied to Noise with Various Kurtosis, by MIZOGUCHI, Satoshi, SAITO, Yuki, TAKAMICHI, Shinnosuke and SARUWATARI, Hiroshi (The University of Tokyo)

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，$k \in \mathcal{K} := \{1, \cdots, K\}$ 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 Kurtosis Discrepancy（KD）を

$$\mathrm{KD}(\boldsymbol{X}, \boldsymbol{Z}) := \sum_{i=1}^{N} \left| \mathcal{K}^{t\in\mathcal{T}'}_{k\in\mathcal{K}_i}(X_{k,t}) - \mathcal{K}^{t\in\mathcal{T}'}_{k\in\mathcal{K}_i}(Z_{k,t}) \right| \tag{4}$$

で定義する [9]．ここで，$\mathcal{K}^{t\in\mathcal{T}'}_{k\in\mathcal{K}_i}(X_{k,t})$ 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 (3) によって計算する．

式 (4) 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．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，SKD を

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

によって定義する．これは，[4] 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，

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

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，非音声区間はターゲットの音声より決定する．また，$\mathcal{K}$ の分割 $\mathcal{K}_i$ を任意に固定する．

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 [10] 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 [11] 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 Table 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 [12] より任意に選んだ発話音声の前に 6.25 秒の非音声区間を付与した 200 文に対し，入力 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．

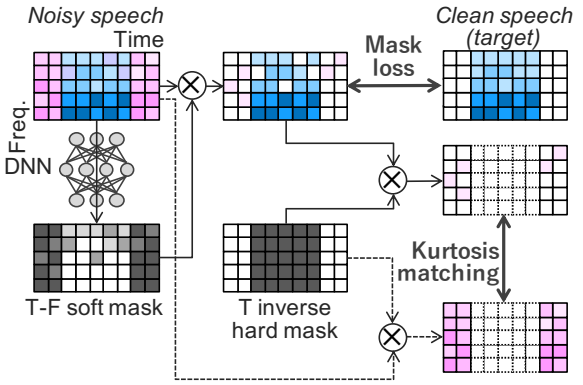

Fig. 1　提案手法の概要．Hard mask は clean speech より直接決定する非音声区間判定マスクである．この hard mask を用いて，非音声区間のみの SKD を雑音抑圧の損失関数に加えて学習を実行する．Mask loss は式 (1)，kurtosis matching は式 (5) にそれぞれ対応する．

Table 1　実験に用いた雑音の種類とカートシスの一覧

ラベル	雑音の種類	カートシス
GA	GAUSS	3.00
PS	PSTATION	5.56
PR	PRESTO	12.1
NF	NFIELD	13.3
SP	SPSQUARE	29.8
TB	TBUS	35.8

DNN のアーキテクチャには，中間層 12 層の U-Net [13] を用いた．U-Net の構造は [14] 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，$N = 4$，$\mathcal{K}_1 = \{1, \cdots, 127\}$，$\mathcal{K}_2 = \{128, \cdots, 255\}$，$\mathcal{K}_3 = \{256, \cdots, 383\}$，$\mathcal{K}_4 = \{384, \cdots, 512\}$，$\lambda = 1 \times 10^{-4}$ とした．勾配には Adam [15] 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）改善量を Fig. 2 に，22 次までのケプストラム歪み（Cepstral Distortion: CD）を Fig. 3 に示す．

まず，雑音抑圧性能（Fig. 2）に関して述べる．い

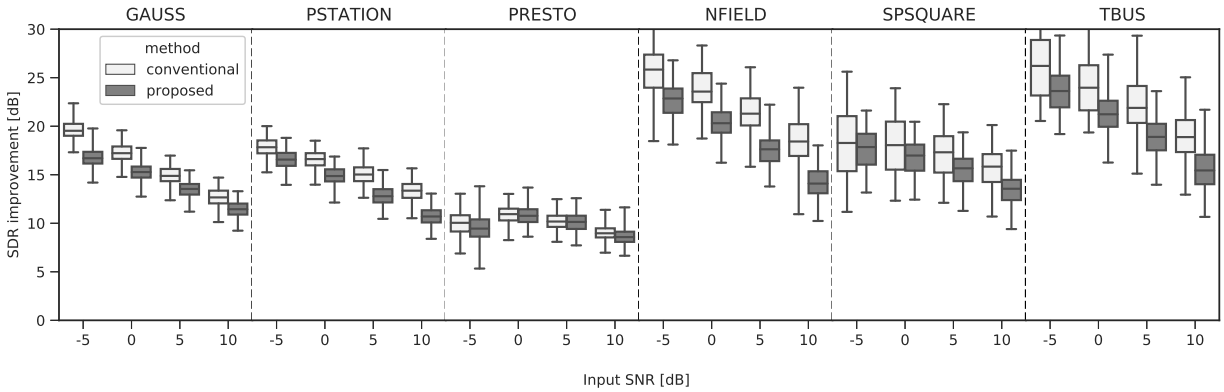

Fig. 2　従来手法と提案手法における SDR 改善量の箱ひげ図．この値が大きいほど雑音抑圧性能が良いことを表している．

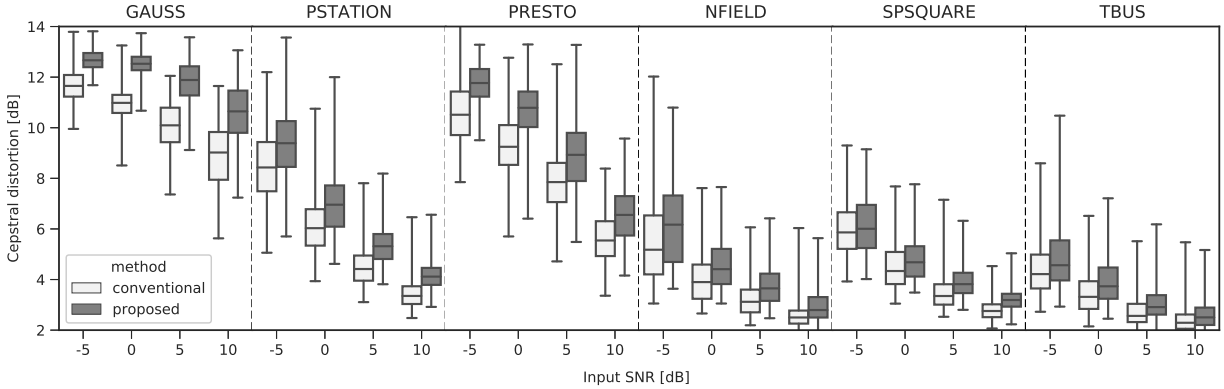

Fig. 3　従来手法と提案手法におけるケプストラム歪みの箱ひげ図．この値が小さいほど強調音声における音声歪み発生量が小さく，品質が良いことを表している．

ずれの種類の雑音，いずれの入力 SNR についても，SDR 改善量の差は従来手法と提案手法の間でわずかである．したがって，提案手法が雑音抑圧性能を低下させることは少ないと考えられる．

次に，音声歪み発生量（Fig. 3）について述べる．いずれの種類の雑音，いずれの入力 SNR 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，あるいはわずかに増加していることがわかる．したがって，提案手法は，わずかに音声歪み発生量を増大させると考えられる．

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 Fig. 4 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．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 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唆している．

4.4　ミュージカルノイズ発生量の主観評価

提案手法と従来手法の強調後のミュージカルノイズ発生量を比較するため，24 人の参加者に，強調後の信号の冒頭 6.25 秒の非音声区間の対を聴いて，どちらの雑音が自然である（人工的に聴こえない）と感じたかの回答を求めた．なお，この評価は雑音の各種類，各 SNR と対して別々に実施した．その結果

Table 2　雑音の自然性の評価スコアと p-値．p-値が 10^{-3} を下回る場合を有意とし，そのときの評価が良い方のスコアを太字によって示す

ラベル	入力 SNR	従来法	提案法	p-値
GA	−5 dB	**0.688**	0.313	$< 10^{-10}$
	0 dB	**0.804**	0.196	$< 10^{-10}$
	5 dB	**0.725**	0.275	$< 10^{-10}$
	10 dB	**0.863**	0.138	$< 10^{-10}$
PS	−5 dB	0.408	**0.592**	5.34×10^{-5}
	0 dB	0.417	**0.583**	2.45×10^{-4}
	5 dB	0.404	**0.596**	2.36×10^{-5}
	10 dB	0.271	**0.729**	$< 10^{-10}$
PR	−5 dB	0.517	0.483	4.66×10^{-1}
	0 dB	0.421	**0.579**	4.98×10^{-4}
	5 dB	0.354	**0.646**	$< 10^{-10}$
	10 dB	0.483	0.517	4.66×10^{-1}
NF	−5 dB	0.250	**0.750**	$< 10^{-10}$
	0 dB	0.221	**0.779**	$< 10^{-10}$
	5 dB	0.200	**0.800**	$< 10^{-10}$
	10 dB	0.192	**0.808**	$< 10^{-10}$
SP	−5 dB	0.367	**0.633**	2.94×10^{-9}
	0 dB	0.383	**0.617**	2.34×10^{-7}
	5 dB	0.225	**0.775**	$< 10^{-10}$
	10 dB	0.250	**0.750**	$< 10^{-10}$
TB	−5 dB	0.238	**0.763**	$< 10^{-10}$
	0 dB	0.258	**0.742**	$< 10^{-10}$
	5 dB	0.167	**0.833**	$< 10^{-10}$
	10 dB	0.238	**0.763**	$< 10^{-10}$

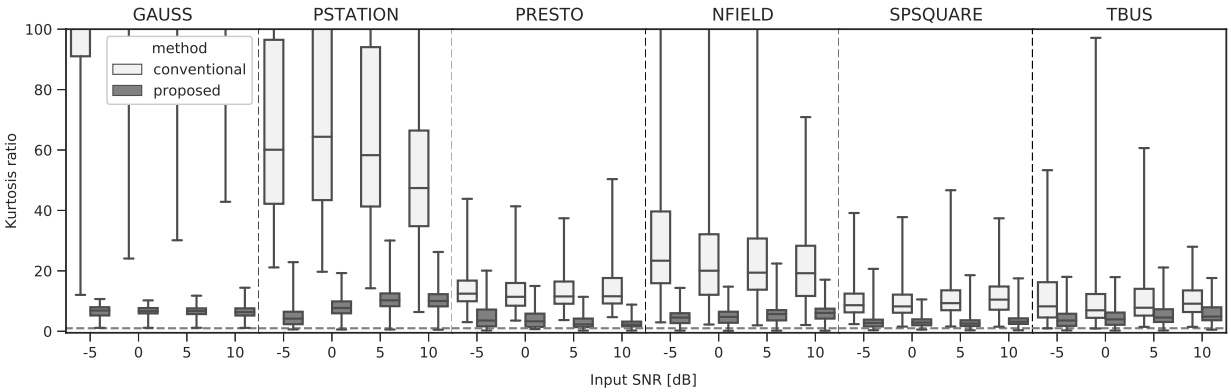

Fig. 4　従来手法と提案手法における非音声区間のカートシス比の箱ひげ図．この値が 1 に近いほどミュージカルノイズの発生量が小さく，品質が良いことを表している．

と，二群の t-検定による p-値を Table 2 に示す．この結果により，GAUSS 以外の場合，すなわち自然環境に存在する雑音の場合は提案法が有意にミュージカルノイズ発生量を抑えていることが示唆される．

5　結論

多様なカートシスを持つ雑音に対応した低ミュージカルノイズな音声強調を，SKD による正則化を利用した 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ある程度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．さらに，主観評価によって提案手法が残留雑音の自然性を保つことを示した．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，セコム科学技術振興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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 溝口聡, 齋藤佑樹, 高道慎之介, and 猿渡洋, “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,” 日本音響学会 2018 年秋季研究発表会講演論文集, pp. 177–180, 2018.

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

[11] J. Thiemann, N. Ito, and E. Vincent, “Musical-noise-free speech enhancement based on optimized iterative spectral subtraction,” in *Proceedings of 21st International Congress on Acoustics*, Montreal, Canada, Jun. 2013.

[12] 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.

[13] 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.

[14] 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.

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