様々な雑音抑圧手法における倍音復元による音声の品質評価に関する研究

Evaluation of Speech Quality using Harmonic Regeneration for Various Noise Reduction Techniques

宇根 昌和 † 宮崎 亮一 †
Masakazu Une † Ryoichi Miyazaki †
† 徳山工業高等専門学校 情報電子工学科

1 はじめに

音声対話ロボットやテレビ会議システム，補聴器など，音声通信に関するシステムが増加しており，音声による情報伝達が多く利用されている．しかし，周囲の雑音によって目的音声の品質が劣化してしまうため，高精度で目的音声を抽出する雑音抑圧が必要となる．雑音抑圧手法は主に２つに分けられる．１つは信号の変形が線形な関係で表される線形処理，もう１つは信号の変形が非線形な関係で表される非線形処理である．線形処理に基づく雑音抑圧技術の代表例としては，ビームフォーミングに基づく手法やブラインド音源分離に基づく手法などがある．線形処理に基づく雑音抑圧は出力する音質が良い反面，複数のマイクロホンを必要とするためにシステムの規模やコストが大きくなってしまう問題がある．一方，非線形処理に基づく雑音抑圧技術は，雑音抑圧性能が高く，アルゴリズムの汎用性に優れており，演算量も少ないことから盛んに研究されている技術である．しかし，非線形処理は，出力信号中に「ミュージカルノイズ」と呼ばれる非常に耳障りな歪みが生じる問題がある．また，雑音抑圧量を大きくすると，目的の音声成分も歪んでしまい，かえって聞こえづらくなってしまうこともある．

そこで，音声成分の歪みを改善する方法として倍音復元に基づく雑音抑圧（Harmonic Regeneration Noise Reduction: HRNR）が提案されている [1]．発声された音声には倍音成分が多く含まれており，雑音抑圧を行うと音声中の多くの倍音成分が歪んでしまう．HRNR で歪んだ倍音成分を倍音復元することにより，より品質の良い音声を得ることができる．しかし論文 [1] では，１つの雑音抑圧手法についてのみ言及しており，他の雑音抑圧手法での挙動については述べられていない．

そこで本研究では，３つの雑音抑圧手法（Wiener Filtering: WF [2]，Spectral Subtraction: SS [3]，Minimum Mean-square Error Short-Time Spectral Amplitude: MMSE-STSA 法 [4]）に対して倍音復元を行い，歪みの改善量などの効果について調査した．

2 雑音抑圧手法

2.1 信号の定義

雑音を含む観測信号 $x(t)$ は，元の音声信号 $s(t)$ と雑音信号 $n(t)$ から成り次の式で表される．

$$x(t) = s(t) + n(t) \tag{1}$$

式 (1) を短時間フーリエ変換することで，次の式に表される複素スペクトルを得る．

$$X(p,k) = S(p,k) + N(p,k) \tag{2}$$

ここで，p は短時間フレームのインデックス，k はフレーム内の周波数インデックスを表す．次に，雑音抑圧を行う．観測信号のスペクトル $X(p,k)$ から，次の式に示すような音声信号のスペクトルの推定値 $\hat{S}(p,k)$ を求めることを考える．

$$\hat{S}(p,k) = G(p,k)X(p,k) \tag{3}$$

ここで，$G(p,k)$ はスペクトルゲインと呼ばれるもので，観測信号のスペクトルに適当なスペクトルゲインを乗じることで，推定音声信号のスペクトルを得る．

2.2 WF

WF は，古典的な雑音抑圧手法である [2]．WF のスペクトルゲイン $G_{\mathrm{WF}}(p,k)$ は次の式で表される．

$$G_{\mathrm{WF}}(p,k) = \xi(p,k)/1 + \xi(p,k) \tag{4}$$

$\xi(p,k)$ は事前 SNR と呼ばれ，次の式で表される．

$$\xi(p,k) = \mathrm{E}[|S(p,k)|^2]/\mathrm{E}[|\hat{N}(p,k)|^2] \tag{5}$$

ここで，$|\hat{N}(p,k)|^2$ は推定雑音信号のパワースペクトル，$\mathrm{E}[\cdot]$ は期待値演算子を表す．事前 SNR を求めるには音声信号の情報が必要となるが，実環境で音声信号を事前に知っておくことはできない．そこで，次式で表される decision-directed 法を利用した事前 SNR 取り出す $\xi(p,k)$ を推定する [4]．

$$\hat{\xi}(p,k) = \alpha \frac{|G_*(p-1,k-1)X(p-1)|^2}{\mathrm{E}[|\hat{N}(p,k)|^2]} + (1-\alpha)P[\gamma(p,k)-1] \tag{6}$$

ここで，α は忘却係数と呼ばれ，前フレームの情報をどの程度事前 SNR の推定に利用するかを決めるパラメータである．一般的には $\alpha = 0.98$ と設定するのが音質的には最も良いとされる．また，$G_*(p,k)$ は各抑圧手法でのスペクトルゲインである．$P[\cdot]$ は半波整流関数であり，次の式で定義される．

$$P[x] = \begin{cases} x & (\text{if } x > 0) \\ 0 & (\text{otherwise}) \end{cases} \tag{7}$$

また，$\gamma(p,k)$ は事後 SNR と呼ばれ，次の式で表される．

$$\gamma(p,k) = |X(p,k)|^2 / \mathrm{E}[|\hat{N}(p,k)|^2] \tag{8}$$

2.3 SS

SS は演算量が少なく，高い雑音抑圧性能を持つため，現在でも多く用いられている雑音抑圧手法である [3]．SS は観測信号から推定した雑音信号をパワースペクトル領域で減算し，目的音声を推定する手法である．SS のスペクトルゲイン $G_{\mathrm{SS}}(p,k)$ は，次の式で表される．

$$G_{\mathrm{SS}}(p,k) = \begin{cases} \sqrt{1 - \frac{\beta \mathrm{E}[|\hat{N}(p,k)|^2]}{|X(p,k)|^2}} & (|X(p,k)|^2 > \beta \mathrm{E}[|\hat{N}(p,k)|^2]) \\ 0 & (\text{otherwise}) \end{cases} \tag{9}$$

ここで，β は減算係数と呼ばれ，観測信号から推定雑音信号をどの程度減算するかを決めるパラメータである．

2.4 MMSE-STSA 法

MMSE-STSA 法は，元の音声信号と推定音声信号の振幅スペクトルの平均二乗誤差を最小にする手法である [4]．MMSE-STSA 法のスペクトルゲイン $G_{\mathrm{STSA}}(p,k)$ は次の式で表される，

$$G_{\mathrm{STSA}}(p,k) = \frac{\sqrt{\nu(p,k)}}{\gamma(p,k)} \Gamma\left(\frac{3}{2}\right) M\left(-\frac{1}{2}; 1; \nu(p,k)\right) \tag{10}$$

ここで，$\Gamma(h)$，$M(a;b;z)$ はそれぞれガンマ関数，第一種合流超幾何関数を表し，$\nu(p,k)$ は次の式で表される．

$$\nu(p,k) = \xi(p,k)\gamma(p,k)/1 + \xi(p,k) \tag{11}$$

ここで，事前 SNR $\xi(p,k)$ は WF と同様に，式 (6) により推定する．

3 倍音復元

音声は多くの倍音成分を含み，雑音抑圧を行うと倍音成分が失われる．HRNR は，失われた倍音成分を復元させるために提案された手法である [1]．HRNR で倍音復元した音声信号のスペクトルの推定値 $\hat{S}_{\mathrm{HRNR}}(p,k)$ を得るためのスペクトルゲイン $G_{\mathrm{HRNR}}(p,k)$ は以下に示すように，HRNR の事前 SNR $\hat{\xi}_{\mathrm{HRNR}}(p,k)$ と事後 SNR $\gamma(p,k)$ の関数で表される．

$$G_{\mathrm{HRNR}}(p,k) = \upsilon(\hat{\xi}_{\mathrm{HRNR}}(p,k), \gamma(p,k)) \tag{12}$$

ここで，関数 υ には様々なスペクトルゲイン関数を選択できる（WF，MMSE-STSA 法など）．HRNR の事前 SNR $\hat{\xi}_{\mathrm{HRNR}}(p,k)$ は，次の式で表される．

$$\begin{aligned} \hat{\xi}_{\mathrm{HRNR}}(p,k) =& \rho(p,k) \frac{|\hat{S}(p,k)|^2}{\mathrm{E}[|\hat{N}(p,k)|^2]} \\ &+ (1 - \rho(p,k)) \frac{|S_{\mathrm{harmo}}(p,k)|^2}{\mathrm{E}[|\hat{N}(p,k)|^2]} \end{aligned} \tag{13}$$

ここで，$\rho(p,k)$ は，音声の推定に利用する情報をどの程度にするかを決める関数である．この関数は，雑音抑圧に用いたスペクトルゲインとすると良いとされる [1]．$S_{\mathrm{harmo}}(p,k)$ は復元信号のスペクトルと呼ばれ，時間領域での復元信号は次の式で定義される．

$$s_{\mathrm{harmo}}(t) = NL(\hat{s}(t)) \tag{14}$$

ここで，NL は非線形関数である（絶対値，半波整流関数など）．また，$\hat{s}(t)$ は時間領域での雑音抑圧を行った信号である．

4 評価実験

4.1 評価尺度の定義

評価実験では，3 つの手法で雑音抑圧を行うため，各雑音抑圧手法での雑音抑圧量を揃えて評価を行う．そこで SNR 向上比 (Noise Reduction Rate: NRR) [5] を導入する．NRR は次のように計算される．

$$\mathrm{NRR} = 10 \log_{10} \frac{\mathrm{E}[s_{\mathrm{out}}^2] / \mathrm{E}[n_{\mathrm{out}}^2]}{\mathrm{E}[s_{\mathrm{in}}^2] / \mathrm{E}[n_{\mathrm{in}}^2]} \tag{15}$$

ここで s_{in} と s_{out} はそれぞれ入力，出力音声であり，n_{in} と n_{out} はそれぞれ入力，出力雑音である．NRR は，各抑圧の内部のパラメータを変動させることで値を調整する．一般的に，内部のパラメータを大きくすると，雑音抑圧量が大きくなり，NRR の値も大きくなる．評価対象の信号は雑音抑圧後の NRR を揃えた各推定音声信号と，倍音復元した音声信号であり，NRR，カートシス比 (Kurtosis Ratio: KR) [6]，ケプストラム歪み (Cepstral Distortion: CD) [7] の 3 つの客観評価尺度で評価する．KR は，ミュージカルノイズの発生量を測る評価尺度である．カートシス kurt は次式で計算される．

$$\mathrm{kurt} = \mu_4 / \mu_2^2 \tag{16}$$

ここで μ_2，μ_4 はそれぞれ 2 次，4 次のモーメントである．これより KR は次式で計算される．

$$\mathrm{KR} = \mathrm{kurt}_{\mathrm{proc}} / \mathrm{kurt}_{\mathrm{org}} \tag{17}$$

ここで $\mathrm{kurt}_{\mathrm{proc}}$ は雑音抑圧後の雑音区間でのカートシスであり，$\mathrm{kurt}_{\mathrm{org}}$ は観測信号の雑音区間でのカートシスである．KR の値が小さいほど，ミュージカルノイズの発生量が少ないことを表す．

さらに，精度の良い KR の評価を行うため，信号のパワースペクトルがガンマ分布に従うと仮定する [6]．観測されたデータサンプルから，ガンマ分布の形状母数 α_{kurt} は最尤推定によって次の式で推定する [8]．

$$\alpha_{\mathrm{kurt}} = \frac{3 - \gamma_{\mathrm{kurt}} + \sqrt{(\gamma_{\mathrm{kurt}} - 3)^2 + 24\gamma_{\mathrm{kurt}}}}{12\gamma_{\mathrm{kurt}}} \quad (18)$$

ここで γ_{kurt} は次式で計算される．

$$\gamma_{\mathrm{kurt}} = \log(\mathrm{E}[|X(p,k)|^2]) - \mathrm{E}[\log|X(p,k)|^2] \quad (19)$$

式 (18) より，ガンマ分布によるモデリング信号のカートシス $\mathrm{kurt}_{\mathrm{GM}}$ は次の式で得られる [6]．

$$\mathrm{kurt}_{\mathrm{GM}} = \frac{\mu_4}{\mu_2^2} = \frac{(\alpha_{\mathrm{kurt}} + 2)(\alpha_{\mathrm{kurt}} + 3)}{\alpha_{\mathrm{kurt}}(\alpha_{\mathrm{kurt}} + 1)} \quad (20)$$

CD は，雑音抑圧による歪みの量を測る評価尺度である．CD は次式で計算される [7]．

$$\mathrm{CD} = \frac{20}{T\log 10}\sum_{p=1}^{T}\sqrt{\sum_{k=1}^{B} 2(C_{\mathrm{out}}(p,k) - C_{\mathrm{ref}}(p,k))^2} \quad (21)$$

ここで，T は総フレーム数，B は評価に用いるケプストラムの次元数である．$C_{\mathrm{out}}(p,k)$ と $C_{\mathrm{ref}}(p,k)$ はそれぞれ，処理後の音声とクリーンな音声のケプストラム係数である．CD の値が小さいほど，音声の歪みが小さいことを表す．

4.2 実験条件

目的音声には JNAS の音声コーパス [9] より 10 文（男性 5 発話，女性 5 発話，計 10 発話），雑音は駅雑音，道雑音，白色ガウス雑音の 3 種類とし，それらをそれぞれ 0 dB，5 dB の入力 SNR で混合したものを観測信号とした．評価対象はこれらの観測信号に対して 3 つの雑音抑圧手法（WF，SS，MMSE-STSA 法）で雑音抑圧した音声信号と，雑音抑圧後に倍音復元した音声信号の 6 信号とした．式 (12) の倍音復元するスペクトルゲインの式 v は MMSE-STSA 法のスペクトルゲインを用いた．式 (13) のパラメータ $\rho(p,k)$ には，それぞれの雑音抑圧のスペクトルゲインを用いた．また，式 (14) の非線形関数には式 (7) 半波整流関数を用いた．3 つの手法で雑音抑圧を行った音声の NRR が 10 dB になるように，WF，MMSE-STSA 法においては式 (6) の忘却係数 α を，SS においては式 (9) の減算係数 β を調整した．式（21）の次元数 B は 22 とした．

4.3 実験結果

まず，音声と駅雑音を入力 SNR 0 dB で混合した場合における，各種信号のスペクトログラムを図 1 に示す．図 1(a) は音声と駅雑音を入力 SNR 0 dB で混合した信号，図 1(b) は SS で雑音抑圧した信号，図 1(c) は図 1(b) に対し，倍音復元を行った信号，図 1(d) はクリーンな音声信号である．図 1(b) でより雑音抑圧で失われた倍音成分が図 1(c) では倍音復元されていることがわかる．

次に入力 SNR を 0 dB，5 dB で混合した時の NRR，KR，CD の値を図 2，図 3 に示す．図 2(a)，図 3(a) より倍音復元による NRR の変化は小さいことがわかる．

また，図 2(b)，図 3(b) より KR について，倍音復元前の音声より倍音復元後の音声の KR が小さいことから，ミュージカルノイズの発生量が減少していることがわかる．しかし，白色ガウス雑音中での音声を MMSE-STSA 法で雑音抑圧し，倍音復元した場合には，ミュージカルノイズの発生量がわずかに増加していることが確認できる．

さらに，図 2(c)，図 3(c) より CD について，全ての雑音抑圧手法で CD が減少しており，最も音声成分の歪みが改善されている雑音抑圧手法は SS であることがわかる．また，倍音復元後の CD の値はおおよそ 1 dB から 2 dB の間にある．SS で抑圧した場合のように音声が大きく歪んでいると，倍音復元による大きな歪みの改善が期待できる．一方，MMSE-STSA 法で抑圧すると倍音復元しなくとも歪みの量が小さいため，歪みは改善するものの，改善量は小さい．図 1(c) からも，特に周波数の高い部分が倍音復元されており，歪みが改善されていることが見て取れる．

以上をまとめると，NRR の変化は小さいが，ミュージカルノイズの発生量が減少していることや，CD が改善されていることから，どの雑音抑圧手法に対しても，倍音復元を行うことは有効であると言える．

5 まとめ

本稿では，様々な雑音抑圧手法で雑音抑圧を行った音声信号に対し，倍音復元を行うことで得られる信号についての評価を行った．評価実験から，倍音復元を行うことで，雑音抑圧による歪みの改善に加え，ほとんどの場合においてミュージカルノイズの発生量も減少することわかり，どの雑音抑圧手法に対しても倍音復元を行うことは有効であることがわかった．今後は，音声認識実験や主観評価実験を行い，倍音復元の音声認識性能や人間の聴覚への影響を調査する．

6 謝辞

本研究は JSPS 科研費 16K21579 の助成を受けたものです．

参考文献

[1] C. Plapous, “Improved signal-to-noise ratio estimation for speech enhancement,” *IEEE Transactions on Acoustics,Speech and Signal Processing*, vol.14, no.6, pp.2098–2108, 2006.

[2] N. Wiener, “Extrapolation,interpolation and smoothing of stationary time series with engineer-

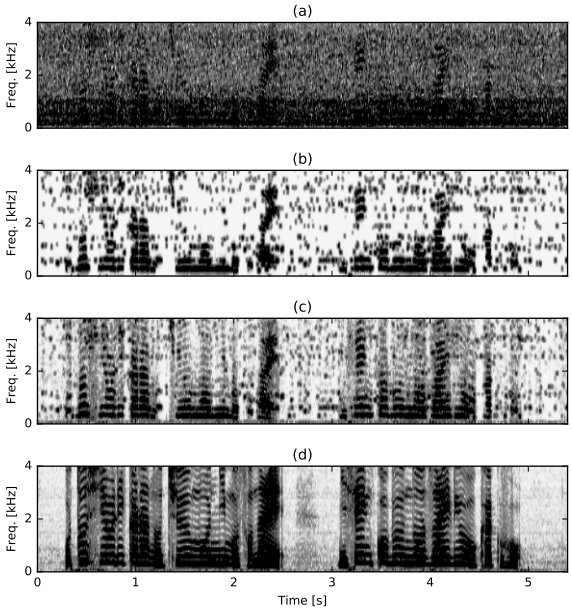

図 1: 各信号のスペクトログラム．(a)：音声と駅雑音を入力 SNR 0 dB で混合した雑音混入信号，(b)：雑音混入信号を SS で雑音抑圧した音声，(c)：雑音混入信号を HRNR で雑音抑圧した音声，(d)：クリーンな音声

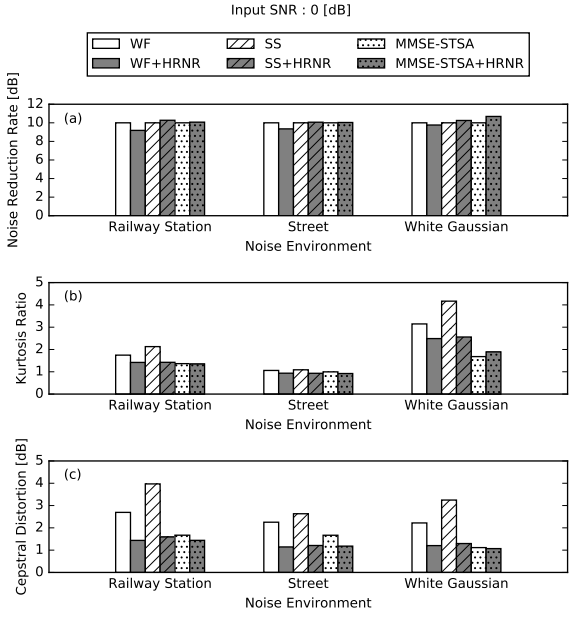

図 2: 入力 SNR 0 dB 混合でそれぞれ雑音抑圧したときの評価．(a)：NRR，(b)：KR，(c)：CD.

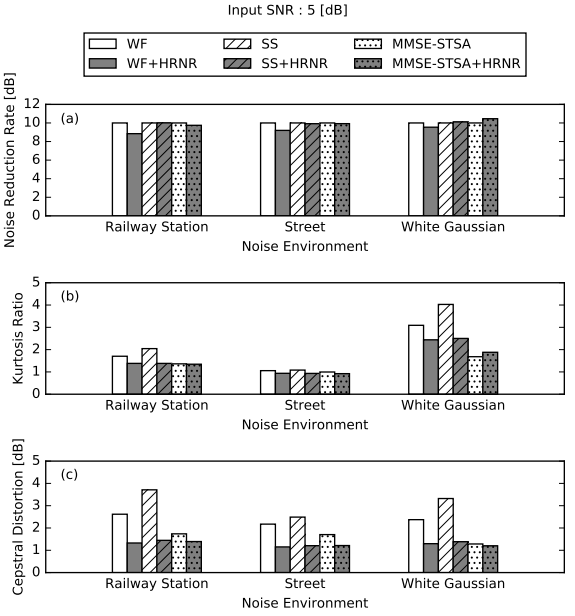

図 3: 入力 SNR 5 dB 混合でそれぞれ雑音抑圧したときの評価．(a)：NRR，(b)：KR，(c)：CD.

ing applications," *Cambridge, MA: MIT Press*, 1949.

[3] S. F. 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, 1979.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.27, no.6, pp.1109–1121, 1984.

[5] R. Miyazaki, *et al.*, "Musical-noise-free speech enhancement based on optimizediterative spectral subtraction," *IEEE Transactions on Audio, Speech and Language Processing*, vol.20, no.7, pp.2080–2094, 2012.

[6] Y. Uemura, *et al.*, "Automaticoptimization scheme of spectral subtraction based on musicalnoise assessment via higher-order statistics," *Proc. IWAENC2008*, 2008.

[7] L. Rabiner and B. Juang, "Fundamentals of Speech Recognition," *Upper Saddle River, NJ: Prentice-Hall*, 1993.

[8] M. Evans, *et al.*, "Statistical Distributions," *2nd ed., Wiley-Interscience*, 1993.

[9] K. Ito, *et al.*, "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," *The Journal of Acoustical Society of Japan*, vol.20, pp.196–206, 1999.