# 雑音と残響の同時抑圧による音声強調*

◎吉岡 拓也，　中谷 智広，　三好 正人
日本電信電話 (株)　NTT コミュニケーション科学基礎研究所

## 1　はじめに

実環境で収音した音声信号は，雑音と残響によって歪む．この観測信号から元の音声信号を回復することは，音声の品質や認識性能を向上させる上で，重要な課題である．従来の雑音残響下における音声強調法に共通するアプローチは，雑音抑圧と残響除去を縦列的に接続するというものである．例えば，[1] は，スペクトル減算により雑音を抑圧した信号に対して，残響除去を適用している．しかし，より精度良い音声強調を行うためには，雑音抑圧と残響除去を協調して動作させるべきであろう．

本稿では，雑音と残響を同時に抑圧する音声強調法を提案する．雑音と残響をともに抑圧する方法がこれまで提案されてこなかったのは，一つには，雑音抑圧が通常，周波数領域で実現されるのに対して，残響除去は，時間領域におけるフィルタ処理課題として取り組まれてきたことによる．中谷らは最近，周波数領域における残響除去法を提案した [2]．本稿では，これに基づいて，観測信号の統計モデルを周波数領域で設定する．提案法は，このモデルのパラメータを最尤推定法により求めた後，推定されたパラメータを用いて原音声信号を回復する．こうして，提案法では，単一の尤度関数を最大化するという一貫した方法で，雑音抑圧と残響除去が統合される．本稿では，単一マイクロホンを想定して記述するが，複数マイクロホンの場合にも容易に拡張できる．

## 2　観測信号のモデル

本稿では，すべての信号を時間周波数領域で表現する．本稿を通じて，時間インデクスを $t$ $(0 \le t \le T-1)$ で，角周波数インデクスを $l$ $(0 \le l \le N-1)$ で表す．ただし，すべての信号は，時間軸上及び単位円上で等間隔に標本化されているものとする．すなわち，時間周波数解析としては，短時間フーリエ変換（STFT）やポリフェーズフィルタバンク等を用いる．

原音声信号と観測信号を，それぞれ $S_{t,l}$，$Y_{t,l}$ と表す．$Y_{t,l}$ は，図 1 に示す系によって，$S_{t,l}$ から生成される．図 1 の残響音声信号と雑音信号を，それぞれ $X_{t,l}$，及び $D_{t,l}$ としよう．このとき，$Y_{t,l}$ は次式にしたがって生成されるとする

$$Y_{t,l} = X_{t,l} + D_{t,l} \tag{1}$$

$$X_{t,l} = \sum_{k=1}^{K_l} g^*_{k,l} X_{t-k,l} + S_{t,l} \tag{2}$$

上付き添え字 $*$ は複素共役を表す．(2) では，室内伝達系は周波数ごとに自己回帰系として表せるということを，仮定している．$g_{1,l}, \cdots, g_{K_l,l}$ は，周波数 $l$ における回帰系のタップ重み係数である．

以下の条件を仮定する．

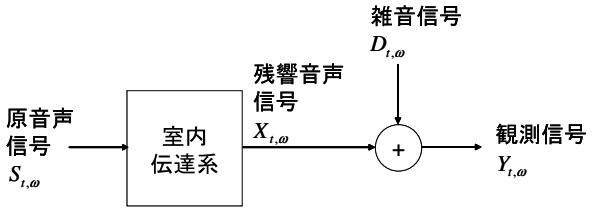


図 1　観測信号の生成過程

(1) $\omega \in [-\pi, \pi)$ を角周波数として，時間 $t$ における音声のパワースペクトル密度 ${}_s\lambda_t(\omega)$ は，$P$ 次の全極型スペクトル密度で表される．

$${}_s\lambda_t(\omega) = \frac{{}_s\sigma_t^2}{|A_t(e^{j\omega})|^2} \tag{3}$$

$$A_t(z) = 1 - a_{t,1}z^{-1} - \cdots - a_{t,P}z^{-P} \tag{4}$$

$a_{t,k}$ と ${}_s\sigma_t^2$ はそれぞれ，線形予測分析における予測係数及び予測残差のパワーと等価である．

(2) $S_{t,l}$ は平均 0，分散 ${}_s\lambda_t(2\pi l/N)$ の複素正規分布にしたがう．

$$p(S_{t,l}) = \mathcal{N}_{\mathbb{C}}\{S_{t,l}; 0, {}_s\lambda_t(2\pi l/N)\} \tag{5}$$

ただし，$\mathcal{N}_{\mathbb{C}}\{\mathbf{x}; \mu, \Sigma\}$ は，平均 $\mu$，共分散行列 $\Sigma$ の複素正規分布の確率密度関数である．

(3) 雑音は定常である．そのパワースペクトル密度を ${}_d\lambda(\omega)$ として（時間 $t$ に依存しないことに注意），$D_{t,l}$ は平均 0，分散 ${}_d\lambda(2\pi l/N)$ の複素正規分布にしたがう．

$$p(D_{t,l}) = \mathcal{N}_{\mathbb{C}}\{D_{t,l}; 0, {}_d\lambda(2\pi l/N)\} \tag{6}$$

(4) 任意の $(t, l, t', l')$ について，$S_{t,l}$ と $D_{t',l'}$ は統計的に独立である．また，$(t,l) \neq (t',l')$ ならば，$S_{t,l}$ と $S_{t',l'}$，及び $D_{t,l}$ と $D_{t',l'}$ は各々独立である．

以上の仮定から，$\mathcal{Y} = \{Y_{t,l}\}_{0\le t\le T-1, 0\le l\le N-1}$ と $\mathcal{X} = \{X_{t,l}\}_{0\le t\le T-1, 0\le l\le N-1}$ の同時分布の確率密度関数は，次式のように書ける．

$$\begin{aligned} p(\mathcal{Y}, \mathcal{X}; \Theta) \propto & \Big(\prod_{l=0}^{N-1} {}_d\lambda(2\pi l/N)^{-T}\Big)\Big(\prod_{t=0}^{T-1} ({}_s\sigma_t^2)^{-N}\Big) \\ & \times \exp\Big\{-\sum_{t=0}^{T-1}\sum_{l=0}^{N-1}\Big(\frac{|Y_{t,l} - X_{t,l}|^2}{{}_d\lambda(2\pi l/N)} \\ & + \frac{|A_t(e^{j\frac{2\pi l}{N}})|^2}{{}_s\sigma_t^2}|X_{t,l} - \sum_{k=1}^{K_l} g^*_{k,l}X_{t-k,l}|^2\Big)\Big\} \end{aligned} \tag{7}$$

*Simultaneous Suppression of Noise and Reverberation for Speech Enhancement, by Takuya Yoshioka, Tomohiro Nakatani, and Masato Miyoshi. (NTT Communication Science Laboratories)

ただし，$\Theta = \{{}_s\Theta, {}_g\Theta, {}_d\Theta\}$ であり，${}_s\Theta = \{a_{t,1}, \cdots, a_{t,P}, {}_s\sigma_t^2\}_{0\le t\le T-1}$ は原音声信号に関するパラメータの集合，${}_g\Theta = \{g_{1,l}, \cdots, g_{K_l,l}\}_{0\le l\le N-1}$ は室内伝達系に関するパラメータの集合，${}_d\Theta = \{{}_d\lambda(2\pi l/N)\}_{0\le l\le N-1}$ は雑音信号に関するパラメータの集合である．

## 3 原音声信号の MMSE 推定

今，仮に，パラメータ $\Theta$ の真値が既知と考える．すると，$S_{t,l}$ の最小平均二乗誤差（MMSE）推定値 $\hat{S}_{t,l} = \langle S_{t,l}\rangle_{p(S_{t,l}|y;\Theta)}$ は，以下の通り求められる．ただし，表記 $\langle f(x)\rangle_{p(x)}$ は，確率変数 $x$ の確率密度関数が $p(x)$ であるときの，関数 $f(x)$ の期待値を表す．

$\mathbf{Y}_l = [Y^*_{T-1,l}, \cdots, Y^*_{0,l}]^H$ とおく．$\mathbf{S}_l$ や $\mathbf{X}_l$ も同様に定義される．まず，$\mathbf{X}_l$ の事後分布 $p(\mathbf{X}_l|y;\Theta)$ を計算する．$p(\mathbf{X}_l|y;\Theta)$ は複素正規分布に従い，その共分散 $\Sigma_l$ と平均 $\mu_l$ は，それぞれ次式で計算される．

$$\Sigma_l = (B_l B_l^H + G_l A_l A_l^H G_l^H)^{-1} \quad (T \text{ 次正方行列}) \tag{8}$$

$$\mu_l = \Sigma_l (B_l B_l^H)\mathbf{Y}_l \quad (T \text{ 次列ベクトル}) \tag{9}$$

$G_l$ は第一列が $[1, -g^*_{1,l}, \cdots, -g^*_{K_l,l}]^H$，第一行が $[1, 0, \cdots, 0]$ である T 次の Toeplitz 行列，$A_l$ と $B_l$ は，それぞれ ${}_s\lambda_{T-t}(2\pi l/N)^{-1/2}$ 及び ${}_d\lambda_{T-t}(2\pi l/N)^{-1/2}$ を t 番目の対角要素にもつ T 次の対角行列である．すると，$\mu_{T-t,l}$ を $\mu_l$ の第 t 要素として，$\hat{S}_{t,l}$ は次式により求められる．

$$\hat{S}_{t,l} = \mu_{t,l} - \sum_{k=1}^{K_l} g^*_{k,l}\mu_{t-k,l} \tag{10}$$

実際には，$\Theta$ の真値は未知であるから，これを推定しなければならない．${}_d\Theta$ は，観測信号の無音声区間から予め推定しておく．これを ${}_d\hat{\Theta}$ とおく．一方，$\Psi = \{{}_s\Theta, {}_g\Theta\}$ は $y$ から推定される．

## 4 パラメータの最尤推定法

本稿では，最尤推定法によりパラメータ $\Psi$ の真値を推定する．すなわち，$p(y;\Psi, {}_d\hat{\Theta})$ を最大化する $\hat{\Psi}$ を，$\Psi$ の推定値とする．

統計モデル (7) は潜在変数 $\mathcal{X}$ を含むから，EM アルゴリズムを用いる．すなわち，$\mathcal{X}$ の条件つき事後分布 $r(\mathcal{X}) = p(\mathcal{X}|y;\hat{\Psi}, {}_d\hat{\Theta})$ と $\hat{\Psi}$ を，交互に繰り返し更新する．EM アルゴリズムでは，本来，補助関数 $q(\Psi) = \langle \log p(y, \mathcal{X}; \Psi, {}_d\hat{\Theta})\rangle_{r(\mathcal{X})}$ を最大化する $\tilde{\Psi}$ として，$\hat{\Psi}$ を更新する．しかしながら，$q(\Psi)$ の最大化課題を解析的に解くことはできないため，${}_s\Theta$ と ${}_g\Theta$ をそれぞれ，一方を固定しながら他方を最適化する．これは ECM アルゴリズムと呼ばれ，EM アルゴリズムと同様に，解の局所最適性が保証される．

結局，$\Psi$ を推定するアルゴリズムは，以下のようにまとめられる．

(1) パラメータの推定値 $\hat{\Psi} = \{{}_s\hat{\Theta}, {}_g\hat{\Theta}\}$ を初期化する．

(2) (E-step)
$r(\mathcal{X}) = \prod_{l=0}^{N-1} p(\mathbf{X}_l|y;\hat{\Psi}, {}_d\hat{\Theta})$（を規定するパラメータ）を計算する．右辺にある $\mathbf{X}_l$ の条件つき事後分布は，前章で与えたものと同じである．

(3) (CM-step1)
$q(\Psi)$ を ${}_g\Theta = {}_g\hat{\Theta}$ の条件下で最大化する ${}_s\tilde{\Theta}$ を計算し，${}_s\hat{\Theta} \leftarrow {}_s\tilde{\Theta}$ とする．${}_s\tilde{\Theta}$ は，各時間 t について，パワースペクトルが $\langle |X_{t,l} - \sum_{k=1}^{K_l} \hat{g}^*_{k,l} X_{t-k,l}|^2\rangle_{r(\mathcal{X})}$ で与えられる時系列を線形予測分析して求められる．$r(\mathcal{X})$ による期待値は解析的に書けるが，紙数の制約により省略する．

(4) (CM-step2)
$q(\Psi)$ を ${}_s\Theta = {}_s\hat{\Theta}$ の条件下で最大化する ${}_g\tilde{\Theta}$ を計算し，${}_g\hat{\Theta} \leftarrow {}_g\tilde{\Theta}$ とする．$\mathbf{g}_l = [g^*_{1,l}, \cdots, g^*_{K_l,l}]^H$ とすると，各 $\tilde{\mathbf{g}}_l$ は次のように求められる．

$$\tilde{\mathbf{g}}_l = {}_xR_t^{-1}\ {}_x\mathbf{r}_t \tag{11}$$

$${}_xR_t = \sum_{t=0}^{T-1} \frac{\hat{A}_t(e^{j\frac{2\pi l}{N}})}{{}_s\hat{\sigma}_t^2}\langle \mathbf{x}_{t-1,l}\mathbf{x}_{t-1,l}^H\rangle_{r(\mathcal{X})} \tag{12}$$

$${}_x\mathbf{r}_t = \sum_{t=0}^{T-1} \frac{\hat{A}_t(e^{j\frac{2\pi l}{N}})}{{}_s\hat{\sigma}_t^2}\langle \mathbf{x}_{t-1,l}X^*_{t,l}\rangle_{r(\mathcal{X})} \tag{13}$$

ただし，$\mathbf{x}_{t,l} = [X^*_{t,l}, \cdots, X^*_{t-K_l+1,l}]^H$，$\hat{A}_t(z)$ は (4) で $a_{t,k} = \hat{a}_{t,k}$ としたものである．

(5) 収束していれば終了．さもなければ (2) に戻る．

## 5 実験結果とまとめ

提案法の雑音と残響の抑圧性能を評価する実験を行った．JNAS データベースに含まれる 10 名の発話について，各々先頭 3 秒の音声を抜粋し，8 kHz にダウンサンプリングして用いた．これら音声信号に，残響時間が約 0.5 秒の部屋で測定したインパルス応答を畳み込んで残響音声信号を合成し，これに SN 比が 10 dB になるように白色雑音信号を加算したものを観測信号とした．ただし，ここでの SN 比は，残響音声信号と雑音信号の比率である．時間周波数解析には STFT を用いた．フレーム長は 256 点，シフト幅は 128 点，窓関数はハニング窓とした．回帰次数は $K_l = 30$ $(0 \le l \le N-1)$, 極の個数は $P = 12$，ECM アルゴリズムの繰り返し回数は 5 回とした．

性能指標として，次式で定義される，振幅スペクトルの SN 比の平均値（SASNR と呼ぶ）を用いた．

$$\frac{1}{T}\sum_{t=0}^{T-1} 10\log_{10}\frac{\sum_{l=0}^{N-1}|S_{t,l}|^2}{\sum_{l=0}^{N-1}||S_{t,l}| - |\hat{S}_{t,l}||^2} \quad (\text{dB}) \tag{14}$$

提案法は，SASNR を平均で 7.72 dB 改善した．処理後の音声は，高域で雑音がやや過剰に抑圧され，くぐもっていたものの，雑音と残響がともによく抑圧されていた．一方，提案法において残響除去処理，あるいは雑音抑圧処理を行わなかった場合，SASNR の平均改善値は，それぞれ 4.46 dB，1.49 dB に低下した．

以上の結果から，提案法は雑音抑圧と残響除去を効果的に統合し，これによって雑音と残響をともに抑圧できたことが確認された．非定常な雑音のレベルを適応的に推定することは，今後の課題である．

## 参考文献


[1] K. Kinoshita, *et al.*, *Proc. Interspeech'07*, pp. 854–857, 2007.
[2] T. Nakatani, *et al.*, *Proc. ICASSP'08*, *accepted*, 2008.