音声のスパース性と非負制約つき畳み込みモデルに基づくパワースペクトル領域残響除去*

○亀岡弘和，中谷智広，吉岡拓也 (NTT コミュニケーション科学基礎研究所)

1 序 論

室内で発せられた音声信号を，話者から離れた位置にあるマイクロホンで収音する際，マイクロホン収音信号には不可避的に原音声信号の残響成分が混入し，これが原因となって音声の明瞭度が低下する。収音信号から残響成分だけを除去し，音声の明瞭度を向上する技術は，実環境における音声認識システムの性能を向上する目的，テレビ会議システムなどの通信システムにおいて音声の明瞭度を向上する目的，録音音声を明瞭に再生する目的など幅広い用途に有用である。

音響信号から残響を除去する方法として，これまでさまざまなアプローチが提案されている。マイクロホンアレーを用いたアプローチでは，ターゲット音の到来方向を推定してそれ以外の到来方向からの音を抑制するもの [1, 2] や，事前に測定しておいた室内インパルス応答をもとに観測信号に逆フィルタ処理を施してターゲット音を復元するもの [3] 等がある。

一方，単一マイクロホン入力を対象としたアプローチでは，クリーン音声に関する仮定やモデル (調波性, 自己回帰モデル，自己相関関数コードブック等) に基づいて，復元音声ができるだけクリーンな音声らしさを有するように室内インパルス応答の逆フィルタを推定するもの [4, 5, 6, 7] が検討されている。一般に室内インパルス応答は音源位置に応じて時々刻々と著しく変化することがあるため，これらのアプローチにおいては，短い観測信号からいかに頑健に逆フィルタを構成できるかが重要課題となっている。

これと並行した試みとして，サブバンドごとのパワーエンベロープに逆フィルタ処理を行うアプローチも検討されている [8, 9]。このアプローチは，室内インパルス応答のサブバンド信号の中でも音源位置に応じて著しく変化するのは特に位相の方であり，パワーエンベロープに関しては比較的変動しにくいという仮説を基礎としている。音源と室内インパルス応答のサブバンドパワーエンベロープ同士の畳み込みによるモデル化はパワースペクトルの加法性など様々な近似仮定の上に導かれるため，残響除去精度に関してはある程度の限界があることが予想されるが, 上記の室内インパルス応答の逆フィルタ処理に基づくアプローチに比べて音源位置などの環境変化に対してある程度頑健に動作する可能性がある。

本稿では，この利点に着目してサブバンドごとのパワーエンベロープから残響成分を除去するアプローチに準拠し，非負値行列分解 (NMF)[10] と呼ぶ原理をヒントにした新しい解法を提案する。提案法は，音声および室内インパルス応答のサブバンドパワーエンベロープを各離散時刻で自由度をつようにモデル化する ([8, 9] のように特定のパラメトリックな関数を仮定しない) 点，音声のスパース性を尺度に最適化規準を設計する点，両サブバンドパワーエンベロープの非負性を保証した乗法更新アルゴリズムと呼ぶ効率的な反復計算法が導ける点，乗法更新アルゴリズムが FFT(Fast Fourier Transform) を効果的に利用してさらに高効率化ができる点，等が特徴である。

2 提案法の原理

2.1 残響音声のサブバンドパワーエンベロープ

クリーン音声および室内インパルス応答の k 番目 ($k = 1, \cdots, K$) のサブバンド信号をそれぞれ $s_{k,t}$, $h_{k,t}$ とすると，残響音声のサブバンド信号 $x_{k,t}$ は，各サブバンド信号が STFT 型のフィルタバンクからの出力である場合には，特定の仮定の下で

$$x_{k,t} = \sum_{\tau} s_{k,\tau} h_{k,t-\tau} \tag{1}$$

のように近似できる [11]。ただし，$t = 1, \cdots, T$ は時刻に対応するインデックスである。式 (1) より，残響音声のサブバンドパワーエンベロープは

$$\begin{aligned}\left|x_{k,t}\right|^2 &= \sum_{\tau}\sum_{\tau'} s^*_{k,t-\tau} h^*_{k,\tau} s_{k,t-\tau'} h_{k,\tau'} \qquad (2)\\ &= \sum_{\tau}\sum_{\tau'} s^*_{k,t-\tau} s_{k,t-\tau'} \left|h_{k,\tau}\right|\left|h_{k,\tau'}\right| e^{-j\phi_{k,\tau}} e^{j\phi_{k,\tau'}}\end{aligned}$$

と表される。ただし，$h_{k,\tau} = |h_{k,\tau}|e^{j\phi_{k,\tau}}$ である。ここで，$\phi_{k,\tau}$ を $D = [-\pi, \pi)$ 上の一様分布に従う独立な確率変数とすると，$|x_{k,t}|^2$ の期待値は,

$$\begin{aligned}&\mathbb{E}\left[|x_{k,t}|^2\right]\\ &= \sum_{\tau} \left|s_{k,t-\tau}\right|^2 \left|h_{k,\tau}\right|^2 \int_D \frac{1}{2\pi} e^{j\phi_{k,\tau}} e^{-j\phi_{k,\tau}} \mathrm{d}\phi_{k,\tau}\\ &\quad + \sum_{\tau\neq\tau'} s^*_{k,t-\tau} s_{k,t-\tau'} \left|h_{k,\tau}\right|\left|h_{k,\tau'}\right|\\ &\qquad \int_D \frac{1}{2\pi} e^{-j\phi_{k,\tau}} \mathrm{d}\phi_{k,\tau} \int_D \frac{1}{2\pi} e^{j\phi_{k,\tau'}} \mathrm{d}\phi_{k,\tau'}\\ &= \sum_{\tau} \left|s_{k,t-\tau}\right|^2 \left|h_{k,\tau}\right|^2 \qquad (3)\end{aligned}$$

と書ける。以上のように，残響音声のサブバンドパワーエンベロープは，期待値の意味で，クリーン音声と室内インパルス応答のサブバンドパワーエンベロープの畳み込みで表される。

* Power spectrum domain dereverberation based on speech sparseness and non-negative convolution model. by KAMEOKA Hirokazu, NAKATANI Tomohiro and YOSHIOKA Takuya (NTT Communication Science Laboratories)

2.2　音声スパース性に基づく残響除去の問題設定

表記の簡単のため，以後 $S_{k,t} \equiv |s_{k,t}|^2$, $H_{k,t} \equiv |h_{k,t}|^2$ と置く。2.1 節の議論より，残響音声の k 番目のサブバンドのパワーエンベロープ $X_{k,t}$ を

$$X_{k,t} \equiv \sum_{\tau} S_{k,\tau} H_{k,t-\tau} \tag{4}$$

のようにモデル化する。ここで，S, H のスケールの任意性を除くため，便宜的に $\sum_t H_{k,t} = 1$ を仮定する。

音声には時間周波数領域においてエネルギーがまばらにしか存在しない (スパース性) という性質があるが，式 (4) のモデルにおいて，スパース性の性質を $S_{k,t}$ を推定するための手がかりにすることが提案法の狙いである。具体的には，観測信号のサブバンドパワーエンベロープが $Y_{k,t}$ のとき，$X_{k,t} \simeq Y_{k,t}$ となるような非負の $S_{k,t}$, $H_{k,t}$ の中で，できるだけ $S_{k,t}$ がスパースになるような解を求めるのがここでの目的である。今，Y と X との間に，

$$Y_{k,t} = X_{k,t} + \epsilon_{k,t} \tag{5}$$

なる関係があるとする。このモデル化誤差 $\epsilon_{k,t}$ は，2.1 節において立てられた近似仮定に起因するあらゆる誤差を含む。この $\epsilon_{k,t}$ を $\mathcal{N}(0, \sigma^2)$ に従う Gauss 性白色雑音と仮定すると，$Y \equiv (Y_{k,t})_{K \times T}$ に対する $S \equiv (S_{k,t})_{K \times T}$, $H \equiv (H_{k,t})_{K \times T}$ の尤度は,

$$P(Y|S,H,\sigma^2) = \prod_{k,t} \frac{1}{\sqrt{2\pi}\sigma} e^{-(Y_{k,t}-X_{k,t})^2/2\sigma^2} \tag{6}$$

となる。ここで S の事前確率 $P(S)$ を一般化正規分布

$$P(S) = \prod_{k,t} \frac{1}{2\Gamma(1+\frac{1}{p})b} e^{-\frac{{S_{k,t}}^p}{b^p}} \tag{7}$$

とし，$P(H_{k,1}, \cdots, H_{k,T})$ を k ごとに独立な一様分布 ($\sum_t H_{k,t} = 1$ の制約があるため，正確にはパラメータがすべて 1 の Dirichlet 分布) とすると，事後確率は

$$P(S,H,\sigma^2|Y) \propto P(Y|S,H,\sigma^2)P(S) \tag{8}$$

と表される。ただし，p, b は一般化正規分布の形状を規定する定数であり，$0 < p < 2$ のとき $P(S)$ は優ガウス的となり，スパース性を測るための適切な尺度になる [12]。式 (8) より，$\log P(S,H,\sigma^2|Y)$ を S, H に関して最大化する問題は，p, σ^2, b をあらかじめ定めておく定数とすると，

$$f(S,H) \equiv \sum_{k,t} \left(Y_{k,t} - X_{k,t}\right)^2 + 2\lambda \sum_{k,t} |S_{k,t}|^p \tag{9}$$

を最小化することと等しい。S, H はいずれも非負な量であるので，$S_{k,\tau} \geq 0$, $H_{k,\tau} \geq 0$ となる制約のもとで $f(S,H)$ を最小化する解を求めたい。ただし，λ は p, σ^2, b に依存して決まる適当な定数であり，モデル化誤差に対するコストと第 2 項のスパース性コストの比重に相当する。以上より，最適化問題は以下のようにまとめられる。

$$\begin{aligned} &\text{minimize} \quad f(S,H) \\ &\text{subject to} \quad \sum_t H_{k,t} = 1,\ H_{k,t} \geq 0,\ S_{k,t} \geq 0 \end{aligned} \tag{10}$$

2.3　非負性を保証する乗法更新アルゴリズム

本節では，非負値行列分解の効率的な解法として知られる乗法更新アルゴリズム [10] をヒントにし，S, H の非負性を保証しながら $f(S,H)$ を反復的に小さくできるアルゴリズムを導出する。

まず，S の乗法更新式を導く。1 ステップ前の S と H の更新値をそれぞれ S', H' として,

$$m_{k,t,\tau} = \frac{{S'}_{k,t} H'_{k,t-\tau}}{\sum_{\tau} {S'}_{k,t} H'_{k,t-\tau}} \tag{11}$$

とすると,

$$\begin{aligned} f(S,H') \leq &\sum_{k,t,\tau} \frac{(m_{k,t,\tau} Y_{k,t} - S_{k,t} H'_{k,t-\tau})^2}{m_{k,t,\tau}} \\ &+ \sum_{k,t} \left(p{S'_{k,t}}^{p-2} S_{k,t}^2 + 2|S'_{k,t}|^p - p|S'_{k,t}|^p \right) \end{aligned} \tag{12}$$

が成り立つ。上不等式右辺を $\bar{f}(S)$ と置く。証明は省略するが，$\bar{f}(S)$ を最小化するように $S_{k,t}$ を更新すれば $f(S,H)$ の非増加性が保証される。$\frac{\partial \bar{f}(S)}{\partial S_{k,\tau}} = 0$ を解くと，$f(S,H)$ の非増加性が保証される更新式

$$S_{k,\tau} = S'_{k,\tau} \frac{\sum_t H'_{k,t-\tau} Y_{k,t}}{\sum_t H'_{k,t-\tau} X'_{k,t} + \lambda p |{S'}_{k,\tau}|^{p-1}} \tag{13}$$

を得る。ただし,

$$X'_{k,t} = \sum_{\tau} S'_{k,\tau} H'_{k,t-\tau} \tag{14}$$

である。式 (13) のとおり，S の更新値が 1 ステップ前の S' と更新係数との積となるため，このような形の更新式を乗法更新式という [10]。また，式 (13) より，S' および H' の要素がすべて非負値であれば更新係数は非負となるため，S の要素はすべて非負値に更新される ($S'_{k,t} = 0$ であれば $S_{k,t} = 0$ となる)。

次に H の乗法更新式を導く。同様に，1 ステップ前の S と H の更新値をそれぞれ S', H' として，式 (11) を用いると,

$$\begin{aligned} f(S',H) \leq &\sum_{k,t,\tau} \frac{(m_{k,t,\tau} Y_{k,t} - S'_{k,t} H_{k,t-\tau})^2}{m_{k,t,\tau}} \\ &+ 2\lambda \sum_{k,t} |S'_{k,t}|^p \end{aligned} \tag{15}$$

が成り立つ。上不等式右辺を $\widetilde{f}(H)$ と置く。同様に $\frac{\partial \widetilde{f}(H)}{\partial H_{k,\tau}} = 0$ を解くと，乗法更新式

$$H_{k,\tau} = H'_{k,\tau} \frac{\sum_t S'_{k,t-\tau} Y_{k,t}}{\sum_t S'_{k,t-\tau} X'_{k,t}} \tag{16}$$

を得る。ただし，上記更新式の導出においては，$\sum_t H_{k,t} = 1$ の拘束は考慮していないので，式 (16) の更新後に規格化する必要がある。

2.4　畳み込み NMF との関係

以上のアルゴリズムは畳み込み NMF[13] の特殊型と解釈できる。通常の NMF では 1 次元配列の基底を想定するのに対し，2 次元配列の基底を想定するのが畳み込み NMF の基本的な考え方である。具体的には，観測行列を $\mathbf{Y}$ とすると，

$$\mathbf{Y} \simeq \sum_{j=1}^{J} \mathbf{W}_j \overset{j\rightarrow}{\mathbf{U}} \tag{17}$$

となるように $\mathbf{W}_1, \cdots, \mathbf{W}_J$ と $\mathbf{U}$ を求めるのが畳み込み NMF である。ただし，$\overset{j\rightarrow}{(\cdot)}$ は行列の成分をすべて $j-1$ 個分右にシフトする演算子とする。例えば，

$$\overset{1\rightarrow}{\mathbf{A}} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}, \quad \overset{2\rightarrow}{\mathbf{A}} = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0 & 5 & 6 & 7 \end{pmatrix},$$

$$\overset{3\rightarrow}{\mathbf{A}} = \begin{pmatrix} 0 & 0 & 1 & 2 \\ 0 & 0 & 5 & 6 \end{pmatrix}, \quad \overset{4\rightarrow}{\mathbf{A}} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 5 \end{pmatrix} \tag{18}$$

となる。$\mathbf{W}_j = (\mathbf{w}_{1,j}, \mathbf{w}_{2,j}, \cdots, \mathbf{w}_{I,j})$ とすると，$\mathbf{w}_{i,1}$, $\mathbf{w}_{i,2}$, $\cdots$, $\mathbf{w}_{i,J}$ を並べたものが i 番目の 2 次元配列の基底ということになる。

ここで，式 (4) を以上の表記の流儀に従って表す。$\mathbf{H}_t$ を，$H_{1,t}, H_{2,t}, \cdots, H_{K,t}$ を対角要素にした行列

$$\mathbf{H}_t \equiv \begin{pmatrix} H_{1,t} & & 0 \\ & \ddots & \\ 0 & & H_{K,t} \end{pmatrix} \tag{19}$$

とし，$\mathbf{S}$ を

$$\mathbf{S} \equiv \begin{pmatrix} S_{1,1} & \cdots & S_{1,T} \\ \vdots & & \vdots \\ S_{K,1} & \cdots & S_{K,T} \end{pmatrix} \tag{20}$$

と置くと，

$$\mathbf{X} = \sum_{t=1}^{T} \mathbf{H}_t \overset{t\rightarrow}{\mathbf{S}} \tag{21}$$

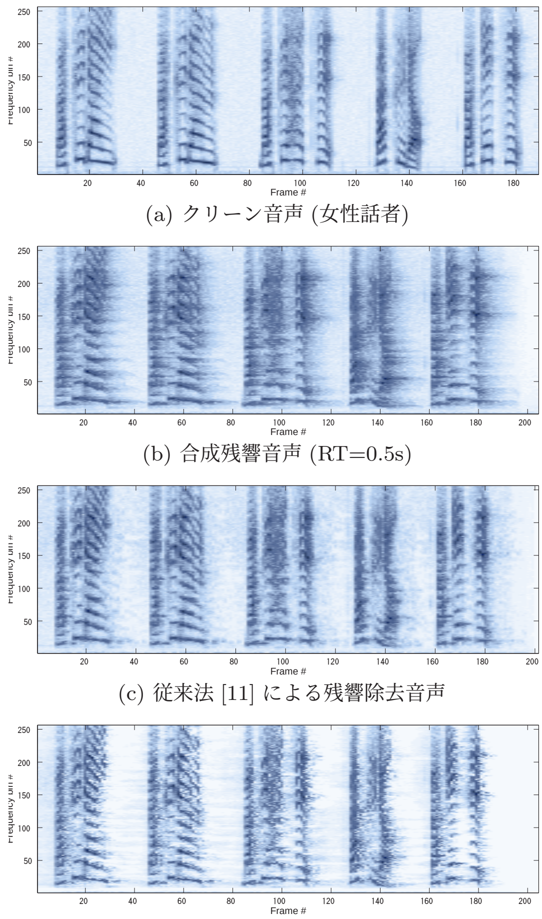
は式 (4) の行列表現となっており，式 (17) と対比すると同型のモデルであることが分かる。ただし，式 (17) に対し，$\mathbf{H}_t$ が対角行列である点で式 (4) の残響音声モデルは畳み込み NMF モデルの特殊型と言える。

式 (13), (16) の乗法更新式はそれぞれ同様に行列表現にすると，

$$\mathbf{S} = \mathbf{S}' \odot \frac{\sum_t \mathbf{H}'_t \overset{t\leftarrow}{\mathbf{Y}}}{\sum_t \mathbf{H}'_t \overset{t\leftarrow}{\mathbf{X}'} + \lambda p \mathbf{S}'^{p-1}} \tag{22}$$

$$\mathbf{H}_t = \mathbf{H}'_t \odot \mathbf{I} \odot \frac{\overset{t\rightarrow}{\mathbf{S}'} \mathbf{Y}^{\mathrm{T}}}{\overset{t\rightarrow}{\mathbf{S}'} \mathbf{X}'^{\mathrm{T}}} \tag{23}$$

と書かれる。ただし，$(\cdot)^p$ は行列要素ごとに p 乗する演算子とする。また，$\odot$ は Hadamard 積 (行列要素ごとの積をとる演算)，行列の除法は要素ごとの商をとる演算とする。



(a) クリーン音声 (女性話者)

(b) 合成残響音声 (RT=0.5s)

(c) 従来法 [11] による残響除去音声

(d) 提案法による残響除去音声

Fig. 1　合成残響に対する提案法の動作結果例

2.5　FFT による乗法更新式の高速計算

式 (13), (16) の更新式には，S' と H' の畳み込み，H' と Y の相関関数，H' と X' の相関関数，S' と Y の相関関数，S' と X' の相関関数の計算を要するが，これらはすべて FFT を使って非常に効率良く計算することができる。提案法は畳み込み NMF の特殊型であるがゆえに 1 回の更新を式 (22), (23) のとおりに計算するよりはるかに効率的に計算できるのである。

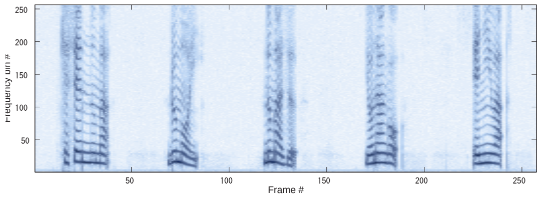
3　動作実験

提案法をさまざまな条件の残響音声に対して適用し動作確認を行った。ここにその動作結果を従来法によるそれとともにいくつかを例示する。以下，従来法は [11] の手法を単一チャネル入力で行ったものをさす。実験データの標本化および周波数分析の条件は Table 1 に示すとおりとした。また，提案法の条件に関しては，各パラメータは

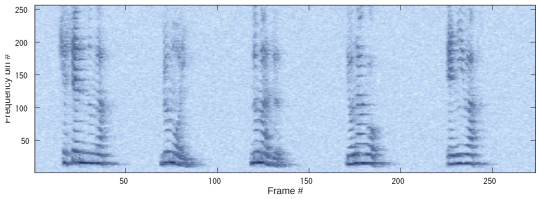
$$p = 1.2,\ \lambda = \frac{E^2}{E^p},\quad E = \sum_{k,t} Y_{k,t} \times 10^{-8}$$

に設定し，反復計算回数は 20 とした。

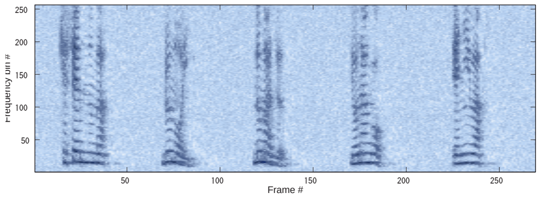
図 1〜4 に，各種実験データに対して提案法および従来法を適用して得られた残響除去音声 (提案法の場合，$S_{k,t}$) のスペクトログラムを，観測信号のスペクト



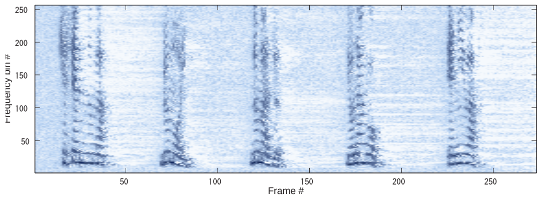
(a) クリーン音声 (女性話者)



(b) 雑音つき合成残響音声



(c) 従来法による残響除去音声



(c) 提案法による残響除去音声

Fig. 2　雑音重畳合成残響に対する提案法の動作結果

Table 1　標本化・周波数分析条件

標本化周波数		8kHz
STFT 条件	フレーム長	64ms
	フレーム周期	32ms
	窓関数	Hanning 窓

ログラムと併せて図示する。図 1 は，ATR 単語データベースの女性話者音声に，可変残響室で測定したインパルス応答 (残響時間 (RT) は 0.5s) を畳み込んで作成した合成残響音声データを，図 2 は，同じく ATR 単語データベースの女性話者音声に，同様の残響時間のインパルス応答を畳み込んで合成した信号にさらに信号対雑音比が 30dB となるように定常の白色 Gauss 雑音を重畳させた雑音つき残響音声データを，図 3, 4 は，それぞれ，男性話者と女性話者が可変残響室内を移動しながら発した音声を実際に収録した実環境音声データを対象とした結果を示している。

4　おわりに

本稿では，音源位置などの環境変化に対して頑健に動作する残響除去法の実現を目的とし，サブバンドごとのパワーエンベロープから残響成分を除去する問題に対し，非負値行列分解 (NMF) の原理をヒントにした新しい解法を提案した。残響音声のサブバンドパワーエンベロープのモデルについての妥当性，音声のスパース性に基づく最適化規準，そのもとで導かれるモデルの制約つき最適化アルゴリズム，等を中心に論じ，さまざまな実験データに対する提案法の動作結果例を示した。

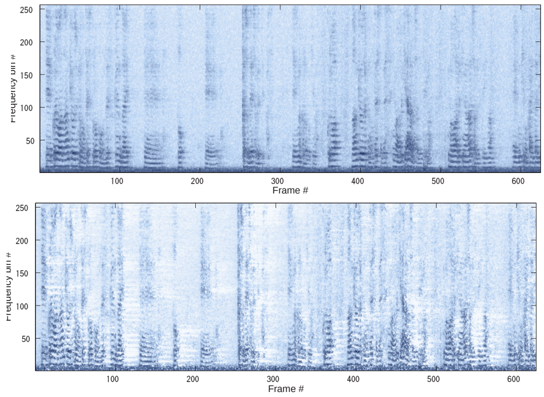


Fig. 3　移動音源 (男性話者) の実環境収録音声 (上段) に対する提案法の動作結果 (下段)

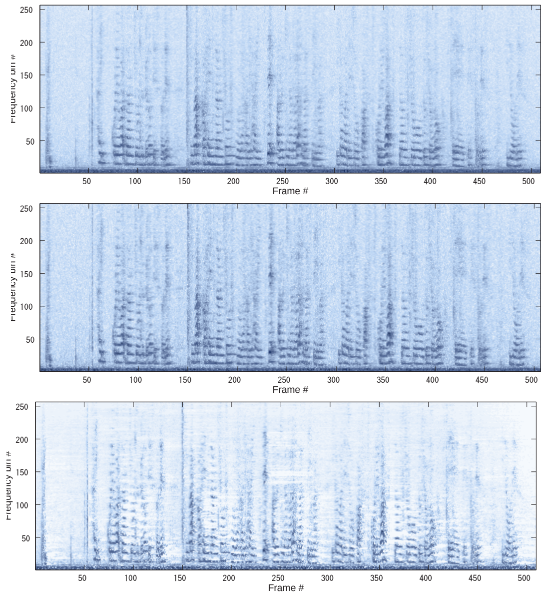


Fig. 4　移動音源 (女性話者) の実環境収録音声 (上段) に対する従来法の動作結果 (中段) と提案法の動作結果 (下段)

参考文献

[1] Flanagan et al., J. Acoust. Soc. Am., 78, 1508–1518, 1985.
[2] Schmidt, IEEE Trans. AP, 34(3), 276–280, 1986.
[3] Miyoshi and Kaneda, IEEE Trans. ASSP, 36(2), 145–152, 1988.
[4] 中谷他, 信学論, J88-D-II(3), 509–520, 2005.
[5] Kinoshita et al., Proc. ICASSP'06, 1, 817–820, 2006.
[6] Yoshioka et al., Proc. IWAENC'06, 2006.
[7] Nakatani et al., Proc. ICASSP'07, 193–197, 2007.
[8] 広林他, 信学論, J81-A(10), 1323–1330, 1998.
[9] Unoki et al., ICASSP'03, 1, 840–843, 2003.
[10] Lee and Seung, Nature, 401, 788–791, 1999.
[11] Nakatani et al., Proc. ICASSP'08, 85–88, 2008.
[12] Karvanen et al., Proc. ICA'03, 125–130, 2003.
[13] Smaragdis et al., Proc. ICA'04, 494–499, 2004.