車室内コミュニケーション用低遅延音源分離手法の検討*

☆上田哲也，井上翔太，牧野昭二，松本光雄，山田武志（筑波大学）

1 はじめに

情報通信技術 (Information and Communications Technology) が急速に発展しており，近年では情報通信技術を車室内に活用したサービスの提供が検討されている [1,2]．車室内では各座席が同じ方向を向いており，エンジン音や他の話者の話し声といった複数の非目的音が存在することから，特に前方の座席と後部座席間での会話が成り立ちにくくなる．そこで我々は運転手の音声を強調して，離れた同乗者に遅延なく提示することで会話を補助する車室内コミュニケーションの実用化を検討している．これを実現する有効なアプローチの一つにブラインド音源分離（BSS）がある．BSS とは観測された信号のみから個々の信号を推定する技術であり，音源に関する情報や音源とマイクロホン間の伝達関数等の事前情報を必要としない利点を持つ．

マイクロホンの数が音源数を上回る優決定条件下の BSS においては，音源信号間の独立性を最大化するように分離フィルタを推定する独立成分分析 (Independent Component Analysis: ICA) [3] が有用であることが知られており，近年では周波数領域での分離手法が数多く提案されている [4–6]．これらの手法は，周波数領域で成立する音源に関する様々な仮定やマイクロホンアレーの周波数応答に関する仮定を有効に活用できるという利点がある．例えば，独立低ランク行列分析 (Independent Low-Rank Matrix Analysis: ILRMA) [5,6] は，各音源信号のパワースペクトログラムを非負値行列とみなし，非負値行列因子分解 (Non Negative Matrix Factorization: NMF) [7] で近似表現する手法である．これは，各時間フレームにおけるパワースペクトルを時間的に変化する振幅によってスケーリングされた基底スペクトルの線形和で近似することに相当する．

周波数領域での分離手法は，観測信号が入力されてからその信号が短時間フーリエ変換 (STFT) されるまでの待機遅延が不可避である．待機遅延は STFT の窓長に依存して生じる遅延であり，CPU の高性能化やアルゴリズム改善により削減不可能である．周波数領域での分離手法の車室内コミュニケーションへの適用には低遅延が要求されるため，STFT の窓長を短くする必要がある．一方で，ILRMA は周波数領域において瞬時混合近似に基づいた定式化がなされているため，STFT の窓長に対して観測信号の残響時間が長い条件下においては分離性能が低下する．一般に車室内の残響時間は短いが，STFT の窓長を短くした場合，その影響は無視できなくなる．この影響は STFT の窓長を大きくすることで緩和できるが，STFT の待機遅延が生じ，低遅延での動作が困難となる．

一方，長い残響環境下を想定した BSS 手法として，周波数領域における畳み込み混合近似に基づいた ILRMA が提案されている [8]．本稿ではこの手法を畳み込み混合に基づく ILRMA と呼ぶ．[8] では音源分離実験によって，残響時間が STFT の窓長を上回る状況においての周波数領域の畳み込み混合近似の有効性が示されている．本研究ではこのアプローチを特に車室内の短い残響環境下での音源分離に適用することで，低遅延化を図る．

2 畳み込み混合に基づく ILRMA

2.1 瞬時混合近似に基づいた定式化

I 個のマイクロホンで J 個の音源から到来する信号を観測する場合を考える．i 番目のマイクロホンで観測される信号と j 番目の音源信号の時間周波数成分をそれぞれ $x_i(f,n)$ と $s_j(f,n)$ とする．ただし，f と n は周波数と時間フレームのインデックスである．以下では優決定条件下において $I=J$ とする．

まず，音源とマイクロホンの間の室内インパルス応答長が STFT における窓長よりも短い場合を考える．音源信号 $\boldsymbol{s}(f,n)=[s_1(f,n),\ldots,s_J(f,n)]^{\mathsf{T}}\in\mathbb{C}^J$ と観測信号 $\boldsymbol{x}(f,n)=[x_1(f,n),\ldots,x_I(f,n)]^{\mathsf{T}}\in\mathbb{C}^I$ の関係性は瞬時混合系を用いると

$$\boldsymbol{s}(f,n)=\boldsymbol{W}^{\mathsf{H}}(f)\boldsymbol{x}(f,n), \tag{1}$$

$$\boldsymbol{W}^{\mathsf{H}}(f)=[\boldsymbol{w}_1(f),\ldots,\boldsymbol{w}_I(f)]\in\mathbb{C}^{I\times I} \tag{2}$$

と表せる．ここで，$\boldsymbol{W}^{\mathsf{H}}(f)$ は分離フィルタを表し，$(\cdot)^{\mathsf{T}}$ は行列の転置であり，$(\cdot)^{\mathsf{H}}$ はエルミート転置である．

次に，観測信号が生成されるプロセスを生成モデルにより記述する．音源 j の複素スペクトログラム $s_j(f,n)$ を

$$s_j(f,n)\sim\mathcal{N}_{\mathbb{C}}(s_j(f,n)|0,v_j(f,n)) \tag{3}$$

*Low-latency blind source separation for in-car communication by Tetsuya Ueda, Shota Inoue, Shoji Makino, Mitsuo Matsumoto, Takeshi Yamada (University of Tsukuba).

のように平均が 0，分散が $v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$ の複素正規分布に従う確率変数と仮定する．各音源が統計的に独立である場合，$\boldsymbol{s}(f,n)$ は

$$\boldsymbol{s}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}(f,n)|0, \boldsymbol{V}(f,n)) \tag{4}$$

に従う．ここで，$\boldsymbol{V}(f,n)$ は $v_1(f,n),\ldots,v_I(f,n)$ を対角成分に持つ対角行列である．さらに，分散 $v_j(f,n)$ を基底行列 $H = \{h_{j,k}(f)\}_{j,k,f}$ 及びアクティベーション行列 $U = \{u_{j,k}(n)\}_{j,k,n}$ の二つの非負行列の積で表せると仮定すると $v_j(f,n)$ は

$$v_j(f,n) = \sum_{k=1}^{K} h_{j,k}(f) u_{j,k}(n) \tag{5}$$

のように表せる．ここで，$k = 1, ..., K$ は NMF の基底インデックスを表す．式 (1) と式 (4) より，観測信号 $\boldsymbol{x}(f,n)$ は

$$\boldsymbol{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{x}(f,n)|0, (\boldsymbol{W}^{\mathsf{H}}(f))^{-1}\boldsymbol{V}(f,n)\boldsymbol{W}(f)^{-1}) \tag{6}$$

に従う．従って，観測信号 $\mathcal{X} = \{\boldsymbol{x}(f,n)\}_{f,n}$ が与えられたときの分離行列 $\mathcal{W} = \{W(f,0)\}_f$ 及び各音源のパワースペクトログラム $\mathcal{V} = \{H, U\}$ についての尤度関数は以下のように記述できる．

$$\mathcal{I}(\mathcal{W}, \mathcal{V}|\mathcal{X}) \stackrel{c}{=} -2N \log|\det \boldsymbol{W}^{\mathsf{H}}(f)| + \sum_{f,\,n,\,j} \left(\log v_j(f,n) + \frac{|\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{y}(f,n)|^2}{v_j(f,n)} \right) \tag{7}$$

ここで，$\stackrel{c}{=}$ はパラメータに依存しない項を除いた等号を表す．

この瞬時混合近似に基づいた手法は STFT の窓長が残響時間より十分に長い必要がある．そのため，この手法を用いて STFT の窓長を短くすると瞬時混合近似の仮定を十分に満たさなくなり分離性能が低下する．

2.2 畳み込み混合近似に基づいた定式化

周波数領域での瞬時混合近似が十分に成り立たない環境下を想定した BSS 手法として畳み込み混合近似に基づいた手法が提案されており，音源分離実験によって分離性能の向上が示されている [8,9]．音源信号と観測信号の関係は次式のような有限インパルス応答フィルタの形式として記述できる．

$$\boldsymbol{s}(f,n) = \sum_{n'=0}^{N'} \boldsymbol{W}^{\mathsf{H}}(f,n')\boldsymbol{x}(f,n-n') \tag{8}$$

ここで，$\boldsymbol{W}(f,n')$, $0 \le n' \le N'$ は $I \times I$ 行列の分離フィルタであり，音源分離とともにフレーム外に及ぶ残響成分を除去する役割を担ったパラメータである．$\boldsymbol{W}^{\mathsf{H}}(f,0)$ は式 (1) の分離フィルタに対応する．N' はフィルタ $\{\boldsymbol{W}^{\mathsf{H}}(f,n')\}_{f,n'}$ の次数である．$\boldsymbol{W}^{\mathsf{H}}(f,0)$ を正則な行列であると仮定すると，残響除去後の観測信号 $\boldsymbol{y}(f,n) = [y_i(f,n),\ldots,y_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I$ と音源信号 $\boldsymbol{s}(f,n)$ は

$$\boldsymbol{y}(f,n) = \boldsymbol{x}(f,n) - \sum_{n'=1}^{N'} \boldsymbol{D}^{\mathsf{H}}(f,n')\boldsymbol{x}(f,n-n'), \tag{9}$$

$$\boldsymbol{s}(f,n) = \boldsymbol{W}^{\mathsf{H}}(f,0)\boldsymbol{y}(f,n) \tag{10}$$

のように書き直せる．ここで，$\boldsymbol{D}^{\mathsf{H}}(f,n') = -(\boldsymbol{W}^{\mathsf{H}}(f,0))^{-1}\boldsymbol{W}^{\mathsf{H}}(f,n')$, $1 \le n' \le N'$ である．式 (9) は観測信号 $\boldsymbol{x}(f,n)$ に含まれる残響成分を除去するプロセスに相当し，$\mathcal{D} = \{\boldsymbol{D}^{\mathsf{H}}(f,n')\}_{f,n'}$ は長さ N' の残響除去フィルタと見なせる．式 (10) は残響を除去した信号 $\boldsymbol{y}(f,n)$ に対して，周波数ごとの音源分離を行うプロセスとして解釈できる．この関係式に 2.1 節の $\boldsymbol{S} = \{s_j(f,n)\}_{j,f,n}$ の生成モデルを組み込むことで，分離行列 $\mathcal{W}$，残響除去フィルタ $\mathcal{D}$ の尤度関数を得ることができる．観測信号 $\mathcal{X}$ が与えられたときの残響除去フィルタ $\mathcal{D}$，分離行列 $\mathcal{W}$ 及び各音源のパワースペクトログラム $\mathcal{V}$ の尤度関数は以下のように記述できる．

$$\mathcal{I}(\mathcal{D}, \mathcal{W}, \mathcal{V}|\mathcal{X}) \stackrel{c}{=} -2N \log|\det \boldsymbol{W}^{\mathsf{H}}(f)| + \sum_{f,n,j} \left(\log v_j(f,n) + \frac{|\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{y}(f,n)|^2}{v_j(f,n)} \right) \tag{11}$$

2.3 最適化アルゴリズム

本節では，観測信号 $\mathcal{X}$ が与えられた下で，残響除去フィルタ $\mathcal{D}$，分離行列 $\mathcal{W}$ 及び各音源のパワースペクトログラム $\mathcal{V}$ についての対数尤度関数 (11) を最小化するアルゴリズムについて述べる．この最適化問題の大域最適解は解析的に求めることはできないが，局所最適解は

$$\hat{\mathcal{V}} \leftarrow \underset{\mathcal{V}}{\operatorname{argmin}}\, \mathcal{I}(\mathcal{D}, \mathcal{W}, \mathcal{V}|\mathcal{X}), \tag{12}$$

$$\hat{\mathcal{W}} \leftarrow \underset{\mathcal{W}}{\operatorname{argmin}}\, \mathcal{I}(\mathcal{D}, \mathcal{W}, \mathcal{V}|\mathcal{X}), \tag{13}$$

$$\hat{\mathcal{D}} \leftarrow \underset{\mathcal{D}}{\operatorname{argmin}}\, \mathcal{I}(\mathcal{D}, \mathcal{W}, \mathcal{V}|\mathcal{X}) \tag{14}$$

を繰り返すことで数値探索することができる．

$\mathcal{V}$ の更新は以下の通りである．

$$h_{j,k}(f) = h_{j,k}(f) \sqrt{\frac{\sum_n |s_j(f,n)|^2 u_{j,k}(n) v_j^{-2}(f,n)}{\sum_n u_{j,k}(n) v_j^{-1}(f,n)}}, \tag{15}$$

$$u_{j,k}(n) = u_{j,k}(n) \sqrt{\frac{\sum_f |s_j(f,n)|^2 h_{j,k}(f) v_j^{-2}(f,n)}{\sum_f h_{j,k}(f) v_j^{-1}(f,n)}}, \tag{16}$$

$\mathcal{W}$ の更新には反復射影法 (Iterative Projection: IP)

$$\boldsymbol{w}_j(f) \leftarrow (\mathbf{W}^{\mathsf{H}}(f)\boldsymbol{\Sigma}_j(f))^{-1}\boldsymbol{e}_j, \tag{17}$$

$$\boldsymbol{w}_j(f) \leftarrow \frac{\boldsymbol{w}_j(f)}{\sqrt{\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{\Sigma}_j(f)\boldsymbol{w}_j(f)}} \tag{18}$$

を用いることができる．ただし，$\boldsymbol{\Sigma}_j(f) = \frac{1}{N}\sum_n \frac{\boldsymbol{x}(f,n)\boldsymbol{x}^{\mathsf{H}}(f,n)}{v_j(f,n)}$ であり，$\boldsymbol{e}_j$ は $I \times I$ の単位行列 $\boldsymbol{I}$ の第 j 列のベクトルである．

$\mathcal{D}$ の更新は $\boldsymbol{D}(f,n')$ の第 i 列のベクトルを $\boldsymbol{d}_i(f,n')$ として

$$\begin{aligned}\boldsymbol{d}(f) &= \mathrm{vec}(\{\boldsymbol{D}(f,n')\}) \\ &= [\boldsymbol{d}_1^{\mathsf{T}}(f,1),\ldots,\boldsymbol{d}_I^{\mathsf{T}}(f,1),\boldsymbol{d}_1^{\mathsf{T}}(f,2),\ldots,\boldsymbol{d}_I^{\mathsf{T}}(f,2), \\ &\quad \ldots,\boldsymbol{d}_1^{\mathsf{T}}(f,N'),\ldots.,\boldsymbol{d}_I^{\mathsf{T}}(f,N')]^{\mathsf{T}} \in \mathbb{C}^{I^2N'}\end{aligned} \tag{19}$$

のようにベクトル形式に変形することで解析的に求めることができ，$\boldsymbol{d}^*(f)$ の更新式として

$$\begin{aligned}\boldsymbol{d}^*(f) \leftarrow &\left(\sum_n \boldsymbol{X}^{\mathsf{H}}(f,n)\boldsymbol{\Sigma}_{w/v(f,n)}\boldsymbol{X}(f,n)\right)^{-1} \\ &\times \left(\sum_n \boldsymbol{X}^{\mathsf{H}}(f,n)\boldsymbol{\Sigma}_{w/v(f,n)}\boldsymbol{x}(f,n)\right)\end{aligned} \tag{20}$$

が得られる．ここで，$\boldsymbol{\Sigma}_{w/v(f,n)} = \sum_j \frac{\boldsymbol{w}_j(f)\boldsymbol{w}_j^{\mathsf{H}}(f)}{v_j(f,n)}$ であり，$(\cdot)^*$ は複素共役を表す．また，

$$\begin{aligned}\boldsymbol{X}(f,n) = [&\boldsymbol{I}\otimes\boldsymbol{x}^{\mathsf{T}}(f,n-1),\boldsymbol{I}\otimes\boldsymbol{x}^{\mathsf{T}}(f,n-2),\ldots, \\ &\boldsymbol{I}\otimes\boldsymbol{x}^{\mathsf{T}}(f,n-N')] \in \mathbb{C}^{I\times I^2N'}\end{aligned} \tag{21}$$

である．$\boldsymbol{I}$ と $\otimes$ はそれぞれ $I \times I$ の単位行列とクロネッカー積を表す．

従って，各変数の最適化までの流れは以下のようにまとめられる．

1. $\mathcal{V}$，$\mathcal{W}$ と $\mathcal{D}$ を初期化する．
2. 各 j，f，n について下記の更新を繰り返す．
   (a) 式 (15)，(16) を用いた $h_{j,k}(f), u_{j,k}(n)$ の更新．
   (b) 式 (17)，(18) を用いた $\boldsymbol{w}_j(f)$ の更新．
   (c) 式 (20) を用いた $\boldsymbol{d}^*(f)$ の更新．

2.4 残響除去手法の低遅延化の応用

車室内コミュニケーションでは，発話者音声の観測信号が入力されてから分離信号が出力されるまでの間で遅延時間が生じ，その時間が 12 ms より大きいと発話者音声と分離音声が反響音声に聞こえ，音声聴取に影響を与えることが示されている [2]．その為，STFT の窓長を短くする必要があるが，窓長を短くすると瞬時混合近似の仮定を十分に満たさなくなり分離性能が低下する．

それに対し [8] では，上述の手法を用いることで窓長を超える残響成分を含む信号の分離精度が向上することが示されている．我々はこの点に注目し，STFT に起因する待機遅延の低減への応用を検討する．本稿では，特に車室内の短い残響環境下の音源分離アプローチとして周波数領域での畳み込み混合を用いる．

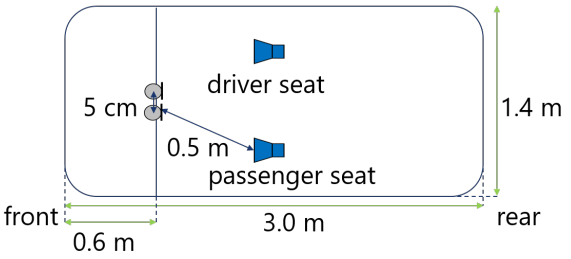
3 評価実験

3.1 実験の概要

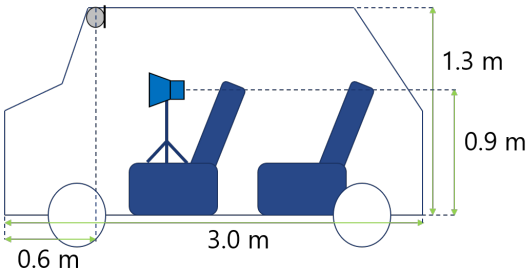
車室内環境下で畳み込み混合に基づく ILRMA を低遅延化に適用した手法の有効性を評価するために，音源分離実験を行った．STFT の窓長が残響時間を下回る条件での分離性能の変化を確認するため，STFT の窓長 L，残響除去フィルタ長 N'，分離性能の関係を調査した．実験に用いた多チャンネルの観測信号は実測の車室内インパルス応答とドライソースを畳み込むことで生成した．分離性能の評価には，signal-to-distortion ratio (SDR) [10] を用いる．SDR は値が高い程高い分離性能を示す．

3.2 実験条件

本実験では，ATR デジタル音声データベースのセット B に収録されている，全 503 文の音素バランス文の男性 6 話者，女性 4 話者計 10 話者分のデータを使用した．このデータセットの中からランダムに選択した 2 人の異なる話者の発話に対し，実測の車室内インパルス応答を畳み込むことで，2 チャンネルの観測信号を 10 パターン作成した．観測信号のサンプリング周波数は 8 kHz とした．インパルス応答は車室内で録音した時間引き延ばしパルス（Time-Stretched Pulse: TSP）を用いて測定した．音源を自動車の運転席と助手席に配置し，マイクロホンを車室内前方のマップランプに取り付けて TSP 信号を録音した．音源とマイクロホンの配置を Fig. 1 に示す．車室内の残響時間（T_{60}）は 58 ms であった．実験では L を {2，4，8，16, 32} ms とした場合の残響除去フィルタ長の変化に対する SDR の平均値を評価した．残響除去フィルタ長 N' を大きくすると残響除去フィルタの更新時間が大きくなるため本実験では $0 \le N' \le 10$ とした．$N' = 0$ の場合は瞬時混合に基づく ILRMA を用いた場合と等価である．また，窓長 L が残響時間に比べて十分に長い条件での比較対象として，$L =$ 128 ms での瞬時混合モデルに基づく ILRMA を用いた場合の SDR の平均値も示す．畳み込み混合に基づく ILRMA を用いる際に，STFT の移動長は窓長の 4 分の 1 とした．また，反復更新を 50 回行った．ILRMA での音声信号分離では基底数が多い場合に分離性能が劣化することが確認されているため [6]，基



(a) Plan view



(b) Side view

Fig. 1: Sound source and microphone layout in experiment

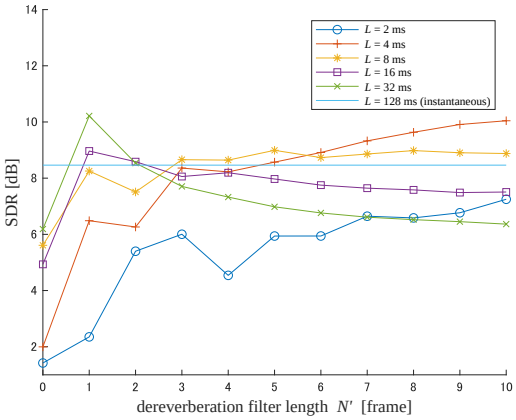


Fig. 2: Average SDR according to STFT frame length L and dereverberation filter length N'

底数を 1 とした．

3.3 実験結果と考察

各窓長 L における残響除去フィルタ長 N' の変化に対する SDR の平均値を Fig. 2 に示す．図より，残響除去フィルタ長 N' を適切に選択することで，いずれの L においても SDR の向上が確認できる．特に，$4 \leq L \leq 32$ ms において SDR の平均値が $L = 128$ ms 及び $N' = 0$ の場合を上回る結果となった．以上より，周波数領域での畳み込み混合モデルに基づく ILRMA を車室内における音源分離に用いることで，SDR を維持しつつ，STFT に起因する待機遅延の削減が可能であることが示された．

4 まとめ

本稿では，畳み込み混合に基づく ILRMA の車室内環境下での低遅延な音源分離への応用を検討した．実測のインパルス応答を用いた音声の分離実験を実施し，車室内において分離性能を維持しつつ STFT に起因する待機遅延の削減が可能であることを示した．

謝辞 本研究は科研費 19H04131，SECOM 科学技術振興財団，サポインの助成を受けた．

参考文献

[1] R. Landgraf *et al.*, “Can you hear me now? reducing the lombard effect in a driving car using an in-car communication system,” in Proc. Speech Prosody, pp. 479–483, 2016.

[2] A. Theiss *et al.*, “Instrumental evaluation of in-car communication systems,” in Proc. ITG, pp. 1–4, 2014.

[3] A. Hyvärinen *et al.*, “Independent component analysis,” Wiley, 2001.

[4] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” in Proc. WASPAA, pp. 189–192, 2011.

[5] H. Kameoka *et al.*, “Statistical model of speech signals based on composite autoregressive system with application to blind source separation,” in Proc. LVA/ICA, pp. 245–253, 2010.

[6] D. Kitamura *et al.*, “Determined blind source separation with independent low-rank matrix analysis,” *Audio Source Separation*, Springer, pp. 125–155, 2018.

[7] D. D. Lee *et al.*, “Algorithms for non-negative matrix factorization,” in Proc. NIPS, pp. 556–562, 2001.

[8] H. Kagami *et al.*, “Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization,” in Proc. ICASSP, pp. 31–35, 2018.

[9] T. Yoshioka *et al.*, “Blind separation and dereverberation of speech mixtures by joint optimization,” in IEEE Trans. ASLP, vol. 19, no. 1, pp. 69–84, 2010.

[10] E. Vincent *et al.*, “Performance measurement in blind audio source separation,” IEEE Trans. ASLP, vol. 14, no. 4, pp. 1462–1469, 2006.