

伝達関数ゲイン基底 NMF を用いた遠方雑音抑圧の実環境での評価

Multiple far noises suppression under real environments with transfer-function-gain NMF

松井裕太郎[†]

牧野昭二[†]

小野順貴[‡]

山田武志[†]

[†]筑波大学

[‡]国立情報学研究所 / 総研大

Yutaro MATSUI[†]

Shoji MAKINO[†]

Nobutaka ONO[‡]

Takeshi YAMADA[†]

[†]University of Tsukuba

[‡]National Institute of Informatics / SOKENDAI

アブストラクト 到来方向が特定できないような、いわゆる背景雑音を効果的に抑圧できる手法として、伝達関数ゲイン基底 NMF を適用した遠方雑音の抑圧手法がある。この手法では、背景雑音が遠方から到来することを仮定し、時間周波数領域における振幅情報のみに着目する。これにより様々な方向から到来する複数の遠方雑音を一つの混合雑音源としてモデル化し、その振幅の混合モデルを伝達関数ゲイン基底 NMF に適用することで遠方雑音の抑圧を行う [1]。これまで本手法を用いた遠方雑音抑圧の効果は、シミュレーション実験により確認されていた。本論文では、非同期録音された、背景雑音の混入した音声に本手法を適用し、実環境における性能を調査した。その結果、本手法により、実環境においても距離の異なる複数の遠方雑音を効果的に抑圧できることを確認した。

1 はじめに

会話や会議の録音などにおける音声品質向上を目的とした雑音抑圧技術について、特定方向からの雑音に対し、指向性を制御することによって雑音抑圧を行う手法がこれまで提案されてきた。このような雑音抑圧手法では、雑音の到来方向が特定可能な場合は高い効果を発揮するが、到来方向が特定できない、いわゆる背景雑音の処理は難しかった。そこで、様々な方向から到来する複数の遠方雑音をひとつの混合雑音源としてモデル化し、その振幅モデルを伝達関数ゲイン基底 NMF に適用して遠方雑音抑圧を行う手法が提案された [1]。本手法では、背景雑音が遠方から到来すると仮定することで、様々な方向から到来する複数の遠方雑音を一つの混合雑音源としてモデル化することが可能となる。本手法は、シミュレーション実験において有効性が確認されている。本論文では、実環境において複数の雑音を非同期録音し、その雑音群に対して伝達関数ゲイン基底 NMF を適用し、実環境におけ

る本手法の有効性を検証する。

非同期録音された音声に対して適用可能な、振幅領域における雑音抑圧手法としては、SN 比最大化ビームフォーマ [2,3] を振幅情報のみで行う振幅スペクトルビームフォーマ [4] のような手法がある。振幅スペクトルビームフォーマは観測音声に含まれるすべての音源を事前に単一音源区間から学習しておく必要がある。これに対して伝達関数ゲイン基底 NMF はすべての音源に対するそのような学習は必要でない。従って、本論文では背景雑音に対して有効な振幅領域に置ける雑音抑圧手法として、伝達関数ゲイン基底 NMF を適用した。

伝達関数ゲイン基底 NMF による雑音抑圧手法は、これまで罰則付き伝達関数ゲイン基底 NMF [5] や、伝達関数ゲインを事前情報により学習する教師あり伝達関数ゲイン基底 NMF [6]、半教師あり伝達関数ゲイン基底 NMF [1] などが提案されている。本論文では、背景雑音が常時存在する環境下で雑音抑圧を行うことを想定し、雑音の伝達関数ゲインを事前情報により学習したもので固定し、目的音の伝達関数ゲインとアクティベーションを NMF により推定する半教師あり伝達関数ゲイン基底 NMF を適用して雑音抑圧を行う。

2 遠方音源の混合モデル

2.1 振幅領域での混合モデル [1, 6]

強調を行う目的音源と K 個の雑音源を、 M 個のマイククロホンで録音する。非同期録音では、録音機器間のサンプリング周波数ミスマッチにより、観測信号間の位相ずれが発生する。この位相ずれのサンプル数が信号処理における STFT フレーム長より十分に小さいと考えると、伝達関数の振幅は時不変と仮定できる。そこで、観測信号の振幅の加法性を仮定したモデルを導入し、観測信号

の位相情報を用いない混合モデルを式 (1) のように表す.

$$\bar{\mathbf{X}}(\omega) \approx \bar{\mathbf{X}}^{\mathbf{S}}(\omega) + \bar{\mathbf{X}}^{\mathbf{N}}(\omega) \quad (1)$$

ここで ω は離散時間フーリエ変換の角周波数, $\mathbf{X}^{\mathbf{S}}$ は目的音の観測信号, $\mathbf{X}^{\mathbf{N}}$ は雑音の観測信号を表し, 各記号のオーバーラインはその記号が示す要素の絶対値を表すものとする. $\bar{\mathbf{X}}(\omega), \bar{\mathbf{X}}^{\mathbf{S}}(\omega), \bar{\mathbf{X}}^{\mathbf{N}}(\omega)$ は $M \times N$ の行列であり, それぞれ m 行 n 列に $\bar{X}_{mn}(\omega), \bar{X}_{mn}^{\mathbf{S}}(\omega), \bar{X}_{mn}^{\mathbf{N}}(\omega)$ の要素を持つ. N は時間フレーム数である. さらに, 目的音 $\bar{\mathbf{X}}^{\mathbf{S}}(\omega)$ は

$$\bar{\mathbf{X}}^{\mathbf{S}}(\omega) = \bar{\mathbf{a}}^{\mathbf{S}}(\omega) \bar{\mathbf{s}}^{\mathbf{S}}(\omega) \quad (2)$$

と表される. ここで $\bar{\mathbf{a}}^{\mathbf{S}}(\omega)$ は $M \times 1$ の列ベクトルであり, 要素 $\bar{a}_m^{\mathbf{S}}(\omega)$ は目的音から m 番目のマイクロホンまでの伝達関数の絶対値を表す. $\bar{\mathbf{s}}^{\mathbf{S}}(\omega)$ は $1 \times N$ の行ベクトルであり, 要素 $\bar{s}_n^{\mathbf{S}}(\omega)$ は目的音の n 番目の時間フレームにおける時間周波数成分の絶対値を表す. また, 雑音 $\bar{\mathbf{X}}^{\mathbf{N}}(\omega)$ は K 個の雑音源の足し合わせによって

$$\bar{\mathbf{X}}^{\mathbf{N}}(\omega) \approx \sum_{k=1}^K \bar{\mathbf{a}}^{\mathbf{N}^k}(\omega) \bar{\mathbf{s}}^{\mathbf{N}^k}(\omega) \quad (3)$$

と表される. ここで $\bar{\mathbf{a}}^{\mathbf{N}^k}(\omega)$ は $M \times 1$ の列ベクトルで, 要素 $\bar{a}_m^{\mathbf{N}^k}(\omega)$ は k 番目の雑音から m 番目のマイクロホンまでの伝達関数の絶対値を表す. $\bar{\mathbf{s}}^{\mathbf{N}^k}(\omega)$ は $1 \times N$ の行ベクトルであり, 要素 $\bar{s}_n^{\mathbf{N}^k}(\omega)$ は k 番目の雑音の n 番目の時間フレームにおける時間周波数成分の絶対値を表す.

以上のように振幅領域における混合モデルを導入する. このようなモデルを伝達関数ゲイン基底 NMF [5, 6] に適用する場合, 音源数が既知でなければならない. しかし, 本論文で取り扱う背景雑音は雑音源の数 K が未知であるため, 雑音抑圧を行うことができない. 従って, 次節では背景雑音を抑圧するための混合モデルについて示す.

2.2 背景雑音抑圧のための遠方音源の混合モデル [1]

前節のような観測において, 点音源 k における 1 番目のマイクロホンと m 番目のマイクロホンとの音圧レベルの差は

$$L_1 - L_m = 10 \log \frac{P_1^2}{P_0^2} - 10 \log \frac{P_m^2}{P_0^2} \quad (m = 2, \dots, M) \quad (4)$$

と表せる. ここで P_m, L_m はそれぞれ点音源 k における各マイクロホンの音圧, 音圧レベルを表す. P_0 は人間の最小可聴音圧の基準値 $P_0 = 20 \times 10^{-6} [\text{Pa}]$ である. このとき, m 番目のマイクロホンの音圧 P_m は 1 番目のマイクロホンの音圧 P_1 を用いて

$$P_m^2 = \frac{\beta_m}{\beta_1} \times \frac{r_m^2}{r_1^2} \times P_1^2 \quad (5)$$

と表せる. ここで β_m, r_m はそれぞれ m 番目のマイクロホンの感度, 点音源 k までの距離を表す. 式 (4), (5) より, マイクロホン間の音圧レベルの差は

$$L_1 - L_m = -10 \log \frac{\beta_m}{\beta_1} - 20 \log \frac{r_1}{r_m} \quad (m = 2, \dots, M) \quad (6)$$

と表せる. ここで点音源 k がマイクロホン間隔に対して十分遠方から到来すると仮定できるならば $r_1 \approx r_m$ となり, マイクロホン間のゲイン差は音源の音量・方向に依存せず, マイクロホンの感度のみ依存していることが分かる. すなわち音源が遠方にあると見なせる場合には, 全ての雑音で伝達関数ゲイン $\bar{\mathbf{a}}_k^{\mathbf{N}}(\omega)$ が類似する.

そこで, 遠方音源を構成する全ての雑音の伝達関数ゲインを代表とする伝達関数ゲインで近似することによって, 遠方から到来する全ての音源を一つの遠方音源として近似する. このとき, 振幅領域における遠方音源の観測信号, 伝達関数ゲイン, 絶対値振幅をそれぞれ $\widehat{\mathbf{X}}^{\mathbf{N}}(\omega), \widehat{\mathbf{a}}^{\mathbf{N}}(\omega), \widehat{\mathbf{s}}^{\mathbf{N}}(\omega)$ とすると, これらは式 (7)-(9) のように表される.

$$\widehat{\mathbf{X}}^{\mathbf{N}}(\omega) \approx \widehat{\mathbf{a}}^{\mathbf{N}}(\omega) \widehat{\mathbf{s}}^{\mathbf{N}}(\omega) \quad (7)$$

$$\widehat{\mathbf{a}}^{\mathbf{N}}(\omega) \approx \bar{\mathbf{a}}^{\mathbf{N}^1}(\omega) \approx \dots \approx \bar{\mathbf{a}}^{\mathbf{N}^K}(\omega) \quad (8)$$

$$\widehat{\mathbf{s}}^{\mathbf{N}}(\omega) \approx \sum_{k=1}^K \bar{\mathbf{s}}^{\mathbf{N}^k}(\omega) \quad (9)$$

ここで $\widehat{\mathbf{a}}^{\mathbf{N}}(\omega)$ は $M \times 1$ の列ベクトルであり, 要素 $\widehat{a}_m^{\mathbf{N}}(\omega)$ は (一つの) 遠方音源から m 番目のマイクロホンまでの伝達関数ゲインを表す. $\widehat{\mathbf{s}}^{\mathbf{N}}(\omega)$ は $1 \times N$ の行ベクトルであり, 要素 $\widehat{s}_n^{\mathbf{N}}(\omega)$ は n 番目の時間フレームにおける, 遠方音源を構成する背景雑音の絶対値振幅の総和を表す.

以上により, 遠方音源と見なせる背景雑音が存在する場合の, 振幅領域での混合モデルは式 (10)-(12) のようになる.

$$\bar{\mathbf{X}}(\omega) \approx \bar{\mathbf{a}}(\omega) \bar{\mathbf{s}}(\omega) \quad (10)$$

$$\bar{\mathbf{a}}(\omega) \approx [\bar{\mathbf{a}}^{\mathbf{S}}(\omega) \widehat{\mathbf{a}}^{\mathbf{N}}(\omega)] \quad (11)$$

$$\bar{\mathbf{s}}(\omega) \approx \begin{bmatrix} \bar{\mathbf{s}}^{\mathbf{S}}(\omega) \\ \widehat{\mathbf{s}}^{\mathbf{N}}(\omega) \end{bmatrix} \quad (12)$$

本論文では, 以上のような観測において背景雑音抑圧を行う. 具体的には, 非同期の録音機器を配置し, 目的音源との距離が最も近いマイクロホンを 1 番目のマイクロホンとして目的音, 背景雑音を録音し, $\bar{a}_1^{\mathbf{S}}(\omega)$ が $\bar{a}_j^{\mathbf{S}}(j = 1, \dots, M)$ 中で最大になるようにする. そして, 目的音が全マイクロホン中最も高い SN 比で録音される 1 番目のマイクロホンの観測信号 $\bar{X}_{1n}(\omega)$ に対する時間周波数マスクを作成することで背景雑音を抑圧する.

以降では同様の処理を各周波数ビンごとに行うため、周波数ビンを表す ω は各記号中から省略する。

3 伝達関数ゲイン基底 NMF を用いた遠方雑音抑圧 [1]

3.1 伝達関数ゲイン基底 NMF を用いた雑音抑圧

伝達関数ゲイン基底 NMF を用いた雑音抑圧は、前章で述べた背景雑音の振幅領域における混合モデルに対して、時間チャンネル領域の NMF を適用し雑音抑圧を行う [5,6]. NMF は非負の行列を、2つの非負行列により

$$\bar{\mathbf{X}} \approx \tilde{\mathbf{X}} = \tilde{\mathbf{a}}\tilde{\mathbf{s}} \quad (13)$$

のように低ランク近似する手法である [7]. ここで $\tilde{\mathbf{X}}, \tilde{\mathbf{a}}, \tilde{\mathbf{s}}$ は観測に対するそれぞれの推定解を表し、要素の次元は式 (10)-(12) と同じである. このような低ランク近似では、行列の要素が非負であるという制約により、解がスパースなものに限定される. このため、適切な条件の下では基底 $\tilde{\mathbf{a}}$ が伝達関数ゲイン $\bar{\mathbf{a}}$ の同定となり、アクティベーション $\tilde{\mathbf{s}}$ が絶対値振幅 $\bar{\mathbf{s}}$ の推定となるような行列分解となる.

また、低ランク近似を行うには元の行列と推定行列との距離を最小化する必要がある. この最小化の距離尺度には様々なものを用いることができるが、本論文では I ダイバージェンス基準を用い、この際の更新式は式 (14),(15) のようになる.

$$\tilde{a}_m^i \leftarrow \tilde{a}_m^i \frac{\sum_n \bar{x}_{mn} \tilde{s}_n^i}{\sum_n \tilde{s}_n^i} \quad (14)$$

$$\tilde{s}_m^i \leftarrow \tilde{s}_m^i \frac{\sum_n \bar{x}_{mn} \tilde{a}_n^i}{\sum_n \tilde{a}_n^i} \quad (15)$$

さらに、NMF ではそれぞれの周波数ビンにおいて分離信号の各周波数成分が異なる順番で現れるパーミュテーション問題が発生する. この問題を解決するため、目的音源が 1 番目のマイクロホンに最も近い距離にあり、 $\bar{a}_1^S(\omega)$ が $\bar{a}_j^S(j = 1, \dots, M)$ 中で最大であるという仮定を利用し、目的音の基底 $\bar{\mathbf{a}}^S$ の初期値を

$$\bar{a}_m^S = \begin{cases} 1 - (1 - M)\alpha & (m = 1) \\ \alpha & (\text{otherwise}) \end{cases} \quad (16)$$

とする. ここで α は $0 < \alpha < 1/(M - 1)$ を満足する任意の実数である. 背景雑音の基底 $\bar{\mathbf{a}}^N$ の初期値は

$$\bar{a}_m^N = \frac{1}{M} \quad (m = 1, \dots, M) \quad (17)$$

とする.

目的音強調信号 $\tilde{\mathbf{Y}}_n$ は、SN 比が最も高い 1 番目のマイクロホンの観測信号 X_{1n} と、推定解 $\tilde{\mathbf{a}}, \tilde{\mathbf{s}}$ から作成したウィナーマスクを用いて

$$\tilde{\mathbf{Y}}_n = X_{1n} \frac{(\tilde{a}_1^S \tilde{s}_n^S)^2}{(\tilde{a}_1^S \tilde{s}_n^S)^2 + (\tilde{a}_1^N \tilde{s}_n^N)^2} \quad (18)$$

と表される.

3.2 半教師あり伝達関数ゲイン基底 NMF

前節で示した伝達関数ゲイン基底 NMF によるパラメータ推定は、音源数とマイクロホン数が近い場合は低ランク近似の拘束力としては弱く、十分な雑音抑圧性能を得られない [8]. この解決法として、事前に伝達関数のゲインを学習しておくことによって最適解に近いアクティベーションを推定する手法がある [1].

教師あり伝達関数ゲイン基底 NMF は、目的音、雑音それぞれの伝達関数ゲインを事前に学習しておく手法である. まず、目的音のみ、雑音のみの単一音源区間を録音し、単一音源信号 $\mathbf{X}^S, \mathbf{X}^N$ を得る. 次に、それらの単一音源に NMF を適用し、それぞれの伝達関数ゲイン基底 $\tilde{\mathbf{a}}^S, \tilde{\mathbf{a}}^N$ を求める. そして、雑音抑圧を行う観測 (すなわち、目的音と雑音の混合信号) に対して NMF を適用する際に、式 (16),(17) の初期値の代わりに学習で得た $\tilde{\mathbf{a}}^S, \tilde{\mathbf{a}}^N$ を用い、式 (14) の更新を行わずに基底を初期値のまま固定し、式 (15) のアクティベーション更新のみを行う.

以上のように、教師あり伝達関数ゲイン基底 NMF は目的音・雑音の双方に対して学習を行う. しかし、背景雑音のような雑音が常駐する状況下では目的音の単一音源信号を得ることが難しく、一方で雑音の単一音源信号は得るのが容易である. 従って、本論文では雑音のみ事前学習によって求めた基底で固定する半教師あり伝達関数ゲイン基底 NMF [1] を適用して雑音抑圧を行う. この場合、目的音に関しては式 (14),(16) を適用する.

4 評価実験

4.1 実験概要

前章までで示した混合モデルと半教師あり伝達関数ゲイン基底 NMF によって、遠方音源の抑圧ができるかどうかの実環境における検証を行う. 実験条件を Table 1 に示す. 実験では Fig. 1 のように遠方から到来する複数の静止した雑音源による混合遠方雑音を抑圧できるか確かめる. 中央のマイクロホンの詳細は Fig. 2 のようになる. 雑音源の配置を Table 2 に示す. 評価に用いる音声は、Fig. 1 に示したそれぞれのスピーカから再生した音声を非同期録音することによって作成した. 目的音のスピーカに最も近い Mic. 2 を、ウィナーマスクをかける観測信号を録音するマイクロホンとする. また、実験に使用したマイク

Table 1: Experimental conditions

目的音源数	1(音声)
遠方雑音源数	8(音声)
マイクロホン数	3, 6, 9
録音のサンプリング周波数	16 kHz
STFT フレーム長	4,096 samples
STFT フレームシフト幅	2,048 samples
雑音抑圧区間	10 s
学習に用いる信号の区間	10 s
距離尺度	I-divergence
初期パラメータ α	0.1
NMF 更新回数	200
残響時間	640 ms
暗騒音	35.3~35.8 dB

Table 2: Arrangement of noises

No.	目的音源に対する角度	マイクロホン群の中心からの距離
1	25°	250 cm
2	75°	300 cm
3	115°	250 cm
4	155°	200 cm
5	185°	350 cm
6	225°	300 cm
7	265°	350 cm
8	305°	200 cm

Table 3: Channels and microphone numbers

チャンネル数	3	6	9
マイク番号	2, 5, 9	1, 2, 3, 5, 7, 9	6 以外

ロホンのチャンネル数と使用マイクロホンの関係を Table 3 に示す。評価尺度には SDR・SIR を用いた [9]。SDR は強調信号の歪みを、SIR は雑音信号の抑圧度合を表し、いずれも値が大きいほど性能が高いことを示す。評価対象は半教師あり伝達関数ゲイン基底 NMF を適用した信号 (SSNMF) と未処理の信号 (unpro) である。

4.2 実験結果

実験結果を Fig. 3, Fig. 4 に示す。Fig. 3 は SSNMF と unpro の SDR を示し、Fig. 4 は SIR を示している。これらの結果より、SNR が $-5, 0, 5$ [dB] いずれの場合も SDR, SIR の値が大きくなり目的音強調がなされていること、また、マイクロホン数が増加するにつれて目的音強調の性能が向上することが分かる。SDR がいずれの場合もマイクロホン数が 3 から 6 に増加した場合には大きく SDR が

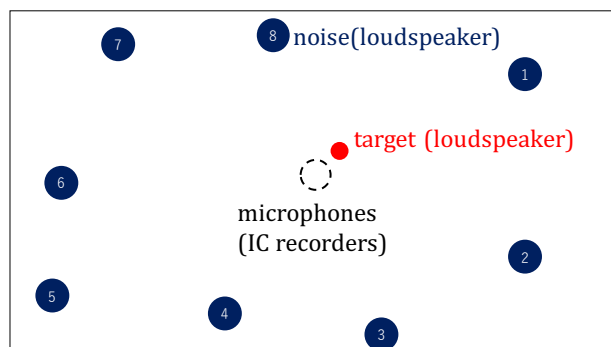


Fig. 1: Arrangement of target, multiple far noises, and microphones

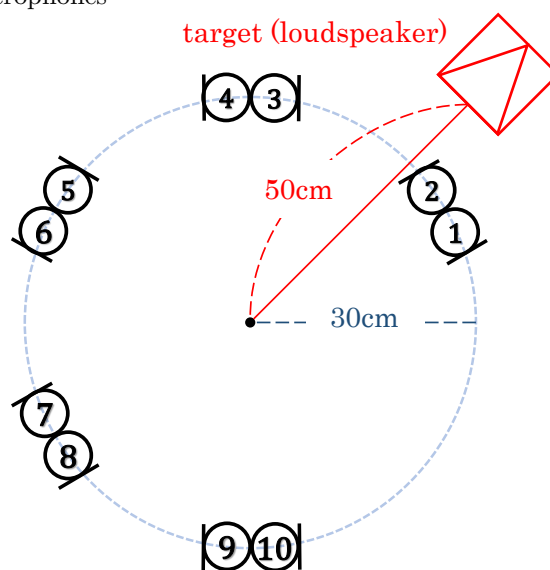


Fig. 2: Arrangement of microphones (IC recorders)

向上しているのと比べて、6 から 9 に増加した場合には SDR の向上値が小さい。これは Fig. 2 に示すマイクロホンが 3 チャンネルから 6 チャンネルになった場合では 3 番、7 番のマイクロホンの増加によって配置の状態が大きく変わったのに対し、6 チャンネルから 9 チャンネルになった場合にはあまり変わっていないからであると考えられる。

4.3 シミュレーション実験 [1] との比較

前節までで示した実験とシミュレーション実験で同じ点として次のような条件が挙げられる。

- 雑音源の数
- 録音のサンプリング周波数及び伝達関数ゲイン基底 NMF を適用する際のパラメータ

異なる点としては次のような条件が挙げられる。

- 残響時間
- サンプリング周波数ミスマッチ

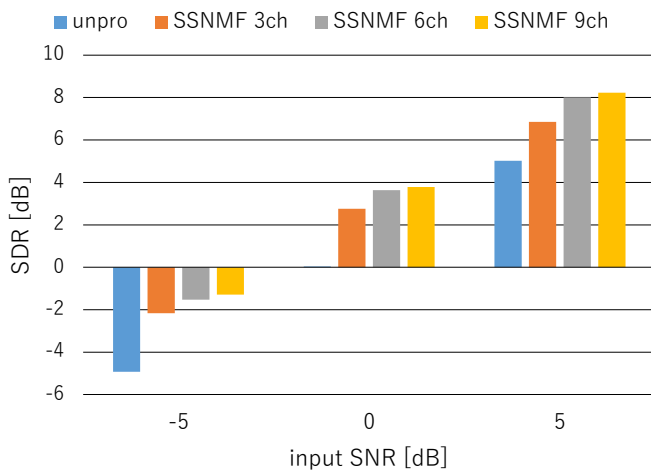


Fig. 3: Results of SDR

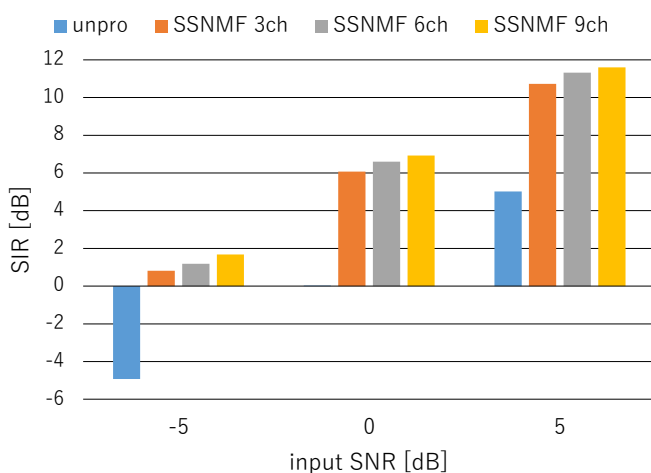


Fig. 4: Results of SIR

- マイクロホンアレーの配置
- 雑音源の配置

シミュレーション実験, 実環境ともに 8 個の雑音源に対して 3 チャンルの音声に伝達関数ゲイン基底 NMF を適用して SDR の改善が見られることから, 実用的な範囲のマイクロホン数で背景雑音を抑圧する際に本手法は有効であると考えられる。一方で, 残響時間やサンプリング周波数 mismatches はいずれも本実験の方が悪条件であり, 本実験の SDR, SIR 改善量はいずれもシミュレーション実験よりも小さくなった (シミュレーション実験では SNR=0 で 4~5[dB], SNR=5 で 2~3[dB] の SDR 改善が見られた)。これらの条件と本手法の性能との関係性についてはさらに調査する必要がある。

5 まとめ

本論文では, 伝達関数ゲイン基底 NMF を適用して, 実環境において距離の異なる複数の遠方雑音を抑圧できる

か検証した。その結果, 実環境において目的音源と雑音源それぞれの伝達関数ゲインに差がある, すなわち目的音源に対して雑音源が遠方であると見なせる状況で, シミュレーション実験 [1] と同様に遠方雑音を抑圧することができた。本手法はシミュレーション実験においてはその有効性が確認されていたが, 実環境においても背景雑音抑圧に対して有効であることを確認した。

参考文献

- [1] 村瀬 慶和, 千葉 大将, 小野 順貴, 宮部 滋樹, 山田 武志, 牧野 昭二, “非同期分散マイクロフォンアレーによる伝達関数ゲイン基底 NMF を用いた拡散雑音抑圧,” 音講論集, pp. 557–560, Mar. 2015.
- [2] H. L. Van Trees, ed. *Optimum Array Processing*, Wiley, 2002.
- [3] 荒木 章子, 澤田 宏, 牧野 昭二, “話者分類と SN 比最大化ビームフォーマに基づく会議音声強調,” 音講論集, pp. 571–572, Mar. 2007.
- [4] 加古 達也, 小林 和則, 大室 伸, “非同期分散マイクロフォンアレーのための振幅スペクトルビームフォーマの提案,” 音講論集, pp. 829–830, Mar. 2013.
- [5] M. Togami, Y. Kawaguchi, H. Kokubo, and Y. Obuchi, “Acoustic echo suppressor with multichannel semi-blind non-negative matrix factorization,” *Proc. APSIPA*, pp. 522–525, Dec. 2010.
- [6] 千葉 大将, 小野 順貴, 宮部 滋樹, 山田 武志, 牧野 昭二, 高橋 祐, “伝達関数ゲイン基底 NMF による分散配置非同期録音における目的音強調の検討,” 音講論集, pp. 757–760, Mar. 2014.
- [7] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” *Proc. NIPS*, pp. 556–562, Dec. 2001.
- [8] 村瀬 慶和, 千葉 大将, 小野 順貴, 宮部 滋樹, 山田 武志, 牧野 昭二, “伝達関数ゲイン基底 NMF におけるマイク数・マイク配置と目的音強調性能の関係,” 音講論集, pp. 523–526, Sept. 2014.
- [9] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1462–1469, Jun. 2006.