人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B302-03

# Semi-Blind Infinite NMF を用いた動作雑音抑圧手法の提案とその評価

**Semi-Blind Infinite Non-negative Matrix Factorization for Ego-motion Noise Suppression**

◯手塚太貴 [1] 吉田尚水 [1*] 中臺一博 [1,2]

**Taiki TEZUKA　Takami YOSHIDA　Kazuhiro NAKADAI**

1 東京工業大学　情報理工学研究科, 2 (株) ホンダ・リサーチ・インスティチュート・ジャパン

1 **Graduate School of Information Science and Engineering, Tokyo Institute of Technology,**

2 **Honda Research Institute Japan Co, Ltd.**

tezuka@cyb.mei.titech.ac.jp nakadai@jp.honda-ri.com

## Abstract

ロボット聴覚機能の実現における課題としてロボットの動作雑音が挙げられる. 本論文では, ノンパラメトリックベイジアンモデルの一種である Semi-Blind Infinite Non-negative Matrix Factorization (SB-INMF) を提案し, これを用いた新たな動作雑音抑圧手法を報告する．SB-INMF は動作雑音抑圧に関節角の情報が不要，動作雑音と目的音を線形分離するため，スペクトル減算よりも歪みが少ないといった特長がある．二種類の実ロボットを用いて評価実験を行った結果, 関節角情報を用いた従来の雑音テンプレートによる動作雑音抑圧よりも良好な雑音抑圧結果を得ることができた.

## 1　はじめに

ロボットの聴覚機能，つまり「ロボット聴覚」[1] は人とロボットの音声コミュニケーションを実現する上で重要な技術であり，2000 年に提唱されて以降精力的に研究が行われている．この研究の狙いはロボットに取り付けたマイクロホンを用いて聴覚機能を実現することである．これまでにもバイノーラル [2–7], マイクアレイ [8–11], マルチモーダル [2, 12] そしてユビキタスセンサ [13–15] など様々なアプローチによるロボット聴覚システムが報告されている．ロボット聴覚の重要な要素に音源定位や音源分離，音声認識技術が挙げられるが，ロボットの動作雑音はこれらの実現における大きな課題である．ロボットは荷物の運搬やダンスなど様々な動作中であっても，人間と音声を介したコミュニケーションが可能でなければならない．しかし，ロボットの動作中は必ず動作雑音が発生し，ロボットの聴覚機能を妨げてしまう．そのため，動作雑音を抑圧することの重要性が論じられるとともにこれまでに様々な手法が提案されてきた.

動作雑音抑圧手法には大きく二種類のアプローチがある．１つ目はセンサを用いて動作雑音を測定するアプローチ, ２つ目は関節角のような動作雑音と相関のある情報から動作雑音を推定するアプローチである．１つ目のアプローチとして，中臺らは，ロボット外装の内外にマイクロホンを設置することで，ロボットの内外を区別する音響的身体性を構築し，内部雑音が小さい時間のみ音源定位を行う手法を報告している [1]．しかし, 音源分離のように，内部雑音の有無にかかわらず常に処理が必要な場合には対応が難しい．その他には，ロボットの機体内部に内部雑音の検知用のセンサを取り付け，*Frequency-Domain Blind Signal Separation(FD-BSS)* を適用することで動作雑音を推定する手法が提案されている [16]．しかしこれらの手法は音声収録用のマイクロホン以外のセンサを必要とし, 性能の面で重要となるセンサの配置について議論されていない．そして，これらのアプローチでは動作雑音を推定するために追加したマイクロホンやセンサが必要となるが，これによりシステムが複雑になり，結果として計算コストが大きくなってしまうという問題がある.

２つ目のアプローチは，関節駆動により発生する動作雑音と関節角に強い相関関係があるという事実に基づいている．例えば, 伊藤らは， Sony AIBO を用いて，関節角度や位置を入力として動作雑音を推定するニューラルネットワークを構築し，推定雑音をスペクトル減算 [18] することにより，音声認識性能が向上できることを報告している [17]．しかし，シミュレーション実験しか行っておらず, 残響など実環境ならではのファクターが加わった場合の有効性は不明である．また，西村らは，動作コマンドに対応した雑音テンプレートを構築し，これをスペクトル減算をする手法を提案した．さらに, スペクトル減算によって生じる歪みに対応するため，ミッシングフィーチャ理論を用いて音声認識性能を向上させた [3]．しかし，この手法は

* 現在 株式会社東芝　研究開発センター　知識メディアラボラトリー勤務

Table 1: 変数表記

| 意味 | 表記 |
|---|---|
| マイクロホンの数 | $N_{mic} \in \mathbb{N}$ |
| 周波数ビン数 | $N_f \in \mathbb{N}$ |
| 因子数 | $N_k \in \mathbb{N}$ |
| 基底数 | $K \in \mathbb{N}$ |
| 観測信号のサイズ | $N_s \in \mathbb{N}$ |
| 観測信号のパワースペクトル | $Y \in \mathbb{R}_{+0}^{N_f+1}$ |
| 動作雑音のパワースペクトル | $X \in \mathbb{R}_{+0}^{N_f+1}$ |
| 基底 | $F \in \mathbb{R}_{+0}^{N_f \times N_k}$ |
| アクティベーション | $Z \in \mathbb{R}_{+0}^{N_s \times N_f}$ |
| ゲイン | $\theta \in \mathbb{R}_{+0}^{N_k}$ |

$\mathbb{R}_{+0}$ は正の実数, $\mathbb{N}$ は自然数を表す.

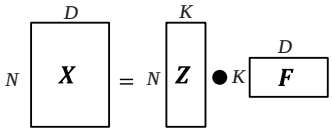



Figure 1: 因子モデル

動作コマンド単位の手法であるため，複数の動作を組み合わせた複雑な動作への対応は難しい．Ince らは，雑音テンプレートを処理フレーム単位で構築する手法を考案し，音源定位，分離，音声認識それぞれに提案手法が有効性であることを報告している [19]．しかし，この手法は動作が同じであれば, 必ず同じ動作雑音が発生すると仮定しており，実環境においてこの仮定が必ずしも成り立つとは言えない．このため, 動作雑音の推定に誤りが生じ，スペクトル減算によって音声認識性能が悪化する原因となる．そこで，本稿ではノンパラメトリックベイジアンモデルである *Semi-Blind Infinite Non-negative Matrix Factorization (SB-INMF)* を提案し，関節角など他のセンサの情報を使わない新たな動作雑音抑圧手法の実現を試みる．本研究で提案する SB-INMF では収録した音信号から直接動作雑音の推定を行うため前述のテンプレート法で発生する推定誤差は生じない．また，線形過程により動作雑音と目的音を分離することが可能であり，スペクトル減算を用いる手法よりも歪みが少ない動作雑音抑圧が可能である．

## 2 SB-INMF による動作雑音抑圧

表 1 に本論文で用いる記号の定義を示す．動作雑音は主に関節が駆動することで発生する．本稿では，動作雑音は各関節からの動作雑音を組み合わせて表現できるものと考える．このような場合，図 1 の様な因子モデル (LFM) を用いて動作雑音を表現することができる．ここで，基底 $\boldsymbol{F}$ は動作雑音の周波数方向の特徴を表し, アクティベーション $\boldsymbol{Z}$ は動作雑音の時間方向の特徴を表す．

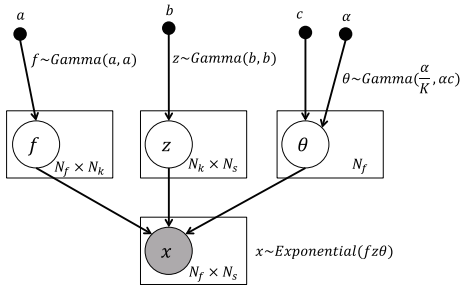



Figure 2: INMF のグラフィカルモデル

具体的には LFM として，モノラル信号の音源分離手法でよく用いられる非負値行列因子分解 (NMF) を用いる．各々の関節から生じる動作雑音パワースペクトルは非負値，かつ加法性であると見なし，収録信号に含まれる動作雑音と目的音を NMF を用いて線形分離する．NMF による音源の分離を行う際, 観測信号の確率分布として一般に以下の指数分布が用いられる [20].

$$x_{nd} \sim \text{Exponential}\left(\sum_k z_{nk} f_{kd}\right), \tag{1}$$

ここで $x_{nd} \geq 0, z_{nk} \geq 0, f_{kd} \geq 0$ は各々行列 $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{F}$ の要素である．NMF による動作雑音の因子モデル推定を行う際に，次に示す問題点を考慮する必要がある．①動作雑音と関節角の情報との間に完全な相関関係は成り立たない．②NMF によって推定された動作雑音の基底の数は関節の数と必ずしも一致しない，すなわち, 適切な基底の数は未知である．これらの問題を扱うために本研究では NMF による因子モデルを拡張した無限因子モデル (ILFM) を導入する．つまり，NMF を無限因子モデルが扱えるモデルように, *Infinite Non-negative Matrix Factorization(IMNF)* [21] に拡張する．これにより因子モデルが基底の候補数を上限なく保有する事を可能とし，かつ目的とするデータを表現するのに最適な基底の数を機械的に推定することが可能になる．

本論文ではノンパラメトリック確率過程の一つであり，スパースな学習を可能とするガンマ過程を用いて無限因子モデルを構築する．動作雑音の INMF を以下の様に定式化する．

$$f_{kd} \sim \text{Gamma}(a, a), \tag{2}$$
$$z_{nk} \sim \text{Gamma}(b, b), \tag{3}$$
$$\theta_k \sim \text{Gamma}(\alpha/K, \alpha c), \tag{4}$$
$$x_{nd} \sim \text{Exponential}\left(\sum_k \theta_k z_{nk} f_{kd}\right) \tag{5}$$

ここで，$a, b, c$ はガンマ分布のパラメータである．このモデルは各基底に対応した非負値のゲイン $\theta_k$ を導入し，不必要な基底のゲイン値を 0 にすることで，最適な基底のみを用いた推定が可能である．なお，このモデルでは非負値行列の各要素の分布がガンマ過程に基づくものであると仮定している．そのため，このモデルは *Gamma Process Non-negative Matrix Factrization(GaP-NMF)* [22] と呼ばれる．

図 2 に式 (2)–(5) に対応したグラフィカルモデルを示す．白と灰色の円はそれぞれ隠れ変数と観測変数を表す．黒い点は与えられたパラメータを表し，複数ノードがある場合はプレートを用いて表す．$x$ が観測信号とした動作雑音に対応する．

このモデルでは, 動作雑音を入力すれば INMF によって，動作雑音の因子モデルを得る事ができる．しかし, 実際の入力信号には音声を始めとする目的音が含まれているため，INMF を適用すると動作雑音と目的音の基底が得られる．このため得られた基底のうち, どの基底が動作雑音に対応し，どの基底が目的音に対応するのかを判別することが難しい．この問題を解決するために *Semi-Blind INMF(SB-INMF)* を提案する．

$$x_{nd} \sim \text{Exponential}\left(\sum_k \theta_k z_{nk} f_{kd} + \sum_l \tilde{\theta}_l \tilde{z}_{nl} \tilde{f}_{ld}\right), \quad (6)$$

ここで $\theta_k$，$z_{nk}$，および $f_{kd}$ は動作雑音に対応し，$\tilde{\theta}_l$，$\tilde{z}_{nl}$，および $\tilde{f}_{ld}$ は目的音に対応する．動作雑音の基底 $f_{kd}$ を予め与えることで，動作雑音のアクティベーション $z_{mk}$，ゲイン $\theta_k$，及び目的音の因子モデル $(\tilde{\theta}_l,\ \tilde{z}_{nl},\ \tilde{f}_{ld})$ を推定する．これにより，動作雑音と音声が混合した信号入力に対しても，動作雑音が抑圧された音声を得ることが可能になる．この手法による動作雑音抑圧はスペクトル減算のような非線形過程を踏まない雑音抑圧のため，目的音の歪みが少ない．INMF のパラメータ推定には GaP-NMF と同様の変分ベイズを用いる [22].

$$\begin{aligned}
q(z, f, \theta) &\approx q(z)q(f)q(\theta) && (7)\\
q(z) &= \text{GIG}(z; \gamma^{(z)}, \rho^{(z)}, \tau^{(z)}) && (8)\\
q(f) &= \text{GIG}(f; \gamma^{(f)}, \rho^{(f)}, \tau^{(f)}) && (9)\\
q(\theta) &= \text{GIG}(\theta; \gamma^{(\theta)}, \rho^{(\theta)}, \tau^{(\theta)}) && (10)\\
\text{GIG}(y; \gamma, \rho, \tau) &= \left(\frac{\rho}{2\tau}\right)^{\gamma/2} \frac{\exp\{(\gamma-1)\log y - \rho y - \tau/y\}}{\mathcal{K}_\gamma(2\sqrt{\rho\tau})} && (11)
\end{aligned}$$

GIG は一般化逆ガウス分布であり，そして $\gamma$, $\rho$, 及び $\tau$ は GIG のパラメータである．$\mathcal{K}_\gamma$ は第二ベッセル関数である．なお $z$，$f$，$\theta$，$\gamma$，$\rho$ 及び $\tau$ の添字は簡単化のため省略した．

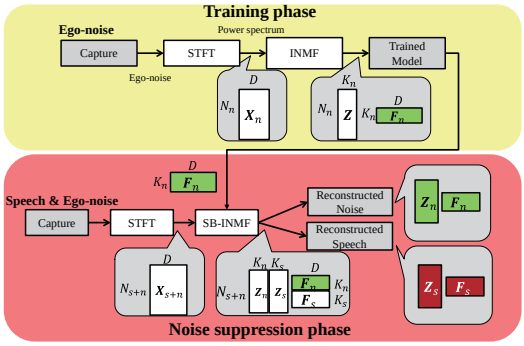



Figure 3: 動作雑音の推定及び抑圧システム

## 3 動作雑音抑圧システム

前節にて提案した SB-INMF に基づく動作雑音抑圧システムを図 3 に示す．図 3 の上部は INMF に基づく動作雑音の学習過程を表す．マイクロホンで収録した動作雑音に短時間フーリエ変換 (STFT) を行う．得られた動作雑音信号のスペクトル $\boldsymbol{X}_n(N_s \times N_f)$ に対して INMF を行い，$N_{k1}$ 個の動作雑音の基底 $\boldsymbol{F}_n$ を得る．また，SB-INMF による動作雑音推定及び抑圧の過程を表す．目的音と動作雑音を含む信号入力に対し，STFT を行うことでスペクトル $\boldsymbol{X}_{s+n}$ を得る．次に，事前に学習した動作雑音のモデル $\boldsymbol{F}_n$ を既知の基底として SB-INMF を行う事で $\boldsymbol{X}_{s+n}$ に含まれた目的音の基底 $\boldsymbol{F}_s$ 及びアクティベーション $\boldsymbol{Z}_s$ を得る．最終的に $\boldsymbol{F}_s$，$\boldsymbol{Z}_s$ より目的音のスペクトルを得る．また，動作雑音も同様に，与えた基底 $\boldsymbol{F}_n$ 及びそのアクティベーション $\boldsymbol{Z}_n$ によって得ることができる．

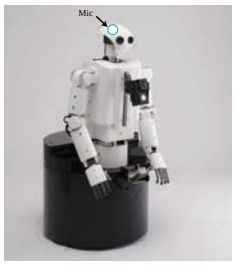
## 4 評価実験

### 4.1 音響データ収録

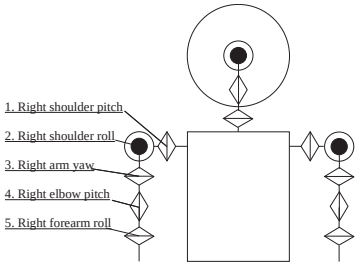
提案手法の評価を行うために 2 つのヒューマノイド・ロボットを用いて音響データの収録を行った．

#### 4.1.1 ヒューマノイドロボット

実験には Hearbo と Robovie-W の 2 つのヒューマノイド・ロボットを使用した．Hearbo は頭部に 8 個のマイクロホンが取り付けてあるが，収録にはそのうちの額にあるマイクロホン 1 つを用いた．また，Hearbo は計 34 自由度を持っているが，本研究の実験ではそのうち右腕の 5 つの関節: 右肩ピッチ角 (J1)，右肩ロール角 (J2)，右腕ヨー角 (J3)，右肘ピッチ角 (J4)，右前腕ロール角 (J5) を駆動させた．Robovie-W は市販されている小型ロボットであり, 頭部には音声収録用に 8 個のマイクロホンをつけた帽子が取り付けてあるが，実験ではそのうち 1 つを使用した．Robovie-W は計 17 自由度を持っているが, 本実験ではそのうち 5 つの関節: 腰ヨー角, 肩ロール角 (左右), 肘ピッチ角 (左右) を駆動させた．Robovie-W は各関節に指令値を送ることで制御できるが, 機体から関節角の状態

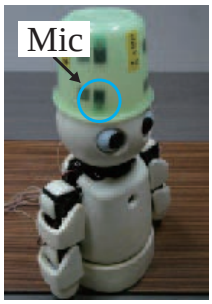



a) 写真

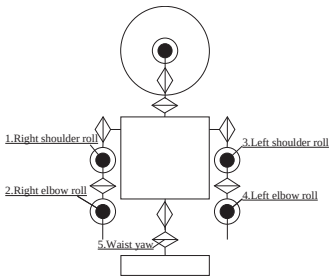



b) ジョイントとリンクの構成

Figure 4: ヒューマノイド・ロボット Hearbo




a) 写真




b) ジョイントとリンクの構成

Figure 5: ヒューマノイド・ロボット Robovie-W

を読み取ることは不可能である．従って, このロボットでは動作情報から動作雑音を推定する動作雑音抑圧手法を適用することが不可能である．

#### 4.1.2　収録条件

Hearbo を用いた実験では以下の条件 D1-D6 の動作雑音及び音声データを収録した．音声及び動作雑音の収録は 4m×7m×3m の部屋で行った．Hearbo は部屋の中心に, 発話者 (スピーカ) は Hearbo の正面から 1.2m 離れた位置に配置した．

D1 J1 から J5 までの 5 個の関節をそれぞれ 20 秒ずつ駆動させ, 動作雑音を発生させた．

D2 J1 を 40 秒間駆動させ, 後半の 20 秒間にスピーカーから音声を流した．

D3 2 つの関節 (J1 と J2) を 40 秒間駆動させ, 後半の 20 秒間にスピーカーで音声を流した．

D4 3 つの関節 (J1,J2,J3) を 40 秒間駆動させ, 後半の 20 秒間にスピーカーで音声を流した．

D5 4 つの関節 (J1 – J4) を 40 秒間駆動させ, 後半の 20 秒間にスピーカーで音声を流した．

D6 5 つの関節 (J1 – J5) を 40 秒間駆動させ, 後半の 20 秒間にスピーカーで音声を流した．

D7 Hearbo を動かさずに 20 秒間スピーカーから音声を流した．

また,Robovie-W を用いた実験では以下の条件 R1,R2 の動作雑音及び発話を収録した．音声及び動作雑音の収録は 10m×10m×3m の部屋で行った．Robovie-W は円形のテーブル載せ，発話者 (人間) は Robovie-W の正面から 2.0m 離れた位置に配置した．

R1 5 つの関節を 5 秒間ランダムに動かし, 動作雑音を発生させた

R2 R1 の動作を行った状態で 2 秒間発話を行う

### 4.2　評価方法と結果

提案手法を以下の 4 つの実験により評価を行った．

1. 提案手法のパラメータ評価実験
2. 動作雑音推定・抑圧実験

まず実験 1 では SB-INMF に使用されているパラメータの変化が動作雑音抑圧結果に対しどのような影響を与えるか評価し，得られた結果から実験 2 で使用するパラメータを決定する．実験 2 では次の 3 段階に分けて実験を行う．①INMF による動作雑音の再構成実験，②動作雑音と音声の合成音を対象とした動作雑音抑圧実験，③実収録した動作雑音を含む音声を対象とした動作雑音抑圧実験．なお, 実験 2 ～ 4 では提案手法との比較に動作雑音性能が高いことが報告されている Ince らが提案したテンプレートベースの雑音抑圧手法 [19] を用いた．

#### 4.2.1　提案手法のパラメータ評価実験

**実験 1**：式に示した基底のパラメータ $a$ とアクティベーションのパラメータ $b$ を $10^{-4}$～$10^{0}$ の間で変化させながら動作雑音抑圧を行う．Hearbo と Robovie-W について実験を行い，学習データには各々D1 と R1，テストデータには D1 と D7 を足し合わせ音響信号と R1 にクリーンな音声を足し合わせた音響信号を使用した．そして得られた結果を *Signal-to-Inference Ratio(SIR)*, *Signal-to-Distortion Ratio(SDR)*, そして *Signal-to-Noise Ratio(SNR)* により評価を行った．

SNR(SNR$_1$) を以下のように定義する．

$$\mathrm{SNR}_1 = 20\log_{10}\left(\frac{\sum_f\sum_\omega |X(\omega,f)|^2}{\sum_f\sum_\omega ||Y(\omega,f)| - |X(\omega,f)| - |\hat{N}(\omega,f)||^2}\right) \tag{12}$$

ここで $\omega$ と $f$ はそれぞれ周波数ビンと時刻フレームを表し，$X$,$Y$ は雑音を含まない音声及び動作雑音と音声の合成音のスペクトルを表す．また,SIR は目的音以外の妨害音 (動作雑音，環境雑音) による歪みを評価する指標，SDR は目的音の線形歪み，非線形歪み，そして上記の妨害音による歪みを総合的に評価する指標となっている．推定した音声信号 $\hat{x}(t)$ が式 (13) の様に分解できると仮定する．

$$\hat{x}(t) = x_{true}(t) + e_{noise}(t) + e_{artif}(t) \tag{13}$$

ここで $x_{true}(t)$ を真の音声成分，$e_{noise}(t)$ を動作雑音の成分，$e_{artif}(t)$ をいずれにも寄与しない成分とする．この時 SIR，SDR は式 (14),(15) で与えられる．

$$\mathrm{SIR} = 10\log_{10}(\frac{||x_{true}(t)||^2}{||e_{noise}(t)||^2}), \tag{14}$$

$$\mathrm{SDR} = 10\log_{10}\left(\frac{||x_{true}(t)||^2}{||e_{noise}(t)+e_{artif}(t)||^2}\right) \quad (15)$$

なお，SIR と SDR の計算は MATLAB のツールボックス”BSS Eval[1] ”を使用した．

**結果**： 得られた結果を図 6,7 に示す. 縦軸, 横軸ともにログスケールとなっている．Robovie の結果を見ると，SNR, SIR，SDR 全て共通して基底のパラメータによる影響の大きいことがわかる．しかし，各指標のカラーマップを比較すると SIR と SDR の分布は近い傾向にあるのに対し，SNR は全く異なる分布であることがわかる．また，Hearbo の結果を見ると SNR と SIR には Robovie-W と同じような傾向が見られるのに対し，SDR は異なりほぼ均一な値となってしまっていることがわかる．これは D7 には音声の他に環境雑音含まれており，推定した音声では環境雑音が抑圧されている (後の実験 3 参照) ため，SIR の値が高くても SDR の値が向上しなかったと考えられる．これらの結果を踏まえ，実験 2 〜 4 で用いるパラメータは SNR, SIR, SDR の各々の値がある程度補償されるものを使用した．具体的には Robovie の場合は $a = 10^{-2}$,$b = 10^{-2}$, Hearbo の場合は $a = 10^{-1.5}$,$b = 10^{-1.5}$ とした．

#### 4.2.2 動作雑音推定・抑圧実験

Hearbo を用いた実験では D1 を用いて INMF による動作雑音の基底の学習を行った．また, 比較手法の雑音テンプレートのデータベース作成に実験 2,3 では D1 を，実験 3 では D2-D6 の最初の 10 秒間を使用した．
Robovie-W を用いた実験では R1 を用いて INMF による動作雑音の基底の学習を行った．

**実験 2.1**： Hearbo では D1,Robovie-W では R1 をテストデータとして与え, 動作雑音の推定が可能であるか評価した．指標には *Noise Estimation Error(NEE)* を用いて評価を行った．

$$\mathrm{NEE} = 20\log_{10}\left(\frac{\sum_f\sum_\omega |N(\omega,f)|^2}{\sum_f\sum_\omega (|N(\omega,f)| - |\hat{N}(\omega,f)|)^2}\right) \quad (16)$$

$N$ 及び $\hat{N}$ はそれぞれオリジナルと推定した動作雑音スペクトルを表す.

**実験 2.2**: Hearbo では D1 に D7 を, Robovie-W では R1 に雑音を含まない音声を加算した音響信号をテストデータとして与え動作雑音抑圧を行った．評価指標には実験 1 と同様に $\mathrm{SNR}_1$,SIR,SDR を用いた．

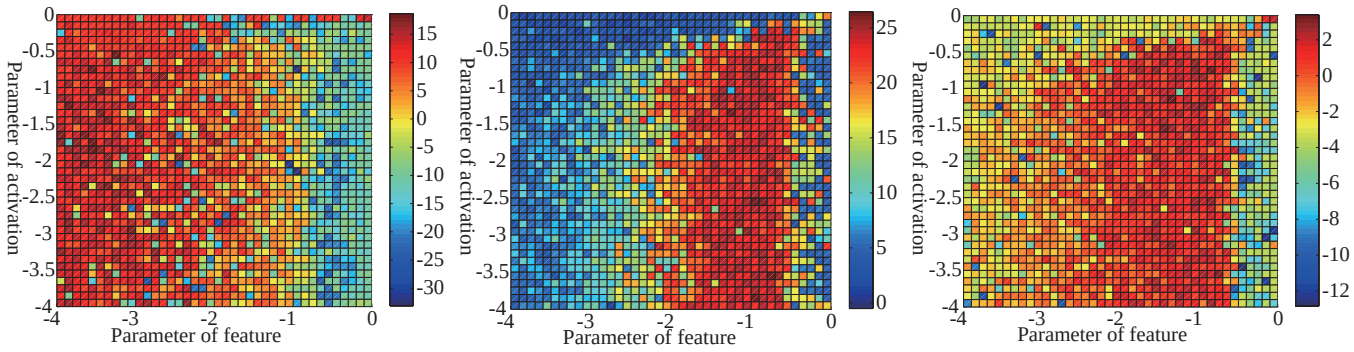
**実験 2.3**： Hearbo の場合は D2-D6 を,Robovie-W の場合は R2 をテストデータに用いた．評価は SNR, 実験 2.3 では正解となる音声及び動作雑音は分からない．そこで SNR を改良した $\mathrm{SNR}_2$ を以下に定義し，評価にはこれを用いた．

[1] http://bass-db.gforge.inria.fr/bss eval/

$$\begin{aligned}\mathrm{SNR}_2 &= 20\log_{10}\left(\frac{\sum_f\sum_\omega \left|\hat{S}(\omega,f)\right|^2}{\sum_f\sum_\omega \left||Y_N(\omega,f)| - |\hat{N}(\omega,f)|\right|^2}\right) \\ &\quad -20\log_{10}\left(\frac{\sum_f\sum_\omega |Y_S(\omega,f)|^2}{\sum_f\sum_\omega |Y_N(\omega,f)|^2}\right) \end{aligned} \quad (17)$$

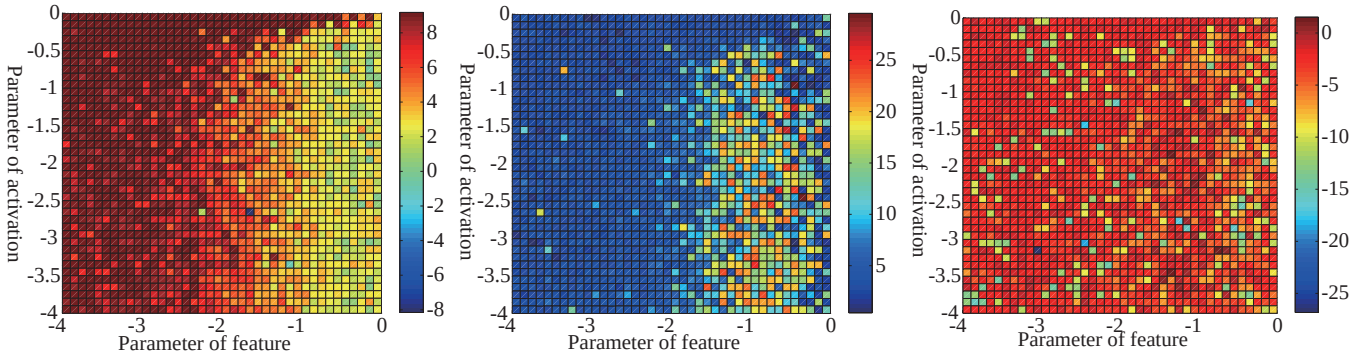
$Y_S$ 及び $Y_N$ は各々入力信号の雑音部分と音声部分を示す．また,$\hat{S}$ は提案手法により推定した音声を示す．なお, テンプレートを用いた手法では $\hat{S}$ の推定ができないため, $|Y_S(\omega,f)| - |\hat{N}(\omega,f)|$ を代わりに使用した．

**結果**： 実験 2.1,2.2 の結果を表 2 に示す．表 2 の提案手法における基底の数はテンプレートを用いた雑音抑圧手法におけるテンプレートの数に対応する．提案手法とテンプレートを用いる手法で推定結果を比較すると，提案手法では推定に用いる基底の数は Hearbo の場合は 7 つ,Robovie-W の場合は 5 つのみであるが，テンプレート数が 300 個を超える場合よりも NEE の値が良いことがわかる．
Hearbo を用いた実験 2.2 において提案手法では音声に対し基底数を 2 つ，雑音の基底と合わせて 9 つの基底が推定された．その抑圧性能を $\mathrm{SNR}_1$ で評価すると，テンプレートの数が 1,022 の場合とほぼ同等の結果であった．また，SIR，SDR による評価では，テンプレート数に関わらずテンプレートを用いた手法よりも提案手法が良い結果が得られた．基底やテンプレートの数が多くなればなるほど，計算コストは大きくなる．そのため，少ない基底数で雑音の推定と抑圧が可能であるということは大きな利点である．図 8 に Hearbo を用いた実験 2.2 で推定した各信号及びその元信号のスペクトログラムを示した．図 8a) は,8b) に示した動作雑音と 8c) の音声を足し合わせた合成音で, 右が各信号を提案手法を用いて再構成した信号である．スペクトログラムからも動作雑音, 音声が推定できている事がわかる．また，実験 1 でも記述したが再構成した音声には環境雑音が含まれていない．これは動作雑音の基底の学習の際，動作雑音に含まれた環境雑音によって環境雑音の基底が学習され，既知の基底に含まれていたためと考えられる．Robovie-W は関節角情報を得られないためテンプレートを用いた手法との比較は不可能であるが，表 2 及び図 9 から Hearbo の場合と近い結果が得られていることがわかる．これらの結果から，動作雑音及び音声は提案手法によって推定可能であることがわかる．
実験 2.3 の結果を表 3 に示す．表 3 の”#of templates”はテンプレートを用いた動作雑音抑圧において，$\mathrm{SNR}_2$ が最大となった時に使用したテンプレートの数を示している．表 3 より,D2-D6 全ての場合において提案手法のほうがテンプレートを用いた手法より $\mathrm{SNR}_2$ が良い結果が得られた．学習データとは異なるテストデータを用いたオープン




a)SNR b)SIR c)SDR

Figure 6: 実験 1 （Robovie-W）




a)SNR b)SIR c)SDR

Figure 7: 実験 1 (Hearbo)

テストでも効果が確認できたことから，今回の提案手法がロバストであることがわかった．Robovie-W を用いた実験の結果は $SNR_2$ は 3.9dB であった．この値は Hearbo の場合よりも小さい値となったが，Robovie-W と話者の距離が Hearbo の場合よりも離れていたことが原因と考えられる．また，テンプレートベースの手法を用いた場合，実験 2.2 でのクローズテストでは非常に多くのテンプレート数を用いた時良い結果が得られた．しかし，実験 2.3 のオープンテストではテンプレートの数が少ないほうが良い結果が得られた．これは，同じ動作であっても常に同じ動作雑音が発生するとは限らない，すなわち完全な相関関係では無いためである．

## 5 おわりに

本論文では単一のマイクで関節角や他のセンサーの情報を用いない新たな動作雑音抑圧手法を提案し，その評価を行った．具体的には，*Infinite Non-negative Matrix Factrization* によって予め動作雑音の基底を学習し，*Semi-Blind INMF* によって動作雑音とた目的音を分離する手法を提案した．そして実際に 2 種類のロボットを用いた動作雑音抑圧実験を行い，提案手法の有効性を示した．今後は提案手法によって動作雑音が抑圧された音声による音声認識を行う予定である．

## 謝辞

本研究の一部は科研費（24118702，22700165）の補助を受けた．

## 参考文献

[1] K. Nakadai *et al.* Active audition for humanoid. *AAAI 2000*, pp. 832–839.

[2] K. Nakadai *et al.* Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots. *Speech Communication*, Vol. 44, pp. 97–112, 2004.

[3] Y. Nishimura *et al.* Speech recognition for a humanoid with motor noise utilizing missing feature theory. *Humanoids 2006*,IEEE pp. 26–33. IEEE.

[4] T. Rodemann *et al.* Sound localization for humanoid robots — building audio-motor maps based on the hrtf. *IROS 2006*,IEEE/RSJ pp. 1171—1176..

[5] J. Hornstein *et al.* Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping. *IROS 2006*,IEEE/RSJ pp. 860—865.

[6] J. Hornstein *et al.* Spectral cues for robust sound localization with pinnae. *IROS 2006*,IEEE/RSJ pp. 386-391.

[7] A.Portello *et al.* Active binaural localization of intermittent moving sources in the presence of false measurements. *IROS 2012*,IEEE/RSJ pp. 3294—3299.

[8] J.-M. Valin *et al.* Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach. In IEEE, editor, *ICRA 2004*,IEEE-RAS, pp. 1033—1038.

[9] Y.Sasaki *et al* . Spherical microphone array for spatial sound localization for a mobile robot. *IROS 2012*,IEEE/RSJ pp. 713-718.

Table 2: 実験２，３の結果

| Robot | Hearbo | | | | | | | Robovie W |
|---|---|---|---|---|---|---|---|---|
| | Proposed | Template-based | | | | | | Proposed |
| # of feat. /templ. | 7 (ego-noise) 2 (speech) | 31 | 98 | 303 | 1,022 | 3,115 | 8,431 | 5(noise) 3(speech) |
| NEE (dB) | 9.4 | 8.0 | 8.2 | 8.0 | 9.9 | 12.3 | 24.6 | 9.3 |
| $SNR_1$ (dB) | 7.2 | 6.1 | 6.3 | 5.9 | 7.4 | 8.7 | 14.6 | 7.6 |
| SIR (dB) | 11.0 | 1.4 | 2.6 | 2.6 | 3.0 | 2.3 | 2.2 | 17.0 |
| SDR (dB) | 1.3 | -0.8 | 1.1 | 1.1 | 1.3 | 0.8 | 0.8 | 2.0 |

Table 3: 実験４の結果

| Robot | Dataset | Proposed | Template-based | |
|---|---|---|---|---|
| | | $SNR_2(dB)$ | $SNR_2(dB)$ | # of templates |
| Hearbo | D2(J1) | 5.9 | 2.5 | 21 |
| | D3(J1+J2) | 5.2 | 3.1 | 8 |
| | D4(J1–J3) | 6.3 | 3.2 | 12 |
| | D5(J1–J4) | 3.5 | -0.76 | 244 |
| | D6(J1–J5) | 5.3 | 2.4 | 45 |
| Robovie-W | R2 | 3.9 | N/A | N/A |

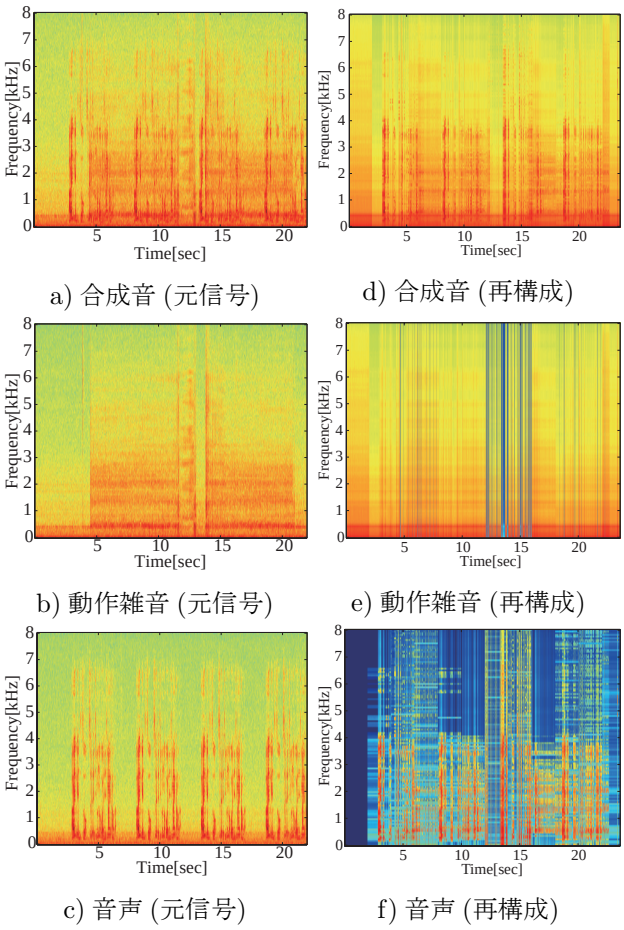



Figure 8: Hearbo での動作雑音抑圧

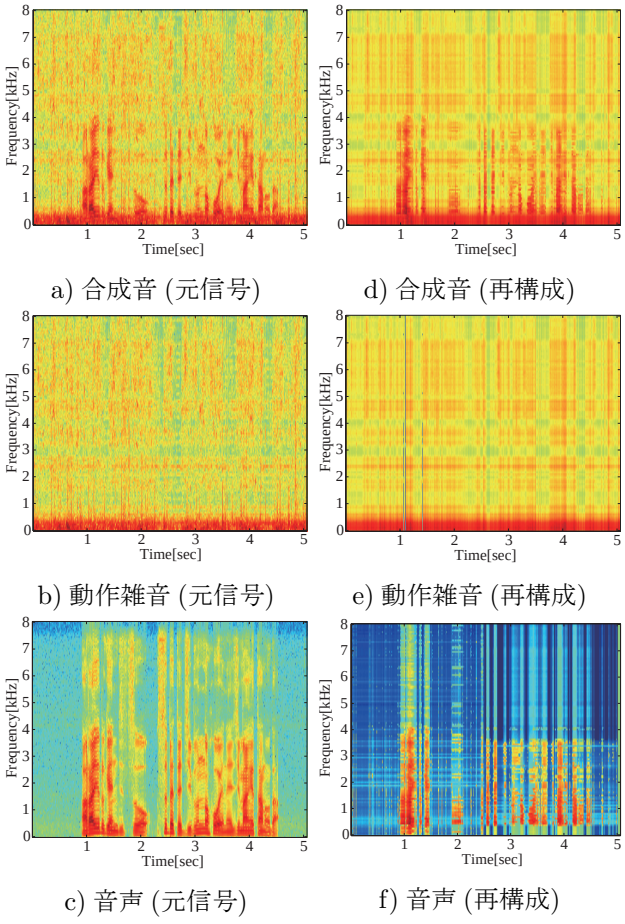



Figure 9: Robovie-W での動作雑音抑圧


[10] S.Yamamoto *et al.* Design and implementation of a robot audition system for automatic speech recognition of simultaneous speech. *ASRU-2007*, pp. 111–116. IEEE, Dec. 2007.

[11] H.Saruwatari *et al.* Two-stage blind source separation based on ica and binary masking for real-time robot audition system. *IROS 2005*, pp. 209–214. IEEE, 2005.

[12] T.Yoshida and K.Nakadai . Active audio-visual integration for voice activity detection based on a causal bayesian network. *Humanoids 2012*,IEEE pp. 370-375.

[13] K.Nakadai *et al.* Sound source tracking with directivity pattern estimation using a 64ch microphone array. *IROS 2005*,IEEE/RSJ pp. 196-202.

[14] F. Perrodin *et al.* Design and calibration of large microphone arrays for robotic applications. *IROS 2012*,IEEE/RSJ pp. 4596-4601.

[15] J.Even *et al.* "Combining laser range finders and local steered response power for audio monitoring. *IROS 2012*,IEEE/RSJ pp. 986-991.

[16] J.Even *et al.* "Semi-blind suppression of internal noise for hands-free robot spoken dialog system. *IROS 2009*,IEEE/RSJ pp. 658-663.

[17] A. Ito *et al.* Internal noise suppression for speech recognition by small robots. *Eurospeech 2005*, pp. 2685–2688.

[18] S. F. Boll. A spectral subtraction algorithm for suppression of acoustic noise in speech. *ICASSP 1979)*, IEEE, pp. 200–203.

[19] Gokhan Ince *et al.* Incremental learning for ego noise estimation of a robot. *IROS 2011*, IEEE/RSJ, pp. 131–136.

[20] S. Abdallah and M. D. Plumbley. Polyphonic music transcription by non-negative sparse coding of power spectra. *ISMIR 2004*, pp. 10–14, 2004.

[21] M. N. Schmidt, M.Mørup. Infinite non-negative matrix factorization. *EUSIPCO*, 2010.

[22] M. D. Hoffman *et al.* Bayesian nonparametric matrix factorization for recorded music. *ICML 2010*, pp. 439–446.