空気／体内伝導信号の非負値テンソル分解に基づく体内伝導微弱音声に対する雑音抑圧法*

☆田尻祐介 (奈良先端大), 亀岡弘和 (NTT), 戸田智基 (名大／奈良先端大), 中村哲 (奈良先端大)

1 はじめに

周囲に発話内容を聴取されることなく，音声通話を可能にする技術として，非常に微弱なささやき声を非可聴つぶやき（Non-Audible Murmur: NAM）マイクと呼ばれる体表密着型マイクにより収録する枠組みが提案されている [1]．音声の体内伝導収録は，マイクの構造上，外部雑音に対して比較的頑健であるものの，微弱な音声を収録する場合，その影響を避けることは困難となる．そこで，発声音の微弱性を利用して外部雑音を通常の空気伝導マイクでモニタリングし，NAM マイクに混入した外部雑音を抑圧する手法が提案されている [2]．ただし，従来法の有効性は，雑音源が一つの点音源である条件下でのみ確認されており，実環境下で使用する場合，雑音抑圧性能が著しく低下すると予想される．

本稿では，雑音源の増加や移動に頑健な雑音抑圧法として，空気／体内伝導収録信号のパワースペクトログラムに対する非負値テンソル分解（Non-negative Tensor Factorization: NTF）の適用を提案する．また，実験的評価結果より，実環境で収録した外部雑音に対する雑音抑圧性能が，従来法によるものと比較して大幅に改善されることを示す．

2 空気伝導マイクを用いた外部雑音モニタリングに基づく雑音抑圧

雑音環境下では，マイクを口唇付近に設置しない限り，微弱なささやき声を空気伝導音として収録することは困難である．このような発声音の微弱性に着目し，Fig. 1 に示すような位置に空気伝導マイクを配置することで，口唇から放射される音声の集音を極力抑え，かつ，NAM マイクへの混入音と近い音響特性を持つ外部雑音を収録する．空気伝導マイクで収録した信号を，混入する外部雑音の参照信号として利用することで，体内伝導信号に対する雑音抑圧を行う．

雑音源が一つの点音源の場合，各種適応アルゴリズム [3] やセミブラインド信号分離（Semi-Blind Source Separation: Semi-BSS）[4] に基づき，各マイクで収録される信号間の特性を補正する線形フィルタを推定することで，高い雑音抑圧効果が得られる．一方で，二つ以上の雑音源がマイクに対して異なる方向に存在する場合，混合雑音と体内伝導雑音間の特性を補正する処理となるため，時々刻々と急激に変化する線形フィルタの推定が必要となり，十分な雑音抑圧効果が得られない．

3 非負値テンソル分解の適用

Fig. 1 のマイク配置では，ユーザと雑音源の距離が近接している場合を除き，音源から各マイク間で生じる信号の減衰量の差は，多少の音源位置の変化に対して比較的安定している．一方で，到達時間の差は無視できないことが多く，空気伝導信号に対するフィルタリング処理によって，NAM マイクに混入する外部雑音を精度よく推定することは困難である．そこで，信号の位相情報はあえて信頼せずに，振幅情報のみを重視した観測モデルを検討する．

NAM マイクおよび空気伝導マイクによる観測複素

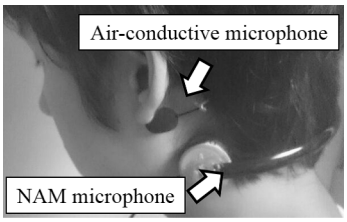

Fig. 1 Setting position of air- and body-conductive (NAM) microphones

スペクトログラムを $\boldsymbol{y}_{\omega,\tau} = [y_{1,\omega,\tau}, y_{2,\omega,\tau}]^\top$，各音源の複素スペクトログラムを $s_{i,\omega,\tau}$ とすると，時間周波数領域における瞬時混合モデルは以下のように表される．

$$\boldsymbol{y}_{\omega,\tau} = \sum_i \boldsymbol{a}_{i,\omega} s_{i,\omega,\tau} \tag{1}$$

ここで，i は音源インデックス，ω は周波数インデックス，τ は時刻フレームインデックスを表す．$\boldsymbol{a}_{i,\omega}$ は音源から各マイクまでの伝達特性を表すベクトルで

$$\boldsymbol{a}_{i,\omega} = \begin{bmatrix} |a_{1,i,\omega}| & 0 \\ 0 & |a_{2,i,\omega}| \end{bmatrix} \begin{bmatrix} e^{j\phi_{1,i,\omega}} \\ e^{j\phi_{2,i,\omega}} \end{bmatrix} \tag{2}$$

のように絶対値と偏角に分解できる．ここで，新たに

$$\boldsymbol{A}_{i,\omega} = \begin{bmatrix} |a_{1,i,\omega}| & 0 \\ 0 & |a_{2,i,\omega}| \end{bmatrix} \tag{3}$$

$$\tilde{\boldsymbol{s}}_{i,\omega,\tau} = \begin{bmatrix} e^{j\phi_{1,i,\omega}} \\ e^{j\phi_{2,i,\omega}} \end{bmatrix} s_{i,\omega,\tau} \tag{4}$$

と定義する．ここで，信頼度の低い偏角については確率変数とみなし，1) $\phi_{c,i,\omega}$ は区間 $[0, 2\pi)$ で一様分布に従う，2) $\phi_{c,i,\omega}, \phi_{c',i,\omega}\,(c \neq c')$ は互いに独立である，3) $s_{i,\omega,\tau}$ は複素正規分布 $\mathcal{N}_{\mathbb{C}}(0, p_{i,\omega,\tau})$ に従う，と仮定して，偏角を周辺化すると，次式が導出される．

$$\tilde{\boldsymbol{s}}_{i,\omega,\tau} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, p_{i,\omega,\tau}\boldsymbol{I}) \tag{5}$$

結果，式 (1) の観測複素スペクトログラムも同様に，次の複素正規分布で表せる．

$$\boldsymbol{y}_{\omega,\tau} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \sum_i p_{i,\omega,\tau} \boldsymbol{A}_{i,\omega} \boldsymbol{A}_{i,\omega}^{\mathrm{H}}) \tag{6}$$

ここで，$p_{i,\omega,\tau}$ は，各音源のパワースペクトログラムの期待値を表すパラメータであり，非負値行列因子分解（Non-negative Matrix Factorization: NMF）[5] で用いられる構造 $p_{i,\omega,\tau} = \sum_k w_{i,\omega,k} h_{i,k,\tau}$ で近似できると仮定すると，パラメータ $\boldsymbol{A}' = (|a_{c,i,\omega}|^2)_{C\times I\times\Omega}$，$\boldsymbol{W} = (w_{i,\omega,k})_{I\times\Omega\times K}$，および $\boldsymbol{H} = (h_{i,k,\tau})_{I\times K\times T}$ の最尤推定問題は，観測パワースペクトログラム $|y_{c,\omega,\tau}|^2$ を要素に持つテンソル $\boldsymbol{Y}'$ に対し，板倉齋藤擬距離基準の NTF を行うことと等価となる．したがって，目的関数はパラメータ $\Theta = \{\boldsymbol{A}', \boldsymbol{W}, \boldsymbol{H}\}$ に依らない項を省略すると以下で表される．

$$D_{\mathrm{IS}}(\Theta) \overset{\boldsymbol{A}',\boldsymbol{W},\boldsymbol{H}}{=} \sum_{c,\omega,\tau} \left(\frac{|y_{c,\omega,\tau}|^2}{x_{c,\omega,\tau}} + \log x_{c,\omega,\tau} \right) \tag{7}$$

ただし，$x_{c,\omega,\tau} = \sum_i |a_{c,i,\omega}|^2 \sum_k w_{i,\omega,k} h_{i,k,\tau}$ とする．

---

* A noise suppression method for body-conducted soft speech based on non-negative tensor factorization of air- and body- conducted signals. by TAJIRI, Yusuke (NAIST), KAMEOKA, Hirokazu (NTT), TODA, Tomoki (Nagoya Univ./NAIST), NAKAMURA, Satoshi (NAIST)

補助関数法を用い，式 (7) の上限関数を設計することで，最終的に次の更新アルゴリズムが導かれる．

$$|a_{c,i,\omega}|^2 \leftarrow |a_{c,i,\omega}|^2 \left(\frac{\sum_{k,\tau} |y_{c,\omega,\tau}|^2 w_{i,\omega,k} h_{i,k,\tau} / x_{c,\omega,\tau}^2}{\sum_{k,\tau} w_{i,\omega,k} h_{i,k,\tau} / x_{c,\omega,\tau}} \right)^{\frac{1}{2}} \quad (8)$$

$$w_{i,\omega,k} \leftarrow w_{i,\omega,k} \left(\frac{\sum_{c,\tau} |y_{c,\omega,\tau}|^2 |a_{c,i,\omega}|^2 h_{i,k,\tau} / x_{c,\omega,\tau}^2}{\sum_{c,\tau} |a_{c,i,\omega}|^2 h_{i,k,\tau} / x_{c,\omega,\tau}} \right)^{\frac{1}{2}} \quad (9)$$

$$h_{i,k,\tau} \leftarrow h_{i,k,\tau} \left(\frac{\sum_{c,\tau} |y_{c,\omega,\tau}|^2 |a_{c,i,\omega}|^2 w_{i,\omega,k} / x_{c,\omega,\tau}^2}{\sum_{c,\tau} |a_{c,i,\omega}|^2 w_{i,\omega,k} / x_{c,\omega,\tau}} \right)^{\frac{1}{2}} \quad (10)$$

以上の更新を収束するまで繰り返し，求められたパワースペクトログラムから構成したウィナーフィルタを適用することで雑音を抑圧する．

また，本枠組みにおいて，目的信号である微弱音声は NAM マイクでのみ観測されるとみなせるため，$\boldsymbol{a}_{1,\omega} = [1, 0]^\top$ と固定する．さらに，マイク間距離が比較的短いことに着目し，空気伝導信号に含まれる複数の雑音を一つの雑音として考える．よって，雑音源の数に関わらず音源インデックスは $I = 2$ とし，$\boldsymbol{a}_{2,\omega} = [a_{1,2,\omega}, 1]^\top$ と固定する．結果，パラメータ $\boldsymbol{A}'$ に関しては，$|a_{1,2,\omega}|^2$ のみが更新される．

4 実験的評価

4.1 実験条件

男性話者 1 名の非常に微弱なささやき声を，NAM マイクおよび空気伝導マイクで同時収録する．収録文は ATR 音素バランス文 A セット中の 50 文とする．また，次の 7 種類の雑音（3 種類は防音室にて収録雑音を一つのスピーカから再生することで収録，4 種類は実環境で収録）をささやき声とは別に収録し，重畳することで収録信号を生成する．ただし，実環境雑音については，アンプのゲイン設定がささやき声の収録時と一致していないため，体内伝導信号における SNR が 5 dB もしくは 0 dB となるように重畳する．

- **crowd60dB**：60 dB の人混みの雑音
- **booth70dB**：70 dB の展示場の雑音
- **station80dB**：80 dB の駅構内の雑音
- **crowd5dB_SNR**：人混みの雑音（実環境）
- **traffic5dB_SNR**：高架下の雑音（実環境）
- **restaurant0dB_SNR**：飲食店の雑音（実環境）
- **station0dB_SNR**：駅構内の雑音（実環境）

収録信号のサンプリング周波数は 16 kHz，FFT 分析のフレーム長は 64 ms，シフト長は 32 ms とする．比較対象には，従来法である Semi-BSS，(体内伝導) 観測パワースペクトログラムに対する板倉齋藤擬距離基準の NMF を用いる．NTF および NMF における各音源の基底数は 20，パラメータの更新回数は 50 回とする．また，両者ともに微弱なささやき声の基底は事前に学習したものを使用し，一個抜き交差検証を行う．Semi-BSS のステップサイズは 0.01，更新回数は 200 回とする．

4.2 防音室で収録した雑音に対する結果

Fig. 2 に雑音抑圧前後の SNR を示す．提案法である NTF は，**crowd60dB** に対しては，従来法である Semi-BSS よりも高い SNR を得られることがわかる．ただし，雑音レベルの増加に伴い，結果は逆転する傾向にあり，**station80dB** に対しては，NMF より高いものの，Semi-BSS より低い SNR となっている．したがって，雑音源が一つの固定された点音源であるという限られた条件下では，提案法の NTF よりも Semi-BSS のほうが有効であるといえる．

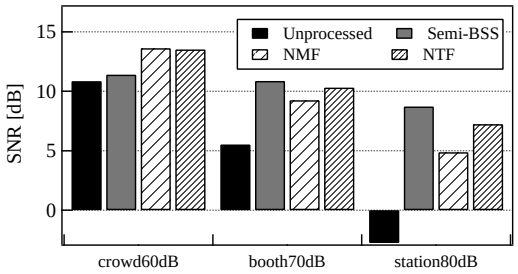

Fig. 2 Average SNR of estimated body-conducted signals (in a soundproof room)

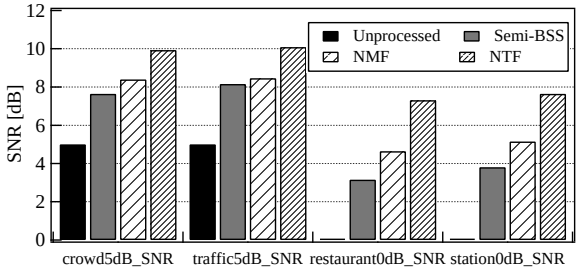

Fig. 3 Average SNR of estimated body-conducted signals (in real environments)

4.3 実環境で収録した雑音に対する結果

Fig. 3 に雑音抑圧前後の SNR を示す．提案法である NTF は，4 種類の雑音全てに対して，Semi-BSS よりも高い SNR が得られることがわかる．また，体内伝導信号に対する NMF の結果を大幅に上回っており，空気伝導マイクにより得られた外部雑音情報を有効的に活用できていることがわかる．

5 おわりに

複数音源の存在およびその移動に頑健な体内伝導微弱音声の強調法として，空気／体内伝導信号の非負値テンソル分解（NTF）に基づくに基づく雑音抑圧法を提案した．実験的評価結果より，実環境で収録した雑音に対する雑音抑圧性能が，従来法である Semi-BSS やシングルチャネルの NMF と比較して大幅に改善されることを示した．ただし，提案法は目的信号の基底を事前に学習する半教師ありの雑音抑圧法であり，ロンバード効果 [6] として知られるような雑音聴取に伴う発話様式変化が含まれる場合，性能が低下すると予想される．また，バッチ処理のため円滑なサイレント音声通話を実現するのは困難である．これらの対策として，今後は NTF のオンライン化に取り組む．

謝辞 本研究の一部は，JSPS 科研費 15K12064 および 26280060 の助成を受け実施したものである．

参考文献

[1] 中島 他, 信学論, Vol. 87, No. 9, pp. 1757-1764, 2004.
[2] 田尻 他, 信学技報, Vol. 115, No. 302, pp. 41-46, 2015.
[3] Haykin, *Adaptive filter theory*, Prentice Hall, fourth edition, 2002.
[4] Miyabe *et al.*, *Proc. ICASSP*, pp. 109-112, 2006.
[5] Lee and Seung, Nature, Vol. 401, pp. 788-791, 1999.
[6] Toda *et al.*, *Proc. INTERSPEECH*, pp. 632-635, 2009.