非可聴つぶやき強調のためのセグメント特徴量正則化 NTF *

◎田尻祐介 (名大), 亀岡弘和 (NTT), 戸田智基 (名大)

1 はじめに

サイレント音声通話の実現に向けて，微弱なささやき声を非可聴つぶやき（Non-Audible Murmur: NAM）マイク [1] と呼ばれる体表密着型マイクで収録し，統計的手法 [2] により空気伝導音声へと変換する枠組みが提案されている．音声の体内伝導収録は，外部雑音に対して比較的頑健であるものの，微弱音声の収録では十分な SN 比を確保できず，変換音声が著しく劣化する．そこで，発声音と外部雑音の音量差を利用した空気伝導マイクによる雑音モニタリング法および雑音抑圧法が提案されている [3, 4]．従来法では，非負値テンソル分解（Non-negative Tensor Factorization: NTF）の適用により，実環境雑音下における SN 比の改善を実現している．ただし，後段における変換性能の向上に直結するとは限らず，雑音処理によって生じる特徴量空間でのひずみが性能低下を引き起こす可能性がある．

本研究では，NTF に基づく雑音抑圧において，変換処理時の入力であるセグメント特徴量の事前情報を活用した正則化を行い，その有効性を検証する．実験的評価結果より，実環境雑音下におけるスペクトル特徴量の変換精度が改善されることを示す．

2 空気／体内伝導信号の非負値テンソル分解に基づく雑音抑圧法 [4]

NAM マイクおよび空気伝導マイクによる観測複素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 $s_{i,\omega,\tau}$ とし，以下のような時間周波数領域における瞬時混合モデルを仮定する．

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

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．$\boldsymbol{a}_{i,\omega,\tau}$ は音源から各マイクまでの伝達特性を表すベクトルで，

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

のように絶対値成分 $\boldsymbol{A}_{i,\omega}$ と偏角成分 $\boldsymbol{\psi}_{i,\omega,\tau}$ に分解できる．なお，絶対値成分 $\boldsymbol{A}_{i,\omega}$ は音源移動などに対して比較的安定していると仮定し，時不変な変数として扱う．このとき，$s_{i,\omega,\tau}$ が複素正規分布 $\mathcal{N}_{\mathbb{C}}(0, p_{i,\omega,\tau})$ に従うと仮定すると，式 (1) は以下で表される．

$$\boldsymbol{y}_{\omega,\tau} \sim \mathcal{N}_{\mathbb{C}}\Big(\boldsymbol{0}, \sum_i p_{i,\omega,\tau} \boldsymbol{A}_{i,\omega} \boldsymbol{\psi}_{i,\omega,\tau} \boldsymbol{\psi}_{i,\omega,\tau}^{\mathsf{H}} \boldsymbol{A}_{i,\omega}^{\mathsf{H}}\Big) \tag{3}$$

さらに，偏角を確率変数をみなし，以下の仮定をおくことで周辺化する．

- $\phi_{c,i,\omega,\tau}$ は区間 $[0, 2\pi)$ で一様分布に従う
- $\phi_{c,i,\omega,\tau}, \phi_{c',i,\omega,\tau'}$ $(c \neq c' \text{ or } \tau \neq \tau')$ は互いに独立

結果，チャネル毎に以下の観測モデルが得られる．

$$y_{c,\omega,\tau} \sim \mathcal{N}_{\mathbb{C}}\Big(0, \sum_i p_{i,\omega,\tau} |a_{c,i,\omega}|^2\Big) \tag{4}$$

ここで，各音源のパワースペクトログラム $p_{i,\omega,\tau}$ に対して，非負値行列因子分解（Non-negative Matrix Factorization: NMF）[5] の構造 $p_{i,\omega,\tau} = \sum_k h_{i,\omega,k} u_{i,k,\tau}$ を仮定すると，パラメータ $\boldsymbol{A}' = (|a_{c,i,\omega}|^2)_{2\times I\times\Omega}$，$\boldsymbol{H} = (h_{i,\omega,k})_{I\times\Omega\times K}$，および $\boldsymbol{U} = (u_{i,k,\tau})_{I\times K\times T}$ の最尤推定問題が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 $\boldsymbol{Y}'$ に対する板倉齋藤擬距離規準 NTF として定式化される．

また，本枠組みでは空気伝導マイクを NAM マイク付近に配置し，口から放射される微弱音声と外部雑音の音量差を利用することで雑音をモニタリングする [3]．よって，音源インデックス $i = 1$ が微弱音声に対応する場合，$s_{1,\omega,\tau}$ は NAM マイクでのみ観測されるため，$\boldsymbol{a}_{1,\omega,\tau} = [1, 0]^\top$ と固定する．さらに，空気伝導マイクで観測した複数の雑音を一つの雑音として考えることで，音源インデックス数を $I = 2$ とし，$\boldsymbol{a}_{2,\omega,\tau} = [|a_{1,2,\omega}| e^{j\phi_{1,2,\omega,\tau}}, 1]^\top$ 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．

3 セグメント特徴量正則化

従来法 [4] では事前に学習した目的音声の基底スペクトルおよび空気伝導マイクで観測した雑音信号が分離を行う際の手がかりとなっている．ただし，音声の基底スペクトルは雑音のスペクトルを表現し得るため，目的関数であるスペクトル距離を最小化したとしても個々の音源を完全に推定できるとは限らない．このような問題の対策として，特徴量空間の確率分布を正則化規準とする手法が提案されている [6]．本稿では，後段の変換処理で使用するメルケプストラムセグメント特徴量に着目し，推定された目的音声のスペクトログラムがその特徴量空間において，事前に学習した分布に則するよう制約を課す規準を設ける．

まず，従来法における目的関数は以下で表される．

$$\mathcal{D}_{\mathrm{IS}}(\boldsymbol{Y}'|\boldsymbol{X}) = \sum_{c,\omega,\tau} \left(\frac{|y_{c,\omega,\tau}|^2}{x_{c,\omega,\tau}} - \log \frac{|y_{c,\omega,\tau}|^2}{x_{c,\omega,\tau}} + 1 \right) \tag{5}$$

ただし，$x_{c,\omega,\tau} = \sum_i |a_{c,i,\omega}|^2 \sum_k h_{i,\omega,k} u_{i,k,\tau}$ とする．次に，目的とする体内伝導微弱音声のパワースペクトログラムに対して以下の規準を考える．

$$\mathcal{K}(\boldsymbol{X}) = \log \prod_\tau \left\{ \sum_m w_m \prod_n \mathcal{N}(\mathcal{X}_{n,\tau}; \mu_{n,m}, \sigma_{n,m}^2) \right\} \tag{6}$$

ここで，n はセグメント特徴量の次元インデックスを表す．w_m，$\mu_{n,m}$ および $\sigma_{n,m}^2$ はそれぞれ，m 番目の正規分布の重み，平均，分散であり，クリーンな目的音声を用いて事前に学習する．$\mathcal{X}_{n,\tau}$ は前後 L フレームから算出されるセグメント特徴量で，対数パワースペクトルに対する線形変換で表される．

$$\mathcal{X}_{n,\tau} = \sum_{\omega,l} \mathcal{A}_{n,\omega,l} \log p_{1,\omega,\tau-L+l} + \mathcal{B}_n \tag{7}$$

ここで，$\mathcal{A}_{n,\omega,l}$ は DFT 行列，周波数軸変換行列および主成分分析行列の積で構成される変換用テンソル，$\mathcal{B}_n$ はバイアスベクトルであり，それぞれ事前に計算される．提案法では，式 (5) および式 (6) の二つの

---

* Non-negative tensor factorization with segment feature regularization for nonaudible murmur enhancement. by TAJIRI, Yusuke (Nagoya Univ.), KAMEOKA, Hirokazu (NTT), TODA, Tomoki (Nagoya Univ.)

規準を考慮した目的関数

$$\mathcal{O}(\boldsymbol{A}', \boldsymbol{H}, \boldsymbol{U}) = \mathcal{D}_{\mathrm{IS}}(\boldsymbol{Y}'|\boldsymbol{X}) - \lambda\mathcal{K}(\boldsymbol{X}) \qquad (8)$$

を最小化することで，各パラメータを推定する．ここで，λ は正則化パラメータである．

補助関数法を用い，先行研究 [6] と同様の手順で式 (8) の上界関数を設計し，時刻フレームインデックスに関して変数変換を行うことで更新式が導出される．なお，正則化項はパラメータ $u_{1,k,\tau}$ の更新にのみ影響するため，その他のパラメータの更新については，従来法と同じ更新式を用いる．

4 実験的評価

4.1 実験条件

男性話者 1 名の微弱なささやき声を，NAM マイクおよび雑音モニタリング用の空気伝導マイクで同時収録する．また，変換処理時の目標音声として，通常音声を口元に配置した空気伝導マイクで収録する．収録文は ATR 音素バランス文 A セット中の 50 文とし，40 文を学習，残りの 10 文を評価に用いる．また，同じマイク配置で収録した以下の 4 種類の実環境雑音を重畳することで収録信号を生成する．

- **station_0dB**：駅構内の雑音（SNR = 0 dB）
- **restaurant_0dB**：飲食店の雑音（SNR = 0 dB）
- **crowd_5dB**：人混みの雑音（SNR = 5 dB）
- **traffic_5dB**：高架下の雑音（SNR = 5 dB）

サンプリング周波数は 16 kHz，FFT 分析フレーム長は 32 ms（窓長 25 ms），シフト長は 5 ms とする．NTF における各音源の基底数は 20，パラメータの更新回数は 5000 回，正則化パラメータ λ は 1 とし，求められたパラメータから構成されるウィナーフィルタを適用することで目的音声を推定する．目的音声の基底は事前に学習したものを使用し，固定する．変換処理における特徴量として，入力には FFT 分析による 0〜24 次のメルケプストラム係数から得られる 50 次元のセグメント特徴量（前後 4 フレーム使用）を用いる．出力には STRAIGHT 分析 [7] による 0〜24 次のメルケプストラム係数から得られる 50 次元の結合静的・動的特徴量を用いる．混合正規分布モデルの混合数は，正則化用に 128（対角共分散），スペクトル変換用に 32（全共分散）とする．また，残留雑音の影響を軽減するため，雑音抑圧後の後処理として既知雑音重畳による雑音の均質化 [8] を行う．既知雑音には白色雑音を使用し，学習および変換時の入力に対して SN 比 5 dB で重畳する．

4.2 雑音処理後の SN 比

Fig. 1 に結果を示す．NTF に基づく雑音抑圧を行うことで，未処理の場合（**Unprocessed**）から SN 比が改善していることがわかる．また，正則化なし（**NTF w/o reg**）と正則化あり（**NTF w/ reg**）を比較すると，正則化によりさらに 2 dB 前後改善しているのがわかる．

4.3 変換処理後のメルケプストラムひずみ

Fig. 2 に結果を示す．なお，メルケプストラムひずみは出力変換音声および目標通常音声の 1〜24 次のメルケプストラム係数から算出している．また，グラフの下限はクリーンな収録信号を使用した場合の理想値（約 5.45 dB）を表す．結果より，正則化を行うことで変換後のメルケプストラムひずみが改善しているのがわかる．改善度合いの違いから，提案法は低レベルの雑音環境において特に有効であると考えら

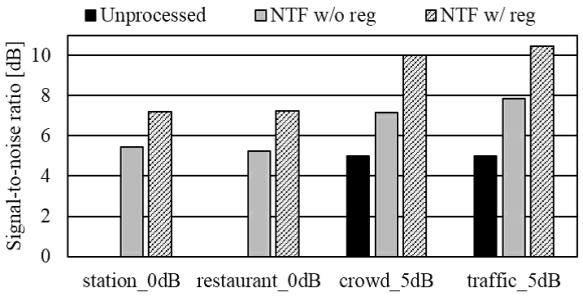

Fig. 1 Signal-to-noise ratio of estimated body-conducted soft speech

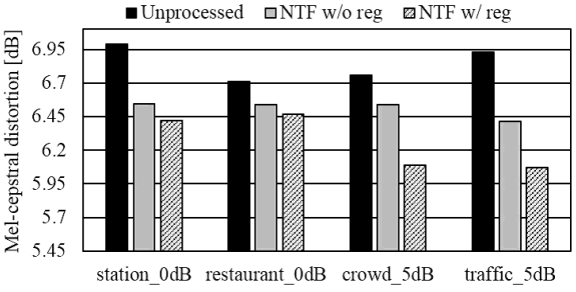

Fig. 2 Mel-cepstral distortion of converetd speech

れる．ただし，事前 SN 比が 5 dB の場合であっても理想値と比較して，ひずみの値に 0.6 dB 以上の差があることがわかる．

5 おわりに

従来の NTF に基づく体内伝導微弱音声の強調法では，後段処理で使用する特徴量を強調する保証がなく，前段処理としての有効性が低いという問題があった．本研究では後段処理として統計的声質変換に基づく体内伝導音声強調を行うことを前提とし，NTF によって推定された体内伝導音声のスペクトルが，メルケプストラムセグメント特徴量空間において，事前に学習した分布に則するよう正則化項を設計した．実験的評価結果より，SN 比が 5 dB の雑音環境では正則化を行うことで，スペクトル変換精度が大幅に改善されることを示した．今後は静穏環境での変換精度に近づけるため，さらなる改良に取り組む．

謝辞 本研究の一部は，JSPS 科研費 15K12064, 26280060 および 16J08977 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] Toda *et al.*, *IEEE Trans.*ASLP, Vol. 15, No. 8, pp. 2222-2235, 2007.
[3] Tajiri *et al.*, *Proc. ICASSP*, pp. 5935-5939, 2016.
[4] 田尻 他, 信学技報, Vol. 115, No. 523, pp. 117-122, 2016.
[5] Lee and Seung, Nature, Vol. 401, pp. 788-791, 1999.
[6] Li *et al.*, *Proc. INTERSPEECH*, pp. 3753-3757, 2016.
[7] Kawahara *et al.*, *Speech Commun.*, Vol. 27. No. 3-4, pp. 187-207, 1999.
[8] 山出 他, 信学論, Vol. 87, No. 4, pp. 933-941, 2004.