外部雑音モニタリングを用いた非可聴つぶやきに対する雑音抑圧法*

☆田尻祐介, 戸田智基, Graham Neubig, Sakriani Sakti, 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，音声入力システムの使用を可能にする技術として，体内伝導音声の一種である非可聴つぶやき（Non-Audible Murmur: NAM）を用いた音声認識が提案されている [1]．また，体内伝導収録に伴う音質劣化を統計的声質変換により改善し，音声通話へと利用する技術も提案されている [2]．ただし，従来の研究では，防音室のような静穏環境下で収録された NAM に対してのみ，その有効性が確認されている．専用マイクによる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はあるものの，NAM のような微弱信号の収録では，その影響を無視することはできない．したがって，これらの技術を実環境へ適用するには，外部雑音の影響を考慮する必要がある．

本稿では，NAM の微弱性を利用して外部雑音をモニタリングし，体内伝導収録信号に混入する雑音成分を抑圧する手法を提案する．また，実験的評価結果より，セミブラインド信号分離 [3] の枠組みを適用することで，高い雑音抑圧性能が得られることを示す．

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

NAM は周囲の人物が聴取困難なほど微弱なささやき声であり，マイクを口唇付近に設置しない限り，雑音環境下では空気伝導音声として収録するのは困難である．そこで，NAM の微弱性を利用し，Fig. 1 に示すようにマイクを口唇から離れた位置に設置することで，外部雑音のみの収録を行う．このとき，時刻 t における目的信号の NAM を $s_1(t)$，外部雑音を $s_2(t)$ とすると，体内伝導マイクおよび空気伝導マイクにおける観測信号は

$$x_1(t) = s_1(t) + h(t) * s_2(t) \quad (1)$$
$$x_2(t) \approx s_2(t) \quad (2)$$

で表されると考えられ，エコーキャンセラやノイズキャンセラと同様の問題として扱うことができる．ここで，$h(t)$ はマイク間の伝達特性を表し，最小二乗平均法（Least Mean Square: LMS）などの代表的な適応アルゴリズムにより推定できる．

3 セミブラインド信号分離の適用

エコーキャンセラでは，除去すべき回り込み音声に加えて，送話者の音声がマイクに入力されること（ダブルトーク）が，フィルタ推定の安定性において問題となる．提案する枠組みでは NAM の発話時が，ダブルトーク状態に相当するため，ブラインド信号分離（Blind Source Separation: BSS）を適用することで，この問題を解消する．

時間周波数領域での入力信号を $\boldsymbol{s}(f,\tau) = [s_1(f,\tau), s_2(f,\tau)]^\top$，観測信号を $\boldsymbol{x}(f,\tau) = [x_1(f,\tau), x_2(f,\tau)]^\top$ とすると，観測モデルは

$$\boldsymbol{x}(f,\tau) = \boldsymbol{H}(f)\boldsymbol{s}(f,\tau) \quad (3)$$

で表される．ここで，f は周波数ビン番号，τ は時刻フレーム番号を表す．また，$\boldsymbol{H}(f)$ は音源からマイクまでの伝達特性を表す混合行列であり，本稿では時間的に変化しないものとみなす．BSS では，$\boldsymbol{H}(f)$ の逆行列である分離行列 $\boldsymbol{W}(f)$ を，各音源の独立性に基づいて推定する．また，式 (2) より本枠組みでは外部雑音が既知となるため，セミブラインドな問題となり，分離行列の要素の一部を次のように固定できる．

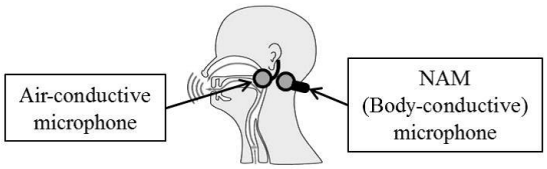


Fig. 1 Setting position of air- and body-conductive microphones

$$\boldsymbol{W}(f) = \begin{bmatrix} w_{11}(f) & w_{12}(f) \\ 0 & 1 \end{bmatrix} \quad (4)$$

したがって，自然勾配法 [4] に基づく独立成分分析（Independent Component Analysis: ICA）により分離ベクトル $\boldsymbol{w}(f) = [w_{11}(f), w_{12}(f)]$ を推定する場合，その更新式は次式で表される．

$$\Delta\boldsymbol{w} = \eta\{\boldsymbol{w}(f) - \langle\varphi(y_1(f,\tau))\boldsymbol{y}^{\mathrm{H}}(f,\tau)\rangle_\tau \boldsymbol{W}(f)\} \quad (5)$$
$$\boldsymbol{w}(f) \leftarrow \boldsymbol{w}(f) + \Delta\boldsymbol{w} \quad (6)$$

ここで，$\boldsymbol{y}(f,\tau) = [y_1(f,\tau), y_2(f,\tau)]^\top$ は分離信号，η はステップ幅，$\varphi(y_1(f,\tau))$ はスコア関数と呼ばれる非 2 次的関数，$\langle\cdot\rangle_\tau$ は時間平均演算を表す．スコア関数には，目的信号の分布がスーパーガウシアンであることを仮定し，次式を用いる．

$$\varphi(y_1(f,\tau)) = \tanh(|y_1(f,\tau)|)\exp(j\arg(y_1(f,\tau))) \quad (7)$$

周波数ビンごとのスケールの不定性を解消するため，逆投影法（Projection Back: PB）を用いると．最終的な推定信号は，次式で表される．

$$\hat{s}_1(f,\tau) = x_1(f,\tau) + \frac{w_{12}(f)}{w_{11}(f)} x_2(f,\tau) \quad (8)$$

4 実験的評価

4.1 実験条件

男性話者 1 名の NAM を，体内伝導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．このとき，シミュレーション用に，次の 3 種類の雑音を NAM とは別に各マイクで同時収録する．

- 60 dB の人混み雑音（crowd60dB）
- 70 dB の展示場の雑音（booth70dB）
- 80 dB の駅構内の雑音（station80dB）

ただし，雑音は話者前方に配置したスピーカーから提示し，音量は話者の頭部位置で測定した値とする．また，実際の雑音環境下での発話を想定して，上述した 3 種類の雑音をスピーカーで提示しながら NAM を各

*Noise suppression method for non-audible murmur using external noise monitoring. by TAJIRI, Yusuke, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani, NAKAMURA, Satoshi (NAIST)

マイクで同時収録する．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（1024 点），シフト長は 32 ms とする．セミブラインド信号分離（semi-BSS）におけるステップ幅 η は 0.01 とし，分離ベクトルの更新回数は 5，10，20，50，100，200 と変化させる．

比較対象として学習同定法 (Normalized LMS: NLMS)[5]，アフィン射影法（Affine Projection Algorithm: APA）[6]，再帰最小二乗法（Recursive Least Squares: RLS）[7] を用いる．ただし，これらの手法については，ダブルトーク対策として，口唇付近で収録したクリーンな空気伝導 NAM に対する音声区間検出（Voice Activity Detection: VAD）の結果を正解ラベルとして与え，ダブルトーク時には，フィルタの更新を停止する．各手法におけるパラメータ設定は Table 1 に示す通りである．また，線形時不変なフィルタによる雑音抑圧効果の上限値を検証するため，雑音のみの収録信号に対して，最小二乗法（Least Squares: LS）で推定したフィルタの性能も評価する．

Table 1　各手法におけるパラメータ設定

NLMS
フィルタ長：64 ms（1024 点）
ステップ幅：0.01，0.05，0.1，0.25，0.5
APA
フィルタ長：64 ms（1024 点）
ステップ幅：0.01，0.05，0.1，0.25，0.5
拘束条件数：2
RLS
フィルタ長：16 ms（256 点）
忘却係数：0.99，0.995，0.999，1

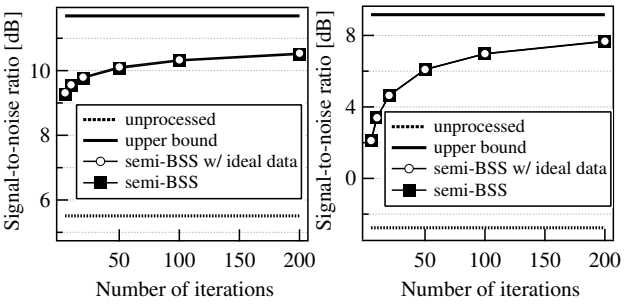
4.2　客観評価実験結果

NAM に雑音を重畳して生成した信号に対する semi-BSS の雑音抑圧効果を Fig. 2 に示す．Fig. 2 より，5 回程度の更新であっても，SN 比は処理前（unprocessed）と比較して大幅に改善され，更新を繰り返すことにより，上限値（upper bound）に近い値が得られることがわかる．また，雑音モニタリング用のマイクに NAM の成分が全く含まれない場合（semi-BSS w/ ideal data），すなわち $x_2(t) = s_2(t)$ が成立する場合の結果との差がほとんどないこともわかる．

次に，他の手法との比較結果を Fig. 3 に示す．ただし，semi-BSS における分離行列の更新回数は 200 回とし，他の手法におけるパラメータは，Table 1 内で最も SN 比の値が高くなるものを選択する．Fig. 3 より，60 dB の人混み雑音（crowd60dB）および 70 dB の展示場の雑音（booth70dB）に対しては，音声区間の正解ラベルが与えられる場合の NLMS，APA，RLS と同等の SN 比の改善がみられる．また，80 dB の駅構内の雑音（station80dB）に対しては，他の手法よりも高い SN 比が得られている．

4.3　主観評価実験結果

雑音環境下での発話を想定して収録した信号に対する雑音抑圧効果を検証するため，推定した NAM の音質を 5 段階 MOS（1：非常に悪い〜5：非常に良い）で評価する．被験者は日本人 10 名で，1 名あたり手法毎に 18 サンプル，合計 90 サンプルを受聴する．Fig. 4 に結果を示す．Fig. 4 より，クリーンな NAM（clean）には及ばないものの，処理前（unprocessed）と比較して，semi-BSS により音質が大幅に改善されることがわかる．80 dB の雑音に対する結果が 70 dB の雑音に対する結果を上回っているのは，雑音の音量増加に伴うロンバード効果 [8] により，NAM の発話自体が変化した影響であると考えられる．



(a) booth70dB　(b) station80dB

Fig. 2　Improvement in signal-to-noise ratio by using semi-BSS based noise suppression

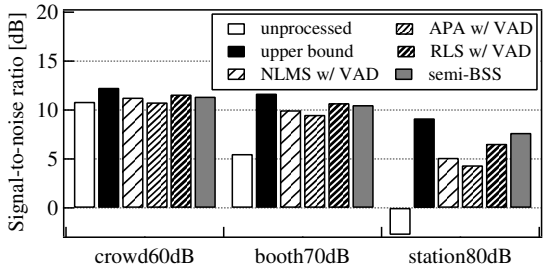


Fig. 3　Signal-to-noise ratio of estimated NAM

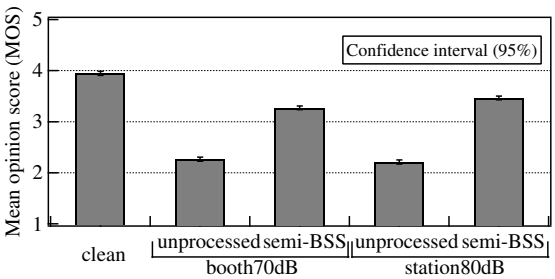


Fig. 4　Quality of clean and estimated NAM

5　おわりに

非可聴つぶやきの微弱性に着目した外部雑音モニタリングに基づき，体内伝導収録信号に混入する雑音成分を抑圧する手法を提案した．実験的評価結果から，代表的な適応アルゴリズムと比較して，セミブラインド信号分離の適用が有効であることを示した．今後は，拡散性雑音や移動音源が存在する状況での雑音抑圧効果の検証や，統計的声質変換との統合による，外部雑音に頑健なサイレント音声通話技術の構築に取り組む．

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] T. Toda *et al.*, *IEEE Trans.*ASLP, Vol. 20, No. 9, pp. 2505-2517, 2012.
[3] S. Miyabe *et al.*, *Proc. ICASSP*, pp. 109-112, 2006.
[4] S. Amari *et al.*, *Advances in neural information processing systems*, pp. 757-763, 1996.
[5] J. Nagumo *et al.*, *IEEE Trans.*AC, Vol. 12, No. 3, pp. 282-287, 1967.
[6] K. Ozeki *et al.*, *Electronics and Communication in Japan*, Vol. 67-A, No. 5, pp. 19-27, 1984.
[7] S. Haykin, *Adaptive filter theory*, Prentice Hall, fourth edition, 2002.
[8] T. Toda *et al.*, *Proc. INTERSPEECH*, pp. 632-635, 2009.