異種センサを用いて収録された非可聴つぶやき音声におけるブラインド音声抽出*

☆糸井三由希, 宮崎亮一，戸田智基，猿渡洋，鹿野清宏 (奈良先端大・情報)

1　はじめに

近年，音声検索システムやディクテーションソフトなど，音声を用いた携帯端末用アプリケーションが普及しつつある．これらは日々多機能化する携帯端末を直感的かつハンズフリーに扱うことができるため，今後もその需要は拡大していくと考えられる．一方で，実環境においては，発声が敬遠される静粛な環境や，他人に聞かれたくない情報を入力したい場合など，音声入力インタフェースの使用を躊躇する状況が多々存在する．そのため，秘匿性が高く，周囲に迷惑をかけないサイレント音声インタフェースの実現が望まれており，様々な研究が進められている [1] ．サイレント音声インターフェースの一つとして，非可聴つぶやき（Non-Audible Murmur: NAM）を用いた音声認識（NAM 認識）[2] が提案されている．NAM は，発話内容を周囲の者が聴受困難なほどの微弱な信号であり，体表に直接圧着させる専用のマイクロホン（NAM マイクロホン）を用いて，体内を伝導する音声として収録される．声道内の極めて微弱な共振音を収録できる反面，発話の際に話者が動くと，NAM マイクロホン の圧着面が変動することにより雑音が生じる．この雑音は NAM 認識性能を大きく低下させる要因となる．この問題に対して，複数の NAM マイクロホンでステレオ NAM 信号を収録し，ブラインドマルチチャネル雑音抑圧を行う手法が提案されている [3].

本報告では NAM マイクロホンだけでなく，スロートマイクロホン及び粘着式の NAM マイクロホンも用いて NAM 信号の 6 チャネル同時収録を行い，収録した信号群に対してブラインド空間的サブトラクションアレー (Blind Spatial Subtraction Array: BSSA)[4] を適用する．また，BSSA の雑音推定部に音声と拡散性雑音のスパース性を利用した ICA[5] の考え方に基づくスパース信号抽出 (Sparse Signal Extraction: SSE) を適用する．その際に使用する信号ペアとして，どのマイクロホンのペアの信号を用いれば，高い雑音抑圧精度を達成するのかを調査する．

2　非可聴つぶやき認識とユーザ動作雑音

2.1　非可聴つぶやき

NAM の音響学的な定義は,「声帯振動ではなく気道の乱流雑音を音源とする無声呼気音が，発話器官の運動による音響的フィルタ特性変換により調音されて，人体頭部の主に軟部組織を伝達したもの」である．NAM は，Fig. 1 のように，専用のマイクロホンを耳介後方下部に直接圧着させて収録される．NAM は微小な信号であるため，専用のアンプを用いて増幅される．Figure 2 に，収録された NAM 信号波形とそのスペクトログラムを示す．NAM の周波数成分は，高域になるに従い急速に減衰するため，約 4 kHz 以上の帯域では暗騒音の成分に埋もれてしまい，観測が困難であることが分かる．これは，口からの放射特性の影響が無いこと，および軟部組織伝達による高域遮断特性の影響を受けることに起因する [6].

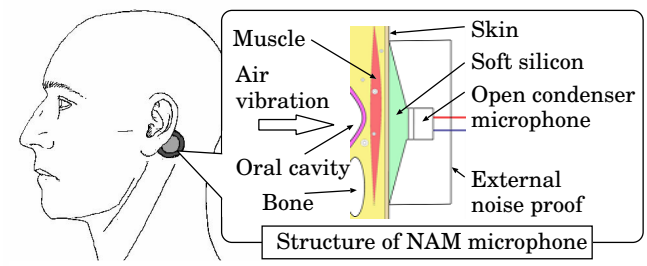

Fig. 1　Setting position and structure of NAM microphone.

2.2　NAM 認識とユーザ動作雑音の影響

従来の NAM 認識の研究 [2, 6, 7] では，NAM 収録の際，極力体を静止するよう話者に指示している．しかし，人が発話行為を行う際，体を完全に静止していることはなく，発話以外の何らかの動作を少なからず行っている．発話中，話者が頭を動かすなどの動作を行った場合，NAM マイクロホンの圧着面の皮膚の伸縮及び筋肉の隆起などにより，NAM マイクロホンの圧着状況が変化し，雑音が生じる．Fig. 3 に，話者が小さく首を横に振った状態での NAM 信号を示す．同図はネックバンドタイプの NAM マイクロホン [6] を用いて収録されたものであり，NAM マイクロホンは圧着位置に押し付けられる形で固定されるにも関わらず，ユーザ動作により非定常な雑音が生じることが分かる．この雑音により，NAM 認識の性能は大幅に劣化する [3].

3　他種マイクロホンの導入

本報告では先行研究で用いられていた NAM マイクロホンの他に Fig. 4 のスロートマイクロホンと Fig. 5 の粘着式 NAM マイクロホンを使用した．スロート

* “Blind speech extraction for non-audible murmur speech recorded by various microphones,” by Miyuki Itoi, Ryoichi Miyazaki, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano (Graduate of Information Science, Nara Institute of Science and Technology).

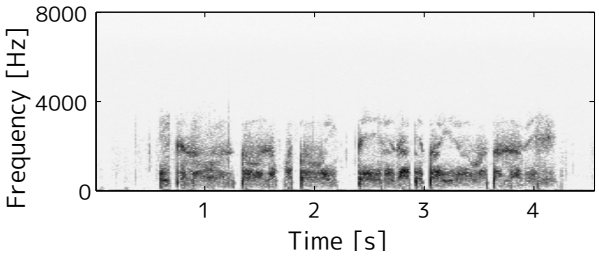

Fig. 2　Example of spectrogram of clean NAM signal.

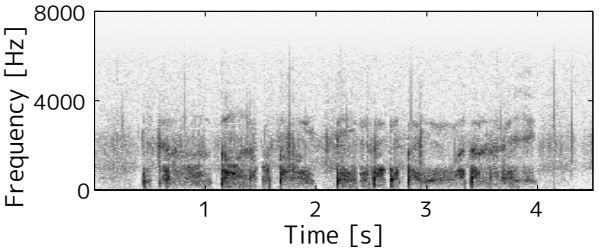

Fig. 3　Example of spectrogram of NAM signal when speaker moves during speaking.

マイクロホンは喉を包み込むようにして圧着させ，体内で発生した音声を皮膚を通して収録する．これは会話における通常音声をターゲットとした市販のマイクロホンである．粘着式NAMマイクロホンは，従来のNAMマイクロホンと同じく，発せられた音声を体内の軟部組織を伝導した音として拾うことができる．本マイクロホンの表面は粘着物質でできており，体表に粘着させることで固定する．粘着式マイクロホンを用いることで従来圧着させていた耳介後方下部以外の場所に圧着させることができ，圧着位置を限定しないNAM収録が可能である．

4　ステレオNAM信号を用いたブラインド雑音抑圧法

4.1　NAMと雑音の混合過程

ユーザ動作により生じる雑音を抑圧する手法として，各種マイクロホンによって収録されるステレオNAM信号を用いたブラインド雑音抑圧法について述べる．Fig. 6にそのブロック図を示す．この手法はBSSAに基づいており，雑音推定部および雑音抑圧部から成る．

ユーザ動作を伴うステレオNAM信号の時間周波数領域表現 $\boldsymbol{x}(f,\tau) = [x_1(f,\tau), x_2(f,\tau)]^\top$ （$\top$ は行列の転置を表す）は次式により近似的に表すことができる．

$$\boldsymbol{x}(f,\tau) \simeq \boldsymbol{a}(f) s_1(f,\tau) + \boldsymbol{n}(f,\tau) \qquad (1)$$

ここで，f は周波数，τ は時間フレーム番号を示し，$s_1(f,\tau)$ は体内伝導前のNAM信号であり未観測な信号である．また，$\boldsymbol{a}(f) = [a_1(f), a_2(f)]^\top$ は各チャネルごとの伝達関数を示し，NAMマイクロホンの圧着位置やアンプ設定などに依存する線形時不変フィルタで表される．一方で，ユーザ動作により生じるステレオ雑音信号は，各チャネルで異なる雑音源を持つものとして，$\boldsymbol{n}(f,\tau) = [n_1(f,\tau), n_2(f,\tau)]^\top$ と表される．

Fig. 4　Throat microphone.

Fig. 5　Adhesive NAM microphone.

4.2　雑音推定部

雑音と音声のスパース性を利用して雑音推定を行うSSEについて述べる．この手法では，音声と拡散性雑音の平均と分散の比率を利用して分離を行うことで，パーミュテーション問題の解決を回避することができ，高精度な音声推定が可能となる．

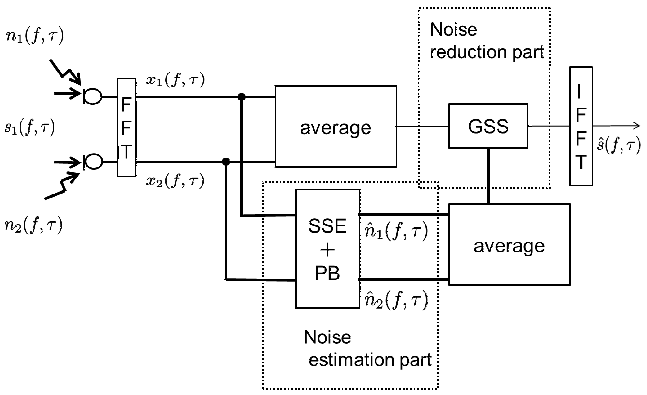

Fig. 6　Block diagram of blind noise reduction method for NAM.

推定分離音 $y(f,t)$ は拘束条件

$$\mathrm{E}\{|y(f,\tau)|^2\} = 1 \tag{2}$$

の基で抽出ベクトル $\boldsymbol{\omega}(f)$ を観測信号 $\boldsymbol{x}(f,\tau)$ に適用することで得られる．

$$y(f,\tau) = \boldsymbol{\omega}(f)\boldsymbol{x}(f,\tau) = \boldsymbol{\omega}(f)\boldsymbol{A}(f)\boldsymbol{s}(f,\tau) \tag{3}$$

ここで，$\boldsymbol{A}(f)$ は瞬時混合行列を表し，$\boldsymbol{s}(f,\tau)$ は発せられた信号の要素を表す．$\boldsymbol{s}(f,\tau)$ の第一要素である $s_1(f,\tau)$ は目的音声要素とする．抽出ベクトル $\boldsymbol{\omega}(f)$ は，下記条件を満たすコスト関数を最小化するように更新される．

$$J(\boldsymbol{\omega}(f)) = (\mathrm{E}\{|y(f,\tau)|\} - \gamma)^2 \tag{4}$$

ここで $\gamma \geq 0$ は散在さを制御するパラメータである．拘束条件式 (2) は $\mathrm{var}\{|y|\} + \mathrm{E}\{|y|\}^2 = 1$ と表すこともでき，

$$\mathrm{E}\{|y(f,\tau)|\} = \gamma \quad \text{and} \quad \mathrm{var}\{|y|\} = 1 - \gamma^2 \tag{5}$$

となる．γ が小さい時，抽出要素は平均が小さく分散の大きい係数を持つ．つまり，抽出要素の係数のほとんどが 0 に近く，少数のみが大きな値を持つことになる．拡散性背景雑音と目的音声に関して適用すると，音声係数は拡散性背景雑音の係数よりもまばらになることにより．目的音声の要素が抽出されたときに本コスト関数が最小となる．このため，SSE ではパーミュテーション問題の解決する必要はない．抽出ベクトル $\boldsymbol{\omega}(f)$ は以下の最急降下アルゴリズムで更新される．

$$\boldsymbol{\omega}^{[k+1]}(f) \quad = \quad \boldsymbol{\omega}^{[k]}(f) - \mu^{[k]} \frac{\partial J(\boldsymbol{\omega}(f))}{\partial \boldsymbol{\omega}(f)}\Big|_{\boldsymbol{\omega}(f)=\boldsymbol{\omega}^{[k]}(f)} \tag{6}$$

$\boldsymbol{\omega}_k(f)$ と μ は k 回目の反復における抽出ベクトルと更新ステップである．コスト関数の勾配は以下で与えられる．

$$\frac{\partial J(\boldsymbol{\omega}(f))}{\partial \boldsymbol{\omega}(f)} \quad = \quad 2\mathrm{E}\left\{\boldsymbol{x}(f,t)\frac{y(f,\tau)^{\mathrm{H}}}{|y(f,\tau)|}\right\}(\mathrm{E}\{|y(f,\tau)|\} - \gamma) \tag{7}$$

推定雑音は，観測信号から抽出された要素 $y(f,\tau)$ の直交射影を減算することによって得られる．目的音声の完全な抽出ができたと仮定すると，$\boldsymbol{\omega}(f)\boldsymbol{A}(f) = \lambda\boldsymbol{e}_1$ となり，ここで $\boldsymbol{e}_1$ は第一恒等列ベクトルであり，$|\lambda|^2\mathrm{E}\{|s_1|\}^2 = 1$ という拘束を持つ．この時，$\boldsymbol{A}(f)_{(1,:)}$ を $\boldsymbol{A}(f)$ の 1 行目の要素とすると，$y(f,\tau)$ の射影は

$$\hat{\boldsymbol{s}}(f,\tau) = \mathrm{E}\{\boldsymbol{x}(f,\tau)\}y(f,\tau) = \boldsymbol{A}(f)_{(1,:)}s_1(f,\tau) \tag{8}$$

となり，想定される信号の数を N とすると雑音要素は以下を行うことによって求めることができる．

$$\hat{\boldsymbol{n}}(f,\tau) \quad = \quad \boldsymbol{x}(f,\tau) - \hat{\boldsymbol{s}}(f,\tau) = \sum_{j=2}^{N} \boldsymbol{A}(f)_{(j,:)}s_j(f,\tau) \tag{9}$$

Table 1　Channel information

Channel 1	Left throat microphone
Channel 2	Right throat microphone
Channel 3	Left NAM microphone
Channel 4	Right NAM microphone
Channel 5	Left adhesive NAM microphone
Channel 6	Right adhesive NAM microphone

4.3　雑音抑圧部

雑音抑圧部では，雑音推定部で推定した雑音信号を用いて，混合信号に対して一般化スペクトル減算法 (generalized spectral subtraction: GSS) を適用する．BSSA のポスト処理に GSS を用いた時の推定 NAM 信号 $\hat{s}^{[\mathrm{GSS}]}(f,\tau)$ は次式で得られる．

$$\hat{s}^{[\mathrm{GSS}]}(f,\tau) \quad = \quad \begin{cases} \sqrt[2\xi]{|x(f,\tau)|^{2\xi} - \beta|\hat{n}(f,\tau)|^{2\xi}}e^{j\arg(x(f,\tau))} \\ \qquad (\text{if } \ |x(f,\tau)|^{2\xi} > \beta|\hat{n}(f,\tau)|^{2\xi}) \\ \eta \cdot x(f,\tau) \qquad (\text{otherwise}) \end{cases} \tag{10}$$

ここで，β は減算係数，η はフロアリング係数，ξ は指数乗ドメインパラメータを示す．

5　評価実験

5.1　実験条件

抽出対象信号は一般的な女性話者一名の NAM 信号とする．NAM 信号として，スロートマイクロホン 2 チャネル，NAM マイクロホン 2 チャネル，粘着式 NAM マイクロホン 2 チャネルの計 6 チャネルで同時収録したデータを使用する．スロートマイクロホンは喉頭部へ，NAM マイクロホンは従来の耳介後方下部に圧着させ，粘着式 NAM マイクロホンは本実験では鎖骨部に圧着させた．チャネル番号と各マイクロホンとの対応を表 1 に示す．首を横に振った時に生じた雑音と縦に振った時に生じた雑音を収録し，それぞれ SNR が 0 dB になるように混合した．各マイクロホンにおいて雑音を混合した信号に対し，様々なマイクロホンペアを用いて，BSSA を適用した．本稿では，SSE により推定した雑音のチャネル平均を，混合信号のチャネル平均から GSS を用いて減算した．なお，SSE の初期更新ステップ μ_0 は 0.1，γ は 0.001 に設定した．GSS の指数上ドメインパラメータは 0.1 に固定し，雑音抑圧量 (noise reduction rate: NRR)[8] が 15 dB になるように GSS の減算係数 β を調整した．フロアリング係数 η は 0.01 に設定した．

5.2　実験結果

各チャネルのペアにおいて BSSA を適用し，NRR が 15 dB を達成した時のケプストラム歪みを算出した結果を Figs. 7，8 に示す．付加した雑音に関しては，Fig. 7 は首を横に振った時の雑音，Fig. 8 は首を縦に振った時の雑音である．それぞれに対して BSSA

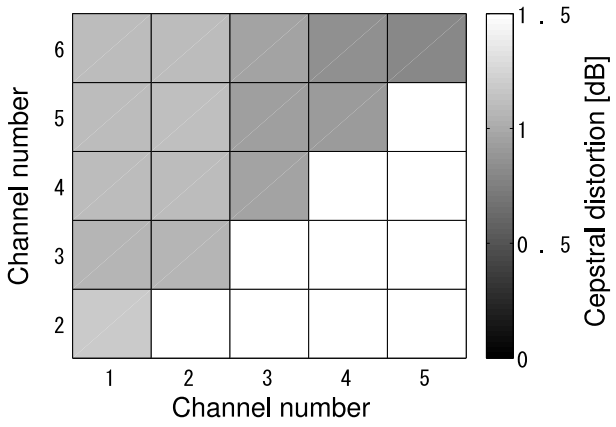

Fig. 7 The result of experiment with noise generated from “shaking” his/her head.

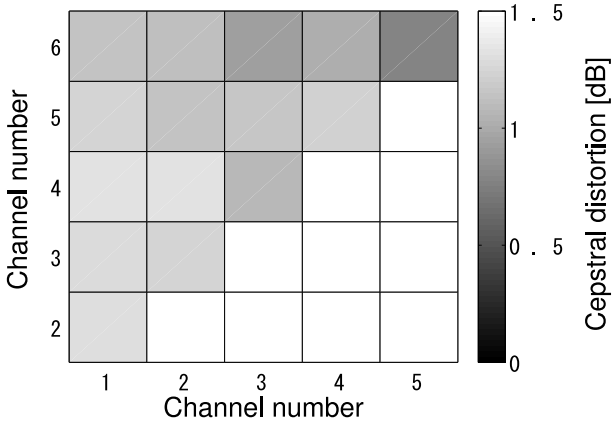

Fig. 8 The result of experiment with noise generated from “nodding” his/her head.

を適用し，各チャネルペアのデータを用いた時のケプストラム歪みの大きさを色の濃淡で示している．縦軸，横軸をそれぞれ表 1 に対応付けられるチャネル番号とし，ケプストラム歪みが小さいほど濃い色で塗りつぶされている．すなわち，色の濃いペアほど，高い雑音抑圧精度が達成できるマイクロホンペアであると言える．Fig. 7 においては，チャネル 5 とチャネル 6 ペア（粘着式 NAM マイクロホン同士）のケプストラム歪みが 0.69 dB となり，最小となった．また，Fig. 8 においても同ペアのケプストラム歪みが 0.67 dB となり最小値を示している．一方で，チャネル 1 とチャネル 2 のペア（スロートマイクロホン同士）に関しては，Fig. 7 では 1.11 dB，Fig. 8 では 1.25 dB となり，最大値を示した．すなわち，粘着式 NAM マイクロホン同士のペアが最も高い雑音抑圧精度を達成し，スロートマイクロホン同士のペアが最も低い雑音抑圧精度であることを確認した．

6 まとめ

本稿では，NAM マイクロホン，スロートマイクロホン，粘着式 NAM マイクロホンから得られた NAM 信号に対して BSSA を適用し，同種のマイクロホンペアのみならず，様々なマイクロホンペアを用いて推定される雑音の精度の比較を行った．その結果，粘着式 NAM マイクロホン同士のペアで最も高精度な雑音抑圧が可能であることを確認した．

謝辞 本研究の一部は，JST Core Research of Evolutional Science and Technology (CREST) により実施したものである．

参考文献

[1] B. Denby, et al., “Silent speech interfaces,” *Speech Communication*, vol.52, no.4, pp.270–287, 2010.

[2] Y. Nakajima, et al., “Non-audible murmur (NAM) recognition,” *IEICE Trans. Information and Systems*, vol.E89-D, no.1, pp.1–8, 2006.

[3] S. Ishii, et al., “Blind noise suppression for non-audible murmur recognition with stereo signal processing,” *Proc. ASRU*, pp.494–499, 2011.

[4] Y. Takahashi, et al., “Blind spatial subtraction array for speech enhancement in noisyenvironment,” *IEEE Trans. on Audio, Speech and Language Processing*, vol.17, no.4, pp.650–664, 2009.

[5] J. Even, H. Saruwatari, and K. Shikano, “Blind signal extraction based speech enhancement in presence of diffuse backgraund noise,” *2009 IEEE 15th Workshop on Statistical Signal Processing*, pp.513–516, 2009.

[6] T. Hirahara, et al., and K. Shikano, “Silent-speech enhancement using body-conducted vocal-tract resonance signals,” *Speech Communication*, vol.52, no.4, pp.301–313, 2010.

[7] T. Toda, et al., “Technologies for processing body-conducted speech detected with non-audible murmur microphone,” *Proc. INTERSPEECH*, pp.632–635, 2009.

[8] H. Saruwatari, et al., “Blind source separation combining independent component analysisand beamforming,” *EURASIP Journal on Applied Signal Processing*, vol.2003, no.11, pp.1135–1146, 2003.