非可聴つぶやき認識における ユーザ動作に伴う雑音に起因する性能低下の抑制*

☆石井隼太，戸田智基，猿渡洋，Sakriani Sakti，中村哲（奈良先端大・情報）

1 はじめに

静粛な環境など，発声行為自体を躊躇する状況においても音声入力を可能とする技術として，微弱な体内伝導音声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]．NAM マイクロフォンを体表に圧着することで NAM の収録が可能となる一方で，ユーザの動作によっては NAM マイクロフォンの圧着環境が大きく変動するため，収録信号に雑音が混入する．

本稿では，ユーザ動作に起因する雑音が NAM 認識に与える影響を調査し，2 個の NAM マイクロフォンで収録されるステレオ信号を用いた雑音抑圧法を提案する．また，実験的評価により，提案法の有効性を示す．

2 非可聴つぶやき認識

2.1 非可聴つぶやき

NAM は，周囲の者が発話内容を聴受困難なほど微弱な無声音声であり，耳介後方下部の皮膚に密着させた NAM マイクロフォンにより収録される．NAM 信号の例を Fig. 1 に示す．空気伝導の通常音声と比べて，その音響的特徴は大きく異なるため，NAM 認識を行うためには専用の音響モデルが必要となる．

2.2 ユーザ動作により生じる雑音

首の動作などのユーザの動きにより，NAM マイクロフォンの圧着状況が変化すると雑音が混入する．首を左右に振る動作をしながら発声した際の NAM 信号の例を Fig. 2 に示す．Fig. 1 と比較すると，非定常な雑音が重畳していることが分かる．なお，雑音の音響的特徴は，NAM マイクロフォンの圧着位置やアンプのゲイン設定，動作に伴う圧着面の変化など，様々な要因に依存する．

3 NAM のステレオ収録による雑音抑圧

NAM マイクロフォンを左右の耳介後方どちらに装着しても，NAM 信号の収録は可能である．一方で，ユーザ動作に伴う NAM マイクロフォンの圧着状況の変化は，必ずしも左右同一ではないため，生じる雑音信号は左右で異なる．そこで，左右二つの NAM マイクロフォンにより収録される NAM のステレオ信号を利用した雑音抑圧手法を提案する．

3.1 NAM と雑音の混合過程

ユーザ静止時に収録される NAM のステレオ信号は，チャンネル間で異なる音響的特性を持つが，その違いは静的なものと考えられる．そこで，周波数領域における 2 チャンネルの NAM 信号 $\boldsymbol{s}(f,\tau)$ を次式でモデル化する．

$$\boldsymbol{s}(f,\tau) \quad = \quad \boldsymbol{a}^{(s)}(f) s_0(f,\tau) \qquad (1)$$

ここで，f は周波数,τはフレーム番号，$s_0(f,\tau)$ は体内伝導前の NAM 信号である．また，$\boldsymbol{a}^{(s)}(f) = \left[a_1^{(s)}(f), a_2^{(s)}(f)\right]^\top$ は各チャンネルの伝達関数を表し，NAM マイクロフォンの圧着位置やアンプ設定などに依存する．ユーザ動作時に発生する雑音信号は両チャンネル間で異なるため，$\boldsymbol{n}(f,\tau) = [n_1(f,\tau), n_2(f,\tau)]^\top$ でモデル化する．本稿では，観測されるステレオ信号は NAM 信号と雑音信号の加算で表現されるものとし，次式でモデル化する．

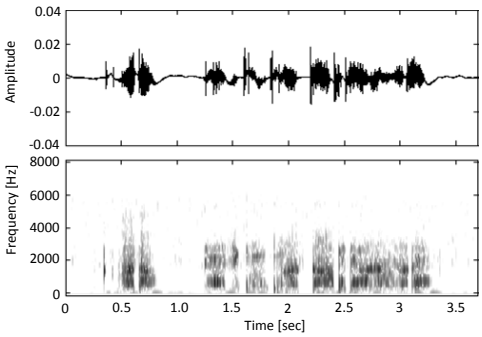

Fig. 1 Waveform and spectrogram of NAM signal

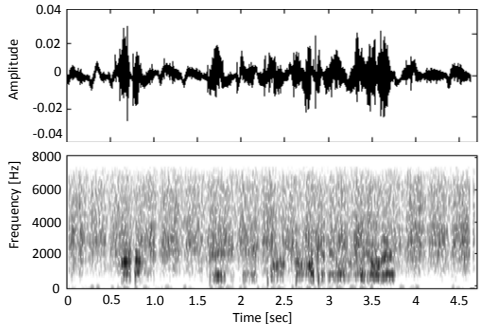

Fig. 2 Waveform and spectrogram of NAM signal when the speaker moves during speaking.

$$\boldsymbol{x}(f,\tau) \quad \simeq \quad \boldsymbol{a}^{(s)}(f) s_0(f,\tau) + \boldsymbol{n}(f,\tau) \qquad (2)$$

3.2 ブラインド空間サブトラクションアレー

式 (2) の混合過程において，ビームフォーミングなどの線形処理により，目的信号を高精度に抽出することは困難である．また，伝達関数 $\boldsymbol{a}^{(s)}$ の観測も容易ではない．そこで，ブラインド非線形処理による雑音抑圧法として，ブラインド空間サブトラクションアレー（Blind Spatial Subtraction Array: BSSA）[2] を適用する（Fig. 3）．まず，雑音推定部で，独立成分分析により，目的信号と雑音信号の分離フィルタ $\boldsymbol{W}_{\mathrm{ICA}}$ を学習する．

$$\boldsymbol{o}(f,\tau) \quad = \quad \boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{x}(f,\tau) \qquad (3)$$

ここで，$\boldsymbol{o}(f,\tau)$ は出力信号である．分離フィルタにより，雑音を抑圧することは困難であるが，目的信号を抑圧することは可能である．そこで，出力信号から推定雑音成分のみを取り出した信号 $\boldsymbol{o}^{(n)}(f,\tau)$ に射影

*Reduction of performance degradation of non-audible murmur recognition caused by noise generated depending on speaker's movements. by ISHII, Shunta, TODA, Tomoki, SARUWATARI, Hiroshi, SAKTI, Sakriani, and NAKAMURA, Satoshi (Nara Institute of Science and Technology)

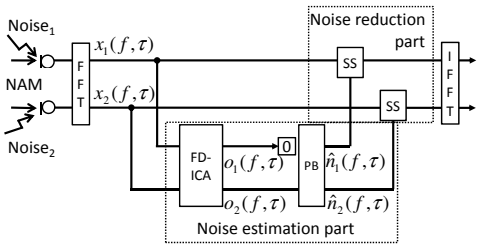

Fig. 3　Block diagram of BSSA.

法（Projection Back: PB）[3] を適用し，観測点でのパワーを持つ推定雑音信号 $\hat{\boldsymbol{n}}(f,\tau)$ を求める．

$$\hat{\boldsymbol{n}}(f,\tau) \;=\; \boldsymbol{W}^{+}_{\mathrm{ICA}}(f)\boldsymbol{o}^{(n)}(f,\tau) \qquad (4)$$

ここで，$\boldsymbol{M}^{+}$は$\boldsymbol{M}$のムーアペンローズの擬似逆行列である．その後，雑音抑圧部において，次式の一般化スペクトル減算法 (Generalized Spectral Subtraction: GSS)[4] により推定 NAM 信号 $\hat{\boldsymbol{s}}(f,\tau) = [\hat{s_1}(f,\tau), \hat{s_2}(f,\tau)]^{\top}$ を抽出する．

$$\hat{s}_c(f,\tau)=\begin{cases}\sqrt[2n]{|x_c(f,\tau)|^{2n}-\beta|\hat{n}_c(f,\tau)|^{2n}}\,e^{j\arg(x_c(f,\tau))} \\ \qquad (\text{if}\ \ |x_c(f,\tau)|^{2n} > \beta|\hat{n}_c(f,\tau)|^{2n}) \\ 0 \qquad (\text{otherwise})\end{cases} \qquad (5)$$

ここで，c はチャンネル番号，β は減算係数，n は指数乗ドメインを示す．[2] とは異なり，複数チャンネルの信号を同相化してモノラル信号を抽出するのは容易ではないため，チャンネル毎に GSS を行う．

3.3　チャンネル選択

ユーザ動作に伴う雑音は各チャンネルにおいて異なるため，BSSA における雑音推定性能は十分ではなく，推定 NAM 信号には GSS による歪みが生じる．チャンネル間における歪みの大小関係はフレーム毎に異なると考えられるため，より歪みの小さいチャンネルの選択をフレームレベルで行う．選択尺度として，各チャンネルにおいて，観測信号 $\boldsymbol{x}$ と推定雑音信号 $\hat{\boldsymbol{n}}$ からフレーム毎に次式で計算される信号対雑音比 $\mathrm{SNR}_{c,\tau}$ を用いる．

$$\mathrm{SNR}_{c,\tau} = 10\log_{10}\frac{\sum_f |x_c(f,\tau)|^2 - \sum_f |\hat{n}_c(f,\tau)|^2}{\sum_f |\hat{n}_c(f,\tau)|^2} \qquad (6)$$

各チャンネルの音響特徴量系列から，選択フレームを抽出することで，一つの音響特徴量系列を構築し，音声認識処理を行う．

4　実験

4.1　実験条件

提案法の有効性を示すために，大語彙連続音声認識実験を行った．話者は成人男性 1 名とした．NAM 発声時に首を横に振る動作をすることで，雑音が混入された信号（実混合信号）を収録した．また，同じ動作をした際の雑音のみの信号と，静止時に発声した NAM 信号も別途収録した．これらを足し合わせることで，式 (2) に示す混合過程に基づく信号（擬似混合信号）を作成した．サンプリング周波数は 16 kHz であり，DFT 点数を 1024，窓長を 512，シフト長を 256 として分析を行った．音響特徴量は 12 次元の MFCC，Δ MFCC，1 次元のΔパワーを用いた．

音響モデルは，通常音声の不特定話者モデルを初期モデルとして，最尤線形回帰 [5] による適応を行うことで作成した．適応データとして，208 発話を用い，二つのチャンネルで収録された NAM 信号を同時に用いた（計 416 発話相当）．言語モデルは，新聞記事から学習した 6 万語彙のトライグラムを用いた．評価データとして，143 発話を用いた．評価尺度は単語正解精度とした．

Table 1　Word accuracy [%]

	Simulated		Real	
	ch1	ch2	ch1	ch2
Unprocessed	53.6	52.1	53.6	52.5
GSS	55.5	52.9	56.7	54.6
BSSA	61.4	61.6	57.7	55.6
BSSA+selection	63.3		58.6	
Clean	69.2	67.3	69.2	67.3

音声認識実験は下記の信号に対して行った．

- **Unprocessed**: 未処理の混合信号
- **GSS**: GSS を適用した信号
- **BSSA**: BSSA を適用した信号
- **BSSA+selection**: BSSA 及びフレーム毎のチャンネル選択を適用した信号
- **Clean**: 静止状態での NAM 信号

なお，GSS を適応した信号（**GSS，BSSA，BSSA+selection**）に対しては，既知雑音重畳処理 [6] を施した．

4.2　実験結果

Table 1 に擬似混合信号 (Simulated) の認識実験結果を示す．ユーザ静止時と比較し，ユーザ動作により雑音が混入すると，大幅に認識性能が劣化することが分かる．GSS は雑音の定常性を仮定するため，その性能は限られる．一方で，提案法である BSSA は，大幅に認識性能低下を抑えることができ，チャンネル選択によりさらなる改善が得られる．

実混合信号（Real）の認識実験結果も同表に示す．提案法においてチャンネル選択を行うことで最も高い性能が得られることから，ステレオ信号を用いた雑音抑圧法の有効性が確認できる．なお，擬似混合信号の結果と比較すると，BSSA の性能が大きく劣化することが分かる．よって，実混合信号では，式 (2) の混合過程における NAM の伝達関数 $\boldsymbol{a}^{(s)}(f)$ もユーザ動作の影響を受けて変化すると考えられる．

5　おわりに

本稿では，NAM 認識において，ユーザ動作に起因する非定常雑音が大幅な認識性能低下を招くことを示した．また，2 つの NAM マイクロフォンにより収録されるステレオ信号を用いたブラインド雑音抑圧手法を提案した．BSSA およびチャンネル選択処理を導入することで，認識性能低下を効果的に抑制できることが分かった．

参考文献

[1] Y. Nakajima *et al.*, *IEICE Trans. Information and Systems*, E89-D(1), 1-8, 2006.
[2] Y. Takahashi *et al.*, *IEEE Trans. on Audio, Speech and Language Processing,* 17(4), 650-664, 2009.
[3] S. Ikeda and N. Murata, *Proc. ICA*, 365-370, Aussions, France, Jan. 1999.
[4] B. L. Sim *et al.*, *IEEE Trans. on Speech and Audio Processing,* 6(4), 328-337, 1998.
[5] M.J.F. Gales, *Computer Speech and Language,* 12(2), 75-98, 1998.
[6] S. Yamade *et al.*, *Proc. INTERSPEECH,* 1493-1496, Geneva, Switzerland, Sep. 2003.