非可聴つぶやき認識のための ブラインド雑音抑圧におけるステレオ信号統合法*

☆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

静粛な環境などの発話行為自体を躊躇する状況でも音声認識を行う技術として，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 は他人に聴受されないほど小さな無声音声であり，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って収録される．一方，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，収録信号に非定常な雑音が混入し，NAM 認識性能が著しく低下する．この問題に対して，ステレオ NAM 信号を用いたブラインド雑音抑圧法が提案されており，その有効性が確認されている [2]．本稿では，同手法にブラインドステレオ信号統合による目的音声強調処理を導入することによって，雑音抑圧性能が向上することを示す．

2 従来法 : ステレオ NAM 信号を用いたブラインド雑音抑圧

左右の耳介後方下部に NAM マイクロフォンを圧着させてステレオ NAM 信号を収録し，ブラインド空間的サブトラクションアレー（Blind spatial subtraction array: BSSA）[3] およびチャンネル選択に基づく雑音抑圧法 [2]（以下, 従来法）を用いることで，NAM 収録時のユーザ動作により生じる非定常な雑音を抑圧する．

2.1 NAM と雑音の混合過程

ユーザ動作時のステレオ 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）は次式でモデル化される．

$$\boldsymbol{x}(f,\tau) \simeq \boldsymbol{a}(f)s_0(f,\tau) + \boldsymbol{n}(f,\tau) \qquad (1)$$

f は周波数，τ はフレーム番号を示し，$s_0(f,\tau)$ は体内伝導前の NAM 信号であり未観測な信号である．$\boldsymbol{a}(f) = [a_1(f), a_2(f)]^\top$ は各チャンネルの伝達関数を示し，NAM マイクロフォンの圧着位置などに依存する時不変な線形フィルタである．ステレオ雑音信号は各チャンネルで異なる雑音源を持つものとして，$\boldsymbol{n}(f,\tau) = [n_1(f,\tau), n_2(f,\tau)]^\top$ と表す．

2.2 雑音抑圧処理

従来法（Fig.1）は，BSSA（雑音推定部，雑音抑圧部）とチャンネル選択部で構成される．雑音推定部では，周波数領域独立成分分析（FD-ICA）を用いて分離行列 $\boldsymbol{W}_{\mathrm{ICA}}$ を学習し，混合信号から推定 NAM 成分 $o_1(f,\tau)$，推定雑音成分 $o_2(f,\tau)$ を求める．

$$\boldsymbol{o}(f,\tau) = [o_1(f,\tau), o_2(f,\tau)]^\top = \boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{x}(f,\tau) \quad (2)$$

先行研究 [3] により推定精度が高いとされる $o_2(f,\tau)$ のみを抽出し，射影法（Projection Back: PB）[4] を適用することにより ICA の Scaling 問題を解決し，推定雑音信号 $\hat{\boldsymbol{n}}(f,\tau) = [\hat{n}_1(f,\tau), \hat{n}_2(f,\tau)]^\top$ を得る．

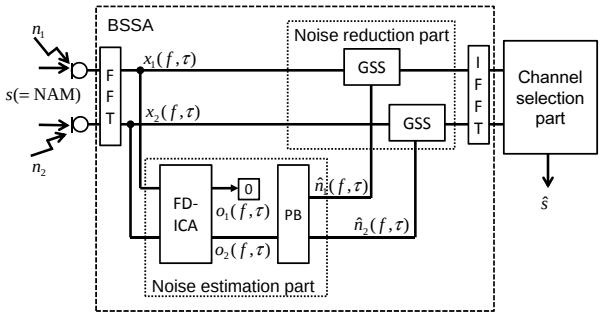

Fig. 1 Block diagram of blind noise suppression method for NAM recognition.

$$\hat{\boldsymbol{n}}(f,\tau) = \boldsymbol{W}^{+}_{\mathrm{ICA}}(f)\,[0, o_2(f,\tau)]^\top \qquad (3)$$

$\boldsymbol{M}^+$ は $\boldsymbol{M}$ のムーアペンローズの擬似逆行列を示す．雑音抑圧部では，$\hat{\boldsymbol{n}}(f,\tau)$ を用いて，$\boldsymbol{x}(f,\tau)$ に対して各チャンネルで一般化スペクトル減算法（Generalized spectral subtraction: GSS）[5] を適用することで，雑音を抑圧する．更に，$\boldsymbol{x}(f,\tau)$，$\boldsymbol{n}(f,\tau)$ を用いて，各チャンネルでフレーム毎に推定信号対雑音比（Signal-to-noise Ratio: SNR）を式 (4) で求める．

$$\mathrm{SNR}_{c,\tau}=10\log_{10}\frac{\sum_f |x_c(f,\tau)|^2 - \sum_f |\hat{n}_c(f,\tau)|^2}{\sum_f |\hat{n}_c(f,\tau)|^2} \quad (4)$$

各チャンネルの推定 SNR をフレーム毎に比較し，より推定 SNR が高いチャンネルの音響特徴量に切り替えることで，一つの音響特徴量系列を生成する．

2.3 問題点

従来の BSSA[3] では，ICA にて推定した目的信号の到来方向の情報を用いて，遅延和アレー（Delay and sum: DS）により多チャンネル混合信号を統合することで，目的信号を強調し，SNR を向上させている．一方，NAM の音響特性は NAM マイクロフォンの設置位置やアンプのゲイン設定などの影響を受けるため，左右のチャンネルで異なる伝達特性を持つ．また，[2] では，ユーザ動作により ICA の分離フィルタの精度が低下する事が示されており，目的信号の到来方向を正確に推定するのは困難である． これらの理由から，単純な DS によるチャンネル統合は容易ではない．

3 提案法:ステレオ NAM 信号の統合を適用したブラインド雑音抑圧

雑音抑圧性能をさらに向上させるために，各チャンネルの伝達特性を揃えてステレオ信号を統合するこ

*Stereo signal integration in blind noise suppression for Non-Audible Murmur recognition. by ISHII, Shunta, TODA, Tomoki, SARUWATARI, Hiroshi, SAKTI, Sakriani, and NAKAMURA, Satoshi (Nara Institute of Science and Technology)

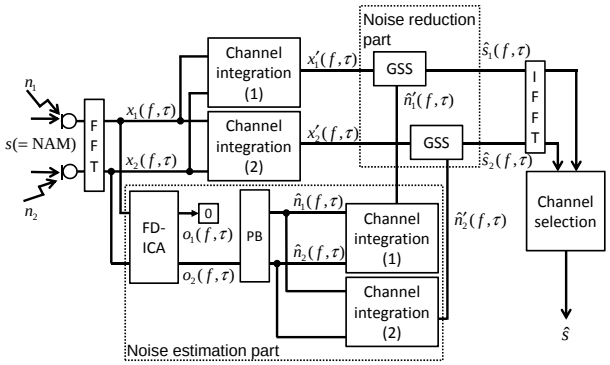

Fig. 2 Block diagram of blind noise suppression with channel integration for NAM recognition.

とで目的音声信号強調を行う手法として，ブラインドチャンネル統合法を提案する．提案法（Fig.2）では，次式で示す分離行列 $\boldsymbol{W}_{\mathrm{ICA}}$ において，推定 NAM 信号 $o_1(f,\tau)$ を求めるフィルタ係数 $[W_{11},\ W_{12}]$ を用いて，目的音声強調フィルタを設計する．

$$\boldsymbol{W}_{\mathrm{ICA}}(f) = \begin{bmatrix} W_{11}(f), & W_{12}(f) \\ W_{21}(f), & W_{22}(f) \end{bmatrix} \tag{5}$$

$[W_{11}(f),\ W_{12}(f)]$ を定数倍することで，チャンネル 1 の信号を変化させないフィルタ $[1,\ W_{12}(f)/W_{11}(f)]$ を得られる．さらにパワーの正規化を行うことで，チャンネル 1 に特性を揃える統合フィルタ $\boldsymbol{I}_1(f)$ を得る．

$$\boldsymbol{I}_1(f) = \frac{1}{1+\left|\frac{W_{12}(f)}{W_{11}(f)}\right|} \begin{bmatrix} 1, & \frac{W_{12}(f)}{W_{11}(f)} \end{bmatrix} \tag{6}$$

同様に，チャンネル 2 に特性を揃える統合フィルタ $\boldsymbol{I}_2(f)$ は次式で得られる．

$$\boldsymbol{I}_2(f) = \frac{1}{1+\left|\frac{W_{11}(f)}{W_{12}(f)}\right|} \begin{bmatrix} \frac{W_{11}(f)}{W_{12}(f)}, & 1 \end{bmatrix} \tag{7}$$

これらのフィルタにより，各チャンネルに特性を揃えた混合信号 $\boldsymbol{x}'(f,\tau) = [x'_1(f,\tau), x'_2(f,\tau)]^\top$ と推定雑音信号 $\hat{\boldsymbol{n}}'(f,\tau) = [\hat{n}'_1(f,\tau), \hat{n}'_2(f,\tau)]^\top$ を得る．対応する信号同士で GSS を行い推定 NAM 信号 $\hat{s}_1(f,\tau)$, $\hat{s}_2(f,\tau)$ を得た後，式 (4) と同様にチャンネル選択を行うことにより，更に雑音抑圧性能が向上できる．

4 評価実験

4.1 実験条件

一般的な成人男女各 1 名による NAM 信号を使用し，大語彙連続音声認識実験を行う．サンプリング周波数は 16 kHz とし，DFT 点数を 1024，窓長を 512，シフト長を 256 としてフレーム分析を行う．音響特徴量として，12 次元の MFCC および Δ MFCC，Δ パワーを用いる．音響モデルは，通常音声用不特定話者音響モデルを初期モデルとして，新聞記事を静止状態で読み上げたステレオ NAM データ（計 416 発話相当）を用いて最尤線形回帰 [6] を繰り返すことで構築する．評価データは各話者で 143 発話とし，首を左右に動かす動作をした際の NAM 信号を用いる．言語モデルは新聞記事から生成した 6 万語彙のトライグラムを用いる．評価尺度は単語正解精度とする．また，従来法と提案法の比較のため，下記に示す 6 つの信号を用いる．

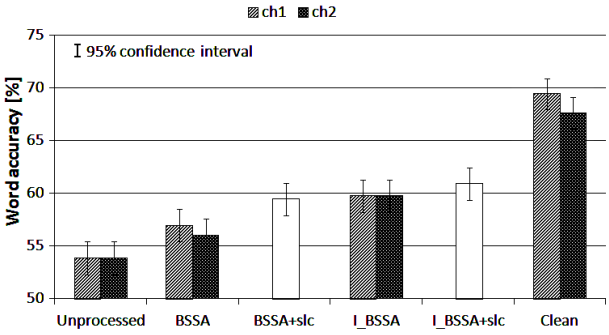

Fig. 3 Experimental result

- **Unprocessed**: 未処理の混合信号
- **BSSA**: BSSA を適用した信号（従来法）
- **BSSA+slc**: BSSA 及びチャンネル選択を適用した信号（従来法）
- **I_BSSA**: チャンネル統合 BSSA を適用した信号（提案法）
- **I_BSSA+slc**: チャンネル統合 BSSA 及びチャンネル選択を適用した信号（提案法）
- **Clean**: 静止状態での NAM 信号

4.2 実験結果

Fig.3 に実験結果を示す．ユーザ動作により認識性能が低下する中で，従来法（BSSA，BSSA+slc）により，単語正解精度の改善が見られる．チャンネル統合処理を行う提案法（I_BSSA，I_BSSA+slc）では，目的音声強調が行われるため，さらなる改善が見られる．

5 おわりに

本報告では，NAM 認識のためのユーザ動作雑音のブラインド抑圧手法に対して，さらにチャンネル統合処理による目的音声強調を導入する手法を提案した．雑音推定部で学習した分離行列の係数を用いて，NAM マイクロフォンの各チャンネルに特性を揃えてチャンネル統合を行うフィルタを設計し，目的音声強調を実現した．これにより，従来法よりも認識性能が向上することを示した．

謝辞: 本研究の一部は，科研費補助金基盤研究（A）により実施したものである．

参考文献

[1] Y. Nakajima *et al.*, *IEICE Trans. Information and Systems*, E89-D(1), 1-8, 2006.
[2] S. Ishii *et al.*, *ASRU*, Hawaii, USA, 494–499, 2011.
[3] Y. Takahashi *et al.*, *IEEE Trans. on Audio, Speech and Language Processing*, 17(4), 650–664, 2009.
[4] S. Ikeda and N. Murata, *Proc. ICA*, 365–370, Aussions, France, Jan. 1999.
[5] B. L. Sim *et al.*, *IEEE Trans. on Speech and Audio Processing*, 6(4), 328–337, 1998.
[6] M.J.F. Gales, *Computer Speech and Language*, 12(2), 75–98, 1998.