Vol.2011-SLP-88 No.1
2011/10/28

非可聴つぶやき認識のための ステレオ信号を用いたブラインド雑音抑圧法

石井隼太[†1] 戸田智基[†1] 猿渡洋[†1]
Sakuriani Sakti[†1] 中村哲[†1]

静粛な環境などの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（NAM 認識）が提案されている．NAM は多人に聴受されないほど小さなささやき声であり，体表に直接圧着させる NAM マイクロフォンによって収録される．その一方で，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．本報告では，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 つの 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．

Blind Noise Suppression for Non-Audible Murmur Recognition with Stereo Signals

SHUNTA ISHII,[†1] TOMOKI TODA,[†1] HIROSHI SARUWATARI,[†1]
SAKRIANI SAKTI[†1] and SATOSHI NAKAMURA[†1]

Recently, speech recognition with Non-Audible Murmur (NAM) was proposed in order to enable to use speech interfaces in quiet environments where we hesitate to speech. NAM is a very soft wispered voice detected with NAM microphone, which is one of the body-conductive microphones. The detected NAM signal suffers from noise caused by speaker's movement because the setting condition of NAM microphone is changed. In this paper, we investigate the effect of the noise on NAM recognition and propose a blind noise suppression method using a stereo signal detected with two NAM microphones. Experimental evaluations are conducted to show the effictiveness of the proposed method.

1. はじめに

近年，音声検索システムや音声翻訳ソフトなど，音声を用いた携帯端末用アプリケーションが注目を浴びている．音声認識システムはハンズフリーかつ直感的な端末操作が可能であることから，多機能化が進む携帯端末の普及と共に，その需要は拡大していくと考えられる．しかし，実環境においては，声を出すことをためらうような静粛な環境や，他人に聞かれたくない情報を入力したい場合など，音声認識システムの使用を躊躇する状況が多々存在する．そのため，使用する場所を選ばない音声認識システムの実現が求められる．

秘匿性が高く，周囲に迷惑を掛けない音声インタフェースの実現を目指し，様々なサイレント音声インタフェースの研究が進められている[1)]．その中の一つとして，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フォン（NAM マイクロフォン）を用いて，体内を伝導する音声として収録される．NAM 認識に関する従来研究（例えば[3)]）において，その認識性能が評価されているが，それらは全て，話者が静止した状況下におけるものである．NAM は体表に圧着した NAM マイクロフォンで収録されるため，発話時に話者が動くとその圧着面が変動し，それに伴い発生する雑音が NAM 認識性能に影響を与えると懸念される．NAM を用いたインターフェースの実用化を目指す上で，ユーザ動作を許容するシステムの実現は必要不可欠である．

本報告では，ユーザ動作により生じる雑音が，NAM 認識性能に与える影響を調査すると共に，ステレオ NAM 信号を用いた雑音抑圧手法を提案する．2 つの NAM マイクロフォンを用いてステレオ収録を行い，ブラインド空間サブトラクションアレー（Blind spatial subtraction array: BSSA）[4)] によりチャンネル毎に雑音抑圧を行った後，推定した信号対雑音比（Signal-to-noise ratio: SNR）がより高いチャンネルを選択する．提案法の性能を評価するため，大語彙連続音声認識実験を行い，有意な認識性能の改善が得られることを示す．

2. 非可聴つぶやき認識

2.1 非可聴つぶやき（NAM）

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，

†1 奈良先端科学技術大学院大学
Nara Institute of Science and Technology

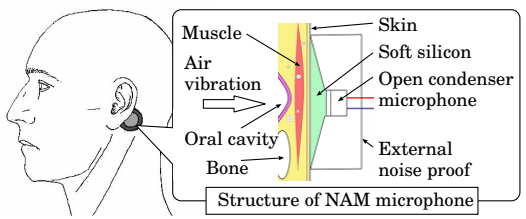

図 1 NAM マイクロフォンの構造と圧着位置
Fig. 1 Setting position and structure of NAM microphone.

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[2]．NAM は，図 1 のように，専用のマイクロフォ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．図 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 のスペクトログラムでは，約 4 kHz 以上の周波数成分が観測されないことが分かる．これは，口からの放射特性の影響が無いこと，かつ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[5]．

2.2 NAM の従来研究

NAM 認識に関わる従来研究として，混合正規分布を出力確率密度関数とする隠れマルコフモデル（Hidden Markov Model: HMM）に基づく音響モデルの構築が行われている．NAM は通常音声と比較してデータ量が少ないため，予め学習された通常音声用不特定話者音響モデルを初期モデルとして，最尤線形回帰（Maximum Likelihood Regression: MLLR）[6] によるモデル適応を繰り返すことで，NAM 用特定話者音響モデルを構築する．その際に，話者適応学習（Speaker Adaptive Training: SAT）により他の話者の NAM データを利用して初期モデルを改善することで，より高精度な音響モデルを構築できる．その結果，大語彙連続音声認識実験において，様々な話者に対して平均 70%以上の単語正解精度が得られている[3]．その一方で，これらの従来研究の結果は，話者の発話以外の動作を極力抑えた状態でのものであり，動作時に生じる雑音の影響については言及されていない．

2.3 ユーザ動作が NAM 収録に与える影響

話者が頭を動かすなどの動作を行った場合，NAM マイクロフォンの圧着面の皮膚の伸縮，筋肉の隆起が生じる．それにより，NAM マイクロフォンの圧着状況が変化する．図 2 に，話者が首を振った状態において，ネックバンドタイプの NAM マイクロフォン[3] を用いて収録した NAM 信号を示す．NAM マイクロフォンは圧着位置に押し付けられる形で固定されるものの，ユーザ動作により非定常な雑音が生じることが分かる．

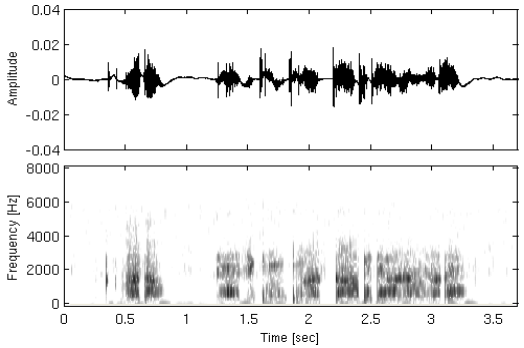

図 2 NAM の波形及びスペクトログラム
Fig. 2 Example of waveform and spectrogram of NAM signal.

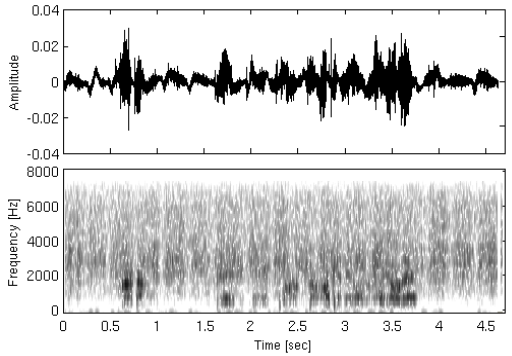

図 3 ユーザ動作を伴う NAM の波形及びスペクトログラム
Fig. 3 Example of waveform and spectrogram of NAM signal when the speaker moves during speaking.

3. ステレオ NAM 信号を用いたブラインド雑音抑圧

本報告では，ユーザ動作に伴う雑音を抑圧するため，2 つの NAM マイクロフォンを用いて収録されたステレオ信号を用いる手法を提案する．ステレオ NAM 信号は，NAM マイクロフォンを左右の耳介後方下部に圧着させて収録される．本節では，ステレオ NAM 信号と雑音信号の混合過程をモデル化し，その混合過程に適すると思われる雑音抑圧手法の適用に

Vol.2011-SLP-88 No.1
2011/10/28

ついて述べる．

3.1 NAM と雑音の混合過程

ユーザ静止状態において収録されたステレオ NAM 信号を図 4 に示す．各チャンネルの NAM 信号は互いに異なる音響特性を持つが，各チャンネル間で高い相関がある事が分かる．そこで，チャンネル 1 及びチャンネル 2 で収録されるステレオ NAM 信号の時間周波数領域表現 $\boldsymbol{s}(f,\tau) = [s_1(f,\tau), s_2(f,\tau)]^\top$ （$\top$ は行列の転置を示す）を次式でモデル化する．

$$\boldsymbol{s}(f,\tau) = \boldsymbol{a}(f) s_0(f,\tau) \tag{1}$$

ここで，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．また，$\boldsymbol{a}(f) = [a_1(f), a_2(f)]^\top$ は各チャンネルごとの伝達関数を示し，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時不変な線形フィルタで表される．なお，予備実験により，本モデル化の妥当性は確認している[⋆1]．

首を左右に振った時に生じる雑音のステレオ信号を図 5 に示す．動作に応じて雑音が生成されるものの，各チャンネルの雑音信号は完全に同期しているわけではなく，相関が低いことが分かる．従って，ステレオ雑音信号を次式のようにモデル化する．

$$\boldsymbol{n}(f,\tau) = \boldsymbol{b}(f,\tau) n_0(f,\tau) \tag{2}$$

ここで，$n_0(f,\tau)$ は未知である雑音の原信号，$\boldsymbol{b}(f,\tau) = [b_1(f,\tau), b_2(f,\tau)]^\top$ は各チャンネルにおける NAM マイクロフォンの圧着状況の変化に依存する時変の伝達関数であり，互いに独立である．すなわち，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，$\boldsymbol{n}(f,\tau) = [n_1(f,\tau), n_2(f,\tau)]^\top$ と表せる．

NAM 信号に雑音信号が加算的に重畳されると仮定すると，ユーザ動作時のステレオ NAM 信号は，

$$\boldsymbol{x}(f,\tau) \simeq \boldsymbol{a}(f) s_0(f,\tau) + \boldsymbol{n}(f,\tau) \tag{3}$$

で表される．ここでは混合過程を簡単化するため，NAM の伝達関数 $\boldsymbol{a}(f)$ はユーザの動作に依存しないとする．この仮定の真偽については 4.2 節で考察する．

3.2 ブラインド空間サブトラクションアレー

式 (3) の混合過程において，雑音信号は指向性を持たないため，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，NAM マイクロフォンの圧着状況は話者によって異なることから，NAM 伝達関数 $\boldsymbol{a}(f)$ 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り高い雑音抑圧精度が得られる方法として，ブラインド空間サブトラクションアレー（Blind spatial subtraction array: BSSA）[4] を適用する．BSSA は独立成分分析（Independent component analysis: ICA）を用いた適応ビームフォーマにより，目的信号を消去し雑音信号の推定を行う雑音推定部と，推定した雑音信号を用いて一般化スペクトル減算（Generalized spectral subtraction: GSS）[7] を行う雑音抑圧部の 2 つから構成される．BSSA はビームフォーミングでは困難である拡散性雑音の抑圧にも対応できる．またブラインドな雑音抑圧法であり，NAM マイクロフォンの圧着状況などに依存する伝達関数の情報などが不要である．これらのことから，BSSA は NAM と雑音の混合過程に適した雑音抑圧法であると考えられる．図 6 に，BSSA のブロック図を示す．

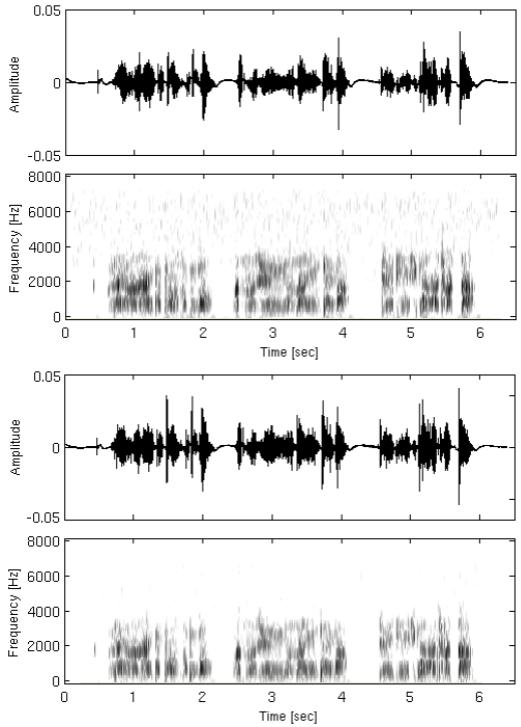

図 4 ステレオ NAM 信号の波形とスペクトログラム（上： チャンネル 1，下： チャンネル 2）
Fig. 4 Example of waveform and spectrogram of stereo NAM signal (top: the 1st channel, bottom: the 2nd channel).

⋆1 エコーキャンセル技術のように，線形フィルタをかけた片方のチャンネルの NAM 信号を用いて，他方のチャンネルの NAM 信号を大幅に抑圧できることを確認した．

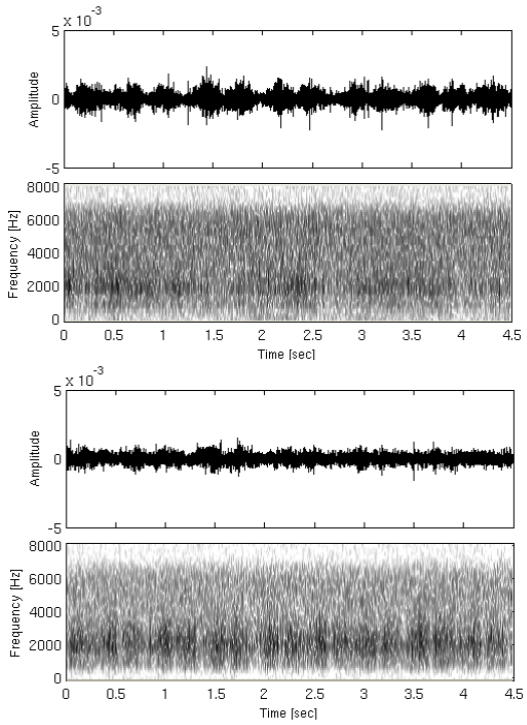

図 5 ステレオ雑音信号の波形とスペクトログラム（上： チャンネル 1，下： チャンネル 2）
Fig. 5 Example of waveform and spectrogram of stereo noise signal caused by speaker's movement (top: the 1st channel, bottom: the 2nd channel).

3.2.1 雑音推定部

雑音推定部では，周波数領域での ICA（FD-ICA）を用いて雑音を推定する．ICA では，出力ベクトル $\boldsymbol{o}(f,\tau) = [o_1(f,\tau), o_2(f,\tau)]^\top$ が互いに独立になるよう学習した分離行列 $\boldsymbol{W}_{\mathrm{ICA}}(f)$ を用いて，混合信号の分離を行う．

$$\boldsymbol{o}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{x}(f,\tau) \tag{4}$$

$\boldsymbol{W}_{\mathrm{ICA}}(f)$ は，出力ベクトルの結合確率密度関数 $p(\boldsymbol{o}(f,\tau))$ と，周辺確率密度関数 $p(o_1(f,\tau))p(o_2(f,\tau))$ の Kullback-Leibler 距離を最小化するように，次式のように学習される．

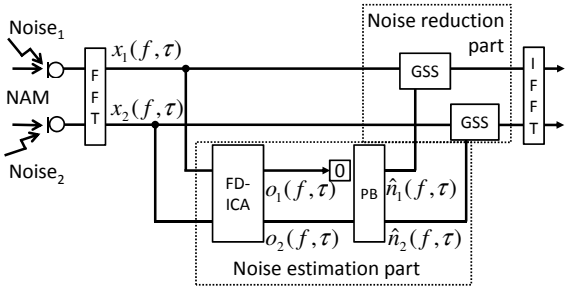

図 6 NAM 認識のための BSSA のブロック図
Fig. 6 Block diagram of BSSA for NAM recognition.

$$\boldsymbol{W}_{\mathrm{ICA}}^{[i+1]} = \boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f) + \alpha\left[\boldsymbol{I} - \langle\boldsymbol{\Phi}(\boldsymbol{o}(f,\tau))\boldsymbol{o}^H(f,\tau)\rangle_\tau\right]\boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f) \tag{5}$$

ここで，α は更新係数，$[i]$ は更新回数，$\boldsymbol{I}$ は単位行列，$\langle\cdot\rangle_\tau$ は時間平均，H は複素共役転置，$\boldsymbol{\Phi}(\cdot)$ は非線形関数を示す[8]．なお，分離行列は発話毎に学習を行う．

式 (3) の混合過程において，指向性を持たない雑音信号は，ICA が学習する線形フィルタによる適応ビームフォーマによって除去することは難しいため，NAM 信号成分 $s_0(f,\tau)$ の推定精度は低い．その一方で，指向性を持つ目的信号は同ビームフォーマで効果的に除去できるため，雑音信号成分 $n_0(f,\tau)$ の推定精度は高いことが知られている[4]．従って，出力ベクトルから推定 NAM 信号成分を除去し，推定雑音信号成分を抽出する．

$$\boldsymbol{o}^{(n)}(f,\tau) = [0, o_2(f,\tau)]^\top \tag{6}$$

この時，ICA のパーミュテーション問題を解決するため，$o_2(f,\tau)$ が推定雑音信号となるように分離行列 $\boldsymbol{W}_{\mathrm{ICA}}$ の初期値を適切に設定する[8]．そして，射影法 (Projection Back: PB)[9] により Scaling 問題を解決し，観測点での推定雑音信号 $\hat{\boldsymbol{n}}(f,\tau) = [\hat{n_1}(f,\tau), \hat{n_2}(f,\tau)]^\top$ を得る．

$$\hat{\boldsymbol{n}}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}^{+}(f)\boldsymbol{o}^{(n)}(f,\tau) \tag{7}$$

ここで，$\boldsymbol{M}^+$ は $\boldsymbol{M}$ のムーアペンローズの擬似逆行列を示す．以上の処理で得られる $\hat{\boldsymbol{n}}(f,\tau)$ の推定精度は十分に高いものではなく，後段において時間領域での雑音抑圧処理に用いるのは困難であるが，周波数領域（パワースペクトル領域など）での雑音抑圧処理においては有効に利用できる．なお，雑音推定精度を向上させるため，観測信号における無音声区間の雑

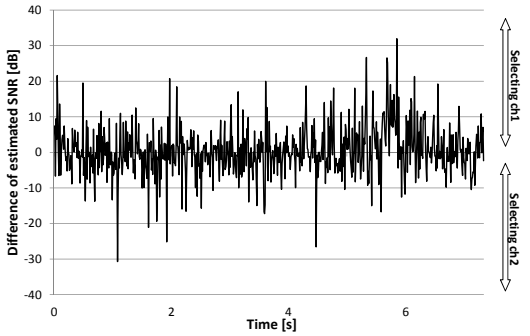

図 7 各フレームにおけるチャンネル間の推定 SNR の差分
Fig. 7 Difference of SNRs estimated frame by frame in individual channels (*i.e.*, $\mathrm{SNR}_{1,\tau} - \mathrm{SNR}_{2,\tau}$ in Eq. (9)).

音パワースペクトルを用いて推定雑音信号のパワースペクトルの補正を行う．

3.2.2 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，観測信号に対して GSS を適用することで，雑音を抑圧する．文献 4) では，観測信号と推定雑音信号に遅延和アレー（Delay and sum: DS）を適用し，モノラルの観測信号と推定雑音信号を生成してから GSS を行う．しかし，NAM の場合，各チャンネルの伝達関数は互いに異なっており，到来方向の情報だけでそれらを求めることは困難である．従って提案法では，それらの信号を DS によって同相化することはせず，各チャンネルそれぞれで GSS を行う．推定 NAM 信号 $\hat{\boldsymbol{s}}(f,\tau) = [\hat{s}_1(f,\tau), \hat{s}_2(f,\tau)]^\top$ は次式で得られる．

$$\hat{s}_c(f,\tau) = \begin{cases} \sqrt[2\xi]{|x_c(f,\tau)|^{2\xi} - \beta|\hat{n}_c(f,\tau)|^{2\xi}}\, e^{j\arg(x_c(f,\tau))} & (\text{if } \ |x_c(f,\tau)|^{2\xi} > \beta|\hat{n}_c(f,\tau)|^{2\xi}) \\ 0 & (\text{otherwise}) \end{cases} \tag{8}$$

ここで，c はチャンネル番号，β は減算係数，ξ は指数乗ドメインパラメータを示す．

推定後の NAM 信号には，GSS で抑圧しきれなかった残留雑音成分が存在し，また，GSS により人工的な歪が生じる．従って，推定 NAM 信号は，静止した状態での NAM 信号とは異なった音響特性を持つ．そこで，推定 NAM 信号と音響モデル作成のための適応 NAM データに既知雑音重畳処理[10] を行う．重畳する雑音は予め定めた定常雑音信号を用い，一定の SNR で対象信号に加算される．これにより，GSS により生じる音響特性差の影響を緩和することができ，雑音の種類ごとに異なる音響モデルを用意することが不要となる．

表 1 比較信号
Table 1 Compared signals.

信号名	詳細	信号処理区分
Unprocessed	未処理の混合信号	-
GSS	GSS を適用した信号	モノラル
BSSA	BSSA を適用した信号	ステレオ
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	ステレオ
Clean	静止状態での信号	-

3.3 チャンネル選択

3.1 節で示したとおり，ユーザ動作に伴う雑音は非定常であり，各チャンネル間で非同期である．従って，雑音の影響が大きいチャンネルは短時間毎に切り替わることが予想される．このことから，雑音の影響がより小さいチャンネルを時間フレーム毎に選択することにより，認識性能が向上すると考えられる．提案法では，選択尺度として観測信号 $\boldsymbol{x}(f,\tau)$ と推定雑音信号 $\hat{\boldsymbol{n}}(f,\tau)$ から得られる観測信号の推定 SNR を用いる[⋆1]．

$$\mathrm{SNR}_{c,\tau} = 10\log_{10}\frac{\sum_f |x_c(f,\tau)|^2 - \sum_f |\hat{n}_c(f,\tau)|^2}{\sum_f |\hat{n}_c(f,\tau)|^2} \tag{9}$$

各チャンネルの推定 SNR を比較し，フレーム毎にチャンネル 1，2 の音響特徴量を切り替えることで，一つの音響特徴量系列を生成する．図 7 に，ある発話の各フレームにおけるチャンネル間の推定 SNR の差分 $(\mathrm{SNR}_{1,\tau} - \mathrm{SNR}_{2,\tau})$ を示す．縦軸に正値をとるフレームではチャンネル 1 が，負値をとるフレームではチャンネル 2 が選択されることを示しており，各フレーム毎に選ばれるチャンネルが異なることが分かる．

4. 評 価 実 験

4.1 実 験 条 件

一般成人男性 1 名による NAM 信号を使用する．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．また，式 (3) の混合

⋆1 他の選択尺度として雑音抑圧後の推定 SNR なども考えられるが，予備実験の結果，式 (9) の選択尺度により最良の認識精度が得られたため，本報告ではこれを採用する．

過程における NAM 伝達関数の時不変性を確かめるため，2 つの混合信号で実験を行う．一つは首を左右に振る動作をした際の雑音のみの信号と，静止時での NAM 発声により収録される NAM 信号とを足し合わせた信号（擬似混合信号）であり，もう一つは同じ動作をしながら NAM 発声を行った際の信号（実混合信号）である．

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，音響モデルは Left-to-right の 3 状態トライフォン HMM で，共有状態数は 2189，出力確率分布は混合数 16 の GMM を使用する．通常音声用不特定話者音響モデルを初期モデルとして，MLLR 適応を 10 回繰り返すことで，NAM 特定話者用音響モデルを構築する．適応データとして，新聞記事 208 文を静止状態で読み上げた NAM データを用いる．チャンネル選択を行う手法では，2 チャンネル分 416 発話で適応した音響モデルを使用し，チャンネル選択を行わない手法では，各チャンネル 208 発話で適用した音響モデルをそれぞれ別に使用する．評価データは 143 発話とする．言語モデルは新聞記事から作成した 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．

雑音抑圧手法間の比較のため，表 1 に示す 5 つの信号を用いる．なお，GSS を適用した信号（GSS，BSSA，BSSA+selection）に対しては，3.2.2 節で述べた既知雑音重畳処理を適用する．雑音は白色雑音とし，適応データと雑音抑圧後の信号の SNR が 30 dB となるよう重畳する．指数乗ドメインパラメータは 1/3 とする．

4.2 実験結果

擬似混合信号に対して，減算係数を変化させた時の，BSSA による雑音抑圧後の NAM 信号のケプストラム歪（Cepstral distortion: CD）を図 8 に，雑音抑圧量（Noise reduction rate: NRR）を図 9 に示す．減算係数を大きくするにつれ，NRR が大きくなり，より高い雑音抑圧効果が得られることが分かる．一方で，同時に CD も大きくなるため，NAM の音響特徴量の歪も大きくなることが分かる．そのため，最良の認識性能を得るには，これら 2 つの要因を考慮して減算係数を決める必要がある．減算係数を変化させた時の，擬似混合信号における BSSA の単語正解精度を図 10 に，実混合信号におけるそれを図 11 に示す．最も高い単語正解精度が得られる減算係数は，擬似混合信号ではチャンネル 1，チャンネル 2 共に 0.5，実混合信号ではチャンネル 1，チャンネル 2 共に 0.1 であることが分かる．一般の通常音声の場合の結果と比較すると[4]，最適な減算係数は小さくなる傾向が見られる．また，実混合信号においては，その傾向がさらに顕著となる．以降の実験では，減算係数は上記の最適値に設定する．

図 12 に，擬似混合信号における実験結果を示す．Clean のチャンネル 1 とチャンネル 2 の

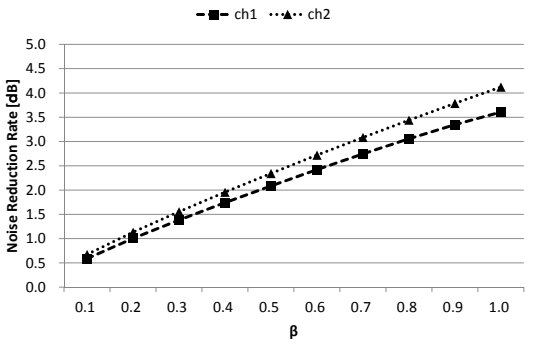

図 8 雑音減算量と減算係数の関係
Fig. 8 Noise reduction rate as a function of oversubtraction parameter β.

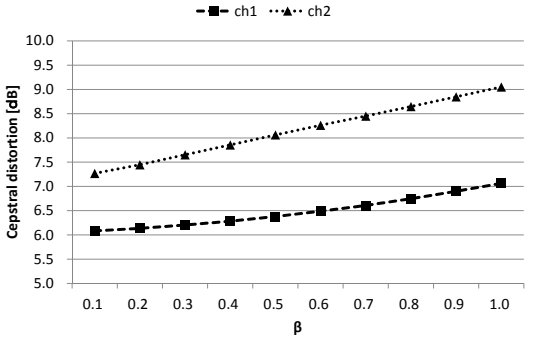

図 9 ケプストラム歪と減算係数の関係
Fig. 9 Cepstral distortion as a function of oversubtraction parameter β.

それぞれの単語正解精度 69.2%，67.3% と比較し，Unprocessed の各チャンネルの単語正解精度は，それぞれ 53.6%，52.1%と大きく低下している．このことから，ユーザ動作に伴う雑音は，認識性能に大きく影響をおよぼすことが分かる．モノラル信号処理である GSS を適用した場合，各チャンネルで 55.5%，52.9%と僅かな改善しか見られない．これは，GSS では定常雑音抑圧を仮定しており，フレーム毎の雑音推定を行っていないためである．非定常雑音抑圧に対応したステレオ信号処理である BSSA は 61.4%，61.6%と有意な改善が見られている．さらにチャンネル選択も行った BSSA+selection では，63.3%の単語正解精度が得られる．なお，Clean の認識精度には及ばない理由として，雑音推定精度が十分に高くな

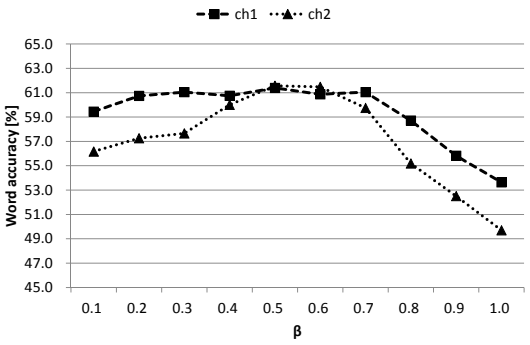

図 10 単語正解精度と減算係数の関係 (擬似混合信号)
Fig. 10 Word accuracy as a function of oversubtraction parameter β in *simulated mixed-signals*.

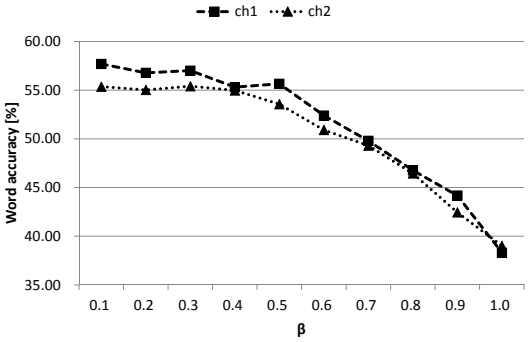

図 11 単語正解精度と減算係数の関係（実混合信号）
Fig. 11 Word accuracy as a function of oversubtraction parameter β in *real mixed-signals*.

いことが考えられる．

図 13 に実混合信号の結果を示す．擬似混合信号の結果と比較すると，BSSA の認識性能が 57.7%，55.6%と大きく低下しており，GSS の認識性能 56.7%，54.6%と有意な差は見られない．それに伴ない BSSA+selection の認識精度も低下しているが，依然として他手法より高い認識性能である 58.6%を得ている．これは，GSS と BSSA のチャンネル 1 の結果と比較すると有意差は認められないが，チャンネル 2 の結果と比較すると有意な改善が認められる．GSS や BSSA では，最終的にいずれかのチャンネルを選択しなければならない．認識性能がより高いチャンネルは，NAM マイクロフォンの圧着状況など，様々な条件に依存

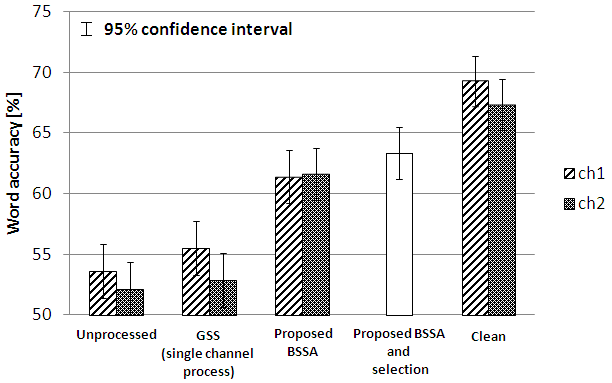

図 12 擬似混合信号の実験結果
Fig. 12 Result for *simulated mixed-signals*.

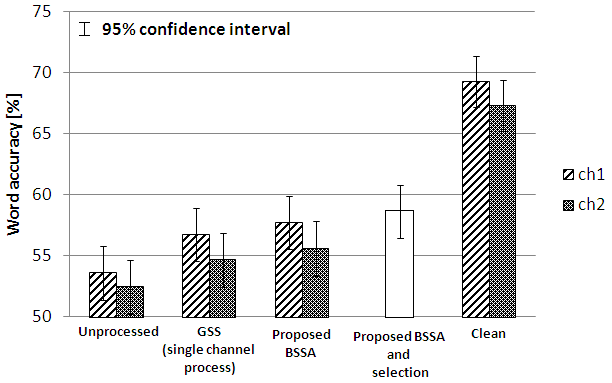

図 13 実混合信号の実験結果
Fig. 13 Result for *real mixed-signals*.

して決まる．常に片方のチャンネルの認識性能が高くなるとは限らない．そのため，GSS や BSSA において，認識性能の高いチャンネルを自動的に選択するのは容易ではない．一方で，BSSA+selection ではチャンネル選択を自動的に行うことができるため，他手法に比べて優位である．

実混合信号において BSSA の認識性能が大きく劣化する原因を明らかにするために，擬

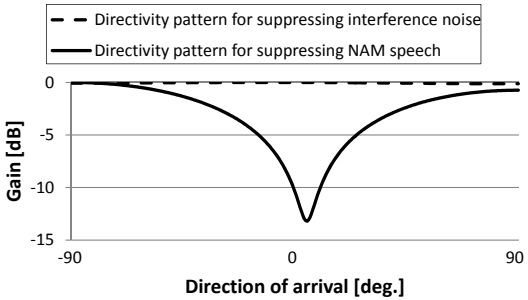

図 14 ICA の分離フィルタの指向特性（擬似混合信号）
Fig. 14 Directivity patterns given by unmixing matrix in *simurated mixed-signals*.

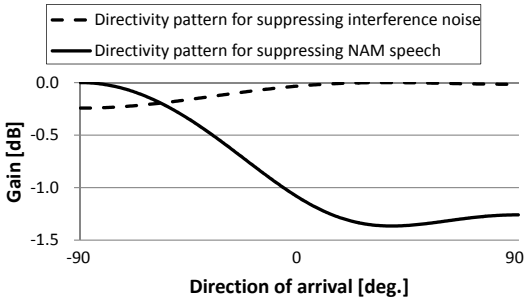

図 15 ICA の分離フィルタの指向特性（実混合信号）
Fig. 15 Directivity patterns given by unmixing matrix in *mixed-signals*.

似混合信号，実混合信号に対して，ICA で学習した分離フィルタの指向特性をそれぞれ図 14，15 に示す．擬似混合信号ではおよそ 0° 方向に深い谷が形成されるにもかかわらず，実混合信号ではそのように顕著な指向特性が得られていないことが分かる．このことから，式 (3) の NAM 信号の伝達関数 $\boldsymbol{a}(f)$ はユーザ動作によって変化し，時不変な線形フィルタでは抑圧できなくなると考えられる．結果，ICA における雑音推定精度は低くなるため，BSSA の認識性能は大幅に低下する．

5. おわりに

本報告では，NAM 収録中のユーザ動作により生じる非定常な雑音が NAM 認識に悪影響を及ぼすことを示し，その雑音を抑圧する方法として，ステレオ NAM 信号を用いたブラインド雑音抑圧法を提案した．独立成分分析および一般化スペクトル減算に基づく BSSA を用いて雑音抑圧を行い，さらに雑音の影響が小さいチャンネルを選択することにより，認識性能の向上が得られることを示した．一方で，ユーザ動作により NAM の伝達関数が変化するため，線形アレー処理では十分な雑音推定性能が得られず，BSSA においても高い雑音抑圧性能を得るのは困難であることも示した．今後の課題として，NAM と雑音の混合過程モデルの見直しと，それに適した雑音抑圧手法の提案が挙げられる．また，NAM の伝達関数の変動に対応した音響モデルの適用も検討する必要がある．

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) B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, and J.S. Brumberg. Silent speech interfaces. *Speech Communication*, Vol. 52, No. 4, pp. 270–287, 2010.
2) Y. Nakajima, H. Kashioka, N. Cambell, and K. Shikano. Non-Audible Murmur (NAM) recognition. *IEICE Trans. Information and Systems*, Vol. E89-D, No. 1, pp. 1–8, 2006.
3) T. Toda, K. Nakamura, T. Nagai, T. Kaino, Y. Nakajima, and K. Shikano. Technologies for processing body-conducted speech detected with non-audible murmur microphone. *Proc. INTERSPEECH*, pp. 632–635, Brighton, UK, Sep. 2009.
4) Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano. Blind spatial subtraction array for speech enhancement in noisy environment. *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 17, No. 4, pp. 650–664, 2009.
5) T. Hirahara, M. Otani, S. Shimizu, T. Toda, K. Nakamura, Y. Nakajima, K. Shikano. Silent-speech enhancement using body-conducted vocal-tract resonance signals. *Speech Communication*, Vol. 52, No. 4, pp. 301–313, Apr. 2010.
6) M.J.F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, Vol. 12, No. 2, pp. 75–98, 1998.
7) B.L. Sim, Y.C. Tong, J.S. Chang, and C.T. Tan. A parametric formulation of the generalized spectral subtraction method. *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 4, pp. 328–337, 1998.
8) H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano. Blind source separation combining independent component analysis and beamforming. *EURASIP Journal on Applied Signal Processing*, Vol. 2003, No. 11, pp. 1135–1146, 2003.
9) S. Ikeda and N. Murata. A method of ICA in time-frequency domain. *Proc. ICA*, pp. 365–370, Aussions, France, Jan. 1999.
10) S. Yamade, A. Lee, H. Saruwatari, and K. Shikano. Unsupervised speaker adaptation based on HMM sufficient statistics in various noisy environments. *Proc. INTERSPEECH*, pp. 1493–1496, Geneva, Switzerland, Sep. 2003.