社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－SLP－74（1）
2008/12/9

スパース性に基づくブラインド音源分離を用いた 2チャンネル入力音声認識

西亀 健太† 和泉 洋介† 渡部 晋治†† 西本 卓也† 小野 順貴†
嵯峨山茂樹†

† 東京大学情報理工学系研究科システム情報学専攻，〒113-8656 東京都文京区本郷 7-3-1
†† 日本電信電話 (株) NTT コミュニケーション科学基礎研究所，〒619-0237 京都府相楽郡精華町光台 2-4
E-mail: †{nishiki,izumi,nishi,onono,sagayama}@hil.t.u-tokyo.ac.jp, ††watanabe@cslab.kecl.ntt.co.jp

あらまし 本稿ではスパース性に基づくブラインド音源分離をフロントエンドに用いた雑音残響下における2チャンネル入力音声認識を提案する．2チャンネルブラインド音源分離により観測音からターゲット音声が分離される．本稿で用いた音源分離手法ではEMアルゴリズムによって設計された時間周波数マスキングを行うことにより残響などの拡散性雑音下でも精度よく音源分離を行うものである．音源分離後に残った歪みや，新たに生じた歪みに関してはCepstral Mean Nomalizationによる抑圧を行う．提案手法に対し，複数妨害音および残響の存在下における連続数字音声認識タスクにおいて提案手法の有効性を確認し，特に残響下で比較手法より高い認識性能を実現した．
キーワード スパース性，2チャンネルブラインド音源分離，残響，音声認識

Two-channel input speech recognition using sparsness-based blind source separation

Kenta NISHIKI†, Yousuke IZUMI†, Shinji WATANABE††, Takuya NISHIMOTO†, Nobutaka ONO†, and Shigeki SAGAYAMA†

† Department of Information Physics and Computing, University of Tokyo 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan.
†† NTT Communication Science Laboratories 2-4 Hikaridai Seika-cho, Soraku-gun, Kyoto, 619-0237, Japan.
E-mail: †{nishiki,izumi,nishi,onono,sagayama}@hil.t.u-tokyo.ac.jp, ††watanabe@cslab.kecl.ntt.co.jp

Abstract This paper discusses a two-channel input speech recognition using a sparsness-based blind source separation. The target speech is extracted from observed signals under diffusive noises (e.g. reverberation) by the source separation technique where a time-frequency mask is dynamically designed for speech separation using the EM algorithm. Cepstral Mean Normalization is exploited to reduce a remaining distortions or a newly introduced distortions in separated speech features. In a connected digit recognition task with multiple noise sources, the proposed method drastically improved the word accuracy in anechoic and reverberant environments. The proposed method achieved higher performance especially in a reverberant environment than conventional methods.
Key words sparsness, 2-channel blind source separation, reverberation, speech recognition

1. はじめに

議事録の自動作成システム，PDA・モバイルPCの音声インターフェース等の実環境における遠隔発話音声認識において，雑音や残響の影響により，音声認識性能は著しく低下する．

それに対し，従来は加法性雑音に対するSpectrum Subtraction (SS) [1]や，Cepstral Mean Normalization (CMN) [2]などフロントエンドでモノラル信号処理による音声強調を行うものが多かった．SSは入力された雑音重畳音声のスペクトルから雑音の平均スペクトルを減算することにより，音声のスペクトルを推定する方法である．CMNは，チャネル歪みなどの乗法性歪みがケプストラム領域で加法性になることを利用し，ケ

プストラムの平均を差し引くことで乗法性の歪みを抑圧し音声特徴量を正規化するものである．

上記の手法は簡便で効果が高いため広く用いられているが，非定常な雑音に対する処理が難しいため，近年多チャンネル信号処理を行う手法も提案されてきている [3] [4] [5]．Matassoni らによる Delay & Sum Beamformer を用いたもの [3]，阪本らによる Griffith-Jim 型適応アレーを用いたもの [4]，辻川らによる ICA に基づく周波数領域におけるブラインド音源分離 (Blind Source Separation, BSS) を用いたもの [5] などがある．ここで，実際の応用においては以下の状況が現実的には起こり得る．

- 話者の正確な方向情報は未知
- 音源数がマイクロフォン数より多く存在
- 妨害音だけでなく，背景雑音や残響が同時に存在

従って，話者の位置情報を必要とする Delay & Sum Beamformer やそのままではマイクロフォン数より多い音源数を扱うのが難しい ICA とは異なり，音源の方向情報を必要とせず，マイクロフォン数より多い音源数を扱えるような多チャンネル信号処理の枠組みが必要となる．特に，IC レコーダやラップトップ PC ではステレオ入力が標準的であり，2 チャンネルで動作する手法が現実的である．従って，Griffith-Jim 型適応アレーのようにマイクロフォン数が少ない場合に背景雑音や残響の抑圧能力が低くなってしまう手法は 2 チャンネルでの実現に適さない．

そこで，本稿では我々の研究室で和泉らによって提案されたスパース性に基づく 2 チャンネルブラインド音源分離 (2ch BSS) [6] をフロントエンドとして用いた音声認識手法を提案する．従来のスパース性に基づく音源分離は，各時間周波数成分を個々の音源に帰属させるクラスタリングの問題ととらえ，時間周波数マスクを設計することにより分離を行う手法が多かったが (e.g. [7])，残響や背景雑音の存在下においてはこうしたクラスタリングが難しくなることが大きな問題であった．和泉らによる 2ch BSS は，各時間周波数成分は各音源に確率的に帰属するというモデルに基づき音源分離を行うことで，残響や背景雑音の存在下でも高い分離性能を実現した．そのため，提案手法ではより実際に即した環境での音声認識が実現可能となる．本稿では 2ch BSS をフロントエンドとして用い，フロントエンド処理で残った歪みに対して CMN により音響的ミスマッチのさらなる解消を行う．提案手法に対し，雑音残響下における音声認識実験による検討を行い報告する．

2. スパース性に基づく 2ch BSS

本節でスパース性に基づく 2ch BSS [6] の概要を説明する．2 個のマイクロフォンにより観測された信号の時間周波数表現を $\boldsymbol{M}(\tau,\omega) = (M_L(\tau,\omega), M_R(\tau,\omega))^T$ と表す．ただし，τ はフレーム番号，ω は角周波数，T は転置，L，R の添字はそれぞれ左右のマイクロフォンで取得された信号であることを示す．目的は最尤推定によって観測信号のみから音源信号とその音源方向を得ることである．ここで，以下の 2 点，1) 音源信号は十分にスパースであり，各時間周波数成分においてアクティブな音源はただ 1 つ，および 2) それぞれの音源信号は平面波として到来，を仮定する．これらの仮定により，観測モデルは

$$\boldsymbol{M}(\tau,\omega) = S_k(\tau,\omega)\boldsymbol{b}_k(\omega) + \boldsymbol{N}(\tau,\omega), \tag{1}$$

と表される．ただし，$S_k(\tau,\omega)$ 時間周波数成分 (τ,ω) においてアクティブな音源のスペクトル，k は音源信号のインデックス，$\boldsymbol{b}_k(\omega) = (1, \exp(j\omega\delta_k))^T$ は音源からマイクロフォンへの伝達関数 (δ_k はマイクロフォン間の時間遅れ)，$\boldsymbol{N}(\tau,\omega) = (N_L(\tau,\omega), N_R(\tau,\omega))^T$ は残響や背景雑音を含めた観測誤差である (音源信号から独立であると仮定する)．本稿では簡単のために，曖昧さがないときには (τ,ω) や (ω) を省略し，S_k，$\boldsymbol{M}$，$\boldsymbol{N}$，$\boldsymbol{b}_k$ などと書くことがある．

従来の音源分離は，各時間周波数成分を個々の音源に帰属させるクラスタリングの問題ととらえ，1/0 の値を持つバイナリマスクを設計することが多かった．それに対し和泉らは，そもそも寄与する音源のインデックス k は未知の隠れ変数であり，各時間周波数成分は各音源に確率的に帰属するというモデルに基づき，最尤推定によって雑音パワー，音源方向および帰属率を統合的に推定する手法を提案した．これにより，バイナリマスクだけではなく帰属率に従う値をとる連続値マスクも設計できる．以下でその方法について説明する．

$\boldsymbol{N}$ が平均 $\boldsymbol{0}$，共分散行列 V のガウス分布に従うとする．ここで，あらゆる方向から確率的に等しく平面波が到来する拡散音場モデル [8] に基づいて雑音共分散行列 $V = E[\boldsymbol{N}\boldsymbol{N}^H]$ を

$$V = \sigma^2 \begin{pmatrix} 1 & \mathrm{sinc}(\omega D/c) \\ \mathrm{sinc}(\omega D/c) & 1 \end{pmatrix}, \tag{2}$$

のように書く．ただし，H はエルミート転置，σ^2 は雑音パワー，D はマイク間距離，c は音速を表す．時間差が δ_k となる方向から $k(\tau,\omega)$ 番目の音声信号 S_k が到来して $\boldsymbol{M}(\tau,\omega)$ が観測される尤度は

$$p(\boldsymbol{M}|\ \delta_k, \sigma^2, S_k) = \frac{1}{2\pi|V|^{1/2}} \times \exp\left(-\frac{1}{2}(\boldsymbol{M} - S_k\boldsymbol{b}_k)^H V^{-1}(\boldsymbol{M} - S_k\boldsymbol{b}_k)\right), \tag{3}$$

と与えられる．ただし $|V|$ は V の行列式を表す．このもとで，音源方向 δ_k，雑音パワー σ^2 および音源信号 S_k は下記の対数尤度

$$J = \sum_{(\tau,\omega)} \log p(\boldsymbol{M}(\tau,\omega)|\ \boldsymbol{\theta}), \tag{4}$$

を最大化するパラメータとして統合的に最尤推定することができる．ただし $\boldsymbol{\theta} = \{\boldsymbol{\delta}, \sigma^2, \boldsymbol{S}\}$ はパラメータセットであり，K を音源数として $\boldsymbol{S} = (S_1, \cdots, S_K)$，$\boldsymbol{\delta} = (\delta_1, \cdots, \delta_K)$ である．本稿では，音源数 K が既知であることおよび雑音共分散行列 V が各時間周波数において一定であることを仮定する．

最尤推定問題には時間周波数成分 (τ,ω) が帰属する音源のインデックス $k(\tau,\omega)$ を未知の隠れ変数とした EM アルゴリズムを適用する．このとき，$p(\boldsymbol{M}(\tau,\omega)|\ \boldsymbol{\theta})$ は，

$$p(\boldsymbol{M}(\tau,\omega)|\ \boldsymbol{\theta}) = \sum_k p(\boldsymbol{M}(\tau,\omega), k(\tau,\omega)|\ \boldsymbol{\theta}), \tag{5}$$

のように周辺化され，これより我々の問題における EM アルゴリズムの Q 関数は

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(i)}) = \sum_{\tau,\omega,k} m^{(i)}_{\tau,\omega,k} \log r_k^{(i)} p(\boldsymbol{M}|\ \delta_k^{(i)}, (\sigma^2)^{(i)}, S_k^{(i)}), \tag{6}$$

と導出される．ここで，r_k は k 番目の音源がアクティブになる事前分布に相当するパラメータであり，$\sum_k r_k = 1$ を満た

す．$\boldsymbol{\theta}^{(i)}$ は i ステップにおけるパラメータで，以下の式で更新される．

$$S_k^{(i)} = \frac{(\boldsymbol{b}_k^H)^{(i-1)}(V^{-1})^{(i-1)}\boldsymbol{M}}{(\boldsymbol{b}_k^H)^{(i-1)}(V^{-1})^{(i-1)}\boldsymbol{b}_k^{(i-1)}}, \quad (7)$$

$$\delta_k^{(i)} = \underset{\delta_k}{\operatorname{argmax}}\, Q(\delta_k; \delta_k^{(i-1)}), \quad (8)$$

$$\left(\sigma^2\right)^{(i)} = \frac{1}{2C}\sum_{\tau,\omega,k}\frac{m_{\tau,\omega,k}^{(i-1)}}{1-\operatorname{sinc}^2(\omega D/c)} \times\left(\boldsymbol{M}^H(V^{-1})^{(i-1)}\boldsymbol{M} - \frac{(\boldsymbol{b}_k^H)^{(i-1)}(V^{-1})^{(i-1)}\boldsymbol{M}}{(\boldsymbol{b}_k^H)^{(i-1)}(V^{-1})^{(i-1)}\boldsymbol{b}_k^{(i-1)}}\right), \quad (9)$$

$$r_k^{(i)} = \frac{\sum_{\tau,\omega} m_{\tau,\omega,k}^{(i-1)}}{\sum_{\tau,\omega,k'} m_{\tau,\omega,k'}^{(i-1)}}, \quad (10)$$

$$m_{\tau,\omega,k}^{(i)} = \frac{r_k^{(i)} p(\boldsymbol{M}(\tau,\omega)|\ \delta_k^{(i)},\ (\sigma^2)^{(i)}, S_k^{(i)})}{\sum_{k'} r_{k'}^{(i)} p(\boldsymbol{M}(\tau,\omega)|\ \delta_{k'}^{(i)},\ (\sigma^2)^{(i)}, S_{k'}^{(i)})}, \quad (11)$$

ただし，C は時間周波数 bin の総数である．式 (11) は EM アルゴリズムにおける E ステップに対応し，式 (7)，(8)，(9)，(10) は M ステップに対応する．式 (7) は未知数 S_k として最尤値を用いたものであり，2 チャンネルの信号から推定された音源位置に基づいてビームフォーミングを行っていることに相当する．また，式 (8) において $\operatorname{argmax}_{\delta_k} Q(\delta_k; \delta_k^{(i-1)})$ は解析的に得ることができないため，離散全探索によって求める．$m_{\tau,\omega,k}^{(i)}$ は分配関数と呼ばれ，ある時間周波数成分 $\boldsymbol{M}(\tau,\omega)$ において k 番目の音源がアクティブである確率を表す．

音源分離は時間周波数マスキングによって行われる．本稿で用いられる 2ch BSS では，各時間周波数成分がある音源に帰属する方向尤度を推定するものであり，その方向尤度に基づいたバイナリマスクおよび連続値マスクの 2 種類のマスク方式の設計が可能である．バイナリマスクは $m_{\tau,\omega,k}$ が最大の音源のみがアクティブであるとハードに決定し，アクティブでない音源は 0 と推定するものである．バイナリマスクにより，k 番目の音源信号のスペクトル $\hat{S}_k$ は下式で計算される．

$$\hat{S}_k = \begin{cases} \dfrac{\boldsymbol{b}_k^H V^{-1}\boldsymbol{M}}{\boldsymbol{b}_k^H V^{-1}\boldsymbol{b}_k} & (m_{\tau,\omega,k} > m_{\tau,\omega,k'}) \\ 0 & (\text{otherwise}) \end{cases}. \quad (12)$$

一方，連続値マスキングは二乗誤差最小化という意味での最適推定値である音源信号の期待値を求めるものである．連続値マスクによって $\hat{S}_k$ は

$$\hat{S}_k = m_{\tau,\omega,k}\frac{\boldsymbol{b}_k^H V^{-1}\boldsymbol{M}}{\boldsymbol{b}_k^H V^{-1}\boldsymbol{b}_k}, \quad (13)$$

のように計算され，分配関数 $m_{\tau,\omega,k}$ はそれ自身がマスクとして用いられる．音源分離の例を図 1 に示す．時間周波数マスキング処理によって残響下においても雑音が重畳した観測信号からターゲットの音声信号を分離できていることがわかる．

3. 分離音声に含まれる歪みの補償

従来の隠れマルコフモデル (Hidden Markov Model，HMM) に基づく音声認識では，あらかじめ用意したクリーンな音声の特徴量から音響モデルを学習し，入力音声の特徴量と音響モデルの確率的なマッチングを行う手法が一般的である．2ch BSS により観測信号から大幅に雑音を低減することができるが，雑音・残響の消し残りや過剰な抑圧により生じた音声の歪みが依然存在する．こうした歪みのため，音声特徴量と音響モデルの間に新たなミスマッチが生じるため，音声強調による音声品質の改善が音声認識性能の向上につながらないケースも報告されており [9]，音声認識との高い親和性を確保するためには，音響モデル適応法や特徴量の補正が必要になる(注1)．

2ch BSS による分離音声に含まれる歪みには，

a. ビームフォーミングによるもの
b. 時間周波数マスキングによるもの
c. 音源分離の誤りによるもの
d. 残響によるもの

が存在すると考えられる．a. は式 (7) に対応するものである．b. は式 (12) や式 (13) で音声のスペクトルに対し 0 ～ 1 の間の値を乗じていることに起因する．c. は時間周波数 bin を各音源に帰属させる確率の推定誤りに起因し，妨害音の消し残りやターゲット音の削除という形で現れる．こうした歪みへの対処法は Maximum Likelihood Linear Regression (MLLR) [10] などの音響モデル適応や Missing Feature Mask (MFM) を生成する手法 [11] などさまざまな方法がありうるが，本稿では簡易に実行可能であり計算やデータ収集におけるコストの小さい方法として CMN [2] を用いる．本来，CMN は定常な乗法性歪みを補償する手法であり，時間周波数マスキングによる歪みのような通常は非定常と考えられる乗法性歪みや，妨害音の消し残りなどの加法性の雑音に対して使われるものではないが，本稿では a. ～ d. に生じている可能性のある固有の定常乗法性歪みを取り除くために用いる．

4. 認識実験

4.1 実験条件

評価は CENSREC-4 [12] 準拠の数字 HMM に基づく日本語数字音声認識タスクにより行う．学習および認識は HTK(Ver3.4) を用いる．HMM は数字ごとに 11 モデル (いち，に，さん，よん，ご，ろく，なな，はち，きゅう，ぜろ，まる) と sil，sp の 13 モデルからなり，数字は 18 状態 (出力分布を持つのは 16 状態)，sil は 5 状態 (同じく 3 状態)，sp は 3 状態 (同じく 1 状態) である (始状態と終状態は出力分布を持たない)．学習データは CENSREC-4 のクリーン音声 8,440 発話 (男女 55 名) を用いて学習した．

CENSREC-4 に含まれる残響インパルス応答データはシングルマイクで収録したものであるので，2 チャンネルマイクロフォンを用いた本手法の評価系には適していない．従って，本実験では図 2 のようにマイクロフォンを配置し (音源数 3，ターゲット音源は S_1)，鏡像法によって球面波伝播および残響をシミュレートした．ターゲット音源は CENSREC-4 のクリーン音声 2,002 発話 (男女 52 名) を用いた．また，妨害音 S_2，S_3 はそれぞれ SMILE2004 データベース [13] の音源データから選択し，表 1 にある 2 種類の雑音環境を作成した．雑音条件 1 における 2 種類の雑音はともに「生活音」カテゴリから選択されたものである．雑音条件 2 における読み上げ音声は「音声」カテゴ

(注1)：例えば，文献 [9] では，音響モデルの分散パラメータ補正と適応学習を組み合わせてこれに対処している．

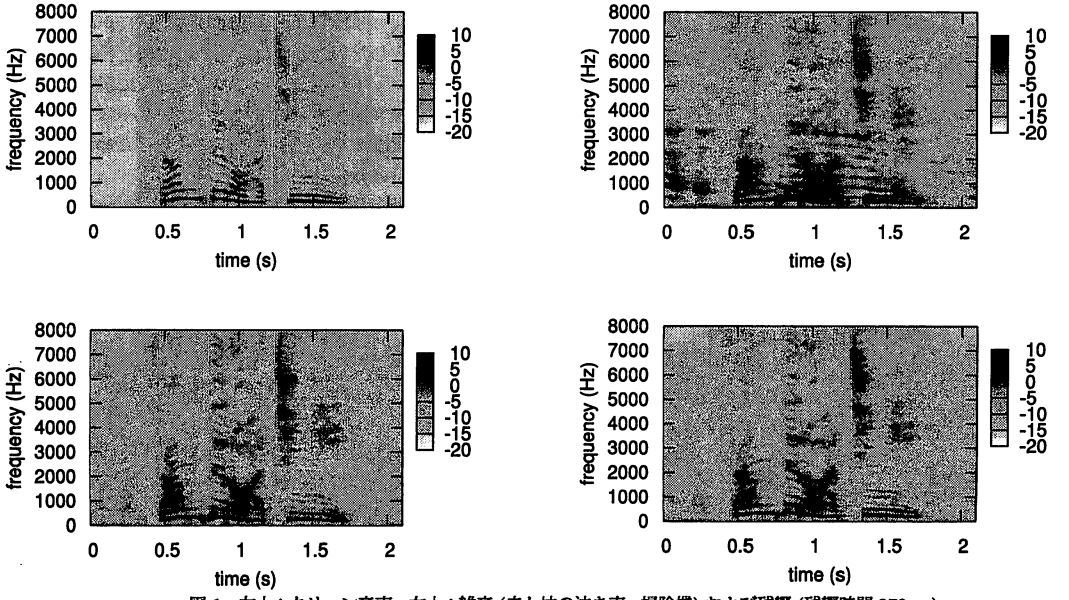

図 1　左上：クリーン音声，右上：雑音 (赤ん坊の泣き声・掃除機) および残響 (残響時間 270ms) が重畳した信号 (左チャンネル)，左下：バイナリマスクにより分離した信号，右下：連続値マスクにより分離した信号

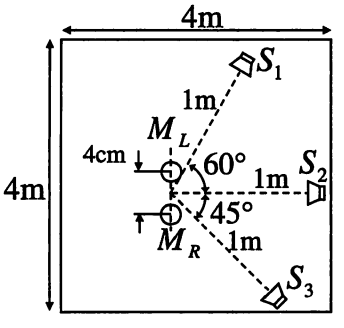

図 2　シミュレーションにおけるマイクロフォンと音源の位置関係．部屋の高さは 3m，マイクロホンおよび音源の高さは全て 1.5m.

表 1　使用した音源の条件

	S_1	S_2	S_3
雑音条件 1	ターゲット音声	赤ちゃんの泣き声	掃除機
雑音条件 2	ターゲット音声	読み上げ音声	突発性雑音

リから「日本語男声朗読」および「日本語女性朗読」を連結し作成した．突発性雑音は，「楽音」カテゴリから「拍手 (男声)」「拍手 (女性)」，「機械音」カテゴリから「自動ドア」，「生活音」カテゴリから「プリンタ印刷」，「固体音」カテゴリから，「ボイド 280 素面・椅子引摺り」，「音風景」カテゴリから「風鈴 (明珍火箸：鍛鉄)」からそれぞれ雑音を含む 5 秒程度を切出し，0.5 秒程度の空白を挟んで連結した．妨害音の長さを音声データに合わせるために，元の妨害音ファイルから発話毎に音声ファイルと同じ長さだけ切出しを行った．その際，切出しの始点はランダムに指定した．サンプリングレートを CENSREC-4 の音声データに揃えて 44.1kHz から 16kHz にダウンサンプリングした．音声認識のフレーム分析は，$1-0.97z^{-1}$ のプリエンファシス，フレーム長 25ms，フレームシフト 10ms，Hamming 窓によって行った．特徴量は以下の 2 種類を用いた．

（1）　MFCC (12 次元) + log power，およびその Δ，ΔΔ の 39 次元

（2）　上記特徴量から log power を除いた 38 次元

本稿では以下，(1) の特徴量を「39 次元特徴量」，(2) の特徴量を「38 次元特徴量」と呼ぶ．CENSREC-4 評価スクリプトで使用されているのは 39 次元特徴量であるが，2ch BSS においては，音源分離のミスによりパワーの大きな雑音が残りそれが悪影響を及ぼすケースも考えられるため，38 次元特徴量も用いた．この条件のもとで，クリーン音声の認識においては 39 次元特徴量を使用した際には 99.6%の単語正解精度 (Word Accuracy, WA(%)) が達成される．また，2ch BSS のフレーム分析は分析フレーム長 64ms，フレームシフト 32ms，Hamming 窓によって行った．音源分離によって分離された音源は 3 つ存在するが，そのどれがターゲットの音声であるかは既知とした．

4.2　分離性能

分離結果を表 2 に示す．SN 比は全ての雑音残響条件に関する平均値によって算出した．BMask および CMask は提案手法であり，BMask は式 (12) で示したバイナリマスクによる音源分離，CMask は式 (13) で示した連続値マスクによる音源分離を表している．DUET は文献 [7] に基づいた手法であり，1) パワーで重みづけした強度と時間差の 2 次元ヒストグラムを作成し，2) これに矩形関数を畳み込むことでスムージングを行い，3) そのピークをそれぞれの信号の強度・時間差として推定し，4) 次に各時間周波数成分に対する尤度が最大になる音源を通過させるマスクを作成し分離信号を得るものである．BMask と

表 2　各 2ch BSS 手法の分離性能.

	処理後の SN(dB)	SN 比 (dB) の改善値
DUET	5.87	11.34
BMask	8.13	13.60
CMask	8.29	13.76

CMask の SN 比の改善値は同程度であるが，わずかに CMask の方が分離性能が高く，どちらも DUET に比べ分離性能が高い．これは本稿で用いた音源分離手法の分離性能における優位性を示していると言える．

4.3　認識実験結果

2 チャンネル信号処理によるフロントエンドなし (NoMask)，提案手法で連続値マスクを用いたもの (CMask)，提案手法でバイナリマスクを用いたもの (BMask)，従来手法 (DUET) の単語正解精度 (WA(%)) の比較を各種の雑音残響環境で行った．雑音条件 1 と 2 については，残響がない場合 (残響時間 0ms) および残響時間 270ms の場合について評価を行い，雑音条件 1 についてはさらに残響時間が 468ms の場合についても評価を行った．図 3 より，NoMask，NoMask+CMN では著しく認識性能が劣化していることがわかる．しかし，DUET+CMN，BMask+CMN，CMask+CMN では NoMask+CMN に比べて認識性能が大きく向上している．音源分離により雑音が大きく抑圧され，さらに残った歪みや新たに生じた歪みが CMN によって軽減されているためと考えられる．この中で，BMask+CMN と DUET+CMN の性能は同程度であり，CMask+CMN の性能が最も高い．これより提案手法のうち CMask+CMN が，音源数がターゲット音声と妨害音を合わせマイクロフォン数以上であり，同時に残響が存在する環境における音声認識手法として有効であることが確認できた．ただし，SN 比を指標とした分離性能と認識性能の傾向は大きく異なることもわかった．表 2 では BMask と CMask が同程度で DUET に比べ高性能であるが，図 3 では DUET+CMN と BMask+CMN が同程度であり CMask+CMN が最も認識性能が高い．この理由については 4.5 節で考察する．

次に，残響時間の違いによる認識性能の違いについて検討する．表 3 より残響時間が長くなればなるほど認識性能が劣化する傾向にある．特に残響時間が最も長い 468ms になると，認識性能は最高でも 74.7% (CMask+CMN，38 次元特徴量) であり実用上十分な性能とは言えない．この原因は，残響時間が長くなることにより分離性能が低下し，音声に生じる歪みが大きくなっているためと考えられる．

最後に，特徴量の違いによる認識性能の違いについて検討する．図 3 の左右のグラフを比較すると，CMask+CMN では 38 次元特徴量を用いた方が 39 次元特徴量を用いるより認識性能が高く，提案手法においては特徴量から log power を除いた方が有効であるといえる．

4.4　マッチドモデル

本手法の上限性能を示すものとして，残響時間 270ms・雑音条件 1 の環境についてマッチドモデル との併用を行い，同じ残響雑音環境について CMask+CMN との比較を行った．マッチドモデルは残響時間 270ms・雑音条件 1 の環境で発声した音声を提案手法 (CMask) を用いて処理し，学習を行ったものである．学習データにはクリーン音響モデルのときと同じ 8,440 発話を用いた．また，特徴量は 38 次元特徴量を用いた．CMask のマッチドモデルとの併用における認識率 (WA(%)) は，98.7%であり，CMask+CMN の 89.9%に比べ高い．これより，提案法はマッチドモデルとの併用によりクリーン音声の認識と同程度の性能を実現できる．ただし，環境に応じたマッチドモデルの構築は大変コストがかかるため，データを用いた適応手法などによるフロントエンド処理後の音響的ミスマッチの更なる解消がのぞまれる．

表 4　時間周波数マスキングが音声認識に与える影響 (WA(%)). 特徴量は 38 次元特徴量.

BMask	29.8
BMask+CMN	78.9
CMask	42.1
CMask+CMN	98.4

4.5　時間周波数マスクが音声認識に与える影響

BMask と CMask の音声認識性能の差がどのように生じるかについて検討する．両者の違いは時間周波数マスクの値のみであり，3 節で扱った歪みのうち，b. の時間周波数マスクによる歪みのみが異なる．そこで，以下の方法で時間周波数マスキングが音声認識に与える影響を純粋に取り出すことを考える．音源分離で得た時間周波数マスクを観測音にではなく雑音・残響が重畳していない音声にかけ，時間周波数マスクの歪みのみが重畳した音声を得る．ただし，マスクの設計は雑音残響環境等に依存するため，3 節で扱った歪みのうち b. 以外の影響も間接的には受ける．そのようにして得た音声に対する認識実験を行い，結果を表 4 に示す．表 4 で例えば BMask となっているものは，音源からマイクロフォンへの遅延を計算したのち，その信号のスペクトルに音源分離の際に得たバイナリマスクをかけたものである．ただし，遅延を計算する際のマイクロフォンの位置は左右のマイクロフォンの中心の位置とした．認識率はマスキング方式ごとに，全ての雑音残響環境に関する平均により算出した．バイナリマスク，連続値マスク双方とも認識性能を大きく劣化させるが，CMN により劣化した認識性能が大幅に改善していることがわかる．特に，連続値マスクの方がバイナリマスクに比べ改善率が高く，このことからも音声認識との親和度は連続値マスクの方が高いと考えられる．バイナリマスクはターゲット音声に帰属しないと推定された時間周波数 bin を完全に 0 にするものであり，音声のスペクトルを削ることの音声認識への悪影響が大きく，そのため同程度の SN 比でも認識性能に差が生じると考えられる．本実験結果から，CMN を施すことにより 3 節で考察した歪みのうち，時間周波数マスクによる影響 (b.) は大幅に低減でき，特に連続値マスクのときにそれが顕著であることがわかった．これは時間周波数マスクによる歪みが定常な乗法性歪みで表現されることを示している．これらの現象の数理的な解明は今後の課題である．

5.　ま　と　め

今回，スパース性に基づくブラインド音源分離をフロントエンドとして用いた 2 チャンネル入力音声認識手法を提案した．CENSREC-4 データベースに準拠した連続数字認識タスクにおいて，マイクロホン数より多い妨害音および残響が存在する環境下における頑健性を確認した．特に，連続値マスクによる音源分離に CMN を組み合わせ，log power を除いた特徴量を用いた際に非常に高い性能を得た．

今後の課題は，実環境実験によって提案手法の有効性を確かめることが挙げられる．また，分離音源中のどれがターゲット

表 3　各環境での単語正解精度 (%)．左：39 次元特徴量，右：38 次元特徴量．時間は残響時間を示している．また，「雑音 1・雑音 2」とあるのは表 1 における「雑音条件 1・雑音条件 2」のことである．

	0ms		270ms		468ms	
	雑音 1	雑音 2	雑音 1	雑音 2	雑音 1	平均
NoMask	10.3	- 0.4	7.8	0.5	6.9	5.0
NoMask+CMN	25.6	- 1.7	12.1	- 2.6	9.5	8.6
DUET	36.3	41.1	22.6	15.6	8.8	24.9
DUET+CMN	**87.5**	**90.0**	61.5	53.6	20.6	62.6
BMask	31.1	36.7	17.4	22.8	13.5	24.3
BMask+CMN	75.8	74.6	43.0	56.8	25.6	55.2
CMask	46.7	40.1	37.6	27.9	27.1	35.9
CMask+CMN	86.3	70.3	**66.3**	**61.5**	**50.9**	**67.1**

	0ms		270ms		468ms	
	雑音 1	雑音 2	雑音 1	雑音 2	雑音 1	平均
NoMask	19.9	12.2	11.6	9.1	10.3	12.6
NoMask+CMN	45.1	34.5	26.3	25.0	18.7	29.9
DUET	23.5	27.7	19.9	14.4	12.9	19.6
DUET+CMN	63.7	70.3	57.9	46.3	30.7	53.8
BMask	23.8	34.5	14.4	19.2	11.5	20.7
BMask+CMN	62.2	85.6	37.4	56.9	24.1	53.2
CMask	39.6	40.8	32.1	29.4	23.7	33.1
CMask+CMN	**93.2**	**95.7**	**89.9**	**88.6**	**74.7**	**88.4**

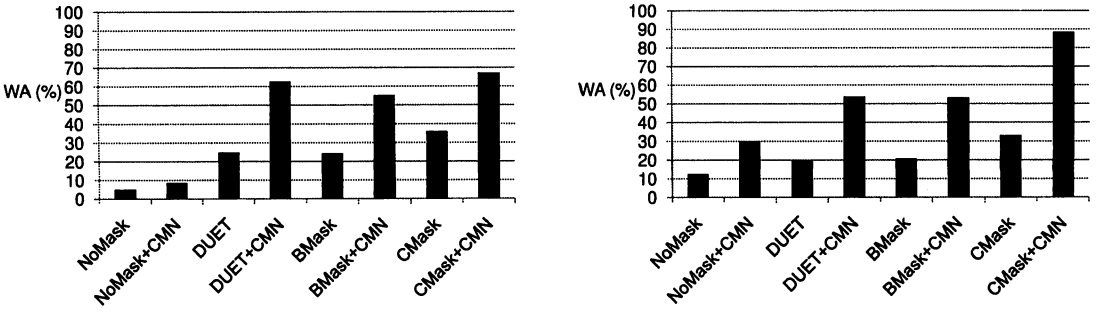

図 3　各手法の単語正解精度の全雑音残響環境についての平均 (%)．左：39 次元特徴量，右：38 次元特徴量．

であるかを推定することで，より多くのアプリケーションに適用可能な技術とすることも課題である．さらに，特に残響時間が長いとき，残響存在下でよりよい認識性能を達成する必要もある．そのため，missing feature 理論や音響モデル適応手法との併用を考えている．

謝辞　本研究は東京大学と NTT の共同研究として行われた．また，情報提供や有益な議論等に御協力いただいた NTT コミュニケーション科学基礎研究所の信号処理研究グループのメンバーおよび東京大学嵯峨山・小野研究室のメンバーに感謝する．本研究は CENSREC-4 の音声データおよび評価スクリプト，SMILE2004 データベースの音データを利用・改変して行われた．

文　献

[1] M. Berouti, Schwartz. R, Makhoul. J, "Enhancement of speech corrupted by acoustic noise," in *Proc. ICASSP*, vol. 4, Apr. 1979, pp. 208-211.

[2] F. H. Liu, R. M. Stern, X. Huang, A. Acero: "Efficient cepstral normalization for robust speech recognition, " Proc. ARPA Workshop on Human Language Technology, pp.69-74, Princeton, USA, March 1993.

[3] M. Matassoni, M. Omologo, D. Giuliani: "Hands-free speech recognition using a filtered clean corpus and incremental HMM adaptation, " Proc. ICASSP2000, pp.1407-1410, Apr., 2000.

[4] 阪本浩一，李晃伸，猿渡洋，鹿野清宏："Griffith-Jim 型適応アレーと環境適応法を統合したハンズフリー音声認識，" 日本音響学会 2003 年秋季発表講演論文集 3-6-6，2003.

[5] 辻川剛範，磯健一："ブラインド音源分離と 2 段スペクトル減算法によるハンズフリー音声認識，" 日本音響学会 2004 年秋季発表講演論文集 1-1-1，2004.

[6] Y. Izumi, N. Ono, S. Sagayama: "Sparseness-based 2ch BSS using the EM Algorithm in Reverberant Environment," Proc. WASPAA, pp.147-150, Oct., 2007.

[7] O. Yilmaz, S. Rickard: "Blind Separation of Speech Mixtures via Time-Frequency Masking," IEEE Transaction on Signal Processing, Vol. 52, No. 7, pp.1830-1847, 2004.

[8] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, M. C. Thompson: "Measurement of correlation coefficients in reverberant sound fields," JASA, Vol. 27, No. 6, pp.1072-1077, 1955.

[9] M. Delcroix, S. Watanabe, T. Nakatani: "Combined Static and Dynamic Variance Adaptation for Efficient Interconnection of Speech Enhancement Pre-Processor with Speech Recognizer, " Proc. of ICASSP, pp.4075-4076, Apr., 2008.

[10] M. J. F. Gales, P. C. Woodland: "Mean and variance adaptation within the MLLR framework, " Computer Speech and Language, vol. 10, pp. 249-264, 1996.

[11] 山本俊一，中臺一博，中野幹生，辻野広司，Jean-Marc Valin，駒谷和範，尾形哲也，奥乃博："音源分離との統合によるミッシングフィーチャマスク自動生成に基づく同時発話音声認識，" 日本ロボット学会誌，25(1)，pp.92-102，Jan.，2007.

[12] M. Nakayama, Masato Nakayama, Takanobu Nishiura, Yuki Denda, Norihide Kitaoka, Kazumasa Yamamoto, Takeshi Yamada, Satoru Tsuge, Chiyomi Miyajima, Masakiyo Fujimoto, Tetsuya Takiguchi, Satoshi Tamura, Tetsuji Ogawa, Shigeki Matsuda, Shingo Kuroiwa, Kazuya Takeda and Satoshi Nakamura, "CENSREC-4: Development of evaluation framework for distant-talking speech recognition under reverberant environments," *Proc. Interspeech*, Sep. 2008, pp. 968-971.

[13] K. Kawai, K. Fujimoto, T. Iwase, H. Yasuoka, T. Sakuma, Y. Hidaka , "Development of a sound source database for environmental/architectural acoustics: Introduction of SMILE 2004 (Sound Material in Living Environment 2004)," in *Proc. ICA*, 2004, pp. 1561-1564.