音声スパース性に基づく 2ch BSS を用いた雑音・残響下での音声認識*

◎西亀 健太，和泉 洋介，小野 順貴, 西本 卓也，嵯峨山 茂樹 (東大情報理工)，渡部 晋治 (NTT)

1 はじめに

近年，会議録の自動作成や音声インターフェースの利用等で，実環境における遠隔発話音声認識の必要性が高まっている．実環境では，雑音や残響の影響により，音響モデルと入力音声特徴量のミスマッチが生じ音声認識性能は著しく低下する．それに対し，従来のフロントエンドは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が，多チャンネル信号処理によって，より高い性能を実現することができる．しかし，実際の応用においては話者の正確な方向情報は未知であることが多く，また妨害音も複数存在することがあり得るため，音源の方向情報を必要とせず，また，マイクロフォン数以上の音源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 チャンネル信号処理の手法が現実的である．そこで，本稿では我々の研究室で和泉らが提案した音声スパース性に基づく 2 チャンネルのブラインド音源分離手法 (2ch BSS)[1] 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．2ch BSS[1] は，残響などの拡散性雑音存在下でも高い音源分離性能を持つため，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 や人工残響を用いたマルチコンディションモデル [2] 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，音声認識実験による検討を行い，報告する．

2 音声スパース性に基づく 2ch BSS

本章で音声スパース性に基づく 2ch BSS[1] 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 $(M_L(\tau,\omega), M_R(\tau,\omega))^T$ 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．

音声のスパース性から，各時間周波数成分において観測信号に寄与する音源が 1 個だけであると仮定すると観測モデルは，

$$\begin{pmatrix} M_L(\tau,\omega) \\ M_R(\tau,\omega) \end{pmatrix} = S_n(\tau,\omega) \begin{pmatrix} 1 \\ e^{-j\omega\delta_n} \end{pmatrix} + \begin{pmatrix} N_L(\tau,\omega) \\ N_R(\tau,\omega) \end{pmatrix} \quad (1)$$

と表される．$S_n(\tau,\omega)$ は n 番目の音源信号である．$(N_L(\tau,\omega), N_R(\tau,\omega))^T$ は残響や拡散性雑音を含む観測誤差であり，平均 0 のガウス分布に従うと仮定する．$(1, e^{-j\omega\delta_n})^T$ は音源 n の位置に対応するベクトルを表す．方向性雑音のスパース性も仮定すると，問題はターゲット音源が寄与する時間周波数平面内の領域を求めることである．従来は，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 n 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，最尤推定によって誤差分散，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．最尤問題は，EM アルゴリズムによって解いている．詳細は文献 [1] を参照されたい．

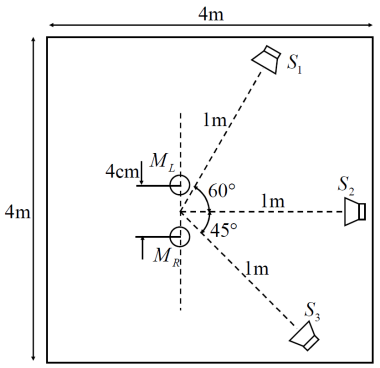

Fig. 1 シミュレーションにおけるマイクロフォンと音源の位置関係

3 認識実験

本実験で，提案手法の雑音残響下での頑健性を検証する．

3.1 実験条件

評価は CENSREC-4[3] 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．学習データは CENSREC-4 のクリーン音声 8440 発話 (男女 55 名) である．Fig. 1 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)，球面波伝播と残響を鏡像法によってシミュレートした．ここで，ターゲット音源の大まかな方向は既知であるとする．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，1,001 発話ずつ) を用いた．また，方向性雑

*Speech Recognition under Noisy Reverberant Environment Using Speech-Sparseness-Based 2-Channel Blind Source Separation by NISHIKI Kenta, IZUMI Yosuke, ONO Nobutaka, NISHIMOTO Takuya, SAGAYAMA Shigeki and WATANABE Shinji

音 S_2, S_3 はそれぞれ SMILE2004 データベース [4] から赤ん坊の泣き声と掃除機の音を選択した．サンプリングレートを CENSREC-4 の音声データに揃えて 96kHz 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，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 MFCC(12 次元+log power) およびその Δ，ΔΔ の 39 次元である．ただし，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．

3.2 実験結果

実験結果を Fig. 2，Fig. 3 に示す．2 チャンネル信号処理によるフロントエンドなし (no front-end)，提案手法でマスクに連続値マスクを用いたもの (proposed(CM))，提案手法でマスクにバイナリマスクを用いたもの (proposed(BM))，従来手法 (DUET[5])，音源信号を既知としたときに各時間周波数成分をその成分に対し最も寄与が大きい音源に帰属させるバイナリマスク (0dB mask) の単語正解精度 (WA(%)) の比較を，残響がない場合とある場合について行った．残響がない場合はクリーンモデルと CMN の比較を行った．Fig. 2 より，proposed(CM) はフロントエンドに何も用いていない場合に比べて認識率が 35.2%，proposed(CM) と CMN を併用すると 81.5%向上し CMN のみを用いた場合と比べても 61.7%向上している．CMN との併用において提案手法は特に有効であるが，それは音源分離によって生じた歪みが CMN によって軽減されているためと考えられる．ただし，残響がない場合は，従来手法や 0dB マスクに比べて proposed(CM) の方が認識性能が高いものの，ほぼ同等の性能である．Fig. 3 より残響のある場合は全ての手法において認識性能が下がり，proposed(CM) と CMN の併用でも 70.7%と十分な認識性能とは言えない．CMN はフレーム内に収まる乗法性歪みの抑圧を行うため，残響によるフレームをまたいだ歪みが十分に除去できなかったためと考えられる．そこで，残響による歪みに対処するため，人工残響インパルス応答を用いた残響マルチコンディションモデル (multi-rev)[2] を併用した．その場合の認識率は 81.0%であり proposed(CM) と CMN との併用に比べて認識率が 10.3%向上している．特に 0dB マスクと CMN の併用 (認識率 75.4%) に比べても認識性能が高く，各時間周波数成分に対し最も寄与が大きい音源が未知であっても高い認識性能を実現可能である．また，本手法の上限性能を示すものとして，マッチドモデル (matched) との併用も行った．matched は音源配置等の条件を全て観測環境に揃え，提案手法 (連続値マスク) を用いて処理した音声を用いて学習を行ったものである．proposed(CM) の matched との併用は認識率が 94.6%であり，proposed(CM) と multi-rev の併用に比べても認識率が 13.7%高く，フロントエンド処理後の音響的ミスマッチ解消には改善の余地がある．

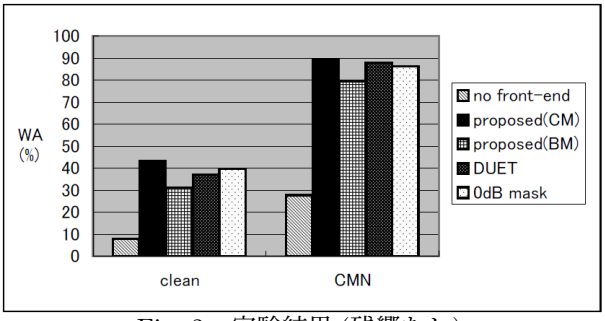

Fig. 2　実験結果 (残響なし)

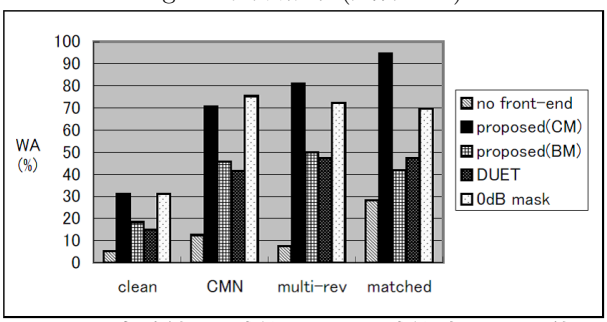

Fig. 3　実験結果 (残響があり，残響時間 0.27 秒)

また，提案手法のマスク方法の比較では，全ての状況において proposed(CM) の方が proposed(BM) に比べて認識性能が高く，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．

4　まとめ

今回，音声のスパース性に基づいた 2ch BSS をフロントエンドとして用いることで雑音および残響に対して頑健な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その有効性を確認した．今回は音源数を既知としたが，一般には音源数は未知である．その場合はある程度多い音源数を与えることでこの問題に対処できる．

今後の課題としては以下が挙げられる．まず，分離音源からターゲット音源を推定することでターゲット音源の位置情報が全くない状況下で動作させることである．また，マッチドモデルを用いずに残響存在下でよりよい認識性能を達成することである．それらにより，より実環境に即した音声認識が可能になると期待される．

謝辞

本研究は東京大学と NTT の共同研究の一部として行われた．

参考文献

[1] Izumi, *et al.*, Proc. WASPAA, pp.147-150, Oct., 2007.
[2] 西亀他，信学技報，vol. 108, No. 66, pp.43-48, 2008.
[3] 北岡他，情処研報，vol. 2007, No. 129, pp.7-12, 2007.
[4] DVD 版 建築と環境のサウンドライブラリ，日本建築学会編 (SMILE2004)
[5] Yilmaz, *et al.*, IEEE Trans. on Signal Processing, Vol. 52, No. 7, pp.1830-1847, 2004.