学術・技術論文

MFTを用いたロボットの動作中における音声認識

西村義隆[*1] 石塚満[*1] 中臺一博[*2]
中野幹生[*2] 辻野広司[*2]

Noise robust automatic speech recognition method for the robot with motor noise using missing feature theory

Yoshitaka Nishimura[*1], Mitsuru Ishizuka[*1], Kazuhiro Nakadai[*2], Mikio Nakano[*2] *and* Hiroshi Tsujino[*2]

Automatic speech recognition (ASR) is essential for human-humanoid communication. One of the main problems with ASR by a humanoid is that it is inevitably generates motor noises. These noises are easily captured by the humanoid's microphones because the noise sources are closer to the microphones than the target speech source. Thus, the signal-to-noise ratio (SNR) of input speech becomes quite low (sometimes less than 0 dB). However, it is possible to estimate these noises by using information on the humanoid's motions and gestures. This paper proposes a method to improve ASR for a humanoid with motor noises by utilizing its motion/gesture information. The method consists of noise suppression and missing-feature-theory-based ASR (MFT-ASR). The proposed noise suppression technique is based on spectral subtraction, and a white noise is added to blur distortion of suppression. MFT-ASR improves ASR by masking unreliable acoustic features in the input sound. The motion/gesture information is used for obtaining the unreliable acoustic features. Furthermore, we also evaluated with the acoustic model adaptation technique called MLLR (Maximum Likelihood Linear Regression). Un-supervised MLLR was used for the adaptation. We evaluated the proposed method through recognition of speech recorded by using Honda ASIMO in a room with reverberation. The noise data contained 34 kinds of noises: motor noises without motions, gesture noises, walking noises, and other kind of noises. The experimental results show that the proposed method outperforms the conventional multi-condition training technique.

Key Words: Noise robust automatic speech recognition , Missing feature theory, Spectral subtraction, MLLR, Multi condition training, Noise matching

1. はじめに

近年，さまざまなロボットが開発されている．その中でも特にヒューマノイドロボットは家事や介護など人間のパートナーとしての役割が期待されている．このためには人と音声を用いたコミュニケーションを行う能力が不可欠である．ロボットは身体をもっており，動作を行うことができる．動作は人と自然で円滑なコミュニケーションを実現するためには重要な要素である．例えば，ロボットの身体動作は対話、案内、プレゼンなどで有効であることが報告されている [1]〜[4]．しかし，ロボットは，多くのモータ，CPU, および，これらの冷却用にファンを搭載しており，定常時にはモータ音やファン音が，動作中には手足の動作に伴うモータ音が発せられる．これらの音は目的の音声と比較して，マイクに近いところから発せられ，場合によっては雑音が目的音声よりも大きくなり，SNR（Signal-to-Noise Ratio）が著しく低下する．さらに，移動やジェスチャに伴いマイクに混入するモータ音やファン音も変化するため，SNRは不規則に変化する．これらの理由により，ロボットの動作中における音声認識は難しい問題である．人・ロボットコミュニケーションの研究では，高雑音下での音声認識を避けるため，ロボット搭載マイクを用いず，接話マイクを用いることが多い [1]．しかし，常に接話マイクを用いることは利用者にとって煩わしく，ロボット自身のマイクで音声認識を行うことが必要である．

ロボット自身のマイクを用いて音声認識を試みた研究例も報告されている [5] [9]．これらの研究では定常雑音を仮定して音響モデルの雑音適応を行っているが，実際にはロボットの動作音は非定常であるため，こうした手法をそのまま適用することは難しい．しかし，ロボットの動作情報は取得可能であるため，あ

原稿受付
[*1]東京大学大学院情報理工学系研究科電子情報学専攻
[*2](株) ホンダ・リサーチ・インスティチュート・ジャパン
[*1]Graduate school of Information Science and Technology, The University of Tokyo
[*2]Honda Research Institute Japan Co., Ltd.

らかじめ収録した動作音データベースを用いて動作音の推定を行うことが可能である．本稿では，ロボットの動作情報を積極的に用い，動作音に頑健な音声認識手法を提案する．提案手法は，主に，SS (Spectral Subtraction) [11] 処理による入力信号の雑音除去，白色雑音重畳による雑音除去歪みの平坦化，MFT (Missing Feature Theory) [12] の導入による信頼度の低い音声特徴量への対応という 3 つの手法を用いて動作音に対する音声認識の頑健性向上を図る．また，教師なし MLLR (Maximum Likelihood Linear Regression) [7] を本手法に組合せた実験も行った．50 cm, 100 cm, 150 cm, 200 cm 離れた話者からの発話を想定した評価実験の結果，提案手法が有効であること，特に，距離が離れた場合に提案手法効果が大きいことを確認した．

2. ロボットの音声認識の頑健性向上に関する先行研究

音声認識の先行研究では，雑音への頑健性向上に対して数々の手法が提案されている．マルチコンディション学習による音響モデルを用いた音声認識は最も有効な手法の一つである．この手法は，あらかじめ雑音を含んだ音声を音響モデルの学習に用いるため，既知雑音に対しては効果が大きい．しかし，非定常雑音に対する効果的な学習が難しい，高雑音下では，音声の特徴が雑音に埋もれてしまうといった問題がある．

MLLR は，アフィン変換を用いて音響モデルを雑音に適応するアプローチであり，学習時とは異なる雑音環境や話者に音響モデルを適応させることができる．また，教師なし MLLR を用いれば，入力音声に対してオンライン適応を行い，認識性能を対話が進むにつれて向上できるようなシステムの実現も期待できる．実際，MLLR は有効な手法であり，話者適応などに一般的に用いられているが，マルチコンディション学習による手法と同様に動作音のような高雑音・非定常雑音下では期待したほどの効果が得られない．

このように従来の音声認識では，入力信号から雑音を除去するよりも，むしろ音響モデルを雑音に適応させることによって頑健性を向上させるアプローチが研究されてきた．これは，雑音除去を行うと音声に歪みが生じ，結果的に雑音適応を行う方が性能が高くなることが多いためである．しかし，ロボットにおける音声認識では，従来の音声認識で想定していた雑音よりも大きな雑音環境（SNR 0 dB 以下の場合もある）での認識が必要となる．このような環境では，音声の特徴を示す成分は雑音に埋もれてしまい，有効な情報はほとんど残らず，雑音適応による音声認識性能の向上は望めない．したがって，前処理によって雑音を除去するアプローチが必要となる．

実際に，ロボットにおける音声認識では，ビームフォーミング (BF:Beam Forming) [5]，独立成分分析 (ICA:Independent Component Analysis) [8]，幾何学的音源分離 (GSS:Geometric Source Separation) [9] といった主にマイクロホンアレーを用いた前処理を利用するアプローチが盛んに研究されている．BF は計算コストが低い音源分離手法であるが，分離精度はそれほど高くない．より精度の高い適応 BF [6] も提案されているが，計算コストが高くなるという欠点がある．ICA は音源の独立性を仮定するだけで分離を行うことができる反面，実環境ではこの仮定自体が成立しないことが多い，周波数帯域毎に音源信号が入れ替わるパーミュテーション問題が生じてしまうといったことから，期待通りの性能が得られないことが多い．BF と ICA の中間的な手法である GSS もしばしば用いられる．GSS は音源とマイクの位置関係を利用しながら，音源同士が無相関であることを仮定して音源分離を行う．パーミュテーション問題が発生しないといった利点はあるが，実環境では音源位置の正確な抽出が難しく，このため分離性能が劣化する．このようにマイクロホンアレーを用いた手法は，得手・不得手はあるものの，動作音に加え，環境雑音，他の話者からの音声雑音なども扱うことができるため，ロボットでは有効な手法である．しかし，本稿ではロボット自身が発する動作音を対象としており，この動作音は，動作情報から推定可能である．このため，マイクロホンアレーを用いず，単一マイクでも動作音への対応が可能である．

単一マイクを用いて動作音を扱うアプローチとして，伊藤らによる SS 処理を用いた手法が挙げられる [10]．SS 処理は，非発話区間を用いて定常雑音推定を行い，スペクトル領域において推定雑音成分を減算し，音声強調を行う．彼らは，関節角度や位置ごとに雑音の学習を行ったニューラルネットワークを用いて SS 処理の減算に用いる雑音信号の推定を行い，シミュレーション上での認識性能を報告している．しかし，反響のある環境における有効性，マルチコンディション学習による音響モデルを用いた手法に対する優位性には言及していない．一般に，SS 処理は非定常雑音に対しては歪みが大きくなり，性能が劣化するため，実環境で有効とは必ずしもいえない．

非定常雑音に対しても有効な手法として，MFT を用いた手法が挙げられる．MFT は音声信号のうち，雑音や歪みのため信頼度が低くなった特徴量をマスクし，信頼度の高い特徴量のみを用いて音声認識を行うアプローチである．MFT で用いるマスクには，0, 1 のバイナリマスク，もしくは連続値をとるソフトマスクが使われる．狭義には，バイナリマスクを用いる場合を MFT，ソフトマスクを用いる場合をマルチバンド音声認識 [13] [14] と区別するが，本稿では，広義の意味で捉え，共に MFT として扱うものとする．MFT は，信頼度の推定を正確に行うことができれば，他の雑音適応手法と比較して認識性能が大きく向上するが，正確な信頼度推定は難しく，MFT を用いた手法における大きな課題となっている．このため，従来の音声認識では，MFT が用いられることが少なかった．しかし，本稿で対象とするロボットの動作音は動作情報から推定可能であるため，MFT の有効的な利用が可能である．

3. 動作音に対応した音声認識手法

提案手法では，まず，入力信号に対し，SS 処理を適用し，主に定常雑音の除去を行う．次に，SS 処理によって生じる歪みが音声認識に悪影響を与えるため，SS 処理後の音声に白色雑音を重畳し平坦化を行い歪みを抑える．さらに，非定常な動作音成分に対応するため，MFT を利用した音声認識を行う．この際，MFT のマスク生成には推定動作雑音を用い，雑音の多く重畳した箇所は信頼度を低くし，音声認識への関与を低くする．図 1 に提案手法のブロック図を示す．以下，それぞれの処理について示す．

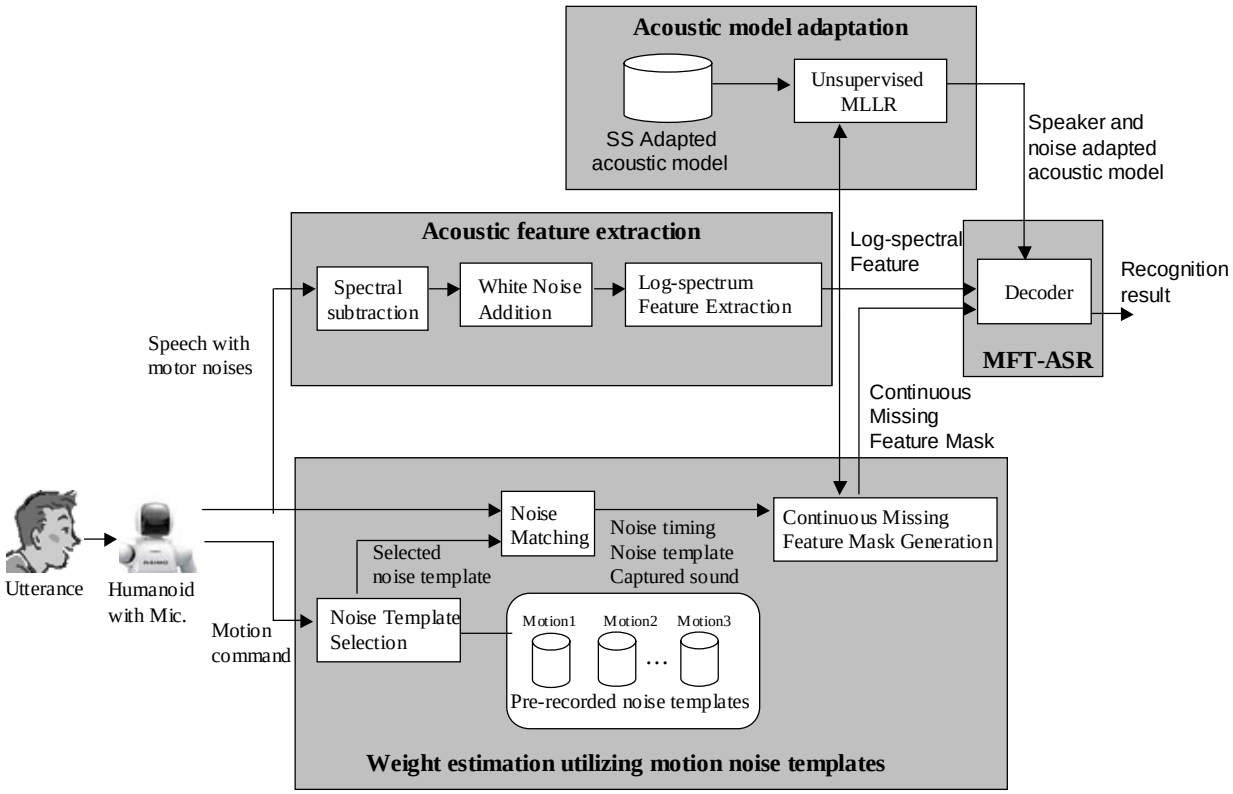


Fig. 1 *Block diagram of the proposed method*

3.1 モータ音の正常雑音除去処理

入力信号の SNR は低い (0dB 以下である場合もある) ため，このような環境で音声認識に有効な音声特徴量を抽出することは難しい．そこで，入力信号の SNR を改善するため雑音除去を行う．雑音除去には式 (1) に示される SS 処理を用いた．

$$|X(f)| = \max\{|X(f)| - \sqrt{\alpha}|\bar{N}|, \sqrt{\beta}|\bar{N}|\} \qquad (1)$$

$X(f)$ は入力信号のスペクトルを示し，$\bar{N}$ は入力信号に重畳している雑音信号の平均スペクトルを示す．α, β は SS 処理を行う際のパラメータであるが，本稿では一般的によく用いられている値 $(\alpha = 1, \beta = 0.1)$ を用いた．

3.2 白色雑音重畳

雑音除去により SNR は向上するが，同時に認識性能に悪影響を及ぼすスペクトル歪みが生じる．特に，ロボットの動作音はパワーが大きく，歪みも大きくなる傾向がある．本稿では，このスペクトル歪みの影響を軽減するため，雑音除去処理の後に薄く白色雑音を重畳させた．同様の方法は，山出 [15] らによって報告されており，定常雑音を加えることで，歪みを平坦化し，認識性能向上が期待できる．白色雑音は，以下の式を用いて重畳した．

$$y'(t) = y(t) + \frac{2p}{T}\sum_{t=1}^{T}|y(t)| \cdot \mathrm{random(t)} \qquad (2)$$

$y(t)$ は雑音除去処理後の信号であり，random(t) は -1 から 1 までの任意の実数値をランダムに返す関数である．本稿では $p = 0.1$ とした．すなわち，平均して入力信号の 1 割程度の大きさの白色雑音を重畳した．

3.3 音響モデルの雑音除去処理への適応

ロボットの音声認識では，定常雑音が重畳した音声データを用いてマルチコンディション学習を行った音響モデルを用いる手法が有効である．ロボットは定常時でもモータ音やファン音を発するため，この雑音を含めて学習することでクリーン音声データのみで音響モデルを学習する場合と比べ，認識性能が向上する．しかし，雑音除去処理を行ったデータを用いてマルチコンディション学習を行った場合，雑音除去処理によって生じるスペクトル歪みの影響で，認識性能が期待するほど向上しないことがある．本稿ではこの問題を解決するため，雑音除去処理後に白色雑音を重畳した音声データを用いて音響モデルの学習を行った．これにより，雑音除去処理による認識性能の低下を抑えることができる．

3.4 対数スペクトル特徴量の抽出

白色雑音を重畳した音声に対して音声特徴量を抽出する．音声認識の特徴量として，一般には，音声スペクトルを DCT (Discrete Cosine Transform) して得られるケプストラム領域の特徴量 MFCC (Mel Frequency Cepstrum Coefficients) が用いられる．しかし，本稿では MFCC ではなく，対数スペクトル特徴量 [16] [17] を用いた．動作音などの加法性雑音は，特定のスペクトル領域にパワーが集中していることが多い．このような雑音が重畳した音声に対し，MFCC を特徴量として使用すると全特徴量に雑音の影響が広がってしまい，すべての特徴量の信頼度が低くなってしまう．したがって，MFT では，ケプストラム領域の音声特徴量よりもスペクトル領域の音声特徴量

の方が適している．なお，MFCC ではケプストラム領域に変換した後，C_0 項の除去，リフタリング，CMS (Cepstrum Mean Subtraction) [18] の 3 つの正規化処理が行われる．これらの正規化処理は音声認識性能を向上させる上で重要であることが知られているため，本稿で使用した対数スペクトル特徴量においても，対数スペクトル領域において同様の正規化処理を施している．

3.5 MFT マスクの生成

MFT マスクはフレーム，および周波数帯域ごと (音声特徴量の次元ごと) に生成される．先見的な情報を利用しないマスクの生成は Raj らの報告 [19] や山本らの報告 [9] がある．しかし，完全に理想的なマスクを生成することは現実的には困難である．本稿では，ロボット自身の動作情報は動作前に取得できるため，これに基づいて動作音の推定を行う．動作情報に基づき，あらかじめ収録したテンプレート雑音から動作音を取得し，入力信号と時間方向にマッチングを行い，入力との同期をとる．最終的に，入力信号と同期を取った動作音を比較し，マスク生成を行う．詳細を以下に示す．

3.6 テンプレート雑音の選択

あらかじめ 34 種類の動作音を収録しテンプレート雑音としてデータベース化した．動作中に音声認識を行う際は，データベースから動作種類に応じたテンプレート雑音を選択する．発生した動作音は，選択されたテンプレート雑音と同じであると仮定し，テンプレート雑音を用いた時間方向の雑音マッチングに基づく雑音推定を行う．

3.7 雑音マッチング

テンプレート雑音が選択された時点では，実際の動作音とテンプレート雑音の時間的な関係は不明である．そこで，動作音とテンプレート雑音の時間方向のマッチングを行う．$T(\omega,t)$, $Y(\omega,t)$ をフレーム t，周波数 ω におけるテンプレート雑音，および入力信号ののパワースペクトルとする．また，テンプレート雑音における各周波数のパワースペクトルの最大値を $M(\omega)$ とする．

ここで，入力信号 $Y(\omega,t)$ について，$M(\omega)$ を超えるものは音声信号が重畳しており，ミスマッチの要因となるとして，パワースペクトル値を 0 として，マッチング用に $Y'(\omega,t)$ を求める．

$$Y'(\omega,t) = \begin{cases} Y(\omega,t) & if \quad Y(\omega,t) \leq M(\omega), \\ 0 & otherwise. \end{cases} \tag{3}$$

$Y'(\omega,t)$ と $T(\omega,t)$ の相互相関をとることによりマッチングを行った．最も相関が高いフレーム $s(\omega)$ は

$$s(\omega) = \underset{\tau}{\operatorname{argmax}} \sum_{f=0}^{N-1} Y'(\omega,t)T(\omega,t-\tau) \tag{4}$$

である．

このとき，$Y(\omega,t)$ と $T(\omega,t)$ の時間差は，$s(\omega)$ を最大にする周波数を ω_{max} とした場合，$s_{match} = s(\omega_{max})$ で与えられるものとした．よって，推定された雑音 $N(\omega,t)$ は，

$$N(\omega,t) = T(\omega,t-s_{match}) \tag{5}$$

と表すことができる．

3.8 マスクの生成

まず，推定雑音 $N(\omega,t)$ を対数スペクトルに変換する．変換した対数スペクトル雑音を $n(k,t)$ とする．k は特徴量の次元 (対数周波数軸方向) を示す．同様に，入力信号 $Y(\omega,t)$ の対数スペクトルを $y(k,t)$，入力信号に対し，雑音除去，白色雑音重畳を行った信号の対数スペクトルを $p(k,t)$ とする．この場合，音声信号は以下のように推定できる．

$$c'(k,t) = y(k,t) - n(k,t) \tag{6}$$

また，マスク $m(k,t)$ は以下のように計算する．

$$m(k,t) = \begin{cases} m'(k,t) & if \quad m'(k,t) < M_{th}, \\ M_{th} & otherwise. \end{cases} \tag{7}$$

$$m'(k,t) = \frac{|C'(k,t) - \mathrm{median}_k(C'(k,t))|}{P(k,t) - C'(k,t)} \tag{8}$$

$\mathrm{median}_k(a(k))$ は $a(k)$ の k についての中央値を得る関数である．$P(k,t)$ および $C'(k,t)$ は対数スペクトル $p(k,t)$ および $c'(k,t)$ に 3.4 節で述べた正規化処理を施したものである．

式 (8) の右辺の分母は，入力信号とクリーン音声の対応する特徴量同士が近い値であるほど，その特徴の信頼度は高いという考えを反映したものである．また，予備実験により，MFT を用いた音声認識では，音声特徴量として対数スペクトルを用いた場合，スペクトルの山と谷にあたる特徴量が，他の部分よりも認識精度に貢献するという知見が得られている．式 (8) の右辺の分子では，この知見を反映して，対数スペクトル特徴量の山と谷の部分に重みをかける処理を行っている．

式 (7) は，マスク値が必要以上に大きな値になることを防ぐ閾値処理を行い，マスクの値域は 0 から M_{th} に制限している．なお，閾値 M_{th} は，実験的に 5.0 とした．

次に，各フレームごとに，マスク値の合計が次元数 K になるように正規化を行う．この正規化は，MFT を用いた音声認識を行う際に，挿入ペナルティなどの最適パラメータが変化することを抑えるために行う．最終的に得られる $w(k,t)$ を MFT マスクとして，音声認識で用いる．

$$w(k,t) = \frac{m(k,t)}{\sum_{k=1}^{K} m(k,t)} \tag{9}$$

3.9 MFT に基づく音響尤度の計算

MFT を用いた音声認識では，音響尤度の計算以外の処理は，一般的な MFT を用いない音声認識と同様である．一般的な音声認識の処理は様々な文献で詳細に解説されているため，音響尤度の計算方法の違いのみを述べる．MFT を用いない一般的な音声認識では，入力音声特徴量 $s(i,t)$ の音素モデル q_l に対する尤度は以下の式によって与えられる．

$$L(s(i,t)|q_l) = \sum_{i=1}^{M} L(s(i,t)|q_l) \tag{10}$$

一方，MFT では，信頼性の高い特徴成分に対しては大きな重みを，信頼性の低い特徴成分に対しては小さな重みを用いて尤

度の計算を行う．このため，入力には，入力音声特徴量に対応するマスク情報も必要となる．式 (9) で得られるマスク値を用いれば，入力音声特徴量 $s(i,t)$ の音素モデル q_l に対する尤度は以下のように定義できる．

$$L(s(i,t)|q_l) = \sum_{i=1}^{M} w(i,t) L(s(i,t)|q_l) \qquad (11)$$

4. 実 験 条 件

Honda ASIMO を用いて評価実験を行った．ロボット頭部の左右に設置されたマイクのうち，左マイクを用いて音声の収録を行い，孤立単語認識による評価を行った．評価用データには ATR 音素バランス単語を用いた．音素バランス単語には男性 12 話者，女性 13 話者の合計 25 話者の音声データが含まれ，1 話者あたりの発話数は 216 である．各発話は「いきおい」,「いよいよ」などの単語発声である．

音響モデルの構築には，このうち男性 9 話者女性 10 話者の合計 19 話者の音声データ (学習セット A_1) を用いた．このデータは無響室で 100 cm の距離から収録を行い，音圧変化に柔軟に対応できるように，5 dB, 10 dB, 15 dB の 3 種類の SNR で学習を行った．

テスト用のデータは学習データとは異なる男性 3 話者女性 3 話者の合計 6 話者の音声データ (テストセット R_1) を用いた．収録には，7 m (W) × 4 m (D) × 3 m (H)，残響時間が約 0.2 s の部屋を用いた．話者とロボットのマイクの距離は 50 cm, 100 cm, 150 cm, 200 cm の 4 種類について調べた．

ロボットの動作雑音については，34 種類の動作音を用いて認識実験を行った．この動作音は ASIMO の電源を投入した状態で，動作を行っていない定常雑音（1 種類）,「バイバイ」など手の動きを主とするジェスチャ雑音 (15 種類),「うなずき」など頭部の動きを主とするジェスチャ雑音 (5 種類),「お辞儀」など手・頭部の同時動作を主とするジェスチャ雑音（5 種類），および「直進」や「回転」など足を用いた動きを主とする歩行雑音（8 種類）の 5 つの雑音グループから構成される．テストセット R_1 に動作音を重畳したものをテストセット R_2 とする．

提案手法とマルチコンディション学習を用いた従来手法の比較を行うため，マルチコンディション学習用のデータを用意した．マルチコンディション学習は A_1 のデータに加え，ASIMO の電源を投入したときのモータ音やファン音などの定常雑音を重畳したデータ A_2，動作雑音 (動作 $1 \leq N \leq 34$) を重畳したデータ $A_{3(N)}$（A_2 を含む）を用い，以下の音響モデルを用意した．

AM-1 学習セット A_1 を用いたモデル (クリーンモデル)
AM-2 学習セット A_1 と A_2 を用いたモデル
AM-3 学習セット A_1 と $A_{3(N)}$ を用いたモデル
AM-4 学習セット A_1 と A_2 に雑音除去処理を施した A_4 を用いたモデル
AM-5 学習セット A_1 と A_4 に白色雑音を重畳した音声データを用いたモデル

音響モデル **AM-3** は雑音ごとに作成しているため，全部で 34 種類のモデルを作成した．また，音響モデル **AM-5** は白色雑音重畳率 (式 (2) の p) を $0.05, 0.1, 0.2, 0.4$ と変化させて 4 種類のモデル **AM-5a** – **AM-5d** を作成した．

評価実験において比較した方法を表 1 にまとめる．まず，ベースラインの性能を調べるため，テストセット R_2 に対して，前処理を行わずに，3 種類の条件 (**A** – **C**) で認識実験を行った．**A** では，一般的に用いられるクリーンモデル **AM-1** を用い，**B** および **C** では，雑音に頑健な手法として一般的に用いられるマルチコンディション学習による音響モデルを使用した．**B** では定常雑音のみを重畳した音声データを用いて，また，**C** ではロボットの動作音，すなわち非定常雑音も重畳した音声データで音響モデル学習を行った．

次に，白色雑音重畳の効果を検証するため，条件 **D** – **I** について，MFT の効果を検証するため，条件 **J** – **L** について実験を行った．また，MLLR との組合せの効果を検証するために，条件 **C'**，**J'** についても実験を行った．

4.1 白色雑音重畳の効果検証

D は，テストセット R_2 に対して，モータ音に含まれる定常雑音の除去を目的として SS 処理を行った音声データを認識した．音響モデルには，**AM-2** を用いた．**E** はテストセットに **D** と同じ処理を施して，認識を行ったが，SS 処理後の音声データを用いて学習した音響モデル (**AM-4**) を用いた．**F** – **I** はテストセットに SS 処理を行った後，表 1 に示すように白色雑音重畳率 p を変化させて白色雑音を重畳した音声データの認識を行った．音響モデルは，p の値が認識用のデータと対応するよう，それぞれ，**F** – **I** に対して **AM-5a** – **AM-5d** を用いた．

4.2 MFT の効果検証

J – **L** は，**G** と同様の処理をテストセットに施した後，MFT による音声認識行う実験条件である．**J** –**L** はマスク計算時の雑音マッチングの条件が異なる．**J** は，実環境を想定した提案手法にあたる条件である．雑音マッチングの入力信号を生成する際に，テンプレート雑音とは別に録音した雑音を音声に重畳した．つまり，入力信号に含まれる雑音の波形と選択されたテンプレート雑音の波形は完全には一致しない．さらに，雑音区間検出が完全にはできないことを想定し，マスク生成の際には，テンプレート雑音を s_{match} に対して 0 ms から 200 ms の範囲でランダムにシフトさせた．**K** は，雑音マッチングの入力信号として，**J** と同様の雑音信号を用いたが，雑音部分が完全に検出できたことを仮定し，音声に重畳させず雑音信号のみでテンプレート雑音と入力信号の雑音のマッチングを行った．このため，**J** よりも理想的で雑音マッチングの容易な条件である．**L** は，雑音波形，および雑音検出が完全にできることを仮定し，雑音テンプレートに含まれる雑音信号を雑音マッチングに用いた．これは，提案手法の認識性能上限を調べるために用意した最も理想的な条件である．

いずれの場合も，白色雑音重畳率は 0.1 とした．白色雑音重畳率の最適値は，距離や動作によって異なるため，0.1 は中間的な値となっている．また，この p 値に対応して，音響モデルは **AM-5b** を使用した．

4.3 MLLR の効果検証

C'，**J'** は，それぞれ **C**，**J** に対して教師なし MLLR を行った音響モデルを用いる条件である．この実験により，マルチコ

Table 1 *Experimental Conditions*

Condition	**A**	**B**	**C**	**D**	**E**	**F**	**G**
Noise Suppression (SS)				✓	✓	✓	✓
White Noise Addition						$p = 0.05$	$p = 0.1$
MFT (voice+noise matching)							
MFT (only noise matching)							
MFT (known noise)							
Unsupervised MLLR							
Acoustic Model	**AM-1**	**AM-2**	**AM-3**	**AM-2**	**AM-4**	**AM-5a**	**AM-5b**

Condition	**H**	**I**	**J**	**K**	**L**	**C'**	**J'**
Noise Suppression (SS)	✓	✓	✓	✓	✓		✓
White Noise Addition	$p = 0.2$	$p = 0.4$	$p = 0.1$	$p = 0.1$	$p = 0.1$		$p = 0.1$
MFT (voice+noise matching)			✓				✓
MFT (only noise matching)				✓			
MFT (known noise)					✓		
Unsupervised MLLR						✓	✓
Acoustic Model	**AM-5c**	**AM-5d**	**AM-5b**	**AM-5b**	**AM-5b**	**AM-2**	**AM-5b**

ンディション学習，マルチコンディション学習と MLLR の組合せ，提案手法と MLLR を組合せについての比較を行う．

教師なし MLLR 適応は，テストデータセットに対して，**C'**，および **J'** と同様の処理を行った認識用データを一旦，音声認識し，その結果を正解ラベルとして，教師あり MLLR を行うことで実現した．

実験では，話者ごと，雑音の種類ごとに教師なし MLLR 適応を行い，同じ話者，同じ種類の雑音のデータに対して認識を行った結果を集計した．

5. 実 験 結 果

5.1 白色雑音重畳の効果

34 種類の動作音について，雑音グループごとに結果を平均した実験結果を Fig.2a) – d) に示す．**D** は SS 処理を施し，雑音を含んだ音声を用いてマルチコンディション学習を行った音響モデルを用いた音声認識結果である．SS 処理を用いることで，マルチコンディション学習で適応した雑音が入力信号から除去されてしまうため，認識性能が低い．これに対し，**E** は SS 処理を施した音声データを用いて，音響モデルの学習を行っているため，**D** に比べて認識性能が高い．この結果から，SS 処理と雑音を重畳した音声を用いてマルチコンディション学習を行った音響モデルの組合せでは，性能が低く，雑音除去効果が薄れてしまうが，雑音除去後の音声データを用いて音響モデルの学習を行うことで，高い雑音除去効果が得られることが確認できる．

F から **I** は SS 処理により生じた歪みを軽減するため，白色雑音重畳を行った．音響モデルは，雑音除去処理後，白色雑音重畳を行った音声データを用いて学習を行い，認識時にも同様の処理を入力信号に施している．**E** よりも，**F** から **I** に性能が高い結果が多く含まれており，白色雑音重畳により，認識性能が向上することが確認できる．ただし，**F** から **I** の中では認識性能が最もよい条件を一意に定めることはできないため，認識性能を最大にする白色雑音重畳率は雑音環境によって異なることがわかる．

5.2 MFT の効果

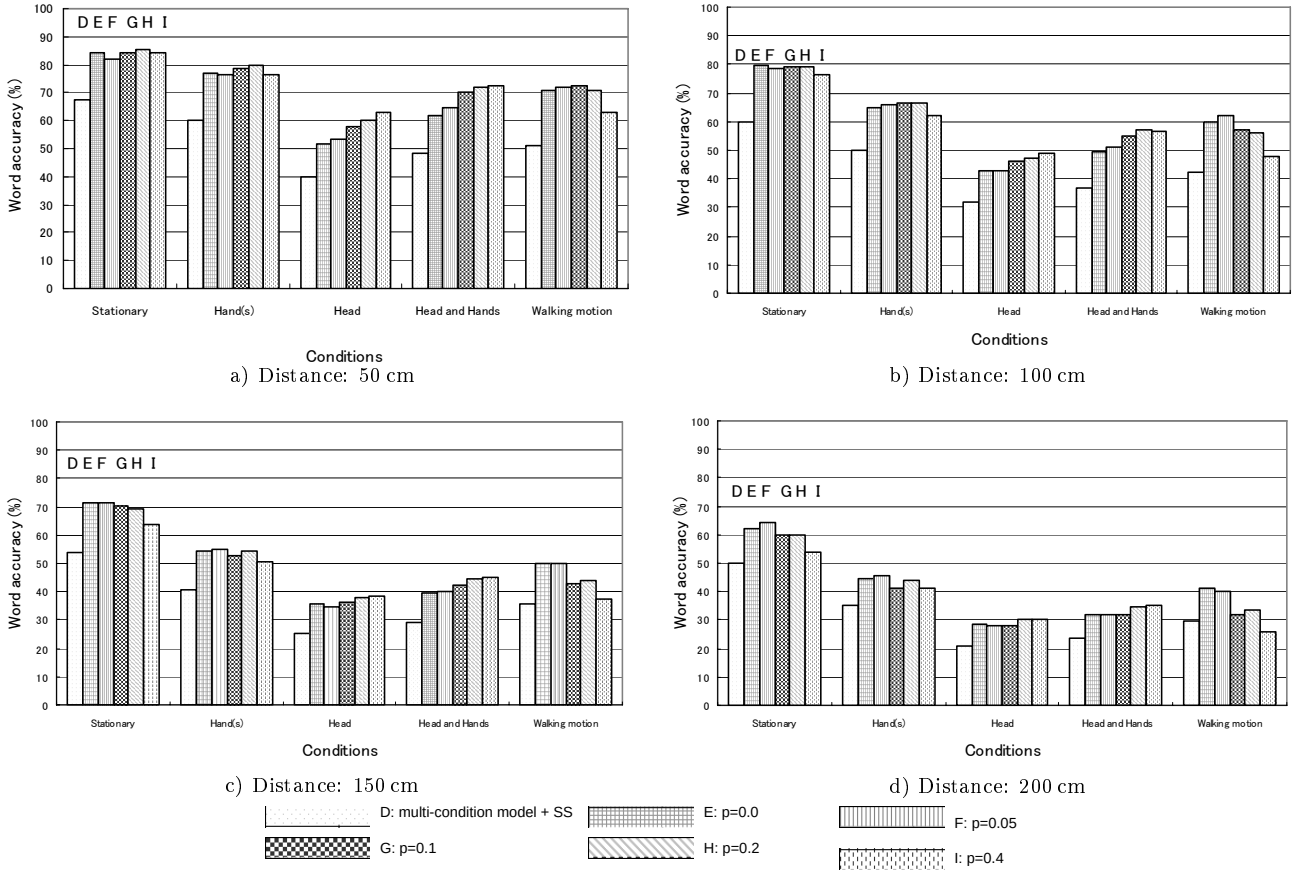
Fig.3a) – d) に実験結果を示す．これらの図でも，結果は雑音グループごとの平均となっている．ここでは，ベースラインとして求めた音声認識結果に加えて，提案手法の認識結果を示した．クリーン音響モデルを用いた **A** と比較し，マルチコンディション学習による音響モデルを用いた **B** および **C** の性能が高く，マルチコンディション学習の有効性が確認できる．**B** と **C** のどちらがより有効であるかは環境によって異なるが，総じて **C** の性能が高いといえる．そこで，**C** を従来手法として提案手法との比較を行う．

MFT を用いた **J** から **L** の中で，最も実環境に即した状況を扱っているのは **J** である．そこで，危険率 p 値 [20] を用いて，**C** に対する **J** の優位性を確認した．**C** に対して **J** が 5% 水準で有意に優位な結果となるのは，距離 50 cm では 34 雑音環境中 28 雑音（Fig.3a)），100 cm で 32 雑音（Fig.3b)），150 cm で 31 雑音（Fig.3c)），200 cm で 33 雑音（Fig.3d)）であった．なお，白色雑音重畳には，$p = 0.1$ を用いた．

実験結果では，雑音環境，距離の違いに関わらず，提案手法 (**J**) が従来手法 (**C**) よりも高い性能を示した．これにより，提案手法の有効性，特に MFT がロボットの動作音に対して有効に働くことが確認できる．

5.3 教師なし MLLR の効果

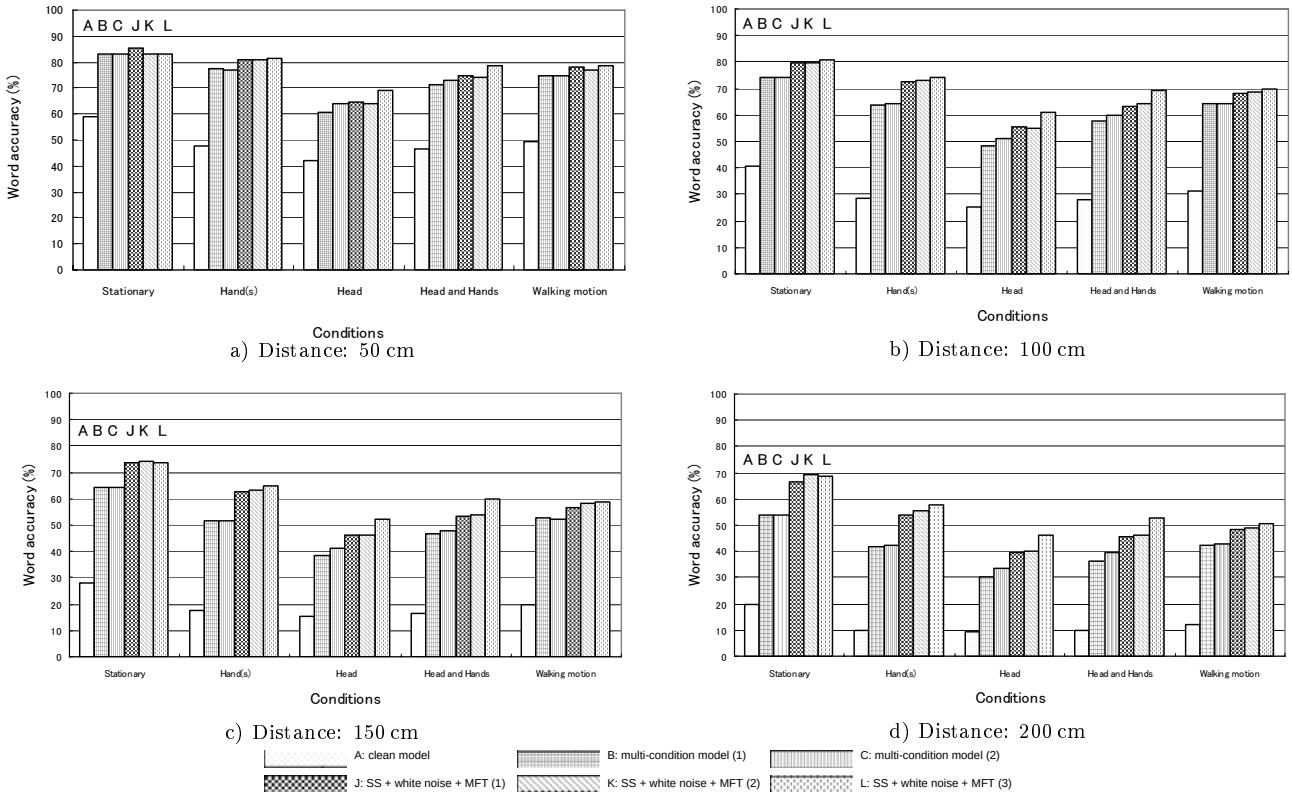
Fig.4a) – d) にマルチコンディション学習と教師なし MLLR の組合せ **C'**，および提案手法と教師なし MLLR の組合せ **J'** に対する実験結果を示す．これらの図でも，雑音グループごとに結果を平均した．距離 50 cm の定常雑音のみ，**J'** が **C'** よりも認識性能が低くなっているが，その他の環境では **J'** が優位となっている．特に 200 cm の距離では，**J'** の有効性が大きく現れている．実験結果より，MLLR との組み合せた場合も，提案手法は従来手法より高い認識性能を示すことが確認できた．



a) Distance: 50 cm　b) Distance: 100 cm

c) Distance: 150 cm　d) Distance: 200 cm

Fig. 2 Result of white noise addition



a) Distance: 50 cm　b) Distance: 100 cm

c) Distance: 150 cm　d) Distance: 200 cm

Fig. 3 Result of MFT

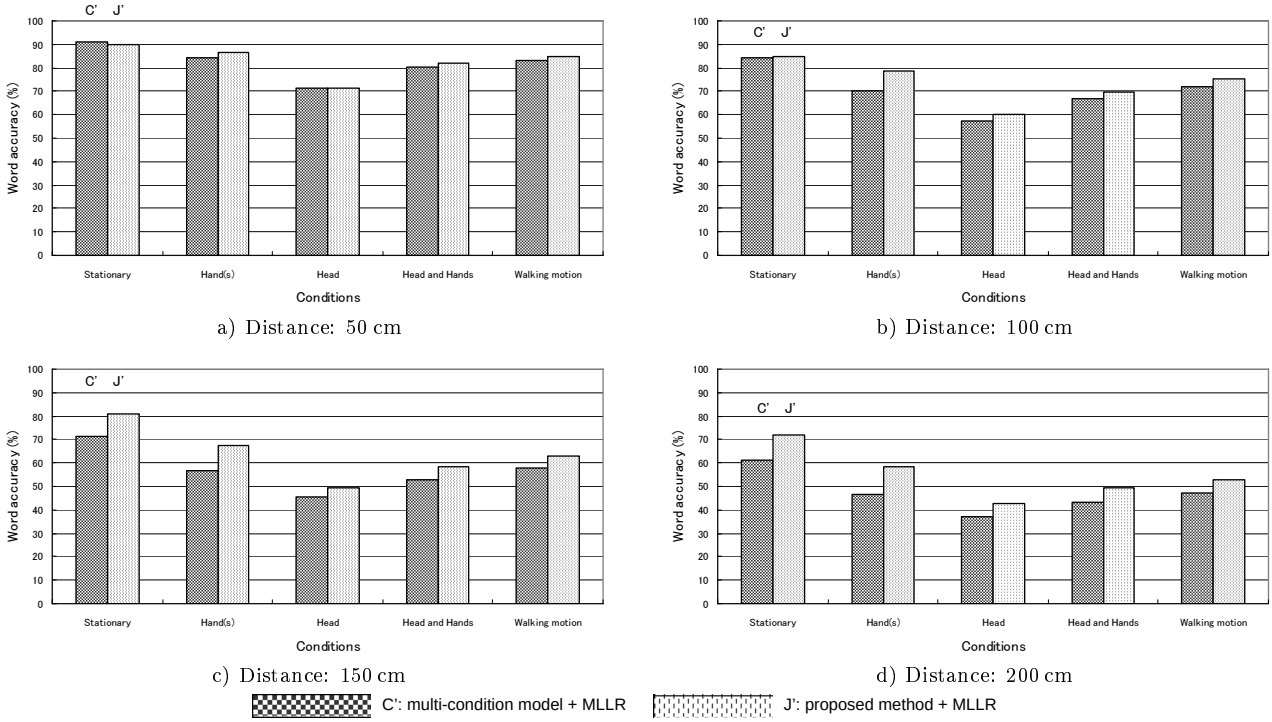


Fig. 4 Result of unsupervised MLLR

6. 考　　察

6.1 雑音除去処理と白色雑音重畳の効果

SS 処理は雑音除去では有効であるが，マルチコンディション学習による音響モデルを用いた音声認識と組み合わせた場合は，SS 処理によって学習時と認識時の音声データが乖離してしまい，認識性能が低下する．実際に，実験でも **D** の認識性能が低いことが示され，これらの単純な組合せで認識性能が低下することが実験結果より明らかとなった．

本稿では，マルチコンディション学習と SS 処理という二種類の雑音に頑健な手法を両方とも有効に活用するため，音響モデル学習の際に，SS 処理を行った音声データを用いた．実際に，**D** と比べて **E** は，環境によらず 10%–20%程度性能が改善され，SS 処理を行った音声データを音響モデル学習に用いることで，二種類の雑音に頑健な手法が両立できることが確認できた．

次に，SS 処理後に白色雑音を重畳することにより SS 処理による歪みを抑え，さらなる認識性能の向上を図った．実際に，白色雑音を重畳した **F** から **I** は白色雑音を重畳しない **E** と比べ，高い認識性能を示しているものがほとんどであり，白色雑音重畳によりスペクトル歪みが軽減され，認識性能が向上することが確認できた．

しかし，実験した全ての環境で最適となる白色雑音重畳率を一意に見出すことはできなかった．雑音グループごとに見ると，頭部動作を含む雑音については白色雑音の重畳を大きくした方がよいことが分かる．頭部動作は，他の動作と比べ雑音源がマイクに近いため，マイクに混入する雑音のパワーが大きいものの，動作時間が短いものが多い．SS 処理は平均雑音を用いて雑音除去を行うため，継続時間の短い動作の平均雑音パワーは小さくなる．よって，実際には SS 処理に用いられる平均雑音パワーよりも大きな雑音が重畳しており，SS 処理の雑音除去性能が低下する．頭部動作では，重畳する白色雑音のパワーを大きくすることで，除去しきれない雑音成分を平坦化し，認識性能向上につながったと考えられる．

他の雑音グループについては，距離が離れるほど白色雑音重畳率を小さくした方が効果的であるという傾向がある．一見，距離が離れると SNR が低下し，SS 処理による歪みが大きくなるため，白色雑音重畳率を大きくした方が効果的と考えられる．しかし，距離が離れた環境では入力信号と比較して雑音信号が大きく，SS 処理のフロアリング (式 (1) において雑音除去後のパワースペクトル値が負にならないよう右辺右側の項が選ばれること) が働くことが多くなる．このフロアリングにより歪みの発生が軽減され，白色雑音重畳率が小さくても高い性能を保つことができたと考えられる．白色雑音重畳率を決定する際に，スペクトルのパワー値だけでなく，フロアリングや雑音の持続時間を考慮に入れることが必要と考えられる．

6.2 MFT の効果

雑音除去処理，白色雑音重畳，および MFT のすべてを用いた提案手法 (**J**) は従来手法 (**C**) と比べてほぼ全ての環境で高い性能を示し，有効であることがわかった．また，MFT を用いない **G** と比べて MFT を用いる **J** は，ほぼ全ての環境で高い性能を示しており，MFT 単体での効果も確認できた．

雑音検出を既知とした **K** および雑音検出・雑音波形を既知とした **L** は提案手法の条件 **J** と比べると理想的な条件であるた

め，認識性能は向上するものの，その差は僅かであった．これにより，提案した雑音マッチングに基づく雑音推定の有効性が確認できた．距離が 50 cm の場合は，**L** が **J** より性能が低い場合も見受けられる．これは，MFT マスク生成がうまくできなかったためと考えられる．本稿で用いた MFT マスク生成手法は，スペクトルの山と谷の重み，および雑音が小さいと推定される部分の重みを大きくしている．しかし，これらの特徴がすべての入力に対して必ずしも正しく推定できるとは限らない．特に，入力に歪が含まれていることが前提となっている場合は難しい．**L** と **J** の認識性能が逆転したのは，特徴の推定が失敗し，**L** で生成された MFT マスクが音声認識にとって最適マスクにならなかったケースであると考えられる．しかし，全体として MFT により認識性能は向上しており，提案したマスク生成手法は有効であるといえる．

提案手法では，マルチコンディション学習は定常的な雑音に対して効果が高いと考え，**B** をベースとした音響モデルを用いている．すなわち，あらかじめロボットの定常雑音を収録しておき，この雑音と音声との重畳を行う．得られた雑音を含む音声に SS，および白色雑音重畳を行った音声データを用いて音響モデルを学習する．しかし，Fig.3c) – d) の結果を見ると，**B** よりも **C** の方が認識性能が高い場合が多く見られる．提案手法においても **C** をベースとした音響モデル，すなわち，ロボットの定常雑音だけでなく，動作音を含む音声データを用いて音響モデル学習を行うことで，認識性能のさらなる向上が期待できよう．

6.3 教師なし MLLR に対する提案手法の有効性

教師なし MLLR と組み合わせた場合においても，提案手法は従来手法より高い性能を示すことが，**J'** と **C'** の結果を比較することで確認できた．MLLR は音響モデルの適応手法として一般的に用いられる実用的な手法であり，マルチコンディション学習との併用も広く用いられている．提案手法と，MLLR を組み合せることによって，より高い性能を得られることからも提案手法のメリットは大きいといえる．

我々は，これまでに，ロボットによるプレゼンテーションを行うソフトウェアを開発 [3] した．プレゼンテーションの場面では話者からの質問が想定される．このような場合に提案手法を教師なし MLLR と組み合わせ，オンラインで音響モデル適応を行い，対話が進むにつれ，質問に対する音声認識性能を向上させていくことが可能であろう．また，案内ロボットについても同様に提案手法を用いて，対話における音声認識を向上させることが可能であろう．

7. ま と め

本稿ではロボットの動作時の音声認識を目的とした雑音に頑健な音声認識手法の提案を行った．提案手法では，SS 処理に基づく雑音除去処理と，歪みを平坦化するための白色雑音の重畳，MFT を用いた非定常雑音への対応を行った．

白色雑音重畳は SNR を低下させるため，一見，認識性能を下げるように思えるが，音声認識に悪影響を与える SS 処理によって発生する歪みを平坦化する効果があり，認識性能を改善することができる．SS 処理は定常雑音の除去には有効であるが，非定常雑音に対しては十分な効果が得られない．このため，信頼度の低い音声特徴量が音声認識に関与する割合を低くすることができる MFT を用いて音声認識性能の向上を図った．

提案手法を用いることにより，従来から音声認識の対雑音性能向上に用いられるマルチコンディション学習を用いた手法よりも認識性能が高いことを示した．さらに，対話中のオンライン適応を想定し，教師なし MLLR を組み合わせた実験を行い，マルチコンディション学習と MLLR を組み合わせた場合よりも認識率が高くなることを確認し，提案手法の有効性を示した．

8. 今 後 の 課 題

今後の課題としては，白色雑音重畳に際し，フロアリングも考慮に入れた重畳方法の検討や，マスクの計算手法について，より最適な手法の検討などが考えられる．また，雑音推定については，特定動作だけではなく，より小さい単位の動作の組合せによる任意の動作への対応や，複数の動作の組合せに対する雑音推定を検討したい．

謝辞 MFT を用いるにあたり貴重なアドバイスを頂いた東京工業大学教授古井貞熙氏，および岩野公司氏に感謝する．また，本実験を行うにあたり貴重なアドバイスを頂いた HRI-JP の船越孝太郎氏および雑音の収録にあたりお手伝いいただいた京都大学山本俊一氏に感謝する．

参 考 文 献

[1] C. Breazeal, *Designing Sociable Robots*, MIT press, 2002.

[2] H. Miwa, T. Okuchi, K. Itoh, T. H., and A. Takanichi, "A new mental model for humanoid robots for human friendly communication – introduction of learning system, mood vector and second order equations of emotion –," in *Proc. of IEEE-RAS International Conference on Robotics and Automation (ICRA 2003)*. 2003, pp. 3588 – 3593, IEEE.

[3] Y. Nishimura, K. Kushida, H. Dohi, M. Ishizuka, J. Takeuchi, and H. Tsujino, "Development and psychological evaluation of multimodal presentation markup language for humanoid robots," in *Proc. 5th IEEE-RAS International Conference on Humanoid Robots (Humanoids-2005)*, 2005, pp. 393–398.

[4] M. Nakano, Y. Hasegawa, K. Nakadai, T. Nakamura, J. Takeuchi, T. Torii, H. Tsujino, N. Kanda, and H. G. Okuno, "A two-layer model for behavior and dialogue planning in conversational service robots," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2005)*, 2005, pp. 1542–1547.

[5] I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa, and K. Yamamoo, "Robust speech interface based on audio and video information fusion for humanoid HRP-2," in *Proc. of IEEE/RAS International Conference on Intelligent Robots and Systems (IROS-2004)*. 2004, pp. 2404–2410, IEEE.

[6] S. Araki, S. Makino, R. Mukai, Y. Hinamoto, T. Nishikawa, and H. Saruwatari, "Equivalence between Frequency Domain Blind Source Separation and Frequency Domain Adaptive Beamforming," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP-2002)*, 2002, pp.1899–1902, IEEE.

[7] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[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. Yamamoto, K. Nakadai, J. M. Valin, J. Rouat, F. Michaud, T. Ogata, H. Komatani, and H. G. Okuno, "Making a robot recognize three simultaneous sentences in real-time," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2005)*, IEEE, Ed., 2005, pp. 897–892.

[10] A. Ito, T. Kanayama, M. Suzuki, and S. Makino, "Internal noise suppression for speech recognition by small robots," in *Proc. of European Conference on Speech Communication and Technology (Eurospeech-2005)*, 2005, pp. 2685–2688.

[11] S. F. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-79)*. 1979, pp. 200–203, IEEE.

[12] J. Barker, M. Cooke, and P. Green, "Robust asr based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise," in *Proc. of 7th European Conference on Speech Communication Technology (Eurospeech-2001)*. 2001, vol. 1, pp. 213–216, ESCA.

[13] A. Hagen and A. Morris, "Comparison of HMM experts with MLP experts in the full combination multi-band approach to robust ASR," in *Proc. of International Conference on Spoken Language Processing (ICSLP-2000)*, 2000, vol. 1, pp. 345–348.

[14] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. of International Conference on Spoken Language Processing (ICSLP-1996)*, 1996, vol. 1, pp. 426–429.

[15] 山出慎吾, 馬場朗, 芳澤伸一, 李晃伸, 猿渡洋, 鹿野清宏, "実環境における頑健な音声認識のための音韻モデルの教師なし話者適応," **電子情報通信学会論文誌**, vol. J87-D-II, no. 4, pp. 933–941, 2004.

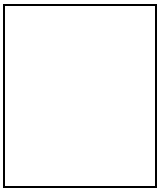
[16] Y. Nishimura, T. Shinozaki, K. Iwano, and S. Furui, "Noise-robust speech recognition using multi-band spectral features," in *Proc. of 148th Acoustical Society of America Meetings*, ASA, Ed., 2004, p. 1aSC7.

[17] 西村義隆, 篠崎隆宏, 岩野公司, 古井貞熙, "周波数帯域ごとの重みつき尤度を用いた雑音に頑健な音声認識," **電子情報通信学会技術研究報告**, *SP2003-116*, 2003, pp. 19–24.

[18] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of Acoust. Soc. America*, Vol.55, No.6, pp.1304-1312, 1974.

[19] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

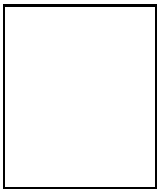
[20] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-89)*, IEEE, Ed., 1989, pp. 532–535.



西村 義隆（Yoshitaka Nishimura）

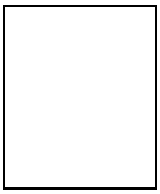
1980 年 1 月 19 日生．2002 年東京工業大学工学部電気・電子工学科卒業．2007 年東京大学大学院情報理工学系研究科電子情報学専攻博士課程修了．同年，明治大学法科大学院入学．博士（情報理工学）．電子情報通信学会，日本音響学会，情報処理学会，日本ロボット学会各学生会員．

（日本ロボット学会学生会員）



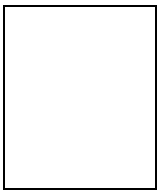
石塚 満（Mitsuru Ishizuka）

1948 年 4 月 23 日生．1971 年東京大学工学部電子工学科卒業．1976 年同大学院博士課程修了．工学博士．同年 NTT 入社，横須賀研究所勤務．1978 年東京大学生産技術研究所 助教授，(1980 年-81 年 Purdue 大学客員准教授)，1992 年東京大学工学部電子情報工学科 教授，2001 年東京大学大学院情報理工学系研究科・電子情報学専攻 教授，2005 年同創造情報学専攻（電子情報学専攻兼任）教授．人工知能，Web インテリジェンス，次世代 Web 情報基盤，生命的エージェントによるマルチモーダルメディアの研究に従事．IEEE, AAAI, 人工知能学会（前会長），電子情報通信学会，情報処理学会，映像情報メディア学会，画像電子学会 会員．



中臺 一博（Kazuhiro Nakadai）

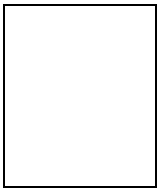
1970 年 10 月 21 日生．1993 年東京大学工学部電気工学科卒業，1995 年同大学院工学系研究科情報工学専攻修了．同年日本電信電話株式会社入社，1997 年 NTT コムウェア（株）出向後，1999 年退職．同年，JST ERATO 北野共生システムプロジェクト研究員．2003 年より（株）ホンダ・リサーチ・インスティチュート・ジャパン，シニア・リサーチャ．博士（工学）．2006 年 4 月より，東京工業大学大学院情報理工学研究科 客員准教授兼務．主にロボット聴覚，実時間情報統合，音環境理解の研究に従事．IROS 2001 BEST Paper Nomination Finalist，2002 年第 2 回船井情報科学振興賞など受賞．日本人工知能学会，日本音響学会，ヒューマンインタフェース学会，IEEE 各会員．　（日本ロボット学会正会員）



中野 幹生（Mikio Nakano）

1965 年 8 月 7 日生．1990 年東京大学大学院理学系研究科相関理化学専攻修士課程修了．1990 年-2004 年日本電信電話 (株) に勤務．2004 年より (株) ホンダ・リサーチ・インスティチュート・ジャパン，シニア・リサーチャ．音声言語理解・対話の研究に従事．博士（理学）．情報処理学会，言語処理学会，電子情報通信学会，人工知能学会，ACM，IEEE，ISCA 会員．

（日本ロボット学会正会員）



辻野 広司（Hiroshi Tsujino）

1960 年 10 月 26 日生．1984 年東京工業大学理学部情報科学科卒業．1986 年同大学院情報科学専攻修士課程修了．1987 年（株）本田技術研究所入社．2003 年より（株）ホンダ・リサーチ・インスティチュート・ジャパン　チーフ・リサーチャ．脳型コンピュータ，知能システム，ヒューマンロボットインターフェース，画像認識などの研究に従事．IEEE，SFN，INNS，日本ロボット学会，人工知能学会，日本ソフトウェア科学会各会員．人工知能学会理事．　（日本ロボット学会正会員）