学術・技術論文

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

山本俊一[*1]　Jean-Marc Valin[*2]　中臺一博[*3]　中野幹生[*3]
辻野広司[*3]　駒谷和範[*1]　尾形哲也[*1]　奥乃博[*1]

Simultaneous Speech Recognition Based on Automatic Missing Feature Mask Generation by Integrating Sound Source Separation

Shunichi Yamamoto[*1], Jean-Marc Valin[*2], Kazuhiro Nakadai[*3], Mikio Nakano[*3], Hiroshi Tsujino[*3], Kazunori Komatani[*1], Tetsuya Ogata[*1] *and* Hiroshi G. Okuno[*1]

Our goal is to realize a humanoid robot that has the capabilities of recognizing simultaneous speech. A humanoid robot under real-world environments usually hears a mixture of sounds, and thus three capabilities are essential for robot audition; sound source localization, separation, and recognition of separated sounds. In particular, an interface between sound source separation and speech recognition is important. In this paper, we designed an interface between sound source separation and speech recogniton by applying Missing Feature Theory (MFT). In this method, spectral sub-bands distorted by sound source separation are detected from input speech as missing features. The detected missing features are masked on recognition not to affect the system badly. Therefore, this method is more flexible when noises change dynamically and drastically. It is the most important issue how distorted spectral sub-bands are detected. To solve the issue, we used speech feature apropriate for MFT-based ASR, and developed automatic missing feature mask generation. As a speech feature, we used a Mel-Scale Log Spectral (MSLS) feature instead of Mel-Frequency Cepstrum Coefficient (MFCC) which is commonly used for ASR. We presented a method of generating missing feature mask automatically by using information from sound source separation. To evaluate our method, we implemented it in a humanoid robot *SIG2*, and performed the experiments on recognition of three simultaneous isolated words. As a result, our method outperformed conventional ASR with MSLS feature.

Key Words: Automatic Missing Feature Mask Generation, Missing Feature Theory, Sound Source Separation, Automatic Speech Recognition, Robot Audition

1. はじめに

将来，様々な面で人間をサポートするようなヒューマノイドロボットを実現するためには，人間と同等の認識能力を有する必要があろう．特に，人間同士のコミュニケーションにおいて音声は重要な位置を占めることから，実環境における音声認識はヒューマノイドロボットの基本的な聴覚機能といえる．

一般に，実環境においてヒューマノイドロボットに搭載されたマイクを用いて音声認識を行う場合，様々な音源からの音が混在した混合音を扱う必要がある．しかし，現在の音声認識技術のほとんどは単一音源を仮定しているため，十分な認識精度が得られないという問題がある．この問題に対処するためには，音源定位，音源分離，分離音認識という主に 3 つの機能が必要である．このうち，音源定位と音源分離については，信号処理や音環境理解 (*Computational Auditory Scene Analysis*, CASA) の分野で研究が行われてきたが，分離音認識はこれまでほとんど扱われていなかった．

このため，実環境での音声認識が必要とされるヒューマン・ロボット・コミュニケーションの分野では，音声だけを収音するために口元に設置された接話型マイクを利用するのが一般的である．例えば，MIT の Kismet は耳介付近に 2 本のマイクを有しているが，音声認識には接話型マイクを利用している [1]．

音声に非音声雑音が混在している混合音については，AURORA プロジェクト [2] [3] などで，研究が行われている．こう

原稿受付
[*1]京都大学情報学研究科知能情報学専攻
[*2]CSIRO ICT Centre
[*3](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 Informatics, Kyoto University
[*2]CSIRO ICT Centre
[*3]Honda Research Institute Japan Co., Ltd.

した状況に対応する一般的な手法として，雑音を含んだ音声に対して HMM パラメータを学習するマルチコンディション学習が挙げられる [4] [5]．この手法で得られた音響モデルには，特定条件下の雑音が反映されているため，想定条件の範囲内の雑音には効果的であり，実際に，カーナビや電話サービスといった音声認識アプリケーションで用いられている．

一方，実環境では音声に音声雑音が混在している混合音を扱わなければならない場合もある．このような問題を扱う研究としては，マルチコンディション学習を同時発話認識に応用した研究 [6] があり，話者・方向依存の音響モデルを構築する際に，音源分離による歪みを含んだ音声も学習データに利用することで，高精度な分離音認識を実現している．しかし，認識の対象となる話者の複数同時発話音声が学習データとして必要になるため制約が大きい．音源分離に重点を置いた研究としては，マイクロフォンアレイを用いたビームフォーミングによる音声分離が挙げられる．例えば，澤田らは，8 ch のマイクロフォンアレイで同時発話音声を分離し，音響モデル適応による分離音声認識を報告している [7]．また，非定常性雑音に対処するために，**ミッシングフィーチャ理論**（*Missing Feature Theory*, MFT）も利用されている [8] [9].

音源分離問題は一般に不良設定問題であることから，元の音源を完全に分離抽出することは困難である．つまり，音声認識では，分離音はある程度歪んでいるということを前提にする必要がある．また，音声歪みを前提とした音声認識を行う場合でも，歪み方は音源分離の手法に依存するため，音源分離の特徴を考慮した処理を行う必要がある．逆に，音源分離側でも，音声認識の処理方法を勘案して，できるだけ音声認識処理にとって認識しやすい分離音声を提供する必要がある．

本研究の目的は，実環境下でヒューマノイドロボットに搭載されたマイクにより音声に音声雑音が含まれる混合音である同時発話を認識することである．我々は，これまでに，2 本のマイクを用いた混合音声分離，およびクリーン音声を先見情報として用いて生成された *a priori* マスク生成 [10] による MFT に基づく音声認識を利用して三話者同時発話認識を実装した [11]．本論文では，8 本のマイクを用いたマイクロフォンアレイ音源分離と音源分離からの情報を利用したミッシングフィーチャマスク（*Missing Feature Mask*, MFM）自動生成を提案し，複数同時発話認識に適用した．

以降，第 2 章では混合音声認識における本研究の課題について述べる．第 3 章では音源分離と音声認識の統合ついて説明する．第 4 章ではミッシングフィーチャマスク自動生成について述べ，第 5 章でこれらの評価実験を行う．最後に第 6 章で本研究のまとめをする．

2. 混合音声認識における本研究の課題

実環境において混合音を扱えるロボット聴覚を実現するためには，次のような課題がある．

(1) 混合音の音源分離
(2) 分離音声の認識
(3) 音源分離と音声認識の統合

音源分離による分離音は完全ではなく，ある程度歪んでいたり雑音を含んでいたりするため，分離音声の認識ではこのような歪みや雑音を扱う必要がある．このため，音源分離と音声認識を独立に行うよりもお互いの処理を補完するように統合することが重要である．一般に，音源分離と音声認識の間のインタフェースは，直列型と統合型の 2 種類に分類できる．前者は音源分離を音声認識のフロントエンドとして使い，音声認識側で分離音に対応する方法で，音源分離と音声認識が独立して行われている．このため，音源分離による分離音声の歪による影響をすべて音声認識側で対応しなければならない．一方，後者は音源分離と音声認識の両方を改良して情報統合することによって性能向上を狙っている．我々は音源分離と音声認識の間のインタフェースとして後者の統合型を採用した．

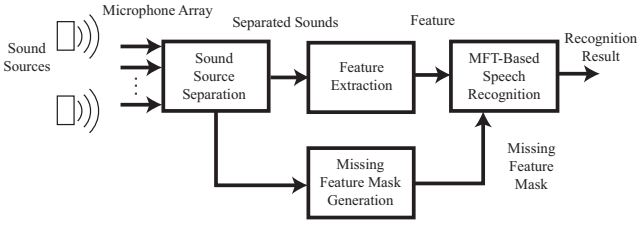

Fig. 1 System overview

山本らはこれまでに，2 本のマイクロフォンによるアクティブ方向通過型フィルタ（*Active Direction-Pass Filter*, ADPF）と *a priori* マスクを利用した MFT に基づく音声認識を統合した混合音声認識システムを開発した [11]．これは，*a priori* マスクを利用しているので理想システムと呼ぶことにする．この理想システムで残された重要な課題は，MFM 自動生成を行うことであった．

本論文では MFM 自動生成とは，*a priori* マスクのようにクリーン音声を用いて MFM を生成するのではなく，実環境で得られる情報のみを利用して MFM を生成することを言う．理想システムでは *a priori* マスクを用いており，MFT に基づく音声認識による性能の上限を測るのには適していたが，実用的ではなかった．また，MFM 自動生成には，各特徴量成分に関して音源分離によって歪んでいるかどうかという情報が必要になる．しかし，2 本のマイクロフォンを利用した ADPF では入力情報が少なく，分離音の音源間の干渉を推定するのが難しいという問題があった．

3. 音源分離と音声認識の統合

3.1 システムの概要

混合音声認識システムは以下の 3 つのモジュールから構成されている．(Fig. 1)．

(1) 音源分離モジュール
(2) MFM 自動生成モジュール
(3) MFT に基づく音声認識モジュール

音源分離モジュールには，ヒューマノイドロボットに設置された 8 本の無指向性マイクロフォンによるマイクロフォンアレイ音源分離 [12] を利用した．このマイクロフォンアレイ音源分離は，幾何学的音源分離（Geometric Source Separation, GSS）と多チャンネル post-filter から構成されており，多チャンネル

post-filter は音源分離した分離音における干渉音を抑制する効果があるだけでなく，特定の時刻，特定の周波数における雑音に関する手がかりを得ることができる．詳細は 3.2 節で述べる．

我々が開発した MFM 自動生成のアイデアは，多チャンネル post-filter から得られる情報をミッシングフィーチャの手がかりとし，MFM を自動生成することにある．多チャンネル post-filter の入出力と多チャンネル post-filter で推定された背景雑音から MFM 推定を行う．

MFT に基づく音声認識では，マイクロフォンアレイ音源分離による分離音声と自動生成された MFM を利用して音声認識を行う．本論文では，MFT に基づく音声認識が行えるマルチバンド版 Julian [13] を利用した．マルチバンド版 Julian は通常の Julian [14] をマルチバンド音声認識のために改良したものである．Julian は単純なモノフォンやトライフォンだけでなく，状態を共有したトライフォンや，分布を共有したモデルなどの音響モデルもサポートしている．Julian は 2 パスによる HMM のデコードを行い，リアルタイムで音声認識ができるように実装されている．

以下，MFM 自動生成に必要なシステムの概要を述べる．

3.2 マイクロフォンアレイ音源分離

本システムに採用した音源分離手法は GSS と多チャンネル post-filter を組み合わせた手法である．本稿では，音声認識にとって認識しやすい分離音声とは，除去された雑音も分かる分離音声であるとする．そこで，本音源分離手法では，分離音声だけではなく除去された雑音の情報を併せて出力するように改良している．音源分離は，基本的には Parra ら [15] によって提案された GSS に基づく線形音源分離法の実装であり，確率的勾配法を適用し，推定に利用する時間幅を短くすることによって高速化を図っている．さらに分離音を強調するために，Ephraim らによって提案された最適化推定 [16] [17] に基づく周波数領域 post-filter を利用している．この post-filter は非定常性の干渉を考慮した Cohen らの提案したものの実装である [18].

GSS は，畳み込み混合モデルが狭帯域信号においては瞬時混合モデルで近似できることを利用して周波数領域で音源分離を行う手法である．各帯域の分離行列は，2 つのコスト関数により推定される．1 つ目は，信号間相関を最小化して無相関化するためのコスト関数である．これは独立性を評価するコスト関数として独立成分分析（Independent Component Analysis, ICA）においても利用されているので，GSS は ICA に似た性質を持っている．2 つ目は，音源からマイクロフォンまでの伝達関数と分離行列の積が単位行列になるようにする．つまり目的方向の利得を 1 に，他の方向の利得を 0 になるように最適化するためのコスト関数である．伝達関数は，与えられた音源方向を利用して音源からマイクロフォンまでの遅延時間をもとに計算するので，音源方向が既知である必要がある．このように幾何的な情報から得られる関数なので幾何制約（geometric constraint）と呼ばれており，このコスト関数を導入したことが GSS と呼ばれる所以である．

多チャンネル post-filter [19] では，Fig. 2 に示すように，GSS のチャンネル出力雑音を定常性雑音と非定常性雑音に分けて推定を行う．定常性雑音は，主に背景雑音であるとし，背景雑音推定

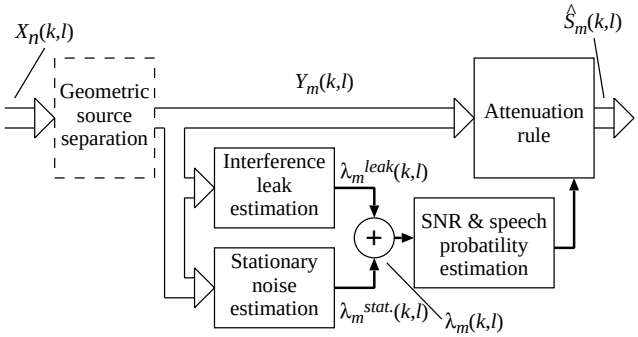

Fig. 2 Overview of multi-channel post-filter

を行う．非定常性雑音は，GSS の過程で他のチャンネルから漏洩したものであると仮定して，適応的に他チャンネルからの干渉成分のスペクトル推定を行う．定常性雑音推定と非定常性雑音推定を統合することにより，最終的な雑音推定を行う．なお，Fig. 2 において，k は周波数帯域のインデックス，l は時刻，$S_m(k,l)$ は音源 m の信号，$X_n(k,l)$ は n 番目のマイクから GSS への入力，$Y_m(k,l)$ は GSS で分離された 音源 m の信号, $\hat{S}_m(k,l)$ は多チャンネル post-filter 処理後の分離音源 m の信号を表している．$G_m(k,l)$ は重み関数であり $\hat{S}_m(k,l) = G_m(k,l)Y_m(k,l)$ と定義される．

この多チャンネル post-filter では，干渉音源はすべて定位されているものとし，残響，音源定位誤り，マイクの周波数応答の相違，近接音場効果などによるチャンネル間の漏洩は一定とする．

ロボットに設置したマイクロフォンにより音源分離を行う場合の特有の問題として，ロボットの手や頭部の動きが原因でロボットの体からの反射が一様ではないことが挙げられる．本手法では，多チャンネル post-filter からの分離音の雑音に関する情報を利用した MFM 自動生成によって，反射による分離誤りに動的に対応することができるという利点がある．ただし，GSS の幾何制約では反射の影響が考慮されていないので，GSS によって推定した分離行列は誤差の影響により，ICA によって推定した分離行列と比較して性能が低下する．一方 ICA には，マイクロフォンの数が音源数より多い場合に 1 音源の信号が複数の成分に分離するという問題がある．反射を考慮した幾何制約を GSS に導入すれば，ICA と同等の性能を持ちつつ，この問題を原理的に有しない GSS を構築できる可能性がある．これは今後の課題である．

3.2.1 雑音推定アルゴリズム

推定された雑音の分散 $\lambda_m(k,l)$ は以下の式で定義される．

$$\lambda_m(k,l) = \lambda_m^{stat}(k,l) + \lambda_m^{leak}(k,l) \qquad (1)$$

ここで，$\lambda_m^{stat}(k,l)$ は音源 m，フレーム l，周波数 k の定常性雑音の推定値であり，$\lambda_m^{leak}(k,l)$ は音源から漏洩した信号の推定値である．

定常性雑音 $\lambda_m^{stat}(k,l)$ は Minima Controled Recursive Average (MCRA) [20] により計算する．

非定常性雑音 $\lambda_m^{leak}(k,l)$ は，他の音源からの干渉は係数 η（一般的には $-10dB < \eta < -5dB$）により除去することがで

きるものとして，$\lambda_m^{leak}(k,l)$ を以下のように定義する．

$$\lambda_m^{leak}(k,l) = \eta \sum_{i=0, i\neq m}^{M-1} Z_i(k,l) \tag{2}$$

ここで，M は音源数，$Z_m(k,l)$ は分離音スペクトル $Y_m(k,l)$ を時間方向に平滑化したスペクトルであり以下の式により再帰的に定義される．α_s は平滑化のための係数であり，実装では $\alpha_s = 0.5$ である．

$$Z_m(k,l) = \alpha_s Z_m(k,l-1) + (1-\alpha_s) Y_m(k,l) \tag{3}$$

3.2.2 音声が存在する場合の抑制規則

音声が存在するという仮定 H_1 のもとでの抑制規則について述べる．以後，特に明示しない限り音源のインデックス m と時刻 l は省略し，各式は変数 m，l のもとに定義されるものとする．音源 $S(k) = A(k)e^{j\alpha_k}$ とすると，本雑音抑制規則は，音源 $S(k)$ の振幅スペクトル $A(k)$ の最小二乗平均誤差推定に基づいている．

振幅の推定量 $\hat{A}(k)$ は，文献 [17] より以下の式で定義される．

$$\hat{A}^{\alpha}(k) = E\left[A^{\alpha}(k)\middle|Y(k)\right] \tag{4}$$

$$= \frac{\int_0^{\infty}\int_0^{2\pi} A(k)^{\alpha} p(Y(k)|A(k),\alpha_k) p(A(k),\alpha_k) d\alpha_k dA(k)}{\int_0^{\infty}\int_0^{2\pi} p(Y(k)|A(k),\alpha_k) p(A(k),\alpha_k) d\alpha_k dA(k)} \tag{5}$$

ここで，α はモーメントの次数を表しており，実装では $\alpha = 1$ である．また，$E[\cdot]$ は期待値を，$p(\cdot)$ は確率密度関数を表している．ガウスモデルを用いると，$p(Y(k)|A(k),\alpha_k)$ と $p(A(k),\alpha_k)$ は以下のように定義される．

$$p(Y(k)|A(k),\alpha_k) = \frac{1}{\pi\lambda(x)} \exp\left\{-\frac{1}{\lambda(k)}|Y(k) - A(k)e^{j\alpha_k}|^2\right\} \tag{6}$$

$$p(A(k),\alpha_k) = \frac{A(k)}{\pi\lambda_s(k)} \exp\left\{-\frac{A(k)^2}{\lambda_s(k)}\right\} \tag{7}$$

ここで，$\lambda(k)$ は式 (1) で推定された雑音の分散，$\lambda_s(k) = E\left[|S(k)|^2\right]$ である．さらに，振幅の推定量は以下のように変形できる（詳細は [16] を参照）．

$$\hat{A}(k) = \frac{\sqrt{\upsilon(k)}}{\gamma(k)} \left[\Gamma\left(1+\frac{\alpha}{2}\right) M\left(-\frac{\alpha}{2}; 1; -\upsilon(k)\right)\right]^{\frac{1}{\alpha}} |Y(k)| \tag{8}$$

$$\gamma(k) \triangleq \frac{|Y(k)|^2}{\lambda(k)},\quad \xi(k) \triangleq \frac{\lambda_s(k)}{\lambda(k)},\quad \upsilon(k) \triangleq \frac{\gamma(k)\xi(k)}{\xi(k)+1} \tag{9}$$

ここで，$M(a;c;x)$ は合流型幾何関数，$\gamma(k)$ は事後信号対雑音比（S/N 比），$\xi(k)$ は事前 S/N 比である．よって，音声が存在する場合の利得 $G_{H_1}(k)$ は以下の式で求められる．

$$G_{H_1}(k) = \frac{\hat{A}(k)}{|Y(k)|} \tag{10}$$

音声が必ずしも存在しないということを考慮して，事前 S/N 比 $\xi(x)$ は以下の式で再帰的に推定される [20]．

$$\hat{\xi}(k,l) = \alpha_p G_{H_1}^2(k,l-1)\gamma(k,l-1) + (1-\alpha_p)\max\{\gamma(k,l)-1, 0\} \tag{11}$$

ここで，α_p は雑音除去と過渡歪みの間のトレードオフを制御する重み係数である．

3.2.3 音声存在確率を考慮した振幅推定

前節で述べた，音声が存在すると仮定した場合の振幅推定を一般に拡張し，音声存在確率を考慮した振幅推定について述べる．音源 m において，音声が存在するという仮定 H_1 と音声が存在しないという仮定 H_0 とすれば，式 (4) は次のように変形できる．

$$\hat{A}(k) = \Big(p(k)E[A^{\alpha}(k)|H_1, Y(k)] + \{1-p(k)\}E[A^{\alpha}(k)|H_0, Y(k)]\Big)^{\frac{1}{\alpha}} \tag{12}$$

$$= \left(p(k)G_{H_1}^{\alpha}(k) + (1-p(k))G_{min}^{\alpha}\right)^{\frac{1}{\alpha}} |Y(k)| \tag{13}$$

ここで，$p(k)$ は周波数 k における音声存在確率である．

よって，最適な利得は次の式から得られる．

$$G(k) = \left(p(k)G_{H_1}^{\alpha}(k) + \{1-p(k)\}G_{min}^{\alpha}\right)^{\frac{1}{\alpha}} \tag{14}$$

ここで，$G_{H_1}(k)$ は，式 (10) で定義され，G_{min} は音声が存在しない場合に許される最小利得であり，$G_{min} = 0$ とする．$\alpha = 1$ の場合，次のようになる．

$$G(k) = p(k)G_{H_1}(k) \tag{15}$$

$G_{min} = 0$ とすると，音声が存在しない確率が高い場合には，利得が 0 に近づく傾向がある．干渉が定常性雑音ではなく音声である場合には，残された多重漏話雑音はミュージカルノイズの原因になるので，この傾向は特に重要である．

音声存在確率 $p(k)$ は文献 [20] の式 (9) より，次の式で定義される．

$$p(k) = \left\{1 + \frac{\hat{q}(k)}{1-\hat{q}(k)}(1+\xi(k))\exp(-\upsilon(k))\right\}^{-1} \tag{16}$$

ここで，$\hat{q}(k)$ は周波数 k に音声が存在しない場合の事前確率の推定値であり，文献 [20] の式 (28) より，以下のように定義される．

$$\hat{q}(k) = 1 - P_{local}(k)P_{global}(k)P_{frame} \tag{17}$$

ここで，P_{local}，P_{global}，P_{frame} は，それぞれ，[20] の式 (25) で定義されており，現在のフレームにおける周波数帯域 k 周辺の狭い周波数帯域での音声らしさ，広い周波数帯域での音声らしさ，全周波数帯域での音声らしさを表している．これらの音声らしさは，前のフレームの事前 S/N 比 $\hat{\xi}(k)$ から求まるので，$\hat{q}(k)$ は，前のフレーム，または現在のフレームの周波数帯域 k の周辺の帯域に音声が存在しない場合に大きくなるような関数である．

3.3 ミッシングフィーチャ理論に基づく音声認識

MFT に基づく音声認識は一般の音声認識と同様に，隠れマルコフモデル（*Hidden Markov Model*，HMM）に基づいている．一般の HMM に基づく音声認識システムでは，状態遷移確率と出力確率から与えられた信号系列を最も高い確率で出力する状態遷移系列を求める．

時刻 T までの入力音声の観測ベクトル列を X として以下のように表す．

$$X = \{x_t | t = 1, 2, \cdots, T\} \tag{18}$$

ここで，x_t は時刻 t に観測されたベクトルである．音声認識においては，K 個の音素 HMM などの音響モデルを

$$\{H_k | k = 1, 2, \cdots, K\} \tag{19}$$

として，

$$\hat{H} = \underset{i}{\operatorname{argmax}}\, P(H_i|X) \tag{20}$$

となる音響モデル $\hat{H}$ を選択すればよい．この確率はベイズ則によって以下のように求められる．

$$P(H_i|X) = \frac{P(X|H_i)P(H_i)}{P(X)} \tag{21}$$

ここで，条件付き確率 $P(X|H_i)$ は音響モデル H_i によって得られる確率である．$P(X)$ は H_i に無関係であり $P(H_i)$ は音響モデル H_i の示す音素が出現する確率なので，通常，大量の言語コーパスを用いて学習される．そして，$P(X|H_i)$ を求めるために，一般的に HMM が用いられる．

HMM に基づく音声認識では，各音素を確率状態遷移機械（マルコフモデル）で表現している．HMM は，観測事象が離散シンボルである場合だけでなく，音声のような連続信号の場合においても有効で，確率密度関数を用いて連続信号を直接モデル化する．つまり，HMM では音声の時間的な変化を状態遷移として捉え，その各状態での特徴パラメータの出力を確率分布として表現している．

MFT に基づく音声認識システムでは，このうち出力確率の計算方法が一般の音声認識とは異なっている．特徴ベクトル x，状態 S_j の時の正規分布の確率密度関数を $f(x|S_j)$，L を混合正規分布の混合数，$P(l|S_j)$ を混合係数，N を特徴量の次元数とする．このとき，通常の連続分布型 HMM では出力確率は以下のように定義される．

$$b_j(x) = f(x|S_j) = \sum_{l=1}^{L} P(l|S_j) f(x|l, S_j) \tag{22}$$

しかし，MFT に基づく音声認識では，出力確率 $b_j(x)$ は信頼できる特徴量ほど出力確率に大きく貢献し，信頼できない特徴量ほど出力確率に貢献しないように設計する．つまり，信頼できる特徴だけが出力確率の計算に用いられ，信頼できない特徴による影響を除去しなければならない．これを実現するために，特徴量の各成分に対する信頼度を表す MFM ベクトル $M(i)$ を用いて以下のように定義する．

$$b_j(x) = \sum_{l=1}^{L} P(l|S_j) \exp\left\{ \sum_{i=1}^{N} M(i) \log f(x(i)|l, S_j) \right\} \tag{23}$$

この式によると，信頼できない特徴量に対するすべての音素 HMM の尤度が等しくなるので認識には影響しない．さらに，正解の音素 HMM の尤度を低下させるような信頼度の低い特徴量をマスクすることにより，正解の音素 HMM の尤度が相対的に低くなるのを防ぐことができる．

3.4 MFT のための音声認識特徴量

一般に音声認識システムでは，音声の特徴として MFCC が用いられる．MFCC は入力音声がクリーンな場合は有効であるが，入力スペクトルに歪みがあると，それがたとえ特定の周波数領域での歪みであっても， MFCC の全係数に影響を与えてしまい，ロバスト性が低下する．また，音源分離手法の多くは，周波数領域において分離処理を行うので，スペクトル歪みが少なからず生じる．このため，分離音声の認識で，特徴量として MFCC を利用した場合は，スペクトル歪みが全 MFCC に広がり，MFM を推定することは困難である．一方，スペクトル特徴量としては，ガンマトーンフィルタバンクの出力が用いられることも多い．しかし，対ノイズロバスト性を向上させるために，MFCC 算出時に行われるような特徴量の正規化が難しく，ロバスト性の面でパフォーマンスを確保することが難しい．

本稿で扱う MFT ベースの音声認識システムでは，音声認識の特徴量としてスペクトル特徴量を用いる．本質的には，MFCC を逆離散コサイン変換することによって得られる**メルスケール対数スペクトル**（*Mel-Scale Log Spectrum*, MSLS）を用いる [13]．

周波数領域の特徴量を利用することにより，ビームフォーミングの後処理である多チャンネル post-filter とも親和性が高いというメリットもある．多チャンネル post-filter は，周波数領域で背景雑音推定や，他の音源からの干渉成分のスペクトル推定を行っており，これらの情報から MFM 自動生成が期待できる．

以下に，MFCC で行われるのと同等の正規化を行ったメル周波数領域対数スペクトルの導出の手順を示す．

(1) 音響信号を 16 ビット，16 kHz でサンプリングし，窓幅 25 ms，シフト幅 10 ms の FFT を行う．
(2) メル周波数領域で等間隔に配置した 24 個の三角形窓によりフィルタバンク分析を行う．
(3) 24 個のフィルタバンクの出力の対数を取り，24 次元のメル周波数対数スペクトルを得る．
(4) 対数スペクトルを離散コサイン変換し，24 次元のケプストラム係数を得る．
(5) ケプストラム係数のうち 0 次と高次の項を除去し，1-12 次の項を用いる．
(6) ケプストラム平均除去（CMS）を行う．
(7) 逆離散コサイン変換を行って，スペクトル領域に戻す．
(8) 次元毎に一次回帰係数を計算する．
(9) 一次回帰係数と合わせて，計 48 次元の特徴量として抽出し，メル周波数対数スペクトル特徴量を得る．

分離音声から MSLS 特徴量を抽出した場合の例を Fig. 3，4 に示す．Fig. 3 は「いよいよ」という分離音声のスペクトログラムであり，Fig. 4 は抽出した MSLS 特徴量である．横軸はフレーム数，縦軸は特徴量を表しており，下半分が MSLS 特徴量であり，上半分がその一次回帰係数を表している．

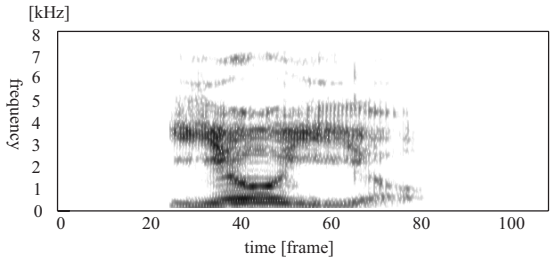

Fig. 3 Spectrogram of separated sounds (speech: i y o i y o)

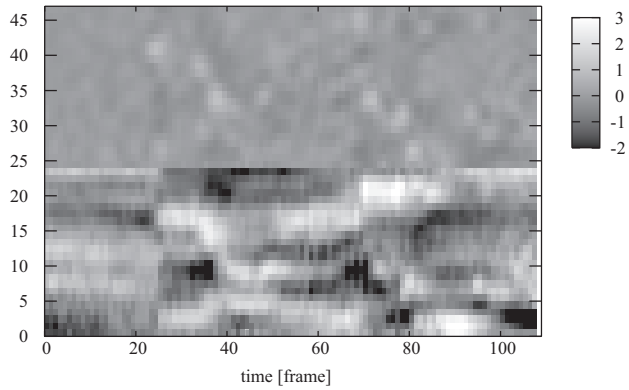

Fig. 4 Mel-scale log spectrum feature (speech: i y o i y o)

4. 音源分離からの情報を利用した MFM 自動生成

MFM を自動生成するには，分離音声のスペクトルのうち，どの周波数帯域が歪んでいるかという情報が必要である．先見的情報を与えず，音源分離処理から得られるデータのみを利用して，このような情報を得るために，多チャンネル post-filter の入力および，出力音響信号，推定された背景雑音のスペクトルを利用する．多チャンネル post-filter は，ビームフォーマーの出力音響信号を入力として雑音推定を行い，雑音を抑制した音響信号を出力するフィルタである．詳細なアルゴリズムは，3. 2 節に記述した．

まず，多チャンネル post-filter から得られる入力，出力，背景雑音のスペクトルをメル周波数対数スペクトル特徴量の抽出と同様にメルスケールに変換する．フレーム k，サブバンド i のときの多チャンネル post-filter への入力を $Y(k,i)$，多チャンネル post-filter からの出力を $\hat{S}(k,i)$，多チャンネル post-filter で推定された背景雑音を $BN(k,i)$ とする．

一次回帰係数でない特徴量に対応する MFM $\{M(k,i)|i = 1,\cdots,\frac{N}{2}\}$ は，以下のように 2 値のマスク（信頼できるとき 1，信頼できないとき 0）として定義する．

$$M(k,i) = \begin{cases} 1, & m(k,i) > T \\ 0, & otherwise \end{cases} \tag{24}$$

$$m(k,i) = \frac{\hat{S}(k,i) + BN(k,i)}{Y(k,i)} \tag{25}$$

閾値 T を変えると単語正解率も変わり，$T = 0.25$ 付近で単語正解率が最大となり，その周辺では数%単語正解率が低下する．T を 0 に近づけると，ほとんどの特徴量がマスクされてしまい単語正解率が低下し，一方，T を 1 に近づけると，ほとんどの特徴量がマスクされなくなり単語正解率が低下する．従って，閾値 T は実験的に求め，0.25 とした．

また，特徴量の一次回帰係数に対する MFM $\{M(k,i)|i = \frac{N}{2}+1,\cdots,N\}$ は，以下のように定義する．この場合も，2 値のマスクとなる．

$$M(k,i) = \prod_{i=k-2, i\neq k}^{k+2} M\left(i, i - \frac{N}{2}\right) \tag{26}$$

特徴量とその一次回帰係数に対応したマスクからなる MFM の次元数は，スペクトル特徴量と同じ 48 となる．

多チャンネル post-filter への入力 $Y(k,i)$ とは，GSS からの出力でもある．つまり，目的方向以外の音源からの漏洩，背景雑音，各マイクロフォンの周波数特性の違いなどによる音源分離誤りがある．特に大きな影響があるものは，目的方向以外の音源からの漏洩と背景雑音である．多チャンネル post-filter はこれらの音源分離誤りを抑制するので，理想的には出力 $\hat{S}(k,i)$ からは音源分離誤りが除去されていると考えられる．簡単に言うと，

$$\begin{aligned} Y(k,i) \;=\; & (\text{目的方向の音源}) \\ & +(\text{目的方向以外の音源からの漏洩}) \\ & +(\text{背景雑音}) \end{aligned} \tag{27}$$

となる．

直感的には MFM 生成のために多チャンネル post-filter の利得，つまり $\hat{S}(k,i)/Y(k,i)$ を特徴量の信頼度とすることが考えられる．しかし，利得が 0 に近い場合に音声のパワーが弱い帯域がほとんど信頼されなくなるという問題がある．利得が 0 に近いのは雑音が支配的な場合であり，無音区間などの音声のパワーが弱い帯域である．この音声のパワーが弱い帯域も音声認識のためには重要であるので，雑音のうち漏洩雑音が支配的であれば信頼できないとするが，背景雑音が支配的であれば信頼できるとする．これは，多チャンネル post-filter による漏洩雑除去の信頼性の方が背景雑音除去の信頼性よりも低いという仮定に基づいている．

自動生成された MFM の例を Fig. 5 に示す．Fig. 5 a) c) e) は三話者同時発話の各方向の分離音声のスペクトル特徴量，Fig. 5 b) d) f) は a) c) e) それぞれの特徴量に対して自動生成された MFM である．

自動生成した MFM の妥当性を検討するために，*a priori* マスクを正解として自動生成された MFM の正解率を計算した．音声が存在する区間の全フレーム，全特徴量成分に対して，単純に *a priori* マスクと比較した場合だけでなく，信頼できる特徴量を信頼できないと判定した場合と，信頼できない特徴量を信頼できると判定した場合のペナルティを考慮した正解率も計算した．信頼できる特徴量を信頼できないと判定した場合は，リスクが少ないので正解と判定し，信頼できない特徴量を信頼できると判定した場合はリスクが大きいので不正解と判定した．

三話者の間隔を 10 度から 10 度毎に 90 度まで変えて録音した三話者同時発話 216 発話 9 セットに対して本手法により MFM を自動生成した．これらの三話者同時発話は 5 章の実験で利用す

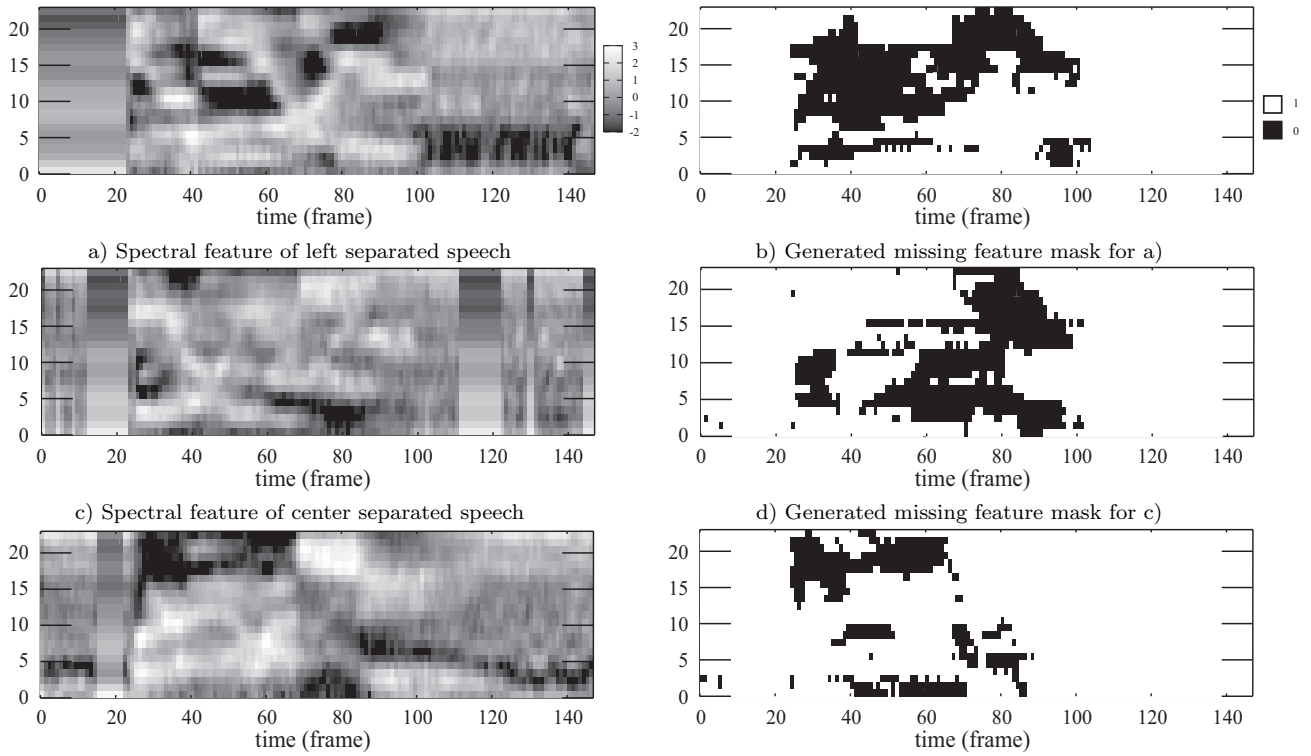

Fig. 5 Examples of missing feature mask

るものと同一である．これらの MFM に対して，単純に *a priori* マスクと比較して計算した正解率は 56.9%，ペナルティを考慮して計算した正解率は 69.9%であった．

5. 三話者同時発話認識実験

システムの評価を行うためにヒューマノイド *SIG2* に 8 本のマイクを取り付け，三話者同時発話認識実験を行った．3 体のスピーカから異なる組み合わせで ATR 音素バランス単語を再生して三話者同時発話を録音し，孤立単語認識により評価した．

5.1 評価項目

評価実験では，具体的には次の 3 点を明らかにする．

(1) 本システムの MFM 自動生成の有効性
(2) 分離音認識に対応する他の解決法との比較
(3) 本システムの MFM 自動生成の上限の同定と本システムの性能の比較

(1) では，MFM 自動生成を利用しない場合と比較して，本手法である MFM 自動生成によって単語正解率が向上したかどうかを評価する．音響モデルはクリーン音声による音響モデルを利用し，自動生成した MFM を利用する場合と利用しない場合の実験を行った．自動生成した MFM を利用しない場合では，特徴量が MFCC と MSLS のそれぞれの場合で実験を行った．

(2) では，分離音認識に対応する他の解決法の一つであるマルチコンディション学習 [6] との比較を行う．マルチコンディション学習を利用した音声認識は，特徴量が MFCC と MSLS のそれぞれの場合で実験を行った．

(3) では，本システムの MFM 自動生成の上限を同定するために，*a priori* マスクを利用した MFT に基づく音声認識 [11] を行い，本システムの性能と比較する．

具体的に行った実験は，次の 7 種類である．

①クリーン音声による音響モデル（MFCC）で Julian による音声認識
②クリーン音声による音響モデル（MSLS）で Julian による音声認識
③自動生成したマスクを利用してクリーン音声の音響モデル（MSLS）でマルチバンド版 Julian による音声認識
④マルチコンディション学習による音響モデル（MFCC）で Julian による音声認識
⑤マルチコンディション学習による音響モデル（MSLS）で Julian による音声認識
⑥自動生成したマスクを利用してマルチコンディション学習による音響モデル（MSLS）でマルチバンド版 Julian による音声認識（③と⑤の組み合わせ）
⑦*a priori* マスクを利用してクリーン音声による音響モデル（MSLS）でマルチバンド版 Julian による音声認識

評価項目 (1) のために実験①，②，③で比較を行い，評価項目 (2) のために実験③，④，⑤，⑥で比較を行い，評価項目 (3) のために実験③，⑦で比較を行った．

5.2 実験条件

実験を行った部屋は 4 m × 5 m の大きさで，残響時間は 0.3 - 0.4 秒 (RT_{20}) である．実験で利用した *SIG2* を Fig. 6 に示す．8 本の無指向性マイクロフォンが，両肩に 2 本，両胸に 2 本，背中に 2 本，腰に 2 本設置されている．マイクロフォンア

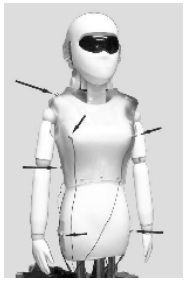

Fig. 6 *SIG2* with eight microphones

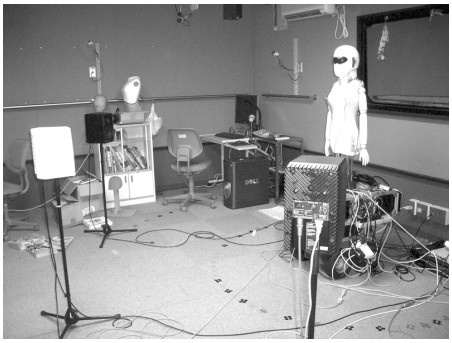

Fig. 7 *SIG2* and three loudspeakers in experiment room

Table 1 P-values of null hypothesis

a) Experiment ① and Experiment ③

speaker interval	left	center	right
10°	0.000	0.000	0.078
20°	0.000	0.000	0.597
30°	0.000	0.000	0.108
40°	0.000	0.000	0.002
50°	0.000	0.000	0.016
60°	0.000	0.087	0.154
70°	0.000	0.000	0.018
80°	0.000	0.004	0.059
90°	0.000	0.000	0.028

b) Experiment ② and Experiment ③

speaker interval	left	center	right
10°	0.000	0.001	1.000
20°	0.019	0.000	0.780
30°	0.009	0.033	0.228
40°	0.001	0.027	0.033
50°	0.005	0.028	0.036
60°	0.009	0.458	0.076
70°	0.036	0.000	0.043
80°	0.000	0.000	0.001
90°	0.000	0.000	0.002

レイ音源分離では，8 本のマイクロフォンの位置の平均を中心として全方位の音源定位を行うため，マイクロフォンは全方位に満遍なく配置されている．また，音源分離の精度を向上させるために，マイクロフォン間の距離ができるだけ長くなるように設置している．

SIG2 と 3 体のスピーカの距離は 2 m で，左 10 度，中央，右 10 度のように 3 方向に設置し，スピーカの間隔を 10 度間隔から 90 度間隔まで 10 度毎に変化させた場合の 9 パターンで録音した．音源分離の際には，事前に音源の到来方向を与えるものとする．Fig. 7 に，実験時の *SIG2* とスピーカの配置の様子を示す．

単語の組み合わせに関しては，発話長や発話内容とは関係なく無作為に組み合わせた．利用した発話のうち，最も発話長の短いもので約 0.4 秒，最も発話長の長いもので約 1.5 秒であった．同時に再生した発話の開始のタイミングはほぼ同時刻とした．しかし，実際の発話は必ずしも同じタイミングで始まるわけではない．また，終了のタイミングは発話長が異なるので組み合わせ毎に異なっていた．

すべての実験において，マイクロフォンアレイ音源分離による三話者同時発話の分離音声を入力とするが，音響モデルと音声認識エンジンが異なっている．音響モデルは 3 状態の HMM で，混合数が 1，2，4，8，16 の場合で試した結果，4 混合の場合で最良の結果が得られたので，実験で利用する音響モデルの混合数は 4 とした．

まず，本手法で利用する音響モデルはクリーン音声で学習した 3 状態 4 混合のトライフォンである．研究用 ATR 日本語音声データベースを用いて学習させた．学習に用いた音声は無響室で録音された合計 22 人の音素バランス単語 216 語の音声で，男性 10 人，女性 12 人のクリーン音声である．

音響モデルをマルチコンディション学習するのに利用した音声は以下の通りである．

(1) クリーン音声（男性 10 人，女性 12 人の合計 22 人分）

(2) 三話者同時発話の分離音声
左が女性話者 f102，正面が男性話者 m103，右が男性話者 m104 の三話者同時発話と，左が女性話者 f103，正面が男性話者 m105，右が男性話者 m106 の三話者同時発話を利用した．スピーカ間隔は 10 度間隔から 90 度間隔までの 9 セットである．

マイクロフォンアレイ音源分離によって分離した音声によって学習した 3 状態 4 混合のトライフォンである．これらの音声すべてを利用して一つの音響モデルを学習した．

テストデータの三話者同時発話を録音するために，女性 f101（左），男性 m101（正面），男性 m102（右）の 3 話者を利用したので，本実験は話者オープンテストである．また，マルチコンディション学習のための学習データは異なる組み合わせの単語による三話者同時発話を利用して，分離音声に特有の歪のみを学習するようにしている．

音声認識エンジンとしては，本手法ではマルチバンド版 Julian を，MFM を利用しない他の手法では通常の Julian を利用した．孤立単語認識では ATR 音素バランス単語 216 語から 200 語を使用した．

5.3 実験結果と考察

三話者同時発話認識結果の単語正解率を Fig. 8 a)–c) に示す．各図は左，中央，右の各方向の単語正解率を示しており，図の横軸は 3 体のスピーカの間隔 (deg.) を，縦軸は単語正解率を表している．

評価項目 (1) について考察する．実験①，②，③の単語正解率を比較すると，すべての場合で本手法により単語正解率が向上している．単語正解率が有意に向上していることを確認するために，実験①と実験③，実験②と実験③をそれぞれ McNemar 検定 [21] により評価した．帰無仮説をそれぞれ「実験①と実験③の単語正解率には差がある」,「実験②と実験③の単語正解率には差がある」としたときの有意確率を Table 1 に示す．有意水準を 0.05 とすると，ほとんどの場合で本手法による単語正解率の向上には有意な差が認められる．

評価項目 (2) について考察する．Fig. 8 の実験③，④，⑤の結果から提案手法はマルチコンディション学習と同程度の性能が得られていることが分かる．提案手法は事前に分離音特有の歪を学習しなくても良いので，マルチコンディション学習による手法よりも事前情報が少なくて済むという点で有利である．また，提案手法とマルチコンディション学習は対立する手法ではなく，両方同時に利用することもできる．実験⑥の結果の通り，本手法とマルチコンディション学習をそれぞれ単独で利用した場合よりも組み合わせた手法の方が良い結果が得られている．この結果から，両手法が互いの欠点を補い合うことができることが分かる．

評価項目 (3) について考察する．Fig. 8 の実験⑦の結果を見

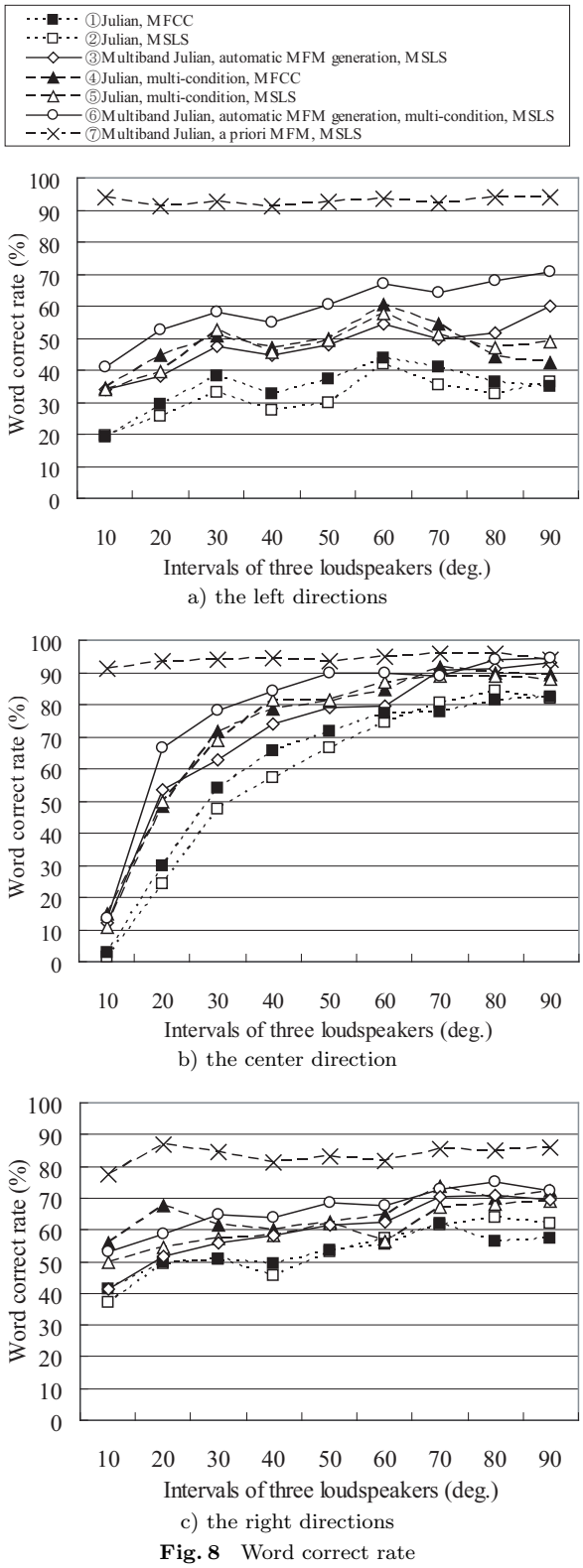

Fig. 8 Word correct rate

ると，実験③の結果よりも高い性能が得られている．つまり，本システムの MFM 自動生成を改良することでさらに性能を改善することができる可能性があることを示している．

評価項目以外について実験結果から分かったことについて考察する．各方向の単語正解率の違いに着目すると，中央の話者の場合スピーカの間隔が大きいほど単語正解率が高く，間隔が狭くなるにつれて単語正解率は下がっている．右方向の場合にも同様の傾向が見られる．しかし，左方向と右方向では，スピーカの間隔と単語正解率には大きな相関は見られなかった．このように，スピーカの間隔によって単語正解率が変化するのは，マイクロフォンアレイ音源分離の分離精度と関係があると考えられ，中央の話者の場合特にその傾向が強い．また，左右の方向で単語正解率が複雑に変動するのは，マイクロフォンアレイの内部が自由空間ではなく *SIG2* のボディがあることが要因の一つであると考えられる．

6. お わ り に

本稿では，実環境において音声に音声雑音が混在する混合音を扱うためには，混合音声の音源分離と音声認識の間のインタフェースが重要であることに着目して，MFT を応用した混合音認識システムを設計した．また，同時発話の分離音声の歪みの原因が対象音源以外の音源に由来することに注目し，音源分離から得られる情報を利用して MFM 自動生成手法を提案した．

評価実験では，分離音に対し通常の音声認識を行うよりも，自動生成した MFM を利用したほうが，三話者同時発話の孤立単語認識の単語正解率が向上した．結果として，本手法により単一音響モデルによる音声認識の有効範囲が広がり，環境変化へのロバスト性が増した．

本論文では，実環境での音声認識を目標とし，複数話者の同時発話認識を扱った．今後の課題として，実環境での音声認識の実現のために，話者数の同定，話者の定位，音声の同定（音声であるかどうかの判定），音声区間の検出などの様々な課題を解決する必要がある．また，実環境での音声認識を利用したヒューマン・ロボット・コミュニケーションのために必要とされる音声認識の精度を達成する必要がある．

謝辞 本研究の一部は，科学研究費補助金（基盤研究（A），特定領域研究「情報学」，特別研究員）および 21 世紀 COE の援助を受けた．また，マルチバンド版 Julian の利用を許可していただいた，東京工業大学の古井研究室と東京大学の西村義隆氏に感謝いたします．また，本研究に関して議論や助言を頂いた，京都大学奥乃研究室，および（株）ホンダ・リサーチ・インスティチュート・ジャパンの皆様，日東紡音響エンジニアリング（株）の中島弘史氏に感謝いたします．また，本論文の構成について適切なご助言をいただいた匿名の査読者に感謝いたします．

参 考 文 献

[1] C. Breazeal. Emotive qualities in robot speech. In *Proc. IROS-2001*, pp. 1389–1394. IEEE.
[2] AURORA. http://www.elda.fr/proj/aurora1.html" "http://www.elda.fr/proj/aurora2.html.
[3] D. Pearce. Developing the ETSI AURORA advanced distributed speech recognition front-end & what next. In *Proc. of Eurospeech-2001*. ESCA, 2001.
[4] R. P. Lippmann, E. A. Martin, and D. B. Paul. Multi-styletraining for robust isolated-word speech recognition. In *Proc. of ICASSP-87*, pp. 705–708. IEEE, 1987.

[5] M. Blanchet, J. Boudy, and P. Lockwood. Environmentadaptation for speech recognition in noise. In *Proc. of EUSIPCO-92*, Vol. VI, pp. 391–394, 1992.
[6] K. Nakadai, D. Matasuura, H. G. Okuno, and H. Tsujino. Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots. *Speech Communication*, Vol. 44, No. 1-4, pp. 97–112, October 2004.
[7] 澤田知寛, 関矢俊介, 小川哲司, 小林哲則. 階層的音源分離に基づく混合音声の認識. AI チャレンジ研究会（第 18 回）, pp. 27–32, 2003.
[8] 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 Eurospeech-2001*, pp. 213–216. ESCA, 2001.
[9] P. Renevey, R. Vetter, and J. Kraus. Robust speech recognition using missing feature theory and vector quantization. In *Proc. of Eurospeech-2001*, pp. 1107–1110. ESCA, 2001.
[10] K. Palomaki, G.J. Brown, and J. Barker. Missing data speech recognition in reverberant conditions. In *Proc. of ICASSP-2002*, pp. 65–68. IEEE, 2002.
[11] 山本俊一, 中臺一博, 辻野広司, 奥乃博. ミッシングフィーチャー理論による音源分離と音声認識のインターフェースと複数ロボットへの適用. 日本ロボット学会誌, Vol. 23, No. 6, pp. 743–751, 2005.
[12] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat. Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach. In *Proc. of ICRA 2004*, pp. 1033–1038. IEEE.
[13] 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音声認識の検討. 日本音響学会 2004 年春季研究発表会講演論文集, pp. 117–118. 日本音響学会, 2004.
[14] T. Kawahara and A. Lee. Free software toolkit for japanese large vocabulary continuous speech recognition. In *International Conference on Spoken Language Processing (ICSLP)*, Vol. 4, pp. 476–479, 2000.
[15] L. C. Parra and C. V. Alvino. Geometric source separation: Mergin convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 6, pp. 352–362, 2002.
[16] Y. Ephraim and D. Malah. Speech enhancement using minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-32, No. 6, pp. 1109–1121, 1984.
[17] Y. Ephraim and D. Malah. Speech enhancement using minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-33, No. 2, pp. 443–445, 1985.
[18] I. Cohen and B. Berdugo. Microphone array post-filtering for non-stationary noise suppression. In *ICASSP-2002*, pp. 901–904, 2002.
[19] J.-M. Valin, J. Rouat, and F. Michaud. Enhanced robot audition based on microphone array source separation with post-filter. In *Proc. of IROS 2004*, pp. 2123–2128. IEEE, 2004.
[20] I. Cohen and B. Berdugo. Speech enhancement for non-stationary noise environments. *Signal Processing*, Vol. 81, No. 2, pp. 2403–2418, 2001.
[21] L. Gillick and S. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proc. of ICASSP-89*, pp. 532–535. IEEE, 1989.

山本 俊一 (Shunichi Yamamoto)

1981 年生．2003 年京都大学工学部情報学科卒業．2005 年 3 月京都大学大学院情報学研究科知能情報学専攻修士課程修了．同年 4 月同博士課程進学．日本学術振興会特別研究員．主にロボット聴覚の研究に従事．IEEE Robotics and Automation Society Japan Chapter Young Award，IEEE 関西支部学生研究奨励賞受賞．情報処理学会，人工知能学会, IEEE 各会員．（日本ロボット学会学生会員）

Jean-Marc Valin

中臺 一博 (Kazuhiro Nakadai)

1970 年生．1993 年東京大学工学部電気工学科卒，1995 年 同大学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，科学技術振興事業団（現：科学技術振興事業機構） ERATO 北野共生システムプロジェクト研究員．2003 年 5 月より，(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）．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．2001 年度日本人工知能学会研究奨励賞，IROS 2001 BEST Paper Nomination Finalist，2002 年電気通信普及財団賞奨励賞，第 2 回船井情報科学振興賞等受賞など受賞．本学会ロボット聴覚研究専門委員会幹事．日本人工知能学会，日本音響学会，AAAI，IEEE 各会員．（日本ロボット学会正会員）

中野 幹生 (Mikio Nakano)

（日本ロボット学会正会員）

辻野 広司 (Hiroshi Tsujino)

（日本ロボット学会正会員）

駒谷 和範 (Kazunori Komatani)

尾形 哲也 (Tetsuya Ogata)

（日本ロボット学会正会員）

奥乃 博 (Hiroshi G. Okuno)

1972 年東京大学教養学部基礎科学科卒業．日本電信電話公社, NTT, 科学技術振興事業団北野共生システムプロジェクト, 東京理科大学理工学部情報科学科を経て, 2001 年 4 月より京都大学大学院情報学研究科知能情報学専攻 教授．博士 (工学)．この間, スタンフォード大学客員研究員, 東京大学工学部客員助教授．人工知能, 音環境理解, ロボット聴覚の研究に従事．1990 年度人工知能学会論文賞, IEA/AIE-2001 最優秀論文賞, IEEE/RSJ

IROS-2001 Best Paper Nomination Finalist, 第 2 回船井情報科学振興賞等受賞. 情報処理学会, 人工知能学会, 日本ソフトウェア科学会, 日本認知科学会, ACM, AAAI, IEEE, ASA 各会員. 本学会評議員. 本学会ロボット聴覚研究専門委員会委員長. 著編書:『インターネット活用術』(岩波書店),『*Computational Auditory Scene Analysis*』(共編, LEA),『*Advanced Lisp Technology*』(共編, Taylor & Francis) 他. （日本ロボット学会正会員）