4R-4

# マイクロフォンアレイによる分離音声認識のための ミッシングフィーチャーマスク自動生成

**山本 俊一** † **Jean-Marc Valin**‡ **中臺 一博** * **駒谷 和範** † **尾形 哲也** † **奥乃 博** †

† 京都大学情報学研究科知能情報学専攻 ‡Université de Sherbrooke

* (株) ホンダ・リサーチ・インスティチュート・ジャパン

## 1. はじめに

ロボットと人間が自然なインタラクションを行う上で，音声による対話は重要な機能の一つである．実環境では，通常，単一音源からの音ではなく，複数の音が混在した混合音が聞こえるので，ロボット聴覚は混合音を扱える必要がある．我々は，これまでに，2 本のマイクを用いた混合音声分離，および先見的情報を使った**ミッシングフィーチャー理論（MFT）**による分離音声認識を実装・評価した [1]．本稿では，クリーン音声などの先見的情報を与えず，マイクロフォンアレイによる音源分離処理から得られるデータのみを利用したミッシングフィーチャーマスク（MFM）の自動生成手法を報告する．

## 2. 混合音認識システム

混合音声認識システムは以下の 4 つのシステムから構成されている（図 1）．

(1) 幾何学的音源分離（*Geometric Source Separation*, GSS）の一種として実装されているビームフォーマ
(2) 多チャンネル post-filter
(3) MFM の計算
(4) MFM を利用した MFT による分離音声認識

マイクロフォンアレイ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で構成されており，steered beamformer による音源定位 [2] を行う．

音源分離は，GSS に基づく線形音源分離法を用い，さらに，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している [3].

多チャンネル post-filter[3] は，GSS の post-filter 処理 [4] を複数音源を扱えるように拡張した手法である．この手法では，雑音を定常性雑音と非定常性雑音に分けて推定することにより，目的音源の強調を行っている．

多チャンネル post-filter は分離音における干渉音を抑制するだけでなく，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．そこで，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 を自動生成する．

MFT に基づく音声認識では，通常の音声認識とは隠れマルコフモデルにおける出力確率の計算方法が異なり，次のように定義される．特徴ベクトル $x$，状態 $S$ の時の正規分布の確率密度関数を $f(x|S)$，$L$ を混合正規分布の混合数，$P(k|S)$ を混合係数，$N$ を特徴量の次元数とすると，マスクされたときの出力確率 $o(x|S)$ は次の式で求める．

$$o(x|S) = \sum_{l=1}^{L} P(l|S) \exp\left\{ \sum_{i=1}^{N} M(i) \log f(x_i|l,S) \right\}$$

Automatic Missing Feature Mask Generation for Automatic Recognition of Speech Separated by Microphone Array
*by Shunichi Yamamoto (Kyoto Univ.), Jean-Marc Valin (Univ. de Sherbrooke), Kazuhiro Nakadai, Hiroshi Tsujino (HRI-JP), Kazunori Komatani, Tetsuya Ogata and Hiroshi G. Okuno (Kyoto Univ.)*

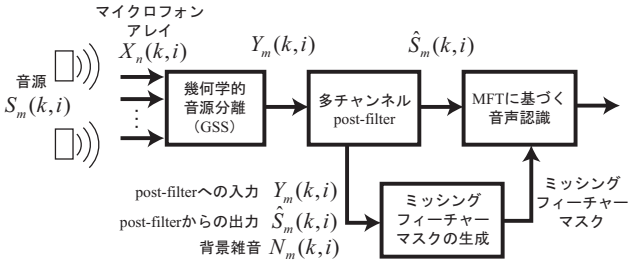



図 1: システムの概要

こうして，信頼できない特徴による影響を除去することができる．これらの実装には，CASA Toolkit を利用した．

## 3. 音声認識特徴量の設計

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，一般の音声認識でよく使われるメル周波数ケプストラム係数（MFCC）ではなく，スペクトル特徴量を用いる．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，周波数領域の特徴量を利用することにより，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．

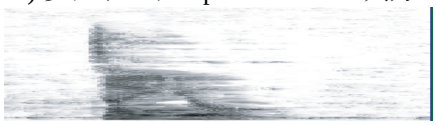
以下に，MFCC で行われるのと同等の正規化を行ったメル周波数領域対数スペクトルの導出の手順を示す．

(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，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する．
(5) ケプストラム係数の 0, 13-23 次の項を 0 にする．
(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う．
(8) 各次元毎に一次微分を計算する．
(9) 微分値と合わせて，計 48 次元の特徴量として抽出する．

## 4. ミッシングフィーチャーマスク自動生成

MFM 自動生成には，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．先見的情報は仮定せずに， 音源分離処理から得られるデータのうち，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．MFM のうち，微分値でない特徴量（$i = 1, \cdots, \frac{N}{2}$）に

a) 多チャンネル post-filter への入力



b) 多チャンネル post-filter からの出力

c) 自動生成された MFM（白 = 1，黒 = 0）

図 2: MFM 自動生成

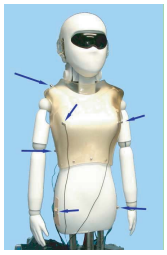
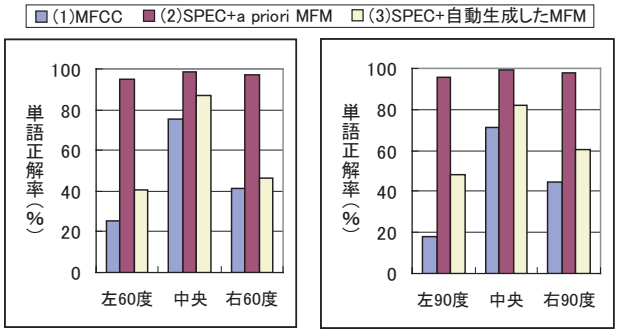


図 3: SIG2 装着された 8 本のマイクロフォン（2 本は見えない）




a) 話者間隔 60 度　　b) 話者間隔 90 度

図 4: 三話者同時発話認識結果の単語正解率

対応するマスク $M(k,i)$ は，メル周波数帯域のフレーム $k$ における多チャンネル post-filter の入力を $Y(k,i)$，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 $N(k,i)$ とした場合，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．また，閾値 $T$ は実験的に求め，0.3 とした．

$$M(k,i) = \begin{cases} 1, & \{\hat{S}(k,i)+N(k,i)\}/Y(k,i) > T \\ 0, & otherwise \end{cases}$$

このように，推定された背景雑音を利用するのは，背景雑音が大部分を占める周波数帯域は信頼度が高くなるようにするためである．つまり音声認識から見ると，背景雑音しか存在しなかった周波数帯域は，無音であることが「信頼できる」領域であるとするということである．

また，MFM のうち，特徴量の一次微分（$i = \frac{N}{2} + 1, \cdots, N$）に対するマスク $M(k,i)$ 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．

$$M(k,i) = M(k-2,j)M(k-1,j)M(k+1,j)M(k+2,j)$$

ここで，$j = i - \frac{N}{2}$ である．特徴量とその一次微分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 $N = 48$ となる．生成された MFM の例を図 2 に示す．

## 5. 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（図 3），三話者同時発話認識実験を行った．3 本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て孤立単語認識実験を行った．実験を行った部屋は 5 m × 4 m の大きさで，残響時間は 0.3 - 0.4 秒 ($RT_{20}$) である．SIG2 とスピーカの距離は 2 m で，左 60 度，中央，右 60 度の場合と左 90 度，中央，右 90 度の場合で録音した．孤立単語認識の語彙サイズは 200 語である．

音響モデルはクリーン音声で学習したトライフォンを利用した．学習データには，合計 25 人の男女の ATR 音素バランス単語 216 語の音声を利用し，3 状態 8 混合の HMM を構築した．

比較のために，GSS と post-filter 処理を行った分離音声に対して以下の 3 通りの音声認識実験を行った．

(1) MFCC による従来の音声認識

(2) スペクトル特徴量と a priori MFM による音声認識

(3) スペクトル特徴量と自動生成した MFM による音声認識

ここで，a priori MFM とはクリーン音声との比較によって得られる理想的な MFM である．単語正解率を図 4 に示す．分離音声を自動生成した MFM を利用して音声認識した場合，分離音声に対して従来の音声認識を行うよりも単語正解率が向上しており，この結果は，post-filter の情報から生成した MFM が分離音声認識に有効であることを表している．また，a priori MFM の単語正解率が高く，MFM 自動生成手法を改良することにより更なる精度向上が期待できる．方向ごとに比較すると，中央が最もよく，左右の単語正解率は中央よりも低くなった．これは，3 方向の音声の再生音量の違いにより，各方向の分離音声の S/N 比が異なることが原因の一つであると考えられる．

## 6. おわりに

本稿では分離音認識に注目し，マイクロフォンアレイによる音源分離から得られる情報を利用して分離音に適した MFM の自動生成手法を報告した．その結果，分離音に対しそのまま通常の音声認識を行うよりも，自動生成した MFM を利用することで，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．今後の予定として，周辺方向の話者の音声認識率の改善，Julius/Julian の MFT への対応による音声認識の高速化を通じたシステム全体の実時間処理実現が挙げられる．なお，本研究の一部は，科学研究費補助金，21 世紀 COE プログラム，および，SCAT 研究助成の支援を受けた．

## 参考文献

[1] Yamamoto, S. *et al.*: Assessment of General Applicability of Robot Audition System by Recognizing Three Simultaneous Speeches, *IEEE/RSJ IROS 2004*, pp. 2111–2116.

[2] Valin, J.-M. *et al.*: Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach, *IEEE ICRA 2004*, pp. 1033–1038.

[3] Valin, J.-M. *et al.*: Enhanced robot audition based on microphone array source separation with post-filter, *IEEE IROS 2004*, pp. 2123–2128.

[4] Cohen, I. *et al.*: Microphone Array Post-Filtering for Non-Stationary Noise Suppression, *ICASSP-2002*, pp. 901–904.