特徴量レベルでの統合法に基づくマルチバンド型モデルによる雑音環境下音声認識

大川茂樹[†] 白井克彦[††]

本論文では，マルチバンド型と呼ばれる音声認識モデルにおいて，部分周波数帯域の情報を統合する新しい方法を提案する．マルチバンド型音声認識は，入力音声信号を複数の周波数帯域に分割し，各帯域を独立に処理した後に再統合するという認識手法であり，近年，特に雑音環境下で良い性能を与えることが示されている．このようなモデル化では，(i) 再統合の方法，(ii) 帯域分割の方法について検討する必要がある．本研究では，まず (i) の問題に対して，すでに Bourlard らにより提案されている HMM の尤度レベルの再統合に基づく方法（LC 法）の追試を行うとともに，新たに音響特徴量レベルの再統合法（FC 法）を提案し，比較評価を行う．また，(ii) の問題に対しては，FC 法において部分帯域の特徴量と音素モデルとの相互情報量を評価基準とした分割の最適化を試みる．偏帯域性雑音が付加された音声を用いた実験の結果，FC 法に基づくシステムは，従来型システムおよび LC 法に基づくシステムの双方に対してより高い認識性能を与えた．また，情報量を基準とした分割点最適化の効果が確認された．

Noisy Speech Recognition by Multi-band Modeling Based on Feature-level Combination

Shigeki Okawa[†] and Katsuhiko Shirai[††]

This paper presents a new approach for sub-band recombination in the framework of multi-band ASR. Recent works suggest that multi-band ASR, which is based on independent processing and recombination of partial frequency bands of input speech, gives more accurate recognition, especially in noisy acoustic environments. In the case, we need to discuss (i) how to recombine the sub-band output, and (ii) how to split the input speech frequencies. We propose and evaluate “feature combination” (FC) approach, as a solution of the above point (i), instead of “likelihood combination” (LC) approach proposed by Bourlard et al. Also for the point (ii), we introduce the mutual information between sub-band features and target phoneme categories to find the optimal splitting frequencies. The experimental results show that the FC-based system can yield better performance both the conventional ASR and the LC-based system for band-limited noisy speech. Also, we could obtain a favorable band-splitting strategy by using the optimization method.

1. はじめに

音声認識技術の実用化において，頑健性すなわち雑音や環境の変動に対する耐性は重要な問題である[1)]．従来より，雑音環境下などにおける音声認識の頑健性を論じた研究は数多く行われており，様々な知見や指針を得ることができる．たとえば，入力音声中の雑音成分を推定して周波数軸上で差し引く方法[2)]や，雑音自体をモデル化する方法[3)]などが代表的なものである．

ところで，伝統的な音声認識では，入力音声の全周波数帯域から計算された音響特徴量（ケプストラムなど）をベクトル特徴として用いることが多い．この場合，もしも周波数帯域の一部が実環境下で多く見られるような偏帯域性雑音☆ などにより汚損されていると，ベクトル特徴のすべての要素が少なからず影響を受けてしまう．この問題を回避するため，近年，複数の部分周波数帯域の特徴量をそれぞれ独立にモデル化するマルチバンド（複合周波数帯域；Multi-band）型の音声認識手法（以下 MB-ASR と記すことがある）が提

† 千葉工業大学情報科学部情報ネットワーク学科
Department of Information and Network Science, Chiba Institute of Technology
†† 早稲田大学理工学部情報学科
Department of Information and Computer Science, Waseda University

☆ 特定の周波数帯域にエネルギー成分が集中した雑音．自動車のエンジン音や計算機のファンの音などもこの性質を持つ．

案され，雑音環境下あるいは不整合な環境下において良好な頑健性を示している[4)~8)]．

本論文では，MB-ASR の枠組みにおいて，部分周波数帯域から得られた情報を特徴量のレベルで統合する方法を提案し，HMM の尤度レベルでの統合法との比較および種々の雑音環境下における有効性を検証する．また，情報量を評価基準とした帯域分割点の最適化について言及する．

以下，2 章では MB-ASR の基本的な考え方とその利点を説明し，3 章では 2 種類の再統合レベルについて検討する．また，4 章では帯域分割の最適化の手法を示す．5 章では実験におけるデータや分析の条件を述べ，6 章で音声認識実験の結果を示し考察する．最後に 7 章で結論を述べる．

2. マルチバンド型音声認識

2.1 基本的な考え方

マルチバンド型音声認識（MB-ASR）は，入力音声信号の周波数帯域全体を複数の部分に分割し，それぞれ独立に分析や音響尤度計算の処理を施し，後段のどこかの時点で個々の出力（尤度など）を再統合するという考え方に基づいている．前章で触れた耐雑音性のみならず，以下で述べるように人間の聴覚機構にも通ずるアプローチであることから，音声認識の新しいパラダイムとして注目され始めている．

図 1 に，MB-ASR の基本的な考え方を示す．図中で，入力音声の低い周波数帯域が局所的に雑音で汚損されている．このとき，(a) 従来型の認識手法では，生成される特徴量ベクトルの全体に局所雑音の影響が及んでしまうのに対して，(b) マルチバンド型モデルでは，低い帯域に対応するベクトル要素のみに雑音の影響を「封じ込め」，他の帯域の情報損失を軽減できる．

2.2 研究の意義

MB-ASR を積極的に追求する意義として，以下にいくつかの根拠をあげる．

- マルチストリーム音声処理：
  複数の情報源から得られた観測あるいは尤度を統合することで，より安定した情報抽出が可能になるという「マルチストリーム音声処理」（Multi-stream ASR；多元情報に基づく音声処理）の考え方に包含される[9)]．
- 偏帯域性雑音：
  我々の身の周りに存在する雑音は，一部の周波数帯域に成分が集中している（偏帯域性である）ことが多い．上で述べたように，部分帯域からの出力を再統合する際に，雑音を含む帯域の影響を低減させる仕組みを導入することが可能である．
- 心理音響的な裏付け：
  人間の聴覚機構は，周波数帯域をいくつかの狭い区間に分けてそれぞれ独立に処理しているといわれている[10)]．
- 「次元の呪い」の回避：
  分析する帯域が広いほど特徴量空間の次数が増大し，いわゆる「次元の呪い（curse of dimensionality）」問題をもたらす．各部分帯域の特徴量空間を高い次元で表現すれば，同じ計算量で全帯域型よりも精密なモデル化が行える可能性がある．

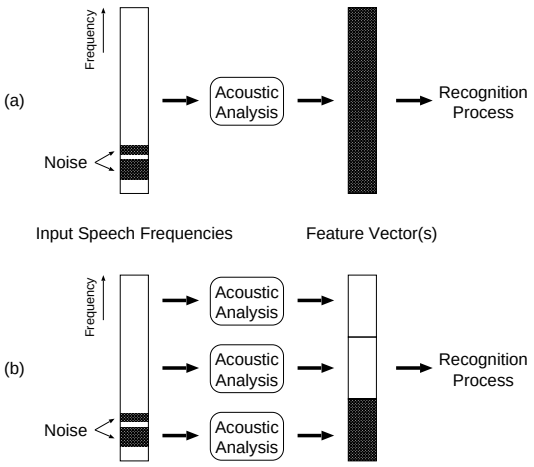

図 1 (a) 全帯域型（従来型）音声認識と (b) マルチバンド型音声認識の考え方
Fig. 1 Schematic diagrams of (a) full-band (conventional) ASR and (b) multi-band ASR.

他にも，異なる情報源を取り扱うマルチストリーム処理に比べ信号処理がしやすい，計算機への実装時に並列処理がしやすいなどの利点がある．

2.3 検討すべき問題

MB-ASR の実現に際しては，従来型の音声認識における様々な条件に加えて，次の 2 つの主要な検討事項がある．

(1) 各部分帯域の処理結果より最終的な決定（認識）をする際,「どの時点」で「どのように」再統合すればよいかという問題．
(2) 部分帯域を定義する際,「いくつの帯域」を「どこで」分割して用いるかという問題．

本論文では，まず (1) の問題に対して，特徴量を生成する時点で再統合する方法を提案・検討し，次に (2) の問題に対して，情報量を評価基準とした分割数および分割点の最適化を試みる．

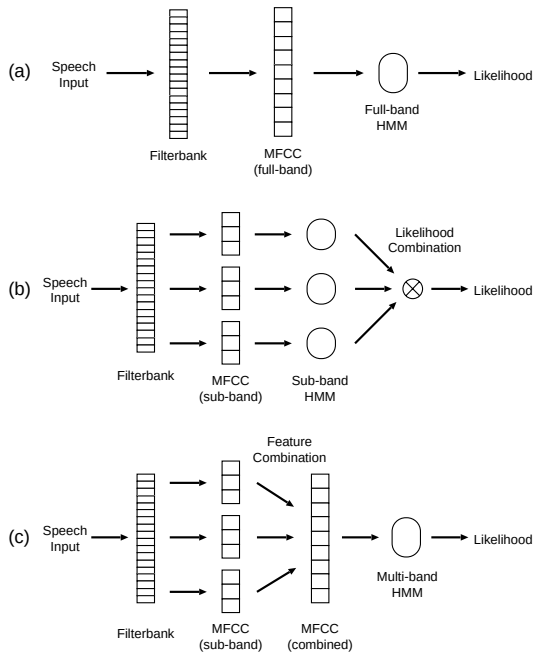

図 2 (a) 全帯域システム（従来型；FB-ASR），(b) LC 法に基づく MB-ASR，(c) FC 法に基づく MB-ASR のデータの流れ．
Fig. 2 Diagrams of (a) full-band (conventional) ASR, (b) MB-ASR (likelihood combination), and (c) MB-ASR (feature combination).

3. 再統合レベルの検討

MB-ASR の原型は，1996 年に Bourlard らと Hermansky らによってほぼ同時に提案された[4),5)]．彼らの研究では，各部分帯域に対して異なる識別器が適用された後，いくつかの統合時点（HMM の状態・音素・単語など）において尤度の統合が試みられている．

本研究では，Bourlard らや Hermansky らの方法（尤度レベルの統合法；以下「LC 法」と呼ぶ）の追試に加え，異なる統合法を導入し，種々の雑音条件下における比較を試みる．新たな方法は，部分帯域から得られた特徴量を識別処理の前に連結し，単一の特徴ベクトルとして用いるものである（特徴量レベルの統合法；以下「FC 法」と呼ぶ）．**図 2** に，(a) 全帯域型（従来型）の音声認識手法（Full-band ASR；以下 FB-ASR），(b)LC 法に基づく MB-ASR，(c) FC 法に基づく MB-ASR におけるデータの流れを示す．

本章では，LC 法および FC 法の概念および特長をそれぞれ説明する．

3.1 尤度レベルの統合法（LC 法）

まず，LC 法では，各部分帯域から得られた特徴量ベクトルを用いて，あらかじめ独立に HMM を学習する．認識時にも，各帯域に対してそれぞれ異なる識別器を独立に適用し，各識別器から認識仮説（音素や単語の候補）とその尤度を得る．その後，すべての識別器からの出力を統合して全体の尤度を計算することで認識結果が得られる（図 2 (b) 参照）．

各帯域の尤度は，それぞれの帯域に対応する HMM により計算・出力される．帯域間の独立性を仮定すれば，すべての出力確率を単純に掛け合わせることで全体の尤度が得られる．

この際，不要と判断された帯域を使用しない方法[11)]や，帯域ごとに何らかの重み係数を設定し，$p(o_i^b|s_j)$ を HMM の出力確率（状態 s_j，時刻 i，帯域 b），w_b を帯域 b に対する重み係数として，

$$p(o_i|s_j) = \prod_b p(o_i^b|s_j)^{w_b}, \tag{1}$$

のように重み付け確率計算を行う方法[12)]なども考えられるが，本論文では，MB-ASR において部分帯域への分割と独立処理そのものが持つ効果に注目するため，重み付けは行わずに議論を進める．すなわち，LC 法における確率計算は，

$$p(o_i|s_j) = \prod_b p(o_i^b|s_j), \tag{2}$$

のように行う．

なお，Bourlard らの報告[4)]によると，尤度レベルで統合を行う単位（尤度の統合計算を行う時点）として，HMM の状態（=分析フレーム）を用いても，言語的にさらに長い単位（たとえば音素や単語）を用いても，ほぼ同等の性能が得られるとされている．HMM の状態を単位とした統合は，明らかに実装が簡単であるので，ここでは LC 法での統合の単位には HMM の状態を用いるものとする．

3.2 特徴量レベルの統合法（FC 法）

次に，本研究で新たに提案する FC 法について述べる．

LC 法では，各部分帯域から得られた特徴量に対してそれぞれ尤度を計算した後で統合したが，FC 法では，各帯域の特徴量を得た直後に（尤度計算以前に）それらを統合する（図 2 (c) 参照）．

具体的には，適当な数に分割した部分帯域の各々について，音響特徴量を独立に計算した後，連結して 1 つのベクトルを構成し，識別器への入力として用いる．LC 法での式 (2) に対して，FC 法での出力確率は，

$$p(o_i|s_j) = p(o_i^1, \ldots, o_i^B|s_j), \tag{3}$$

（B は帯域数）

と表すことができる．この場合，識別器以降の処理は，基本的に従来型の音声認識と同等になるため，認識システムへの実装は比較的容易である．

このFC法は，MB-ASRの一般的な利点とFB-ASR（従来型の音声認識）の利点との両者をあわせ持つと考えられる．すなわち，

- 一部の帯域が汚損されている場合，その影響は，分割の結果として特徴量ベクトルの一部の要素のみに影響し，他の帯域の情報は極力損失させないように作用するであろう（MB-ASRの利点）．
- 特徴ベクトルのすべての要素が統計モデル内で一括表現されるので，複雑な再統合機構を必要としない．また，HMMの出力確率計算時には，一般的な特徴ベクトルと同様に全相関あるいは混合確率密度分布を用いることにより，各帯域間の相関を考慮することが容易である（FB-ASRの利点）．

ただし，FC法では，音素や単語などを単位とした統合は不可能である．本研究では，比較するLC法においても，前述のようにBourlardらの実験結果をふまえてHMMの状態（=分析フレーム）を統合の単位とするが，仮に音素や単語などの単位での統合がより良い結果を与えるような条件が存在する場合，それはLC法においてのみ可能であることを付記しておく．

なお，FC法においても，各帯域が音声認識に対して持つ情報の量に応じてベクトル構成時に重み付けを行う方法が考えられるが，本研究では，LC法の場合と同様の理由により，特に重み付けなどの処理は施さず，各帯域から得られた特徴ベクトルを単純に連結してHMMへの入力として用いている．

4. 情報量に基づく帯域分割の最適化

マルチバンド型モデルを構成する際，どのように周波数帯域を分割するかは重要であるが複雑な問題でもある．特に，音声信号に偏帯域性雑音が含まれるような場合，雑音の性質により最適な分割点は変動するので，入力音声に応じた最適化が必要となる．雑音なし音声の場合も，音声のスペクトル特徴自体が持つ周波数構造（たとえば母音のホルマント位置との関連など）により，帯域により分割の影響が異なると考えられる．

スペクトルの重要な成分のみを選択的に用いて音声認識を行う方法としては文献13)などがあるが，MB-ASRの枠組みの中で，入力に応じた分割点の最適化に言及した研究はほとんどない．このことから，本論文では，音声および雑音の双方の性質をふまえた帯域分割最適化の提案として，FC法において，あらかじめ設定した複数の帯域分割条件に対し，特徴ベクトルと音素HMMの全状態との相互情報量が最大となるものを逐次選択する方法を導入する．

この方法では，全帯域の特徴量ベクトルをそのまま用いる場合も選択候補に含めることができるため，従来型の音声認識が最も良い性能を与える条件でも最適なモデルが選択できるという利点がある．

部分帯域が音声認識に対して与える情報の量は，観測系列の条件付きエントロピーを用いると都合良く定義できる．ただし，条件付きエントロピーは，各帯域間での相対値として扱う必要があるので，事後確率により計算しなければならない．ここで，部分帯域 b に対する時刻（フレーム）i のHMM出力確率を $p(o_i^b|s_j)$ とすると，Bayesの定理により事後確率は，

$$p(s_j|o_i^b) = \frac{p(s_j)p(o_i^b|s_j)}{p(o_i^b)}, \tag{4}$$

となる．ここで，$p(o_i^b)$ は $p(s_j|o_i^b)$ の大小には無関係であり，また言語確率 $p(s_j)$ は，単独の確率変量としては意味を持つものの，部分帯域間の比較においては同じ条件（本論文ではHMMの状態遷移を帯域ごとに変えるような機構は用いていない）であるため，次の近似式により事後確率を計算する．

$$p(s_j|o_i^b) \cong \frac{p(o_i^b|s_j)}{\sum_j p(o_i^b|s_j)}. \tag{5}$$

すべての音素カテゴリおよびHMM状態に対応する条件付きエントロピー $H(S|o_i^b)$ は，観測系列の特徴量ベクトル o_i^b が与えられたとき，次のようになる．

$$H(S|o_i^b) = \sum_j -p(s_j|o_i^b)\log p(s_j|o_i^b). \tag{6}$$

さらに，適当な時間長の複数フレームに対する平均エントロピー $H(S|O^b)$ は次のようになる．

$$H(S|O^b) = \sum_i \sum_j -p(s_j, o_i^b)\log p(s_j|o_i^b). \tag{7}$$

この値は，系列 $O^b = \{o_i^b\}$ が観測されたときに，最も起こりうるHMMの状態を決定する際の曖昧さ（情報量の減少の度合い）を意味する．

同様に，HMM状態の集合 S の自己エントロピーは，事前知識（たとえば言語確率など）に基づき次のように O^b に無関係な定数として定義できる．

$$H(S) = \sum_j -p(s_j)\log p(s_j). \tag{8}$$

こうして得られた $H(S|O^b)$ と $H(S)$ により，状態の集合 S と適当な時間長の観測ベクトルの集合 O^b との間の相互情報量 $I(S;O^b)$ が次式で定義できる☆．

☆ ここで，$H(S)$ は O^b に無関係であるが，$I(S;O^b)$ の計算に必要である．本研究では，学習データ中の音素数から実際に $H(S)$ を計算した4.977という値を用いた．

$$I(S;O^b) = H(S) - H(S|O^b). \qquad (9)$$

この値は，音声認識に対して帯域 b の観測 O^b が持つ情報量を表す．本研究では，様々な分割点または分割点の組合せに対して $\sum_b I(S;O^b)$ を計算したとき，最も大きな値を与える帯域数および分割点を選択し，帯域分割の最適化を行う．

5. 実験条件

本章では，実験に用いたデータや音響分析の条件，認識モデルや帯域分割の条件などについて説明する．

5.1 音声・雑音データ

音声データには，ATIS（Air Travel Information Service）連続音声認識タスク（英語）を用いた[14]．音声は，実験室環境で接話型マイクロフォンを用いて録音されたものである．学習には，19,507 文章（話者 528 名）を用いて，評価には，別の 981 文章（話者 24 名）を用いた．

付加雑音としては，まず各帯域数の条件で最初の（最も低い）周波数帯域のみに白色雑音を加えた 'LPW' 雑音を定義し生成した．これは，いわば MB-ASR における「理想的な」雑音である．帯域を制限した白色雑音を生成する処理は，計算機上で FIR フィルタを設計して行った．次に，実雑音として，NOISEX-92 データベース[15]に含まれる 15 種類の雑音を計算機上でそれぞれ音声データに加えた．いずれの場合も，評価用データには雑音を加えるが，学習用データには加えていない．すなわち，HMM はつねに雑音を含まない条件下で学習される．

5.2 音響分析と認識モデル

認識のフロントエンドとなる音響分析は，フィルタバンク分析および離散コサイン変換（DCT）に基づく MFCC（Mel-Filterbank Cepstrum Coefficients）分析を用いた．

まず 16 kHz で標本化された入力音声をフレーム長 20 ms，分析周期 10 ms のハミング窓で切り出し，FFT 計算によりメルフィルタバンク係数を得る．これを帯域ごとの DCT によりメルケプストラム係数（MFCC）に変換する．この際，各入力に対して，全フレームの MFCC ベクトルから 1 入力あたりの平均ベクトルを差し引く CMS（Cepstrum Mean Subtraction）法を適用している．

認識器は，4 状態 3 ループ，16 混合の文脈独立音素 HMM（音素数 49）とし，言語モデルには単語バイグラムのみを用いる．認識タスクは，連続単語認識であり，以下の結果は単語誤り率で示す．

5.3 帯域分割と特徴ベクトルの構成

フィルタバンク分析は，メル周波数軸上で 31 個の三角窓を等間隔に配置して実現する．比較実験のために設計する全帯域システム（従来型の音声認識手法に相当）では，この 31 次フィルタバンクのすべての次元に対して DCT を適用して MFCC 特徴量を得る．マルチバンド型システムでは，あらかじめ定義した条件（後述）により 31 次フィルタバンクをいくつかの帯域に分割し，それぞれ独立に DCT を適用して複数の MFCC 特徴量を得る（図 2 (b)，(c) 参照）．

帯域数として，当初 1（全帯域），2，3，4 および 6 の場合について実験を行ったが，このうち 4 と 6 の場合は，予備実験の結果，本研究の条件下では十分な性能が得られないことが確認されたため，以下の実験は 1，2 および 3 について行っている．

ちなみに，各帯域数においてメル周波数軸上で等間隔に帯域分割した場合，分割点に相当する周波数（各フィルタに対応する三角窓の中心周波数）は次のようになる（単位は Hz）．

- 2 帯域：0-1772-8000
- 3 帯域：0-1100-2853-8000
- 4 帯域：0-900-1772-3452-8000
- 6 帯域：0-600-1100-1772-2853-4595-8000

MFCC の次数は，全帯域に対して 12 とし，部分帯域のフィルタ数におおむね比例させて決定した．たとえば，等間隔に分割する場合，2 帯域では 6，3 帯域では 4 のようになる．

各フレームの特徴量ベクトルは，この MFCC に dB を単位とする帯域内のフレームパワー，両者の Δ 特徴および ΔΔ 特徴を加えて構成する．FC 法（特徴量レベルの統合法）では，パワー特徴が加わることにより，ベクトルの次元は，たとえば 3 帯域の場合 $3\times3\times(4+1) = 45$，6 帯域の場合 $6\times3\times(2+1) = 54$ のようになる（いずれも分割点の最適化を行わない等間隔分割の場合）．

5.4 分割点最適化

情報量を評価基準とした帯域分割の最適化は，フィルタバンクの分割点を変えることにより実現する．たとえば，3 帯域の場合，(0-a) (a-b) (b-8000) Hz のような分割における a と b のあらゆる組合せ（フィルタバンクを構成する各フィルタが最小単位）に対して，4 章で述べた相互情報量を計算し，最大値を与えた分割点を利用する．

本研究では，評価用データのうち各話者につき 1 文ずつ計 24 文を分割点最適化のための適応化学習に用いることとし，式 (9) に基づき帯域数および分割点を

決定した後，残りの 957 文について認識実験を行った．

6. 実験結果と考察

6.1 LPW 雑音による LC 法と FC 法の比較

まず，最初の帯域のみに白色雑音を付加した‘LPW’雑音を用いて音声認識実験を行う．先にも述べたように，これは，MB-ASR における理想的な雑音であるので，最も効果を確認しやすい条件といえる．

図 3 に，様々な SNR で LPW 雑音を加えた場合と雑音なし（clean）の場合の認識性能の変化を，全帯域（従来型）システム，マルチバンド型システムの LC 法および FC 法について示す．帯域数はすべて 3 とした（等間隔分割）☆．

図 3 から，clean（雑音なし）の場合を除きマルチバンド型システムが従来型システムよりも高い性能を与え，また，非常に低い SNR（0〜5 dB）以外では，FC 法が LC 法よりも高い性能を与えていることが分かる．SNR = 10 dB の場合，誤り改善率は LC 法で 25.3%，FC 法で 34.0%であった．

6.2 各種雑音に対する頑健性

次に，種々の雑音に対する認識性能について比較評価する．ここでは，LPW 雑音のほかに，NOISEX-92 データベースの 15 種類の実雑音を評価用データに SNR = 10 dB で付加した．

図 4 は，雑音を加えた評価用データに対して，全帯域（従来型）システムと，LC 法および FC 法に基づくマルチバンド型システム（帯域数は 2 および 3；等間隔分割）で実験を行った結果をまとめたものである．最も右側の average は，LPW 雑音を除く 15 種類の雑音に対する結果の平均値を示す．

図 4 では，すべての雑音条件に対して FC 法が LC 法よりも良い性能を与えている．また，帯域数 2 および 3 の場合ともに，多くの雑音条件下において全帯域（従来型）システムよりも良い性能を与えている．総合的には，帯域数 2 の FC 法が最も高く，かつ全帯域システムを上回る性能を与えた（全雑音に対する平均値であまり効果が見られないのは，後述するように，一部の雑音に対して MB-ASR がまったく有効に働かなかったためである）．全帯域システムに対する認識性能の向上は，雑音の種類に依存し，destroyer-engine 型雑音では 17.6%もの誤り改善（全帯域で 30.6% → 帯域数 2 の FC 法で 25.2%）が得られた．

個々の雑音について見ると，babble，destroyer-engine，factory2，hfchannel，machinegun の場合，マルチバンド型システムは全帯域システムより高い性能を与える．これらの雑音は，信号のエネルギーが一部の周波数帯域に集中するという共通した性質を持つ．これに対して，buccaneer2，destroyer-ops，pink，white の場合，マルチバンド型システムは全帯域システムよりも性能が低下してしまう（white の場合 19.2%の誤り率増加）．これは，これらの雑音の信号エネルギーが周波数軸上のほとんどすべての領域に対して広がっているためであると考えられる．このことから，周波数帯域全体に影響を及ぼすような雑音に対しては MB-ASR は有効でないといえる．ちなみにこの結果は，Hermansky らの実験結果に符合する[16]．

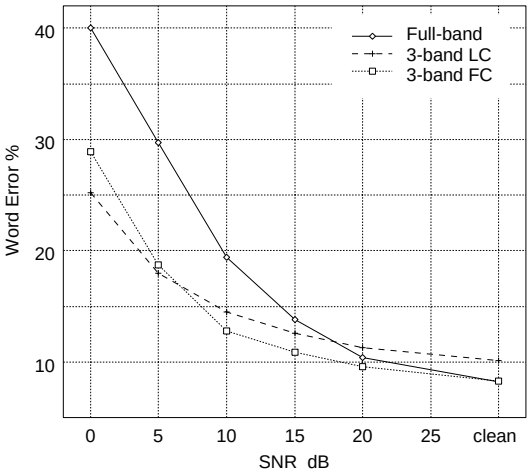

図 3 LPW 雑音を付加した場合の単語誤り率の比較
Fig. 3 Word error rate with ‘LPW’ additive noise.

6.3 分割点最適化に関する予備実験

次に，FC 法に基づき，情報量を評価基準とした帯域分割の最適化を試みる．認識実験に先立って，部分周波数帯域における相互情報量の性質を調べるための予備実験を行う．本論文では，雑音付加音声に対する認識性能向上を目指しているが，ここでは特に音声のスペクトル特性や人間の聴覚特性との関係を調べるために，雑音なし音声についての結果を示す．

図 5 には，低帯域のみ，あるいは高帯域のみを取り出した偏帯域音声において遮断周波数を変化させた場合と，2 帯域システムの帯域分割点を変化させた場合の相互情報量を同時に示す．

偏帯域音声に対するグラフがちょうど交叉する点（約 1,450 Hz）付近で，2 帯域システムにおける相互情報量の最大値が得られている事実は興味深い．なぜなら，偏帯域音声に対する人間の聴覚特性を調査した実験結果[10]によると，本実験と同様に遮断周波数を

☆ LPW 雑音の場合，帯域数により雑音成分を含む帯域幅が異なるため，帯域数を変えた場合の比較はあまり意味を持たない．ここでは，典型的な場合として帯域数 3 の結果を示している．

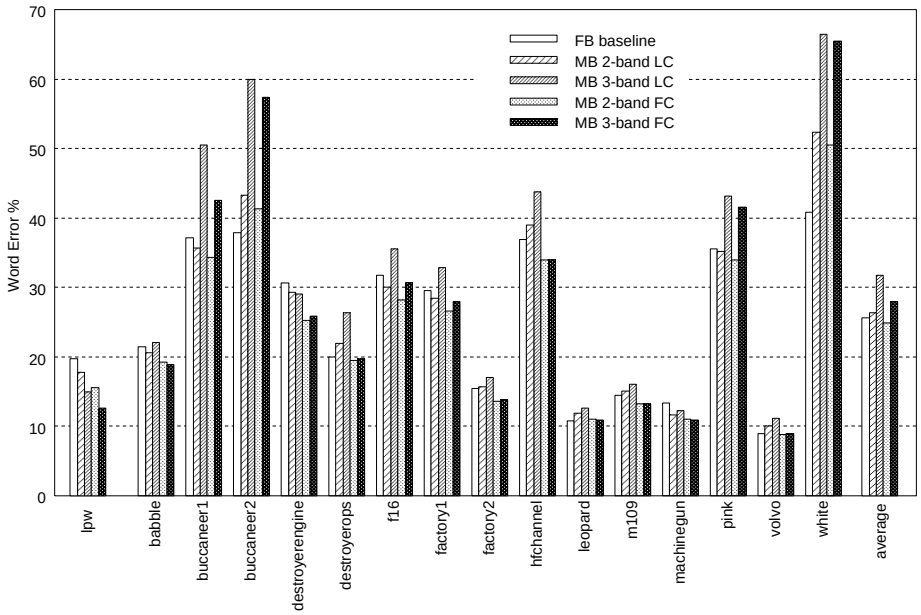

図 4 様々な雑音を付加した場合の単語誤り率の比較
Fig. 4 Word error rate with various additive noise.

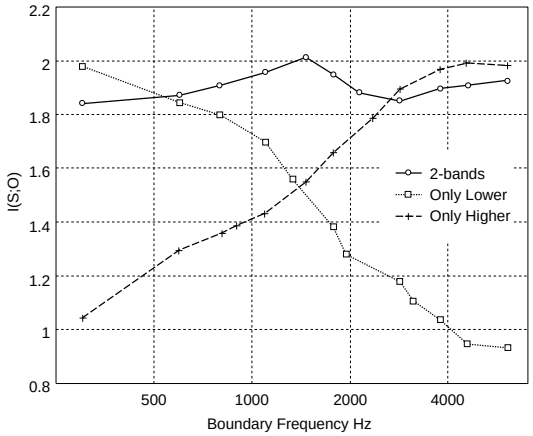

図 5 偏帯域システム（low/high）および 2 帯域システムにおける遮断/分割周波数と相互情報量の関係
Fig. 5 Mutual information for several cut-off/boundary frequencies by partial-band system (low/high) and 2-band system.

変化させた場合，約 1,550 Hz 付近で聴取率が交叉しているからである（文献 10) 中で Articulation Index として説明）．これは，MB-ASR の根拠を議論するうえでの手がかりともなる重要な結果である．

この結果を得た条件の場合，最大相互情報量を与える周波数が 2 帯域における最適分割点となる．

6.4 帯域分割の最適化の効果

最後に，様々な雑音付加条件に対して分割点の最適化実験を行う．付加雑音の種類ごとにそれぞれ 5.4 節で述べた適応化学習を行い，最適な帯域数と分割点を決定する．帯域数としては，先述のとおり 1（全帯域），2，および 3 を用い，あらゆる分割点の組合せの中で最も高い情報量を与える場合を選択した．

図 6 に，全帯域型（FB-ASR；従来型の音声認識）を基準（0%）とした場合の単語誤り数の改善率を示す．図中で “MB equal bandwidth” は帯域分割点をメル尺度上で等間隔に設定した場合（分割の最適化を行わない図 4 における最良値に相当）を，“MB optimized bandwidth” は情報量に基づく最適化学習により帯域分割点を決定した場合を示している．

LPW および destroyer-ops, f16, factory1, hfchannel, leopard, pink の各雑音において，分割点の最適化を施すとさらに性能が向上することが確認できる．また，FB-ASR（従来法）よりかえって性能が劣化してしまう buccaneer2 および white の両雑音では，最適化において帯域数 1（全帯域）での情報量が最大となり，全帯域モデルが選択された．これは，MB-ASR が効果的でない条件においても，少なくとも従来法と同等の性能が得られることを意味している．

図 7 には，15 種類の付加雑音の長時間平均スペクトル（横軸が周波数 [kHz]，縦軸は正規化振幅値）と，本手法による最適分割点の位置（太線）をそれぞれ示す．偏帯域性の強い雑音（babble，destroyer-engine など）において，雑音成分が集中する周波数帯域（の周辺）が良好に分割できていることが確認できる．

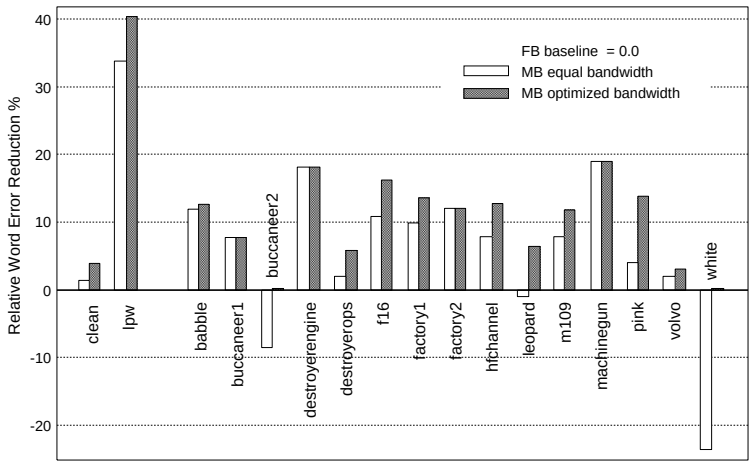

図 6 様々な雑音付加音声に対する分割点最適化の効果
Fig. 6 Effect of optimal band-splitting with various additive noise.

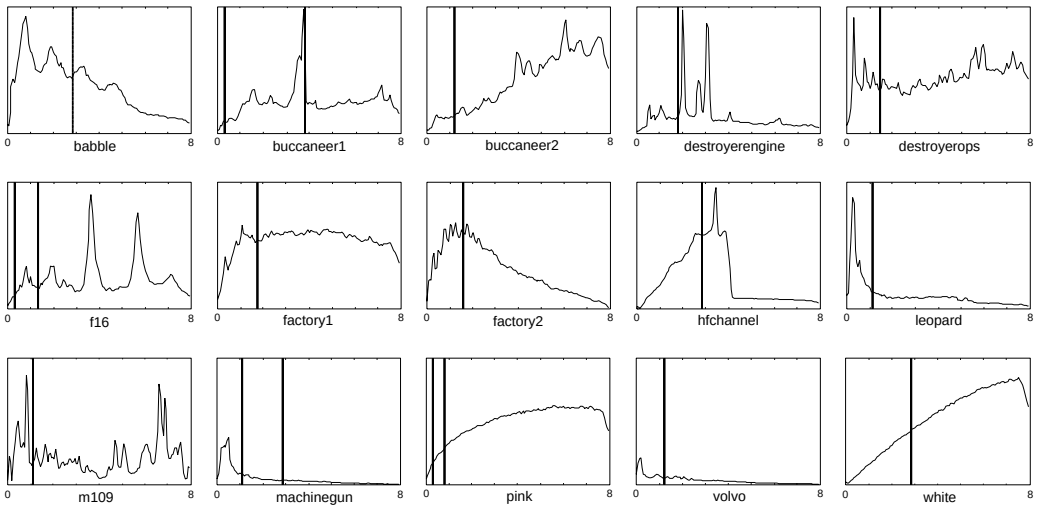

図 7 各種付加雑音の長時間スペクトルと最適分割点
（横軸が周波数 [kHz]，縦軸は高域強調後の正規化振幅値，太線が最適分割点）
Fig. 7 Long-term power spectrum and the optimal splitting frequencies for various additive noise.

7. おわりに

本論文では，マルチバンド型音声認識の方法といくつかの実験結果について述べた．認識手法としては，特に次の 2 つのアプローチを比較検討した．

(1) 尤度レベルの統合法（LC 法）：部分周波数帯域からそれぞれ得られた尤度を，HMM 状態を単位として統合する．

(2) 特徴量レベルの統合法（FC 法）：各帯域に対して音響分析を独立に行い，部分帯域の特徴量ベクトルを連結することにより統合する．

認識実験は，入力音声に対していくつかの異なる種類の雑音を加えて行った．その結果，一般に加えた雑音のエネルギー成分が一部の周波数に集中している場合，マルチバンド型音声認識は従来型の音声認識よりも良い性能を与えることが確認された．また，本論文で提案した FC 法は，一般に LC 法よりも良い性能を与えることが確認された．

さらに，FC 法に基づき，特徴量ベクトルと HMM 状態（音素カテゴリ）との相互情報量を評価基準として，最適な分割点を決定する手法を用いた結果，様々な雑音付加条件において，雑音の周波数特性に応じて最も適した分割点を効果的に決定することができ，さらなる認識性能の改善を得た．

今後は，部分帯域特徴量そのものの改善や，統合時における重み付け基準の導入，さらに人間の聴覚機構との関係などの問題について検討を進める予定である．

謝辞 本研究を進めるにあたり，実験環境の提供と有益な助言をいただいた米国 AT&T 研究所の Enrico Bocchieri，Alex Potamianos（現在 Lucent Technologies），David Roe（現在 Computer Motion Inc.）の各氏ならびに早稲田大学理工学総合研究センター音声言語研究室の諸氏に感謝します．

参 考 文 献

1) Junqua, J.-C. and Haton, J.-P.: *Robustness in Automatic Speech Recognition — Fundamentals and Applications*, Kluwer Academic Publishers, Boston (1996).
2) Van Compernolle, D.: Increased Noise Immunity in Large Vocabulary Speech Recognition with the Aid of Spectral Subtraction, *Proc. ICASSP*, pp.1143–1146 (1987).
3) Varga, A. and Moore, R.: Hidden Markov Model Decomposition of Speech and Noise, *Proc. ICASSP*, pp.845–848 (1990).
4) Bourlard, H. and Dupont, S.: A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands, *Proc. ICSLP*, pp.426–429 (1996).
5) Hermansky, H., Tibrewala, S. and Pavel, M.: Towards ASR on Partially Corrupted Speech, *Proc. ICSLP*, pp.1579–1582 (1996).
6) Mirghafori, N. and Morgan, N.: Combining Connectionist Multi-band and Full-band Probability Streams for Speech Recognition of Natural Numbers, *Proc. ICSLP*, pp.743–746 (1998).
7) Okawa, S., Bocchieri, E. and Potamianos, A.: Multi-band Speech Recognition in Noisy Environments, *Proc. ICASSP*, pp.641–644 (1998).
8) Cerisara, C. and Fohr, D.: Multi-band Automatic Speech Recognition, *Computer, Speech & Language*, Vol.15, pp.151–174 (2001).
9) 武田一哉：頑健な音声処理手法，電子情報通信学会技術研究報告，SP2000-75, pp.1–6 (2000).
10) Allen, J.B.: How Do Humans Process and Recognize Speech?, *IEEE Trans. on Speech and Audio Processing*, Vol.2, No.4, pp.567–577 (1994).
11) 河村尊良，武田一哉，板倉文忠：特定帯域不使用モデルを用いた雑音環境下の音声認識，日本音響学会講演論文集，2-Q-16, pp.121–122 (2000).
12) Okawa, S., Nakajima, T. and Shirai, K.: A Recombination Strategy for Multi-band Speech Recognition Based on Mutual Information Criterion, *Proc. European Conf. on Speech Communication and Technology*, pp.603–606 (1999).
13) 金寺 登，荒井隆行，船田哲男：変調スペクトルの重要な成分のみを選択的に用いた雑音に強い音声認識，電子情報通信学会論文誌，Vol.J84-D-II, No.7, pp.1261–1269 (2001).
14) Dahl, D.A., et al.: Expanding the Scope of the ATIS Task: The ATIS-3 Corpus, *ARPA Spoken Language Technology Workshop*, pp.3–8 (1994).
15) Varga, A. and Steeneken, H.M.: Assessment for Automatic Speech Recognition, NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Automatic Speech Recognition Systems, *Speech Communication*, Vol.12, No.3, pp.247–251 (1993).
16) Tibrewala, S. and Hermansky, H.: Sub-band Based Recognition of Noisy Speech, *Proc. ICASSP*, pp.1255–1258 (1997).

(平成 13 年 11 月 14 日受付)
(平成 14 年 4 月 16 日採録)

大川 茂樹（正会員）

1992 年早稲田大学理工学部電気工学科卒業．1996 年同大学院理工学研究科電気工学専攻博士後期課程修了．1996〜1997 年米国 AT&T 研究所客員研究員．1998 年千葉工業大学情報ネットワーク学科助教授．博士（工学）．音声言語処理，ヒューマンインタフェース等に関する研究に従事．IEEE，ISCA，電子情報通信学会，日本音響学会，人工知能学会等会員．

白井 克彦（正会員）

1963 年早稲田大学理工学部電気工学科卒業．1968 年同大学院理工学研究科電気工学専攻博士後期課程修了．同年同大学電気工学科専任講師．1975 年同教授．1991 年より同大学情報学科教授．1998 年同大学常任理事・副総長．工学博士．音声言語処理，画像処理，ヒューマンインタフェース，教育用マルチメディアシステム，信号処理用アーキテクチャ設計等に関する研究に従事．IEEE，電子情報通信学会，電気学会，日本音響学会，人工知能学会等会員．