音声言語情報処理 29－5
(1999. 12. 20)

マルチバンド型音声認識のための部分帯域特徴量の情報量評価

中島雄大† 大川茂樹‡ 白井克彦†

†早稲田大学 理工学部 情報学科
‡千葉工業大学 情報ネットワーク学科
†〒169-8555 東京都新宿区大久保3-4-1
Tel. (03)5286-3118
E-mail : nakajima@shirai.info.waseda.ac.jp

あらまし：本稿では，マルチバンド型音声認識の性能向上および適用可能性の検討を目的とし，部分周波数帯域より得られる特徴量の情報量を基準とした評価方法を提案する．部分帯域システムを設計する場合に検討すべき種々の構成要素の中で，本稿では特に部分帯域の分割周波数に着目し，その違いにより得られる特徴量の条件付エントロピーと認識性能との相関を考察する．提案した評価方法に基づき，最適な分割周波数を設定した上で音素認識実験を行なった結果，全帯域システムに対し最大11.2%の認識誤り率の減少を達成し，本手法が背景雑音の有無に依存しない評価方法となる可能性が高いことが示唆された．
キーワード：マルチバンド型音声認識，実環境下音声認識，部分帯域特徴量，条件付エントロピー

Evaluation of Sub-Band Features Based on Information Criterion for Multi-Band Speech Recognition

Takehiro NAKAJIMA† Shigeki OKAWA‡ Katsuhiko SHIRAI†

†Dept. of Information and Computer Science, Waseda University
‡Dept. of Network Science, Chiba Institute of Technology
†3-4-1, Okubo, Shinjuku, Tokyo, 169-8555, Japan
Tel. ＋81-3-5286-3118
E-mail : nakajima@shirai.info.waseda.ac.jp

Abstract : This paper proposes an evaluation criterion based on information theory using sub-band features to improve the recognition performance of multi-band speech recognition. Particularly, we discuss the correlation between the recognition performance and the conditional entropies of each sub-band feature derived from various freqency boundaries. As the result of phoneme recognition experiments, we observed an improvement in error reduction by 11.2% at 15dB SNR level by optimizing the sub-band features and we confirmed that our evaluation criterion is effective in noisy environments as well as in clean environment.
Keywords : Multi-Band Speech Recognition, Robust Speech Recognition, Sub-Band Feature, Conditional Entropy

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. これまでにも, 背景雑音が存在する条件における音声認識の頑健性を論じた研究が行なわれているが, それらは (1) 認識モデルや雑音モデルを用いた適応化に関するもの, (2) 雑音等の影響を受けにくい音響特徴量の抽出に関するものに大別される [1].

背景雑音への適応化に関する手法としては, 入力音声のスペクトルから推定雑音のスペクトルを除去するスペクトル・サブトラクション (SS) 法や, 雑音重畳音声を音声認識モデル (HMM) と雑音モデル (HMM) を用いてデコーディングするモデル補償法 (Model Compensation) が知られている. また, 音響特徴量の抽出に関する手法としては, 線形予測分析 (PLP) とスペクトルの時間変化において変動の少ない部分を捉える RASTA(Relative Spectral) を併用した RASTA-PLP 法がある. しかし, これらの手法には (1) 雑音の推定スペクトルに関する知識の必要性, (2) 認識モデル設計の複雑化といった問題がある.

ところで, 近年の音声認識に関する研究において, 狭い部分周波数帯域 (サブバンド) の独立処理・再統合による知覚に基づく考え方がある. サブバンドに分割する意義として, 音声には, (1) 例えば, 有声音 (特に母音) は音韻性を表すフォルマントが低域に存在し, 摩擦音は高域にエネルギーが分布している, (2) 実環境下においては, 雑音成分が狭い帯域に集中することが多い, といった特徴があり, 対象帯域に応じた処理を施すことができるためである.

サブバンドに着目した研究で, いくつかに分割されたサブバンドのそれぞれについて音響特徴量を独立に計算した後, 識別以前のどこかの時点でそれらを再統合する複合周波数帯域 (マルチバンド) 型音声認識 [2-6] が音声認識の新しいパラダイムとして提案され盛んに研究されている. その他にも, サブバンド内に含まれる音声のパワースペクトルを用いてバンド毎のセントロイド周波数を求めるスペクトル・サブバンド・セントロイド (SSC) [7] や, サブバンドのダイナミックレンジを用いて実環境下音声認識の性能を評価する手法 [8] などが提案されている.

サブバンドに着目した音声認識では, 従来型の手法の条件に加えて, (1) バンド数, (2) バンドの分割周波数, (3) バンド毎の分析条件, 加えて, マルチバンド型音声認識には, (4) 再統合の時点, (5) 再統合の方法等の問題を検討する必要がある. 先行研究においては, バンドの分割数は 2~7, バンド分割の基準はメル尺度を使用している場合が多い. 分割基準としてメル尺度を用いることは, 人間の聴覚構造に基づくものとして妥当性があるといえる. しかし, 実環境下においては, 雑音エネルギーの周波数分布が一様でないため, メル尺度により抽出された特徴量が必ずしも音声認識に有効な情報を保持しているとは言い難い. そのため, 各条件に対する定量的な評価方法を検討する必要があると思われる.

そこで本研究では, 実環境下におけるサブバンド特徴量を用いた音声認識の性能向上および適応可能性の検討を目的とし, 情報理論的アプローチに基づくサブバンド特徴量の評価方法を提案する. 本研究では, 音声認識手法としてマルチバンド型音声認識を使用し, 各サブバンド特徴量の条件付エントロピーを評価基準とした. そして, その評価基準をバンドの分割周波数の最適化に適用した場合の効果を, 環境雑音を用いた音素認識実験により検証した結果を報告する.

2 マルチバンド型音声認識

マルチバンド型音声認識では, 複数のサブバンドに対して音響特徴量を独立に計算して認識を行なっている. この手法では, 各サブバンドの特徴量があらかじめ独立にモデル化されており, 認識時においても各サブバンドに対して異なる HMM がそれぞれ独立に適応され, 各 HMM から認識候補とそのスコアが出力される. その後, 全ての HMM の出力を統合して全体の認識スコアが計算され, 認識結果が得られる.

そのため, 1章で述べたように, 尤度の再統合をどの時点で行なうかが問題となってくるが, Bourlard らの報告 [2] では, 尤度の再統合を行なう時点として, HMM の状態を用いた場合でも, さらに高度な単位 (例えば音素や音節, 単語など) を用いた場合でも, ほぼ同等の性能が得られている. HMM の状態での再統合は, 実装が容易であることから, 本研究における尤度の再統合は HMM の状態で行なうものとする.

ここで, フレーム t にサブバンド b より観測されるベクトルを o_t^b, HMM の j番目の状態を s_j とする. 任意の (t,j) において, 各サブバンド b に対するフレーム出力確率 $p(o_t^b|s_j)$ が計算された時, 確率の再統合は, バンド間の独立性を仮定した上で, 式 (1) のように全ての出力確率を掛け合わせることにより実現できる.

$$p(o_t|s_j) = \prod_{b=1}^{B} p(o_t^b|s_j) \qquad (1)$$

3 情報理論的アプローチ

音声は, 連続情報源であることから, 情報理論と密接な関係があるといえる. 特に情報理論に関係が深いのは音声符号化技術であるが, 音声認識もまた, 確率・統計的手法に基づいていることを考えると, 情報理論的なアプローチをとることは自然である.

音声認識では, 手法によらず様々な特徴量が用いられるが, それらの組合せや分析条件によって表現できる情報の内容が異なる. そのため, 音韻や単語等の認識単位に対して, ある特徴量がどれだけ認識に有効であるか, という点が問題になる. この問題は, 音声波から得られる音響特徴量 (例えばスペクトル) の空間を X, 認識目標とする言語的要素 (例えば音韻) の空間を Yとした時, Xの空間要素の分類が, その要素を知った時にYの要素を確定するのにどれだけの情報を持つか, つまり, XとYの情報量 (エントロピー) に関する問題として捉えることができる. このエントロピーを評価の基準とすることにより, 認識に対する様々な特徴量の有効性を, 最終的な認識性能によらず個別に評価できるため, 適切な特徴量の選択が可能であると予測される.

そこで本研究では, エントロピーを音響特徴量に対する評価尺度とし, 特に分割周波数の違いにより得られるサブバンド特徴量に対する適用を試みる. まず, HMM の出力確率 $P(o_t^b|s_j)$ より, 式 (2) を用いて事後確率 $P(s_j|o_t^b)$ の計算を行なう.

$$P(s_j|o_t^b) \cong \frac{P(o_t^b|s_j)}{\sum_j P(o_t^b|s_j)} \tag{2}$$

次に, 観測ベクトル o_t^bが与えられた時の, 全ての音素とそのHMM の全状態に対する条件付エントロピー $H(S|o_t^b)$ を式 (3) より求める.

$$H(S|o_t^b) = \sum_j -P(s_j|o_t^b)\log P(s_j|o_t^b) \tag{3}$$

式 (3) によりフレームあたりのエントロピーが算出されるが, 本研究では, それら全ての平均をとったエントロピー $H(S|O)$ を評価基準として適用した.

$$H(S|O) = \frac{\sum_t H(S|o_t^b)}{T} \tag{4}$$

この $H(S|O)$ は, 観測ベクトル o_t^bより HMM の全状態の中からある状態を決定する, すなわち音響特徴から音韻性を決定する時の曖昧性を表す尺度となる.

本研究では, 認識過程では 2章で述べた尤度レベルの再統合を利用しているが, 後の音素認識実験におけるバンド分割周波数の選択の基準には, 特徴量レベルの再統合 (各サブバンドに対して特徴量をそれぞれ独立に計算した後, それらを連結して 1 つのベクトルとする・FC:Feature Combination) 手法を利用した.

4 実験条件

本章では, 本研究において使用した認識システムと実験試料について述べる. 認識器は文脈独立 HMM を用い, タスクは音素認識実験である. HMM は 4 状態 3 ループ (最終状態はナル遷移) の left-to-right 型の離散出力分布型 HMM である.

4.1 音響分析

本研究では, 音響特徴量として FFT 分析に基づく FFT ケプストラムを用いた. 比較実験のために設計する全帯域 (FB:Full-Band) システムでは, 全周波数の短時間振幅スペクトルを用いて FFT ケプストラムを計算する. 一方, 部分帯域 (SB:Sub-Band) システムでは, 対象とする周波数帯のスペクトルを用いて, それぞれのSB に対して逆 FFT を独立に適用して複数の FFT ケプストラムを算出する.

また, サンプリング周波数 12kHz, 窓長 21.3ms(ハミング窓), 分析周期 5ms である. FFT ケプストラムの次数は, FB システムが 24 次元, SB システムが各帯域で 12 次元とした.

4.2 帯域分割点

本稿では, サブバンド数 Bが 2 である場合の実験結果を示す. その際の帯域分割周波数は, 750, 1125, 1500, 1875, 2250, 2625, 3000, 3375, 3750, 4125, 4500, 4825, 5250Hz の 13 点であるが, 分割点付近におけるオーバーラップは設けていない. なお, 本研究におけるメル尺度等分割周波数は約 1466Hz である.

4.3 音声データ

本研究では, 学習データとして, ATR 日本語データベースの多数話者連続発声 150 文 (男声 20 名) を, 評価データとして ATR 音韻バランス 216 単語 (男声 20 名) を用いた. 使用音素は, 長母音・拗音・無音を除く 25 音素である.

4.4 雑音データ

本研究では, 7 種類の異なる性質の雑音をクリーンな音声データに対して, 計算機上で評価データにのみ加えた. 雑音の種類は, 理想的な定常雑音として, LPW 雑音 (FIR フィルタを用いて低域にのみ白色雑音を音声データに付加する・カットオフ周波数 953Hz) と, NOISEX-92 データベース [9] に含まれている 6 種類の環境雑音

(B:babble, DE:destroyerengine, F16:f16, FA2:factory2, M109:m109, V:volvo) を使用した. これらの雑音のうち, DE は約 1500Hz から約 2500Hz に, F16 は約 1500Hz, 約 3000Hz, 約 4500Hz にエネルギーが集中している, ほぼ定常な雑音で, B, FA2, M109, V は時間的に非定常な雑音である. 評価データに 7 種類の雑音を付加する際の信号対雑音比 (SNR) を 10, 15, 20 dB とした.

5 実験結果

5.1 雑音混入による影響

まず, 雑音を付加したことによる条件付エントロピー $H(S|o_t^b)$ の分布の変化を検討した. 図 1と図 2に定常雑音 LPW, F16 の SNR を 15 dB としてクリーン音声に付加した場合の LB(Low-Band:低域) の $H(S|o_t^b)$ の分布を示す. 帯域分割周波数は, 個々の音声データの LB に雑音成分を含む周波数に設定したため, 図 1が 750Hz, 図 2が 3350Hz と異なっている.

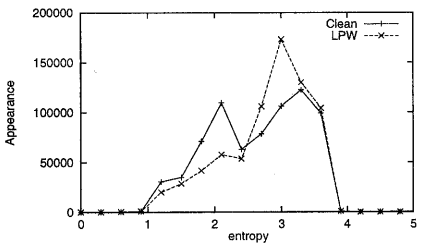


図 1: $H(S|o_t^b)$ の分布 (LPW)

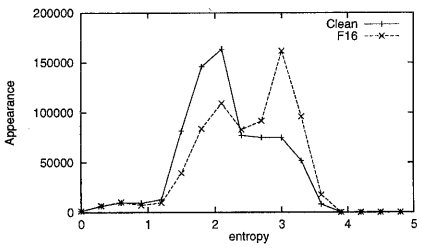
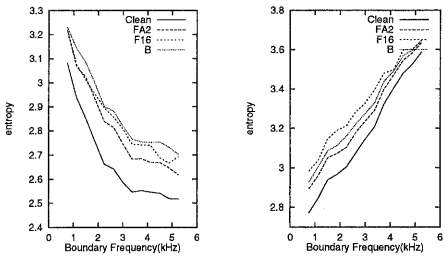


図 2: $H(S|o_t^b)$ の分布 (F16)

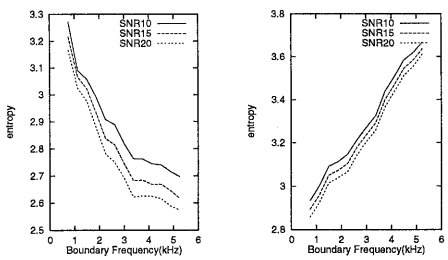
図 1・図 2より, $H(S|o_t^b)$ の分布がそれぞれ右方向にシフトしていることがわかる. 特に $H(S|o_t^b)$ の値が 2 以下の範囲での個数が大幅に減少しており, 雑音によって帯域内の音韻情報が損失していることを示している. このことにより, 対象帯域に応じた処理を施すことは有効であると思われる.

図 3に, クリーン音声と SNR を 15dB として環境雑音 FA2, F16, B を付加した音声の $H(S|O)$ の推移を, 図 4に, SNR を変更して FA2 を付加した音声の $H(S|O)$ の推移を示す. 各図において, 左が LB, 右が HB(High-Band:高域) の推移である.



LB HB

図 3: $H(S|O)$ の推移 (SNR 15dB)



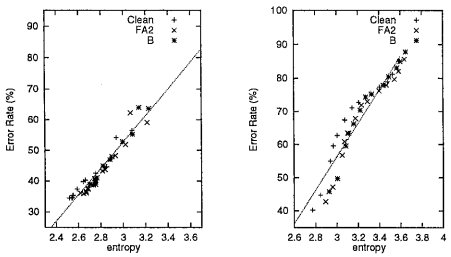
LB HB

図 4: SNR を変更した時の $H(S|O)$ の推移 (FA2)

図 3より, 雑音を付加することによって, クリーン音声と比較して $H(S|O)$ が増加し, 図 4より, SNR 値の減少に伴い, $H(S|O)$ が増加することがわかる. これは, 図 1・図 2における傾向は分割周波数に依存しないことを示していると思われる. また, 図 3より, 雑音の種類によって $H(S|O)$ の推移の傾向が異なっているが, これは, 各雑音のエネルギー分布の違いによると考えられる. 以上より, 情報量が付加雑音の有無やその性質を確認する指標となり得る可能性が高いと思われる.

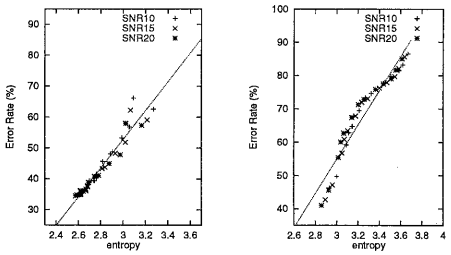
5.2 情報量と認識性能の相関

図 5に, クリーン音声と SNR を 15dB として環境雑音 FA2, F16, B を付加した音声の $H(S|O)$ と音素認識誤り率の分布を, 図 4に, SNR を変更して FA2 を付加した音声の $H(S|O)$ と誤り率の分布を示す. 各図において, 左が LB, 右が HB の分布である.



LB HB

図 5: $H(S|O)$ と誤り率の相関図 (SNR 15dB)



LB HB

図 6: $H(S|O)$ と誤り率の相関図 (FA2)

各図より, 雑音の有無や SNR 値の違いによらず, $H(S|O)$ の減少に伴い誤り率も減少する傾向があり, 情報量と音声認識性能との間には相関があることを示している. また, LB に関しては, $H(S|O)$ が 3 以上の範囲で認識性能にばらつきが見られる. この範囲における分割周波数は 1125Hz 以下であり, 有声音のフォルマント情報損失の影響によると思われる. 逆に, LB の分布が密集している範囲の分割周波数は約 4000Hz 以上であるが, このことは音声の音韻性情報が主として約 4000Hz 以下のスペクトル帯域に存在することを示唆するものと思われる.

5.3 FC による分割周波数の最適化

本研究では, 連結ベクトル FC を分割周波数の最適化の評価基準としているが, (1)FC を構成する際の各バンドの次元数を同一にしている, (2)LB においては分割周波数が 4000Hz 以上の範囲における情報量の変化量が小さい (5.1節), (3) 情報量と認識性能には相関がある (5.2節), といった点をふまえると, FC は分割周波数 4000Hz 以上では音韻的な情報を効率良く表現できているとは言い難い. これより, 本稿における分割周波数の評価基準としての FC の有効範囲は 750Hz から 3750Hz とした.

図 7に各雑音を付加した個々の音声に対して, 750Hz から 3750Hz の間で FC の $H(S|O)$ を求め, その値の小さい周波数の上位 3 点を抽出した際の頻度を示す.

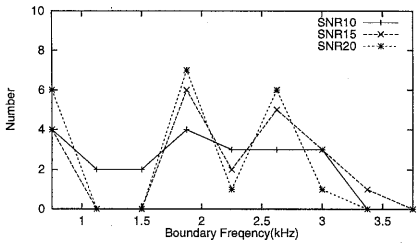


図 7: 最適分割周波数の分布

図 7より, SNR 値が大きくなるほど, いくつかの周波数値に分布が偏る傾向があることがわかる. 特に, 頻度が多くなっている 750Hz, 1875Hz, 2625Hz は, 本研究におけるクリーン音声の FC の $H(S|O)$ の上位 3 点に照合することから, FC の $H(S|O)$ が背景雑音の有無によらない頑健な評価基準である可能性が高い.

5.4 音素認識実験

表 1に, 各雑音の SNR を 15dB としてクリーン音声に付加した場合における, MB システムの FB システムに対する相対的な誤り減少率を示す. 表 1では, Equal が分割周波数を全帯域の半分 (本研究では 3000Hz) とした場合, Optimized-Clean がクリーン音声の FC の $H(S|O)$ の最小値に基づいて最適な分割周波数 (本研究では 1875Hz) を決定した場合, Optimized-Noise が各雑音を付加した音声毎に 750～3750Hz の範囲で最適な分割周波数を決定した場合の実験結果を示している. また, 図 8, 図 9に, 雑音 F16, FA2 の SNR を変化させてクリーン音声に付加した場合の FB システムに対する誤り減少率の推移を示す.

表 1: 認識誤り減少率 (SNR 15dB)

Noise	Equal	Optimized-Clean	Optimized-Noise
Clean	-0.415	4.17	-
B	-2.78	-0.26	0.64
DE	-4.0	-7.76	11.2
F16	0.43	1.00	3.06
FA2	-1.24	1.24	1.76
LPW	-0.01	3.16	3.41
M109	-0.69	2.82	2.82
V	-0.36	4.11	4.11

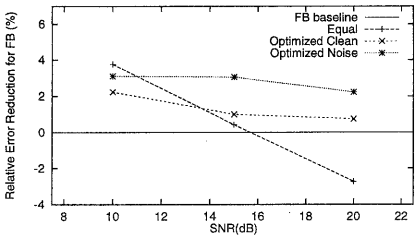


図 8: 認識誤り減少率の推移 (F16)

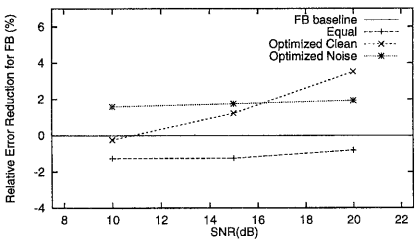


図 9: 認識誤り減少率の推移 (FA2)

表1, 図8, 図9より, FC の $H(S|O)$ を評価基準として分割周波数の最適化を行なうことにより, FB システム・等分割と比較して認識誤りがおおむね減少していることがわかる. また, 各雑音に適応させることによって, クリーン音声に適応させた場合と比較して, 誤り率減少に効果的であることがわかる. FC の $H(S|O)$ を評価基準とする場合に, クリーン音声の $H(S|O)$ を使用することは, 雑音の周波数分布 (雑音 DE で FB に対し 7.76%誤り率増加, 表 1) や, SNR の値によって性能が変化しやすいことから, 雑音により劣化した音声に対して FC の $H(S|O)$ を評価基準として最適化を行なうことが, 雑音の種類や SNR 値に依存せずに認識性能の向上を得られると思われる.

6 まとめ

本稿では, マルチバンド型音声認識の性能向上を目的とし, 情報理論的アプローチに基づくサブバンド特徴量の評価方法を提案した. 本稿では特に, 帯域分割周波数に着目し, その違いによる情報量の変化や認識性能との相関を検証した. また, 本稿で提案した情報量評価基準に基づく帯域分割周波数の最適化の効果を音素認識実験により検討した結果, 雑音入り音声データに対して, 最大 11.2%の認識誤り率の減少を得ることができた.

本稿では, 最適化の基準として連結ベクトルの情報量のみを適用したが, 5.2節の結果を考慮した上で, 各部分周波数帯域との相関を検証する必要がある. 今後は, 本稿で提案した評価基準を基に, 1章で述べたサブバンドを利用した音声認識における諸条件や分割点付近におけるオーバーラップに関する問題に対して, 本稿で提案した評価基準の適用可能性を検証する予定である.

参考文献

[1] 中川, "ロバストな音声認識のための音響信号処理", 音響誌, vol.53-11, pp.864-871, 1997

[2] H.Bourlard, S.Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands", ICSLP, pp.426-429, 1996

[3] H.Hermansky, S.Tibrewala, M.Pavel, "Towards ASR on partially corrupted speech", ICSLP, pp.1579-1582, 1996

[4] N.Mirghafori, N.Morgan, "Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers", ICSLP, pp.743-746, 1998

[5] S.Okawa, E.Bocchieri, A.Potamianos, "Multi-band speech recognition in noisy environments", ICASSP, pp.641-644, 1998

[6] S.Okawa, T.Nakajima, K.Shirai, "A recombination strategy for multi-band speech recognition based on mutual information criterion", Eurospeech, pp.603-606, 1999

[7] 柘植, 深田, Singer, "話者正規化スペクトルサブバンドパラメータを用いた雑音下での音声認識", SLP-24-9, pp.63-68, 1998

[8] 近藤, 武田, 板倉, "帯域別ダイナミックレンジによる音声認識率の予測", SP99-28, pp.15-20, 1999

[9] http://spib.rice.edu/spib.html