社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2003－ＳＬＰ－49　（３）
2003／12／18

周波数特性の変動に頑健な実時間分散音声認識手法

柘植　覚[†]　黒岩　眞吾[†,††]　原　一眞[†]　北　研二[†††]

† 徳島大学 工学部
†† ATR 音声言語コミュニケーション研究所
††† 徳島大学 高度情報化基盤センター
E-mail: †{tsuge,kuroiwa,hara,kita}@is.tokushima-u.ac.jp

あらまし　携帯電話や PDA などの携帯端末の音声認識手法として分散音声認識手法 (DSR: Distributed Speech Recognition) が近年提案された。DSR では、携帯端末とサーバ間で伝送するデータ形式等を共通化する必要があり、現在、ETSI において標準化が進められている。標準化の一環として、2000 年 4 月には ETSI 標準 DSR フロントエンド、2002 年 10 月には雑音対策手法などを追加した ETSI Advanced DSR フロントエンドが勧告された。携帯端末は多種多様であり、使用される入力デバイスの周波数特性には差異が生じる。この差異は特徴パラメータ圧縮時のベクトル量子化歪みを増加させ、音声認識性能劣化の要因の一つとなる。そこで、本稿では、周波数特性を正規化する手法を提案する。提案手法は、各フレーム毎に複数の参照ケプストラムと特徴パラメータを比較し、参照ケプストラムに入力ケプストラムが近付くように周波数特性を正規化する。実際には、入力発声の音素列を推定し、各音素の特徴パラメータの平均が音響モデル学習時の特徴パラメータの平均と一致するように入力特徴パラメータを平行移動させ、周波数特性を正規化する。音声認識実験結果より、提案手法は ETSI Advanced DSR フロントエンドで使用されている Blind Equalization 手法より高い認識性能を示した。特に、提案手法は MIRS フィルタ条件下で ETSI Advanced DSR フロントエンドの単語誤り率を 17.88%削減 (16.67%→13.69%) することが可能であった。

キーワード　分散音声認識、ETSI DSR フロントエンド、乗算性雑音、周波数特性正規化手法

Real-time Frequency Characteristic Normalization for ETSI DSR Front-end

Satoru TSUGE[†], Shingo KUROIWA[†,††], Kazuma HARA[†], and Kenji KITA[†††]

† Faculty of Engineering, Tokushima University
†† ATR-SLT
††† Center for Advanced Information Technology, Tokushima University
E-mail: †{tsuge,kuroiwa,hara,kita}@is.tokushima-u.ac.jp

Abstract　In this paper, we focus on the influence on recognition performance of DSR with acoustic mismatches caused by input devices. DSR employs a vector quantization (VQ) algorithm for feature compression so that VQ distortion is increased by acoustic mismatches. Large VQ distortions increase the speech recognition error rate. To overcome the problem of VQ distortion, we have proposed the Bias Removal Method (BRM) in previous work. However, this method can not be applied in real-time. Therefore, in this paper, we propose a Real-time Bias Removal Method (RBRM). This method estimates the bias using past frames and multiple reference cepstrum vectors instead of one reference which is employed by ETSI advanced DSR front-end. Experimental results on a Japanese newspaper dictation task indicate that the proposed method showed improvement in the recognition performance for blind equalization in ETSI advanced DSR front-end under acoustic mismatched conditions.

Key words　Distributed speech recognition, ETSI DSR front-end, Convolution noise, Frequency characteristic normalization method

1. はじめに

携帯電話やPDA (Personal Digital Assistants) などの携帯端末の発達にともない、モバイル環境の普及が急速に進んでいる。一般にこれらの携帯端末は非常に小型であるため、付属デバイスによる複雑なコマンド入力は困難である。この問題を解消する一手法として、音声インターフェースが熱望されている。現在までに音声入力で、登録電話番号に電話をかける機能 (単語認識) や電話番号を入力する機能 (数字認識) などが携帯電話上で実現されている。しかし、携帯端末のハードウェア (CPU、メモリなど) 面の制約により、電子メイル文入力などの中・大語彙連続音声認識を端末内で実現することは困難である。近年、この問題を解決する手段として、分散音声認識手法 (DSR: Distributed Speech Recognition) が提案された[1]。DSRは、音声認識システムをクライアント (端末) とサーバに分散させ、クライアントの処理量を軽減させる。クライアントでは音響分析のみを行い、分析された特徴パラメータをサーバに伝送し、サーバ側でデコードを行う。伝送するデータが特徴パラメータであるため伝送速度を低減できるうえ、伝送路の制限を受けないため、従来より低域、高域の周波数情報を用いることが可能であり、認識精度を向上できる可能性がある。

DSR方式を広く普及させるためには、音響分析を行うクライアント部と音声認識を行うサーバ部で、圧縮・復元方式、ビットストリーム形式などの共通化が必要である。そのため、欧州電気通信標準化機構 (ETSI: the European Telecommunications Standards Institute) は、そのフロントエンドの標準化を進めている。標準化の一環として、ETSIは2000年4月に雑音に頑健なフロントエンドの研究開発を目的とした『標準DSRフロントエンド (ETSI ES201)』[2]、2002年10月には雑音に頑健な『Advanced DSRフロントエンド (ETSI ES202)』を勧告した[3]。

DSRシステムに使用されるターミナルは多種多様であり、個々のターミナルで使用される入力デバイスの周波数特性には差異があることが予想される。このような周波数特性の差異は特徴パラメータの変動となり、ベクトル量子化 (VQ: Vector Quantization) を用いた特徴パラメータ圧縮部において量子化歪みの増加を引き起こす原因となる。この歪みの増加は音声認識性能を劣化させる要因の一つとなる。従来より、周波数特性の差異を正規化し、認識性能劣化を抑制する有効な手法としてケプストラム平均減算法 (CMS: Cepstral Mean Subtraction) が提案されている[4]。しかし、ETSI DSRフロントエンドでは、CMSを適用しない特徴パラメータで作成されたVQコードブックを使用しているため(注1)、特徴パラメータ圧縮前にCMSを適用した場合、量子化歪みを増加させ認識性能を劣化させる可能性がある。

我々はCMSでは解決ができない入力系の周波数特性の差異によって生じる量子化歪みに着目をし、フロントエンド部でその歪みを減少させる手法に関し研究を行ってきた。実際、入力系の周波数特性の差異による認識性能劣化の問題はDSRにとって大きな問題の一つであるため、Advanced DSRフロントエンドにおいて周波数正規化手法 (Blind Equalization 手法) [5] が導入された。

我々は、周波数特性正規化手法として平均一致化手法を提案し、その有効性を示した[6]。しかし、この手法は一発声終了後にしか適用ができず、実時間処理が不可能であった。そこで、平均一致化手法を実時間処理するため、次の2つの方法を組み込むことを検討した[7]。

(1) 一発声前で計算されたバイアスを用い、現在の発声の周波数特性を正規化する。

(2) 発声開始数秒を用い、バイアスを計算し発声全体の周波数特性を正規化する。

しかし、1では、一発声前と現在の発声の発声環境が異なる場合、本手法を適用することは逆にVQコードブックと特徴パラメータ間のVQ歪みを増加させる原因となる可能性がある。また、2では、発声開始から短時間 (100msec以下) でバイアスを計算することにより実時間処理が可能となるが、バイアス計算に用いるフレーム数が少ないため、適切なバイアスが計算できない可能性がある。逆に、発声開始から長時間を用いバイアスを推定した場合には、適切なバイアスは計算できるが、実時間処理が不可能となる。

そこで、我々は、実時間処理が可能な周波数特性正規化手法として音素平均一致化手法を提案する。本手法は、Blind Equalization手法と類似しており、参照ケプストラムを使用してバイアスを推定する。Blind Equalization手法は1つの参照ケプストラムに対し特徴パラメータの平均が参照ケプストラムに徐々に近付くように周波数特性の正規化を行っている。このため、発声が短い場合や発声された音素に偏りがあった場合には、バイアスは発声された内容に依存するという問題がある。提案手法は、複数の参照ケプストラム (本稿では音素数分) を使用し、各フレームごとで確率計算により各参照ケプストラムに対する重みを計算し、その重み付け平均によりバイアスを計算する。そのため、発声が短い場合や発声された音素に偏りがあった場合においても、適切なバイアスが計算できると考えられる。本稿では、日本音響学会新聞読み上げコーパスを用い、入力デバイスの周波数特性の変動をシミュレーションした実験を行い、提案手法の有効性を示す。

2. 周波数特性正規化手法

マイクなどの入力デバイスの違いに起因する周波数特性の差異は、特徴パラメータを大きく変動させる要因となっている。ETSI DSRフロントエンドでは特徴パラメータの圧縮にVQを用いているため、特徴パラメータの変動はVQ歪みを増加させ、認識性能を劣化させる原因の一つとなる。そこで、周波数特性の差異に対しても、VQ歪みを増加させない手法が必要となる。我々は、特徴パラメータとVQコードブックとの歪みを減少させる、平均一致化手法を提案した。

(注1)：通常CMSによる周波数特性の正規化には1発声全体が必要となるため、実時間処理ができない。そのため、ETSIはCMSを適用していない特徴パラメータでVQコードブックを作成していると予想される。

2.1 平均一致化手法

平均一致化手法は、認識発声の特徴パラメータの平均とVQコードブック作成データの特徴パラメータの平均を一致させるように認識発声の特徴パラメータを平行移動する手法である。以下、本手法をBRM (Bias Removal Method) と呼ぶ。

以下に、本手法の手順を示す。

（1） 前処理: VQコードブック作成データの平均特徴パラメータの計算

$$\boldsymbol{a}_{train} = \frac{\sum_{s=1}^{S}\sum_{n=1}^{N_s}\boldsymbol{x}_{sn}}{\sum_{s=1}^{S}N_s} \tag{1}$$

ここで、$\boldsymbol{a}_{train}$ はVQコードブック作成データの平均特徴パラメータを示し、$\boldsymbol{x}_{sn}$ は発話 s に対する各分析フレームの特徴パラメータを示す。また、S, N_s はVQコードブック作成データ数、発声 s の総分析フレーム数を示す。

（2） 認識発声の平均特徴パラメータの計算

$$\boldsymbol{a}_{test} = \frac{\sum_{n=1}^{N}\boldsymbol{x}_n}{N} \tag{2}$$

ここで、$\boldsymbol{a}_{test}$ は各認識発声の平均特徴パラメータを示し、$\boldsymbol{x}_n$ は各分析フレームの特徴パラメータを示す。また、N は認識発声の総分析フレーム数を示す。

（3） 減算するバイアスの計算

減算するバイアスとして、特徴パラメータの平均と学習データの平均の差を計算する。

$$\boldsymbol{h} = \boldsymbol{a}_{test} - \boldsymbol{a}_{train} \tag{3}$$

（4） BRMの適用 (VQコードブック作成データの平均特徴パラメータと認識発声の平均特徴パラメータの差を減算)

$$\hat{\boldsymbol{x}}_n = \boldsymbol{x}_n - \boldsymbol{h} \tag{4}$$

ここで、$\hat{\boldsymbol{x}}_n$ は本手法を適用した後の特徴パラメータを示す。適用した特徴パラメータを特徴パラメータ圧縮部の入力とすることでVQコードブックとの歪みを減少することが可能である。

実際には、ETSI DSRフロントエンドではVQコードブックのみが与えられるため、VQコードブック作成データの平均特徴パラメータを算出することは困難である。そのため、本稿ではETSIで定義されているVQセントロイドの平均を $\boldsymbol{a}_{train}$ として用いる。

2.2 VQコードブック平均一致化手法の実時間処理

前節で提案した平均一致化手法は、バイアス計算 (式 (2)) に認識発声全体の特徴パラメータが必要となる。そのため、発声終了後でなければ提案手法は適用できず、実時間処理ができない。これは、DSRシステムとしては適していない。そこで、本節では平均一致化手法を実時間処理をさせる2種類の解決策を組み入れる[8]。

（1） 前発声でバイアス計算

一発声前でバイアスを計算し、そのバイアスを用いて現在の発声にBRMを適用する。

（2） 発声開始数フレームでバイアス計算

発声が開始されてからの数フレームを使用して、バイアスを計算する。これは、前節で提案した式 (2) を以下のように変更する。

$$\boldsymbol{a}_{test} = \frac{\sum_{n=1}^{M}\boldsymbol{x}_n}{M} \tag{2'}$$

ここで、$M < N$ であり、M はバイアス計算に使用するフレーム数を示す。この M が小さければ、実時間処理が可能となる。

2.3 BRMの実時間処理の有効範囲と問題点

BRMの実時間処理の有効性を示すため、日本音響学会新聞記事読み上げ音声コーパス (JNAS) [9] を用い、音声認識実験を行った。

2.3.1 実験条件

音響モデルの学習には、IPA学習セットの中から男性話者が発声した音素バランス文 (話者: 103名、発声数: 5,168発声) を使用した。テストセットとして、学習データと同様にIPAで使用されているテストセットの中から男性話者が発声した新聞読み上げ100発声を用いた。

音響モデル学習に用いた特徴パラメータは、ETSI Advanced DSRフロントエンドで分析をしたMFCC 12次元、その一次回帰係数、対数パワーの一次回帰係数の合計25次元を用いた。文献[10]で、VQを行わない特徴パラメータで音響モデルを学習することにより、伝送される特徴パラメータ (VQされた特徴パラメータ) の認識性能が向上できることを報告した。そのため本稿では、音響モデルは圧縮 (VQ) を行っていない特徴パラメータで学習した。認識時の特徴パラメータは、VQにより圧縮された特徴パラメータを用いた。本稿で提案した周波数特性正規化手法は、ETSI Advanced DSRフロントエンドにおけるBlind Equalization手法と変更して、実験を行った。

音響モデルは、各特徴量で学習を行った木構造クラスタリングにより状態共有した3状態16混合の音素環境依存HMM (43音素) の混合連続分布HMMを用いた総状態数は各特徴量ともに約1,000状態である。

周波数特性の差異の影響を検討するため、テストセットに対し以下のフィルタを適用し、人工的に乗算性雑音を加えた音声を作成した。作成した音声データを用い、乗算性雑音に対する提案手法の有効性をシミュレーションした。

- G712フィルタ
- MIRSフィルタ
- 移動平均フィルタ (M/A)

$$\begin{aligned} s_{of}(n) = {} & 0.25 \times (s_{in}(n) + s_{in}(n+1) \\ & + s_{in}(n+2) + s_{in}(n+3)) \end{aligned} \tag{5}$$

G712フィルタ、MIRSフィルタはAURORA2 [11] で使用されているフィルタを使用した。

デコーダにはJuliusを用い、評価は単語誤り率 (WER: Word Error Rate) で行った。各実験のWERは、テストセットに対し最もWERが低くなるようデコード時の最適なパスの広さの設定を行った結果より計算した。

2.3.2 認識実験結果

BRMの実時間処理を用いた音声認識実験結果を表1に示

表1 Recognition results on Japanese speech corpus using the BRM with real-time solutions (Word Error Rate (in %)).

	Filter			
	clean	M/A	G712	MIRS
ES202w/oBEQ	15.22	46.96	18.9	28.73
ES202	13.76	26.74	16.36	16.67
BRM *prev.*	12.36	24.59	14.01	13.25
BRM ($M = 100$)	17.30	27.38	17.12	17.69
BRM ($M = 220$)	12.29	26.13	13.69	14.33
BRM (non real-time)	12.11	24.21	12.37	12.81

す。表中の "ES202" はETSI Advanced DSR フロントエンド (Blind Equlization 手法) の認識結果を示す。"ES202w/oBEQ" はBlind Equalization 手法を適用しない Advanced DSR フロントエンドの認識結果を示す。"BRM *prev.*" はBRM において一発声前でバイアスを計算した結果を示す。"BRM ($M = 100$)"、"BRM ($M = 220$)" は発声開始数フレームを用いる実時間処理手法を組み入れた BRM で、発声開始から 100 フレーム、220 フレームを使用してバイアスを計算した結果を示す。

表 1より、周波数特性正規化手法を用いていない場合 (ES202w/oBEQ)、フィルタによる周波数特性の変化により大きく WER が増加するがしまうことがわかる。これに対し、全ての周波数特性正規化手法は周波数特性の変化に対する WER の増加を低減しており、周波数特性正規化手法は DSR にとって必要な手法であると言える。

一発声前でバイアスを計算した場合 (BRM *prev.*) の WER は "ES202" より低く、実時間処理でない BRM に近い WER を示した。

一方、発声開始から数フレームを用いバイアスを計算する方法は、バイアス計算に使用するフレーム数が多い場合 (BRM ($M = 220$))、"ES202" より低い WER を示すが、少ない場合 (BRM ($M = 100$)) には "ES202" より WER が高いことがわかる。このことより、発声全体の周波数特性を正規化するためにはある程度の長さのフレーム (時間) がバイアス計算に必要であり、実時間処理を行うためには発声が長い場合に限られると言える。

これらの結果より、BRM の実時間解決手法として提案して 2 手法は、安定した発声環境が得られた場合 (発声環境の変動が少ない、発声が長いなど)、Advanced DSR フロントエンドで用いられている Blind Equalization 手法より周波数特性の変動による認識性能劣化を抑制することが可能であることがわかった。しかし、各手法とも発声環境が安定しない場合には、認識性能劣化の抑制効果が低くなると言う問題点が存在することがわかった。

2.3.3 考　　察

2.1で提案した平均一致化手法は、周波数特性の正規化のため、発声全体の特徴パラメータが必要であり、発声終了後でないと適用できず実時間処理の問題が存在した。そのため、**2.2** で平均一致化手法に実時間処理を行う 2 つの解決策を取り入れ、その有効性を音声認識実験にて検証した。音声認識実験結果より、限られた条件下では Blind Equalization 手法より有効に周波数特性の変動による認識性能劣化を抑制できた。しかし、2 つの実時間処理解決策には次の問題点が存在することが明確になった。

(1) 一発声前でバイアスを計算する場合

発声環境が前発声と異なった場合、前発声で計算したバイアスは現在の発声を適切に表現していない可能性が高い。そのため、発声環境が変化した一発声目には適用が不可能であり、周波数特性変動による認識性能劣化の抑制ができず、認識精度が低下する。

(2) 発声開始数フレームでバイアスを計算した場合

実時間処理を行うため、少ないフレーム数でバイアスを計算した場合、発声内容に偏りがある場合や発声時間が短い場合には、発声全体を適切に正規化するバイアスが計算できず、認識性能劣化の低減効果が少ない。発声全体を適切に正規化するバイアスの計算には、多くのフレーム数 (前節の実験では 220 フレーム) が必要であり、実時間処理が困難となる。

3. 実時間周波数特性正規化手法

前節で BRM の実時間処理の有効範囲と問題点を明確にした。そこで、本節では一発声全体の平均を一致させるのではなく音素を推定し、認識発声と VQ コードブック学習データで各音素の平均特徴パラメータを一致させる手法、音素平均一致化手法を提案する。本手法は、実時間処理が可能であり、BRM と同様にバイアスを減算し周波数特性の正規化を行うため、以下では RBRM (Real-time Bias Removal Method) とする。

Blind Equalization 手法は一つの参照ケプストラムに認識時の平均特徴パラメータを近付ける方法である。このため、発声内容 (出現音素数など) に影響され、周波数特性が適切に正規化されない場合があると考えられる。提案手法は、参照ケプストラムを使用しバイアスを計算する点は Blind Equalization 手法と類似する。しかし、本手法は複数の参照ケプストラムを使用し、各フレームごとに各々の参照ケプストラムと特徴パラメータを比較し、発声全体を適切に正規化するバイアスを計算する。これにより、短い発声や音素の出現回数に偏りがある発声に対しても適切なバイアスが計算でき、周波数特性の差異による認識性能劣化を抑制できると考えられる。

3.1 音素平均一致化手法

音素平均一致化手法は以下の手順で入力系の周波数特性の正規化を行う。

(1) 前処理

VQ コードブック作成データを用い、各音素ごとの特徴パラメータの平均ベクトル ($\boldsymbol{a}_p$) を計算する(注2)。

$$\boldsymbol{a}_p = \frac{\sum_{s=1}^{S} \sum_{\tilde{\boldsymbol{x}}_{st} \in p} \tilde{\boldsymbol{x}}_{st}}{\sum_{s=1}^{S} N_{sp}} \tag{6}$$

ここで、$\boldsymbol{a}_p$ は音素 p の VQ コードブック作成データの平均特

(注2)：実際には、コードブック作成データを入手することは困難である。そこで、本稿では音響モデル学習データを用い $\boldsymbol{a}_p$ を計算した。

徴パラメータを示す。$\tilde{x}_{st}$ は発話 s の t フレーム目の特徴パラメータを示す。また、S, N_{sp} はコードブック作成に用いた発声数、発声 s に含まれる音素 p の総分析フレーム数を示す。学習データにおける VQ 歪みを低減させるために、特徴パラメータ ($\tilde{x}_{st}$) には BRM を適用する。これにより、平均特徴パラメータ (a_p) と認識発声の各音素毎の平均特徴パラメータを一致させることにより、周波数特性の正規化と VQ 歪みを減少させるが可能となる。

（2） 入力系の周波数特性の正規化 (提案手法の適用)

$$\boldsymbol{h}_t = (1-\alpha_t)\cdot\boldsymbol{h}_{t-1} + \alpha_t \sum_{p\in phone}(w_{pt}\cdot(\boldsymbol{x}_t - \boldsymbol{a}_p)) \quad (7)$$

$$\tilde{\boldsymbol{x}}_t = \boldsymbol{x}_t - \boldsymbol{h}_t \quad (8)$$

ここで、$\boldsymbol{x}_t$ は入力音声の t フレーム目の特徴パラメータを示す。w_{pt}、α_t は各音素の重み係数、更新係数を示す。$\tilde{\boldsymbol{x}}_t$ は本手法適用後の特徴パラメータを示す。次節において、音素重み係数、更新係数の計算方法について詳しく記述する。

3.1.1 音素重み係数、更新係数の計算方法

提案手法では式 (7) における、各音素の重み係数 w_{pt} と更新係数 α_t の値が重要な要素となる。我々は、以下の方法でこれらの値を計算した。

- 各音素重み係数 (w_{pt}) の計算の計算

$$\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \boldsymbol{h}_{t-1} \quad (9)$$

$$w_{pt} = P(\hat{\boldsymbol{x}}_t|p) = \frac{1}{\sqrt{(2\pi)^D|\Sigma_p|}} \times exp\left\{-\frac{1}{2}(\hat{\boldsymbol{x}}_t - \boldsymbol{a}_p)^T\Sigma_p^{-1}(\hat{\boldsymbol{x}}_t - \boldsymbol{a}_p)\right\} \quad (10)$$

$P(\hat{\boldsymbol{x}}_t|p)$ は音素 p の分布から特徴パラメータ $\hat{\boldsymbol{x}}_t$ が生成される確率を示し、この値を各音素の重み係数とした。本稿では、各音素の特徴パラメータは正規分布で表現されると仮定し、音響モデル学習データより、各音素の正規分布を推定した。

- 音素重み係数の正規化

音素の重み係数の値が高い音素のみをバイアス ($\boldsymbol{h}_t$) の計算に用いる。

– 音素重み係数が高い N 音素:

$$w_{pt} = \frac{w_{pt}}{\sum_{q\in N-phone} w_{qt}} \quad (11)$$

ここで $N-phone$ は音素重み係数の高い上位 N 音素を示す。

– その他の音素:

$$w_{pt} = 0.0 \quad (12)$$

- 更新係数 (α_t) の計算

$$\alpha_t = \frac{1}{t} \quad (13)$$

ここで、t はフレーム番号を示す。これは、現時点までのバイアス ($\boldsymbol{h}_t$) の平均となる。

3.2 認識実験結果・考察

提案した音素平均一致化手法の有効性を示すため、JNAS を用いた音声認識実験を行った。実験条件は **3.1.1** と同様である。

表 2 Recognition results on Japanese speech corpus using the R-BRM (Word Error Rate (in %)).

	Filter			
	clean	M/A	G712	MIRS
ES202	13.76	26.74	16.36	16.67
BRM	12.11	24.21	12.37	12.81
BRM *prev.*	12.36	24.59	14.01	13.25
BRM ($M = 220$)	12.29	26.13	13.69	14.33
RBRM	13.20	25.87	13.94	13.69
RBRM (*ideal*)	11.91	24.21	12.17	12.29

表 2に音声認識実験結果を示す。表中の "ES202" は ETSI DSR Advanced フロントエンド (Blind Equlization 手法) の認識結果を示す。"BRM *prev.*" は BRM において一発声前でバイアスを計算した結果を示す。"BRM ($M = 220$)" は実時間処理手法を組み入れた BRM で、発声開始から 220 フレームを使用してバイアスを計算した結果を示す。これらの結果は表 1の再掲である。

また、"RBRM" は音素平均一致化手法を適用した場合の結果を示す。本実験では、音素平均一致化手法における音素重み係数の計算には上位 1 音素を用いた。"RBRM (*ideal*)" は音素が既知とした場合、すなわち音素重み係数 (式 (10) の w_{pt}) が既知音素ならば 1.0、そうでなければ 0.0 とした場合の認識結果を示す。この結果は提案手法である RBRM の限界値である。

表 2より、全フィルタ環境下で提案した音素平均一致化手法 (RBRM) の WER は "ES202" の WER より低いことがわかる。特に、MIRS フィルタ条件下において、音素平均一致化手法 (表中の RBRM) は Blind Equalization 手法 (表中の ES202) の WER を 17.88%削減 (16.67%→13.69%) できていることがわかる。これは、複数の参照ケプストラムを参照にすることにより、周波数特性を正規化するバイアスが適切に計算できたためであると考えられる。これより、提案手法である RBRM は入力系の周波数特性変動による認識性能劣化を Blind Equalization 手法より高い精度で抑制することが可能であると言える。

また表 2において、音素平均一致化手法 (表中の RBRM) と一発声前をバイアス計算に使用した BRM (表中の BRM *prev.*)、処理遅れを 2.2 秒とした BRM (表中の RBM (M=220)) を比較した場合、フィルタ環境により若干の精度差は見られるが、ほぼ同程度であることがわかる。この表に示した実時間処理を組み込んだ BRM の認識結果は **3.**で述べた問題が生じていない場合の結果である。それらの結果とほぼ同程度であり、かつ実時間処理が可能な音素平均一致化手法は実時間周波数特性正規化手法として有効であると言える。

表 2において実時間処理を考えない場合、"RBRM" の WER は "BRM" より高いことがわかる。しかし、音素平均一致化手法の上限値である "RBRM (*ideal*)" の WER は全環境下で "BRM" の WER より低いことがわかる。"RBRM" と "RBRM (*ideal*)" の違いは音素重み係数の計算方法のみであるため、音素推定精度を向上させることにより、"RBRM" の認識性能を "BRM" と同等以上にできる可能性があると言える。そのため、

表 3 音素数に対する認識結果 (Word Error Rate (in %)).

音素数	Filter			
	clean	M/A	G712	MIRS
1	**13.20**	25.87	**13.94**	**13.69**
2	13.26	25.17	14.20	14.09
3	13.75	25.89	14.14	14.15
5	13.69	**24.41**	14.46	14.32
7	13.43	25.38	14.39	13.69

今後は音素推定精度向上に関する検討が必要である。

3.3 バイアス計算に使用した音素数と認識結果の関連

3.1で述べた通り提案手法の計算には、音素重み係数で重み付けされた参照ケプストラムと各フレームの特徴パラメータの差が使用される。そのため、音素重み係数はバイアス計算に影響を及ぼし、認識性能を変動させる要因の一つとなる。**3.1.1**で提案した音素重み係数は正規化されるため、実際のバイアス計算には音素重み係数の値が高い上位 N 音素のみが使用される。そこで、本節では、バイアス計算に使用される音素数と音声認識精度間の関連性を調べる音声認識実験を行った。実験条件は**2.3.1**と同様である。関連性の検討を行った音声認識実験結果を表3に示す。表中の "音素数" は実際にバイアス計算に使用した上位 N 音素の数を示す。

表より、clean、G712、MIRS フィルタに関しては使用する音素数が 1 の場合が最も高い認識性能を示しており、M/A フィルタに関しては使用する音素数が 5 の場合が最も高い認識精度を示していることがわかる。

clean に関しては学習セットとテストセット間に周波数特性の差異がないため、バイアス計算に適した音素がもっとも高い重み係数値となり、音素数 1 で高い認識精度を示したと推測される。そのため、音素数を増加させることにより適切に推定されていた音素の重み係数値が小さくなり、適切なバイアス計算ができなく、認識精度が若干低下したと考えられる。また、G712、MIRS フィルタに関しては、これらのフィルタにより周波数特性の変動を行った場合の認識精度の劣化は M/A フィルタより少ないため、clean と同様にバイアス計算に用いる音素数 1 で適切なバイアスが計算できたためではないかと考えられる。

しかし、M/A フィルタによる周波数特性の変動は大きく、認識精度の低下が著しい。そのため、1 つの正規分布のみで音素を推定することは困難となり、計算されたバイアスが適切に周波数特性を正規化できなかったのではないかと思われる。そのため、5 つの音素を使用してそれらの重み付けによりバイアスを計算することにより、発声全体の周波数特性の正規化が可能になったと推測される。

4. むすび

我々は、分散音声認識 (DSR: Distributed Speech Recognition) における入力系の周波数特性の差異による認識性能劣化に着目をし、それらの周波数特性を正規化する手法について検討した。周波数特性の差異による認識性能劣化の問題は DSR にとって大きな問題の一つであるため、ETSI Advanced DSR フロントエンドに周波数特性正規化手法として Blind Equalization 手法が導入された。Blind Equalization は 1 つの参照ケプストラムを使用し、その参照ケプストラムに発声全体の平均特徴パラメータを近付けるバイアスを計算し、周波数特性の正規化を行う手法である。

本稿では実時間周波数特性正規化手法として、音素平均一致化手法を提案した。本手法は、複数の参照ケプストラムを使用し、周波数特性の正規化を行うバイアスを計算する手法である。ETSI Advanced DSR フロントエンドを用いた日本音響学会新聞記事読み上げ音声コーパスの音声認識実験より、提案手法は、ETSI Advanced DSR フロントエンドにおける Blind Equalization と比較し、周波数特性の変動による音声認識精度劣化の抑制に有効であることを確認した。実際、提案手法は周波数特性の変動に使用した全フィルタに対しても有効であり、全てのフィルタ条件下で ESTI Advanced DSR フロントエンドより、高い認識精度を示した。特に、提案手法は MIRS フィルタ条件下で ESTI Advanced DSR フロントエンド (Blind Equalization) の単語誤り率を 17.88%削減 (16.67%→13.69%) することが可能であった。

しかし、提案手法は平均一致化手法の認識精度より若干の低いことがわかった。バイアス計算に使用する音素が既知の理想的な条件下では提案手法の認識精度が平均一致化手法より高いため、バイアス計算に使用する音素推定精度の向上が今後の課題である。

謝　辞

本研究の一部は文部科学省科学研究費、若手研究 (B)15700163、基盤研究 (B)(2)14350204、国際コミュニケーション基金および電気通信普及財団の補助を受け行った。

文　献

[1] D. Pearce. Enabling new speech driven services for mobile devices: An overview of the ETSI standards activities for distributed speech recognition front-ends. *AVIOS*, 2000.

[2] ETSI ES 201 108 v1.1.2 distributed speech recognition; front-end feature extraction algorithm; compression algorithm. 2000.

[3] ETSI ES 202 050 v1.1.1 STQ; distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. 2002.

[4] F. Liu, R. Stern, X. Huang, and A. Acero. Efficient cepstral normalization for robust speech recognition. In *Proc. DARPA Workshop*, pp. 69–74, March 1993.

[5] L. Mauuary. Blind equalization in the cepstral domain for robust telephone based speech recognition. *EUSPICO*, pp. 359–363, 1998.

[6] 柘植覚, 黒岩眞吾, 北研二. ETSI 標準分散音声認識フロントエンドのための回線特性正規化手法. 音講論, pp. 117–118, 2002.

[7] S. Kuroiwa and S. Tsuge. Blind equalization techniques for ETSI standard DSR front-end. *Proc. ICASSP*, pp. 392–395, 2003.

[8] T. Kato, S. Kuroiwa, and N. Higuchi. Area code, country code, and time difference information system and its field trial. *Proc. IVTTA'98*, pp. 5–10, 1998.

[9] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄. 音声認識システム. オーム社出版局, 2001.

[10] 柘植覚, 黒岩眞吾. 周波数特性の変動に頑健な分散音声認識手法. *SLP-42*, No. 13, pp. 77–84, 2002.

[11] H. Hirsch and D. Pearce. The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. *ISCA ITRW ASR*, pp. 191–188, 2000.