音声言語情報処理 42−13
（２００２．７．１３）

周波数特性の変動に頑健な分散音声認識手法

柘植 覚 黒岩眞吾

徳島大学 工学部

〒 770-8506 徳島市南常三島町 2-1

Tel.: 088-656-7512 e-mail: {tsuge, kuroiwa}@tokushima-u.ac.jp

あらまし 携帯電話の発展にともない急激に携帯端末によるワイアレスモバイル環境の普及が進んでいる。一般に携帯端末は非常に小型であるため、携帯端末に付属する入力デバイスによる操作は困難である。この問題を解決する一方法として、音声による携帯端末操作が考えられる。しかし、携帯端末内のメモリや CPU などのハードウェアは、中・大語彙の音声認識処理の全てを行うまでには至っていない。そこで、音響分析、特徴パラメータの圧縮を携帯端末内で行いサーバに伝送し、サーバで特徴パラメータの復元、音声認識を行う分散音声認識 (DSR: Distributed Speech Recognition) が提案された。分散音声認識では、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、欧州電気通信標準化機構 (ETSI: the European Telecommunications Standards Institute) において、標準化が進められている。本稿では、ETSI 標準分散音声認識フロントエンドを用い日本語連続音声認識実験を行った結果を報告する。同フロントエンドは、特徴パラメータの圧縮にベクトル量子化を用いるため、入力系の周波数特性の差異はベクトル量子化歪みを増加させ、認識精度を低下させる原因となる可能性が高い。そこで、本稿では、入力系の周波数特性の差異によるベクトル量子化歪みを減少させる手法を提案する。音声認識実験結果より、提案手法は周波数特性の差異による認識精度の劣化を低減することが可能であった。

キーワード 分散音声認識、ETST 標準 DSR フロントエンド、乗算性雑音

Robust Feature Extraction in a Variety of Frequency Characteristic on the Basis of ETSI Standard DSR Front-end

Satoru Tsuge Singo Kuroiwa

Department of Information Science & Intelligent Systems
Faculty of Engineering, Tokushima University

2-1 Minami JosanJima-cho, Tokushima, 770-8506

Tel.: 088-656-7512 e-mail: {tsuge, kuroiwa}@is.tokushima-u.ac.jp

Abstract This paper reports an evaluation of European Telecommunications Standards Institute (ETSI) standard Distributed Speech Recognition (DSR) front-end through continuous word recognition on a Japanese speech corpus and proposes a method, the Bias Removal Method (BRM), that reduces the distortion between feature vector and VQ codebook. Experimental results show that using non-quantized features in acoustic model training procedure can improve the recognition performance of DSR fornt-end features and that the proposed method can improve recognition performances of DSR front-end feature.

key words Distributed speech recognition, ETSI standard DSR front-end, Convolution noise

1 はじめに

携帯電話や PDA (Personal Digital Assistants) などの携帯端末の発展にともない急激にワイアレスモバイル環境の普及が進んでいる。一般に、これらの携帯端末は非常に小さいため、携帯端末に付属する入力デバイスを用いたタイピングによる操作が困難である。これを解消する一方法として、音声による携帯端末操作が考えられる。しかし、携帯端末内の CPU、メモリなどの問題で、携帯端末内で中・大語彙の音声認識処理を全て行うことは困難である。

携帯端末を用いた音声認識方法としては、音声を圧

縮しセンターに送り、センターで音声の復元、音声認識を行う方式が提案されている。しかし、このようなセンター型の音声認識方式には、音声を圧縮・復元するコーデックや回線の影響により十分な認識精度が得られないという問題がある [1]。伝送された情報から音声を復元する際に生じる音声の歪みが音声認識精度にあたえる悪影響を軽減するため、伝送された情報から音声の復元を行わず、直接特徴パラメータを抽出する手法が提案されている [2][3][4][5] 。

また、サーバ型音声認識方式のコーデックによる認識性能劣化の問題を解決するため、携帯端末で音響分析、特徴パラメータ圧縮を行いサーバ側に伝送し、サーバ側で特徴パラメータの復元、音声認識処理を行う、分散音声認識 (DSR: Distributed Speech Recognition) が提案されている [6]。音声認識のコンポーネントをサーバと携帯端末のクライアントとに分離する分散音声認識方式には以下の利点がある。

- 分散音声認識方式は、音声を圧縮して伝送するのではなく、音声認識に必要な特徴パラメータのみを伝送する。そのため、音声圧縮・復元に用いられるコーデックがひきおこす音声の歪みを避けることができ、認識性能の改善が期待できる。
- 音声認識に有効な特徴パラメータのみを伝送するため、伝送速度を低く抑えることができる。
- 従来の電話帯域 (300〜3400Hz) に制限されることなく音響分析処理が可能となるため、低・高域の情報を用いる等により、認識性能を向上できる可能性がある。

分散音声認識方式の場合、音響分析部であるクライアント部と音声認識部であるサーバ部で、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 [7]。

現在、その標準化の一環として、雑音に頑健なフロントエンドを勧告するため、欧米では Aurora プロジェクトにおいて整備された雑音データベース [8] を用いた DSR フロントエンドの評価が数多く行われている [9][10][11] 。しかし、日本語連続音声に対する音声認識実験の結果報告は非常に少ない。そこで、本稿では ETSI 標準分散型音声認識フロントエンド (ETSI STQ WI007 version 1.1.2)[7] を用いた日本語連続音声認識実験結果を報告する。

フロントエンドが標準化された場合、パラメータ圧縮に用いられる VQ コードブックも規定される。入力系の周波数特性の差異はベクトル量子化誤差を増加させ、音声認識精度を低下させる原因となる可能性が高い。ETSI 標準分散型音声認識フロントエンドに対し、背景雑音に頑健な特徴量抽出手法について研究は多く報告されている [6][9][10]。しかし、パラメータ圧縮部に

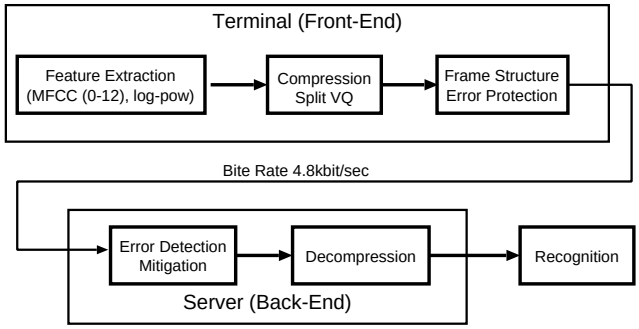

図 1: DSR システム

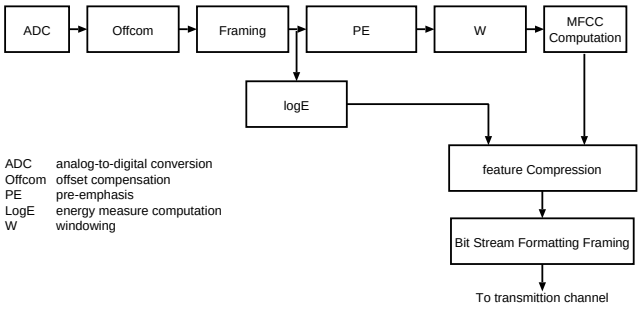

図 2: ETSI 標準フロントエンド

おけるベクトル量子化誤差を考慮した研究はほとんど見られない。そこで、我々は入力系の周波数特性の差異によって生じるベクトル量子化歪みに着目をし、フロントエンド部で歪みを減少させる手法を本稿で提案する。

以下、**2**では分散音声認識手法および ETSI 標準分散型音声認識フロントエンドについて簡単に紹介し、**3**では、本稿で提案する入力系の周波数特性に起因する VQ 歪みを減少させる手法について述べる。**4**では音声認識実験について、**5**において本稿のまとめを述べる。

2 分散音声認識方式

図 1に分散音声認識のブロック図を示す。図に示されるように、分散音声認識システムは、音響分析を行うフロントエンド部と音声認識を行うバックエンド部から構成される。フロントエンド部となるクライアント側では、入力された音声から音声認識に必要な特徴パラメータを分析、圧縮し、伝送路に適するビットストリーム形式に変換を行い、サーバへ伝送する。サーバ側では、伝送されたビットストリームから特徴パラメータを復元し、音声認識を行う。

分散音声認識方式では、フロントエンド部とバックエンド部でビットストリームなどの共通化が必用となる。そのため、広く分散音声認識が利用できるように、ETSI が中心となり、そのフロントエンドの標準化を進めている [7]。現在までに勧告されているフロントエンドの処理過程を図 2 に示す。この図に示すように、フ

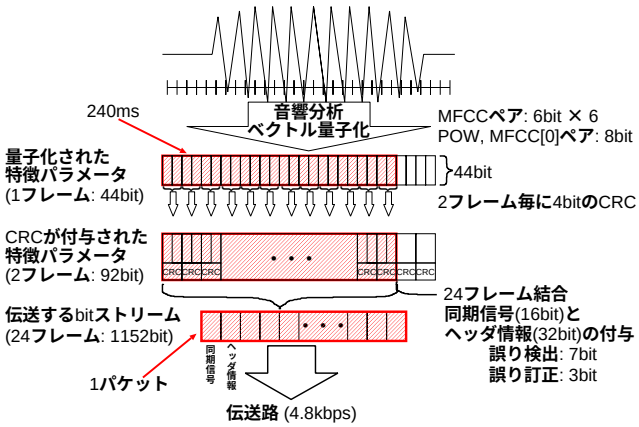

図 3: 特徴パラメータ圧縮部

ロントエンド部は、大きく分類すると特徴パラメータ抽出部、特徴パラメータ圧縮部、ビットストリーム作成部から構成される。以下、各部について簡単に紹介する。詳細については文献 [7] を参考にして頂きたい。

2.1 特徴パラメータ抽出

標準化フロントエンドでは、特徴パラメータとして広く音声認識で利用されている MFCC (Mel Frequency Cepstral Coefficient)、対数パワー、ケプストラム 0 次係数が用いられる。分析条件は次の通りである。標本化周波数は 8, 11, 16kHz の 3 種類が定義されている。分析窓幅は 8kHz, 16kHz の場合 25ms、11kHz の場合 23.27ms であり、切り出しにはハミング窓を用いている。フレーム周期は標本化周波数に関わらず 10ms である。プリエンファシスの係数は 0.97、フィルタバンク数は 23 となっている。その他、以下の特徴がある。

- A/D 変換後の直流成分除去フィルタ
  このフィルタは

$$s_{of}(n) = s_{in} - s_{in}(n-1) + 0.999 * s_{of}(n-1) \quad (1)$$

  で与えられる。ここで、$s_{in}(n), s_{of}(n)$ は入力音声信号、出力音声信号を示す。
- 第一フィルタバンク開始周波数
  第一フィルタバンクの開始周波数は 64Hz であり、64Hz 以下の周波数帯域は分析に利用されない。

2.2 特徴パラメータ圧縮

前節で述べた分析条件で抽出される特徴パラメータは、ベクトル量子化により圧縮される。ベクトル量子化は、特徴パラメータの各次元をペアとした分割ベクトル量子化で行われる。各ペアは、メルケプストラム 1 次と 2 次、3 次と 4 次、...、11 次と 12 次、対数パワーとメルケプストラム 0 次の合計 7 つである。それぞれに対する VQ コードブックとして、標本化周波数 8, 11kHz 用、標本化周波数 16kHz 用の 2 種類が用意されており、コードブックサイズはメルケプストラム係数のペアは 64 (6bit)、対数パワーとケプストラム係数 0 次のペアは 256 (8bit) である。VQ の結果、特徴パラメータは 6bit×6 + 8bit = 44bit で表現される。

2.3 ビットストリーム形式

ベクトル量子化により圧縮された特徴パラメータは、誤り検出情報やヘッダなどの情報を付与したビットストリーム形式へ変換される。ビットストリーム形式に変換する処理を図 3に示す。14 次元の特徴パラメータは、ベクトル量子化により 1 フレームあたり 44bit で表現される。バックエンド部で誤り検出を行うため、2 フレームごとに 4bit の CRC (Cyclic Redundancy Code) が付与される。これらの情報を 24 フレーム (240ms の音声データ分の特徴パラメータ) 結合し、それらに同期信号 (16bit)、ヘッダー情報 (32bit) を付与し、伝送する 1 パッケージとする。1 パッケージあたりは 1,152bit で表現されるため、伝送速度は 4.8kbit/sec となる。

3 VQ 歪み減少化手法

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 標準分散音声認識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ベクトル量子化歪みを増加させ、認識性能を劣化させる原因の一つとなる。

図 4に男性話者 25 発声に対する特徴パラメータ、VQ コードブックを示す。この特徴パラメータは周波数特性を変動させるフィルタリングを行っている。図より、ETSI 標準フロントエンドでは VQ コードブックが規定されているため、VQ コードブック作成データと周波数特性が異なる入力データに対してはベクトル量子化が適切に行えないことがわかる。

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。本稿において、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、VQ 歪み減少化手法を提案する。以下、特徴パラメータと VQ コードブック間の歪みを VQ 歪みとする。

3.1 VQ コードブック平均減算手法

本節では、VQ コードブックと特徴パラメータとの歪みを減少させる一手法として、VQ コードブック平均減算手法を提案する。本手法は、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並行移動する手法である。以下、本手法を BRM1 (Bias Removal Method 1) とする。

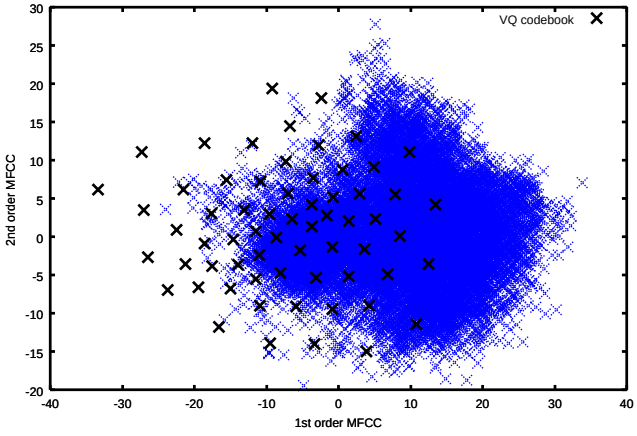

図 4: 特徴パラメータと VQ コードブックとの歪み

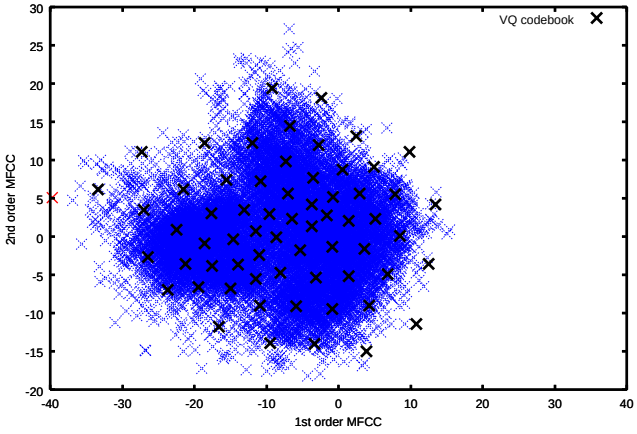

図 5: 提案手法 (BRM1) による特徴パラメータの移動 (MFCC1, 2 次)

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

$$\boldsymbol{a}_{train} = \frac{\sum_{s=1}^{S}\sum_{n=1}^{N_s}\boldsymbol{x}_{sn}}{\sum_{s=1}^{S}N_s} \tag{2}$$

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

$$\boldsymbol{a}_{test} = \frac{\sum_{n=1}^{N}\boldsymbol{x}_n}{N} \tag{3}$$

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. 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

$$\tilde{\boldsymbol{x}}_n = \boldsymbol{x}_n - (\boldsymbol{a}_{test} - \boldsymbol{a}_{train}) \tag{4}$$

ここで、$\tilde{\boldsymbol{x}}_n$ 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特徴パラメータとすることで 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 分散音声認識フロントエンドでは VQ コードブックが与えられるため、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 コードブックの平均を $\boldsymbol{a}_{train}$ として用いる。

図 5に、図 4の特徴パラメータに対し、本提案手法を適用した結果を示す。図より、提案手法を適用することにより特徴パラメータの分布と VQ コードブックの分布が近似し、適切なベクトル量子化が可能となることがわかる。

3.2 歪み誤差最小化手法

本節では、繰り返し演算により VQ 歪みを減少させる歪み誤差最小化手法について述べる。本手法は一般化ロイドアルゴリズムに基づいた手法であり、逐次的に特徴パラメータを平行移動して、VQ コードブックに近似する方法である。以下、本手法を BRM2 (Bias Removal Method 2) とする。

初期特徴パラメータ x_n^0, $(n = 1, \ldots, N$, N は発声内の総フレーム数) と VQ セントロイド判別関数 $q(v)$ が与えられた時、本手法は以下のステップを繰り返すことにより、特徴パラメータと VQ コードブック間の歪みを減少させる。

1. 特徴パラメータ x_n^i と VQ コードブック間の歪み d_n を計算する。

$$d_n = (x_n^i - q(x_n^i))^2 \tag{5}$$

ここで、i は繰り返し回数を示す。$q(v)$ は v に最近の VQ コードを返す関数である。

2. 認識発声全フレームに対し、以下の式で VQ 歪みの総和を計算する。

$$D = \sum_{n=1}^{N} d_n \tag{6}$$

$$= \sum_{n=1}^{N}(x_n^i - q(x_n^i))^2 \tag{7}$$

歪み D が閾値以下ならば、x_n^i を特徴パラメータ圧縮部に送る。

3. VQ 歪みが最小となる差分値 h を計算する。

$$\tilde{D} = \sum_{n=1}^{N}((x_n^i - h) - q(x_n^i)) \quad (8)$$

$$\frac{\partial \tilde{D}}{\partial h} = \frac{\partial(\sum_{n=1}^{N}((x_n^i - h) - q(x_n^i))^2)}{\partial h} \quad (9)$$

$$h = \frac{\sum_{n=1}^{N} x_n^i - q(x_n^i)}{N} \quad (10)$$

4. 3で算出した h を特徴パラメータから減算する。

$$x_n^{i+1} = x_n^i - h \quad (11)$$

$i = i + 1$ とし、1に戻る。

VQ コードブック平均減算手法と同様に図 4の特徴パラメータに本提案手法を適用した結果、VQ コードブック平均減算手法と同様に VQ コードブックと特徴パラメータの分布が近似した。

4 音声認識実験

提案手法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)[12] を用い、

- ETSI 標準分散音声認識フロントエンドの評価
- 認識時に提案手法を適用する実験
- 認識・学習時に提案手法を適用する実験
- 提案手法と CMS を併用する実験

を行った。

4.1 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。

ETSI 標準分散音声認識フロントエンドにより音響分析、ベクトル量子化、復号化を行った特徴パラメータ MFCC12 次元、MFCC の一次回帰係数 12 次元、対数パワーの一次回帰係数の合計 25 次元を特徴ベクトルとして用い、音響モデルの学習を行った。ベクトル量子化による特徴パラメータ圧縮が認識精度へ与える影響を調べるため、ベクトル量子化を行っていない MFCC を用い音響モデルの学習を行った。また、音声を符号化しセンターに伝送し、センターで認識を行う従来の分散音声認識手法との比較のため、音声符号化手法の一つであり IP 電話などのコーデックに使用されている G723.1 (5.3kbps)[13] で圧縮・復号した音声に対し、分析を行った量子化なし特徴ベクトルを用い音響モデルの学習を行った。本実験では、提案手法 BRM2 の繰り返し回数は 100 回とした。

周波数特性の差異の影響を検討するため、JNAS の音声データを以下のフィルタに通し、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。

- 高域追加フィルタ (H/P)

$$s_{of}(n) = s_{in}(n) - 0.9 \times s_{in}(n-1) \quad (12)$$

- 移動平均フィルタ (M/A)

$$s_{of}(n) = 0.25 \times (s_{in}(n) + s_{in}(n+1) + s_{in}(n+2) + s_{in}(n+3)) \quad (13)$$

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[1]。総状態数は各特徴量ともに約 1000 状態である。

デコーダには Julius を用い、評価は式 (14) で与えられる単語誤り率 (WER: Word Error Rate) で行った。

$$\mathrm{WER} = \frac{I + S + D}{N} \cdot 100(\%), \quad (14)$$

I, D, S はそれぞれ、挿入誤り数、削除誤り数、置換誤り数を示す。また、N は全認識単語数を示す。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。

4.2 認識実験結果・考察

4.2.1 ETSI 標準分散音声認識フロントエンドを用いた音声認識実験

ETSI 分散音声認識フロントエンドで分析し、圧縮・復元した特徴パラメータ (以下、*VQ* とする) を音響モデル学習時、認識時に用いた音声認識実験を行った。実験結果を表 1に示す。

表中の特徴パラメータは、以下である。

- *VQ*: VQ を行った特徴パラメータ
- *no VQ*: VQ を行っていない特徴パラメータ
- *G723-noVQ*: G723.1 で圧縮・復元を行った音声を分析した VQ を行っていない特徴パラメータ

[1]分割 VQ により特徴パラメータが離散化するため、離散分布 HMM が適していると考えられるが、26 次元の特徴ベクトルを表す VQ コードブックは膨大となる。そのため、本稿では離散分布 HMM ではなく、連続混合分布 HMM を用いた

表 1: ETSI 標準分散音声認識フロントエンドを用いた認識結果 (WER (%))

	特徴パラメータ		標本化周波数	
	学習時	評価時	8kHz	16kHz
(1)	*noVQ*	*noVQ*	13.89	12.1
(2)	*VQ*	*VQ*	49.33	18.44
(3)	*noVQ*	*VQ*	13.52	12.24
(4)	*G723-noVQ*	*G723-noVQ*	19.86	–

VQ が認識性能に与える影響　表 1の (1), (2) の比較より、ベクトル量子化による特徴パラメータの圧縮は WER を増加させている。特に、標本化周波数 8kHz で *VQ* を音響モデル学習、認識に用いた場合、WER が著しく増加していることがわかる。この原因は、ベクトル量子化により特徴パラメータが離散化されるため、音響モデルの連続混合分布の学習が十分にできていないと考えらる。実際、*VQ* で学習した音響モデルには、学習データ不足のためフロアリングされた分布が存在した (標本化周波数 8kHz: 1041 混合、標本化周波数 16kHz: 130 混合)。

学習時における VQ の影響　音響モデル学習時における特徴パラメータの離散化の影響を調べるため、*noVQ* で学習を行った音響モデルを用い、*VQ* の認識を行った (表 1 (3))。この結果、量子化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、量子化された特徴パラメータの認識性能を量子化を行わない場合 (表 1 (1)) とほぼ同等にすることができた。これは、4.8kbps という低い伝送速度で量子化を行わない音声 (128kbps) とほぼ同程度の音声認識精度を達成できることを示している。

G723.1 との比較　表 1 (3) の WER は、G723.1 を用い圧縮、復元した音声の認識性能 (表 1 (4)) より低いことがわかる。G723.1 では伝送される情報の多くをピッチ予測係数等に割いているため、音声認識に必要な情報が欠落している可能性がある。一方、ETSI 標準分散音声認識フロントエンドでは、伝送するデータとして音声認識に有効な特徴パラメータのみを圧縮し用いている。このような圧縮方法の違いが WER に影響したと考えられる。以上より、特徴パラメータを圧縮・伝送する ETSI 標準分散音声認識フロントエンドが分散音声認識には有効であることがわかった。

さらに、表 1 (3) より、標本化周波数 16kHz の WER が 8kHz より低いことがわかる。これは、伝送速度が同じでも分析帯域を広げることで WER を減少できることを示唆している。

4.2.2　提案手法の有効性

提案手法を認識時に適用する実験を行った。

表 2: 標本化周波数 16kHz での提案手法の有効性 (WER (%))

		フィルタ		
	VQ 歪み減少化手法	なし	H/P	M/A
(1)	適用なし *w/o VQ*	<u>12.1</u>	14.77	29.23
(2)	適用なし *with VQ*	12.24	17.69	32.15
(3)	*BRM1*	12.48	13.43	14.32
(4)	*BRM2*	14.01	<u>12.94</u>	<u>14.01</u>
(5)	*BRM1+2*	13.52	13.94	14.33

表 3: 標本化周波数 8kHz での提案手法の有効性 (WER (%))

		フィルタ		
	VQ 歪み減少化手法	なし	H/P	M/A
(1)	適用なし *w/o VQ*	13.89	<u>21.83</u>	52.34
(2)	適用なし *with VQ*	<u>13.52</u>	23.29	58.17
(3)	*BRM1*	26.02	24.35	26.44
(4)	*BRM2*	24.06	23.22	<u>21.76</u>
(5)	*BRM1+2*	25.11	25.24	25.49

実験条件　前節の実験結果より、量子化を行った特徴パラメータの認識には、量子化を行わない特徴パラメータを音響モデル学習に用いることが有効であることがわかった。そこで、本節の実験では、量子化を行わない特徴パラメータで学習した音響モデルを用いた。

実験結果　音声認識実験結果を表 2、3に示す。表 2は標本化周波数 16kHz、表 3は標本化周波数 8kHz の結果である。これらの表の (1) は提案手法を適用せず、量子化を行わない特徴パラメータを認識した結果、(2) は量子化を行った特徴パラメータを認識した結果を示す。(3), (4) は提案手法 BRM1 と BRM2 をそれぞれ適用した結果、(5) は、BRM1 を適用後、さらに BRM2 を適用した結果を示す。

表より、M/A フィルタにより周波数特性を変更させた音声に対しては、提案手法のいずれにおいても、WER を低減させることができた。また、標本化周波数 16kHz においては、H/P フィルタリングを行った音声に対しても、同様に WER の軽減が可能であった。しかし、フィルタリングを行っていない音声に対しては、提案手法は WER を低減させることができなかった。特に、標本化周波数 8kHz の場合には著しい WER の増加が見られる。提案手法を併用した場合 (認識結果 (5)) においては、それぞれを単独で使用した場合とほぼ同程度の認識性能しか得られなかった。

考察　提案手法は ETSI 標準分散音声認識フロントエンドで与えられた VQ コードブックとの歪みが減少するように特徴パラメータを平行移動する手法である。そ

のため、VQ コードブック作成データと音響モデル学習データが異なっている場合、提案手法は認識時の特徴パラメータと音響モデルのパラメータ間にずれを生じさせ、認識性能の劣化を引き起こす原因となったと考えられる。

実際に本節で示した実験結果のうち、提案手法による WER の低減が可能であった所は、フィルタリングにより大きく認識精度が劣化している所であった (標本化周波数 16kHz H/A, M/A フィルタ, 標本化周波数 8kHZ M/A フィルタ)。逆に、提案手法により WER が増加したところは、フィルタリングを行ってもさほど認識性能が劣化していないことがわかる。これは、提案手法により認識時の特徴パラメータを VQ コードブックに近づけることが音響モデルパラメータとのずれを生じさせ、その結果、WER の増加につながったと考えられる。この結果より、音響モデル学習データと VQ コードブックの歪みを減少させるため、提案手法を音響モデル学習データにも適用する必要があると考えられる。

4.2.3 音響モデル学習時にも提案手法を適用

前節の結果より、認識時にのみ提案手法を適用することは、認識時の特徴パラメータと音響モデルのパラメータ間にずれを生じさせ、認識性能を劣化させる原因となることが推測された。そこで、提案手法によるパラメータのずれを減少するため、提案手法を音響モデル学習データにも適用する認識実験を行った。

実験結果　認識実験結果を表 4, 5に示す。表 4は標本化周波数 16kHz の認識結果であり、表 5は標本化周波数 8kHz の結果である。表中の (1)〜(4) は認識時にのみ提案手法を適用した結果 (表 2, 3の (2)〜(5)) の再掲であり、(5)〜(7) は提案手法を学習データに対しても適用した認識結果を示す。

考察　表より、学習時に提案手法を適用することで全ての結果において、WER を低減していることがわかる。提案手法を認識時のみに適用した場合には、適用なし (表中 (1)) と比較して WER の増加が見られたフィルタリングを行っていない発声に対しても、学習データに提案手法を適用した場合には WER の増加がみられず、むしろ減少することができた。この結果は、ベクトル量子化を行わない 128kbps 音声の認識精度 (表 2, 3 (1)) より低い WER を示した (誤り改善率: BRM1 の場合 10.1%、BRM2 の場合 9.8%)。

また、フィルタリングにより乗算性雑音を付与した音声に対しても、学習データに提案手法を適用することは WER の低減につながることがわかる。この結果はフィルタリング、量子化を行わないクリーンな音声の認識結果 (表 2, 3 (1)) より、低い WER を示している。提案手法を音響モデル学習時、認識時に適用することにより、本稿で用いたフィルタによる乗算性雑音を加えた環境下では、フィルタによる認識性能の劣化をほぼ抑制することが可能であるとこが実験結果から言える。

表 4: 標本化周波数 16kHz における提案手法の学習時適用の有効性 (WER (%))

	適用手法		filter		
	学習時	認識時	なし	H/P	M/A
(1)	適用なし	適用なし	12.24	17.69	32.15
(2)	適用なし	*BRM1*	12.48	13.43	14.32
(3)	適用なし	*BRM2*	14.01	12.94	14.01
(4)	適用なし	*BRM1+2*	13.52	13.94	14.33
(5)	*BRM1*	*BRM1*	10.39	9.95	10.15
(6)	*BRM2*	*BRM2*	10.02	9.96	11.16
(7)	*BRM1+2*	*BRM1+2*	9.25	9.63	9.96

表 5: 標本化周波数 8kHz における提案手法の学習時適用の有効性 (WER (%))

	適用手法		filter		
	学習時	認識時	なし	H/P	M/A
(1)	適用なし	適用なし	13.52	23.29	58.17
(2)	適用なし	*BRM1*	26.02	24.35	26.44
(3)	適用なし	*BRM2*	24.06	23.22	21.76
(4)	適用なし	*BRM1+2*	25.11	25.24	25.49
(5)	*BRM1*	*BRM1*	10.78	10.72	12.3
(6)	*BRM2*	*BRM2*	10.91	11.1	12.37
(7)	*BRM1+2*	*BRM1+2*	10.53	11.35	12.56

さらに、認識時のみに提案手法を適用した場合には有効性が確認できなかった提案手法の併用であったが、学習時より提案手法を適用することにより、若干ではあるが WER を減少させることができた。

4.2.4 CMS と提案手法の併用

本節では、乗算性雑音対策として広く一般に用いられている CMS (Cepstrum Mean Subtraction) との比較を行う。また、提案手法と CMS の併用による有効性について調べた。本実験では、CMS は発声ごとに行った。

ETSI 標準分散音声認識フロントエンドでは、VQ コードブックが規定されている。このコードブックは CMS を行わない特徴パラメータで作成されているため、量子化を行わない特徴パラメータに対し、CMS を行うことは VQ 歪みを増加させる。そこで、CMS は後段のサーバ側で行うと仮定し、量子化を行った特徴パラメータに対し CMS を行った。また、提案手法と CMS の併用は、提案手法を適用後、量子化を行った特徴パラメータに対し、CMS を行う方法とした。認識実験結果を表 6 (標本化周波数 16kHz)、表 7 (標本化周波数 8kHz) に示す。

表 6: CMS を行った場合の提案手法の学習時適応の有効性 (標本化周波数 16kHz, WER (%))

	適用手法		filter		
	学習時	認識時	なし	H/P	M/A
(1)	VQ なし	VQ なし	9.63	9.44	9.32
(2)	適用なし	適用なし	9.58	9.38	11.22
(3)	*BRM1*	*BRM1*	10.46	10.27	10.01
(4)	*BRM2*	*BRM2*	9.63	9.19	9.38
(5)	*BRM1+2*	*BRM1+2*	9.89	9.63	10.53

表 7: CMS を行った場合の提案手法の学習時適応の有効性 (標本化周波数 8kHz, WER (%))

	適用手法		filter		
	学習時	認識時	なし	H/P	M/A
(1)	VQ なし	VQ なし	10.4	10.46	11.67
(2)	適用なし	適用なし	10.78	10.59	14.14
(3)	*BRM1*	*BRM1*	10.84	10.65	11.98
(4)	*BRM2*	*BRM2*	10.65	10.72	11.29
(5)	*BRM1+2*	*BRM1+2*	10.65	10.08	11.67

提案手法を適用した結果 (表 4, 5 (5)〜(7)) との比較により、提案手法は CMS とほぼ同程度の WER を示した。フィルタにより WER が増加する M/A フィルタにおいては、CMS より提案手法が低い WER を示した。これは、後段の CMS では VQ 歪みの増加を低減することが困難であるからだと推測できる。提案手法と CMS を併用することは併用を行わない場合と比較し、ほぼ同等もしくは低い WER を示し、CMS と提案手法の併用は有効であると言える。

5 むすび

本稿では、日本語連続音声データベース (JNAS) に対する、ETSI 標準分散音声認識フロントエンドを用いた音声認識実験結果を報告した。また、量子化を行う際の特徴パラメータと VQ コードブックとの歪みを減少させる、VQ 歪み減少化手法を提案した。

実験結果より、音響モデル学習には量子化を行わない特徴パラメータを用いることで、量子化された特徴パラメータの認識精度を向上させることが可能であった。また、音響分析帯域を広げることにより、認識精度が向上することがわかった。

提案手法を音響モデル学習時、および認識時の双方で適用することにより、量子化による認識性能への影響を軽減することができた。量子化された特徴パラメータを認識する場合には、提案手法は CMS より高い認識性能を示した。また、CMS を用いる場合においても提案手法と併用する方が高い認識性能を得ることができる。

本稿の実験では、一発声が終了した後、提案手法の適用を行っているが、今後はフレーム同期で適用する手法の検討を行う予定である。また、話者認識において、ETSI 標準分散音声認識フロントエンドで分析された特徴パラメータの評価を行う予定である。

参考文献

[1] B. Lilly and K. Paliwal. Effect of speech coders on speech recognition performance. *Proc. ICSLP*, pp. 2344–2347, 1996.

[2] B. Raj, J. Migdal, and R. Singh. Distributed speech recognition with codec parameters. *ASRU*, 2001.

[3] J. Huerta and R. Stern. Speech recognition from GSM codec parameters. *Proc. ICSLP*, pp. 1463–1466, 1998.

[4] H. Kim and S. Member. A bitstream-based front-end for wireless speech recognition on IS-136 communications system. *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 5, pp. 558–568, 2001.

[5] T. Uchibe, S. Kuroiwa, and N. Higuchi. The method to translate codes of Cs-Acelp into acoustic parameters for speech recognition. *Proceedings of the 2000 IEICE General Conference*, Vol. 6, p. 195, 2000. (in Japanese).

[6] 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.

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

[8] 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.

[9] B. Noe, J.Sienel, D.Jouvet, L.Mauuary L. Boves, J. Veth, and F. Wet. Robust feature extraction for distributed speech recognition. *Proc. EuroSpeech*, pp. 433–436, 2001.

[10] Carmen Benitez, Lukas Burget, Barry Chen, Stephane Dupont, Hari Garudadri, Hynek Hermansky, Pratibha Jain, Sachin Kajarekar, and Sunil Sivadas. Robust ASR front-end using spectral-based and discriminant features: experiments on the Aurora tasks,. *Proc. EuroSpeech*, 2001.

[11] D. Macho and C. Nadeu. Comparison of spectral derivative parameters for robust speech recognition. *Proc. EuroSpeech*, 2001.

[12] 音声認識システム. オーム社出版局, 2001.

[13] ITU-T recommendation G.723.1 dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s. 1996.