ETSI 標準分散音声認識フロントエンドにおける 入力系の周波数特性正規化手法

非会員 柘植 覚*
正 員 黒岩 眞吾*

非会員 獅々堀 正幹*
非会員 北 研二**

Channel Compensation Method for ETSI Standard DSR Front-end

Satoru Tsuge*, Non-member, Singo Kuroiwa*, Member, Masami Shishibori*, Non-member,
Kenji Kita**, Non-member

This paper reports an evaluation of European Telecommunications Standards Institute (ETSI) standard Distributed Speech Recognition (DSR) front-end through continuous speech recognition on a Japanese speech corpus and proposes methods, the Bias Removal Methods (BRMs), that reduce the distortion between feature parameters and the VQ codebook. Experimental results show that (1) using non-quantized features in an acoustic model training procedure can improve the recognition performance of DSR front-end features and (2) broadening the analysis band can improve the recognition performance for the same bitrate. The proposed method can improve the recognition performance in DSR condition. Notably, we observed an 18% relative improvement in the error rate using the proposed method under mismatch of channel characteristic conditions.

キーワード：分散音声認識，ETSI 標準分散音声認識フロントエンド，入力デバイスの周波数特性，量子化歪み

Keywords: Distributed speech recognition, ETSI standard DSR front-end, Frequency characteristic of input device, VQ distortion

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 (CPU，メモリなど) 面の制約により、電子メール文入力などの中・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。そこで、携帯端末を用い

た中・大語彙音声認識方法として、通常の電話回線を介し音声センターに伝送し、センターで音声認識を行う方式 (以下、コーデック方式と表記) が実用化されている。しかし、伝送時に音声圧縮・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⁾。

そこで、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を行いサーバに伝送し、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⁽²⁾。DSR には以下の利点がある。

- 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ため、音声圧縮・復元時の音声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 (300～3,400Hz) に制限されることなく音響分析処理が可能となるため、低域・高域の情報を生かす等により、認識性能を向上できる可能性がある。

以上のように利点が多い DSR だが、同方式を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

* 徳島大学 工学部
〒 770-8502 徳島市南常三島町 2-1
Faculty of Engineering, the University of Tokushima
2-1 Minami-Josanjima, Tokushima, 770-8502 Japan

** 徳島大学 高度情報化基盤センター
〒 770-8502 徳島市南常三島町 2-1
Center for Advanced Information Technology, the University of Tokushima
2-1 Minami-Josanjima, Tokushima, 770-8502 Japan

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⁽³⁾。

標準化の一環として、2000 年 4 月に ETSI は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 108 version 1.1.2) を勧告した⁽³⁾。このフロントエンドの評価は、AURORA プロジェクトにおいて整備された雑音データベース⁽⁴⁾を用い、数多く行われている⁽⁵⁾⁽⁶⁾⁽⁷⁾。これらは、欧米中心の言語に対する数字音声認識タスクによる評価であり、小規模なタスクである。そこで、より大規模な連続音声認識タスクでの評価を進めて行く必要があり、AURORA プロジェクトにおいてもデータベースが整備され、評価が開始された⁽⁸⁾。また、欧米中心の言語のみの評価では、それらの言語のみに適した標準化がなされる可能性が高いため、日本語での評価を進める必要がある。実際、日本においても AURORA2 データベースを日本語化したデータベースを作成し、AURORA プロジェクトと同等の評価を行う計画が進んでいる⁽⁹⁾。

そこで本論文ではまず、ETSI 標準 DSR フロントエンド (ETSI ES201 108 version 1.1.2)⁽³⁾を用いた日本語連続音声認識実験を行い、日本語連続音声認識における同フロントエンドの問題点を明確にする。以下、ETSI 標準 DSR フロントエンドを DSR フロントエンドと表記する。

さらに、我々は明確となった問題点のうち入力デバイスの周波数特性の差異の問題に着目し研究を進めた。様々な携帯端末が存在しているため、使用されている入力デバイスの周波数特性は様々である。このような入力デバイスの周波数特性の差異は音声認識精度を低下させる原因の一つとなることが知られている⁽¹⁰⁾。特に DSR フロントエンドでは、ベクトル量子化 (VQ: Vector Quantization) による特徴パラメータ圧縮が行われているため、入力デバイスの周波数特性の差異は VQ 歪みを増大させる原因ともなり、認識性能劣化につながる。そこで、我々は入力系の周波数特性の差異により生じる VQ 歪みに着目し、フロントエンド部で歪みを減少させる手法を本論文で提案する。

以下、2. では DSR フロントエンドの簡単な説明を行い、3. では、日本語連続音声コーパスを用いた DSR フロントエンドの評価を行い、その問題点を明確にする。また、4. では、入力デバイスの周波数特性の差異により生じる VQ 歪みを軽減する手法として周波数特性正規化手法を提案し、その有効性を 5. で検証する。最後に、6. で本論文のまとめと今後の課題を述べる。

2. ETSI 標準 DSR フロントエンド

本章では、DSR システムおよび DSR フロントエンドについて述べる。

2.1 DSR フロントエンド 図 1 に DSR システムのブロック図を示す。DSR システムは、音響分析を行うフ

図 1 DSR システムのブロック図

Fig. 1. DSR System⁽³⁾.

図 2 ETSI 標準 DSR フロントエンド

Fig. 2. ETSI standard DSR front-end⁽³⁾.

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抽出、圧縮し、伝送路に適したビットストリーム形式に変換後、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。

図 2 に、ETSI が勧告した DSR フロントエンドの処理過程を示す。フロントエンドは、図 1 に示した通り、特徴パラメータ抽出部 (Feature Extraction)、特徴パラメータ圧縮部 (Compression)、ビットストリーム構成部 (Frame Structure) 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⁽³⁾を参考にしたい。

2.2 特徴パラメータ抽出 DSR 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。標準化周波数は 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。フレーム周期は標準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

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

で与えられる。ここで、 $s_{in}(n)$, $s_{of}(n)$ は各々入力音声

図 3 特徴パラメータ圧縮部

Fig. 3. Feature compression.

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・3 特徴パラメータ圧縮 DSR では特徴パラメータをサーバに伝送する。そこで，伝送効率を良くするためフロントエンド部において特徴パラメータの圧縮を行っている。DSR フロントエンドは，圧縮手法として VQ を用いている。VQ 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，MFCC 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，標準化周波数 8, 11kHz 用，標準化周波数 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各ペアで 64 (6bit)，対数パワーとケプストラム係数 0 次のペアで 256 (8bit) である。VQ の結果，特徴パラメータは $6\text{bit} \times 6 + 8\text{bit} = 44\text{bit}$ で表現される。

2・4 ビットストリーム形式 VQ 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，VQ 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) の特徴パラメータ 結合し，それらに同期信号 (16bit)，ヘッダ情報 (32bit) を付与し，伝送する 1 パッケージとする。1 パッケージは 1,152bit で表現されるため，伝送速度は 4.8kbit/sec となる。

3. 日本語連続音声コーパスを用いた DSR フロントエンドの評価

DSR フロントエンドの評価のために，日本音響学会新

聞読み上げコーパス (JNAS: Japanese newspaper article sentences speech corpus of ASJ)⁽¹¹⁾ を用いた連続音声認識実験を行った。評価を行った項目は，

- (1) コーデック方式の音声認識との比較 (3・2)
- (2) VQ が認識精度に与える影響 (3・3)
- (3) 音響モデル学習時における VQ の影響 (3・4)
- (4) 入力系の周波数特性の差異の影響 (3・5)

である。

(1) の評価では，従来のコーデック方式 (G723.1) の音声認識と DSR 方式を比較することにより，DSR 方式の有効性を確認する。(2)(3) の評価では，DSR 方式で用いられている特徴パラメータ圧縮手法 (VQ) が認識精度に与える影響を調べ，その問題点を明らかにする。最後に (4) の評価において，入力デバイスの周波数特性の差異が認識性能に与える影響を明らかにする。

3・1 評価実験条件 音響モデルの学習には，JNAS より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。評価には，学習に用いていない男性話者 23 名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，評価に用いた特徴量は，DSR フロントエンドで分析した MFCC12 次元，その一次回帰係数 12 次元，対数パワーの一次回帰係数，合計 25 次元を用いた。これらの特徴量は，1 発声単位で CMS (Cepstrum Mean Subtraction) を行っている。また，コーデック方式の音声認識との比較のため，IP 電話などのコーデックに使用されている G723.1 (5.3kbps)⁽¹²⁾ で圧縮・復元した音声を用いた。

音響モデルは，比較する特徴量毎に木構造クラスタリングを行い，状態共有した 3 状態 16 混合の音素環境依存 HMM (43 音素) を学習した。総状態数は各特徴量とともに約 1,000 状態である。デコードには Julius⁽¹³⁾ を用い，式 (2) に示す単語誤り率 (WER: Word Error Rate) で評価を行った。

$$\text{WER} = \frac{I + S + D}{N} \cdot 100(\%), \dots\dots\dots (2)$$

ここで， I , D , S , N は挿入誤り数，削除誤り数，置換誤り数，正解単語数を示す。本実験で使ったデコードはビームサーチを行っている。そこで，特徴パラメータごとの認識結果は，最高の認識性能を示したビーム幅の結果を用いた。

3・2 コーデック方式音声認識との比較 コーデック (G723.1) により音声を圧縮，伝送する方式と特徴パラメータを伝送する方式の音声認識精度を比較する音声認識実験を行った。実験結果を表 1 に示す。表中の G723.1 は G723.1 により圧縮し，復元した音声に対し音声認識を行った結果を示し，DSR front-end は DSR フロントエンドにより抽出された特徴パラメータを圧縮，復元して認識を行った結果を示す。

表 1 より，DSR フロントエンドを用いた WER は，G723.1 を用い圧縮，復元した WER より低いことがわかる。この結果より，分散音声認識においては，フロントエン

表 1 ETSI 標準 DSR フロントエンドと G723.1 の認識精度比較 (WER (%))

Table 1. Compared recognition performances of ETSI standard DSR front-end and G723.1 (WER in %).

Feature vector	Sampling rate	
	8kHz	16kHz
DSR front-end	15.3	10.7
G723.1	18.6	—

表 2 VQ が認識精度に与える影響 (WER (%))

Table 2. Affect of Vector Quantization (WER in %).

Feature vector	Sampling rate	
	8kHz	16kHz
with VQ	15.3	10.7
w/o VQ	10.4	9.6

ド側で特徴パラメータ抽出を行い、特徴パラメータを圧縮、伝送する方式が、音声圧縮伝送するコーデック方式より、高い認識精度が得られることがわかる。この理由として、次のことが考えられる。G723.1 では、包絡情報 (LSP パラメータ) と励振情報 (ピッチ予測係数など) がサーバに伝送される。伝送される情報の多くはピッチ予測係数等に割かれ、音声認識に必要な包絡情報の一部が欠落している可能性がある。実際、本実験で用いた G723.1 (5.3kbps) では、音声認識に有効とされている包絡情報 (LSP パラメータ) は伝送される情報の内約 15% しか占めていない⁽¹²⁾。一方、DSR フロントエンドでは、音声認識に有効な特徴パラメータのみを圧縮しており、情報の欠落が少ない。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。

3.3 VQ が認識性能に与える影響 DSR フロントエンドでは、特徴パラメータを伝送するため VQ による特徴パラメータ圧縮を行っている。本節では、この VQ が音声認識精度に与える影響を調べる。

表 2 に音声認識実験結果を示す。表中の “with VQ” は VQ により圧縮し、復元した特徴パラメータを音響モデル学習、認識に用いた結果を示す。この結果は表 1 “DSR front-end” の再掲である。また、“w/o VQ” は VQ による圧縮を行わない特徴パラメータで音響モデルを学習し、認識した結果を示す。

この表より、VQ により特徴パラメータを圧縮することが WER を増加させていることがわかる。特に、標準化周波数 8kHz では、VQ が WER を著しく増加させている。この原因は、VQ による歪みの増大もしくは、VQ により特徴パラメータが離散化され、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ためと考えられる。このうち後者は、“with VQ” で音響モデルを学習した際、分散のフロアリングが必要な場合が多々存在することが確認された。これは、各 VQ セントロイド周辺で分散していた特徴パラメータが VQ によりセントロイドに集約され、分散の推定ができなくなりフロアリングされたと考えられる。この結果、正確な分散が推定されないため WER 増加の原因

表 3 ETSI 標準 DSR フロントエンドを用いた認識結果 (WER (%))

Table 3. Recognition performances of ETSI standard DSR front-end (WER in %).

	Feature vector		Sampling rate	
	training	testing	8kHz	16kHz
(1)	w/o VQ	with VQ	10.8	9.6
(2)	with VQ	with VQ	15.3	10.7
(3)	w/o VQ	w/o VQ	10.4	9.6

となったことが予想される。

3.4 学習時における VQ の影響 前節において、VQ は認識精度の劣化の原因となることがわかった。その中でも、音響モデル学習時における特徴パラメータの離散化が大きな原因であると考えられた。そこで、本節では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、VQ を行わない特徴パラメータ (“w/o VQ”) で学習した音響モデルを用い、VQ による圧縮された特徴パラメータ (“with VQ”) の認識を行った。認識実験結果を表 3 に示す。表中の (2)、(3) は表 2 の再掲である。

表 3 の結果より、VQ 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結果 (表 3 (1)) は、学習に VQ を行った特徴パラメータを用いた場合の結果 (表 3 (2)) の WER を大きく削減していることがわかる。この結果は、学習、認識に特徴パラメータの圧縮を行わない結果 (表 3 (3)) とほぼ同等である。これより、認識時の VQ 歪みの問題は少ないことがわかる。さらに、表 3 (1) より、標準化周波数 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。

以上の実験結果より、4.8kbps という低い伝送速度で VQ を行わない音声とほぼ同程度の音声認識精度を達成できることがいえる。また、回線などの分析帯域の制限を受けない DSR では、分析帯域を広げることで認識精度のさらなる向上が期待できる。

DSR フロントエンドでは、2.2 で述べた通り特徴パラメータ抽出方法は広く一般に日本語音声認識に用いられる方法と同様であるが、特徴パラメータ圧縮部で使用されている VQ コードブックは欧州言語中心で作成されているため、日本語音声の特徴パラメータの分布を適切に表現できていない可能性があった。しかし本節の実験結果より、VQ 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ることにより、VQ を行った特徴パラメータの認識精度 (表 3 (1)) を VQ を行わない特徴パラメータの認識精度 (表 3 (3)) とほぼ同等にすることができるとわかった。これより、DSR フロントエンドで使用されている VQ コードブックは日本語音声の特徴パラメータの分布を適切に表現できているといえ、DSR フロントエンドを日本語音声認識に用いる問題はないと結論できる。

以下の実験では、VQ を行わない特徴パラメータで学習を行った音響モデルを用いる。

(a) 1st and 2nd order MFCC (w/o filter)

(b) 1st and 2nd order MFCC (with filter)

図 4 特徴パラメータと VQ コードブックとの歪み

Fig. 4. The distortion between feature vector and VQ codebook.

3.5 入力系の周波数特性の差異の影響 DSR システムが使用される携帯端末の種類は多種多様なため、個々の端末で使用される入力デバイスの違いにより入力系の周波数特性には差異が生じる。本節では、入力系の周波数特性の差異が音声認識精度へ与える影響を調べる。

図 4 (a) に DSR フロントエンドで分析した JNAS 中の男性話者 25 発声の特徴パラメータ (MFCC1 次, 2 次) と ETSI により勧告されている VQ セントロイドの散布図を示す。また、図 4 (b) に移動平均フィルタ (式 (14)) により人為的に周波数特性を変更した音声データから分析した特徴パラメータの散布図を示す。図 4 (a) の場合では、VQ セントロイドが特徴パラメータの分布を比較的適切に表現しているため、VQ 歪みを小さく抑えることが可能である。しかし、図 4 (b) の場合、VQ セントロイドと特徴パラメータの分布間にずれが生じているため、VQ 歪みが增大することがわかる。この周波数特性の差異により引き起こされた VQ 歪みの増大は認識性能劣化の原因の一つとなると考えられる。

3.5.1 音声認識実験 本節では、入力系の周波数特性の差異による VQ 歪みの増大が認識性能に及ぼす影響を調べる実験を行った。入力系の周波数特性の差異の検討

表 4 入力デバイスの周波数特性の影響 (WER (%))

Table 4. Influence of the frequency characteristic in 16 kHz sampling (WER in %)

	Sampling rate			
	8kHz		16kHz	
Feature vector	no	M/A	no	M/A
w/o VQ	10.4	11.7	9.6	9.3
with VQ	10.8	14.1	9.6	11.2

には、式 (14) に示す移動平均 (M/A) フィルタを用いた。

実験結果を表 4 に示す。表中に示す、no はフィルタなしの場合、M/A は M/A フィルタを用いた場合の認識結果を示す。フィルタなしの条件における、“with VQ” は表 3 (1)、“w/o VQ” は表 3 (3) の再掲である。

この結果より、VQ を行わない特徴ベクトル (w/o VQ) の WER は、M/A フィルタによる周波数特性の差異に対しても増加がみられない。これは、サーバ側で特徴パラメータに CMS を行っているため周波数特性が正規化され、フィルタの影響を軽減することができたためである。しかし、伝送するために VQ を行った特徴ベクトル (with VQ) の WER は M/A フィルタにより増加した。これは、サーバ側において CMS を行っているのに関わらず、M/A フィルタにより周波数特性が変動し、VQ 歪みが増大したことが原因であると考えられる。この問題を解決するためには、クライアント側で VQ 歪みを減少させる手法が必要である。

4. 周波数特性正規化手法

3.5.1 において、DSR では入力デバイスの周波数特性の差異は VQ 歪みを増大させ、認識性能を劣化させる原因となることを述べた。このような周波数特性の差異を正規化する手法として CMS が提案されている。しかし、DSR フロントエンドの VQ 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。そのため、CMS を行う場合はサーバ側で行う必要がある。しかし、サーバ側で CMS を行っても、フロントエンド部で生じる VQ 歪みを減少させることはできない。そこで、周波数特性の差異が生じた場合でも、VQ 歪みを増大させない手法としてフロントエンドにおける VQ コードブック平均一致化手法と歪み誤差最小化手法の 2 つの周波数正規化手法を提案する。

4.1 VQ コードブック平均一致化手法 本節では、VQ 歪みを減少させる一手法として、VQ コードブック平均一致化手法を提案する。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平行移動する手法である。以下、本手法を BRM1 (Bias Removal Method 1) とする。

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

$$\mathbf{a}_{train} = \frac{\sum_{s=1}^S \sum_{n=1}^{N_s} \mathbf{x}_{sn}}{\sum_{s=1}^S N_s} \dots\dots\dots (3)$$

ここで、 $\mathbf{a}_{train}$ 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、 $\mathbf{x}_{sn}$ は発声 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

$$\mathbf{a}_{test} = \frac{\sum_{n=1}^N \mathbf{x}_n}{N} \dots\dots\dots (4)$$

ここで、 $\mathbf{a}_{test}$ は各認識発声の平均特徴パラメータを示し、 $\mathbf{x}_n$ 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

$$\tilde{\mathbf{x}}_n = \mathbf{x}_n - (\mathbf{a}_{test} - \mathbf{a}_{train}) \dots\dots\dots (5)$$

ここで、 $\tilde{\mathbf{x}}_n$ 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 歪みを減少することが可能である。

実際には、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、上記の (1) で 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できない。そのため、本論文では ETSI で定義されている VQ コードブックの平均を $\mathbf{a}_{train}$ として用いる。

図 5 に、図 4(b) 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適切に重なり、VQ 歪みを減少可能なことがわかる。

4.2 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 歪みを極小化する方法である。以下、本手法を BRM2 (Bias Removal Method 2) とする。

初期特徴パラメータ $\mathbf{x}_n$ ($n = 1, \dots, N$, 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、VQ 歪みを減少させる。

(1) 特徴パラメータ $\mathbf{x}_n^i$ と VQ コードブック間の歪み d_n を計算する。

$$d_n = |\mathbf{x}_n^i - q(\mathbf{x}_n^i)|^2 \dots\dots\dots (6)$$

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

(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。

$$D = \sum_n d_n \dots\dots\dots (7)$$

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

Fig. 5. Analysis example of BRM1 (1st and 2nd order MFCC).

$$= \sum_n |\mathbf{x}_n^i - q(\mathbf{x}_n^i)|^2 \dots\dots\dots (8)$$

歪み D が閾値以下ならば、 $\mathbf{x}_n^i$ 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 $\mathbf{h}$ を計算する。

$$\tilde{D} = \sum_n |(\mathbf{x}_n^i - \mathbf{h}) - q(\mathbf{x}_n^i)|^2 \dots\dots\dots (9)$$

$$\frac{\partial \tilde{D}}{\partial \mathbf{h}} = \frac{\partial (\sum_n |(\mathbf{x}_n^i - \mathbf{h}) - q(\mathbf{x}_n^i)|^2)}{\partial \mathbf{h}} = 0 \dots\dots (10)$$

$$\mathbf{h} = \frac{\sum_n (\mathbf{x}_n^i - q(\mathbf{x}_n^i))}{N} \dots\dots\dots (11)$$

(4) (3) で算出した $\mathbf{h}$ を特徴パラメータから減算する。

$$\mathbf{x}_n^{i+1} = \mathbf{x}_n^i - \mathbf{h} \dots\dots\dots (12)$$

$i = i + 1$ とし (1) に戻る。

図 4 (b) の特徴パラメータに本提案手法を適用した結果、VQ コードブック平均一致化手法と同様に VQ コードブックと特徴パラメータの分布が近似した。

5. 提案手法の評価実験

5.1 実験条件 提案手法の有効性を検討するため、JNAS を用いた音声認識実験を行った。音声認識実験は、3.1 と同様の実験条件で行った。

3.4 の実験において、VQ 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ることにより VQ し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ることがわかった。そこで、本実験では、VQ 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。提案手法は、VQ 歪み増大による認識性能劣化を抑制する手法であるため、認識には VQ を行った特徴パラメータを用いた。

周波数特性の差異の影響を検討するため、次のフィルタ

図 6 H/P フィルタの周波数特性

Fig. 6. Frequency response of H/P filter.

図 7 M/A フィルタの周波数特性

Fig. 7. Frequency response of M/A filter.

を用い、人工的に乗算性雑音を評価データに加えた。

- 高域通過フィルタ (H/P: High Pass filter)

$$s_{of}(n) = s_{in}(n) - 0.9 \times s_{in}(n-1) \dots\dots\dots (13)$$

- 移動平均フィルタ (M/A: Moving Average filter)

$$s_{of}(n) = 0.25 \times (s_{in}(n) + s_{in}(n+1) + s_{in}(n+2) + s_{in}(n+3)) \dots\dots\dots (14)$$

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。また、図 6 に H/P フィルタ、図 7 に M/A フィルタの周波数特性を示す。

5.2 実験結果 実験結果を表 5、表 6 に示す。表 5 は標準化周波数 16kHz、表 6 は標準化周波数 8kHz の結果である。表中の *no adapt. w/o VQ* は提案手法を適用しない VQ なしの特徴パラメータ、*no adapt. with VQ* は提案手法を適用せず VQ した特徴パラメータ、*BRM1* は提案手法 BRM1 を適用し VQ した特徴パラメータ、*BRM2* は提案手法 BRM2 を適用し VQ した特徴パラメータを認識した結果を示す。この結果より、提案手法である BRM1 は全ての条件下で適用なし (*no adapt. with VQ*) の WER を減少させていることがわかる。BRM2 においても、BRM1 と同様に、ほとんどの条件下において *no adapt. with VQ*

表 5 BRM の有効性 (16kHz) (WER (%))

Table 5. The result of the recognition with proposed methods in 16kHz sampling (WER in %).

adaptation method	filter		
	no filter	H/P	M/A
(1) <i>no adapt. w/o VQ</i>	9.6	9.6	9.3
(2) <i>no adapt. with VQ</i>	9.6	9.4	11.2
(3) <i>BRM1</i>	9.5	9.2	10.7
(4) <i>BRM2</i>	9.1	9.5	9.1

表 6 BRM の有効性 (8kHz) (WER (%))

Table 6. The result of recognition with proposed methods in 8kHz sampling (WER in %).

adaptation method	filter		
	no filter	H/P	M/A
(1) <i>no adapt. w/o VQ</i>	10.4	10.6	11.7
(2) <i>no adapt. with VQ</i>	10.8	10.6	14.1
(3) <i>BRM1</i>	10.2	10.5	11.7
(4) <i>BRM2</i>	10.6	10.5	11.6

の WER を減少させていることがわかる。特に、認識性能の劣化が大きい M/A フィルタに対し、提案手法は改善幅が大きい。実際、BRM1 では、M/A フィルタに対し、*no adapt. with VQ* と比較して標準化周波数 16kHz で 4.5%、8kHz 17.0%の誤り改善率を示した。BRM2 は BRM1 と同様に特に周波数特性の差異が大きい M/A フィルタでは、標準化周波数 16kHz で 18.9%、8kHz で 17.7%の誤り改善率を示した。これらの結果より、VQ 歪みを減少させる提案手法は入力系の周波数特性の差異により引き起こされる認識性能の劣化を抑制できることが確認された。テストセットに対する VQ 歪みを調べたところ、VQ 歪みが BRM1 では約 2.7%、BRM2 では約 4.5%低減していた。この歪みの減少が WER の低減につながったと推測できる。また、本提案手法は入力系の周波数特性の差異がない場合においても認識性能の劣化を引き起こすことはなかった。

6. まとめ

本論文では、DSR フロントエンドの日本語連続音声認識における問題点を明確にすると共に、その解決手法を提案した。

明確となった点として、従来、認識時と学習時の特徴パラメータは同一の条件により抽出した場合が最も高い認識性能が得られていたが、DSR の場合は圧縮を行った特徴パラメータを用いて学習した音響モデルの認識性能が低いことがわかった。これは、VQ 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を学習する際フロアリングを行わなくてはならないことが原因であった。そこで、学習には圧縮を行わない特徴パラメータを用いることにより、この問題を解消できることを示した。この際の認識精度は、VQ を行わない特徴パラメータで認識した場合とほぼ同等であり、音声波形を圧縮・復元した場合の認識精度よりも高いことを示した。また、伝送速度同一

の条件下でも、標準化周波数 16kHz での認識精度が 8kHz を上回った。これは、電話回線等による帯域の制限を受けないフロントエンド側で音響分析を行う場合には、分析帯域幅を広げることにより、認識精度の改善をはかることができることを意味する。

さらに、入力デバイスの周波数特性の差異による認識性能の劣化は従来の CMS では低減させることができないことを示した。これは、周波数特性の差異が VQ 歪みを増大させることが原因である。そこで、入力系の周波数特性の差異により引き起こされる VQ 歪みの増大を低減する手法として、VQ コードブック平均一致化手法、歪み誤差最小化手法を提案した。提案手法の有効性を検討する実験より、提案手法は入力系の周波数特性の差異による認識精度低下を抑制することに有効であることを確認した。

本論文で提案した手法は、1 発声終了後に適用されるため、音声入力から認識結果を出力するまでに処理遅れを伴うという問題が残されている。この問題を解決するため、現在、実時間処理が可能な量子化歪み減少手法の検討を進めている。また、ETSI より加法的雑音に頑健な分散音声認識フロントエンドが勧告されている⁽¹⁴⁾。今後はそのフロントエンドを用いた評価も行う予定である。

(平成 15 年 9 月 29 日受付, 平成 16 年 6 月 30 日再受付)

文 献

- (1) B. Lilly and K. Paliwal: "Effect of speech coders on speech recognition performance", *Proc. ICSLP*, pp.2344–2347 (1996)
- (2) 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)
- (3) ETSI ES 201 108 v1.1.2 distributed speech recognition; front-end feature extraction algorithm; compression algorithm (2000)
- (4) 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)
- (5) B. Noe, J. Siemel, D. Juvet, L. Mauuary L. Boves, J. Veth, and F. Wet: "Robust feature extraction for distributed speech recognition", *Proc. EuroSpeech*, pp.433–436 (2001)
- (6) C. Benitez, L. Burget, B. Chen, S. Dupont, H. Garudadi, H. Hermansky, P. Jain, S. Kajarekar, and S. Sivasdas: "Robust ASR front-end using spectral-based and discriminant features: experiments on the Aurora tasks", *Proc. EuroSpeech* (2001)
- (7) D. Macho and C. Nadeu: "Comparison of spectral derivative parameters for robust speech recognition", *Proc. EuroSpeech*, pp.205–208 (2001)
- (8) N. Parihar, J. PiconeParihar, and J. Picone: "Aurora working group: DSR front end LVCSR evaluation AU/384/02", *European Telecommunications Standards Institute* (2002)
- (9) K. Yamamoto, S. Nakamura, K. Takeda, S. Kuroiwa, N. Kitaoka, T. Yamada, M. Mizumachi, T. Nishiura, and M. Fujimoto: "AURORA-2J/AURORA-3J Corpus and Evaluation Baseline", *SLP-47*, No.19, pp.101–106 (2003) (in Japanese) 山本一公・中村 哲・武田一哉・黒岩眞吾・北岡教英・山田武志・水町光徳・西浦敬信・藤本雅清:「Aurora-2j/aurora-3j データベースとその評価ベースライン」, *SLP-47*, No.19, pp.101–106 (2003)
- (10) S. Kuroiwa, K. Takeda, M. Naito, N. Inoue, and S. Yamamoto: "Error analysis of field trial results of spoken dialogue system for telecommunications applications", *IE-*

ICE Trans. Inf. & Syst., Vol.E78-D, No.6, pp.636–641 (June 1996)

- (11) K. Ito, M. Yamamoto, and et al. Jnas: "Japanese speech corpus for large vocabulary continuous speech recognition", *J. Acoust. Soc. Jap.*, Vol.20, No.3, pp.199–206 (1995)
- (12) ITU-T recommendation G.723.1 dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s. (1996)
- (13) A. Lee, T. Kawahara, and K. ShikaNo: Julius — an open source real-time large vocabulary recognition engine. *Proc. EuroSpeech*, pp.1691–1694 (2001)
- (14) ETSI ES 202 050 v1.1.1 STQ; distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms (2002)

柘 植 覚 (非会員) 1996 年, 徳島大学工学部知能情報工学科卒業, 1998 年, 同大学院工学研究科博士前期課程知能情報工学専攻修了, 2001 年, 同大学院工学研究科博士後期課程システム工学専攻修了, 2000 年, 徳島大学工学部助手・博士(工学)・音声認識, 情報検索等の研究に従事。

黒 岩 眞 吾 (正員) 1986 年, 電通大・電気通信・通信卒。1988 年, 同大学院修士課程了。博士(工学)。同年国際電信電話(株)入社。1988 年～2001 年, 同社研究所において電話音声認識システムの研究・開発に従事。2001 年, 徳島大学工学部助教授。音声認識, 話者照合, 情報検索の研究に従事。電子情報通信学会平成 8 年度学術奨励賞, 日本音響学会第 3 回および第 5 回技術開発賞受賞。日本音響学会, 電子情報通信学会, 人工知能学会各会員。

獅々堀 正 幹 (非会員) 1991 年徳島大学工学部情報工学科卒業。1993 年同大学院博士前期課程修了。1995 年同大学院博士後期課程退学。同年同大学工学部知能情報工学科助手。1997 年同大学工学部知能情報工学科講師。2001 年同大学工学部知能情報工学科助教授。現在, 同大学工学部知能情報工学科助教授。博士(工学)。マルチメディア情報検索, 自然言語処理の研究に従事。著書「情報検索アルゴリズム」共立出版, 情報処理学会第 45 回全国大会奨励賞受賞。電子情報通信学会, 言語処理学会会員。

北 研 二 (非会員) 1981 年, 早稲田大学理工学部数学科卒業。1983 年, 沖電気工業(株)入社。1987 年, ATR 自動翻訳電話研究所出向。1992 年, 徳島大学工学部講師。1993 年, 同助教授, 2000 年, 同教授, 2002 年, 同大学高度情報化基盤センター教授。工学博士。自然言語処理, 情報検索等の研究に従事。1994 年日本音響学会技術開発賞受賞。著書「確率的言語モデル」(東京大学出版会)「情報検索アルゴリズム」(共著, 共立出版)など。