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

2004－ＳＬＰ－51 （２）
2004／5／27

ETSI標準フロントエンドを用いた雑音下音声認識の検討

福士なな子　加藤 正治　小坂 哲夫　好田 正紀

山形大学工学部
〒992-8510 米沢市城南 4-3-16
Tel: 0238-26-3365 FAX: 0238-26-3365

あらまし 本稿では、分散音声認識における問題点である背景雑音に対し、ETSIで標準化されているフロントエンド(Advanced DSR フロントエンド : WI008) をベースとして、日本語連続音声の認識精度の向上を目指す。WI008の開発に用いられたタスクは欧米語連続数字音声であり、これが日本語連続音声にどの程度有効か不明である。そこで音響モデル学習用の雑音重畳データから新たに日本語用コードブックを作成し、WI008の中に組み込み量子化する手法をとり、かつ、フロントエンド内のBlindEqualizationを外すことにより、日本語音声認識での精度向上を試みた。さらに、特徴ベクトルに分散正規化を用いる検討を行い有効性を示した。この結果を踏まえ、雑音下連続数字認識タスクであるAURORA2の日本語版であるAURORA-2Jで同様の評価を行った。雑音除去機能の含まれていないフロントエンドを用いて評価しているベースラインと比較して、全体で55.75 %の改善率を得ることが出来た。

キーワード 分散音声認識, 雑音重畳音声, ETSI標準フロントエンド, マルチコンディション学習, ベクトル量子化

Noisy speech recognition using ETSI standard DSR front-end

N. FUKUSHI, M. KATOH, T. KOSAKA, and M. KOHDA

Faculty of Engineering, Yamagata University
4-3-16 Jonan Yonezawa-shi, 992-8510, Japan
Tel: 0238-26-3365 FAX: 0238-26-3365

Abstract This paper aims at improvement in the recognition accuracy of a Japanese continuous speech against the problem of background noise in distributed speech recognition by using the Advanced DSR Front-End (WI008) standardized in ETSI. It is not clear whether the codebook in WI008 is effective for Japanese continuous speech, because it was developed for recognizing connected digit in European language. In this work, new codebook for Japanese speech was developed by using noisy speech database which was also used for acoustic modeling. The codebook was embedded in the WI008 front-end without BlindEqulization module to improve the performance of Japanese speech recognition. In addition, variance normalization method was tested and it showed effective result. These proposed methods were also tested in AURORA-2J task. It was the Japanese version of AURORA2 where connected digit recognition was performed under noisy conditions. Compared with the baseline front-end in which a noise removal module was not embedded, relative improvement of 55.75 % over the baseline was obtained.

Key words distributed speech recognition, noisy speech, ETSI standard DSR front-end, multicondition training, vector quantization

1. はじめに

近年、モバイルサービスの加入者が急激に増加し、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) が提案されてい

る[1]。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。携帯端末での音声認識では、通常屋外で使用することが多いため背景雑音が問題となる。雑音が混じると認識率が著しく低下することが分かっている。このため雑音除去する手法を取り入れた認識システムが必要である。

欧州電気通信標準化機構 (ETSI) では、2000 年 4 月に分散音声認識用のフロントエンド (WI007) を提案している。但しこのフロントエンドには雑音対策手法は組み込まれていない。本論文では、2002 年 10 月に提案された Advanced DSR フロントエンド (WI008) をベースラインとして用いる[4]。WI008 には数々の雑音除去手法が組み込まれている。日本語音声コーパスにおいては、WI008 を用いた乗算性雑音に対する検討が行われている[3]。本研究では加算性雑音が重畳した場合について、WI008 をベースとしてさらなる改善を目指し、さらに AURORA-2J [6] での評価を行う。

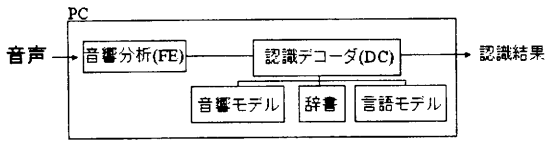

図 1　通常の音声認識システム

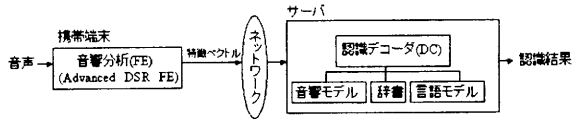

図 2　分散音声認識システム

2. 分散音声認識 (DSR:Distributed Speech Recognition)

DSR システムは図 1 の通常の音声認識システムとは異なり、音響分析を行うフロントエンド部と音声認識のデコードを行うバックエンド部から構成される (図 2 参照)。フロントエンド部となるクライアント側では、音声認識に必要な特徴パラメータを分析後、ベクトル量子化して伝送路に送る。デコードを行うサーバ側では、伝送された特徴パラメータを復元し音声認識を行う。分散型方式をとることにより、音声を圧縮して伝送する携帯電話のコーデックによる歪みを避けることが可能となり認識性能の改善が期待できる。また、音声認識に有効な情報のみを伝送することで、伝送速度を低く抑えることができる。さらに、クライアント側の変更なしに、新しいサービスやアプリケーションに対応でき、多言語対応が可能であるということや、サーバ処理量の軽減になるというメリットもある。ここで、本研究で用いるフロントエンド (Advanced DSR フロントエンド:WI008) の概要について次節以降に示す。

3. Advanced DSR フロントエンド (WI008)

自動音声認識 (ASR) 技術の実用を展開するにあたって、頑健さは不可欠な問題点である。携帯電話のような携帯端末で、異なる音響環境、あるいはチャネルが音声に歪みを与え、認識システムに悪影響を与える。そこで、分散音声認識の標準化においては、実環境利用において十分な性能を達成できる音声認識前処理が必要となる。本研究で用いるフロントエンド (WI008) は ETSI の DSR 標準化活動で開発された雑音に頑健なフロントエンドで、2002 年 10 月に提案された。WI008 には、雑音除去・波形処理・BlindEqualization などの雑音除去手法が組み込まれている (図 3 参照)。このうち、本研究と関連する BlindEqualization について以下の節で述べる。

3.1 BlindEqualization

BlindEqualization とは下記の式 (1)〜(4) を用い、順次参照ケプストラム ($RefCep(i), 1 \leq i \leq 12$) に入力ケプストラム $c(i)$ を近似することにより、乗算性雑音に対し頑健なケプストラム $c_{eq}(i)$ を抽出するブロックである。

$$weightingPar = Min(1, Max(0, \ln E - 211/64)) \quad (1)$$

$$stepSize = 0.0087890625 \times weightingPar \quad (2)$$

$$c_{eq}(i) = c(i) - bias(i), 1 \leq i \leq 12 \quad (3)$$

$$bias(i) += stepsize \times (c_{eq}(i) - RefCep(i)), 1 \leq i \leq 12 \quad (4)$$

$$bias(i) = 0.0, 1 \leq i \leq 12$$

本研究では、日本語から作成されたコードブック使用時に、この BlindEqualization ブロックを外す検討を行う。

4. 耐雑音に関する各種検討

4.1 コードブック

Aurora プロジェクトのタスクは TI-DIGITS の数字認識を対象としたもので、雑音処理のみに焦点を当てるため比較的小さなタスクである。しかし WI008 のコードブックが日本語大語彙連続音声にどの程度有効かは不明である。そこで、日本語音声コーパスからコードブックを作成し、WI008 のプログラムに組み込むことにより、タスクが欧米語の WI008 のコードブックとの比較を行う。また、SNR 別にコードブックを作成し、それぞれの SNR 毎に量子化を行う事によってコードブックの SNR 依存性を検討する。

今回の実験では以下の 3 種類のコードブックを使用した。

- WI008 コードブック

標準化フロントエンドに添付されたもの (タスクは欧米語で連続数字音声)

- 日本語コードブック

音響モデル学習用雑音重畳データ 15,732 文 (5.1 節参照) を WI008 で分析し、LBG アルゴリズムを用い、ETSI の量子化サイズに合わせてコードブックを作成 (タスクは日本語大語彙連続音声)

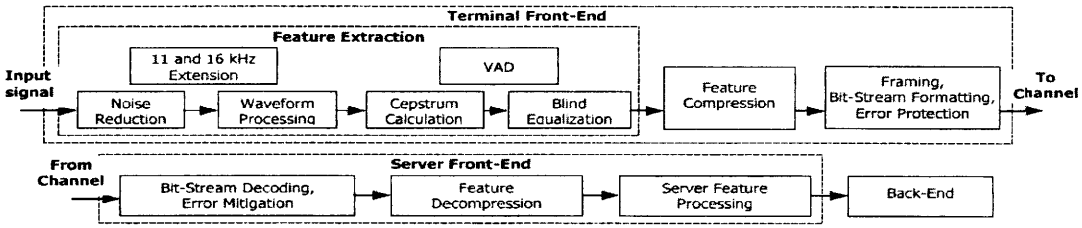

図 3　Advanced DSR フロントエンド (WI008):上部はターミナル側での構成・下部はサーバ側での構成

- 日本語 SNR 別コードブック

音響モデル学習用雑音重畳データ 15,732 文を、SNR 毎に分け (5dB,10dB,15dB,20dB, ∞ dB) × 4 種類の雑音 (走行自動車内・展示会場 (通路)・人込み・列車 (在来線))(表 1 参照)、WI008 で分析し、それぞれの SNR に対応したコードブックを作成する。テストセットの量子化には、それぞれの SNR に対応したコードブックを用いる。(タスクは日本語大語彙連続音声)

コードブックサイズは全て ETSI 標準規格 (表 4) に合わせた。

4.2 BlindEqualization

フロントエンドで BlindEqualization を作成すると入力ベクトルは参照ベクトルに近似される。そのため、コードブックの言語依存性が少なくなり、日本語コードブックの効果が低下する可能性がある。よって WI008 の BlindEqualization 部分を外した場合についても、日本語コードブックの有効性を検討する。

4.3 分散正規化 (Variance Normalization)

雑音重畳音声では雑音によりケプストラムの平均値・分散が変動し悪影響を与える恐れがある。そこでケプストラムの分散正規化 (VN:Variance Normalization) [5] について検討する。以下に分散正規化のアルゴリズムについて示す。

特徴ベクトルの時間 t の d 番目の要素を C_{td} とする。

$(t = 1 \cdots T, d = 1 \cdots D)$

まず初めに MS(平均正規化) を用いる

$$C'_{td} = C_{td} - \mu_d \tag{5}$$

ここで

$$\mu_d = \frac{1}{T}\sum_{t=1}^{T} C_{td} \tag{6}$$

分散正規化では

$$\bar{C}_{td} = \frac{C'_{td}}{\sigma_d} = \frac{C_{td} - \mu_d}{\sigma_d} \tag{7}$$

ここで

$$\sigma_d = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(C_{td} - \mu_d)^2} \tag{8}$$

分散正規化したケプストラムを音響モデル学習用・及び評価用に用いる。なお、分散正規化はケプストラムを抽出した後、⊿や⊿⊿を計算する前に行うこととする。

4.4 マルチコンディション学習

音声認識を行う場合、一般的にその音響モデルには、雑音の入っていない音声データを音響モデルの学習時に用いる、クリーン学習モデルを使用する。しかし、実環境で音声認識を行う場合、当然音声入力段階における雑音の存在を考慮しなければならない。入力されるデータは、単に認識したい音声だけでなく、音声の背景で流れている BGM や、通りすがりの人々の話し声、自動車の騒音など様々の雑音が含まれている。クリーン学習では、このような雑音の混じった音響信号のモデル化ができない。よって、学習時に様々な雑音を音声に重畳したデータで学習することにより、雑音下音声のモデル化ができると考えられる。これをマルチコンディション学習 [2] という。本研究ではマルチコンディション学習の効果の検討も行う。

5. 実験条件

5.1 音声データ・雑音データ

表 1　音響モデル学習用雑音重畳データ

SNR \ 雑音の種類	列車	人ごみ	走行自動車内	展示会場
∞	786 文 ×4			
20	786 文	786 文	786 文	786 文
15	787 文	787 文	787 文	787 文
10	787 文	787 文	787 文	787 文
5	787 文	787 文	787 文	787 文

表 2　評価用雑音重畳データ A(学習用にも用いる雑音)

SNR \ 雑音の種類	列車	人ごみ	走行自動車内	展示会場
20	100 文	100 文	100 文	100 文
15	100 文	100 文	100 文	100 文
10	100 文	100 文	100 文	100 文
5	100 文	100 文	100 文	100 文

表 3　評価用雑音重畳データ B(学習用に用いない雑音)

SNR \ 雑音の種類	駅	工場	幹線道路	エレベータホール
20	100 文	100 文	100 文	100 文
15	100 文	100 文	100 文	100 文
10	100 文	100 文	100 文	100 文
5	100 文	100 文	100 文	100 文

本研究では音声コーパスに日本音響学会新聞記事読み上げコーパス (ASJ＿JNAS) を用いた。学習用データには男性 102 名による計 15,732 文、評価用データには男性 10 名による計 100 文を用いた。また雑音データは電子協騒音データベースより 8 種類を使用した。このうち、「列車 (在来線)」、「人込み」、「1500cc クラス自動車内」、「展示会場 (通路)」の 4 種類は学習用データの作成に用いた。上記 15732 文を 20 分割し、4 種類の雑音 × 5 種類の SNR(5,10,15,20, ∞ dB) の計 20 種類の組み合

わせで雑音を人工的に重畳しマルチコンディション学習用データとした (表 1)。どのサブセットにも 102 名全員のデータを含むように設計してある。 テストセットとしては 2 種類用意した。学習時に用いた 4 種類の雑音を SNR 及び雑音ごとに上記の 100 文に重畳したテストセット A(表 2) と、学習時には用いなかった 4 種類の雑音 (「駅」、「工場」、「幹線道路・交差点」、「エレベータホール」) を同様に重畳したテストセット B(表 3) である。

5.2 音声分析

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) と対数パワーが用いられる。分析条件は標本化周波数 16kHz、フレーム長 25ms、分析周期 10ms のハミング窓、プリエンファシス係数 0.9 である [1]。

なお、認識時に用いる特徴パラメータは MFCC1 次〜12 次と対数パワー、それらのΔと$\Delta\Delta$の計 39 次元とする。MFCC0 次は量子化時に用いるが、認識時には用いない。6.3 節以外では発話毎のケプストラム平均正規化を行う。

5.3 ベクトル量子化

前節で得られた短時間分析を行った特徴パラメータに対し、ベクトル量子化を行い特徴パラメータを圧縮する。ベクトル量子化は隣接する 2 次元を組にして、ETSI で決められた量子化サイズで行う。量子化サイズは表 4 のとおりである。距離尺度としては重みづけユークリッド距離が用いられる。

表 4 量子化サイズ

パラメータ	$\log P_0$	c_1, c_2	c_3, c_4	c_5, c_6	c_7, c_8	c_9, c_{10}	c_{11}, c_{12}
サイズ	256	64	64	64	64	64	32

5.4 認識システム

第 1 パスで triphoneHMnet 及び単語 bigram を用いて単語グラフを生成し、第 2 パスで単語 trigram を用いて単語グラフをリスコアする、2-Pass デコーダを使用する。第 1 パスでの言語重みを 22、挿入ペナルティを-5 とした。音響モデルは、音素カテゴリを 34 音素+無音の計 35 音素とし、各 triphone を 3〜6 状態 HMM でモデル化、状態対応確率に基づく状態クラスタリングを行い、状態数 2000、混合数 16 の HMnet をクリーンデータ、及び、雑音重畳音声データから学習した。学習時に量子化は行わない。言語モデルは N-gram で語彙は 5k、毎日新聞の 45ヶ月分を用いて作成した。

6. 実験結果・考察

6.1 コードブックの検討

まず、BlindEqualization を外さない場合のコードブックの効果について検討した。

テストセット A(学習用データ・評価用データにおいて同じ雑音を用いた場合) の結果を表 5、テストセット B(学習用データ・評価用データにおいて異なる雑音を用いた場合) の結果を表 6 に示す。コードブックの種類に関しては 4.1 節を参照のこと。

表 5 コードブックの検討 : テストセット A(WER[%])

HMnet	clean			multicondition		
CodeBook	WI008	日本語	日本語 SNR 別	WI008	日本語	日本語 SNR 別
クリーン	5.38	5.18	5.18	6.11	6.11	6.52
走行自動車内	6.01	5.98	6.07	6.54	6.38	6.44
人込み	10.12	9.83	9.77	9.15	9.23	8.84
展示会場 (通路)	15.34	15.86	15.64	12.41	13.04	12.44
列車 (在来線)	19.44	19.09	19.01	17.62	17.56	17.39
Average	11.26	11.19	11.13	10.37	10.46	10.33

表 6 コードブックの検討 : テストセット B(WER[%])

HMnet	clean			multicondition		
CodeBook	WI008	日本語	日本語 SNR 別	WI008	日本語	日本語 SNR 別
クリーン	5.38	5.38	5.18	6.11	6.11	6.52
幹線道路・交差点	14.93	18.30	18.01	14.54	15.22	15.94
エレベータホール	23.81	23.85	23.42	18.90	19.77	19.34
駅	23.85	23.54	23.38	19.23	20.42	19.52
工場	23.96	24.00	23.44	18.63	20.97	20.81
Average	18.39	19.01	18.69	15.48	16.50	16.43

テストセット A では 3 種類のコードブックの平均の性能は、クリーン学習の音響モデルを用いた場合も、マルチコンディション学習の音響モデルを用いた場合でも双方に差がほとんど見られない。テストセット B では、音響モデル学習用雑音重畳データから作成したコードブックの方が WI008 のコードブックよりも、性能が悪かった。以上の結果は、BlindEqualization の効果により、日本語コードブックの特徴も、WI008 のコードブックの特徴も、同じ参照ベクトルに近似されてしまい、コードブックを変えた場合の日本語特有の効果が薄れたためと考えられる。

6.2 BlindEqualization の検討

表 7 BlindEqualization の検討:テストセット A(WER[%])

HMnet	clean				multicondition			
CodeBook	日本語		日本語 SNR 別		日本語		日本語 SNR 別	
BE	○	×	○	×	○	×	○	×
クリーン	5.18	5.07	5.18	4.97	6.11	5.80	6.52	5.49
走行自動車内	5.98	5.42	6.07	5.34	6.38	5.86	6.44	5.88
人込み	9.83	9.38	9.77	9.40	9.23	8.12	8.84	8.16
展示会場 (通路)	15.86	14.24	15.64	14.74	13.04	12.38	12.44	12.26
列車 (在来線)	19.09	17.61	19.01	19.03	17.56	16.85	17.39	16.94
Average	11.19	10.34	11.13	10.70	10.46	9.80	10.33	9.75

表 8 BlindEqualization の検討:テストセット B(WER[%])

HMnet	clean				multicondition			
CodeBook	日本語		日本語 SNR 別		日本語		日本語 SNR 別	
BE	○	×	○	×	○	×	○	×
クリーン	5.38	5.07	5.18	4.97	6.11	5.80	6.52	5.49
幹線道路・交差点	18.30	18.14	18.01	18.30	15.22	15.16	15.94	15.51
エレベータホール	23.85	23.50	23.42	23.52	19.77	18.26	19.34	17.45
駅	23.54	23.19	23.38	23.29	20.42	19.24	19.52	18.76
工場	24.00	23.66	23.44	23.94	20.97	20.39	20.81	20.43
Average	19.01	18.71	18.69	18.80	16.50	15.77	16.43	15.53

BE : BlindEqualization

ここでは、BlindEqualization ブロックを WI008 から外すことで、日本語コードブックの効果に変化がないか検討した。上記のコードブックの検討では、日本語音声コーパスからコードブックを作成しても、タスクが欧米語単語数字音声である WI008 のコードブックと差がないか、逆に悪化している。これは、コードブックを日本語音声コーパスから作成しても、BlindEqualization で参照ケプストラムに近似するため、効果が薄れてしまったのではないかと考えられる。

なお、WI008 のコードブック用の音声データは非公開であるため、量子化の際は BlindEqualization を外すことが出来ない。よって WI008 のコードブックを用いる場合は BlindEqualization は外していない。テストセット A の結果を表 7、テストセット B の結果を表 8 に示す。

結果より日本語コードブックには BlindEqualization を外した方が有効であることが分かる。但し、テストセット A では日本語コードブックが WI008 のコードブックより精度が良かったが、テストセット B では WI008 のコードブックには及ばなかった。

また、6.1 および 6.2 節の両方の実験を通して、テストセット A では日本語 SNR 別コードブックが WI008 のコードブックよりも効果があったが、テストセット B では WI008 のコードブックよりも若干悪化したことから、コードブックは SNR よりも雑音の種類に依存するためと考えられる。この結果を踏まえ、次節で分散正規化について検討する。

6.3 分散正規化の検討

表 9 分散正規化の検討:テストセット B(WER[%])

HMnet	multicondition					
CodeBook	WI008		日本語		日本語 SNR 別	
BE	○	○	×	×	×	×
	MS	VN	MS	VN	MS	VN
駅	19.23	19.01	19.24	19.24	18.76	19.17
工場	18.63	16.42	20.39	18.10	20.43	18.28
幹線道路・交差点	14.54	14.33	15.16	13.27	15.51	14.31
エレベータホール	18.90	19.07	18.26	17.72	17.45	18.62
クリーン	6.11	6.11	5.80	6.11	5.49	5.59
Average	15.48	15.00	15.77	**14.89**	15.53	15.19

MS:平均正規化 · VN : 分散正規化

ケプストラムの平均正規化と分散正規化の比較を行う。日本語音声コーパスから作成したコードブックでは BlindEqualization を外した方が効果があることが上記でわかっているため BlindEqualization なしで評価を行う。

結果を表 9 に示す。なお評価は、テストセット B のみで行った。どのコードブックを用いた場合でも平均正規化より、分散正規化を用いた方がよい認識結果を得ることが出来た。また SNR 別で結果を見ると、SNR が低い部分ほど効果がよく見られ、6〜7 % WER を改善できた部分もあった。よって、分散正規化は、雑音下音声に有効であり、かつ、SNR が低い部分ほど効果的であるといえる。

以上の検討より、日本語コードブック、BlindEqualization なし、分散正規化の組み合わせで最良の結果を得ることができた。

7. AURORA-2J を用いた評価

AURORA-2J は、雑音環境下連続英語数字音声認識タスクの共通評価フレームワークである AURORA-2 の日本語版である。JNAS での結果を踏まえて同様の検討を AURORA-2J で行う。ここでは 6 章で得られた最良の組み合わせである、日本語コードブック、BlindEqualization なし、分散正規化を WI008 に適用して実験を行った。AURORA-2J の学習セットとテストセットを表 10 に示す。

学習データには、雑音のない 8,440 発話の Clean データと 4 つの複数雑音環境データ (Subway, Bannle, Car, Exhibition) を 5,10,15,20dB の SNR で重畳された同数の雑音混入データが用意されている。評価データは、3 種類で、①テストセット A : 4 つの既知雑音を含む音声、②テストセット B:4 つの未知雑音 (Restaurant,Street,Airport,Station) を含む音声、③テストセット C : 電話回線特性を付与し、未知雑音を含む音声 (Subway,Street を加算後 MIRS 特性付与) を SNR=-5,0,5,10,15,20dB,Clean で評価する。

量子化の際に用いる日本語コードブックは、学習セットの 8,440 発話を用い、表 4 の量子化サイズに基づいて作成した。認識に用いるデコーダは提供されたものではなく、本研究室のオリジナルのものを用いた。分析条件は標本化周波数が 8kHz 以外は 5.2 節と同様である。HMM は 16 状態 20 混合の数字モデルと、3 状態 36 混合の無音モデルである。

HMM の構造はベースラインと同一で、コードブックの変更や分散正規化を行っているため、大部分の条件は評価カテゴリー 2 (ベースラインスクリプトと同じトポロジーの HMM で、認識時の適応技術を導入している場合) に相当するが、ここでは研究室オリジナルのデコーダを用いたため、評価カテゴリーは 5 (規定なし。提供されるデータベース内であれば、どんな処理でも許される) となる。

7.1 ベースラインとの比較評価

ベースラインの評価に用いられているフロントエンドは雑音除去機能が含まれていない HTK の HCopy [8] である。表 11 に各テストセットで得られた単語正解精度 (%)、表 12 にベースラインに対する誤り改善率を示す。

表 11 より、Clean training の場合、SNR10dB までは 90 %以上の精度、平均しても 80 %以上の精度が得られた。また、Multicondition training の場合では平均して 90 %以上の精度が得られた。また表 12 より、ベースラインと比較した場合、55.75 %の改善がみられた。フロントエンドに WI008 を用いた場合のみの改善率は文献 [7] によると全体で 47.59 %である。よって本実験の検討により、WI008 の有効性と共に提案手法の有効性がわかった。さらに同文献と比較すると、A,B,C の各セット共改善率を上回っている。BlindEqualization を外したため、テストセット C での悪化も予想されたが、逆に改善率は向上した。これは、分散正規化がチャネル歪みに対しても有効であったためと考えられる。

表 12 ベースラインに対する誤り改善率

Relative performance				
	A	B	C	Overall
Clean Training	70.06%	70.93%	70.12%	70.43%
Multicondition training	15.26%	55.70%	31.41%	41.07%
Average	42.66%	63.32%	50.77%	55.75%

8. ま と め

ETSI 標準分散音声認識フロントエンド (Advanced DSR フロントエンド : WI008) を用いて、日本語大語彙連続音声を対象とした雑音重畳音声認識の検討を、コードブック・BlindE-

表 10　AURORA-2J の学習セットとテストセット

学習・テストセット	音声	雑音	チャネル	SNR
Clean training	110 名,8440 発話	—	G.712	clean
Multicondition training	110 名,8440 発話	Subway,Babble,Car,Exhibition	G.712	Clean,20,15,10,5
テストセット A	104 名,4004 発話	Subway,Babble,Car,Exhibition	G.712	Clean,20,15,10,5,0,-5
テストセット B	104 名,4004 発話	Restaurant,Street,Airport,Station	G.712	Clean,20,15,10,5,0,-5
テストセット C	104 名,2002 発話	Subway,Street	MIRS	Clean,20,15,10,5,0,-5

表 11　単語正解精度 (%)

Clean Training (%Acc)

	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.85	99.88	100.00	99.75	99.87	99.85	99.79	99.82	99.69	99.79	99.85	99.78	99.81	99.83
20 dB	98.86	99.49	99.61	99.01	99.24	98.77	98.70	99.49	99.17	99.03	98.74	98.94	98.84	99.08
15 dB	97.21	97.61	97.78	96.98	97.40	96.25	98.25	98.36	97.50	97.59	97.39	97.85	97.62	97.52
10 dB	92.63	93.20	96.15	92.10	93.52	89.16	93.32	93.38	92.35	92.05	94.11	93.71	93.91	93.01
5 dB	78.42	77.81	85.60	77.29	79.78	70.10	79.84	81.80	82.32	78.52	80.32	82.47	81.40	79.50
0 dB	49.62	43.08	58.04	49.15	49.97	37.18	56.29	54.40	57.67	51.39	50.41	56.38	53.40	51.22
-5 dB	21.19	8.59	24.07	20.61	18.62	6.94	24.43	22.19	24.13	19.42	22.49	25.33	23.91	20.00
Average	83.35	82.24	87.44	82.91	83.99	78.29	85.28	85.49	85.80	83.72	84.19	85.87	85.03	84.09

Multicondition Training (%Acc)

	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.85	99.79	99.82	99.69	99.79	99.85	99.79	99.82	99.69	99.79	99.85	99.79	99.82	99.79
20 dB	99.60	99.49	99.79	99.51	99.60	99.51	99.67	99.67	99.63	99.62	99.51	99.67	99.59	99.61
15 dB	99.26	99.37	99.46	98.92	99.25	98.99	99.09	99.34	99.29	99.18	98.99	99.09	99.04	99.18
10 dB	98.59	98.10	98.03	97.69	98.10	96.68	97.19	97.58	97.62	97.27	96.68	97.19	96.94	97.54
5 dB	94.41	80.62	94.84	91.82	90.42	87.90	90.41	91.05	90.56	89.98	87.90	90.41	89.16	89.99
0 dB	78.69	72.40	79.57	76.25	76.73	59.93	73.43	73.30	75.38	70.51	59.93	73.43	66.68	72.23
-5 dB	43.48	31.74	41.40	45.23	40.46	22.54	39.36	38.41	42.49	35.70	22.54	39.36	30.95	36.66
Average	94.11	90.00	94.34	92.84	92.82	88.60	91.96	92.19	92.50	91.31	88.60	91.96	90.28	91.71

qualization・分散正規化の 3 つの観点から行った。

コードブックは ETSI 標準化ワーキンググループ (AURORA) で用いているタスクは欧米語であるため、日本語音声コーパスから作成し、WI008 に組み込んで評価を行った。この場合、BlindEqualization ありとなしで比較したところ、なしの場合が有効であることがわかった。分散正規化については、SNR が低いところほどよく効果を示し、全体でも平均正規化と比べ、若干の改善が見られた。今回の検討では、日本語大語彙連続音声に対しては、コードブックを日本語音声コーパス (学習用雑音重畳データ全体) から作成し、WI008 の BlindEqualization を外して、分散正規化を行った手法で最良の結果が得られた。さらにこの結果を踏まえ、雑音下連続数字認識タスクである AURORA-2 の日本語版である、AURORA-2J でも同様の評価を行った。雑音除去機能がないフロントエンドを用いたベースラインの結果と比較して、全体で 55.75 %の改善が見られた。また、WI008 のみを用いた場合の結果 [7] と比較しても改善率が向上し本手法の有効性が確認できた。

文　献

[1] ETSI standard document :Speech processing,Transmission and Quality aspects(STQ);Distributed speech recognition; Advanced front-end feature extraction algorithm;Compression algorithms,ETSI ES 202 050 v1.1.1 (2002-10).

[2] Davit Pearce and Hans-Gunter Hirsch, "The AURORA Experimental Framework for the Performance Evaluation of Speech Recignition System Under Noisy Conditions", Proc.of ICSLP2000,vol.4,pp.29-32(2000)

[3] 拓殖覚, 原一眞, 黒岩眞吾, 北研二 : "日本語音声コーパスを用いた ETSI STQ DSR Advanced Front-End の評価", 日本音響学会講演論文集 ,pp.57-58 (2003-3)

[4] Dusan Macho,Laurent Mauuary and Bernhard Noe, "Evaluation of a Noise-Robust DSR Front-End on AURORA Databases" ,Proc.of ICSLP2002,pp.17-20(2002)

[5] Chia-Ping Chen , Karim Filali and Jeff A. Bilmes, "Frontend Post-Processing and Backend Model Enhancement on the Aurora 2.0/3.0 Databases", Proc.of ICSLP2002,pp.59-62(2002)

[6] 山本一公他, "AURORA-2J/AURORA-3J データベースとその評価ベースライン", 情報処理学会研究報告,SLP-47-19(2003)

[7] 山田武志他, "AURORA-2J を用いた ETSI STQ Aurora WI008 Advanced DSR Frontend の評価", 信学技報 SP2003-130 ,pp.103-108(2003-12)

[8] http://htk.eng.cam.ac.uk/