TR－SLT－0080

パラレルタイプの日本語音声認識システムにおける雑音除去手法の性能評価

Evaluation of Noise Reduction Methods in Parallel Decoding Japanese Speech Recognition System

山本 遼　　松田 繁樹
Ryo Yamamoto　　Shigeki Matsuda
藤本 雅清　　遠藤 俊樹　　中村 哲
Masakiyo Fujimoto　　Toshiki Endo　　Satoshi Nakamura

2005年1月23日

概要

音声認識システムを実環境で用いるためには，雑音に対して十分に頑健でなければならない．高い頑健性を達成するため，異なる雑音環境に依存した複数の音響モデルをパラレルにデコーディングし，それらの複数の仮説を選択統合するパラレルタイプのシステムが提案され，その有効性が確認されている．パラレルタイプのシステムでは従来の耐雑音手法を併用することができるが，それらが有効であるかは明らかではない．本報告ではパラレルタイプの日本語音声認識システムにおける雑音除去手法の有効性を検討する目的で，Matched 音響モデルを用いた実験を行った．その結果，雑音除去手法は有効とは言えず，パラレルタイプ日本語音声認識システムは雑音除去を必要としないという結論を得た．

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2 TEL：0774－95－1301

Advanced Telecommunications Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai “Keihanna Science City” 619-0288,Japan
Telephone:+81-774-95-1301
Fax　　:+81-774-95-1308

目 次

1 はじめに

1.1 音声認識技術の現在

確率モデルの導入と大規模コーパスの整備が進み，音声認識技術は近年大きく進歩した．しかしながら実環境下でのシステムの利用に際しては，入力音声が雑音により歪み認識性能が著しく低下する問題が解決されていない．雑音環境に対する頑健性は実用化に向けて残された大きな課題の1つとなっており，研究が盛んに行われている．雑音に頑健な特徴量の追求，音響分析の前処理として雑音除去を行う方法，マイクロホンアレーによる指向性の制御，音響モデルの適応など多数の手法が提案されている [13].

1.2 パラレルデコーディングタイプの音声認識システム

また近年では計算機性能の急速な向上により，計算リソースの限定された実システムにおいてもリアルタイムに音声認識処理が行えるようになった．さらに大規模な計算リソースが利用可能なシステムでは，従来単独で用いられてきたデコーダを複数並列に用いてデコーディングを行い，それらの仮説を選択・統合するパラレル出コーディングタイプの音声認識システム (以下パラレルシステム) が利用可能となった．

単独のデコーダを用いるシングルシステムでは，さまざまな雑音環境に対して頑健な音響モデルや言語モデルが必要となる．しかし幅広い雑音環境下の学習データから音響モデルパラメータを学習すると個々の音素モデルの特徴量ベクトル分布の分散が大きくなってしまい，音素分類性能の低下につながる恐れがある．

一方パラレルシステムでは，特定の雑音環境に特化したデコーダを複数並列に用いることが可能であるため，音素分類性能を低下させることなく広い範囲の雑音環境で頑健に音声認識が行える可能性がある．松田ら [14] は，さまざまな雑音環境と2種類の発話スタイル下で学習した音響モデルを持つデコーダを用いたパラレルシステム (図1) により，SN 比が 10dB 以上の通常発声音声に対して 90%以上の，音節強調発声音声に対して約 45%の単語正解精度を得ることに成功した．

1.3 パラレルシステムにおける雑音除去手法

パラレルシステムは従来のデコーダを複数用いる手法であるので，従来のマイクロホンアレー，雑音除去手法，モデル適応などの雑音対策手法が同時に利用可能である．しかしこれらの手法は雑音環境の変動の影響を小さくすることを目的とした，シングルシステムを前提とした手法である．個々のデコーダが特定の雑音環境のみに特化することのできるパラレルシステムにおいては，これらの有効性，必要性は必ずしも自明ではない．

本研究はこのうち雑音除去手法について，パラレルな日本語大語彙連続音声認識システムにおける有効性を検証することを目的とする．

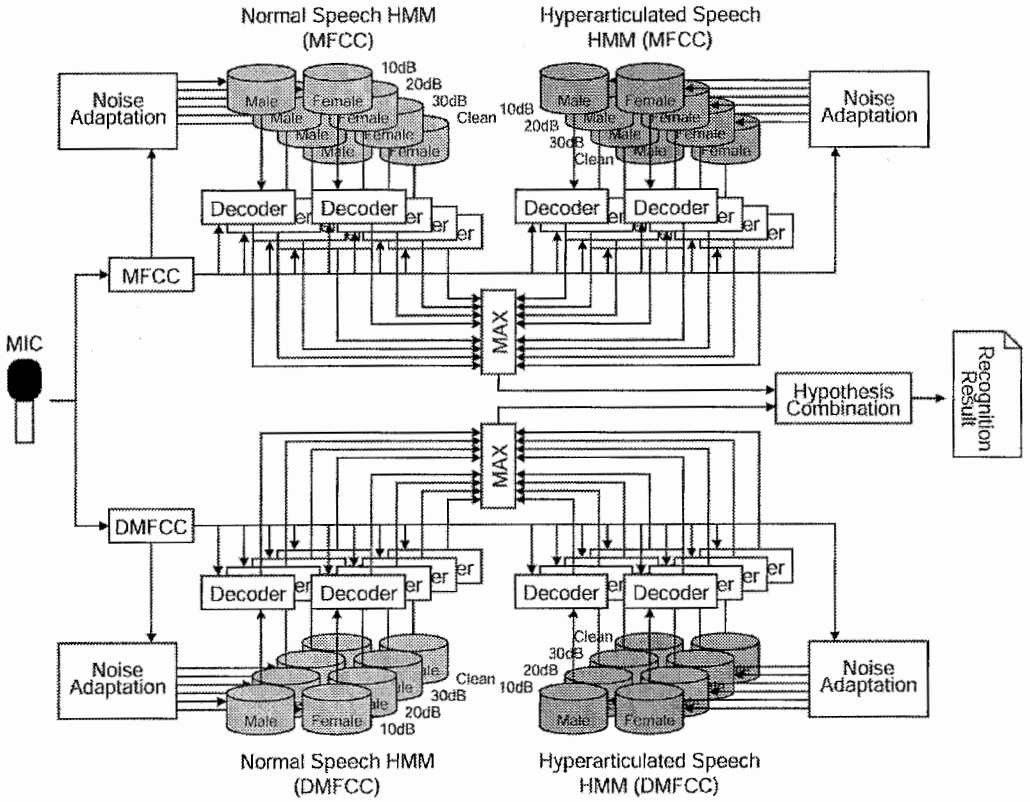

図 1: 松田らの用いたパラレルシステム ([14] より引用). 各雑音環境と発話スタイルごとに音響モデルを作成し，それらを用いたデコーダによる仮説を最大尤度基準で選択する．これを MFCC, DMFCC の各特徴量ごとに行い，それぞれの特徴量において選択された 2 つの仮説を単語単位で統合し出力する.

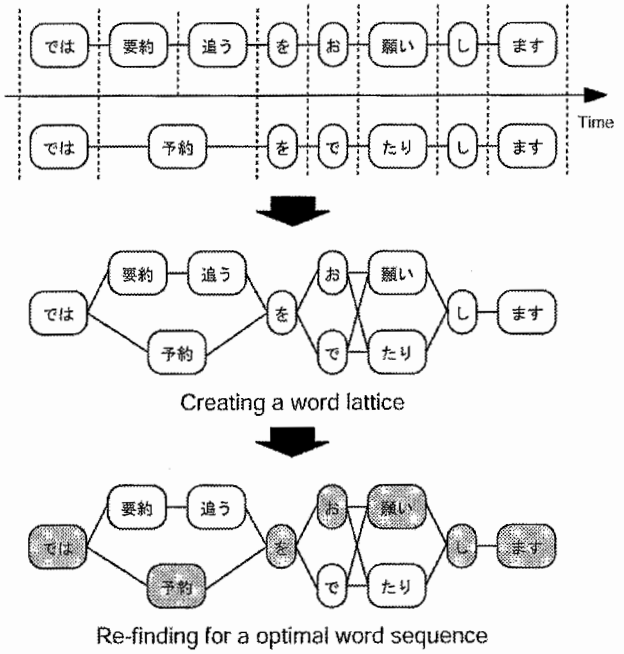

図 2: 仮説統合手法の概念図 ([14] より引用)．与えられた 2 つの仮説から個々の単語の開始および終了時間情報を用いて単語ラティスを再構成し，音響，言語尤度の最も大きな単語列を再探索する．

2 関連手法の概要

2.1 仮説選択，仮説統合

パラレルシステムにおいて複数のデコーダから得られた複数の仮説から 1 つの認識結果を生成する手法には，最大尤度基準で仮説を選択する仮説選択手法の他に，単語単位で得られた仮説を統合する仮説統合手法 [6] が提案されている (図 2)

．仮説選択は最も信頼性のある仮説を選択する手法であるため，デコーダのうち少なくとも 1 つが信頼性のある仮説を出力することが期待できる場合に用いられる．一方仮説統合は仮説ごとの信頼性の高い部分を結合する手法であるから，複数のデコーダの出力する仮説が相補性を持っている場合に適する．さまざまな雑音環境下で学習した複数の音響モデルを用いてデコーディングする場合，認識すべき音声の雑音環境と最も近い雑音環境下で学習した音響モデルの出力が，最も信頼できると考えられるため，認識結果の生成には仮説選択が適する．一方例えば「子音の認識性能がよい」「母音の認識性能がよい」など音素ごとの認識性能の傾向が異なる複数の特徴量をそれぞれ用いるデコーダによってパラレルシステムを構築するような場合は，仮説同士に相補性があり仮説統合が適すると考えられる．

2.2 雑音除去手法

フロントエンドにおける加法性雑音除去手法の基本的なものに Wiener Filter があげられる．Wiener Filter は入力音声のパワースペクトル $S(\omega)$ と雑音のパワースペクトル $N(\omega)$ を推定し，元信号 $x(t)$ との 2 乗誤差を最小化する元信号推定値 $x'(t)$ を求めるフィルタ $F(\omega)$ を

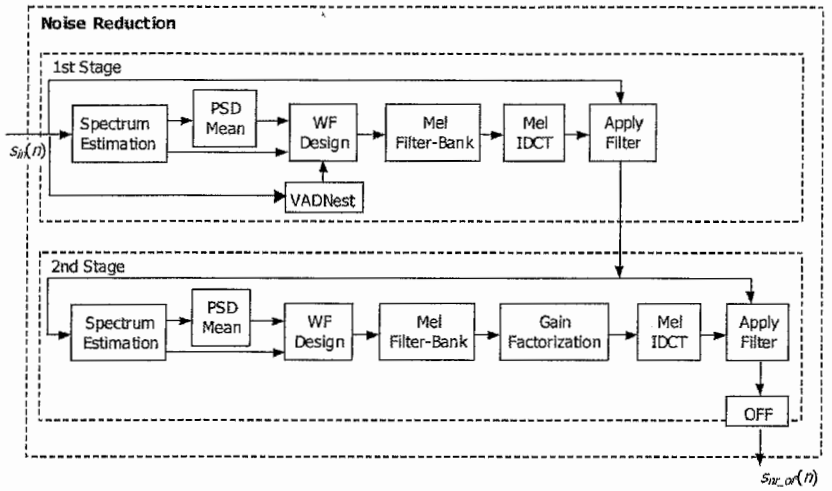

図 3: Advanced Front-End に用いられる雑音除去手法 ([1] より引用).

$$F(\omega) = \frac{S(\omega) - N(\omega)}{S(\omega)} \tag{1}$$

により求める手法である.

実環境の雑音の多くは非定常であるため，この Wiener Filter において雑音のパワースペクトル推定値 $N(\omega)$ を逐次更新する等，これを応用した手法が多く提案されている．欧州電気通信標準化機構 ETSI での分散音声認識標準フロントエンドに採用されている Advanced Front-End[1] の雑音除去部 (以下 AFE) もその 1 つである．AFE は，図 3 のように 2 つの Wiener Filter を段階的に用いる手法である．ここで Wiener Filter のゲイン (式 1) についてはスペクトルを逐次的に推定することにより動的に変更している．1 段階目の Wiener Filter においてはフィルタゲインを時間方向に平滑化し定常性の高い雑音を除去，2 段階目ではフレーム毎の SN 比に応じてフィルタゲインに補正を行い非定常性の強い雑音の除去を試みる.

その他に，音声特徴量の統計的分布を利用する手法 GMM Wiener 法 [5](以下 GMMW) も提案されている．この手法は音声特徴量の分布を事前に混合正規分布 (GMM) として学習し，入力音声が GMM 中のどの正規分布に属するかを帰属度として推定し，それを元に Wiener Filter を設計するものである.

クリーンな音声の MFCC 特徴量の統計量を GMM

$$p(S(i)) = \sum_{k=1}^{K} P_k p_k(S(i)|\mu_{S,k}, \Sigma_{S,k})$$

として事前に与える．$P_k, \mu_{S,k}, \Sigma_{S,k}$ はそれぞれインデックス k のガウス分布の重み，平均ベクトル，共分散行列である．ここで入力信号から推定した雑音の MFCC 特徴量分布

$$p(N(i)) = p(N(i)|\mu_N, \Sigma_N)$$

が得られた場合，この雑音下での音声の MFCC 特徴量分布は上の 2 つの式の合成により得られ，近似的に

$$p(X(i)) = \sum_{k=1}^{K} P_k p_k(X(i)|\mu_{X,k}, \Sigma_{X,k}) \quad (2)$$

$$= \sum_{k=1}^{K} P_k p_k(X(i)|\mu_{S,k} + \log[1 + \exp(\mu_N - \mu_{S,k})], \Sigma_{S,k}) \quad (3)$$

と表せる．ここから入力信号のMFCC $X(i)$ が雑音下の音声特徴量GMM $p(X(i))$ のどのガウス分布に属しているかを，それぞれのガウス分布への帰属度

$$P_{i,k} = \frac{P_k p_k(X(i)|\mu_{X,k}, \Sigma_{X,k})}{\sum_{k'=1}^{K} P_{k'} p_{k'}(X(i)|\mu_{X,k'}, \Sigma_{X,k'})}$$

を求めることができる．今インデックス k のガウス分布に属する音声に対し，雑音 $N(i)$ を除去するためのWiener Filterのフィルタゲインを $\mu_{S,k}, \mu_{X,k}$ の線形スペクトル領域表現 $\mu_{S,k}(\omega), \mu_{X,k}(\omega)$ を用い

$$F_k(\omega) = \frac{\mu_{S,k}(\omega)}{\mu_{X,k}(\omega)}$$

により表し，音声特徴量GMM $p(X(i))$ において各ガウス分布への帰属度が $P_{i,k}$ である入力信号 $X(i)$ に対するフィルタゲインを $F_k(\omega)$ の帰属度による重みつき和

$$F(\omega) = P_{i,k} F_k(\omega)$$

とし，フレーム毎に入力のスペクトルを推定しフィルタゲインを動的に変更しながらWiener Filterを適用するのがGMMWの原理である．

本研究ではここに述べたAFEとGMMWの二つの雑音除去手法について，パラレルシステムにおける有効性を検証することとした．

3 研究の概要

3.1 研究の目的

本研究の目的は日本語音声認識パラレルシステムにおける雑音除去手法の有効性を検証することである．有効性を検証する雑音除去手法はAFEとGMMWの2つとし，日本語大語彙連続音声認識のタスクを対象とする．ここで想定するパラレルシステムとは，さまざまな雑音環境下で学習した音響モデルを並列に用いてデコーディングし，仮説選択を行うシステムであるとする．またパラレルシステムでは仮説統合により雑音除去手法同士の相補性を活用することもできるため，それぞれの雑音除去手法の単独での性能の他に，仮説統合による相補性の評価も行う．

3.2 雑音除去手法の単独での性能の評価方法

パラレルシステムを実際に用いる評価は，シングルシステムの評価の数倍から数十倍といった非常に大きな計算コストが必要となる．しかしさまざまな雑音環境下で学習を行った複数の音響モデルを用いたパラレルシステムの仮説選択においては，入力音声の雑音環境に近い環境で学習したモデルによる仮説が選択されやすいことが実験的に知られている [14]．この性質を考えると，通常の Matched モデルによる評価でパラレルシステムの評価が近似できる．よって各雑音除去手法単独での性能評価実験は，以下のような Matched モデルによる評価を各雑音除去手法を用いた場合について行うこととした．

Well-Matched モデルの評価： Well-Matched とは学習, 評価データの雑音環境を完全に一致させた場合を表す．ここで「雑音環境」は雑音の種類とSN比から決まるとする．パラレルシステムにおいて，パラレルにデコーディングされる音響モデルの数が十分多い場合には，入力音声の雑音環境とほぼ同じ環境で学習したモデルの仮説が選択されると考えることができる．その場合システム全体の性能は学習, 評価時の雑音環境を完全に一致させた Well-Matched モデルの性能に等しい．したがって Well-Matched モデルの評価によって，十分な数の音響モデルが用いられる理想的な場合のシステムの性能，言い換えればパラレルシステムの性能の上限が評価できる．

Noise-Matched, SNR-Matched モデルの評価： 学習/評価時の雑音環境のうち雑音種,SN比の一方のみをそれぞれ一致させた場合を Noise-Matched, SNR-Matched と呼ぶこととする．実際のシステムにおいてはリソースの制約から並列にデコーディングできる音響モデルの数が限られるため，このような場合，各々の音響モデルはある範囲をもった雑音環境に対応する必要がある．各々の音響モデルがどのような範囲の雑音環境をカバーするべきかについて最適な方法は不明であるが，同じ種類の雑音や，同じSN比の雑音の性質は似ていることから，各音響モデルに特定の種類の雑音をカバーさせる，または特定のSN比の雑音をカバーさせる方法は有効な方法のひとつであると考えられる．この場合のシステムの性能は，学習/評価時の雑音種を一致させた Noise-Matched モデルの性能，SN比を一致させた SNR-Matched モデルの性能に近似できる．

3.3 雑音除去手法同士の相補性の評価方法

仮説統合を用いたときの雑音除去手法同士の相補性は，各雑音除去手法を用いた場合のMatchedモデルの認識結果同士を仮説統合し，その認識率により評価できる．

3.4 既存の研究と結果の予測

現在日本語と英語連続数字認識におけるWell-Matchedモデルを用いた評価で雑音除去手法の有効性が確認されている[2][12]．しかし日本語連続音声認識においての実験例はない．福士らは日本語連続音声認識におけるAFEの評価を，Cleanモデルと様々な雑音環境化の学習データから学習を行うMulticonditionモデルにより行った．そしてCleanモデルを用いた場合はAFEによる性能が向上するが，Multiconditionモデルを用いた場合の性能向上は小さいという結果を得た[9]．

この研究は，雑音除去手法は雑音下の音声をよりクリーンに近づけることができるが，雑音環境下の学習データが利用可能な場合は，必ずしも雑音除去手法を用いる必要があるとは言えないことを示唆している．したがって今回実験を行うMatchedモデルの評価においても，雑音除去手法の有効性は小さい可能性がある．

仮説統合による雑音除去手法同士の相補性についての研究例はまだない．関連する研究として松田ら[14]による2種類の音響特徴量の相補性についての研究がある．松田らはこの研究で，MFCCとDMFCC[4]の2種類の特徴量をそれぞれ用いたデコーダの仮説を統合し，音節強調発声音声の認識においてそれぞれの特徴量を単独に用いる場合と比べて大きく認識精度が向上することを確かめた．DMFCCはMFCCと比べて雑音の種類やSN比の変動の影響を受けにくい特徴があるため，MFCCを単独で用いた場合の入力音声と音響モデル間における雑音環境のミスマッチの影響をDMFCCとの仮説統合により軽減できたことがわかる．ただし通常発生音声の認識においてはこの相補性は認められず，相補性の有無はタスクに依存すると考えられる．

雑音除去手法間にも耐雑音性の傾向の違いがある可能性があり，仮説統合による性能の向上が期待できる．

4 実験

4.1 Well-Matched モデルによる評価

Well-Matched モデルの評価は以下の手順で行った

1. 学習用音声データと評価用音声データに，12 雑音環境 (4 雑音種× 3SN 比) の雑音をそれぞれ重畳する．
2. 雑音除去手法を用いない通常のフロントエンド (以下 NoNR) と，AFE を用いたフロントエンド (AFE)，GMMW を用いたフロントエンド (GMMW) の 3 種類について以下の 3~5 を行う
3. 各雑音環境下の学習用音声データと評価用音声データについて，フロントエンドを用いて特徴量を抽出し学習用音声パラメタを得る
4. 各雑音環境下の学習用音声パラメタを用いて音響モデルをそれぞれ学習する
5. 各雑音環境下で学習した音響モデルを用いて，同じ雑音環境下の評価用音声を認識し，単語正解精度を得る．このとき評価用音声のフロントエンドは 3. の学習で用いたものと同じとする

実験では音声認識エンジンとして，当研究所で開発した連続音声認識用デコーダ ATRASR[11] を用いた．言語モデルは，ATR 旅行会話基本表現集 BTEC[7] および，自然発話音声および，自然発話音声・言語データベース SDB, SLDB, LDB[8] に含まれる 6.1M 単語から生成した．辞書サイズは 34k である．第 1 パスは多重クラス複合 2-gram[3] を使用し，第 2 パスでは単語 3-gram を使用した．ビーム幅は 100.0 を用い，最大仮説数を 10000 として探索を行った．

雑音の重畳には単純に音声と雑音の全体平均パワーをもとに重畳するプログラム (addnoise とする) を用いた．サンプリング周波数 16kHz, 分析窓長 20ms, 分析周期 10ms とし，NoNR, AFE, GMMW の各フロントエンドにより MFCC 特徴量を抽出した．MFCC の音響特徴パラメータは 12 次元 MFCC, Δ pow,12 次元Δ MFCC の計 25 次元である．使用した音素は/ɴ, a, b, ʧ, d, e, f, g, h, i, ʒ, k, m, n, o, p, ʔ, r, s, ʃ, t ʦ, u, w, j, z/の 26 種類である．音響モデルの状態共有構造は，MDL-SSS[10] により生成した HMnet を使用した．各状態の混合数は 5 である．学習データとして，ATR 旅行会話データベース TRA, TRABLA(計 30 時間) を用いた．雑音重畳に用いた雑音は ATR 実環境騒音 DB[12] から車 (car1 とする), 工場 (factory), レストラン (restaurant), 展示会場 (exhibition) の 4 種類を用いた．また一部の実験では電子協雑音 DB からの車雑音 (car2) を用いた．雑音重畳音声の SN 比は 0dB, 10dB, 20dB である．評価用音声データは，ATR 旅行会話基本表現集 BTEC testset-01(510 文) を使用した．

次に各フロントエンドについて詳細を述べる．通常フロントエンド NoNR は，ATRASR に含まれる ATRwave2cep, ATRcms, ATRcep2para 各プログラムを，それぞれケプストラム抽出，CMS，パラメータ生成に用いた．AFE と GMMW については，各プログラムを直接利用しケプストラム抽出を行い，CMS とパラメータ生成は上と同じ ATRcms, ATRcep2para を用いた．この時 AFE の出力は音声の終端 6 フレームが欠け，GMMW の出力は始端 11 フレームが欠ける．学習用音声のラベル時刻とのずれを解消するため，AFE については欠けた 6 フレームの特徴量をその 1 つ前のフレームとすべて同じとし，GMMW については欠けた 11 フレームの特徴量をその次のフレームとすべて同じとする補完処理を行った．

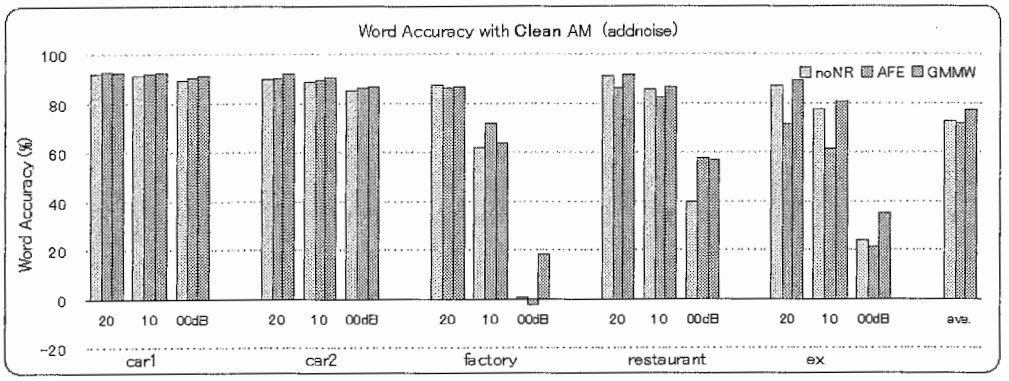

図 4: Clean 音響モデルを用いた場合の単語正解精度 (雑音重畳は addnoise). car1-20dB から exhibition-00dB までの各雑音環境の評価データに対し，3 つのフロントエンドを用いた場合の単語正解精度をそれぞれグラフにとったもの．単語正解精度が高い条件下ではフロントエンド間の差はあまり見られず，正解精度が低い条件下では GMMW の有効性が確認できる．一方 AFE は GMMW ほどの有効性は確認できない．

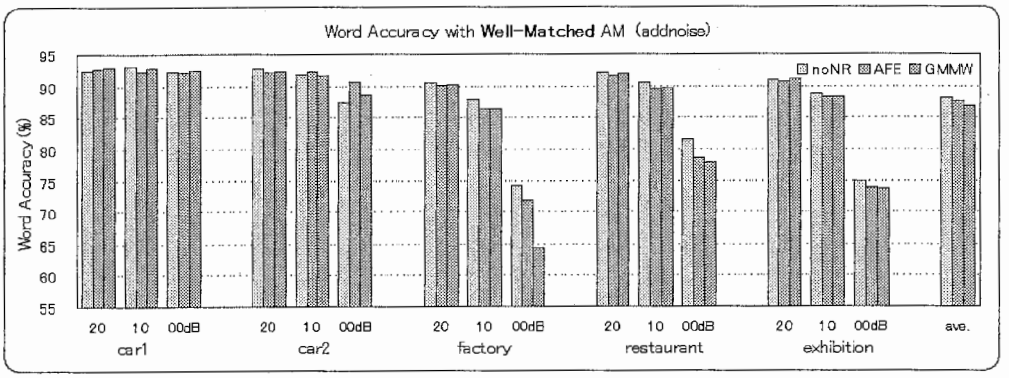

図 5: Well-Matched 音響モデルを用いた場合の単語正解精度 (雑音重畳は addnoise). ほとんどの条件で NoNR の正解精度が最も高い．

図 4 に Clean 音響モデルを用いたベースラインの単語正解精度，図 5 に Well-Matched モデルの単語正解精度を示す．ベースラインの結果は GMMW がほとんどの雑音環境において有効に働き認識率を向上させている一方，AFE はそれほど有効とは言えず，雑音除去手法を用いない場合のほうが平均単語正解精度が高くなっている．また Well-Matched モデルの場合では，SN 比が悪い場合で一様に AFE, GMMW を用いた場合の認識率が雑音除去手法を用いない場合を下回った．

Clean モデルの認識率が AFE を用いることで悪くなったという結果は，福士らの実験結果 [9] と食い違う．福士らの Clean モデルによる実験と本実験の比較を図 6 に示す．図のように福士らの実験では AFE が雑音種によらず有効であったが本実験では AFE の有効性はあまり見られない．この違いは新聞読み上げ, 旅行会話という実験タスクの違いから生じた可能性がある．また福士らの実験における雑音重畳の方法は記されていないが，この重畳方法による影響も考えられる．

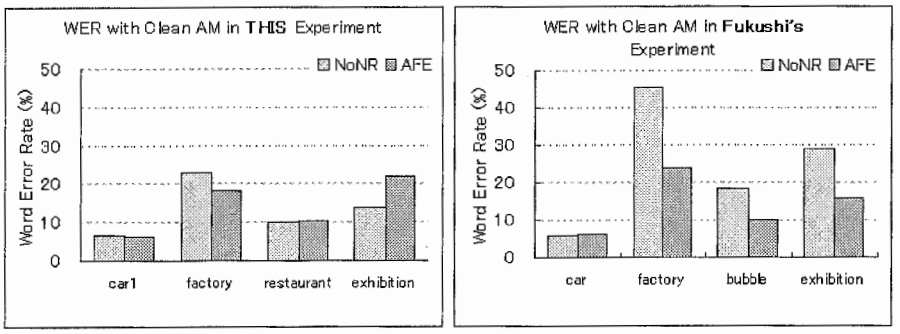

図 6: 福士らの実験 [9] と本実験における，AFE を用いた場合の Clean モデルによる日本語音声認識結果の比較．タスクは福士ら:新聞読み上げ, 男性のみ．本実験:旅行会話, 男性+女性.

また連続数字認識の場合 [12] と異なり，日本語大語彙連続音声認識においては Well-Matched モデルにおいて雑音除去手法が有効ではないという結果を得た．ここで連続数字認識の評価に用いられる AURORA-DB では雑音重畳段階で分散端末をエミュレートするフィルタが音声データにかけられているため，認識性能が異なる原因がこのフィルタにあるという可能性がある．また AURORA-DB の雑音重畳においては SN 比の評価基準も異なっている可能性も高い．AURORA-DB で用いられる雑音重畳プログラムを利用て作成した雑音重畳音声と上の実験で用いた重畳プログラムによる音声を聞いてみると，同じ SN 比であるにもかかわらず前者のほうがかなりノイジーに聞こえる．したがってこの実験結果の違いが単に雑音重畳方法の違いから生じている可能性がある．

認識性能の傾向が異なる原因が雑音重畳プログラムにあるのかどうかを確かめるため，AURORA-DB で用いられる雑音重畳プログラム (filter-addnoise とする) を用いて再実験を行った．上の実験との相違は，filter-addnoise の仕様によりサンプリング周波数を 8kHz とし，比較するフロントエンドを NoNR,AFE の 2 つとした点である．

図 7，図 8 にその結果を示す．Clean モデルを用いたベースライン正解精度では雑音 10dB の条件下で AFE の有効性が確認されたものの，Well-Matched モデルにおいては addnoise, filter-addnoise 間で大きな傾向の違いはない事がわかる．従って Well-Matched モデルにおける雑音除去手法の有効性が，数字認識と日本語音声認識で異なる原因は雑音重畳の方法ではなく，そのタスクの違いにあるのではないかと考えられる．また今回検証実験は行うことができなかったが，AFE で用いられている伝送路特性補償処理 (Blind Equalization) が，その後行った CMS 処理と干渉している可能性も考えられるため，今後検証が必要である．

4.2 Noise-Matched, SNR-Matched モデルによる評価

Noise-Matched, SNR-Matched モデルによる評価実験の手順は，Well-Matched 実験とほぼ同じである．相違点は，Noise-Matched 音響モデルの学習時には各 SN 比の学習データをまとめて用い，SNR-Matched 音響モデルの学習時には各雑音種の学習データをまとめて用いる点である．

図 9，図 10 にその結果を示す．Well-Matched の場合と同様に，Noise-Matched, SNR-Matched 両実験ともに雑音除去手法を用いない場合が最も性能が高かった．

雑音除去手法は雑音環境の変動による影響を小さくし Noise-Matched, SNR-Matched のよ

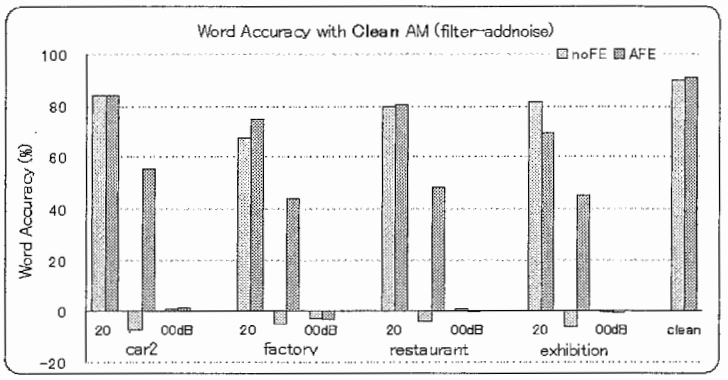

図 7: Clean 音響モデルを用いたベースラインの単語正解精度 (雑音重畳は filter-addnoise). SN 比 20dB,0dB では NoNR, AFE 間で大きな差はない (0dB においてはほとんど認識できない) が，10dB では一様に AFE が高い正解精度となっている.

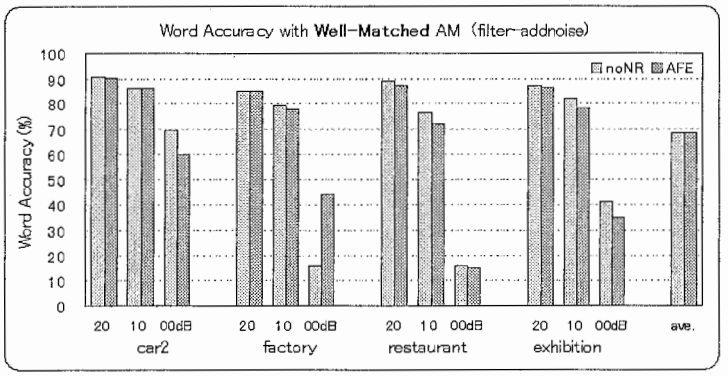

図 8: Well-Matched 音響モデルを用いた場合の単語正解精度 (雑音重畳は filter-addnoise). ほとんどの条件で NoNR の性能が AFE よりも高い.

うに一定の範囲を持った雑音環境を 1 つのモデルでカバーする場合には性能向上に寄与すると思われたが，実験ではその効果は得られなかった．原因としては，雑音の強い非定常性からくる雑音除去による音声への歪みの影響が大きかったことが考えられる.

Well-Matched モデル評価実験の結果とあわせて，日本語連続音声認識で Matched モデルを用いた場合には雑音除去手法は有効とはいえず，したがってパラレルシステムは雑音除去手法を必要としないということがわかった.

4.3 仮説統合による相補性評価

仮説統合による評価実験は，上述の Well-Matched, Noise-Matched, SNR-Matched それぞれの実験について，異なる 2 つのフロントエンドを用いた認識結果同士を仮説統合することで行った．仮説統合は NoNR+AFE, NoNR+GMMW, AFE+GMMW の 3 つのパターンで行った.

図 11, 12, 13 にそれぞれの結果を示す．どの組み合わせの仮説統合も NoNR の認識率を大

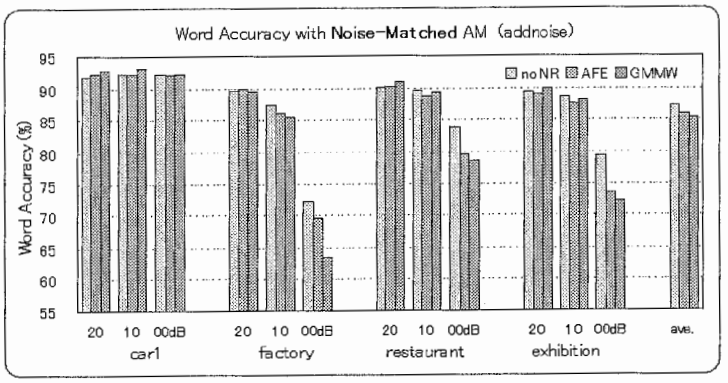

図 9: Noise-Matched 音響モデルを用いた場合の単語正解精 (雑音重畳は addnoise). 多くの場合で NoNR の正解精度が最も高い.

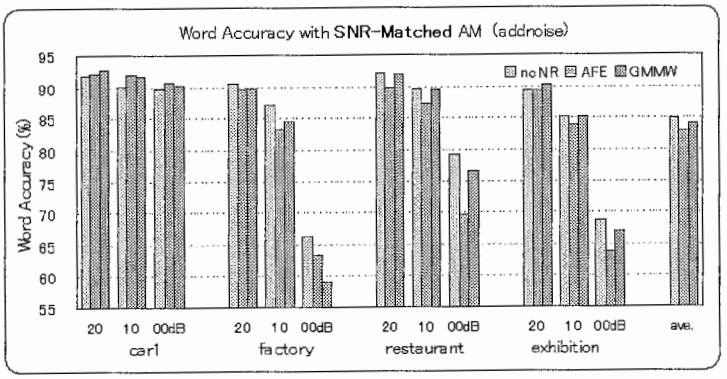

図 10: SNR-Matched 音響モデルを用いた場合の単語正解精 (雑音重畳は addnoise). Noise-Matched 同様に NoNR の性能が最も高い.

きく改善することはできなかった. したがってフロントエンド間に大きな相補性は見られないことがわかった.

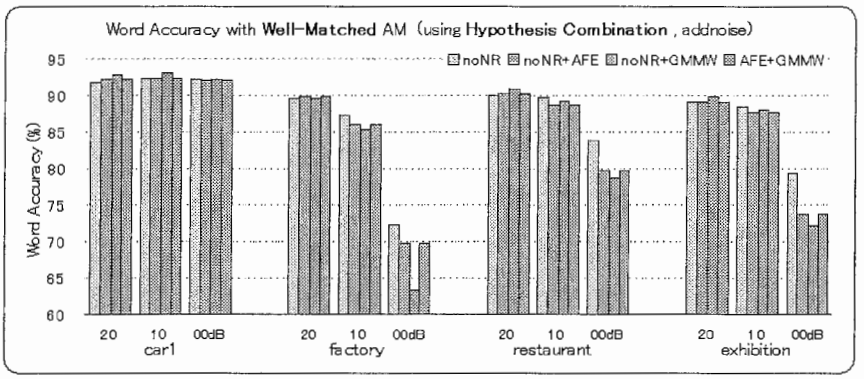

図 11: Well-Matched モデルにおいてフロントエンド同士の仮説を統合した場合の単語正解精度．各雑音環境について，NoNR 単独, NoNR+AFE, NoNR+GMMW, AFE+GMMW の単語正解精度をグラフに表したもの．仮説統合による正解精度の向上は大きくない．

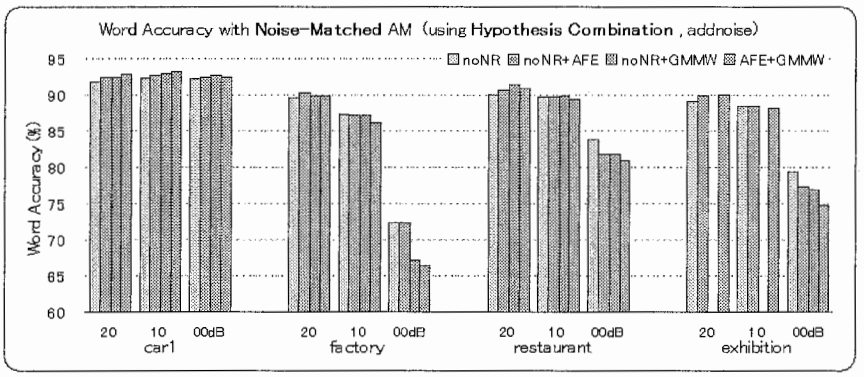

図 12: Noise-Matched モデルにおいてフロントエンド同士の仮説を統合した場合の単語正解精度

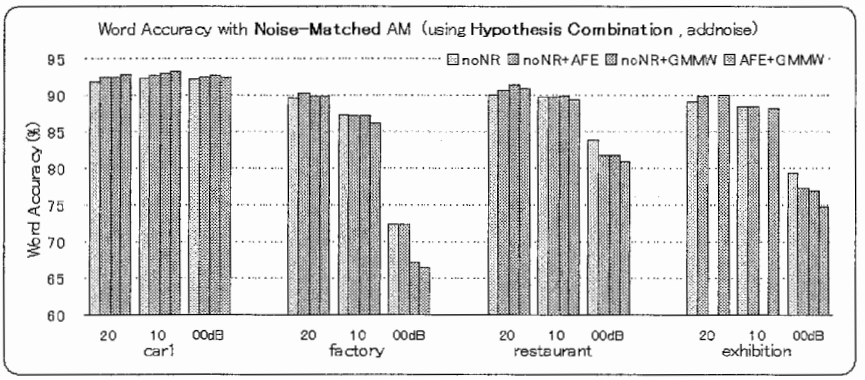

図 13: SNR-Matched モデルにおいてフロントエンド同士の仮説を統合した場合の単語正解精度

5 おわりに

5.1 本研究のまとめ

本研究では，パラレルデコーディングタイプの日本語大語彙連続音声認識システムにおける雑音除去手法の有効性と相補性を検証する目的で，いくつかのMatchedモデルを用いた評価実験を行った．

その結果パラレルシステムにおいて雑音除去手法は有効とはいえず，またその相補性も見られないということを確認し，したがって日本語連続音声認識のパラレルシステムは雑音除去手法を必要としないという結論を得た．

5.2 今後の課題

今後の課題の1つに，前述のAFE内部のBlind EqualizationとCMSの干渉がないかどうかを検証する必要がある．また連続数字認識で認められたMatchedモデルにおける雑音除去手法の有効性が，日本語連続音声認識で認められなかった原因についても検証を行っていく必要がある．このようなタスクの性質と雑音除去手法の性能の関係が明らかになれば，特定タスクに特化した雑音除去手法の研究などに結びつくと考えられる．

6 謝辞

今回，ATRにおける実習の機会を与えて下さいましたASR研の中村室長に心から感謝します．また直接の指導者として手取り足取り教えていただいたASR研の松田さんに感謝します．実習生活全般にわたって御世話いただいた秘書の鈴木さんに感謝します．また同研の藤本さん，遠藤さんをはじめASR研の皆様には研究面のみならず生活面でも大いに御世話になりました．皆様のおかげで充実した実習期間を過ごすことができたことを心から感謝します．ありがとうございました．

参考文献

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

[2] D.Macho, L.Mauuary, B.Noe, Y.M.Cheng, D.Ealey, D.Louvet, H.Kelleher, D.Pearce, and F.jSaadoun. Evaluation of a noise-robust front-end on aurora databases. In *Proc. ICSLP*, Vol. 1, pp. 17–20, 9 2002.

[3] H.Yamamoto and Y.Sagisaka. Multi-class composite n-gram language model based on connection direction. *Proc. ICASSP*, pp. 533–536, 1999.

[4] J.Chen, K.K.Paliwal, and S.Nakamura. Cepstrum derived from differentiated power spectrum for robust speech recognition. *Speech Communication*, Vol. 41, No. 2-3, pp. 469–484, 2003.

[5] J.C.Segra, A.de la Torre, M.C.Benitez, and A.M.Peinado. Model-based compensation of the additive noise for continuous speech recognition. experiments using the aurora ii database and tasks. In *Proc. Eurospeech 2001*, pp. 221–224, 9 2001.

[6] K.Markov, T.Matsui, R.Gruhn, J.Zhang, and S.Nakamura. Noise and channel distortion robust asr system for darpa spine2 task. *IEICE Transaction of Information and System*, Vol. E86-D, No. 3, 2003.

[7] T.Takezawa, E.Sumita, F.Sugaya, H.Yamamoto, and S.Yamamot. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real wor. In *Proc.LREC2002*, pp. 147–152, 2002.

[8] T.Takezawa, T.Morimoto, and Y.Sagisaka. Speech and language databases for speech translation researched in atr. In *Proc. EALEREW*, pp. 148–155, 1998.

[9] 福士なな子, 加藤正治, 小坂哲夫, 好田正紀. ETSI 標準フロントエンドを用いたマルチコンディション学習による雑音重畳音声認識の検討. 日本音響学会講演論文集, pp. 15–16, 9 2003.

[10] 實廣貴敏, 松井知子, 中村哲. MDL 基準を用いた逐次状態分割法. 日本音響学会講演論文集, pp. 41–42, 9 2002.

[11] 伊藤玄, 葦苅豊, 實廣貴敏, 中村哲. 音声認識統合環境 ATRASR の概要と評価報告. 日本音響学会講演論文集, pp. 221–222, 9 2004.

[12] 遠藤俊樹, 中村哲. 実環境雑音 DB の収集及び DSR フロントエンドによる音声認識実験. 日本音響学会講演論文集, pp. 187–188, 9 2004.

[13] 中村哲. 外乱に対良い音声認識を目指して. 日本音響学会誌, Vol. 57, No. 10, pp. 662–667, 2001.

[14] 松田繁樹, 實廣貴敏, コンスタンティンマルコフ, 中村哲. 雑音や発話スタイルの変動に頑健な日本語大語彙連続音声認識. 情報処理学会研究報告, Vol. 2004-SLP-50, pp. 37–44, 2004.