2004－SLP－50　(6)
2004／2／7

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

松田 繁樹　實廣 貴敏　コンスタンティン マルコフ　中村 哲

ATR 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

あらまし

実環境において音声認識を用いるためには雑音や発話スタイルの変動に対して頑健でなければならない．本報告では，お互いに異なる雑音環境や発話スタイル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得られた複数の仮説を最大尤度基準や単語単位での仮説統合を行う手法について検討を行った．個々の音響モデルが利用可能な発話環境の空間が限られていたとしても，複数の音響モデルを用いて認識を行い得られた複数の仮説を効果的に統合することにより，種々の外乱に対して頑健な音声認識を実現することができると考えられる．本報告で構築した音声認識システムに対して日本語大語彙連続音声認識実験を行った結果，SNR が 10dB 以上の通常発声音声に対して 90%以上の単語正解精度が得られた．また，意図的に音節毎に区切って発話した音節強調発声に対して約 45% の単語正解精度が得られた．

キーワード

音声認識, HMM, 頑健性, 雑音, 発話スタイル, パラレルデコーディング, 仮説統合

LVCSR Robust to Noise and Speaking Styles

Shigeki Matsuda　Takatoshi Jitsuhiro
Konstantin Markov　Satoshi Nakamura

ATR Spoken Language Translation Research Laboratories
2–2–2 Hikaridai “Keihanna Science City” Kyoto 619–0288 Japan

Abstract

To make the LVCSR (Large Vocabulary Continuous Speech Recognition) system that is robust to noise and speaking style change, we present our LVCSR system that combines multiple hypotheses obtained by parallel decoding using multiple acoustic models, each trained from a different environment. The performance of the LVCSR system was evaluated by the normal speaking style and hyperarticulated speech data contaminated by various kinds of noises at different SNR levels. Experimental results show that the system could achieve over than 90% word accuracy for the normal speaking style data when the SNR is above 10dB, and about 45% word accuracy for the hyperarticulated speech.

Key words

Speech Recognition, HMM, Robustness, Noise, Speaking Style, Parallel Decoding, Hypothesis Combination

1 はじめに

近年，雑音や発話スタイルに対して頑健な音声認識の研究が盛んに行われている．実環境において音声認識を使用するためには，公共バスや自動車などの乗り物から発せられるエンジン雑音や風切り音，展示会場やオフィス内などの人の声，計算機からのファンの音など多種多様な雑音環境において，高精度な音声認識が実現されなければならない．さらに雑音だけでなく，使用者の年齢や性別，また感情や体調によってその発話スタイルは刻一刻と変化するため，このような変動に対しても雑音と同様頑健でなければならない．従来より雑音や発話スタイルなど個別の変動に対する頑健化手法が数多く提案 [1] されてきた．本報告では，音声の音響的言語的特徴に影響する要因のことを総じて「発話環境」と呼ぶこととする．

雑音に頑健な音響特徴量として，SS(Spectrum Subtraction) 法 [2] を音響分析の前処理として用いる手法や，RASTA(RelAtive SpecTrA)[3], DMFCC (Differential Mel Frequency Cepstrum Coefficient)[4] など，幾つかの音響分析手法が提案されている．SS 法では，雑音重畳音声のスペクトルに対して雑音スペクトルを減算することにより SNR(信号対雑音比) を改善している．RASTA では，個々の周波数バンドの値の変化に対して，音声情報が多く含まれている 1〜12Hz の変調スペクトラム成分を抽出することにより雑音の影響を軽減している．また，DMFCC は FFT によって得られるフーリエ係数に対して隣り合う係数間で差分を取り音声などのピッチを持つスペクトルを強調することによって耐雑音性を改善している．

雑音に頑健な音響モデルの研究としては，PMC(Parallel Model Combination)[5] 法，ヤコビ適応法 [6]，MLLR(Maximum Likelihood Linear Regression)[7] による雑音適応などが提案されている．PMC 法は，HMM の出力確率分布を線形スペクトル領域へ変換し雑音スペクトルを重畳することにより，環境雑音への適応を行う手法である．ヤコビ適応法は，雑音の変化にともなう出力確率分布の非線変換を線形近似することにより，雑音環境へ高速に適応する手法である．MLLR を用いた雑音適応では，クリーン音声と雑音重畳音声の間の分布移動を回帰行列を用いて表現し，音響モデル全体を雑音環境へ適応化する手法である．更に雑音の分布の時間変化を逐次的に推定することにより，非定常雑音に対する認識性能を改善する手法 [9] が提案されている．

発話スタイルに対する頑健性の改善手法としては，発話スタイル依存の音響モデルを用いる手法の他，Lombard 効果によるスペクトルの変形を考慮した手法 [8] や，個々の母音 HMM の最後に無音状態を追加することにより音節強調発声や言い直し発話に頑健な音響モデルを構築する手法 [10] などが提案されている．その他にも，講演音声などの音素継続時間の短い発声を含む音声に対して，分析フレーム周期やウインドウ幅を自動選択することにより認識性能を改善する手法 [11, 12] が提案されている．

これらの頑健化手法は主として，雑音や発話スタイルなどの個別の変動に対する頑健化である．音声認識を実環境で用いるためには，複数の発話環境が刻一刻と変化する状況であっても頑健に音声を認識することのできなければならない．このような種々の外乱に対して頑健な音声認識を実現するための方法は大きく 2 つに分類することができると考えられる．発話環境の変動に頑健な音響モデルや言語モデルを用い，単一のデコーダで認識を行うシングルタイプの方法と，お互いに異なる環境に適応化された複数の音響モデルや言語モデルを使用し，得られた複数の仮説を統合するパラレルタイプの手法である．

シングルタイプの音声認識システムを構築するためには，広い発話環境の音声を頑健に認識する音響モデルや言語モデルが必要である．男性女性両方の学習データから性別独立な音響モデルを推定するなど，複数の発話環境のデータを用いて HMM(Hidden Markov Model) のモデルパラメータ推定を行うことにより頑健性を改善する手法がある．しかし，男性女性などのお互いの音響的特徴が大きく異なる場合ではなく，種々の SNR のデータを用いて学習する場合，個々の音素モデルの分布が過度に広がることにより音素分類性能の低下が懸念される．従って，このようなモデル化法には頑健化の限界があると考えられる．セグメントモデル [13] では，時間的に離れた音響特徴ベクトル間の相関を計算することで音声の非定常な振る舞いのモデル化する手法である．特徴ベクトル間の関係を利用することで，発話環境の変動に強いモデル化を実現できる可能性があるが，効率的な相関の計算方法やモデルパラメータの増大などの問題により十分な性能は得られていない．

一方，パラレルタイプによる音声認識は，個々の音響モデルや言語モデルの利用可能な発話環境

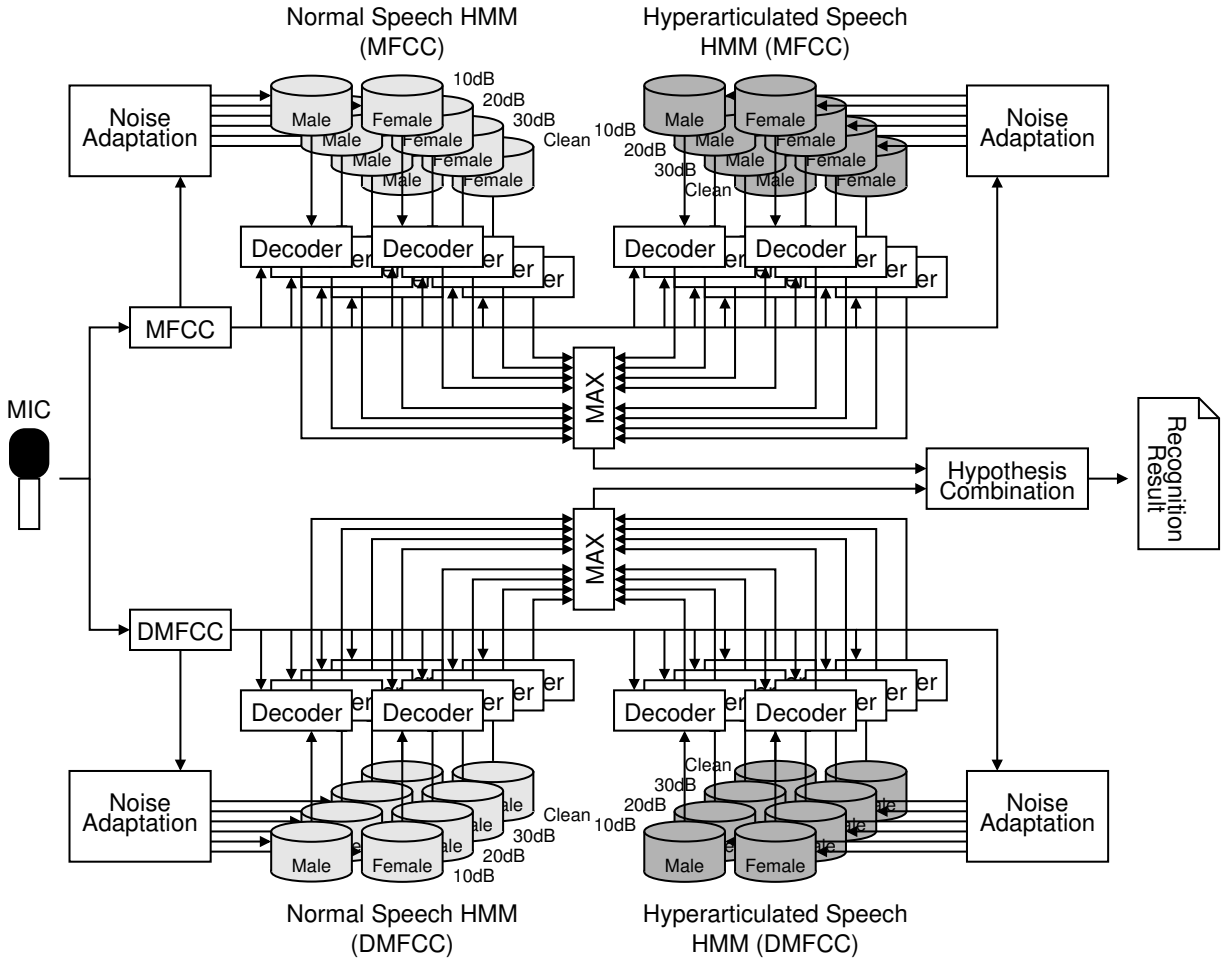

図 1: 統合システム全体の構造

が限られていたとしても，それらを複数個使用しパラレルにデコーディングすることにより，個々の音素間の分類性能を低下させることなく広い発話環境の音声を頑健に認識できる可能性がある．このような音声認識の例としては，SNR に依存した音響モデルを用いて得られた複数の仮説を最大尤度基準で選択する手法や，複数のお互いに異なる音響特徴量を用いて音声認識を行ない，得られた複数の仮説を単語単位で統合する仮説統合法 [15] が提案されている．本報告では，パラレルタイプを基礎とする方法により，雑音などの個別の変動だけでなく発話スタイルの変動に対しても頑健に音声を認識する方法について検討を行った．

第 2 章では，本報告で構築した雑音と発話スタイルに頑健な音声認識システムの構造及び，本システムで使用した頑健化手法について述べる．以降，本システムを「統合システム」と呼ぶこととする．第 3 章では，個々の頑健化手法を組み合わせることによる音声認識性能の改善を調べるため，日本語大語彙連続音声認識実験を行う．第 4 章は，まとめである．

2 統合システム

2.1 統合システムの構造

図 1 に本報告で構築した統合システム全体の構造を示す．図のように本統合システムは，大きく MFCC 特徴量部と DMFCC 特徴量部 [4] に分解することができる．各々の特徴量部には，合計 16 個のデコーダと音響モデルを持ち，得られた 16 個の仮説から最大尤度基準で仮説が選択される．その後，これら 2 つの部分から得られた仮説は Markov らの提案した仮説統合手法 [15] により単語単位で統合される．各特徴量部のデコーダで用いられる音響モデルは，発話スタイルの変動に

対する頑健性を高めるため，通常発声用音響モデルと奥田らの提案した言い直し発話に頑健な音響モデル [10] を用いた．更に，個々の音響モデルを雑音強度や雑音の種類の変動に高速に適応するため，伊田らの提案した高速な雑音適応手法 [14] と複数の SNR の音響モデルを用いてパラレルにデコーディングする手法を用いた．このシステムにより，雑音の種類と SNR の変動及び，言い直し発声に対して頑健な音声認識が実現できると考えられる．次節から，DMFCC 特徴量，個々の頑健化及び仮説統合による高精度化手法について述べる．

2.2 雑音の変動に頑健な音響特徴量

本節では，雑音の変動に対して頑健な特徴量として，Chen らの提案した DMFCC 特徴量 [4] について述べる．DMFCC 特徴量は，式 (1) に示す DPS(Differential Power Spectrum) を基礎とする特徴量である．式中の $Y(i,k)$ は，第 i 番目のフレームにおける第 k 番目のパワースペクトラム係数を表す．同様に $D(i,k)$ は，第 i 番目のフレームにおける第 k 番目の DPS 係数を表す．DMFCC 特徴量は，この DPS 係数に対して DCT(Discrete Cosine Transform) を行うことにより抽出される．

$$D(i,k) = |Y(i,k) - Y(i,k+1)| \qquad (1)$$

有声母音などピッチを含む音声から抽出されたパワースペクトラムは，基本周波数の高調波の影響によって串型の形状を持つ．このようなパワースペクトラムから DPS 係数を計算した場合，隣合うパワースペクトラム係数間の差が大きいため DPS 係数の値も同様に大きなパワーとして計算される．一方，雑音などのピッチを持たない波形のパワースペクトラムから計算される DPS 係数は，隣合うパワースペクトラム係数間の差が小さいため DPS 係数の値も小さくなると考えられる．雑音重畳音声のパワースペクトラムをクリーン音声のパワーと雑音のパワーの和であると仮定した場合，DPS 係数を計算することによって，(音声と比較して) なだらかに変化する雑音のパワー成分を減衰させることができると考えられる．

DMFCC 特徴量は上述の雑音パワーの減衰効果だけでなく，第 2.5 節で述べる仮説統合による認識性能の改善効果が高いことが報告 [15] されている．本統合システムでは，MFCC 特徴量と DMFCC 特徴量で別々にデコーディングを行い，得られた仮説の統合による音声認識性能の改善を試みた．

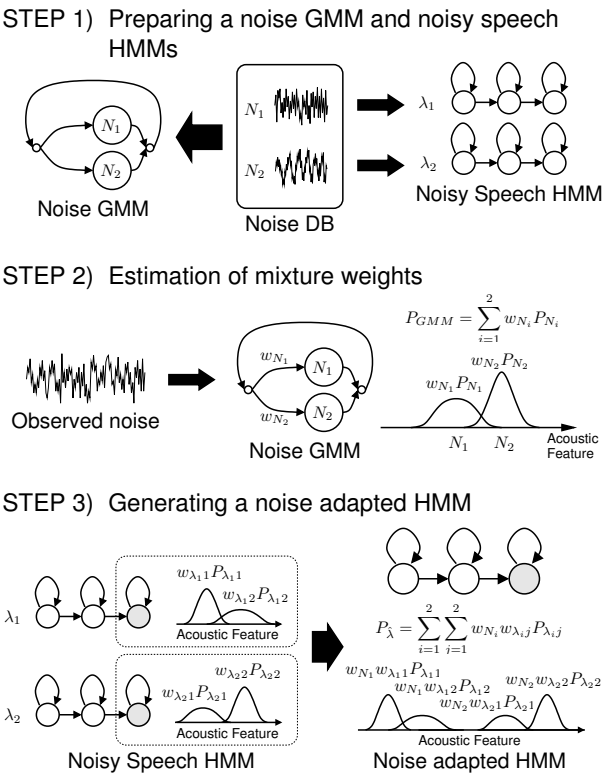

図 2: 雑音 GMM の混合重み適応化による HMM 合成の概念図

2.3 高速な雑音環境適応

雑音環境が頻繁に変動する状況では，音響モデルを高速に雑音環境へ適応できなければならない．本統合システムでは高速な雑音環境適応として，伊田らの提案 [14] した雑音 GMM の混合重み適応化による HMM 合成法を用いた．図 2 に，高速雑音環境適応手法の概念図を示す．この手法は，1) 予め準備した種々の雑音から，個々の雑音を混合成分とする雑音 GMM と，個々の雑音に対して別々に適応化された雑音重畳音声用 HMM を推定し，2) 短時間の未知雑音を用いて雑音 GMM の混合ウエイトのみを推定する．3) 最後に，この混合ウエイトを用いて，雑音重畳音声用 HMM を状態レベルで複数混合化することにより雑音環境への適応化が行われる．図中の N_i は第 i 番目の雑音，λ_i は第 i 番目の雑音に対する雑音重畳音声用 HMM を表す．P_{N_i} と w_{N_i} は雑音 GMM における第 i 番目の雑音の分布と，その分布への混合ウエイトである．更に，$P_{\lambda_i j}$ と $w_{\lambda_i j}$ は第 i 番目の雑音を用いて推定された雑音重畳音声用 HMMλ_i における第 j 番目の混合分布と分岐確率を表す．

この手法の利点として，適応の計算時間が

表 1: 評価実験に使用した雑音

雑音適応用に使用した雑音	
関空国内線ロビー	ANA エアバス
JR 東京駅地下中央通路	自動車内走行音
地下街食品売り場	渋谷はち公前
JR 東京駅西口	JR 東京駅ホーム
ATR 地下	新幹線ひかりグリーン席
北陸本線近くのたんぼ	滋賀県信楽町の森

評価用音声に重畳した雑音		
学園前駅ロータリ	公共バス	建設工事現場

GMM の混合ウエイトの推定時間のみであり大変高速である点と，雑音適応された HMM が複数の雑音環境の分布を含んでおり，単一の雑音から推定された音響モデルよりも雑音の短時間の変動に対する頑健性が高い点を挙げることができる．

第 3 章で行った評価実験では，表 1 上段に示す 12 種類の雑音を準備し，MFCC 特徴量用の雑音重畳音声用 HMM は PMC 法を用いて生成した．DMFCC 特徴量の雑音重畳音声用 HMM は，雑音を重畳した学習データを用いて ML 推定を行った．DMFCC 特徴量に対して PMC 法を適用することができないためである．

2.4 言い直し発話に頑健な音響モデル

本統合システムでは，発話スタイルの変動への対応としてシステムへの言い直し時に頻繁に観測される音節強調発話に対する頑健性の改善を試みた．音声認識ソフトウェアが認識誤りを起こした場合，そのソフトウェアの使用者はもう一度同じ発声を繰り返さなければならない．このような言い直し発話では，母音の後に短時間のポーズが挿入されるなど通常発声とは異なる音響的特徴を持つことが奥田ら [10] によって報告されている．奥田らはこの言い直し発話を頑健に認識するため，図 3 に示す構造を持つ音響モデルを提案した．母音モデルは，母音の後に短時間ポーズを挿入するため，`t-a+sil` の状態パス及び，その母音モデルの後にポーズ状態を追加した状態パスの合計 3 つの遷移を許すマルチパス音響モデルの構造を持つ．更に，子音モデルの前に短時間ポーズの挿入を許すため，通常の子音モデルに加えて `sil-k+i` の状態パスへの遷移が追加されている．このような音響モデルを用いることにより，通常発声の音声以外にも言い直しや音節強調発声などの音声を頑健に認識することが可能となる．

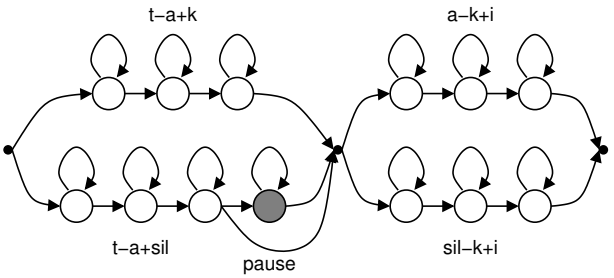

図 3: 言い直し発声に頑健な音響モデルの構造 (図中の `t-a+k` などの表記は，先行音素が/t/，後続音素が/k/，当該音素が/a/ の環境依存音素を表す)

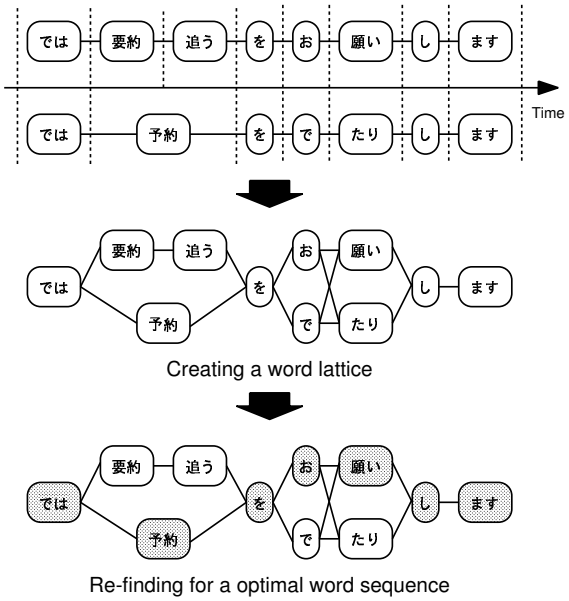

図 4: 仮説統合手法の概念図

2.5 仮説統合

本統合システムでは，複数の仮説の単語単位での統合を行う [15] ことによる音声認識性能の高精度化を試みた．複数の音声認識デコーダから得られた仮説がお互いに相補的である場合，各々の仮説の正しい部分を抽出することにより，より正しい単語列が得られる可能性がある．ここで述べた「相補的」とは，あるデコーダの認識結果の前半は正しいが後半は間違いであったとしても，別のデコーダの認識結果の後半部分は正しいならば，各々の正しい部分をつなぎ合わせることにより，その認識誤りを補償することができると言う意味である．

図 4 に仮説統合手法の概念図を示す．図のように仮説統合は，1) 与えられた 2 つの仮説から，

個々の単語の開始及び終了時間情報を用いて単語ラティスを再構成し，2) 音響と言語尤度の最も大きな単語列を再探索することによって行われる．本統合システムでは，MFCC と DMFCC 特徴量部から得られた仮説に対する仮説統合を試みた．ただし，MFCC の音響モデルから計算される音響尤度と，DMFCC の音響モデルから計算される尤度を直接比較することはできない．そのため，音響モデルの尤度を比較するためには尤度の正規化が必要である．本報告では，認識文全体の音響尤度で個々の単語の尤度を割ることにより正規化を行った．

3 日本語大語彙連続音声認識実験

3.1 雑音適応による頑健化の評価実験

第 2.3 節で述べた雑音適応化手法の評価を行うため，日本語大語彙連続音声認識実験を行った．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 ATRIUMS を用いた．言語モデルは，ATR 旅行会話基本表現集 BTEC[17] 及び，自然発話音声及び，自然発話音声・言語データベース SDB，SLDB，LDB[16] 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[18] 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．実験に使用した音声波形は，サンプリング周波数 16kHz，分析窓長 20ms，分析周期 10ms で分析を行ない，MFCC 及び DMFCC 特徴量を抽出した．MFCC の音響特徴パラメータは，12 次元 MFCC，Δc_0，12 次元 ΔMFCC の計 25 次元である．DMFCC の音響特徴パラメータは，12 次元 DMFCC，Δpow，12 次元 ΔDMFCC 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, ʕ, r, s, ʃ, t ts, u, w, j, z/の 26 種類である．音響モデルの状態共有構造は，ML-SSS[19] により生成した 2100 状態の 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 を用いた．407 名が発声した対話及び，音素バランス 503 文の計 30 時間である．雑音適応元の音響モデルは，表 1 上段に示す 12 種類の雑音を用いて生成した．MFCC の音響モデルは，雑音と SNR 毎に PMC 法を用いてクリーン音声 HMM を適応化することにより生成した．DMFCC の音響モデルは，雑音を重畳した学習データを用いて生成した．雑音重畳音声の SNR は，10dB，20dB，30dB である．MFCC と DMFCC 各々の音響モデルは，男性女性，12 種類の雑音，3 種類の SNR の，$2 \times 12 \times 3 = 72$ 種類とクリーン音声用モデルの計 73 種類である．評価用音声データは，ATR 旅行会話基本表現集 BTEC testset-01(510 文，男性 4 名，女性 6 名，各々51 文の発声データ) を使用し，10dB，20dB，30dB の SNR で雑音を重畳した．評価用に重畳した雑音を表 1 下段に示す．雑音 GMM の混合ウエイト推定には 1 秒間の雑音を使用し，個々の混合ウエイトの上位 4 つの雑音を用いて雑音重畳音声用音響モデルを生成した．

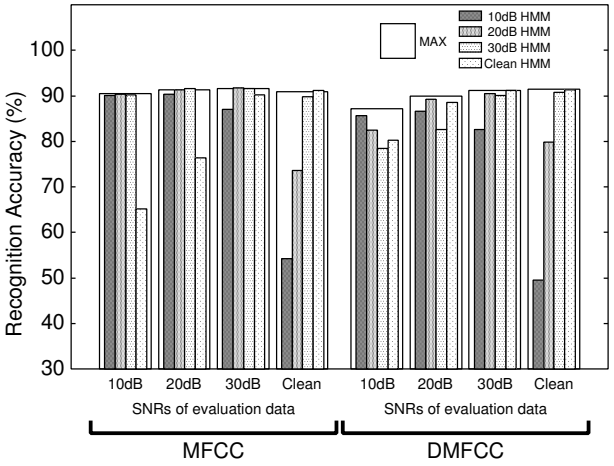

図 5: 雑音重畳音声に対する平均単語正解精度

図 5 に，3 種類の評価用雑音重畳音声データに対する平均単語正解精度を示す．図中の **MAX** は，個々の音響モデル (10dB，20dB，30dB，clean) を用いて得られた仮説を最大尤度基準で選択した場合の単語正解精度である．図に示すように，最大尤度基準による選択を行うことで，実験に用いた SNR 全てにおいて平均 90% 以上の単語正解精度が得られた．DMFCC の音響モデルは，MFCC の音響モデルよりも単語正解精度が低下している．しかし，DMFCC のクリーン音声音響モデルは，雑音重畳音声の単語正解精度が MFCC のクリーン音響モデルよりも高く，雑音の種類や雑音 SNR に対する正解精度への影響が MFCC よりも小さいことがわかる．

3.2 言い直し発話用音響モデルによる頑健化の評価実験

第 2.4 節で述べた言い直し発話に頑健な音響モデルに対して雑音と発話スタイルの変動に対する

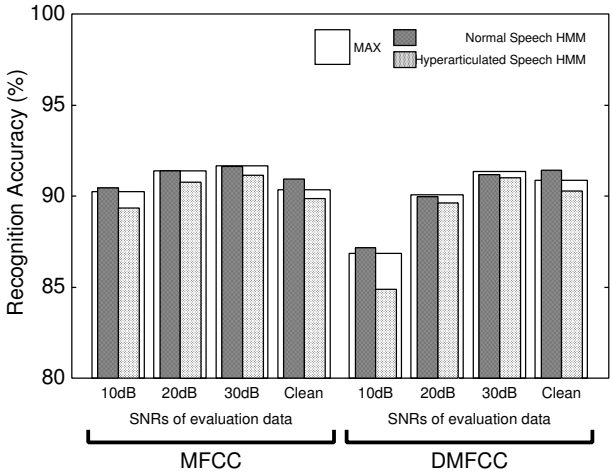

図 6: 雑音重畳した通常発声の音声に対する言い直し発話用音響モデルの平均単語正解精度

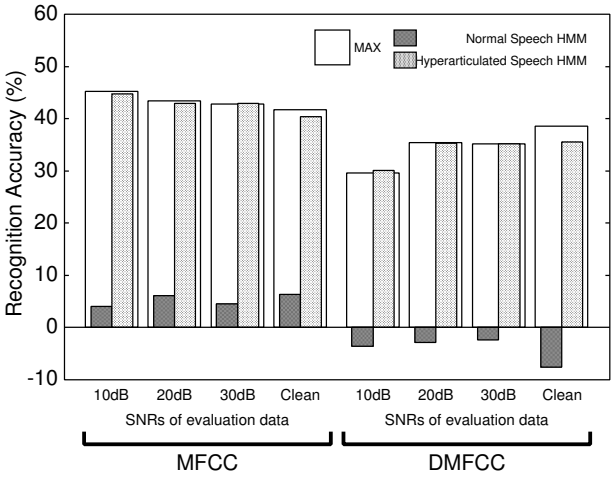

図 7: 雑音重畳した言い直し発話の音声に対する言い直し発話用音響モデルの平均単語正解精度

単語正解精度への影響を調べるため日本語大語彙連続音声認識実験を行った．評価用音声として，第 3.1 節で用いた通常発声の音声と，意図的に音節毎に区切って発声した音節強調発声の音声を用いた．音節強調発声データは，旅行会話文，男性 2 名女性 2 名，各話者 10 文の計 40 文である．評価用音声には，30dB，20dB，10dB の SNR で第 3.1 節で述べた 3 種類の雑音が重畳されている．言い直し発話に頑健な音響モデルは，環境依存音素モデル数が通常発声モデルよりも多い．そのため，探索空間が大きく広がり，通常発声音声に対して単語正解精度の低下が懸念される．そこで，本統合システムでは，言い直し発話用音響モデルと通常発話用音響モデルを別々にデコーディングし最大尤度基準による仮説の選択を行った．

図 6 に，通常発声用音響モデルの場合，言い直し発話用音響モデル単独の場合，2 つの音響モデルをパラレルデコーディングした場合，各々に対する単語正解精度を示す．図に示すように，言い直し発声用音響モデル単独で使用した場合その単語正解精度は若干低下するのに対して，パラレルデコーディングを行うことにより通常発声の音声に対してもほぼ同等の正解精度が得られた．

次に，音節強調発声の音声に対する単語正解精度を図 7 に示す．図に示すように，言い直し発話用音響モデルは通常発声用音響モデルよりも高い単語正解精度が得られた．また，雑音重畳音声に対しても，第 3.1 節で得られた結果同様，10dB の音声に対してもクリーン音声や 30dB の音声と同程度の単語正解精度が得られた．

3.3 仮説統合の評価実験

最後に，MFCC 特徴量と DMFCC 特徴量のデコーダから得られた仮説を統合することによる性能の改善を調べるための評価実験を行った．予備実験から，仮説統合時における言語モデルウエイトを 0.06 とした．

図 8 に，仮説統合を行った場合の単語正解精度を示す．図に示すように，通常発声に対しては MFCC 特徴量の正解精度と同等の結果が得られた．更に，音節強調発声に対しては，MFCC と DMFCC 各々の正解精度以上の性能が得られた．これは，仮説統合により，各々の特徴量の仮説がお互いに相補的であったためと考えられる．

4 まとめ

本報告では，雑音と発話スタイルの変動に頑健な音声認識を実現するための方法について検討を行った．更に，お互いに異なる雑音環境や発話スタイルに適応化された大量の音響モデルをパラレルデコーディングすることにより，より広い発話環境の音声を頑健に認識する音声認識システムの構築を行った．本システムでは，雑音の変動に頑健な音響特徴量としての DMFCC，予め種々の雑音環境へ適応化した HMM を用いて雑音 GMM の混合ウエイトから雑音適応 HMM を高速に生成する雑音適応手法，言い直し発話に頑健な音響モデル，複数の仮説を統合する手法を用いた．

その結果，10dB から 30dB の SNR で雑音を

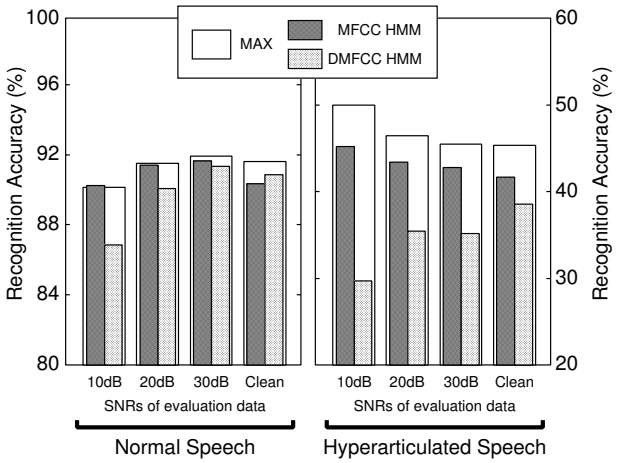

図 8: 仮説統合を行った場合の平均単語正解精度

重畳した通常発声の評価データに対して，平均90%以上の単語正解精度が得られた．また，言い直し発話などの発話スタイルの変動に対しても，通常発声用音響モデルのみを用いた場合よりも高い単語正解精度が得られた．

今後は，個々の音響モデルが高精度に認識可能な発話環境の限界についての検討及び，パラレルデコーディングを前提とした広い発話環境を効果的にカバーする音響モデルの構成方法について検討を行う予定である．

参考文献

[1] 中村, “実音響環境に頑健な音声認識を目指して,” 信学技報, EA2002-12, pp. 31–36, 2002.

[2] S.F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-27, pp. 113–120, 1979.

[3] H. Hermansky and N. Morgan, “RASTA Processing of Speech,” IEEE Trans. Speech and Audio Processing, vol. 2, no. 4, pp. 587–589, 1994.

[4] J. Chen, K.K. Paliwal, S. Nakamura, “Cepstrum Derived from Differentiated Power Spectrum for Robust Speech Recognition,” Speech Communication, vol. 41, no. 2-3, pp. 469–484, 2003.

[5] M. Gales and S. Young, “Robust Continuous Speech Recognition Using Parallel Model Combination,” IEEE Trans. on Speech and Audio Processing, vol. 4, no. 5, pp. 352–359, 1996.

[6] Y. Yamaguchi, S. Takahashi and S. Sagayama, “Fast Adaptation of Acoustic Models to Environmental Noise Using Jacobian Adaptation Algorithm,” Proc. Eurospeech 97, pp. 2051–2054, 1997.

[7] C.J. Leggetter and P.C. Woodland, “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models,” Computer Speech and Language, vol. 9, pp. 171–185, 1995.

[8] J.C. Junqua, “The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizer,” J. Acoustic. Soc. Amer., vol. 93, pp. 510–524, 1993.

[9] K. Yao, B.E. Shi, S. Nakamura and Z. Cao, “Residual Noise Compensation by a Sequential EM Algorithm for Robust Speech Recognition in Nonstationary Noise,” Proc. ICSLP2000, vol. 1, pp. 770–773, 2000.

[10] 奥田, 松井, 中村, “誤認識時の言い直し発話における発話スタイルの変動に頑健な音響モデル構築法,” 信学論, vol. J86-D-II, no. 1, pp. 42–51, 2003.

[11] 奥田, 河原, 中村, “ゆう度基準による分析周期・窓長の自動選択手法を用いた発話速度の補正と音響モデルの構築,” 信学論, vol. J86-DII, no. 2, pp. 204–211, 2003.

[12] 南條, 河原, “発話速度に依存したデコーディングと音響モデルの適応,” 信学技報，SP2001-103, 2001.

[13] M. Ostendorf, V. Digalakis and O. Kimball, “From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition,” IEEE Trans. Speech and Audio Proc., vol. 4, no. 5, pp 360–378, 1996.

[14] 伊田, 中村, “雑音 GMM の適応化と SN 比別マルチパスモデルを用いた HMM 合成による高速な雑音環境適応化,” 信学論, vol. J86-D-II, no. 2, pp. 195–203, 2003.

[15] K. Markov, T. Matsui, R. Gruhn, J. Zhang, S. Nakamura, “Noise and Channel Distortion Robust ASR System for DARPA SPINE2 Task,” IEICE Trans. Inf. & Syst., vol. E86-D, no. 3, 2003.

[16] T. Takezawa, T. Morimoto and Y. Sagisaka, “Speech and language databases for speech translation research in ATR,” Proc. EALREW, pp. 148–155, 1998.

[17] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto and S. Yamamoto, “Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Converstaions in the Real World,” Proc. LREC2002, pp. 147–152, 2002.

[18] H. Yamamoto and Y. Sagisaka, “Multi-class Composite N-gram Language Model Based on Connection Direction,” Proc. ICASSP, pp. 533–536, 1999.

[19] M. Ostendorf and H. Singer, “HMM Topology Design Using Maximum Likelihood Successive State Splitting,” Computer Speech and Language, vol. 11, no. 1, pp. 17–41, 1997.