2009 年度第 6 回情報処理学会東北支部研究会 (山形大学)
資料番号 09-6-A1-5

出力統合を用いた雑音環境下の音声認識の検討

○倉又 俊輔 † 加藤 正治 †† 小坂 哲夫 ††
† 山形大学工学部 , †† 山形大学大学院理工学研究科
〒992-8510 米沢市城南 4 丁目 3-16
Tel & FAX:0238-26-3365

平成 22 年 3 月 5 日 於 山形大学工学部

1　はじめに

音声認識は様々なシステムに用いることができる．しかし，その高い必要性があるにもかかわらず実用化されているとは言いがたい．実用化を妨げている大きな要因として雑音が挙げられる．実際の生活環境（実環境）は必ずといっていいほど雑音が存在する．現在の音声認識システムは雑音があるとその性能は大きく低下してしまう．そのため音声認識システムの実用化には，雑音に頑健な認識システムが必要不可欠である．これまでに雑音に関する研究が多くされてきたが，クリーンな環境と比べると雑音環境では認識率ははるかに劣ってしまうのが現状である．

雑音には準定常雑音と突発性雑音（ドアを閉めるときの音など）などの雑音があり，一般的には準定常雑音や雑音レベルが低い時には連続分布型 HMM(CHMM) が有効で，突発性雑音や雑音レベルが高い時には離散混合分布型 HMM(DMHMM) 有効であることが知られている．

また，複数の認識結果を統合することで認識性能を上げる方法があり，「ROVER による出力統合」 [1]，「単語グラフ統合」 [2] などが挙げられる．これらの方法は特徴が同じシステムを統合すると認識率は変化しないが，それぞれの誤り傾向が違う認識結果を組み合わせると認識率の向上が期待できる．

以上のことから，CHMM と DMHMM の認識結果を出力統合を用いて補完しあえば雑音に頑健なシステムができると考えられ，実際に CHMM，DMHMM それぞれを用いた認識結果より，互いを統合したシステムのほうが認識率が良いことが分かっている [3] [4]．そこで本研究では，システム統合について詳しく検討し， [3] [4] 以上の認識率の向上を目指す．加えて [3] [4] の研究では，突発性雑音下や複合雑音下 (準定常＋突発性) での検討は行われていない．そのため本研究では，突発性雑音下と複合雑音下についても準定常雑音下と同様に認識率が上がるのかを確認し，雑音により頑健なシステムを構築を目指す．

2　連続混合分布 HMM

音声認識では，一般的に音響モデルとして連続分布型 HMM が用いられており，出力確率分布として以下の式に示すような多次元混合正規分布を用いる．

$$b_i(x) = \sum_{m=1}^{M} \lambda_m \mathcal{N}(x|\mu_m, \Sigma_m) \tag{1}$$

ここで，$\mathcal{N}(x|\mu_m, \Sigma_m)$ は m 番目の正規分布，μ_m，Σ_m はその平均と共分散行列，λ_m は m 番目の分布の出現確率 (分岐確率) であり，

$$\sum_{m=1}^{M} \lambda_m = 1$$

の条件がある．また，P は特徴ベクトルの次元数を表す．

共分散行列として全共分散を用いると，パラメータ数が特徴ベクトルの次元数の 2 乗に比例して増加するため，一般には非対角要素を 0 とする対角共分散行列が用いられる．この場合正規分布は次元独立の無相関正規分布になるため，以下のようになる．

$$\mathcal{N}(x|\mu_m, \Sigma_m) = \prod_{p=1}^{P} \frac{1}{\sqrt{2\pi\sigma_{mp}^2}} \exp\left\{-\frac{(x_p - \mu_{mp})^2}{2\sigma_{mp}^2}\right\} \tag{2}$$

ここで，μ_{mp} と σ_{mp} は m 番目の分布の p 番目の次元の平均値と分散で，σ_{mp} は共分散行列 Σ_m の対角要素に一致する．

3　離散混合分布 HMM

現在の HMM の主流は混合連続型 HMM であると前述したが，雑音が重畳した音声認識では，混合分布でも対応できない分布になることも予想される．このような音声に対しては，任意の分布形状が表現できる離散分布型 HMM が有効ではないかと考えられる．離散分布型 HMM の場合，量子化サイズを小さくすると量子化歪みが大きくなり，逆にサイズを大きくすると学習データが不足し，充分にパラメータ推定ができないという問題がある．これに対し量子化サイズが小さくてすむ離散混合分布型 HMM が提案されている [5]．ここでは入力特徴ベクトルをサブベクトルに分割し量子化する方法を示す．

$q_s(\mathbf{o}_{st})$ をサブベクトル s における入力 $\mathbf{o}_{st}$ に対する量子点とすると，DMHMM の出力確率 $b_i(\mathbf{o}_t)$ は以下のようになる．

$$b_i(\mathbf{o}_t) \;=\; \sum_m w_{im} \prod_s \hat{p}_{sim}(q_s(\mathbf{o}_{st})) \tag{3}$$

但し，$\hat{p}_{sim}$ をサブベクトル s，状態 i，混合要素 m における離散確率，w_{im} を混合分布の重み係数 ($\sum_m w_{im} \;=\; 1.0$) とする．

離散出力確率分布の MAP 推定について説明する．通常の最尤推定では事前分布を定数とし，事前分布の影響を無視するが，MAP 推定では事前分布も考慮に入れたパラメータ推定を行う．この方法は CHMM における話者適応などに応用され，その実効性が示されている．

k をコードブックのインデックス，γ_{imt} を時刻 t で状態 i，混合要素 m に存在する確率とすると，離散出力確率の ML 推定値は以下のように求められる．

$$p_{sim}(k) = \frac{\sum_{t=1}^{T} \gamma_{imt} \delta(q_s(\mathbf{o}_{st}), k)}{\sum_{t=1}^{T} \gamma_{imt}} \quad (4)$$

このとき，離散出力確率の MAP 推定値 $\hat{p}_{sim}(k)$ は事前分布をディレクレ分布とした場合，

$$\hat{p}_{sim}(k) = \frac{\tau \cdot p^0_{sim}(k) + n_{im} \cdot p_{sim}(k)}{\tau + n_{im}} \quad (5)$$

となる．ここで τ は事前知識の確からしさに関する係数であるが，今回の実験では $\tau = 10.0$ と定めた．DMHMM に関しては出力分布の平均値だけでなく，混合係数，状態遷移確率についても MAP 推定が可能であるが，これらのパラメータについては，ML 推定により求めた．

3.1 DMHMM の尤度補償

継続時間の短い突発性雑音，例えば咳払い，ドアの開閉音などは，音声認識に悪影響を与えることが知られている．突発性雑音の対処法として，DMHMM の離散分布に閾値を設ける方法を用いた．

DMHMM において，上式 (3) の $\hat{p}_{sim}(q_s(\mathbf{o}_{st}))$ のいずれかのサブベクトルの確率が 0 またはそれに近い値になると，出力確率も 0 に極めて近い値になる．このように出力確率が 0 に近づくと，対数計算を行った場合，僅かな入力値の違いが大きな尤度差となって表れる．これは音声認識に悪影響を与える可能性がある．そこで離散確率に一律に閾値を設け，確率が閾値を下回った場合に閾値に置き換えるという処理を行う．閾値の設定は以下の通りに行う．式 (3) に示す出力確率を以下のように表現する．

$$b_i = \sum_m w_{im} \prod \hat{p}_{sim} \quad (6)$$

このとき，

$$\hat{p}_{sim} = \begin{cases} \hat{p}(q_s(\mathbf{o}_{st}) & if \quad \hat{p}_{sim}(q_s(o_{st}) \geq dth \\ dth & otherwise \end{cases} \quad (7)$$

4 ROVER

ROVER システムは，複数の音声認識システムの出力を統合するためのシステムである [1]．誤り傾向がそれぞれ違う音声認識システムの出力を統合することで，各システムを単独で使用した場合では誤認識した箇所を，他の正しい認識結果を採用することで認識率を良くする．ROVER システムは対応付けモジュール (Alignment Module)，投票モジュール (Voting Module) の 2 つのモジュールから構成されている．図 1 に ROVER システムの概要を示す．

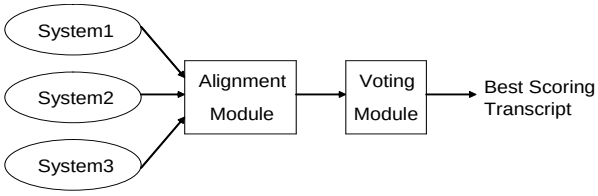

Fig. 1 ROVER システムの概要

ROVER システムは，まず複数の ASR システムから得られた単語遷移ネットワーク (WTN:Word Transition Network) を合成し，合成単語遷移ネットワークを作成する．ここで，各単語遷移ネットワークの対応付けを行うのが対応付けモジュールである．

はじめに, ASR システムの仮説の出力から WTN を作る．ここで得られた初期の WTN は線形である．まず，WTN の 1 番目を基本 WTN(WTN-BASE) とし，次に 2 番目の WTN(WTN-2) を DP Alignment よって基本 WTN に対応付けし，基本 WTN を拡張する．この WTN は WTN-2 から WTN-BASE へアークをコピーすることで得られる．WTN-3 を WTN-BASE' へ結合させる場合も同様に行う．最終的に得られる WTN(WTN-BASE") を図 2 に示す．最後に対応付けモジュールによって作成された合成 WTN に対して，投票モジュールを用いることで最良のスコアを持つ単語列を見つけ出す．

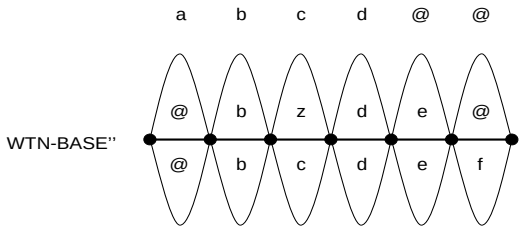

Fig. 2 最終的な WTN

5 単語グラフ統合 (WGC)

単語グラフ統合とは，単語グラフを複数統合することで認識率を上げることを目的としている．単語グラフとは単語間のつながりを表すグラフのことで，各エッジに単語，始

端フレーム，終端フレーム，スコア (音響尤度) の情報がある．単語グラフの統合は，同一単語の始端と終端フレームが一致している場合のみ統合する．以下に単語グラフ統合の手順を示す.

1. 入力音声を音響モデル AM_1〜AM_n，bigram 言語モデル LM でデコードし単語グラフ WG_1〜WG_n を得る．
2. 単語グラフ WG_1〜WG_n を統合し単語グラフ WG_C を得る．
3. 単語グラフ WG_C を音響モデル AM_1〜AM_n，言語モデル LM でリスコアし認識結果を得る．

図 3 に単語グラフの統合の手順を示す.

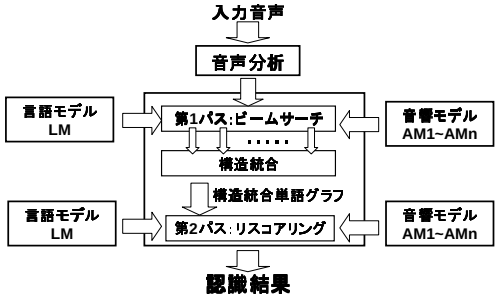

Fig. 3　単語グラフ統合の手順

5.1　WGC の方法

N 個のシステムがあると仮定する。それらから出力された単語グラフを $W_1, W_2, \cdots W_N$ と表す．ここで，システム n は W_n のように表され，そのエッジは，q_n とする．エッジには，始端フレーム時間 t_{start}，終端フレーム時間 t_{end}，t_{start} から t_{end} までに発声された単語 w_i の情報が含まれ，$q_n = [w_i; t_{start}, t_{end}]$ と表現することができる．そのため，W_n は q_n の集合であるといえる．ここで，例としてシステム 1, システム 2 からの出力，単語グラフ W_1 および W_2 があったとき，$q_1 = [w_i; t_{start}, t_{end}]$，$q_2 = [w_j; \tau_{start}, \tau_{end}]$

今、2 つの単語グラフの構成を定義することができるとして

$$W_1 + W_2 \equiv \{q_1 = q_2\} \cup \{q_1 | q_1 \notin W_2\} \cup \{q_2 | q_2 \notin W_1\} \quad (8)$$

ここで $q_1 = q_2$ とは，始端フレーム，終端フレーム，単語が全て一致したときすなわち，

$$q_1 = q_2 \; iff \; w_i = w_j, t_{start} = \tau_{start}, t_{end} = \tau_{end} \quad (9)$$

である．$W_1 + W_2$ は，W_1 または W_2 のいずれかのエッジをすべて含んでいる．式 (8) から、すべてのシステム 1〜N の統合単語グラフ W を表すことができる．

$$W = W_1 + W_2 + \cdots + W_N = \sum_{i=1}^{N} W_i \quad (10)$$

5.2　スコア設定方法

統合処理において，スコア設定方法は対応エッジの平均スコア，対応エッジの平均スコアの重み付けの 2 種類を検討した．N 個のシステムがあり，それら単語グラフには M 個のエッジがあったときに，システム n のエッジ $q(m)$ のスコアを P_m^n で表すとすると，対応エッジの平均スコアの場合，統合単語グラフのエッジ $q(m)$ のスコア $P'_{q(m)}$ は以下の式で表される．

$$P'_{q(m)} = \frac{1}{N} \sum_{k=1}^{N} P^k_{q(m)} \quad (11)$$

また対応エッジの平均スコアの重み付けは以下のように行う．

$$P'_{q(m)} = (1 - \alpha) P^1_{q(m)} + \alpha P^2_{q(m)} \quad (12)$$

ここでの α は重みのことであり $0 \leq \alpha \leq 1$ という条件がある．$P^1_{q(m)}$ は CHMM のスコア，$P^2_{q(m)}$ は DMHMM のスコアである．

6　実験条件

音声データは日本音響学会の新聞記事読み上げコーパスを用いる．トレーニングセットとして 102 名の男性が発声した新聞記事読み上げ文＋音素バランス文，計 15,732 文を使用した．マルチコンディション学習 [6] の場合はこのトレーニングセットを 20 分割し，4 種類の雑音＋ 5 種類の SNR(5,10,15,20, ∞ dB) の計 20 種の組み合わせで雑音を重畳した．マルチコンディション学習とは，あらかじめ音響モデルに何種類かの雑音データを数種類の SNR で重畳した音声データを学習させることで，雑音下での音声認識の性能向上が可能である．重畳する雑音は電子協騒音データベースから，1500cc クラス自動車内，展示会場 (通路)，人ごみ，列車 (在来線) の 4 種を用いた．初期 CHMM(クリーンコンディション CHMM:CC CHMM) をクリーンデータで学習して求めた後，式を用いて DMHMM に変換し初期 DMHMM(クリーンコンディション DMHMM:CC DMHMM) を得る．雑音重畳音響モデルを作成するため，マルチコンディション学習データで CC CHMM を学習し，マルチコンディション CHMM(MC CHMM) を得る．この MC CHMM を変換して得た DMHMM を MAP 推定でもう 1 度学習したものをマルチコンディション

DMHMM(MC DMHMM) とする．また，DMHMM の尤度補償値は 0.00025 とした．

テストセットは準定常雑音下の実験，突発性雑音下の実験，複合雑音下の実験，それぞれ異なる．各テストセットごと，10 名の男性話者が発声した 100 文に対し雑音を重畳したものを用いた．重畳した雑音は準定常雑音は駅，工場，幹線道路，エレベータホールの 4 種，突発性雑音は bank，claps1，whistle3 の 3 種である．複合雑音は定常×突発の 12 種である．また，SNR は準定常雑音で∞，20，15，10，5dB の 5 種，突発性雑音は 0dB，複合雑音では準定常雑音の SNR は 10dB，突発性雑音は 0dB である．学習に用いていない雑音で評価を行うためテストセットは未知の雑音であり，雑音に対してオープンな実験であるといえる．雑音を重畳する際の SN 比は，準定常雑音は音声区間の平均パワーと雑音区間の平均パワーから計算するのに対し，突発性雑音の SN 比は，音声区間の平均パワーと雑音区間の最高パワーから計算した．これは突発性雑音は，ある一点において強いパワーを持った雑音であるため平均すると雑音のパワーが低下するためである．

音声分析条件を表 1 に示す．音響モデルは各 triphone3〜6 状態，総状態数 2000，1 状態あたり 16 混合の CHMM および DMHMM を使用した．言語モデルは語彙 5k で毎日新聞の 45 ヶ月分を使用して作成した．認識システムは第 1 パスで triphone および単語 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 デコーダを用いた．ROVER で統合する場合，第 2 パスにおいて言語重みが 12〜20 まで +2 毎に設定し 5 種類，挿入ペナルティは-40〜0 まで +10 毎に設定し 5 種類，計 5 × 5=25 種類の出力を各音響モデルごとに得る．以下の実験では CHMM,DMHMM の 2 種の出力を統合するため計 50 種の出力を統合する．

Table. 1　音声分析条件

標本化 / 量子化	16kHz / 16bit
フレーム長 / 分析周期	32msec / 8msec
分析窓	ハミング窓
高域強調	$1\text{-}0.97z^{-1}$
特徴ベクトル	1〜12 次の MFCC と対数パワー 及び 1 次と 2 次の回帰係数（計 39 次元）
正規化	発話毎のケプストラムズム平均正規化

7　実験結果

7.1　準定常雑音下の認識実験

MC CHMM,MC DMHMM の 2 種統合の結果を以下に示す．表中の構造統合後とは CHMM，DMHMM の 2 つの単語グラフの構造を統合するがスコア統合は行わない場合の認識結果である．それに対し WGC はスコア統合を行った後の認識結果である．また，単語グラフ統合の比較として ROVER を行い，その認識結果を示す．

表中の WER(word error rate) とは単語誤り率のことであり認識結果を評価する指標となる．

Table. 2　準定常雑音環境下の WER(%)

	統合前		構造統合後	
音響 ＼ モデル SNR	MC CHMM	MC DMHMM	MC CHMM	MC DMHMM
∞	6.21	5.69	5.80	5.38
20	7.79	8.10	7.87	7.87
15	11.52	11.59	11.49	11.34
10	24.43	23.50	22.59	21.64
5	52.23	49.79	50.67	48.50
平均	22.94	**22.21**	22.13	21.34
	スコア統合後		ROVER	
音響 ＼ モデル SNR	WGC 重みなし	WGC $\alpha = 0.9$		
∞	5.80	5.59	5.69	
20	7.89	7.97	7.84	
15	11.36	11.44	11.36	
10	21.87	21.30	22.52	
5	48.86	47.90	49.15	
平均	21.51	**21.18**	21.71	

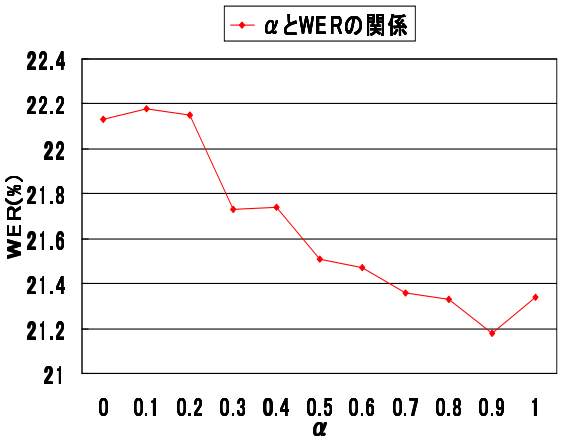

Fig. 4　α と WER の関係

準定常雑音下での音声認識の場合, 統合前の DMHMM の認識率 22.21% に対しスコア統合 ($\alpha = 0.9$) のとき 21.18% となり，WER は 1.03 ポイント改善した．別の統合手法である ROVER と比較しても 0.53 ポイント改善した．よって単語グラフ統合は準定常雑音下において ROVER よりも有効であることが分かった．また，α と WER の関係図を表 4 に示す．これを見ると DMHMM に重みを多く与えたほうが認識結果は良いという結果になった．

7.2 突発性雑音下の認識実験

CC CHMM,CC DMHMM の 2 種統合の結果を以下に示す．突発性雑音は瞬間的な雑音であり，それ以外の部分はクリーンな音声のため CC モデルを使用した．

Table. 3 突発性雑音下の WER(%)

	統合前		構造統合後	
音響＼モデル 雑音	CC CHMM	CC DMHMM	CC CHMM	CC DMHMM
bank	10.46	8.49	8.59	8.28
claps1	12.53	9.21	12.73	10.04
whistle3	37.89	26.71	32.19	26.40
平均	20.29	**14.80**	17.84	14.91
	スコア統合後			
音響＼モデル 雑音	WGC 重みなし	WGC $\alpha = 0.9$		
bank	8.59	8.59		
claps1	11.08	9.63		
whistle3	27.12	26.19		
平均	15.60	**14.80**		

突発性雑音下での音声認識の場合，統合前の CC DMHMM の認識率 14.80% に対し，単語グラフ統合後の最良の値を示した WGC($\alpha = 0.9$) で 14.80% となり同程度の認識結果を得ることができた．認識結果が良くならなかったのは, 突発性雑音ということもあり CHMM と DMHMM の認識率の差がありすぎた点が挙げられる．これは構造統合後 DMHMM と統合前の DMHMM を比較すると認識率が低下していることから推測できる．また，表 5 より，準定常雑音下の実験と同様に DMHMM に重みを多く与えたほうが認識結果は良いという結果になった．

7.3 複合雑音下の認識実験

MC CHMM,MC DMHMM の 2 種統合の結果を以下に示す．表中の雑音の部分は突発性×定常 (4 種) の平均値を示している．

Table. 4 複合雑音下の WER(%)

	統合前		構造統合後	
音響＼モデル 雑音	MC CHMM	MC DMHMM	MC CHMM	MC DMHMM
bank ＋定常	31.42	30.38	30.56	28.91
claps1 ＋定常	34.14	32.92	33.90	31.37
whistle3 ＋定常	52.56	54.19	50.88	50.52
平均	39.37	**39.16**	38.45	36.93
	スコア統合後			
音響＼モデル 雑音	WGC 重みなし	WGC $\alpha = 0.8$		
bank ＋定常	29.19	28.73		
claps1 ＋定常	32.14	31.55		
whistle3 ＋定常	49.51	50.16		
平均	36.95	**36.81**		

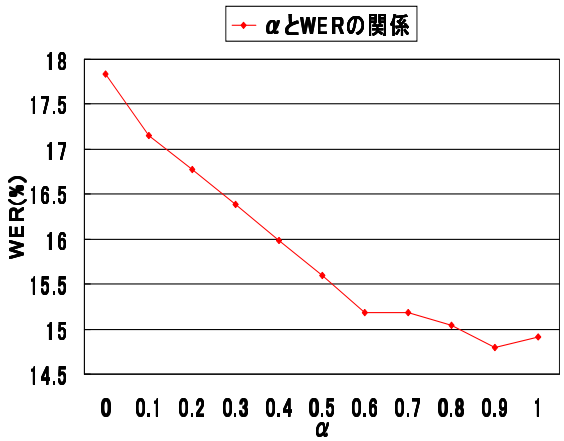

Fig. 5 α と WER の関係

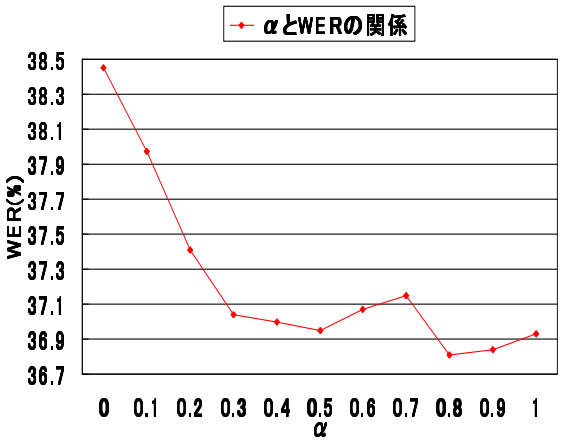

Fig. 6 α と WER の関係

複合雑音下での音声認識の場合, 統合前の DMHMM の認識率 39.16% に対しスコア統合 ($\alpha = 0.8$) のとき 36.81% となり，WER は 2.35 ポイント改善した．複合雑音下では単語グラフ統合は有効であることが分かった．また表 6 より，DMHMM に重みを多く与えたほうが認識結果は良いという結果になった．

7.4 実験結果のまとめ

本研究では準定常雑音の実験に加え突発性雑音下，複合雑音下での認識実験を行い単語グラフ統合の有効性を検討した．実験結果をまとめたものを表 5 に示す．この表を見ると，3 つの雑音環境下全てに対し単語グラフ統合は同等かそれ以上の認識結果を示したことがわかる．

Table. 5 実験結果のまとめ WER(%)

音響モデル 雑音＼	CHMM	DMHMM
準定常雑音	22.94	**22.21**
突発性雑音	20.29	**14.80**
複合雑音	39.37	**39.16**
音響モデル 雑音＼	構造統合 CHMM	構造統合 DMHMM
準定常雑音	22.13	21.34
突発性雑音	17.84	14.91
複合雑音	38.45	36.93
音響モデル 雑音＼	スコア統合 (重みなし)	スコア統合 (重みあり)
準定常雑音	21.51	**21.18**($\alpha = 0.9$)
突発性雑音	15.60	**14.80**($\alpha = 0.9$)
複合雑音	36.95	**36.81**($\alpha = 0.8$)

8 まとめ

本研究では出力統合を用いて準定常雑音下，突発性雑音下，複合雑音下での音声認識を行い，雑音環境での認識率の向上を目指した．

準定常雑音下の認識実験では，統合前と比べ単語グラフ統合を行ったときの認識率は良くなった．また，異なった統合法である ROVER と比較しても認識率は良く，統合を行う場合 ROVER よりも単語グラフ統合のほうが良いことが分かった．突発性雑音下の認識実験では，統合前の認識率と比べ単語グラフ統合後の認識率は同程度であった．複合雑音下の認識実験でも，統合前よりも統合後のほうが認識率は良くなった．以上のことから雑音環境下において単語グラフ統合は有効であり，より雑音に頑健なシステムを構築できると考えられる．また，スコア選択法で重みをつけるときは DMHMM のほうに多く重みを与えることで性能が向上することが分かった．

今後の課題としては，今回比較に用いた ROVER は準定常雑音下の実験のみしか行なっていないので，突発性雑音下，複合雑音下でも ROVER を行い結果を比較する点が挙げられる．また，今回行った単語グラフ統合と ROVER を組み合わせることによる性能向上も研究課題である．

参考文献

[1] J.GFiscus: ”A post-processing system to yield reduced word error rates :Reduction(ROVER)” Proc.of IEEE Workshop on Automatic Speech Recognition and Understanding, pp.347-354 (1979)

[2] I-Fan Chen, Lin-Shan Lee(National Taiwan Univ.): ”A New Framework for System Combination Based on Integrated Hypothesis Space” Proc. of Interspeech2006, pp.533-536 (2006)

[3] 斎藤陽，加藤正治，小坂哲夫: ”離散分布 HMM と連続分布 HMM の出力統合による雑音下音声認識の検討”, 日本音響学会講演論文集, 3-Q-1 pp. 149-152 (2008.9)

[4] 橋本迪明: ”出力統合を用いた音楽雑音下の音声認識の研究”, 山形大学卒業論文 (2009)

[5] T.Kosaka, M.katoh, M.Kohda: ”Robust Speech Recognition Using Discrete-Mixture HMMs” IEICE Transactions Inf. And Syst., Vol.E88-D, No.12,pp.2811-2818 (2005)

[6] D.Pearce and H.-G. Hirsch: ”The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions” Proc. ICSLP2000, vol.4, pp.29-32 (2000)