TR-IT-300

連続音声認識研究用音響モデル (ResearchJ V7)

柘植 覚 Satoru Tsuge
内藤 正樹 Masaki Naito
シンガー・ハラルド Harald Singer
高野 優 Masaru Takano
松井 知子 Tomoko Matsui

1999 年 3 月 31 日

自然発話音声データベース (旅行対話) の整備の終了に伴い、先に正式リリースした音響モデル (TR-IT-0266) と比較して、約 2 倍程度の話者の音声が音響モデルの学習に使用可能となった。そのため、本報告では、学習話者数の増加に応じて、音響モデルのガウス分布の最適な混合数及び、状態数の比較検討を行った。また、朗読発話音声により作成した音響モデルの自然発話音声への発話様式適応に関する検討を行った。
比較実験の結果、自然発話音声のみを用い学習した、ケプストラム平均減算法による正規化を行った MFCC を特徴量とする 1400 状態 5 混合の男性モデル、1400 状態 15 混合の女性モデルを自然発話音声認識用音響モデルとしてリリースをする。

目次

1 はじめに

現在、音声翻訳通信研究所にて、音声認識、翻訳、合成を統合した音声翻訳システムが稼働している。そのシステムの精度を向上を図る上で、音声認識部の認識性能の向上、特に、音響モデルの性能向上が重要な課題の一つであると考えられる。本報告では、自然発話データベース (旅行対話) の整備により利用可能となった、多量の学習データを用い、より高精度な音響モデルをリリースするため、音響モデルの状態数、混合数の検討及び、朗読発話音声音声用音響モデルの自然発話音声認識への発話様式適応について検討を行った。

2 音声データベース

2.1 学習データ

音響モデルの学習サンプルとして、ATR の自然発話音声データベース (SDB)[1] より、男女合計 407 話者の自然発話音声データ (学習セット T_M_0167, T_F_0240) を用いた。この音声データの詳細を表 1に示す。比較のため、同表に、前リリース時まで音響モデルの学習に用いていた学習話者 230 名 (本稿で用いる学習話者 407 名のサブセット) のデータ量も示す。なお、表中に示した発話時間は、トランスクリプションファイルを用いて算出したものである。

図 1に学習話者の音声波形ファイルを例を示す。本稿では、これらの音声波形ファイル (無音も含む) から、トポロジー学習、ラベル学習には図 1に示す **A** の区間 (無音部は含まない) を、連結学習には、**A** に 30msec の無音を前後に付加した **B** の区間を用いた。各音響モデルの時間方向への状態分割は、最大 4 状態であるため、音響モデルのトポロジー学習、ラベル学習時には、自動ラベリングにより 40msec 以上となったデータを用いている[1]。連結学習時には、連結学習が可能なフレーム数をもつ全学習データを用い学習を行っている。このような条件のもとで、学習データベースから実際に音響モデルの各学習に用いられたフレーム数を表 2に示す。表中のトポロジー学習、連結学習に用いたフレーム数は、各音響モデルより抜粋した。

2.2 テストデータ

テストデータとして、現在までにリリースが行われているテクニカルレポート [2][3][4] と同様に、学習データと同一のデータベース内の、学習に用いていない男性 17 名と女性 25 名の合計 42 名からなる 551 発声 (テストセット S1, S2, S4)、総音素数約 2 万 (約 40 分) を用いた。

音声認識性能の評価は、言語体系が TDMT 体系、認識可能語彙数 13000 語のもとで、接続方向を考慮した複合 n-gram を言語モデル [5] として用いた単語認識率により行った。

[1] HMM の最大状態数が 4、フレームシフトが 10msec であるため、40msec 以上のデータとなる。

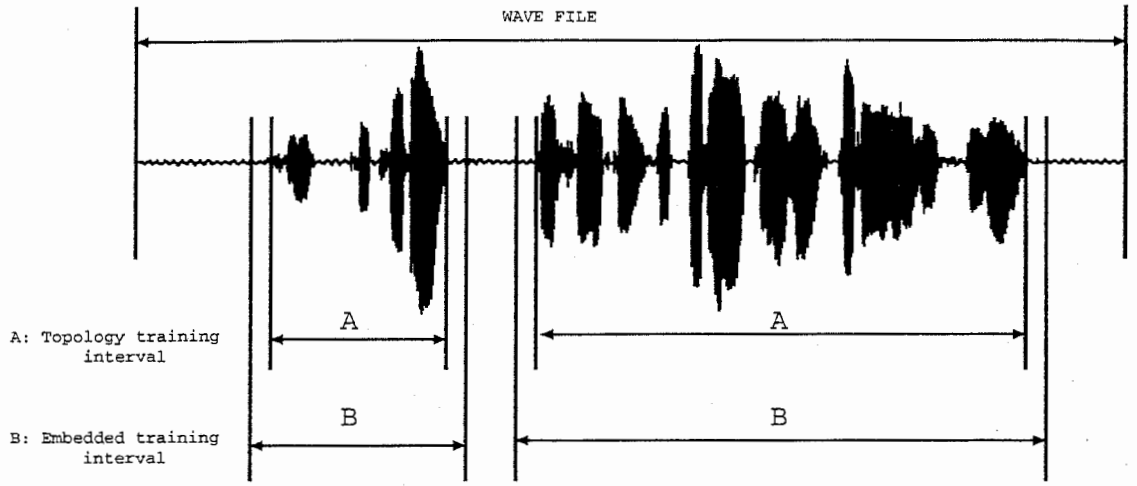

図 1: 各学習に用いる音声データ

表 1: 学習データの詳細 (発声データ)

話者数	(男性 / 女性)	発話数	発話時間 (分)
230	(100/130)	2962 (1245/1717)	135 (55/80)
407	(167/240)	6432 (2583/3849)	299 (118/181)

表 2: 学習データの詳細 (フレーム数 (男性 / 女性))

話者数	ラベル	連結学習
230	307731/466051	343102/504653
407	666085/1046256	740478/1135250

3 音響モデルの作成

音響分析および音響モデルの作成は、ATR-SPREC version r06r01 を用い、DEC 社の Alpha-Station 500/500 上で行った。

3.1 音響分析

表 3に、比較実験に用いた音響分析条件を示す。

特徴パラメータは、TR-IT-0266[4] の検討で高い認識性能が得られた、発話毎に音声区間のケプストラムの平均を計算し、ファイル全体から減算を行う CMS (Cepstrum Mean Subtraction) による正規化を行った MFCC (Mel-Frequency Cepstrum Coefficient) を用いた。対数パワー項も、同様に発話毎に平均を計算し減算を行い正規化を行った。また、一次回帰係数 (ΔMFCC) の計算は、100msec (前後 4 フレーム) の三角窓を用い計算を行った [6]。音響分析のためのコンフィグレーションファイルは、TR-IT-0266 を参考にされたい。

表 3: 音響分析条件

プリエンファシス	0.98
フレーム周期	10 msec
フレーム長	20 msec
分析窓	ハミング窓
フィルタバンク次数	16
MFCC 次数	12
特徴ベクトル	MFCC, 対数パワーと各一次回帰係数 (合計 26 次元) (発話毎のケプストラム平均減算法による正規化)

3.2 音素ラベルファイル

前リリースまでの音素体系では、音素 /f/ は全て音素 /h/ にマッピングを行っていたが、本稿では、音素 /f/ と /h/ を区別した。また、音素“ふ”は、現在トランスクリプションファイルの音素表記に使用されている /h,u/ ではなく /f,u/ とした。音素の区別による認識性能の差は、ほとんど見られない。しかし、このような分割は、新たな外来語等の認識に有効であると考えられる。

音素 /f/ と /h/ の分割が行われたため、新たに音素ラベルファイルの作成を行った。新たに作成を行った音素ラベルファイルは、TR-IT-0266 で検討を行った時の条件と同様に、MFCC を特徴パラメータとした性別依存 800 状態 5 混合モデルを用い、Viterbi セグメンテーションを行った結果を用いた。

3.3 音響モデル作成方法

本稿で用いる音響トポロジー (音響モデルの構造) は、尤度最大化基準逐次状態分割 (ML-SSS) アルゴリズム [7] により、27 状態の初期モデルから各実験に適する状態数まで分割を行ったものを用いた。新たに音素 /f/ を加えたことによる、初期状態の増加は行わず、初期状態では、音素 /f/ と音素 /h/ は区別をせず同一経路の HMM で表現することにした。

分割後のトポロジーに対し、Forward-Backward アルゴリズムを用いてパラメータ学習を行う。その後、学習を行った音響モデルに 1 状態のポーズモデルを結合し、Viterbi 学習にて連結学習を行う。最後に、1 状態のポーズモデルを 3 状態のポーズモデルに変更して認識実験用の音響モデルとする。

この音響モデルは、初期状態で時間方向への状態数は 3 状態であり、時間方向の最大経路長は、4 状態とした。また、無音モデルは、トランスクリプションファイル (*.TRS) の時間情報に基づいて切り出した無音区間を 3 状態 10 混合の HMM で学習したものを用いた[2]。

本稿の音声認識実験では、性別依存音響モデルを並列に使用をし、最尤選択により音響モデルを決定し、認識を行う。この場合、認識ソフトウェアの設計上、並列に使用する音響モデルのトポロジーは共通でなくてはならな

[2]音素モデル作成の連結学習時には、1 状態 10 混合の無音モデルを用いている。

い。このため、音響モデルのトポロジーは男女をあわせた全学習話者 (407 名) で作成し、その後、共通のトポロジーをもとに男女別の性別依存モデルを学習した (図 2)。

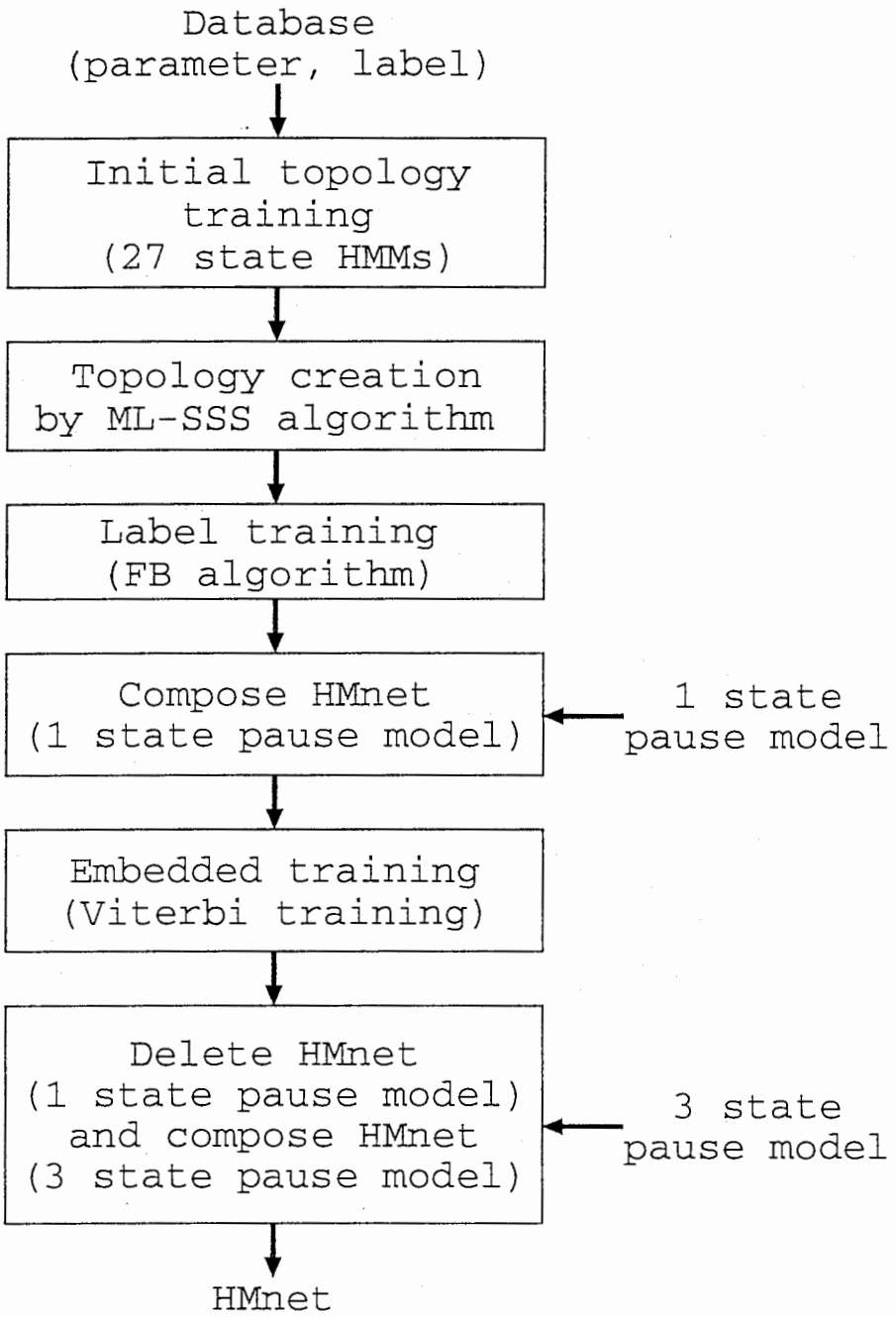

図 2: 音響モデル作成手順

4 音声認識実験

本節で述べる音声認識実験は、ATR-SPREC version r06r01 を用い、DEC 社の Alpha-Station 500/500 上で行った。

4.1 認識条件

本稿で行った音声認識実験は、TR-IT-0284[8] で使用された認識実験条件と同様の実験条件下で行った。しかし、本稿では認識語彙数の比較は行わず、語彙数の多い約 13000 語の辞書のみを用いて、認識実験を行った。認識実験のコンフィグレーションファイルについては、付録 A を参考にされたい。

4.2 学習話者の増加に対する認識性能の比較

本節において、音響モデルの学習データ量による認識性能の比較を行う。

(4.2.1) 前リリース音響モデル (TR-IT-0284) との比較

TR-IT-0284 と同一条件で、学習話者数 407 名の発声を用い音響モデル (800 状態 5 混合) の学習し、その音響モデルを用い認識実験を行った。結果を表 4に示す。比較対象として、TR-IT-0284 で示されている、学習話者数 230 名の音声で学習を行った同状態、同混合数の音響モデルの認識性能を同表に示した。表中に示した C/U は、認識に必要であった認識時間 (CPU time) を実際の発声時間で除した値である。また、err utt は、ワークエリア不足、ビーム幅不足などに起因して認識結果が得られなかった発声数である。

表 4: 学習データ量による単語認識率の比較 (%)

	単語認識率	err utt	C/U
学習話者数	male/female/both	male/female/both	
230	82.84/86.72/85.29	(0/0/0)	3.24
407	85.54/87.46/85.65	(1/0/1)	3.42

表 4より、学習話者を増加による認識性能の向上はほとんど見られなかった。これは、比較実験を行った音響モデルの状態数、混合数や、言語重み等の認識時の設定等が、学習話者 230 名の音声で学習を行った音響モデルに最適化した値であるため、認識性能の差があまり見られなかったと考えられる。

(4.2.2) 学習話者 407 名に適する音響モデルの状態数、混合数の検討

先の実験結果において、認識性能に差があまり見られなかったのは、音響モデルの状態数、混合数や、認識時の設定等が、学習話者 230 名の音声で学習された音響モデルに最適化した値であるためと考えらる。そこで、本節では、学習話者 407 名の音響モデルに適する認識時の設定、状態数、混合数の検討を行う。はじめに、音響モデルの混合数を 5 と固定し、適切な状態数と認識時のビーム幅について検討を行った結果を表 5に示す。結果に示す C/U は、認識に必要であった認識時間 (CPU time) を実際の発声時間で除した値である。また、err utt は、ワークエリア不足、ビーム幅不足などに起因して認識結果が得られなかった発声数である。

表 5: 学習データ量による単語認識率の比較 (単語認識率 % (C/U,err utt))

	ビーム幅			
状態数	110	115	120	130
800	85.65 (3.42,1)			
1000	85.63 (3.31,0)	85.99 (4.08,3)	86.40 (4.78,8)	87.95 (7.03,30)
1400	86.15 (2.92,0)	86.15 (3.60,0)	86.67 (4.37,2)	87.74 (6.22,18)
2000	85.59 (2.68,0)	85.85 (3.19,0)	86.12 (3.85,1)	86.88 (5.40,8)
2400	85.85 (2.46,0)			
3000	84.85 (2.37,0)			

表 5より、認識時の使用メモリの制限下で、認識セット 42 話者の全発話の認識が可能である条件のもと (err utt が 0 の条件下) では、ビーム幅は、 V6 の認識実験と同様に 110 が適切であると考えられる。

次に、認識時のビーム幅を 110 に固定をし、適切な状態数、混合数の検討を行う。今回音響モデルの学習に用いた音声データは、性別により、学習話者数、発声時間が異なる。このため、男性モデルと女性モデルでは、適切な状態数、混合数が異なると予想される。よって、本比較実験では、音響モデルの並列使用による最尤選択は行わず、認識時に男性話者に対しては男性モデルを、女性話者に対しては女性モデルを単独で用い、認識実験を行った。実験結果を表 6に示す。

表 6: 性別依存音響モデルの状態数・混合数の検討 (単語認識率 % (err utt))

男性	状態数			
混合数	800	1000	1400	2000
5	83.47 (1)	83.92 (0)	**85.22 (0)**	84.08 (0)
10	84.06 (2)	84.95 (1)	**85.11 (0)**	83.43 (0)
15	84.38 (2)	82.62 (0)	81.27 (0)	81.10 (0)
20	83.87 (0)	84.14 (0)	83.65 (0)	80.62 (0)
女性	状態数			
混合数	800	1000	1400	2000
5	87.05 (0)	86.70 (0)	86.57 (0)	86.89 (0)
10	87.62 (0)	87.37 (0)	87.21 (0)	86.35 (0)
15	87.46 (0)	87.46 (0)	**87.40 (0)**	85.68 (0)
20	87.50 (0)	**87.72 (0)**	86.73 (0)	84.57 (0)

表 6より、男性モデルは、1400 状態 5 混合、及び 1400 状態 10 混合が、女性モデルは、1000 状態 20 混合、及び 1400 状態 15 混合が高い認識性能を示しており、本稿で用いた学習データ量に対して、適切な混合数、状態数であると考えられる。表 6の認識実験では、男女別の状態数、混合数の検討を行うため、性別既知の条件で音響モデルを選択し、認識を行っていた。しかし、この比較検討により、各性別依存モデルに適した状態数、混合数が得られたため、これらの音響モデルを並列に用い、音響モデルを最尤選択にて決定し、認識を行った。実験結果を表 7に示す。

表 7の結果より、表 6の検討で高い認識性能を示した状態数、混合数をもつ男女モデルの中から、特に、 1400

表 7: 学習データ量による単語認識率の比較 (%)

	男性モデル	
女性モデル	1400 状態 5 混合	1400 状態 10 混合
1400 状態 15 混合	**85.22/87.69/86.77**	85.22/87.59/86.71
	1000 状態 5 混合	1000 状態 10 混合
1000 状態 20 混合	82.89/88.01/86.11	83.74/87.72/86.26

状態 5 混合の男性モデルと 1400 状態 15 混合の女性モデルとの組合せが最も高い認識性能を示した。最も高い認識性能を示した結果の詳細を付録 B に示す。

ここで、男性モデルと女性モデルの混合数に差が生じたのは、女性モデルの学習に用いるデータ量が多いため、混合数の増加による精密なモデル化が可能であったものと推測される。

4.3 自然発話への発話様式適応実験

学習データ量の増加により、認識性能が向上することを前節にて示した。そこで、より多くの音声データを利用可能な朗読発話音声 (音素バランス文) を用い学習した音響モデルを初期モデルとして、自然発話データベースによる適応を行う、朗読発話音響モデルから自然発話発声への発話様式適応の検討を行った。また、朗読発話音声データベースをもちいることにより、性別間の発声コンテキストの偏り少なくなり、トポロジー作成における性別方向への分割を抑制することができると考えられる。前節の検討により、自然発話への発話様式の適応の実験には、男性 1400 状態 5 混合、女性 1400 状態 15 混合の音響モデルを用いた。

(4.3.1) 朗読発話トポロジーを用いた自然発話音響モデル

はじめに、認識実験用の音響モデルとして、朗読発話でトポロジーを作成し、得られたトポロジーをもとに、自然発話音声を用いて、パラメータ学習 (ラベル、連結学習) を行ったものを用いた。音声認識実験結果を表 8に示す。比較のため、自然発話のみから作成した音響モデルを用いた認識実験結果を “自然発話音響モデル” として同表に示す。自然発話音声を用いて音響モデルの状態分割を行った場合、学習サンプル中に含まれる音素環境が性別により偏っている可能性があるため、男女方向へ分割が行われ、音響モデルの精度を低くしている可能性があると考えられた。そのため、音素のバランスがとれた朗読発話音声を用いてトポロジーを作成した音響モデルを用い、実験を行った。

表 8: 学習データ量による単語認識率の比較 (%)

学習話者数	male/female/both
自然発話音響モデル	85.22/87.69/86.77
朗読発話トポロジーモデル	82.67/87.02/85.41

認識実験の結果から、はじめに考えた男女方向への分割による悪影響より、むしろ自然発話サンプルにより分割が有効に働いていると思われる。

また、表 8より、朗読発話から作成を行ったトポロジーを用いた音響モデルの認識性能は、自然発話のみの認識性能より低くなることがわかった。これは、音響モデルのトポロジー学習とパラメータ学習のデータベースが異なるために、パラメータ学習にて、未学習の音素系列が生じ、認識性能が劣化したと考えられる。

(4.3.2) 朗読発話音響モデルの発話様式適応

ここでは、学習用自然発話データベースにおいて出現回数の少ない音素へ対する補完の方法の一つとして、朗読発話発声にて作成を行った音響モデルからMAP-VFS[9]により、自然発話への発話様式適応を行った。適応には、2.1節に示した学習データを全て用いた。適応のために設定を行った各パラメータを表9に示す。パラメータの適応は、平均と分散のみに行った。各パラメータについての詳細は、SPRECのマニュアルを参考にされたい。また、各パラメータの変更実験は、付録Cに示す。

適応後の音響モデルを用い、実験結果を表10に示す。参考のために、適応を前の朗読発話音響モデルを用いた認識結果(朗読発話モデル)、前節の自然発話のみで作成を行った音響モデルを用いた認識結果(自然発話モデル)を同表に示す。

表9: 朗読発話音響モデルの自然発話適応のための適応パラメータ

パラメータ	
number of BW Iterations	10
k neighborhood number	150
smoothing rate	3.0
smoothing type	Gauss
smoothing control	20
training type	viterbi
estimation mode	map
tau of MAP estimation	4

表10: 朗読発話音響モデルの自然発話適応のための適応パラメータ

音響モデル	認識結果 (male/female/both) (err utt)
朗読発話モデル	81.38/86.22/84.43 (0)
自然発話モデル	85.22/87.69/86.77 (0)
発話様式適応モデル	84.95/89.47/87.83 (3)

表10より、発話様式適応を行ったモデルは、err uttを増加させる。認識結果が得られない発声(err utt)は、概ね認識可能である音響モデルを用い認識して場合においても、認識性能が低い。そのため、認識不能会話数を考慮すると発話様式の適応を行っても、自然発話モデルと同等の認識性能しか得られていないことがわかる。この結果より、音響モデルは、認識対象の発話様式の音声を用いて、全学習(音響モデル状態分割(トポロジー学習)、ラベル学習、連結学習)を行っても、問題が少ないことがわかった。そのため、朗読発話音声を用いず作成を行った音響モデルをリリースモデルとする。

5 まとめ

本稿では、新たに音声翻訳通信研究所に対して正式リリースする音響モデルの検討を行った。検討は、音響モデル学習データ量の増加による最適な状態数、混合数の選択、朗読発話音声を用いた音響モデルの自然発話音声認識への適応について行った。比較の結果より、自然発話音声データベースから作成した音響モデルが高い性能を示すことがわかった。

よって、自然発話データベースのみから作成した

- 男性モデル: 1400 状態 5 混合
- 女性モデル: 1400 状態 15 混合

を音声翻訳通信研究所に対し、音響モデルとしてリリースする。リリースする各音響モデル、コンフィグレーションファイル等のディレクトリは、付録 D を参照にされたい。

参考文献

[1] A. Nakamura, S. Matsunaga, T. Shimizu, M. Tonomura, and Y. Sagisaka. Japanese speech databases for robust speech recognition. In *Proc. ICSLP*, pp. 2199–2202, Philadelphia, 1996.

[2] H. Singer, M.Tonomura, Q. Huo, J. Ishii, T. Fukada, and M. Schuster. Baseline acoustic models for the spoken language database(SDB/SLDB). Technical Report TR-IT-0206, ATR, 1997.

[3] 深田俊明, 柘植覚, H. Singer, 内藤正樹. 連続音声認識用音響モデル (version 2.0). Technical Report TR-IT-0241, ATR, 1997.

[4] 柘植覚, 内藤正樹, H. Singer, 深田俊明, 高野優. 連続音声認識用音響モデル (ResearchJ V5). Technical Report TR-IT-0266, ATR, July 1998.

[5] 山本博史, 匂坂芳典. 接続の方向性を考慮した多重クラス複合 n-gram 言語モデル. 信学技報, Vol. SP98-102, pp. 49–54, December 1998.

[6] 嵯峨山茂樹. 音声認識のための音声分析とラベル変換. Technical Report TR-I-0347, ATR, 1993.

[7] 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.

[8] 山本博史, 中嶋秀治. 音声翻訳システムのための日本語音声認識言語モデル ('98 年 11 月版). Technical Report TR-IT-0284, ATR, 1998.

[9] 小坂哲夫. 不特定輪や音声認識の研究. Technical Report TR-IT-0101, ATR, 1995.

付録 A　認識時の コンフィグレーションファイル

```
#I/Ocontrol config :
I/Ocontrol:rpcTable=
I/Ocontrol:rpcNumber=33
I/Ocontrol:outputByteorder=BigEndian
I/Ocontrol:outputFd=stdout
I/Ocontrol:outputParamType=
I/Ocontrol:outputParamSize=
I/Ocontrol:outputFormat=Lattice
I/Ocontrol:inputEOFexit=ON
I/Ocontrol:inputByteorder=BigEndian
I/Ocontrol:inputParamType=float
I/Ocontrol:inputParamSize=26
I/Ocontrol:inputFormat=FrameSync
I/Ocontrol:inputFd=

#ATRlattice config file :
ATRlattice:lexicon=LEX.W.f.h
ATRlattice:amname=AM.M.CMS.bin,AM.F.CMS.bin
ATRlattice:active_model=1,2
ATRlattice:lmscale=8.000000,13.000000
ATRlattice:wdpenalty=0,0
ATRlattice:ngram=Multi-Class-2,multicomp.700.bin
ATRlattice:beam=110.000000,110.000000
ATRlattice:work_area=3800,150
ATRlattice:frame_shift=10
ATRlattice:pause_symbol=-
ATRlattice:dimension=26
ATRlattice:state_skip=ON,75000
ATRlattice:phone_boundary=ON
ATRlattice:word_boundary_skip=2
ATRlattice:word_merge=all
ATRlattice:UTT_START=5
ATRlattice:UTT_END=6
ATRlattice:backward_frame=-1
ATRlattice:amscale=1.000000
ATRlattice:UTT_END_delay=70

#ATRresult config :
ATRresult:merge_list=merge2.list
ATRresult:minimum_utt=0
ATRresult:dp_unit=FILE
ATRresult:dp_weight=1.0,1.0,1.0
ATRresult:pause_symbol=-
ATRresult:UTT_END=6
ATRresult:UTT_START=5
ATRresult:re_beam=
ATRresult:N_best_out=stdout
ATRresult:N_best=10
ATRresult:lattice_out=lattice.out
ATRresult:answer=answer.ANS,answer.TRS
```

付録 B　認識結果の詳細

学習話者数 407 名を用いて学習した男性モデル 1400 状態 5 混合、女性モデル 1400 状態 15 混合を併用し認識実験を行った結果の詳細を表 11に示す。表中の acc は、単語 accuracy、No/Wo は、ラティスのノード数を認識結果の単語数で除した値、cpu/u は認識に要した計算時間を発声時間で除した値、Li/Wo は、ラティスのリンク数を認識結果の単語数で除した値である。また、err は、ワークエリア不足などの原因により、認識結果が得られなかった発話数を示す。

表 11: 男性モデル 1400 状態 5 混合、女性モデル 1400 状態 15 混合を用いた認識結果の詳細

conversation ID	best word	net word	best word accuracy				speed	error
	acc/cor	acc/cor	total	Ins	Del	Sub	No/Wo(cpu/u, Li/Wo)	utt.
TAC70016.A	85.23/ 88.64	96.59/ 97.73	88	3	5	5	6.08(3.76, 9.51)	0
TAC70017.A	96.88/ 98.44	100.00/100.00	64	1	1	0	5.58(3.29, 7.49)	0
TAC70021.A	98.06/ 98.06	100.00/100.00	103	0	1	1	4.99(2.79, 6.51)	0
TAC70022.A	86.92/ 87.69	98.46/ 98.46	130	1	4	12	9.88(3.73, 18.05)	0
TAC70023.A	90.18/ 97.32	94.64/ 97.32	112	8	0	3	7.78(4.09, 11.19)	0
TAC70103.A	93.24/ 95.95	100.00/100.00	74	2	1	2	4.28(3.48, 6.70)	0
TAC70202.A	83.56/ 86.30	100.00/100.00	146	4	4	16	13.79(5.44, 33.50)	0
TAC70304.A	83.61/ 90.16	93.44/ 93.44	61	4	0	6	9.93(6.28, 18.30)	0
TCC70109.A	75.90/ 83.13	92.77/ 95.18	83	6	1	13	9.72(3.94, 15.63)	0
TCC70201.A	60.76/ 68.35	81.01/ 84.81	79	6	8	17	16.16(7.20, 34.06)	0
TCC70212.A	78.71/ 82.58	94.19/ 94.19	155	6	6	21	12.17(3.77, 22.74)	0
TCC70307.A	88.37/ 91.47	94.57/ 96.12	129	4	3	8	9.52(4.08, 18.56)	0
TCC71008.A	78.57/ 80.95	89.29/ 91.67	168	4	7	25	8.92(3.45, 16.66)	0
TCS70034.A	85.33/ 91.33	96.00/ 99.33	150	9	2	11	9.19(3.58, 16.13)	0
TCS70055.A	88.82/ 88.82	98.14/ 98.14	161	0	8	10	5.58(3.03, 9.79)	0
TCS70070.A	81.82/ 87.88	95.45/ 95.45	66	4	1	7	16.71(4.78, 33.84)	0
TCS70074.A	97.44/ 98.72	100.00/100.00	78	1	0	1	5.88(2.59, 9.26)	0
TAC70015.A	91.43/ 93.33	100.00/100.00	105	2	2	5	6.27(3.56, 11.92)	0
TAC70019.A	92.86/ 92.86	97.96/ 97.96	98	0	2	5	4.27(3.21, 5.84)	0
TAC70101.A	98.32/ 98.32	100.00/100.00	119	0	0	2	3.41(2.56, 4.00)	0
TAC70102.A	97.04/ 97.78	99.26/100.00	135	1	1	2	4.27(2.66, 5.92)	0
TAC70201.A	88.89/ 90.48	93.65/ 94.44	126	2	1	11	5.64(2.90, 10.22)	0
TAC70203.A	88.06/ 90.30	93.28/ 94.03	134	3	3	10	6.51(3.24, 10.56)	0
TAC70301.A	85.34/ 88.79	99.14/100.00	116	4	7	6	5.34(2.90, 9.46)	0
TAC70303.A	99.10/ 99.10	99.10/ 99.10	111	0	0	1	4.00(2.31, 5.45)	0
TCC70103.A	91.03/ 92.31	98.72/ 98.72	78	1	0	6	5.17(2.85, 8.08)	0
TCC71001.B	88.11/ 89.19	95.14/ 95.68	185	2	4	16	5.49(3.09, 9.51)	0
TCC71007.A	89.73/ 95.21	94.52/ 97.95	146	8	2	5	4.66(2.46, 5.83)	0
TCC71016.A	80.43/ 84.06	96.38/ 97.10	138	5	4	18	6.93(2.95, 11.34)	0
TCC71035.A	88.76/ 92.13	100.00/100.00	89	3	0	7	5.52(3.25, 8.67)	0
TCS70004.B	80.69/ 82.24	93.82/ 93.82	259	4	15	31	8.25(4.32, 16.00)	0
TCS70010.A	76.53/ 78.57	83.67/ 85.71	98	2	3	18	6.58(3.74, 11.39)	0
TCS70013.A	90.12/ 95.06	97.53/ 98.77	81	4	1	3	5.67(2.89, 9.08)	0
TCS70020.A	83.81/ 87.62	92.38/ 95.24	105	4	5	8	12.93(5.55, 25.22)	0
TCS70023.A	88.95/ 93.68	97.89/ 98.42	190	9	1	11	7.08(2.76, 13.77)	0
TCS70025.A	88.57/ 94.29	94.29/ 99.05	105	6	0	6	5.27(2.75, 9.37)	0
TCS70028.A	96.47/ 97.65	100.00/100.00	85	1	0	2	4.88(2.70, 7.13)	0
TCS70047.A	83.72/ 95.35	94.19/ 98.84	86	10	1	3	7.38(2.81, 11.87)	0
TCS70059.A	94.94/ 94.94	98.73/ 98.73	79	0	2	2	4.90(3.13, 7.63)	0
TCS70082.A	90.00/ 94.67	96.00/ 96.67	150	7	0	8	8.35(3.33, 16.83)	0
TSC71005.B	86.09/ 88.70	93.04/ 93.04	115	3	2	11	6.50(3.10, 10.86)	0
TSC71013.A	74.76/ 82.86	89.52/ 92.86	210	17	4	32	9.30(4.45, 18.50)	0
male	85.22/ 88.63	95.56/ 96.64	1847	63	52	158	9.22(3.95, 16.98)	0
female	87.69/ 90.80	95.58/ 96.69	3143	98	60	229	6.41(3.20, 11.01)	0
both	86.77/ 90.00	95.57/ 96.67	4990	161	112	387	7.45(3.47, 13.24)	0

付録 C　適応条件の比較検討

ここでは、朗読発話モデルを用いた自然発話音声への適応を行うための、適応パラメータの比較検討を行う。適応パラメータは、本文で述べた表 9をもとに検討する。比較検討を行った各結果に示す acc は、単語 Accuracy、err utt は、ワークエリア不足、ビーム幅不足などに起因して認識を正常に行うことができなかった発声数である。

はじめに、 training type を viterbi training から forward-backward アルゴリズムに変更した時の認識性能を表 12に示す[3]。この表より、 training type による認識性能の差は小さいことがわかる。そのため、以下の比較実験では、適応に必要な時間が短い viterbi training を training type として用いる。

表 12: 適応パラメータの比較 (training type の変更)

training type	acc (male/female/both)	err utt (male/female/both)
viterbi	85.22/89.18/87.73	2/0/2
f-b	85.23/89.28/87.81	3/0/3

次に、パラメータ推定方法 (estimation type) の違いによる認識性能の差を調べた。パラメータ推定の方法として、 MAP 推定、最尤推定 (ML) を用いて行った結果を表 13に示す。この結果より、ほぼ両方の認識性能の差がないことがわかる。ここでは、ほんの若干ではあるが MAP 推定の認識性能が高いため、次からの実験も MAP 推定をパラメータ推定方法とする

表 13: 適応パラメータの比較 (estimation type の変更)

estimation type	acc (male/female/both)	err utt(male/female/both)
ML	85.22/89.18/87.73	2/0/2
MAP	84.95/89.47/87.83	3/0/3

最後に、移動ベクトルの平滑化を行う際の平滑化パラメータの値 (smoothing control) を変化させた場合の認識性能の比較した。本稿では、この値をデフォルトの 20 と、 10 の場合について比較を行った。その結果を表 14に示す。

表 14: 適応パラメータの比較 (smoothing control の変更)

smoothing control	acc (male/female/both)	errutt(male/female/both)
20.0	84.95/89.47/87.83	3/0/3
10.0	84.83/89.53/87.81	2/0/2

[3]ここでは、パラメータ推定は、最尤推定を用いている。

付録 D　リリースを行ったディレクトリ構造

```
/dept1/work1/ResearchJ/V7/
        |- TR/: テクニカルレポート (TR-IT-0300)
        |
        |- amodel/: 音響モデル
        |
        |- data/: 認識用音響パラメータファイル
        |       |- CMS_MFCC/: 発話毎に CMS を行った MFCC
        |
        |- lmodel/: 言語モデル
        |
        |- result/: 認識結果
                  |- config/: 認識に用いたコンフィグレーションファイル
                  |- result.b110/: リリース音響モデルを用いた認識実験
```