ＴＲ－ＳＬＴ－００５８

環境依存型発話ノイズモデルによる音声認識
Speech Recognition Using Context-Dependent Utterance Noise Models

濱口　早太　　實廣　貴敏
Souta Hamaguchi　　Takatoshi Jitsuhiro
松田　繁樹　　新美　康永
Shigeki Matsuda　　Yasunaga Nimi
中村　哲
Satoru Nakamura

２００４年２月２３日

概要

人が発声するときには，息継ぎやリップノイズなどの雑音が実際には生じる．音声認識では，周囲雑音だけでなく，こういった雑音において誤認識を起こす可能性が高い．これらの雑音をここでは「発話ノイズ」と呼ぶ．一般には，"filler model"と呼ばれるような比較的簡単なモデル化で済まされる事が多い．そこで，本研究では，視察ラベルを用い，発話ノイズの種類別に環境依存型音響モデルを作成する．さらに，音素モデル自体も発話ノイズ環境依存型に作成しなおしながら，精度向上を図るため，一連の音響モデル作成手順（Viterbi アライメント作成，MDL-SSS 法による自動状態数決定も含んだ HMM 状態共有構造作成，混合分布化，HMM パラメータ推定）を繰返し行う．ATR 擬似旅行会話データベース(TRA)を用いた評価実験により，音素モデルのみに比べ，約 1%の単語錦精度の向上が得られた．

（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 Telecommunication 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

人が発声する際、音声や環境ノイズ以外に息継ぎ音や、喉のなる音、唇を開くときの音（リップノイズ）などの雑音が出現する。それらをここでは「発話ノイズ」と呼ぶ。これらノイズへの対策を行わない場合、発話ノイズ区間で音声モデルのマッチングが起こり、誤認識する可能性が高くなる。これに対し、発話ノイズを認識するモデルを作成し、音素モデルと組み合わせることで、音声と発話ノイズを同時に認識させ誤認識を防ぐ手法が考案されている。

従来では、ＧＭＭのような簡単なモデルを音素モデルに組み合わせることで対処してきた。そこで本研究では発話ノイズ種類別に環境依存型モデルを作成することを目的とする。また、精度向上を図るため、一連の音響モデル作成手順（Viterbi アライメント作成，MDL-SSS 法[2]による自動状態数決定も含んだ HMM 状態共有構造作成，混合分布化，HMM パラメータ推定）を繰返し行い、音響モデルを再作成及び学習する。以上 2 つの手法を検討する。

２節では、発話ノイズ種類、学習データ量、モデルの学習方法の詳細を述べ、さらに手法の拡張として、音響モデル繰り返し作成及び学習について述べる。３節では、２節で提案された手法で作成したモデルを用いて、音素識別実験、連続音声認識実験を行う。音素モデルのみの認識精度と比較し、環境依存型発話ノイズモデルの有効性を検討する。4 節では研究の総括をのべ、5 節では、本研究結果を反映させた、これからの課題を述べる。

２．環境依存型発話ノイズモデルの作成

従来、「発話ノイズ」に対して、ＧＭＭなどの簡単な発話ノイズモデルを音素モデルに組み合わせることで、対処されることが多い。本研究では、より良いモデル作成を目指し、環境依存型モデルを検討する。視察ラベルを用いて，発話ノイズの種類別に環境依存型音響モデルを作成する．さらに，精度向上のため音素モデル自体も発話ノイズ環境依存型に作成しなおしながら，一連の音響モデル作成手順（Viterbi アライメント，MDL-SSS 法[2]による HMM 状態共有構造作成，混合分布作成，HMM パラメータ推定）を繰返し行う．

２．１ 発話ノイズ別環境依存型モデル

学習データはＡＴＲ擬似旅行会話データベース（ＴＲＡ）を使用する。視察ラベルを基に発話ノイズ区間の特徴量を抽出する。本研究で使用する発話ノイズは以下の 3 種類である。

表１　発話ノイズ・データベース

学習データ	テキスト表記	データ量
息継ぎ音	@_br_@	7,164 区間
リップノイズ	@_ls_@	15,106 区間
喉の音	@_vs_@	4,014 区間

発話ノイズは他にも種類はあるが、学習データとしては不十分なデータ量のため、本研究では表に記述する３種類の発話ノイズモデルのみを作成する。

これらの発話ノイズのモデルをＭＤＬ基準を用いた逐次状態分割法により、自動的に構造決定し作成する。作成した環境依存型発話ノイズモデルと音素モデルとを組み合わせて音響モデルを作成する。

次に、発話ノイズ記号を認識辞書内で扱えるように文法を挿入する。まず、認識辞書に発話ノイズ記号を登録する。一般的な認識辞書では単語間、単語後に無音記号が記述されている。この文法を発話ノイズ記号が使用できるように変換する。

本研究では図１の文法を使用した。

例　a ng sh i ng {|-}　→ a ng sh i ng {|-|{|-}{@_br_@|@_ls_@|@_vs_@}{|-}}

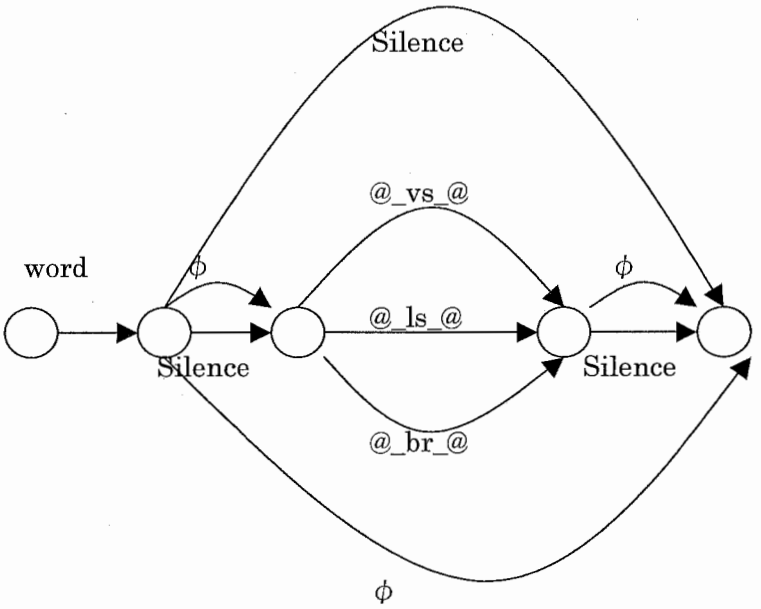


図１　発話ノイズ記号の文法挿入

図１の文法は単語間、もしくは単語後に、単数の発話ノイズ記号を認識する文法となっている。複数の発話ノイズ記号を認識する文法がより好ましいが、音声認識機械の性能の関係により、本研究では以上の簡便な文法を使用している。

２．２　音響モデル繰り返し作成、及び学習

手法の手順を以下に示す

① 特徴量データ作成
② 音素モデル再作成
③ 無音モデル再作成
④ 発話ノイズモデル再作成
⑤ 全モデル結合し音響モデル作成

学習データはＡＴＲ擬似旅行会話データベース（TRA）から特徴量を抽出する。音素区間に関しては音響モデルを使用して学習データのViterbiアライメントを取り、音素区間を推測、特徴量を抽出する。繰り返し１回目の特徴量抽出には、２．１節で作成した環境依

存型発話ノイズモデルを使用する。２回目以降は、一つ前の繰り返しで作成した音響モデルを使用する。また、無音、発話ノイズ区間の特徴量抽出は、視察ラベルを使用する。

抽出した特徴量で、モデルを再作成する。モデルは全て、MDL-SSS 法によるトポロジー学習で、構造を自動決定する。以下に各モデルの状態の分布を示す。

音素モデル　　・・・　５混合
発話ノイズモデル・・・　単一分布
無音モデル　　・・・　１０混合

手法のベースラインモデルの無音モデルは、発話ノイズを含んだ特徴量で学習している。しかし、繰り返し作成、学習する無音モデルは発話ノイズを含まず、無音データの特徴量のみで学習する。

３．実験

発話ノイズモデルを音素モデルと組み合わせた音響モデルを用いて音素識別実験、連続音声認識実験を行う。

３．１ 実験条件

ATR 旅行会話データベースから"The Travel Arrangement Task(TRA)"を用いて評価実験を行った。

評価データとして、TRA から、学習データに含まれない話者である４２人（男性１７人,女性２５人）による、４２片対話、５５１発話を用いた。

分析条件としては 16kHz サンプリング周波数、フレーム長 20msec,フレーム周期 10msec を用いた。特徴量には１２次元ＭＦＣＣ,１２次元ΔＭＦＣＣ,Δ対数パワーを用いている。

言語モデルの学習データとしては TRA を含む 7,389 片対話,129,285 発話を用いた。言語モデルとしては多重クラス複合バイグラム[3]を使用した。

音素は２６種類とした。各音素モデルごとに音響モデル構造を作成するとし、初期モデルは各音素とも３状態３ループのものを用いた。音響モデル構造を各状態単一分布とし、ML-SSS 法[1]、MDL-SSS 法で作成した。前者を、音響モデル繰り返し作成のベースラインモデル１とし、後者をベースラインモデル２とする。ベースラインモデル１は状態数１４００,ベースラインモデル２は状態数７７１である。両モデルを男女別に学習データを分

け、性別依存モデルとした。学習データとして、TRA データベース、話者４０７人（男性１６６人、女性２４１人）による音声データを用いた。音素区間のみで約 5.4 時間、全区間で約 8.3 時間、6,432 発話の音声データであった。

実験に用いた語彙数は 16,528 単語、5,216 複合語であり、評価データの単語は全て含んでいる。

MDL-SSS 法によるトポロジー学習により得られた環境依存型発話ノイズモデルは以下の状態数となった。

表２　環境依存型発話ノイズモデル

発話ノイズ	総状態数	最大状態長
@_br_@モデル	10	3
@_ls_@モデル	3	3
@_vs_@モデル	8	1

３．２　音素識別実験

実験に使用するモデルはベースライン２に対応する音素モデル、音素モデルに環境依存型発話ノイズモデルを組み合わせて作成した音響モデルである。

３．２．１　結果

音素識別実験結果を図２,図３,図４に示す。

図２,図３の凡例の”ベースラインモデル２”は実験条件、ベースラインモデル２に対応する音素モデルの識別率である。これを識別実験のベースラインとする。音響モデルの識別率は”context dependent model”である。それ以外の識別率は音素を発話ノイズと誤識別した割合である。

図４は発話ノイズの識別率である。”phone”は発話ノイズを音素と誤識別した割合である。

図２　音素識別実験結果（母音）

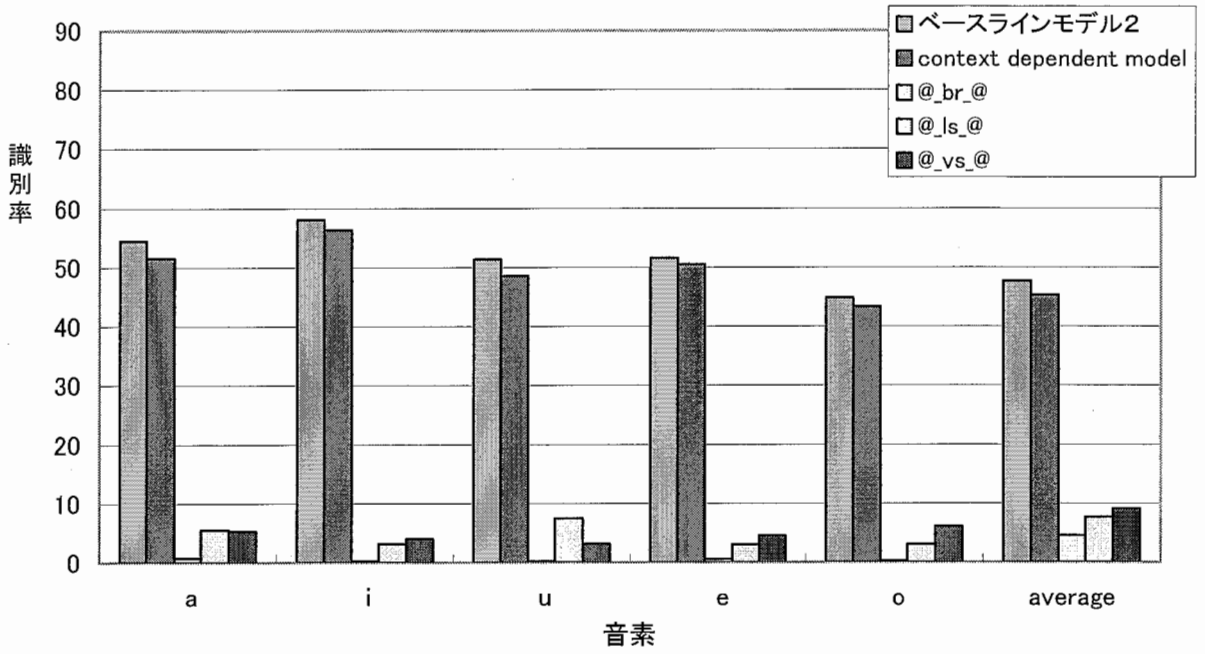


図３　音素識別実験結果（子音）

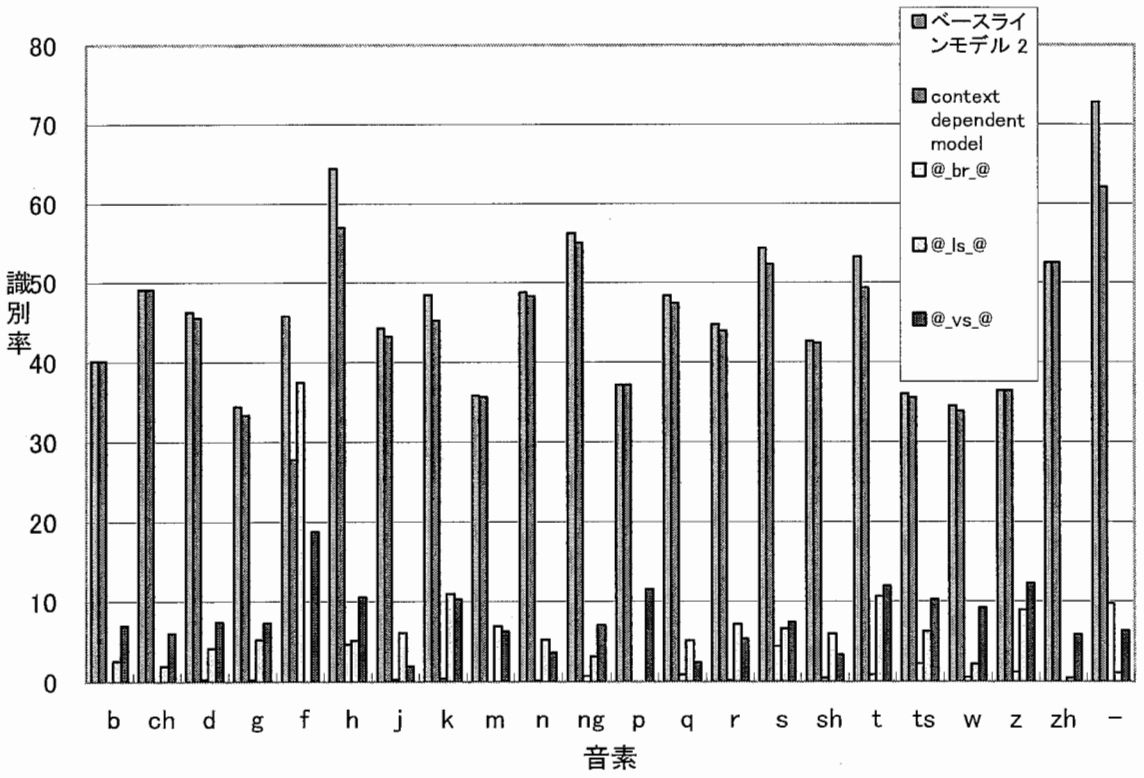
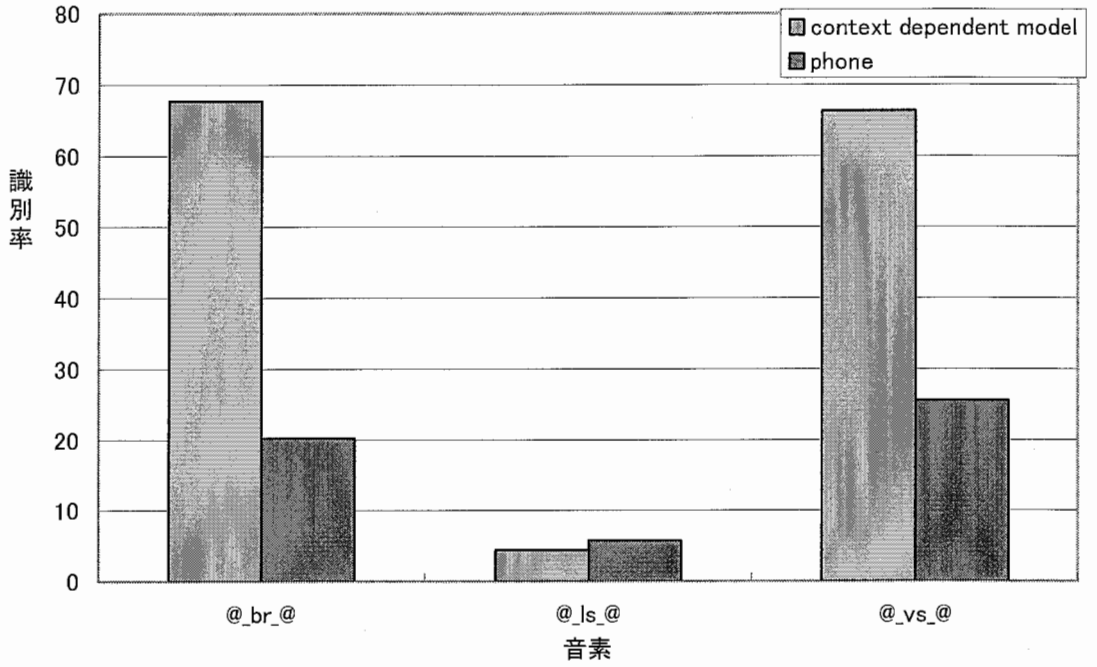


図４　音素識別実験結果（発話ノイズ）



３．２．２　考察

全体的にベースラインと比較し、識別率が低下していることがわかる。識別率低下の傾向が顕著に現れているのが、”f”の識別率であり、音素の大半が”@_br_@”と識別されている。図２より、全音素の約１割が発話ノイズと認識されていることから、識別率低下は発話ノイズモデルの音素の誤認識であると考えられる。

３．３　連続音声認識実験

３．３．１　結果

音響モデル繰り返し作成及び学習を４回行った。繰り返し作成回数に応じた状態数の変化を以下の表３にまとめる。

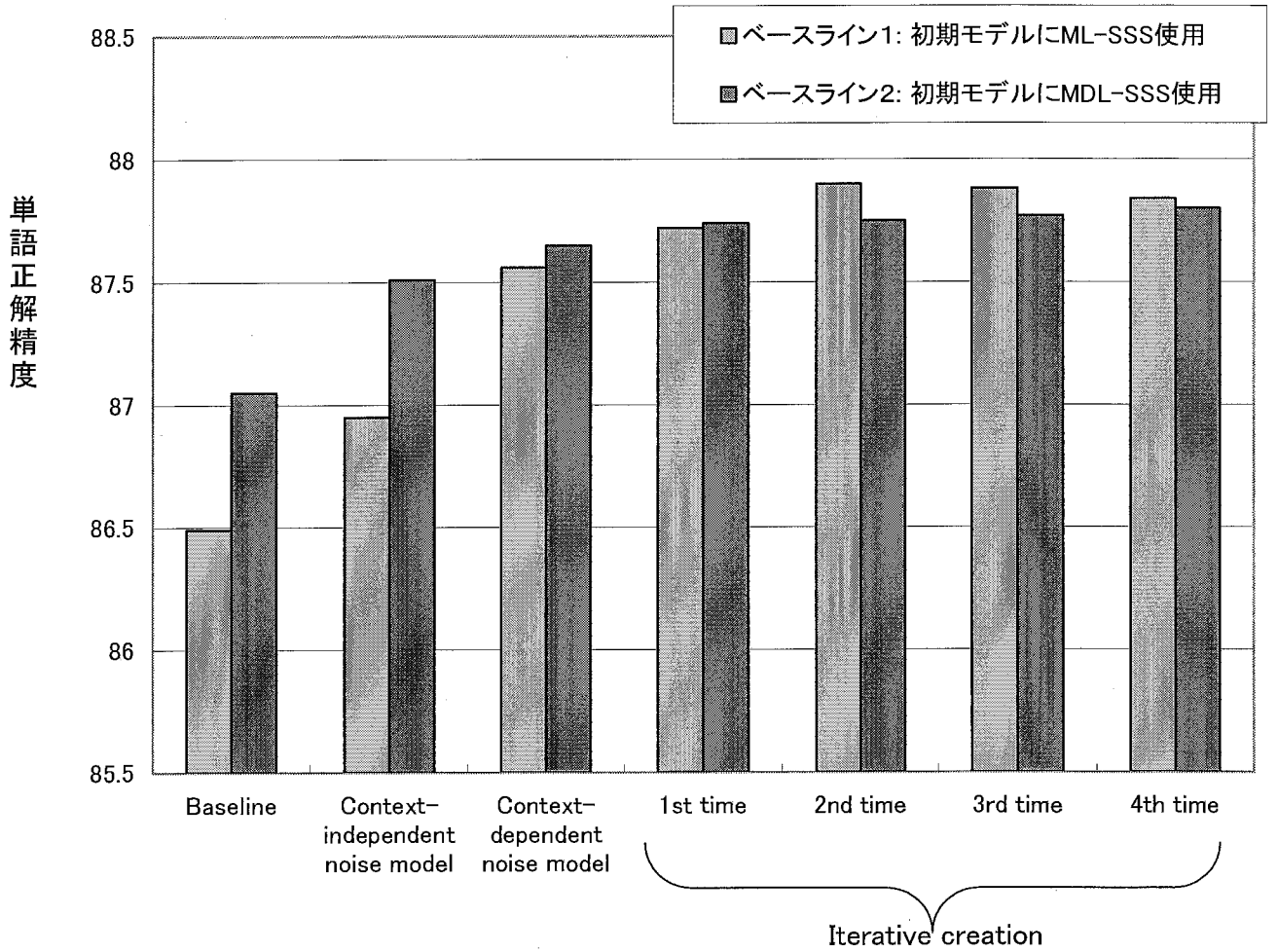
表３　状態数の推移

	Baseline		1st time		2nd time		3rd time		4th time	
ML-SSS	総状態数	最大状態長	総状態数	最大状態長	総状態数	最大状態長	総状態数	最大状態長	総状態数	最大状態長
phone	1400	5	775	5	775	5	775	5	775	5
@_br_@	10	3	13	4	6	4	5	3	7	4
@_ls_@	3	3	5	4	3	2	4	2	5	2
@_vs_@	8	1	8	1	8	1	7	1	8	1
MDL-SSS	総状態数	最大状態長	総状態数	最大状態長	総状態数	最大状態長	総状態数	最大状態長	総状態数	最大状態長
Phone	771	5	768	5	768	5	768	5	768	5
@_br_@	10	3	11	3	6	4	5	3	9	4
@_ls_@	3	3	7	2	5	2	5	2	5	2
@_vs_@	8	1	12	2	8	1	7	1	9	2

連続音声認識実験結果を図５に示す。

項目軸,"baseline","context independent model","context dependent model"は２．１節で作成した音響モデルの単語正解精度である。"context dependent model"は特徴量と共に抽出した発話ノイズ間のトライフォンを考慮する環境依存型発話ノイズモデル、"context independent model"はトライフォンを考慮しない環境独立型発話ノイズモデルを組み合わせた音響モデルである。 "iteration creation"で一括りにしてある項目は２．２節で繰り返し作成および学習した音響モデルの単語正解精度である。

図５　連続音声認識実験結果



※ベースライン１の初期モデルおよびCI noise model以外はすべてMDL-SSS法により構造作成

３．３．２　考察

ベースラインと比較し、最大で約１％、単語正解精度が向上している。よって、提案手法の有効性が確認できる。また、２．１節で作成した音響モデルより、２．２節で作成した音響モデルの単語正解精度が向上しているが、繰り返し作成２回程度で認識率は収束している。

４．まとめ

発話ノイズに対し、視察ラベルを用いて発話ノイズ別環境依存型発話ノイズモデルを作成することで対処した。さらに、精度向上を図るため、音響モデル作成手順を繰返し行い、音響モデルを再作成及び学習した。ATR 擬似旅行会話データベースを用いて、音素識別実験、連続音声認識実験を行う。音素識別実験結果を検討すると、ベースラインと比較して、識別率が平均で約２％低下している。モデルが音素の約１０％を発話ノイズとして認識していることから、発話ノイズモデルの誤認識が原因であると考えられる。次に、連続音声認識実験結果を検討する。ベースラインと比較し、最大約１％の単語正解精度の向上が見られる。正解精度の向上率は、ベースラインモデルによって違う。ML-SSS 法で作成したベースラインモデルが、MDL-SSS 法で作成したベースラインモデルに比べて単語正解精度が向上している。また、正解精度の向上率は手法によっても違う。２．１節の手法より、２．２節の音響モデル繰り返し作成及び学習で作成したモデルの単語正解精度が高い。しかし、繰り返し２回程度で単語正解精度が収束している。

５．課題

音素識別率がベースラインを下回ったことから、発話ノイズと音素の識別率向上が必要である。また、実環境でも使用できるモデル作成を目指すために、発話ノイズの種類数、学習データを増やし、TRA 以上に自然な会話を行っている音声データを集め、発話ノイズモデルを作成する。

6．参考文献

[1]M. Ostendorf, H. Singer,
``HMM Topology Design Using Maximum Likelihood Successive State Splitting,"
Computer Speech and Language, vol. 11, pp. 17--41, 1997

[2]T. Jitsuhiro, T. Matsui, S. Nakamura,
``Automatic Generation of Non-Uniform Context-Dependent HMM Topologies Based on the MDL Criterion,"
Proc. of EUROSPEECH'03, vol. 4, pp. 2721--2724, 2003

[3]H. Yamamoto, Y. Sagisaka,
``Multi-Class Composite N-gram Based on Connection Direction,"
Proc. of ICASSP'99, vol. 1, pp. 533--536, 1999