環境音モデルを用いた頑健な音声認識に関する研究（0221036）

Robust Speech Recognition Using Non-Speech Models

山田武志　筑波大学大学院システム情報工学研究科

Takeshi Yamada　Graduate School of Systems and Information Engineering, University of Tsukuba

研究期間　平成 14 年度～平成 15 年度

研究費総額　8,569,094 円（間接経費、消費税を含む）

概要　本研究の目的は、環境音の隠れマルコフモデル（HMM）を精密に生成し、雑音環境下での音声認識精度を改善することである。まず、環境音モデルの単位や構造を決定する手法として、逐次状態分割法による環境音モデルの生成法、エルゴディックモデルからの環境音モデルの再生成法を考案し、評価実験によりその有効性を確認した。また、このような環境音モデルを有効活用するために、音声と環境音の重畳区間情報を推定する手法、及び推定した重畳区間情報を HMM 合成法に基づく音声認識において利用する方法を考案し、評価実験によりその有効性を示した。さらに、尤度に基づいて音声・非音声判別、既知・未知環境音判別を行うシステムを提案・構築し、その性能を評価すると共に環境音データを大量に収集した。

Abstract　The purpose of this research is to improve the performance of noisy speech recognition by generating non-speech models. This research proposed and evaluated a decision method of the model unit/architecture/parameters, a speech recognition algorithm based on HMM composition and noisy frame detection, and a recording system which always monitors unknown noises.

１．研究目標

近年の音声認識技術の発展は目覚しく、ディクテーションシステム（音声ワープロ）などのアプリケーションが市販されるまでに至っている。しかし、これらの音声認識システムには、周囲雑音の影響によって認識精度が低下するという問題がある。現状では接話マイクロホンの使用によってこの問題を回避しているが、このままでは音声インタフェースとしての利便性を十分に生かすことができない。よって、マイクロホンから離れた位置での発話を可能にする技術（ハンズフリー音声認識の技術）が必要不可欠である。

従来、ハンズフリー音声認識を実現するために様々な研究がなされている。これらの研究では、認識対象である音声以外の音を一律的に雑音とみなしていることが多い。しかし、実世界に存在する多種多様な音（環境音）を雑音として一括りに扱うことには無理があり、雑音の種類によっては十分な認識精度が得られないという問題が生じる。よって、広範な雑音環境下で頑健な音声認識を実現するためには、個々の環境音の特性を十分に考慮する必要があると考えられる。

以上から、本研究では、環境音の隠れマルコフモデル（HMM）を精密に生成し、このような環境音モデルを用いて雑音環境下での音声認識精度を改善することを目的とする。具体的には、モデルの単位や構造を同じ基準の下で自動的に決定する手法を開発すると共に、このような環境音モデルを有効に活用するために、HMM 合成法に代表される音声認識アルゴリズムの改良に取組む。さらに、無数に存在する環境音をあらか

じめ全てモデル化しておくことは不可能なので、未知の環境音を常時モニタリングするシステムを構築し、環境音モデルを逐次更新していく手法を開発する。

２．研究内容

具体的な研究内容は次の通りである。

① モデルの単位や構造を決定する手法の開発と評価

② 音声認識アルゴリズムの改良と評価

③ 未知の環境音を常時モニタリングするシステムの構築と評価

④ 環境音モデルの逐次更新手法の開発と評価

①については、モデルの単位と構造を、出力尤度最大化という共通の尺度の下で同時に最適化する方法を開発する。特に、1993 年に鷹見らにより提案された、逐次状態分割法による隠れマルコフ網の自動生成法に着目し、この方法を環境音の場合に拡張する。また、②の音声認識アルゴリズムにより、その有効性を評価する。

②については、このような環境音モデルを有効に活用するために、HMM 合成法に代表される音声認識アルゴリズムの改良に取組む。

③については、常時環境音をモニタリング（収録）するシステムを構築し、既知の環境音と未知の環境音を判別するアルゴリズムを開発する。

④については、③により収録した既知・未知の環境音データを用いて、環境音モデルを逐次更新する手法を開発する。既知の環境音については、話者適応などに用いられる MLLR 適応や MAP 適応などのアルゴリズムを適用する。また、未知の環境音については、新規に環境音モデルを作成するのか、既存の環境音モデルに含めるのかを判断するアルゴリズムを開発する。

最終的には、①～④の全てを統合したシステムを構築し、広範な雑音環境下でその性能を評価する。

３．研究結果

３．１．モデルの単位や構造を決定する手法の開発と評価

３．１．１．逐次状態分割法による環境音モデルの生成

環境音認識においては、モデルの単位を、その音源の意味内容に基づいて決めることが多い。一方、音声認識においては、必ずしもその必要はないと考えられる。そこで、環境音の音響的・時間的特長を反映するモデルの構造に基づいて環境音をクラスタリングし、このようにして得られた各クラスをモデルの単位とすることを考え、1993 年に鷹見らにより提案された、逐次状態分割法による隠れマルコフ網の自動生成法に着目した。

隠れマルコフ網は、複数の状態のネットワークとして表される。各状態には、状態番号、受理可能なコンテキストクラス、先行・後続状態のリスト、遷移確率、出力確率分布が割り当てられている。隠れマルコフ網では、各コンテキストクラスを表す状態を連結したものが Left-to-Right 型 HMM と等価となるため、通常の HMM と同様に扱うことが可能である。また、複数のコンテキスト間で状態を共有することにより、統計的に安定した学習が行える。これは、学習データの量が制限されることが多い環境音にとって非常に有効である。

逐次状態分割法は、学習データに基づいて自動的に隠れマルコフ網の構造を決定する方法である。逐次状態分割法の処理の流れを以下に示す。

Ⅰ　初期モデルの学習：初期モデルとして、1 状態の隠れマルコフ網を全ての学習データで学習する。

Ⅱ　被分割状態の決定：隠れマルコフ網の全ての状態の中で、最も大きい出力確率分布をもつ状態を被分割状態とする。

Ⅲ　状態の分割：時間方向とコンテキスト方向の分割のうち、学習データに対する尤度の総和が最も高い分割を選択する。

Ⅳ　分布の再学習：分布の再学習を行い、Ⅱ～Ⅳの処理を所望の状態数になるまで繰り返す。

Ⅴ　隠れマルコフ網の学習：隠れマルコフ網の構造が決定された後、各状態の出力確率分布を所望の形状に変更し、再学習する。

逐次状態分割法の利点は、隠れマルコフ網の構造の決定やモデルパラメータの推定を、出力尤度最大化という共通の基準の下で自動的に行うことにある。

逐次状態分割法により、RWCP 実環境音声・音響データベースに含まれる 92 種類の環境音に対してモデルの単位と構造を決定し、環境音認識実験により初期評価を行った。このデータベースには、無響室で収録された 92 種類（各々約 100 サンプル）の環境音が含まれている。その種類は大きく次の 3 つに分けられる。

・　衝突音：物体の単発的な衝突に起因する音。

・　動作音：物体の動作に起因する音。音のみから音源を特定するのが困難。

・　特徴音：特徴的な音色を有する音。音のみから音源を特定するのが容易。

音響分析条件と隠れマルコフ網の条件を表 1 に示しておく。

表 1：音響分析条件（左 2 列）と隠れマルコフ網の条件（右 2 列）

サンプリング周波数	16kHz	初期モデル	1 状態
窓関数	ハミング窓	要因数	1（中心環境音のみ）
フレーム長	25msec	各経路上の最大状態数	3
フレーム周期	10msec	出力確率分布	単一ガウス分布
高域強調	$1-0.97^{-1}$	学習用データ	奇数番号データ
特徴量	MFCC（12 次元）	評価用データ	偶数番号データ

なお、無音モデルは 3 状態 Left-to-Right 型 HMM（単一ガウス分布、対角共分散行列）により別途生成している。

まず、隠れマルコフ網の状態数とクラス数の関係を図 1 に示す。ここで、クラスとは、始端から終端をむすぶ一本の経路が与えられたとき、その経路上の全ての状態で共通に受理可能な環境音の集合である。図 1 から、状態数が 2～36 のとき、状態数とクラス数はほぼ等しく、状態数が 37 以上のとき、クラス数が状態数を 4～8 上回っていることが分かる。

次に、クラス数と認識率の関係を図 2 に示す。図中の HMM とは、各クラスのモデルを 1 状態の HMM で表したものである。HMM の総状態数はクラス数と等しく、隠れマルコフ網の状態数はクラス数とほぼ等しいので、比較のために用いている。実験結果を以下にまとめる。

・　クラス数が 2～24、特に 14～24 のときの認識率は隠れマルコフ網の方が最大で 12%程度高い。これは、状態共有と時間方向への状態数の増加によるものと考えられる。

・　HMM、隠れマルコフ網共にクラス数が増加するにつれて、認識率が下降している。特に認識率の大きく下がっているところが幾つか見られる。この原因としてクラスの分割が適切に行われていないことが

考えられる。

・ クラス数が 25 以上のときの認識率は隠れマルコフ網の方が低くなっている。これは不適切な状態共有が一因であると考えられる。ただし、クラス数が 41 以上のときについては、HMM の状態数が隠れマルコフ網の状態数を 4～8 上まわっていることにより、認識率を直接比較することは難しい。

以上より、逐次状態分割法により環境音の音響的・時間的特長に応じたクラスが概ね形成されているものの、クラスの分割に失敗しているケースもあることが分かった。今後は、その原因を調査すると共に、②の音声認識アルゴリズムを用いて評価する予定である。

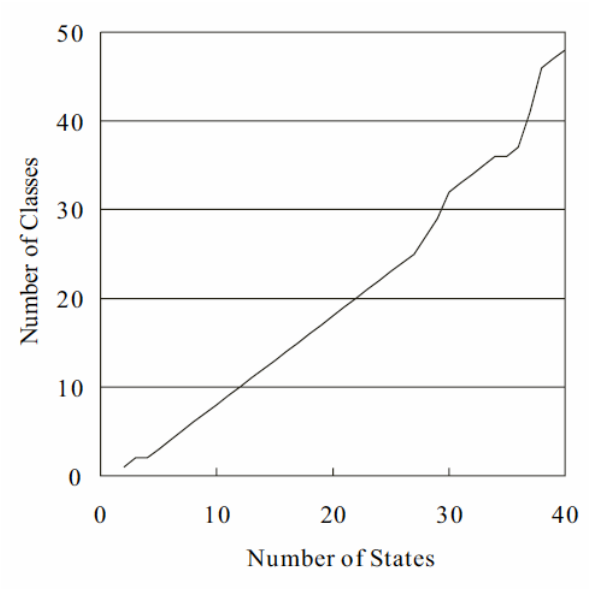

図 1：状態数とクラス数の関係

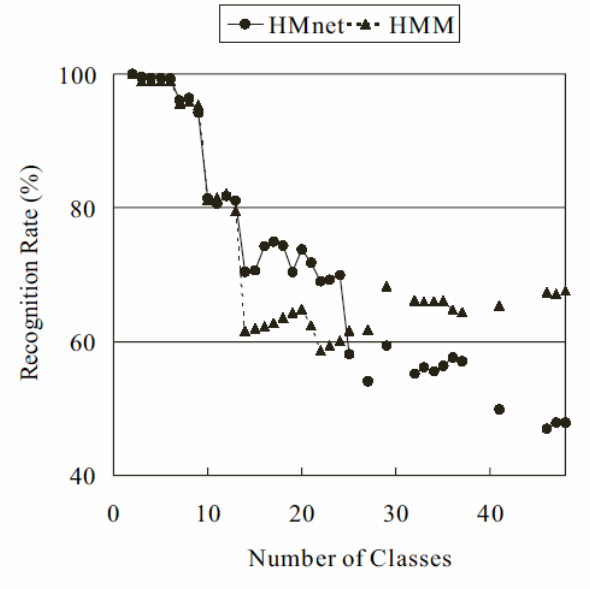

図 2：クラス数と認識率の関係

３．１．２．エルゴディックモデルからの環境音モデルの再生成

前節の手法は、事前に想定した環境音に特化したモデルを生成するため、そのモデルは未知の環境音に対しては必ずしも適していないという問題がある。この問題を解決するために、多種の環境音で学習したエルゴディック構造のモデルから、認識対象の音声に重畳している環境音の特性を適切に表すモデルをその都度再生成する手法を考案した。

提案法の概要を図 3 に示す。提案法では、まず様々な環境で収集した環境音データを用いて、エルゴディック構造のモデルを学習する。認識対象の音声に重畳している環境音のモデルは、このモデルからその都度再生成される。次に、認識対象の音声から非音声部分を切り出し、上述のエルゴディック構造のモデルを用いてビタビアライメントを求める。その結果、当該環境音に対する状態系列が得られる。図 3 には、単純な例として、状態 A から状態 B に状態遷移している様子を示している。最後に、この状態系列から適切な状態を選択し、連結することにより、当該環境音のモデルを再生成する。図 3 には、単純な例として、時間長が最も長い状態 A のみを選択する場合を示している。

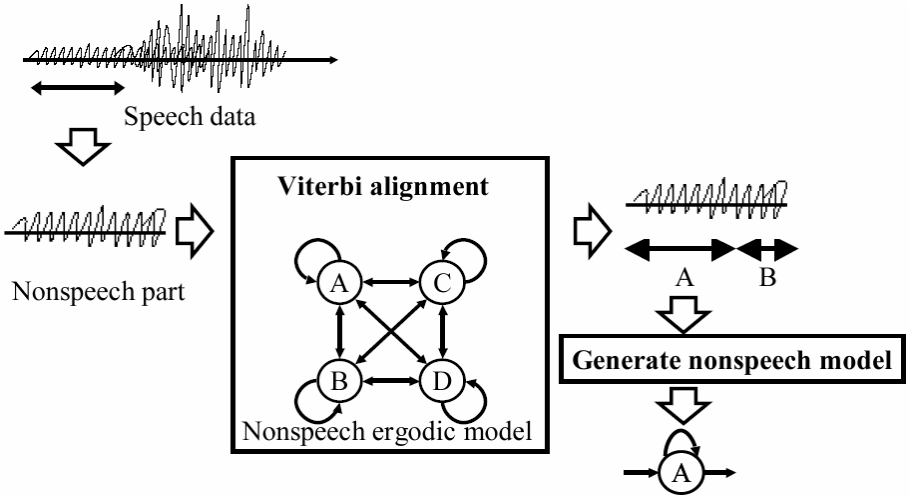

図3：提案法の概要

提案法の有効性を評価するために、雑音下連続数字認識タスクである AURORA-2J を用いて認識実験を行なった。表 2 に AURORA-2J の学習セットとテストセット、表 3 に認識実験の条件を示しておく。なお、本実験では、表 2 の Clean training のみを対象としている。

表 2：AURORA-2J の学習セットとテストセット

学習・テストセット	音声	雑音	チャネル	SNR
Clean training	110 名 8,440 発話	なし	G.712	Clean
Multicondition training	同上	Subway、Babble Car、Exhibition	同上	Clean、20、15、10、5
テストセット A	104 名 4,004 発話	Subway、Babble Car、Exhibition	同上	Clean、20、15、10、5、0、-5
テストセット B	同上	Restaurant、Street Airport、Station	同上	同上
テストセット C	104 名 2,002 発話	Subway、Street	MIRS	同上

表 3：認識実験の条件

窓関数	ハミング窓
フレーム長	25msec
フレーム周期	10msec
特徴量	メルケプストラム係数（12 次元）＋対数パワー（1 次元） ＋Δ係数（13 次元）＋ΔΔ係数（13 次元）
HMM（数字）	16 状態、混合分布数 20
HMM（sil）	3 状態、混合分布数 36
HMM（sp）	1 状態（sil の第 2 状態と共有）

提案法の有効性を評価するために、各数字モデルと環境音モデルの合成モデルを HMM 合成法により生成し、認識に用いる。HMM 合成は、39 次元のうち 12 次元のメルケプストラム係数のみを対象とする。ここで、SN 比は既知としている。提案法におけるエルゴディック構造のモデルの学習には、AURORA-2J のテストセット A に含まれる環境音データと電子協騒音データベースのダイジェスト版の全データを用いる。提案法においてビタビアライメントを求める非音声区間は、各テストデータの冒頭 100msec である。本実験では、無音モデルの代わりに環境音モデルを用いている。

1 状態 1 分布の環境音モデルの生成

提案法は、時間長が最も長い状態を選択し、1 状態 1 分布の環境音モデルを再生成する。ここで、提案法におけるエルゴディック構造のモデルの状態数は 64、状態毎の分布数は 1 である。また、比較のために、次の 3 つの条件で作成した環境音モデルを用いる。ここで、各環境音モデルは 1 状態 1 分布である。

[Ref 1.1] テストデータから切り出した冒頭の 100msec を学習データとし、テストデータ毎に 1 つの環境音モデルを作成する。

[Ref 1.2] 提案法におけるエルゴディック構造のモデルの学習に用いたデータを学習データとし、1 つの環境音モデルを作成する。

[Ref 1.3] 環境音の種類毎に十分なデータを用いて環境音モデルを作成する。テストセット B の環境音を含む。

各環境音モデルを用いたときの単語正解精度の平均を図 4 に示す。ここで、図中の Baseline は HMM 合成を行わない場合である。実験結果を以下にまとめる。

- Ref 1.1 の単語正解精度は、Baseline と比べて低下していることが分かる。これは、学習データが少ないことが原因であると考えられる。
- Ref 1.2 の単語正解精度は、Baseline と比べて 20%程度改善している。しかし、Ref 1.3 と比べると 5%程度の差がある。このことから、環境音の特性を考慮せずに生成した環境音モデルでは、十分な認識性能を得られないことが分かる。
- 提案法の単語正解精度は、Ref 1.3 と比べて 1%弱しか低下していない。

次に、テストセット毎の単語正解精度の平均を図 5 に示す。実験結果を以下にまとめる。

- テストセット A の場合、提案法の単語正解精度は、Ref 1.2 と比べて 5%程度改善しており、Ref 1.3 と同程度である。
- テストセット B の場合、提案法の単語正解精度は、Ref 1.2 と比べて 2%程度改善しているものの、Ref 1.3 と比べて 2%程度低下している。これは、本実験では、提案法におけるエルゴディック構造のモデルの学習に、テストセット B の環境音を用いていないからである。
- テストセット C の場合、提案法の単語正解精度は、Ref 1.2 に比べて 1%程度低下している。この原因については現在調査中である。

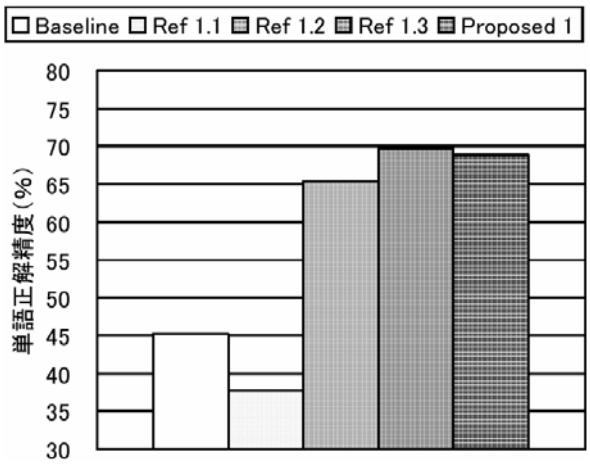

図4：各環境音モデルを用いたときの単語正解精度の平均（1状態1分布）

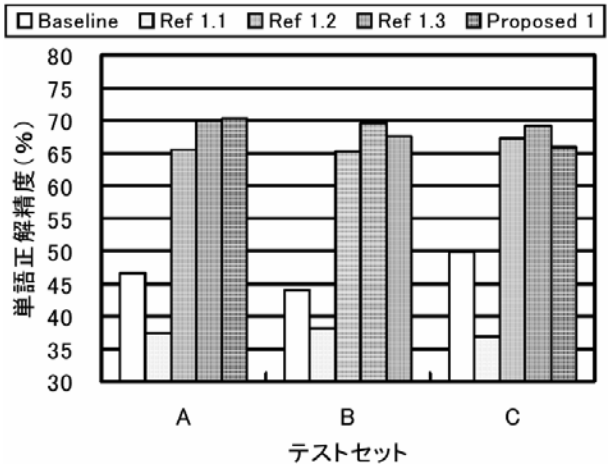

図5：テストセット毎の単語正解精度の平均（1状態1分布）

1 状態 2 分布の環境音モデルの生成

提案法は、時間長が最も長い状態を選択し、1 状態 2 分布の環境音モデルを再生成する。ここで、提案法におけるエルゴディック構造のモデルの状態数は 32、状態毎の分布数は 2 である。また、比較のために、次の 2 つの条件で作成した環境音モデルを用いる。ここで、各環境音モデルは 1 状態 2 分布である。

[Ref 2.2] 提案法におけるエルゴディック構造のモデルの学習に用いたデータを学習データとし、1 つの環境音モデルを作成する。

[Ref 2.3] 環境音の種類毎に十分なデータを用いて環境音モデルを作成する。テストセット B の環境音を含む。

各環境音モデルを用いたときの単語正解精度の平均を図 6 に示す。また、テストセット毎の単語正解精度の平均を図 7 に示す。図より、分布数を増やしたことにより、全体的に単語正解精度が改善しており、その改善量は特にテストセット B の場合に大きいことが分かる。また、各環境音モデルを用いたときの単語正解精度には、前項と同様の傾向があることが分かる。

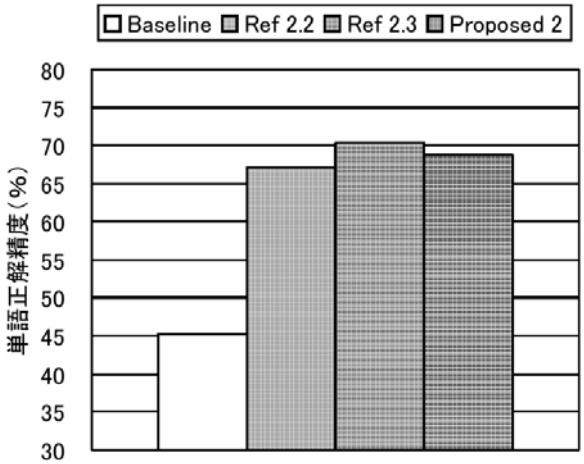

図6：各環境音モデルを用いたときの単語正解精度の平均（1状態2分布）

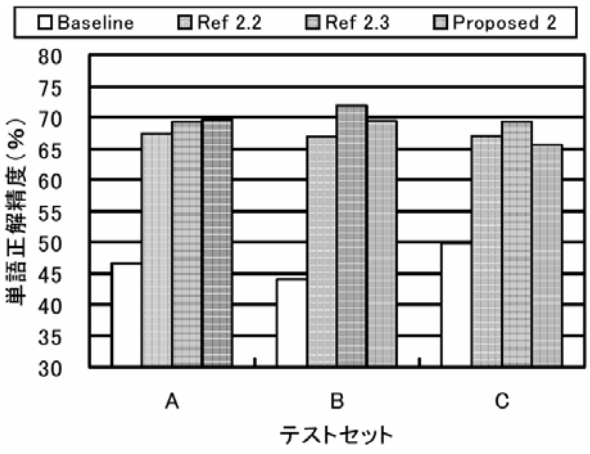

図7：テストセット毎の単語正解精度の平均（1状態2分布）

2 状態LR構造の環境音モデルの生成

提案法は、時間長が長い順に 2 状態を選択し、それらを元の時系列順に並べた 2 状態 LR 構造（状態毎に 1 分布）の環境音モデルを再生成する。ここで、提案法におけるエルゴディック構造のモデルの状態数は 64、状態毎の分布数は 1 である。

各環境音モデルを用いたときの単語正解精度の平均を図 8 に示す。また、テストセット毎の単語正解精度の平均を図 9 に示す。図より、提案法の単語正解精度は、提案法（1 状態 2 分布）と比べて改善していることが分かる。これは、テストセット A の Exhibition が LR 構造に適した環境音であり、この環境音に対する単語正解精度が大きく改善されたことによる。

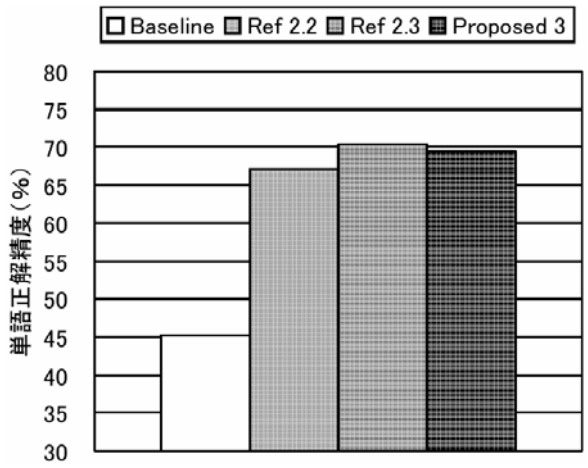

図8：各環境音モデルを用いたときの単語正解精度の平均（2状態LR構造）

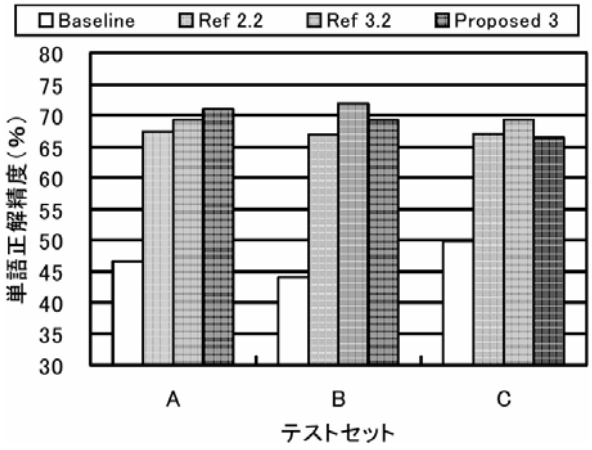

図9：テストセット毎の単語正解精度の平均（2状態LR構造）

提案法により、環境音の種類毎に十分なデータで学習したモデルを用いた場合と同程度の認識性能が得られること、様々な構造のモデルを柔軟に生成できることが分かった。今後は、環境音の特性に応じた構造を自動的に決定する手法を開発する予定である。

３．２．音声認識アルゴリズムの改良と評価

環境音が重畳している音声を精度良く認識するための手法として、HMM 合成法が提案されており、その有効性が広く示されている。しかし、HMM 合成法を適用する際には、音声と環境音が重畳している区間、音声に重畳している環境音の種類とその SN 比、といった情報をあらかじめ推定しておく必要がある。そこで、音声と環境音の重畳区間情報を推定する手法、及び推定した重畳区間情報を HMM 合成法に基づく音声認識において利用する方法を提案した。

重畳区間情報を推定する方法について述べる。まず、環境音が重畳している音声に対して、図 10（左）のような連結モデルを用いてビタビアライメントを求める。ここで、Sは音声モデル、N_1、N_2は環境音モデルである。その結果、SN比が高い区間（音声区間）とSN比が低い区間（環境音区間）については比較的精度良く推定することができる。次に、音声区間とそれと隣接している環境音区間の間には、音声とその環境音の重畳区間が存在すると仮定し、図 10（右）のような連結モデルを用いて再度ビタビアライメントを求める。ここで、S'は音声と当該環境音の合成モデルであり、あらかじめ想定しているSN比の数だけ用意する。その結果、音声と環境音の重畳区間、音声に重畳している環境音の種類とそのSN比、といった情報を推定することができる。

重畳区間情報を利用する方法について述べる。上記の手法で推定した重畳区間情報の例を図 11 に示す。ここで、横軸はフレーム番号であり、各々無音区間、環境音区間、重畳区間、音声区間であると推定されたことを示している。音声認識の際、(1)、(4)、(5) の区間については Clean モデルを参照する。一方、(3) の区間については、音声と当該環境音の合成モデルを参照する。この例では SN 比は 1 通りであるが、実際にはフレーム毎に SN 比の推定結果が得られるので、各々対応する SN 比の合成モデルを用いる。(2) の区間については、環境音区間として厳密に扱うのではなく、SN 比が非常に低い区間であるとみなし、(3) の区間と同様に合成モデルを参照する。これは、語頭の子音などが欠落するのを防ぐためである。

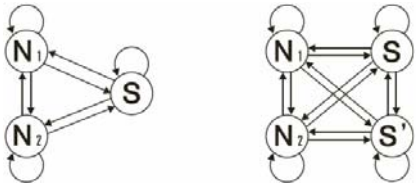

図 10：連結モデル

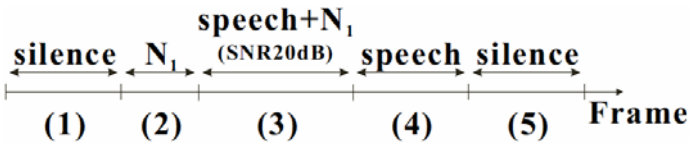

図 11：重畳区間情報の例

生活環境音データベースに含まれる 7 種類の環境音を音声データに重畳し、連続単語認識実験を行った。このデータベースは、家庭内で顕著な音が発生する場面を想定し、複数の家屋において測定した環境音で構成されている。環境音の種類は 11 個であり、各々にいくつかの条件が設定されている。環境音データの長さは、条件毎に 1 分程度である。本実験では、流し台、換気扇、洗面台、ドライヤ、浴室、テレビジョン、掃除機を用いており、各環境音データの前半 15 秒を評価用、残りを学習用としている。重畳区間情報の推定のための音声モデルは 1 状態 64 混合分布（学習データ：ASJ 研究用連続音声データベース)、認識のための音素モデルは 3 状態 16 混合分布（学習データ：ASJ 研究用連続音声データベースと新聞記事読み上げ音声コーパス）である。環境音モデルは 1 状態 1 混合分布であり、環境音の種類毎に作成している。また、合成モデルの SN 比は-20、-10、0、10、20dB の 5 通りである。分析条件は、サンプリング周波数が 16kHz、フレーム長が 25msec、フレーム周期が 10msec であり、12 次元の MFCC 係数を求めている。評価用の音声データとしては、ATR 音声データベース SetA の音韻バランス 216 単語を用いている。話者は MHT と MAU の 2 名である。本実験では、話者毎に単語音声を 6 個接続して連続単語データを 36 個作成し、単語数未知の連続単語認識を行っている。評価用データは、連続単語データと環境音データを計算機上で加算することにより作成している。その際、SN 比が 20、10dB となるように環境音の信号レベルを調整している。具体的な重畳方法は、連続単語データの全区間に渡って環境音データを重畳する場合（全重畳)、連続単語データに 1.25 秒の環境音データを 1 秒おきに重畳する場合（ランダム重畳）の 2 通りである。なお、一つの連続単語データ内では、同じ種類の環境音のみを重畳している。全音声区間に対する重畳区間の割合は、全重畳の場合は 100%、ランダム重畳の場合は 48%である。

全重畳の場合とランダム重畳の場合の単語認識率（全結果の平均）を図 12 と図 13 に示す。

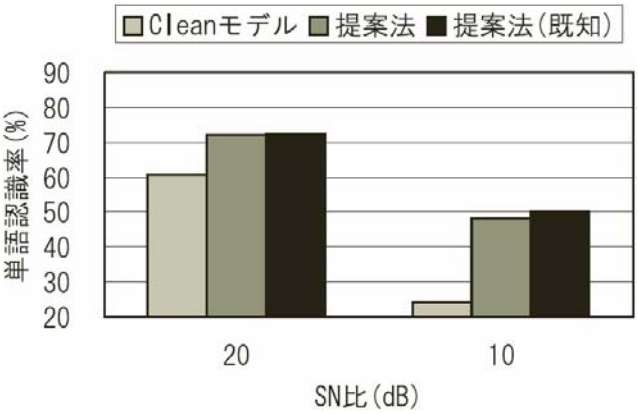

図 12：全重畳の場合の単語認識率

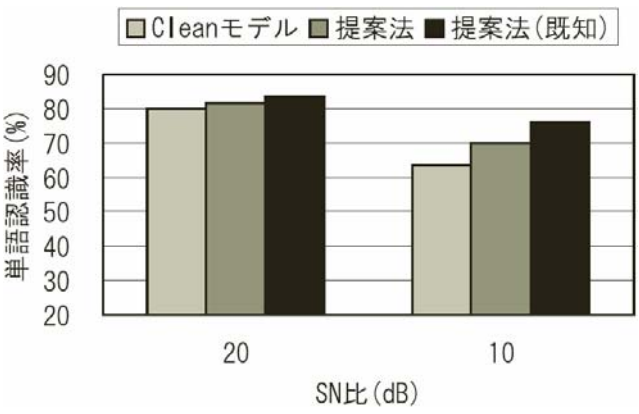

図 13：ランダム重畳の場合の単語認識率

図中の提案法（既知）は、重畳区間と重畳環境音の種類を正確に与えた場合の結果である。ただし、合成モデルの SN 比は一律同じ値を設定している。なお、提案法の重畳区間検出率は 80%強、重畳環境音正解率は 70%前後である。実験結果を以下にまとめる。

- 提案法の単語認識率は、Clean モデルと比べて最大で 23.9%改善している。改善の度合いは SN 比が低いほど大きい。
- 全重畳の場合とランダム重畳の場合を比べると、全重畳の方が単語認識率が低い。また、提案法の Clean モデルに対する改善の度合いは、全重畳の方が高い。これは、音声データに対する環境音の重畳率の違いによる。
- 提案法と提案法（既知）の単語認識率にはあまり差がないものの、ランダム重畳の場合に多少差が広がっていることから、重畳区間検出の精度を改善する必要があると考えられる。

提案法の単語認識率は重畳区間情報の正解を与えた場合と同程度であることを確認したものの、認識性能の改善度は環境音の種類に依存していることが分かった。今後は、①の環境音モデルを用いることにより、認識性能のさらなる改善を図る予定である。

３．３．未知の環境音を常時モニタリングするシステムの構築と評価

従来の HMM 合成法に基づく音声認識では、環境音モデルをあらかじめ作成しておく必要があるため、未知の環境音に対してはその効果が十分に得られないことがある。よって、音声認識の利用時に、環境音データを自動的に収集し、環境音モデルを逐次学習する機能が必要である。さらに、その環境音が既知のものか、未知のものかを判別し、学習の高効率化・高精度化を図る必要がある。そこで、音声・非音声判別、既知・未知環境音判別を行うシステムを提案した。提案システムは、話者認識・話者照合の技術を応用したものであり、尤度に基づいて各種判別を行う。

実環境で収録したデータを用いて性能を評価した結果、音声・非音声判別、既知・未知環境音判別共に、80%以上の判別率が得られた。一方、判別率と判別数の間にはトレードオフの関係があり、判別率を優先すると学習のためのデータ量が不足することになる。今後は、尤度に対する閾値を適切に決定する方法について検討する予定である。また、システムのリアルタイム化に取組む予定である。

４．今後の展開と波及効果

本研究では、以下の研究項目に取組んできた。

① モデルの単位や構造を決定する手法の開発と評価

② 音声認識アルゴリズムの改良と評価

③ 未知の環境音を常時モニタリングするシステムの構築と評価

④ 環境音モデルの逐次更新手法の開発と評価

①については、逐次状態分割法による環境音モデルの生成法、エルゴディックモデルからの環境音モデルの再生成法を考案し、評価実験によりその有効性を確認した。②については、音声と環境音の重畳区間情報を推定する手法、及び推定した重畳区間情報を HMM 合成法に基づく音声認識において利用する方法を考案し、評価実験によりその有効性を示した。③については、尤度に基づいて音声・非音声判別、既知・未知環境音判別を行うシステムを提案・構築し、その性能を評価すると共に環境音データを大量に収集した。今後は、①～③において明らかとなった課題に取組むと共に、④の研究開発を行っていく予定である。さらには、①～④の全てを統合したシステムを構築し、広範な雑音環境下でその性能を評価していきたい。

音声認識において、入力装置であるマイクを特に意識せずに、人に話しかけるような感覚での自然な発話を可能とするためには、雑音の問題を解決せねばならない。本研究は、そのためのアプローチの一つとして提案・実施したものである。本研究の成果が、雑音下音声認識の研究開発や実世界に存在する多種多様な環境音の特性の系統的な分類のための一助となることを期待している。

５．誌上発表リスト

なし

６．口頭発表リスト

[1] Takeshi Yamada、Naoto Isaka、Hiroshi Osuka、Nobuhiko Kitawaki、Futoshi Asano、“Noise robust speech recognition based on HMM composition and noisy frame detection ”、International Congress on Acoustics（ICA2004）（京都府京都市）（2004）

[2] 井坂直人、山田武志、北脇信彦、“HMM 合成法を用いた音声認識のための環境音モデルの生成の検討”、日本音響学会春季研究発表会(神奈川県厚木市)（2004）

[3] 井坂直人、大須賀洋、山田武志、北脇信彦、浅野太、“重畳区間の推定情報を用いた HMM 合成に基づくロバスト音声認識の検討”、電子情報通信学会音声研究会（東京都港区）（2003）

[4] 井坂直人、大須賀洋、山田武志、北脇信彦、浅野太、“環境音モデルと HMM 合成を用いた音声区間検出法の音声認識への適用”、日本音響学会春季研究発表会（東京都新宿区）（2003）

[5] 井坂直人、山田武志、北脇信彦、浅野太、“隠れマルコフ網と逐次状態分割法による環境音モデル化の検討”、日本音響学会秋季研究発表会（秋田県秋田市）（2002）

[6] 山田武志、井坂直人、北脇信彦、浅野太、“隠れマルコフ網と逐次状態分割法を用いた環境音のモデル化の検討”、電子情報通信学会音声研究会（東京都港区）（2002）

７．申請特許リスト

なし

８．登録特許リスト

なし

９.受賞リスト

なし

10.報道発表リスト

なし