6B−5

音声情報検索システム

菅原一秀

日本アイ・ビー・エム株式会社　東京基礎研究所

1.はじめに

音声認識・録音・再生を組み合わせた「音声情報検索システム」を試作し、候補区間の設定法等について検討したので報告する。このシステムは録音された音声データに対し、検索したい単語やフレーズ等をユーザーの発声により指定し、一致度の高い部分を探索していくものである。音声認識の側面から見ると、ＨＭＭを使ったワード・スポッティングの応用である。

2. 構成

全体の構成は図１の通りであり、次の各段階からなる。

1. ラベル・データベース(符号帳と HMM パラメータ)の作成
   1.1　符号帳の作成
   1.2　HMM パラメータの学習
2.　録音音声データの準備
3.　検索

以下、これらについて説明する。

2.1.　ラベル・データベースの作成

2.1.1.　符号帳の作成

音声をある短い時間(10ms程度)で切り出して分析し、その特徴量のなす空間を100-200程度に分類する（特徴抽出・ラベル化）。その分類の基準を決める符号帳を話者の発声からクラスタリングによって作成する。

2.1.2.　HMM パラメータの学習

その部分空間それぞれが特徴量及び継続時間についての揺らぎを持つ確率機械（ＨＭＭ）だとするモデル化を行い、それらの揺らぎのパラメータを話者の発声から forward-backward アルゴリズムにより学習する。

2.2.　録音音声データの準備

検索対象となる音声データを録音すると同時に 2.1 で作成した符号帳を使ってラベル付けを行う。(図２)

2.3.　検索

検索したい単語やフレーズを入力し、録音音声データと同様にラベル付けを行い、検索ラベル列を得る。これと録音音声のラベル列との連続的なマッチングをとり、該当区間の判定を行い、検索結果を区間の候補の表という形で得る。この結果に基づき区間の表示や録音音声の再生などを行う事ができる。次に連続的なマッチングと該当区間の判定について述

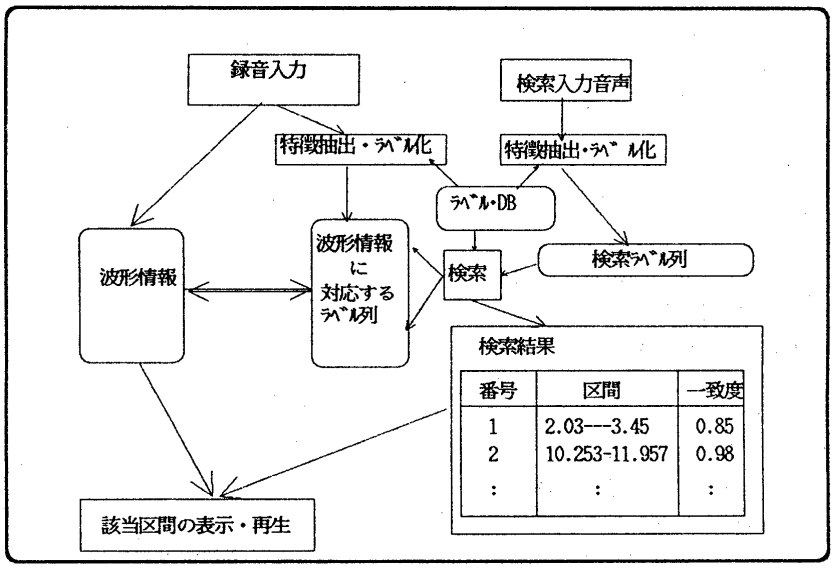

図 1.　音声情報検索装置

Speech Information Retrieve Sytem
Kazuhide Sugawara
Tokyo Research Laboratory, IBM Japan, Ltd.

べる。

2.3.1 連続マッチング

ＨＭＭパラメータ、ラベル出力確率と状態遷移確率をそれぞれ
Out(i, j), Tr(i, k) とする。但し、i は ＨＭＭの番号、j は出力ラベルの番号、 k は状態遷移の番号 (={0, 1, 2}, 図3)とする。検索ラベル列を a(i), i = 0,...,N-1, 録音ラベル列を b(i), i=0,...,N-1 とする。録音ラベル列のすべての点を始端点とし、状態遷移の許す範囲で、(音声の性質を考慮し、遷移 '0'は連続できないとする制限を付ける)最良の累積確率を与えるパスを Ｍ Ｘ Ｎ 個の格子点上で計算する。計算式は次で与えられる。検索入力 i フレーム目、録音入力 j フレーム目にあたる格子点での累積確率 s(i,j)は

$$s(0, j) = 1.0, \text{ for } j=0,\ldots, N-1$$

$$s(i, j) = \max_{i>0} \begin{cases} s(i-1,j)*Tr(a(i-1),0)*Out(a(i-1),b(j)) & \text{前の遷移が'0'でない} \\ s(i-1,j-1)*Tr(a(i-1),1)*Out(a(i-1),b(j-1)) & j>=1 \\ s(i-1,j-2)*Tr(a(i-1),2)*Out(a(i-1),b(j-2)) & j>=2 \end{cases}$$

と計算される。

検索入力の終端まで計算が進むと録音入力の各フレーム j に対して累積確率 s(N, j) が得られる。これの対数をとり、検出ラベル列の長さ N で正規化しマッチング・スコア L(j) とする。また、この計算の途中では始端点のフレーム番号も保持しておき、録音入力のフレーム j で終わるパスの始端点を f(j) とする。

2.3.2 区間の判定

前項で計算されたマッチング・スコア L(j) と始端点のフレーム番号 f(j) から該当区間を算出する。一フレーム当たりの遷移及び出力確率の積の対数の下限を示すいき値 T を定めておく。録音入力のフレーム j で終わるパスについて L(j)>T であれば該当区間である可能性が高いとして処理をする。始端点を共有するパスのうちで最大のマッチング・スコアを持つものを選び、該当区間の候補とする。更に、これらのパスが区間を共有する場合も同様に、最大のマッチング・スコアを持つものを該当区間の候補とする。この処理は録音入力のフレーム順に行う。最終的に得られる結果は互いに重なり合わない、該当区間の候補の始端点、終端点、及びマッチング・スコアの表である。

3. 実験システム

ＰＣ上で、音声認識・録音・再生の機能を持つアダプタ・カードを用い、これらの機能を結合して検索システムを試作した。音声の特徴抽出・ラベル化はアダプタ・カードに搭載されているＤＳＰで処理を行っているが、連続マッチングはＰＣのＣＰＵを使用している。このため現状では検索時間がかなりかかるのが難点である。(録音時間10秒の音声を検索するのに、検索入力時間長の数倍かかる。)

4. まとめ

録音・再生と音声認識の機能を結合して音声情報検索システムを試作した。今後は連続マッチングの簡略化を含めた高速化や現在は固定値となっている検出基準の適応的な設定法、などを検討していく予定である。

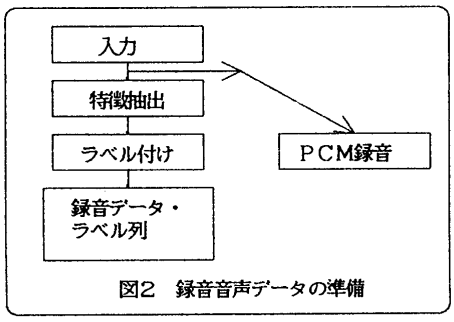

図2 録音音声データの準備

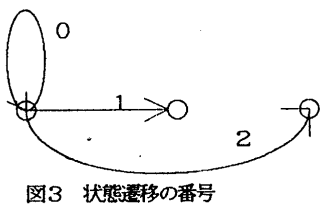

図3 状態遷移の番号