Decision Treeによる日本語音素環境クラスタリングの検討

6G－1

大嶋良明　西村雅史
（日本アイ・ビー・エム　東京基礎研究所）

１．はじめに

HMMを用いた音声認識の手法に興味があり，サブワード単位での，例えば音素連結などによる連続音声の認識を目指している．それには音素認識率を向上する必要があり，連続音声中に見られる大きな音素変形などの問題に対処せねばならない．記述性の高い音響モデルを作成するために，Decision Treeを構成して音響的に類似した環境を階層的にまとめる手法が知られており，日本語連続音声中の音素認識に適用し，効果を調べた．

２．処理の概要

音響処理

下表にまとめた条件で信号処理を行ない，臨界帯域フィルタ出力及びその時間変化量を，ベクトル量子化し，得られたラベル列をもとに，認識装置の訓練と認識実験とをおこなった．

表１　音響処理のまとめ

A/D変換	標本化周波数12kHz 16ビット量子化
FFT分析	窓長256点　シフト96点
静的特徴量	19チャネル臨界帯域フィルタ出力 および正規化対数パワ
動的特徴量	静的特徴量前後4フレームの変化量
VQ符号帳	静的特徴量256 動的特徴量256

認識手法

認識装置の訓練は，ラベルヒストグラムに基づいた最尤推定による．認識アルゴリズムは，以前に報告した単語音声認識装置の予備選択部で用いた方式である[2]．すなわち，静的特徴量と動的特徴量に関するラベルの対数尤度の和を求め，値の上位順に候補とする．

３．Decision Treeによる音素環境クラスタリング

Bahlらは前後５音素分の環境を考慮し，訓練データを階層的にクラスタ化する事で音素モデルの記述性を高める手法を英語音声に適用し，その有効性を報告している[1]．今回我々が使用したのは，同様の方法であるがデータ量が少ない等の理由により，音素環境の考慮範囲を前後１音素分に限定した．

クラスタの分割は，その各要素に対して前後環境に関する最適な質問を行ない，質問の記述に当てはまるものと当てはまらないものとに仕分ける事により実現する．質問とは，例えば「先行音素は/p/であるか」といったものである．最適な質問とは，次節で定義する評価関数を最大化する質問である．クラスタの分割は，その要素数あるいは最適な分割による評価関数値が，予め設定したそれぞれの閾値を下回ると停止する．

クラスタ分割の評価関数

音響ラベル系列yは，その時間構造を無視するとラベルヒストグラム$y_1y_2y_3...y_F$によって統計的性質が記述できる．y_iはyの中でのラベルiの出現回数をあらわす．Fはラベルアルファベットのサイズであり，本報告では256である．これらのy_iがそれぞれ独立のポアソン分布で表現されるとすると，yの尤度Pr(y)は次のようにあらわされる：

$$\Pr(y)=\prod_{i=1}^{F}\frac{\mu_i^{y_i}e^{-\mu_i}}{y_i!}$$

よって要素数N_nであるノードnに含まれるラベル列の集合Y_nの尤度は

$$\Pr(Y_n)=\prod_{y\in Y_n}\prod_{i=1}^{F}\frac{\mu_i^{y_i}e^{-\mu_i}}{y_i!}$$

各ラベルiについて，モデルパラメータμ_iの最尤推定値は

$$\mu_i=\frac{1}{N_n}\sum_{y\in Y_n}y_i$$

となる．質問qによってノードnがlとrに分割された時に，分割の評価関数$m(q,n)$を

$$m(q,n)=\log\{\Pr(Y_l)\ \Pr(Y_r)\ /\ \Pr(Y_n)\}$$

と定義する．Y_l, Y_rはノードl,rに含まれるラベル系列の集合である．

４．認識実験

音声試料

実験に用いた音声試料は，ATRによる研究用連続音声データベースより男性ナレータ１名分（話者MTK）である．音素バランスの取れた約50文ずつをまとめた自由発声全503文[4]のうち，前半250文を訓練用とし，後半253文を認識実験に使用した．

"A Study on Context-dependent Clustering of Japanese Phones by Using Decision Trees," by Y. Ohshima and M. Nishimura (Tokyo Research Laboratory, IBM Japan, Ltd.)

認識の単位は基本的には音素であり，その切り出しには，ATRデータベースで使用されている階層化ラベリング[3]中の，音声記号層の時間情報をそのまま用いた．データ中においては，/N, j/など明確な音素境界を持たない複数の音素よりなる部分も少なくない．本報告ではそれらも音素と同様の一つの認識単位として扱っており，「音素」と述べた時には，かかる複合的なセグメントも含めた認識単位を想定している．

訓練データを用いたクローズドの予備実験においては，/r/の誤認識が顕著であった．これは継続長の短いトークンが多数存在し，前後環境による揺らぎも含めて，音声現象が十分に表現されていないためと推察された．例えば/r/の訓練用593トークン中，継続長1フレームのトークンが203あった．そこで/r/については，特にセグメントを前後に1フレーム分拡張したものを尤度推定および認識に用いた．

最小クラスタサイズ

前述のように，分割を許すクラスタの大きさが音素環境の表現を左右する．そこでクラスタリングによる傾向性を観察するため，最小クラスタのサイズを100, 20, 10, 5と変えて，実験を行った．音素環境クラスタリングの結果生成される終端ノード数との関係を表2に示す．観察のためシングルトンクラスタまで生成すると，終端ノード数は2625となり前後環境も含めた音素環境の異なり数に対応する．クラスタリングを全く行わない場合には，ノード数は中心音素の異なり数となり134である．なお訓練用データに含まれるトークンの総数は12523であった．

表2 最小クラスタサイズと生成ノード数の関係

サイズ	100	20	10	5
ノード数	204	461	705	1,119

実験結果

上記条件にて253文，11918トークンの音素認識実験を行い，図1に示すような傾向性を得た．累積認識率の算出は，中心音素が正答となる候補を見つけ，それよりも上位に選ばれた異なり音素を勘定する事で求めた．

音素環境クラスタリングを行わない場合，第1位候補のみの認識率は70.7%であったが，サイズ100のクラスタリングで79.4%，サイズ5のクラスタリングで80.9%が得られた．第5位候補までの累積認識率では，それぞれ97.0%，97.2%，および97.2%であった．第5位以下も含めると，サイズ5，10，20のクラスタリング結果に基づく認識率もクラスタリングなしの場合の認識率も，数値的に非常に似通ったものとなった．例えば累積認識率が

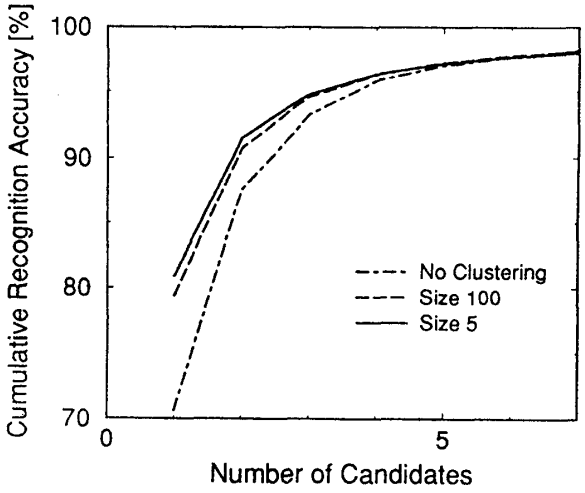

図1 累積認識率

99.5%を越えるのは，第20位候補においてであった．まとめとしては，音素認識実験において，クラスタリングにより，第5位候補までの累積認識率に向上が認められた．

おわりに

Decision Treeによる階層的な音素環境クラスタリングを自由発声データに適用し，音素認識実験において効果を確かめた．現在は，構築した木構造に基づいて，フェノニックな異音ベースフォームによる音声認識の準備を進めており，その場合には前後環境も含めて本手法による認識性能の向上を目指している．なお，今回の実験で検討できなかった問題点には，評価関数の閾値の設定などがある．あわせて有効性の検討を行いたい．

参考文献

[1]Bahl, et. al.,"Decision Trees for Phonological Rules in Continuous Speech," Proc. IEEE ICASSP-91, pp185-188, (1991).

[2]西村，橋本，菅原，「フェノニックマルコフモデルを使った大語彙音声認識」，音響講論，1-1-19,(平成4年3月)．

[3]武田，匂坂，片桐，桑原，「音韻ラベルを持つ日本語音声データベースの構築」，音声研究会資料SP87-19, (1987).

[4]阿部，匂坂，梅田，桑原，「研究用日本語音声データベース利用解説書（連続音声データ編）」，ATRテクニカルレポートTR-I-0166, (1990).