音声の変動記述のための知識ベースについて

3C-10

速水　悟、田中　和世、太田　耕三
電子技術総合研究所

1．はじめに

音声認識システムの性能向上のカギは、そこで用いられる知識の質と量にある。この知識のうち特に重要なものは、音声の変動記述のための知識である。これは、同じ音素に相当する音声の音響的特性が音素コンテキスト、話者、発声速度などの要因により変動し、その変動が性能向上の大きなネックとなっているからである。

本報告の目的は、筆者らが開発中の不特定話者大語彙単語音声認識システム[1-4]において用いられている変動記述のための知識を例として、知識ベースによる変動記述について考察することである。

2．認識システムの構成

はじめに筆者らが開発中の不特定話者大語彙単語音声認識システムについて述べる。全体構成を図１に示す。

特徴ベース（音響的特徴についての情報を集めたもの）は、ラベル付けされた音声データベースから、音素コンテキストごとの音素片（音素より細かい音響的セグメント）を収集し、音響分析を行い、さらにクラスタリングによって集約したものである。単語辞書は、キーボードより入力した認識対象語彙に書換規則・変形規則を適用して、さまざまな発声の変形を表す音素片のネットワーク構造に変換されたものである。ネットワーク中の各音素片のスペクトル情報（メルケプストラム）・継続時間は、音素コンテキストに適合するものが特徴ベースから選ばれる。認識部では、マイクより入力された音声を音響分析し、単語辞書中の音素片ネットワークとのマッチングにより認識結果が得られる。

システムは現在 Cray X-MP 216 上にインプリメントされており、ほぼ実時間で動作する。認識性能として、不特定話者４９２単語で、９６．０％と高い認識率を得ており（男性話者１０名の簡易防音室での録音による）、またその場で単語辞書を作成することによって任意の単語集合の認識が可能となっている。図２に動作例を示す。

3．変動記述

音声認識に対する知識工学的な研究の１つに、音声の専門家のもつ知識をエキスパートシステム化しようとする試みがある[5-8]。たとえば、ＭＩＴのグループのスペクトログラムリーディング・エキスパートシステムでは、人間の専門家がスペクトログラムを読む際に用いる知識をエキスパートシステム化して音素認識を行い、音素レベルの知識を体系的に取り扱おうとしている。これらのシステムは、音素レベルの知識の記号化、明示的規則化における指針を与えたという点で高く評価されるものの、ここでの「変動記述」という立場とは異なるものである。

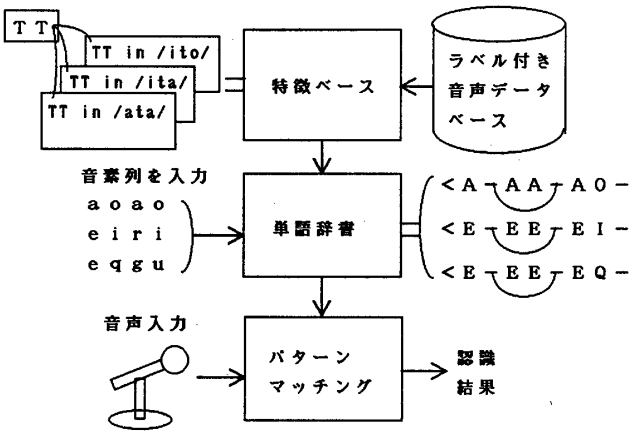

図１　単語音声認識システムの全体構成

ここで「変動」とは、音素コンテキスト、話者、発声速度などの要因によって、同じ音素に相当する音声の音響的特性が変動することを指す[1,9]。エキスパートシステム化の試みにおいても、典型的なパターン認識のわく組みに従って、変動要因によって変わることのない特徴を抽出することで変動に対処しようとしている［不変特徴抽出］。これに対して「変動記述」とは、さまざまな要因による複雑な変動をその変動要因とともに大規模かつ体系的に記述しようとするものである。

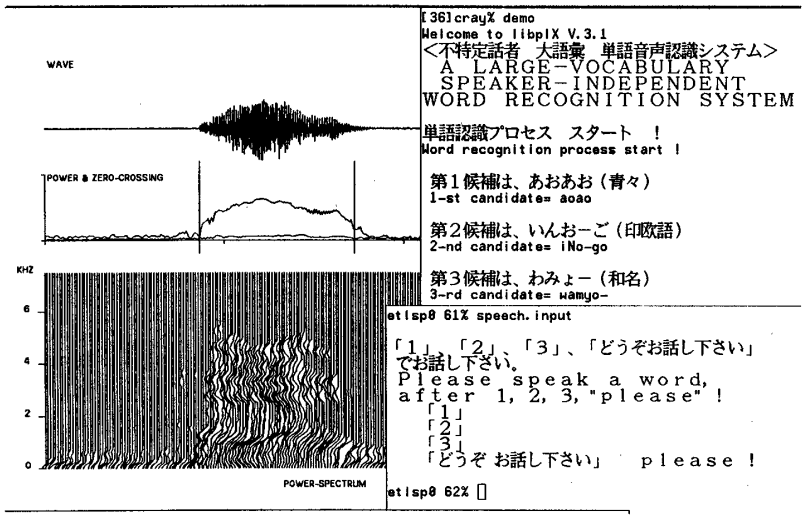

図２　認識システムの動作例

On the knowledge-base for description of acoustic characteristic variations of speech.
Satoru HAYAMIZU, Kazuyo TANAKA, Kozo OHTA,
Electrotechnical Laboratory.

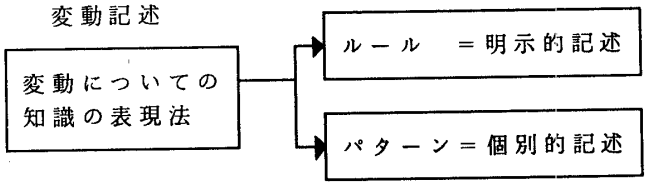

図３　変動についての知識の表現法

4．知識ベースによる変動記述

音声の変動記述を行うためには、変動についての知識およびそれを扱うわく組みを知識ベースシステムと考える必要があろう。変動についての知識をわずかな言葉ですべて書き表すことは困難であり、知識ベースとして大規模かつ体系的に記述する必要がある。たとえば２で述べた単語認識システムには様々な特徴があるが、これらの特徴を変動記述のための知識ベースという考え方に従って整理してみよう。

一般に知識ベースにおいて重要なことは、知識の質と量の問題であり、知識の表現法、獲得法の問題であろう。

知識の質において重要なことは、高精度な変動記述を可能とする知識の表現法である。多くの知識ベースシステムでは記号のみを対象としているが、音声というパターンについての知識を扱う際には、パターンについての知識をどのように表現するかが問題となる。筆者らのシステムでは知識を「ルール」と「パターン」の２レベルで表現している（図３）。具体的には音素コンテキストによる変動のうち明示的に記述できる部分を変形規則によって表現し、それ以外の変動を音素コンテキストごとの音素片パターンによって表現している。これは、「無声音にはさまれた母音/i/ は無声化する」といった変動の一般的な傾向は明示的な規則で表現できるが、/ita/ 中の/t/ の破裂部と単音節中の/t/ の破裂部の音響的な違いなどは高精度な明示的記述が困難であり、パターンによる個別的な記述とならざるを得ないからである。

また知識の量の問題においては、変動についての知識獲得のわく組みが重要である。現在のシステムでは、変動要因として音素コンテキストを対象としている。日本語の場合音素コンテキストは数千種類あり、これを網羅するためには知識獲得のわく組みをあらかじめ十分に考えておく必要がある。このため変動要因を記述し、変動についての知識を実際の音声データベースから収集し、さらに類型化により知識集約を行っている（図４）。具体的には、第１に網羅的に生成した音素コンテキストをできるだけ多く含む発声テキストを選定した。第２に、知識収集のための音声データベースは自動ラベリングシステムによって作成されたものである。第３として、これらの音素片パターンに対してクラスタリングを行い、４２１３種類の音素コンテキスト別の音素片パターンを約１／４の量に削減したものを用いている。

現在のシステムで用いられている変動記述は、変動についての知識が２つのレベルに広範囲に分散して存在していること、また変動についての知識を大規模に収集しさらにその集約化を行うという知識獲得のわく組みを用いていることなど、これを知識ベースによる変動記述と呼んでよいだろう。他の変動要因についても、知識ベースによる変動記述という考え方が適用できる。たとえばアクセントや発声速度を変動要因とし、アクセントの有

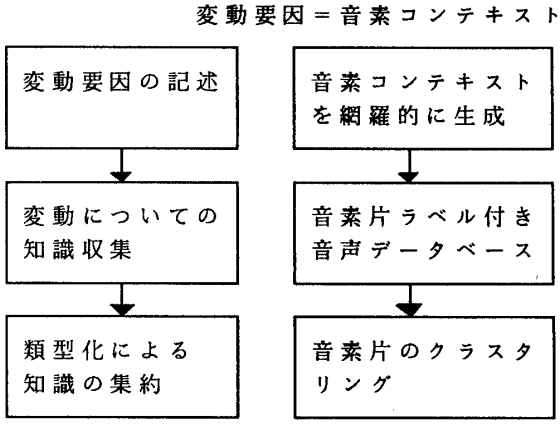

図４　変動についての知識獲得のわく組み

無による区別をつけることや発声速度をいくつかのレベルに量子化することにより、現在の特徴ベースを再構成して変動記述をより精緻化することが可能だろう。

この変動記述というアプローチには、無限の可能性をもつ変動を有限の知識ベースによって記述することによる限界がある。その限界は、変動記述の精緻化による認識性能の向上と知識ベースの量の拡大による認識システムへの負担とのバランスによって決まるものと考えられる。しかしながら、近年の計算機能力の向上、音声データベース整備の進展により、現在のところは変動記述を拡張していくことが有効であると考えられる。

5．まとめ

知識ベースによる音声の変動記述について、単語音声認識システムでの例によって考察した。このような考え方に立つことによって、変動記述を見通しのよいものとすることができよう。

今後は、変動記述を他の要因に拡張するとともに、大規模な記述を可能にする並列マシン上の音声認識システムを考えていく予定である。

謝辞　終わりに日頃ご支援とご討論を頂く、当所知能情報部、中島隆之部長、並びに音声研究室の皆様に深謝致します。

参考文献

(1) 速水、田中、太田：電子情報通信学会論文誌　Vol. J71-D No. 2 pp. 265-273 (1988).
(2) 速水、田中、横山、太田：電総研彙報第49巻第10号 pp. 803-834 (1985).
(3) 田中、速水、太田：日本音響学会誌　第42巻第11号 pp. 860-868 (1986).
(4) Hayamizu, Tanaka, Ohta : IEEE ICASSP-'88 S5.8 pp. 211-214 (1988).
(5) 溝口、田中、福田、辻野、角所：電子情報通信学会論文誌 Vol. J70-D No. 6 pp. 1189-1198 (1987).
(6) 辻野、櫻井、野村、千種、溝口、角所：電子情報通信学会論文誌 Vol. J71-D No. 3 pp. 531-542 (1988).
(7) Zue : Proceedings of the IEEE, Vol. 73 No. 11 pp. 1602-1615 (1985. 11).
(8) Zue and Lamel : IEEE ICASSP'86, 23.2 pp. 1197-1200 (1986).
(9) 嵯峨山：電子情報通信学会 音声研究会 資料 SP87-86 (1987).