大語彙連続音声認識のための読み上げ文コーパスの構築*

5H－10　◎伊藤克亘（電総研）　武田一哉（名古屋大）　竹沢寿幸（ATR 音声翻訳通信研）
松岡達雄（NTT HI研）　鹿野清宏（奈良先端大）

1　はじめに

数万単語を越える語彙を対象とする大語彙連続音声認識は、音声認識技術の応用分野の拡大には極めて重要な課題である。わが国では、要素技術は高水準にあるが、その評価が孤立単語認識による音声制御装置や電話音声応答装置など応用システムに依存して行なわれており、要素技術間の相互比較が極めて困難な状況にある。

このような背景の中で、1995年11月に、情報処理学会音声言語情報処理研究会に「大語彙連続音声認識研究用データベースに関するWG」が発足した。このWGでは、大語彙連続音声認識に含まれる様々な要素技術の性能を迅速かつ厳密に評価しうる基盤を整備することを目的としている。評価基盤を整備することにより、以下の3点の効果が期待される。1) 要素技術間の相互比較が容易になり、先端的な研究領域での技術進歩を加速する。2) 要素技術の共通化により、音声認識技術を応用した製品等の開発の効率化を可能にする。3) 音声認識応用システムの問題点を要素技術に還元することが容易になる。

本稿では、われわれが整備する基盤のうち、最も重要な部分である読み上げ文コーパスの構築について述べる。

2　音声認識評価用読み上げ文コーパス

読み上げ文コーパスとは、音声データ収録のための発声用朗読文のコーパスである。大語彙連続音声の要素技術としては、音響・音声モデルと言語モデルの大きな二つの柱がある。わが国では言語モデル学習用の大規模テキストコーパスの整備が欧米に比べ立ち遅れているため、評価用のコーパスの構築にも学習用のコーパスが確保できることをまず第一に考慮する必要がある。これらの制約から、評価用読み上げ文は「CD-毎日新聞91-94年版」から選択して整備することに決定した[1]。

評価用の読み上げ文の選択には、要素技術の多様な評価が可能なように、i) 語彙の規模 ii) 文長 iii) 文の複雑さ、の3つのパラメータを考慮することにした。語彙の規模を5段階、文長を2段階、文の複雑さを3段階区別しており、全部で30クラスに区別される。この文データは、単独の文から構成されるが、単独の文より広い範囲での言語現象にも対応できるように、4以上14以下の文を含む段落を単位とするデータも用意した。これらのクラスを以下の配分で話者1名当り100文程度のセットを100セット構築することを目標とした。

語彙の規模としては、5000語（中語彙）と20047語（大語彙）を設定し、たとえばMIDというクラスは、中語彙に含まれる形態素のみからなる文を示し、LARGE+は、大語彙に含まれる形態素とそれ以外の1語から構成される文を示す。LARGE++は、大語彙に含まれる形態素とそれ以外の2語から構成される文を示す。文の複雑さは、Lが最も低く、Hが最も高いものとする。語彙クラスについては、たとえばMID+とLARGEの両方の条件を満たす場合もありえるが、そういう場合は、下の表のなるべく上のクラスに含めるようにした。

表1　セット当りの文の内訳

	NORMAL			LONG		
語彙クラス	L	M	H	L	M	H
MID	2	6	2	1	3	1
MID+	2	6	2	1	3	1
LARGE	4	12	4	2	6	2
LARGE+	2	6	2	1	3	1
LARGE++	2	6	2	1	3	1

また、さらに広い範囲での言語現象にも対応できるように、上記のセットとは別に、35文以上からなる記事を単位とするデータも用意した。このセットでも、1名当り100文程度になるように、1名当り3記事を割り当てることにし、全部で5セット構築することを目標とした。

3　読み上げ文コーパスの構築

読み上げ文コーパス構築の手順は、次の通りである。1) テキストの形態素解析。(RWCPのテキストデータベースを利用。)2) 読み上げに適さない文や表現の削除[2]。3) 形態素の頻度リストの作成。4) 文の複雑さを計算するための言語モデルの作成。5) 仮セットの構築（文の分類）。6) 仮セットの検査。7) 本セットの構築。8) 本セットの検査。

この手順の中で大きく問題になるのが、形態素解析である。形態素解析結果については、次の二つの問題がある。i) 現状で、広く共通に利用できる形態素解析結果は、95% 程度の精度であり、解析誤りが相当数ある。ii) 異なる形態素解析システムを適用した場合には、形態素の頻度リストの内容も変わる可能性がある。このうち、i) については、最終的には人間が検査する方針を採用した。また、ii) については、作業量を現実的なものにするという観点から、今回は、単独の形態素解析システムの結果を利用するにとどまった。

読み上げ文コーパスの構築には、頻度リストや言語モデルの構築用には、毎日新聞CD-ROM 91年1月から

Development of a sentence corpus for Large Vocaburary Continuous Speech Recognition, ITOU, K. (ETL) et, al.

94年9月までの東京版記事を利用し、読み上げ文は94年10月から94年12月までの東京版記事から選択した。

[形態素の頻度リスト] RWCの形態素解析結果としては、形態素の表記・原形・品詞名が付与されているので、その三つ組が等しいものだけが同じ形態素であるとして、頻度リストを構築した。全部で形態素は、290,939種類のべ65,347,098個出現していた。被覆率は、中語彙が85.8%、大語彙が95.7%であった。(ちなみに、被覆率が90%になるのは上位8129語、97%になるのは27634語であった。)

[言語モデルの作成] 前処理をおこなった記事データ(段落単位になっている)を、引用句外の句点どうしの間を文と定義して文に分割する。そのように前処理したデータから、大語彙に含まれる20047語以外は未知語としてバックオフバイグラム言語モデルを作成した。言語モデルは、CMUツールキット[3]を利用し、バイグラムのカットオフは2として作成した。バイグラムの種類は2,402,695であった。

[文の分類] 文の長さについては、読みやすさを考慮して、句読点などの記号も含めて39形態素以下とした。また、短すぎる文には、形態素解析誤りや、前処理では排除しきれなかった文ではない表現が多いため、6形態素以上の文を対象とした。同様に、パープレキシティの大きな文も形態素解析誤りしている文が多いと考えられるため、399以下の文だけを対象とした。

文の複雑さは、上記の言語モデルを利用し、文ごとにパープレキシティを計算した。計算には、文頭と文末を示すシンボルを付与し、句点や引用句を示す括弧などの記号も全く取り除かずに計算している。また、句点と文末の間の遷移確率も取り除いていない。

表1に示した割合で分類できるように、文長と複雑さの値を設定する必要があるが、語彙クラスによって文長と複雑さの分布がかなり異なる。具体的には、小さい語彙クラスでは、パープレキシティは小さく文長は短くなる。逆に大きい語彙クラスでは、パープレキシティは大きく、文長は長くなる。したがって、同じ境界値で分類することは困難なので、MIDとLARGEで違う境界値を設定し、2の割合にあったような分類になるように境界値を設定した。その結果、文長は、MIDでは、$5 \leq \mathrm{NORMAL} \leq 19$、$20 \leq \mathrm{LONG} \leq 39$、LARGEでは、$5 \leq \mathrm{NORMAL} \leq 29$、$30 \leq \mathrm{LONG} \leq 39$となった。パープレキシティについては、MIDでは$0 < \mathrm{L} < 40$、$40 \leq \mathrm{M} < 85$、$85 \leq \mathrm{H} < 400$とし、LARGEでは、$0 < \mathrm{L} < 70$、$70 \leq \mathrm{M} < 130$、$130 \leq \mathrm{H} < 400$となった。

段落データに関しては、2のいずれかのクラスの文だけからなる4以上14以下の文からなる段落のみを対象とすることにした。

この分類に基いて、94年10月から12月のデータを分類した。194372文中、131507文が、いずれかのクラスに分類された。人間による検査によって捨てられる文の数も考慮にいれ、各クラス2倍のマージンをとって一セット当り270文+9段落で、122セット作成した。このセットには機械的に読みを付与し、仮セットとした。

[文の検査・読みの付与] 仮セットの読みを人間によって検査し、誤っているものは正しい読みに修正した。この際、形態素解析が明らかに誤っている表現を含む文・段落や、前処理で排除しきれなかった文以外の表現や、内容的に読み上げ作業者に精神的な苦痛を与えるような表現を含む文などは排除した。これを各セットあたり4人の検査者で検査し、最終的に、90文+3(4)段落の読み上げ文セットを150セット(13500文+451段落)と、読み上げセット補充用のコーパス11120文+160段落を構築した。

[読み上げ文セットの諸元] 構築された読み上げ文セットの文当りの平均音韻数は、以下の通りである。

表2 文当りの音韻数

	NORMAL		LONG	
分類	平均	分散	平均	分散
MID群	40.4	14.8	88.3	21.8
LARGE群	63.0	24.5	113.1	22.2

セット全体に出現した形態素は、読みまで考慮して異なるものを数えると20668種類であった。中語彙のうち出現したものは、4940種類であり、大語彙では15256種類であった。中語彙のうち欠けているものは、解析誤りが1/3程度で、後は記号や固有名詞をはじめとする名詞であった。各セットごとに出現する形態素の種類は、平均903で、セット当りの平均形態素数は2114である。

4 今後の予定

このようにして構築した読み上げ文セットをもとに、日本音響学会 音声データベース調査委員会(主査:板橋秀一)において、読み上げ音声を収録し、CD-ROMを作成・配布する予定である。

謝辞

本研究は、情報処理学会音声言語情報処理研究会の「大語彙連続音声認識研究のためのデータベース整備WG」での活動の成果である。委員の皆様の尽力に感謝いたします。また、読み上げ文セットの構築には、日本音響学会音声データベース調査委員会の方々はじめ多くの方々に協力して頂きました。心から感謝いたします。

参考文献

[1] 武田他. 大語彙連続音声認識研究のためのテキストデータ整備. 情処研資, Vol. 96, No. 55, pp. 49–54, 1996.

[2] 伊藤他. 大語彙連続音声認識研究のためのテキストデータ処理. 音響講論, pp. 105–106, 1996/9.

[3] R. Rosenfeld. The CMU Statistical Language Modeling Toolkit and its use in the 1994 ARPA CSR Evaluation. In *Proc. ARPA SLS Technology Workshop*, pp. 47–50, 1995.