音声言語情報処理 11－9
（1996．5．28）

大語彙連続音声認識研究のための テキストデータ整備

武田一哉　名古屋大学工学研究科　〒464-01 名古屋市千種区不老町
伊藤克亘　電子技術総合研究所 音声研究室　〒305 つくば市梅園 1-1-4
松岡 達雄　NTT HI 研究所　〒180 東京都武蔵野市緑町 3-9-11
竹沢寿幸　ATR 音声翻訳通信研究所　〒619-02 京都府相楽郡精華町光台
鹿野清宏　奈良先端科大　〒630-01 生駒市高山町 8916-5

http://www.itakura.nuee.nagoya-u.ac.jp/~takeda/DIC/

情報処理学会・音声言語研究連絡会に発足した「大語彙連続音声認識研究のためのデータベース整備ワーキンググループ」における、テキストデータベースの整備活動について報告する。WG では、音声認識の要素技術の評価に必要な研究資源を研究機関の間で共用可能な形に整備することを目的に発足したものであり、現在言語モデルの構築・評価に用いるためのテキストデータの整備を中心に活動している。本稿では、新聞記事データからテキストデータベース及び評価文を作成する上での問題点と、本WGにおける作業の指針について議論する。

キーワード　大語彙連続音声認識，音声言語処理，言語モデル．テキストコーパス，形態素解析

Developing Text Database for Large Vocabulary Continuous Speech Recognition Technologies

Kazuya TAKEDA　takeda@nuee.nagoya-u.ac.jp
Katsunobu ITO　kito@etl.go.jp
Tatsuo MATSUOKA　matsuoka@splab.hil.ntt.jp
Toshiyuki TAKEZAWA　takezawa@itl.atr.co.jp
Kiyohiro SHIKANO　shikano@is.aist-nara.ac.jp

Recent activities of "Database Developing Working Group for large vocabulary continuous speech recognition research" is reported focusing on developing text database. The main purpose of the WG is to utilize research fundamentals for large vocabulary continuous speech recognition, especially for dictation of news paper reading, including text corpus, standard phoneme models, standard word dictionary and utterances for evaluation. Several issues related on the text processing are discussed as well as the brief summary of the current status of the project.

key words　Continuous speech recognition, spoken language processing, language model, text corpus, morphological analysis.

1 はじめに

文として発声された音声を連続的に認識する、「連続音声認識」技術は、音声認識技術の中心的な課題である。特に数万単語をこえる語彙を扱う大語彙連続音声認識は、日常人間が発声する音声言語のほぼ全てを処理対象とすることから、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。

近年、米国や欧州では、音声／言語のデータやモデルを共通化した上で、大語彙連続音声認識システムの性能比較を行ない、これを通じて個別の要素技術の評価を進めるにより、システムの性能が飛躍的に向上してきた。近年の報告によれば、２万単語の語彙を持つ新聞読みあげ音声の認識を、汎用ワークステーションを用いて実時間の数倍で実行し、単語誤り率１５%程度の認識精度を達成することが、標準的なシステムの性能である。

一方わが国では、要素技術の研究は高水準にあるが、その評価が（例えば、孤立単語認識による音声コントロール装置や、電話音声応答装置に代表される）応用システムに依存して行なわれているのが現状であり、大規模なタスクへの応用は少なく、特に言語モデルの研究の立ち後れが目立っている。また、共通データベースがなく、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、情報処理学会音声言語研究会において、鹿野清宏奈良先端大教授をグループリーダとして、大語彙連続音声認識研究用データベースに関するワーキンググループが昨年１１月に発足した。ワーキンググループ 設立の目的は、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である。特に、欧米に比べて立ち遅れの目立つ、言語モデル学習用の大規模テキストデータベースの整備を、短期間に行なうことが当面の活動目的となる。さらに、音響・音声モデルの研究用に標準的な言語モデルを提供し、言語モデルの研究には標準的な音響・音声モデルを提供することも重要な目的である。これに伴い、大語彙連続音声認識のみならず、音声認識技術全般において、以下の効果が期待される。

- 要素技術間の相互比較が容易になることで、先端的な研究領域における技術進歩が加速される。
- 要素技術が共通化されることで、音声認識技術を応用した製品やサービスの開発を迅速に行なうことが可能となる。
- 音声認識応用システムの実用上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、発足以来ワーキンググループにおいて中心的に取り組んできた、新聞記事を利用した大規模テキスト作成状況を報告するとともに、作業の経緯において明らかとなった、連続音声認識用テキストデータ整備における幾つかの問題を議論する。

2 国外の動向

欧米では新聞の読み上げ文の大語彙連続音声認識（ディクテーション）を通じた技術評価がここ数年さかんに行なわれてきた。

米国では、1992 年頃に DARPA (Defence Advanced Research Projects Agency) 主導のもとに、新聞記事 (Wall Street Journal) を対象とした大語彙連続音声認識の研究が開始された [1]。語彙サイズを 5k、20k に設定し、基本的な評価条件を Hub（車輪の中心）、音響モデル／言語モデルの適応化による評価など、付帯的な評価条件を Spoke（車輪の輻）と呼び、さまざまな評価条件のもと精力的に研究が進められている。1992 年には、5k 語彙／未知語なしの新聞記事読み上げタスクで、16.6%の単語誤り率が報告されている [2]。その後二年間で、誤り率は 1/3 から 1/4 にまで削減された。最近では、語彙制限なしの（20k 語彙に一語以上の未知語が含まれた）タスクで、6.6%の誤り率が報告されている [3]。これらの結果は n-gram といった単純なものでも、言語モデルを導入することによって数万単語の語彙をもつ連続音声が認識可能であることを示している。

一方欧州でも、SQALE プロジェクト [4] において、英語（米語・英語）、仏語、独語に関する連続音声認識の評価を行なっている。このプロジェクトでは、言語モデルや音素モデルの学習データ量を言語間でほぼ均等にして相互比較を容易するとともに、研究機関のあいだで単語の音素辞書を共通して用いることも行なわれている。SQLAE の結果は英語以外の欧米語においても、英語と同様の方法による大規模な連続音声認識が可能なことを示している。

わが国では近年の音声認識の大規模化に対応して、音響学会において連続音声データベースが構築され音素モデル学習用データベースとして広く用いられるにいたっている。しかし連続音声認識用のテキストデータに関しては整備が進んでいないのが現状であり、ARPA やSQALE に匹敵する数千万文のデータを単独の研究機関で整備することには大きな労力が必要である。

3 テキスト整備における問題点

本節ではテキストデータ整備における問題点を挙げ、それらに対し本WG がどのように対処する方針であるかについて述べる。

3.1 コーパスの選択

語義や文体がある程度統一された大量のテキストが利用可能であり、電子的な出版も行なわれていることから、大語彙連続音声認識の対象として新聞記事を取り上げることは現実的な選択と言える。しかし記事は書き言葉であり、記事の読み上げ音声はわれわれが日常話している話し言葉とは異なるものであることに注意する必要がある。すなわち新聞記事を用いて研究を進めることは、大語彙連続音声認識の中でも言語モデルの役割を強調することになる。

新聞記事をもとにしたテキストデータを研究機関での共同利用を目指して整備するには、記事の処理・再配布と同等の作業を行なうことになる。新聞社によっては、研究目的に利用範囲を限定した場合でも記事の処理・再配布には厳しい制限を課しており、利用許諾権の観点から毎日新聞記事を元にテキストデータベースの整備を行なうこととした。後述のとおり、毎日新聞の文字数が経済誌に比べて少ない、ことも明らかになっているが、共同利用というWG の目的を遂行するためには毎日新聞の記事を利用することが最も適当であると判断した。

3.2 記事の選択

新聞記事の中には、当該記事の読み上げを音声認識する場合に言語モデルの効果が及ばないと予想されるものがある。下のスポーツ記事の例では、試合結果の読み方が一様でないため「読み方」まで考慮に入れたテキストを作成することは困難である。

プロ野球 巨人１１－３阪神 巨人、再び貯金「１」

（１６）東京ドーム（８勝８負）
阪神 ０００２００００１＝３
巨人 ０６２１１１００ｘ ＝１１
＜勝＞斉藤 ９勝４負登板１３
＜ｓ＞ 広田 １勝２ｓ登板１４
＜負＞仲田 ８勝５負登板１５
.....
巨人が１４安打を浴びせ圧勝、再び貯金１とした。
巨人は二回、無死一塁で

一方下の人事情報に関する記事は、固有名詞の羅列から構成されており、これらの系列が特に言語的に意味のある系列を構成している訳ではない。

「人事」中小企業信用保険公庫＜１月１１日＞
中小企業信用保険公庫（１１日）総裁 大永勇作
▽退職（総裁）片山石郎＝退
職は１０日付

同様な記事は、アンケートの集計や宝クジの当選番号などにも見られる。WG ではこれらの文を対象から除外する方向で検討しているが、これらの文を全て自動的に選別することは容易ではない。現在、句点「。」もしくはコラム記事の文区切り記号「▲」を含む段落のみをテキストデータの対象とする方針で整備を行なっている。

表 1: SQALE プロジェクトにおける各国の使用テキスト規模 [4] と、日経新聞のデータ [5] との比較。

	WSJ	LeMande	FR DE	日経
総文数（百万文）	37.2	37.7	36	180M/5year
異なり語数（千語）	165	280	650	600/5year
5k Coverage	90.6 %	85.2 %	82.9 %	90.3 %
20k Coverage	97.5 %	94.7 %	90.0 %	97.5 %

5/20k Coverage は頻出上位 n 語が全体の単語数に占める割合。日経新聞では、それぞれ 7/30 k の結果を示している。

3.3 特殊記号等の前処理

記事の中には、様々な記号が使用されている。このうち、○◎●◇◆★▽といった記号は、主として読者の注意を視覚的に引き付けたり区切りの便宜をはかるために使用されており、読み上げにおいてもスキップされることが多いため、記号を取り除いたテキストをデータベース化する方針で検討している。一方「」()【】といった句や読みの挿入に使われる記号の場合、記号のみを取り除いたテキストには部分的に特殊な単語の系列が出現する。例えば下の例では

> 「昨年より売上が一割は増えそう」（販売担当者）と嬉しい悲鳴も聞かれた。

という記事が、括弧を取り除くことで「昨年より売上が一割は増えそう販売担当者と嬉しい悲鳴も聞かれた。」となり、意味の通った文ではなくなってしまう。「聞きとり」よりも「文理解」の要素をより重視するためには、文として意味の通じる系列をより多く学習に利用することが望まれるため、これらに関しては丸括弧の中身だけを削除して「昨年より売上が一割は増えそうと嬉しい悲鳴も聞かれた」と変形したテキストを利用する方針である。

その他、〒%¥など読み方にバリエーションはあるが、記号自体い意味があり読み上げられる記号も存在するが、これらに関しては記号を残す方向で検討している。

3.4 形態素解析

標準的な言語モデルを提供することは WG の重要な目的の 1 つであり、形態素解析に関しては以下の 2 つの方向から検討を進めている。

- 形態素解析のフリーソフトウエアである JUMAN[6] を用いて、WG 独自に毎日新聞記事に対して形態素解析処理を行なう。
- 新情報処理開発機構 (RWCP) により作成が進んでいる、毎日新聞の形態素タグつきコーパスを利用する。

前者は、後述する読みの問題などを考慮にいれた上で音声認識に適した形態素タグを、研究機関が独自に付与することが可能となるような基盤を作成するための作業である。JUMAN はソースコードで提供されるフリーソフトウエアであり、辞書や文法が公開されている唯一の形態素解析システムであることから、WG では JUMAN をベースに作業を行なうとともに、JUMAN のバージョン情報の提供や辞書の配布を行なっていく予定である。

一方後者は、RWCP の研究用データベースの一貫として作成が進められている、形態素タグ付けされた新聞記事データベースを WG の作業に利用する考え方である。当該データベースは、IBM 社の形態素解析システム JMA を用いて解析された記事を、後処理によって RWCP が採用する品詞体系に変換したものであり、一部の記事に関しては人手による修正作業も行なわれている。

これらのことから、RWCP のデータベースには高い解析精度が期待される一方で、JUMAN を用いて新聞記事を解析するためのツールを整備することで、連続音声認識のための品詞体系や解析辞書の研究基盤を整えることが期待される。そこで当面は、RWCP と同一の解析結果を出力するように JUMAN の標準辞書の改良作業を行なう予定である。

3.5 読みの付与

テキストを読み上げた場合、同一の形態素に異なる読みが対応する場合があり、連続音声認識ではこの「読みの揺れ」が問題となる。(例えば、市場(しじょう、いちば)などがその典型的な例である。)この問題は、規則音声合成のためのテキスト処理などで研究されてきた分野であるが、JUMAN では解析において読みを考慮に入れた処理は行なわれておらず、原則として同品詞かつ同表記の形態素には対応する読みは1つだけしか与えられない。

当面形態素解析段階では「読みの揺れ」は考慮にいれず、形態素の音素辞書を複数化するなどでこの問題に対応する方針であるが、読みを考慮にいれた形態素解析を JUMAN を用いて行なう方法を将来的な課題として研究して行く予定である。

3.6 評価データ

大語彙連続音声認識の評価を行なうためには、評価用の音声データも必要である。SQALE プロジェクトの評価用の音声データは、1)文の長さ、2)文のパープレキシティ、3)語彙外単語の出現、の3つの要素を考慮して設計されているが、本 WG でも基本的に同様の考え方で評価用音声の作成を検討している。(実際には、WG で評価用文を選定し、日本音響学会データベース委員会において評価文の読み上げ音声を収録する予定である。)

語彙外単語が出現する場合の性能評価を行なうためには、標準の語彙単語(形態素)セットを決定する必要があり、その方法として語彙数を基準に決定する方法と、被覆率を基準に決定する方法の2つの考え方が提案されている。語彙数によって決定する方法では、例えば出現頻度上位5000単語と2万単語のような基準でいくつかの形態素セットが作成されるのに対して、被服率を基準にする場合には、累積被服率を例えば90%と97%に設定するとことで単語セットが作成される。このようにして作成された単語セットに含まれる単語のみから構成される文と、単語セットに含まれない単語(未知語)を含む文とを一定の割合で評価セットに含めることで、システムの未知語に対する頑健さを評価することが可能となる。

4 作業の進捗状況

上記の問題点を把握するために、日外アソシエーツが販売する94年版の毎日新聞記事1年分の CDROM(データ版)を解析し結果の分析を行なった。以下に結果の概略をまとめる。

4.1 記事データの読みだし

表2には、一年分間の記事に含まれるデータの総量を示した。CDROM には東京版と大阪版のデータが含まれており、一部の記事は東京版の内容と大阪版の内容が同一であるため、WG では東京版の記事に限って処理を進めている。

文単位への分割を、「。」(句点)と「▲」(コラム記事の文区切り記号)により行なった結果、CDROM に納められた記事に含まれる総文数は、表1に示したデータの1/30〜1/40とかなり少ないことが明らかになった。一文に含まれる文字数の分布を東京版通年について調べた結果、38文字であった。400文字をこえる長さの文の数は120であり、記事から文への切り出しはほぼ正しく行なえることが分かる。

4.2 形態素解析

東京版通年の全記事を JUMAN を用いて形態素解析した結果、延べ数で 23 M、異り数で 144k の形態素が得られた。一文当たりの形態素数の分布と、全形態素の累積頻度分布をそれぞれ図1、2に示す。

表2：毎日新聞（９４年）記事のデータ量

	全体			東京版のみ		
	通年	1〜10月	11,12月	通年	1〜10月	11,12月
記事数 (K)	89	74	14	71	58	12
段落数 (K)	560	462	97	469	387	82
文章数 (K)	1145	953	192	946	784	162
文字数 (M)	45	37	7.3	36	30	6.1

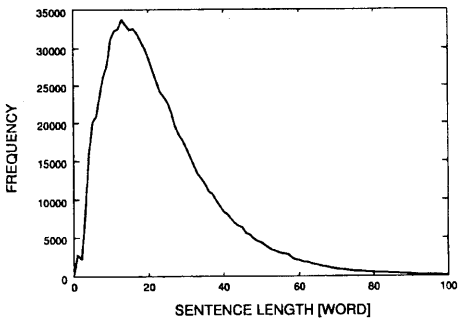

Figure 1: 一文あたりの形態素数の分布

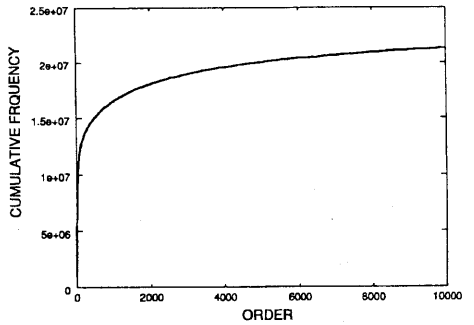

Figure 2: 形態素の累積頻度分布

5 今後の活動予定

今後、WGでは形態素解析の結果を元に出現形態素の頻度リストを作成し、評価用の文章セットの選定を行なう。評価用の文章セットは日本音響学会データベース委員会において、読み上げ音声の形で配布可能な形式に整備される予定である。

さらにWGでは、整備されたテキストから作成した標準的な統計的言語モデルを希望する研究機関に提供するとともに、出現形態素の音素辞書や標準的な音素モデルの公開も行なって行く。

謝辞

新聞記事データの形態素解析処理結果を提供して下さった、名古屋大学工学研究科博士前期課程の小川厚徳君に感謝いたします。

参考文献

[1] D. S. Pallett, "DARPA February 1992 pilot corpus CSR DRY RUN benchmark test results," Proc. DARPA Speech and Natural Language Workshop, pp. 382-385, Feb. 1992

[2] H. Murveit, J. Butzberger, and M. Weintraub, "Performance of SRI's DECIPHER speech recognition system on DARPA's CSR task," Proc. DARPA Speech and Natural Language Workshop, pp. 410-414, Feb. 1992

[3] P. C. Woodland, M. J. F. Gales, D. Pye, and V. Valtchev, "The HTK large vocabulary recognition system for the 1995 ARPA H3 task," Proc. ARPA Speech Recognition Workshop, to be published, Feb. 1996

[4] Lamel, L. Adda-Decker, M. and Gauvain, J.L: "Issues in large vovabulary, multilinghal speech recognition", *Proc. EurosSpeech 95*, Vol.1 pp.185-188, 1995

[5] 大附, 森, 松岡, 古井, 白井："新聞記事を用いた大語彙連続音声認識の検討", 信学技報, Vol. SP95, No.90, pp.63-68, 1995

[6] 松本, 黒橋, 宇津呂, 妙木, 長尾："日本語形態素解析システム JUMAN 説明書 version 2.0", 京都大学工学部長尾研究室, 奈良先端科学技術大学院大学松本研究室 (1994) (http://cactus.aist-nara.ac.jp/staff/matsu/misc/nlc.htlm/)