方言音声データベースの作成と利用に関する研究

A Study on Making and Using Speech Corpora of Dialects

田原広史、江川清、杉藤美代子、板橋秀一
Hirosi TAHARA, Kiyoshi EGAWA, Miyoko SUGITO, Syuichi ITAHASHI

ＪＣＭＤ作成委員会、大阪樟蔭女子大学日本語研究センター内
Committee of Making Japanese Speech Corpora of Major City Dialects,
in Japanese Language Research Center of Osaka Shoin Women's College,
4-2-26 Hishiyanishi, Higashi-Osaka-City 577 JAPAN

キーワード：韻律的特徴、主要都市方言
検索、言語学

Keywords: prosodic features, major-city-dialects, searching, Linguistics

あらまし：この研究は、全国１３主要都市約２５０名の方言音声データをデータベース化するための作業および流通化の方策についての研究である。研究の要点としては、1)「方言音声データベース」を作成すること、2)検索、分析のためのツールを開拓あるいは開発すること、3)当該分野における利用者を開拓し、利用のためのルール作りをおこない、そのルールに基づいて流通化を促進していくこと、以上の３点が柱となっている。

Summary: This Research is based on one of the results of "Integrated Studies on Prosodic Features of Current Japanese Language with Application to Spoken Language Education", funded by "Grant-in-Aid for Scientific Research on Priority Areas by Ministry of Education, Science and Culture", 1989-1992. The results are thousands of recorded DATs(Digital Audio Tapes), which contain vocal sounds of approximately 500 items, such as words, sentences, short-story-readings, a set of Japanese phoneme, numbers, etc. The speakers are selected from 13 Japanese major cities, about 70-100 people per city, 5 old-males, 5 old-females, 5 middle-males, 5 middle-females, 5 young-males, 5 young-females, 10 junior-high-males 10 junior-high-females, 10 elementary-males 10 elementary-females. We are now making speech corpora from this data from 1993, named "Japanese Speech Corpora of Major City Dialects" funded by "Grant-in-aid for Database-making by MESC". This project continues to 1997, and we will make two types of Speech corpora, one is reading of Weather Forecast Report(about 1 minute per speaker) made of Compact Disk, the other is word-reading, made of CD-Rom. Under these conditions, we take the next three aims in our study. 1) making this corpora more complete, 2) developing software to search and analyze the data, 3) increasing the number of the users by advertising our study, and making rules of utilization.

1．研究の背景

・研究の前身

この研究は、重点領域研究「日本語音声における韻律的特徴の実態とその教育に関する総合的研究」（平成元年～４年度、代表者杉藤美代子、以下「日本語音声」）の中で収集された全国各地の方言音声資料を整備（データベース化）し、より効率的な利用、流通を目指すものである。

「日本語音声」期間中に収集された音声資料のうち大きなものとしては、「全国共通項目調査」と「主要都市調査」と呼ばれる二つがある。「全国共通項目調査」では、単語、文、文章、五十音、数字など約1000項目に及ぶ項目を、全国100地点の高年齢層話者各１名についてディジタル録音したものである。「日本語音声」期間中に１９枚のＣＤと３枚のＣＤ-ＲＯＭとして刊行された。「主要都市調査」では、１３主要都市（札幌、弘前、仙台、新潟、名古屋、東京、富山、大阪、高知、広島、福岡、鹿児島、那覇）において、一都市につき、５世代男女計７０名、約500項目についてディジタル録音された。

この資料に関しては、「日本語音声」終了後、平成５年度より新たに成果公開促進費（データベース科研）を受け、現在「日本主要都市方言音声データベース」（同作成委員会）として５年計画でＣＤ、ＣＤ-ＲＯＭ化のための編集作業をおこなっている。平成５、６年度でＣＤ各２枚、計４枚を刊行し、７年度はＣＤ１枚およびＣＤ－ＲＯＭ１枚を刊行の予定である。

・音声データベースをとりまく環境

音声データの編集については、上記「データベース科研」において鋭意作業中であるが、音声そのものを収録するＣＤと異なり、ＣＤ－ＲＯＭ化にあたっては二つの問題が生じた。それは、音声ファイル形式の問題と、検索システムの開発の問題である。ＣＤはＣＤプレーヤがかなり普及しており、一般研究者でも使える状況にあるが、ＣＤ－ＲＯＭはパーソナルコンピュータがなければ分析はおろか聞くことさえできない。「日本語音声」期間中はＣＤ－ＲＯＭの作製はおこなったが実際に音声を聞いた人はほんの一握りに過ぎず、このような研究を進める環境になかった。

ところが、それから４～５年の間に飛躍的にパソコンの普及が進み、ＣＤ－ＲＯＭドライブを標準搭載したパソコンも出回ってきている。研究のための環境は整ってきたといえる。そのような状況の中で、平成７年度重点領域研究「人文科学とコンピュータ」の一公募班として本研究はスタートした。現在、「方言音声データベース」のよりいっそうの整備、より汎用性のある音声データ形式の模索とデータ変換、７年度予算で購入したマッキントッシュによる検索ツールの開発などに取り組んでいる。

・流通に関する試み

また、本作成委員会が所在する大阪樟蔭女子大学日本語研究センターが中心となり「西日本国語国文学データベース研究会」（ＤＢ-Ｗｅｓｔ）を開催している（年２回、平成７年１２月で７回目を迎えた）。この研究会は国語国文学分野におけるデータベースに関連するノウハウの啓蒙、研究、発表をおこなっているのみならず、データベースに関する情報交換の拠点となっており、作成中のデータベースに関しても流通化のためのルール作り、モニター利用の試み等をおこなっている。

2．研究の目的

このような研究背景をふまえ、本研究では研究の目的として次の三つを設定している。

1)「方言音声データベース」そのものを
　より整備されたものにすること。
2)検索、分析のためのツールを開発すること。
3)当該分野（言語学、音声学、国語学、日本

語教育学等）における利用者を開拓し、利用のためのルール作りをおこない、流通化をよりいっそう促進していくこと。

この３点について研究を進めている。進め方は1)2)3)の順にステップアップしていくのではなく、1)2)3)同時進行で進めることが必要である。その理由は、それぞれの段階が密接に関連しており、フィードバックをおこなうことによって、データベースそのものもよくなるし、使用環境も整備されていくと考えるからである。1)に関しては上に述べたとおり、別途データベース科研を受け、編集作業をおこなっているが、製品化（主にＣＤ-ＲＯＭ化）するに当たって、試作品の作成、手直し等の研究を本研究においておこなっている。

・音声データベースの現状

この分野における「音声データベース」は、上記「日本語音声」において作成されたものが始めてであり、きわめて立ち遅れた状況にある。日本語の音声研究・音声教育では、抽象化した音声的特徴を実際の発音と結びつけ、かなりの音声情報を捨て去った形で研究が進められてきた。

現在では、より生の音声に近いものを対象とした、実験音声学、音響音声学の分野が見直されつつあるが、これには、従来の研究に飽きたらず、意欲的に新しい分野に踏み出して行った研究者、教育者たちの努力によるところが大きいことに加え、ハードウェアの面で音声技術、情報工学の飛躍的な発展があったことも忘れることができない。

近年のディジタルオーディオ技術の進歩によって、高品質の録音資料の収集が手軽にできることになったことは言うにおよばず、ＤＡＴ、ＣＤ、ＣＤ-ＲＯＭのような媒体の登場によって、音声が半永久的に劣化しない形で保存でき、さらに進んでパーソナルコンピュータの普及によって検索等も飛躍的に簡単におこなえるようになった。

本研究では、このような時代の流れの中で、高品質の音声データベースを、全国の研究者が容易な形で利用できるようＣＤ、ＣＤ-ＲＯＭの形に整備し、保存、管理、流通の方法を含め、当該分野における音声データベースというものを総合的に研究している。この研究により、今後、当該分野の音声データベースに関して、作成、利用、流通などを含め、水路づけがなされることになると考えている。

近年、マルチメディアを合い言葉に世界のコンピュータ事情は一変つつあるが、日本でも音声自動認識の性能比較を重要な目的として、音声データベースの検討が続けられ、単語音声についてはＪＥＩＤＡ日本語共通音声データやＡＴＲ音声データベースが公開されている。

ただし、これらはいずれも音声情報処理の分野での利用を前提としたものであり、共通語を対象とした「正しい日本語」の「単語読み」である。したがって、音声の韻律的特徴の実態の把握や、教育への応用については、まったく考慮されていない。この研究で扱っている「方言音声データベース」は、日本語方言の韻律的研究を前提に収録されたものであり、その点で一線を画している。

３．研究の現状

この研究で扱う分野は、大きく次の４つに分けられる。

1)検索用テキストデータの入力、整理、データベース化
2)音声信号データの編集、評価、製品化
3)検索用ツールの開拓、開発
4)流通化に関する調査研究

1)検索用テキストデータの入力、整理、データベース化

検索用テキストデータには、「発声内容に関するデータ」（読み、表記、アクセント型など）と「発声者に関するデータ」（話者の年齢、性別、出身地など）の二つがあり、平成７年度までに「発音内容に関するデータ」

１３地点分、総計８４２８項目、「発話者に関するデータ」約１２００人分について、すべての情報の電子化を終えた。この作業には、市販の日本語データベースシステム『桐』（管理工学研究所）を利用している。

音声データと連結して検索作業をおこなうためには、これらの入力されたデータの整備、改良、実際の検索作業に向けての試行錯誤が必要であり、現在はこの作業を中心におこなっている。

2)音声データの編集、評価、製品化

・文章項目のＣＤ化

現在、データベース科研により編集中である。方法は次のとおり。

1)ＤＡＴに録音された音声資料から目的の部分を別のＤＡＴにダビング編集し、すべての収録者（１地点70人から100人）の音声について検聴をおこなう。

2)機械雑音、環境雑音、読み間違えの回数、声質、方言の程度などについて評価をおこなった結果から、最終的に20人について、ＣＤ化する音声をピックアップする。

3)元テープに帰って採用された音声を再度ダビング編集し、ＣＤ作製業者に送る。その際、ＣＤのレーベル、リーフレット、トレイカードのデザインもおこなう。

4)業者はこの音声を一度アナログに変換した上、左右チャンネルのバランス、全体の音量を調整、マスターテープ、原盤を作製し、ＣＤにプレスする。

成果の一部としてこれまでに、天気予報の朗読文章を以下の４枚のＣＤ（音楽用ＣＤ）として発表しており、今年度も引き続き出していく予定である。

『天気予報 Vol.1 富山市・大阪市』
『天気予報 Vol.2 高知市・福岡市』
『天気予報 Vol.3 名古屋市・仙台市』
『天気予報 Vol.4 札幌市・弘前市』

・短文、単語項目のＣＤ－ＲＯＭ化

次の段階として、文章以外の項目、短文、単語などについて、ＣＤ－ＲＯＭとして実用化する計画を立てている。このための作業は平成５年度から進めているが、具体的な手順を以下に示す。

1)上記ＣＤに採用された人についてのみ編集をおこなう。ＤＡＴからディジタル信号のまま、テープの最初からパソコンに取り込みファイル化する（１ファイル３ＭＢ程度）。

2)１人分終わったら、３０程度になったファイルを再度一つづつ読み込み、短文あるいは単語単位に編集し、それぞれファイルとして書き出す（１人分１日６時間で２～３日かかる）。

3)１人分が終了したら光ディスクに書き込む。

この作業の繰り返しである。現在、編集作業はＮＥＣ製のパーソナルコンピュータで、編集用ソフトウェアは、『音声工房』（ＮＴＴアドバンステクノロジ社）を使い、16ビット、16KHｚで編集している。

ＣＤ－ＲＯＭ化にあたっては、このようにして編集してきた数多くの項目（１人の話者につき500発話以上）から、どの項目をデータベース化するかについての検討、個別の音声の評価、ＣＤ-ＲＯＭ内におけるファイル構造の検討などをおこなう必要がある。その成果を、今年度中に、一枚の試作品ＣＤ－ＲＯＭとして発表する予定である。

3)検索用ツールの開拓、開発

検索のためのツール開発に関しては、「日本語音声」期間中に作成されたＣＤ-ＲＯＭ用に開発した検索プログラムがあるが、汎用性がまったくないものである。そこで、このプログラムの設計思想はこのまま生かし、汎用性のあるものに全面的に作りなおす計画を持っている。ただし、もちろんこの目的に合った検索ツールがすでに市販されていれば、それに越したことはないので、ソフトウェアの開拓をおこなっていく必要もある。

パソコンの機種戦争はまだまだ続くものと思われ、単一機種用でしか使えないようなシステムは望ましくない。少し前までは、音声

ファイルと検索用データベースを連動させた形で利用するという目的には、マルチメディア性が高く、インターフェイス、ソフトウェアが充実しているアップル社のＭａｃｉｎｔｏｓｈが有利であった。ところが、ＤＯＳ／Ｖマシン、ＮＥＣといったＷｉｎｄｏｗｓ側も９５の発売と共にいっそうマルチメディア性が高まっていくと考えられ、今後色々なソフトウェアが発売されていく可能性もある。

このような現状を考え、音声データ形式については、Ｍａｃ－ＯＳ、Ｗｉｎｄｏｗｓがにサポートしている形、具体的にはＷＩＮＤＯＷＳ(.ＷＡＶ)形式を採用することにした。この形式であれば、世界中のほとんどすべてのパーソナルコンピュータで再生可能であろう。

今年度の作業は、これまで３年間にわたって編集してきた大量の音声データファイル（約100人分、ファイル数約50000個、容量約３ＧＢにものぼる）をＷＩＮＤＯＷＳ形式に変換する作業を中心におこない、検索ツールの開発については十分な時間がさけなかった。今後は変換されたファイルを用いて具体的な検索ツール作成、実用化していく予定である。

4)流通化に関する調査研究

データベース科研により作成したＣＤをモニター（データベースを使用、評価してくれる人）に配布し、利用方法、利用状況など流通化に関する調査研究をおこなっている。現段階でのモニターの数は１３０名程度であり、平成７年度は使用目的、音質、話者の妥当性などについてのアンケートを実施した。

現段階におけるこのデータベースに関する規程は、以下のような簡単なものである。

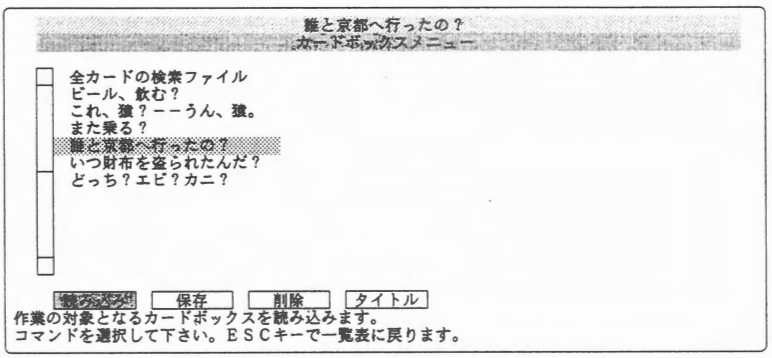
・個人的な使用の範囲を越えてダビングをおこなわないこと。
・研究等に利用した場合は、論文中にその旨を明記すること。
・研究、教育以外の利用はしないこと。
・これらの取り決めを伝えた上で、別の人が利用することは構わないが、モニターアンケート時に報告すること。

現段階における配布媒体はＣＤのみであるが、今後はＣＤ-ＲＯＭに関しても同様の形で調査研究を進めていく予定である。

４．今後の研究

上に述べてきたことを、引き続き進めていく予定である。現状では、まだ研究ではなく作業が中心と言えるかも知れない。研究の目的、計画はほぼ固まっているので、今後はそれぞれの分野を進めながら、適宜フィードバックをおこない、研究を発展させていきたいと思う。

検索結果

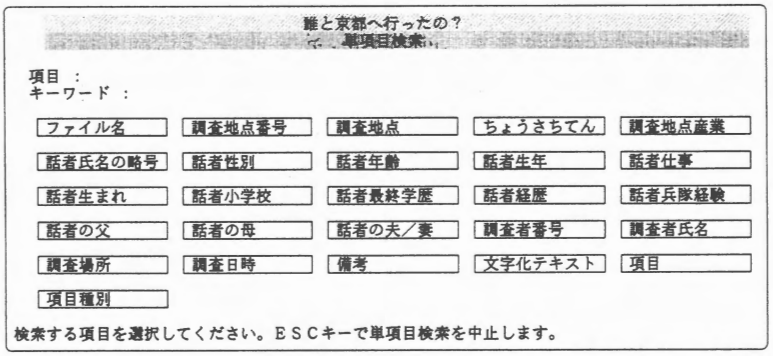


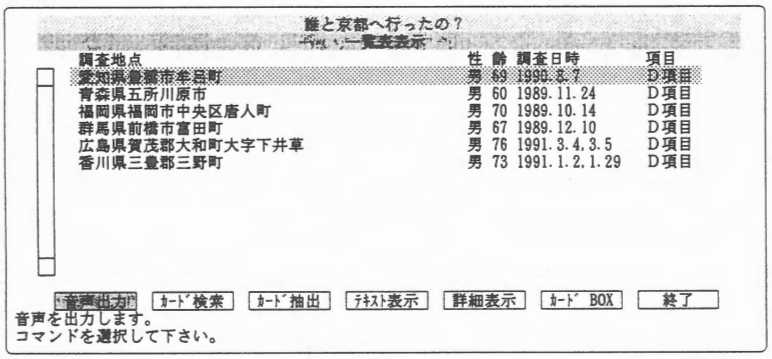
← 基本メニュー

種々の検索項目を登録し、メニュー形式で検索できます。
ここでは「誰と京都へ行ったの？」という単文音声を選択しました。

単項目検索メニュー →

「誰と京都へ行ったの？」という項目からさらに右の話者情報によって絞り込みができます。



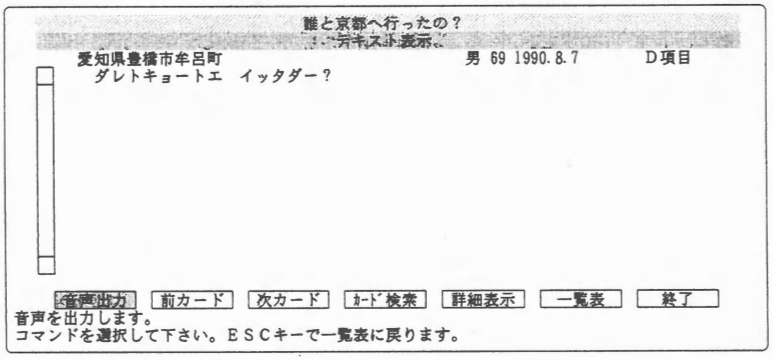


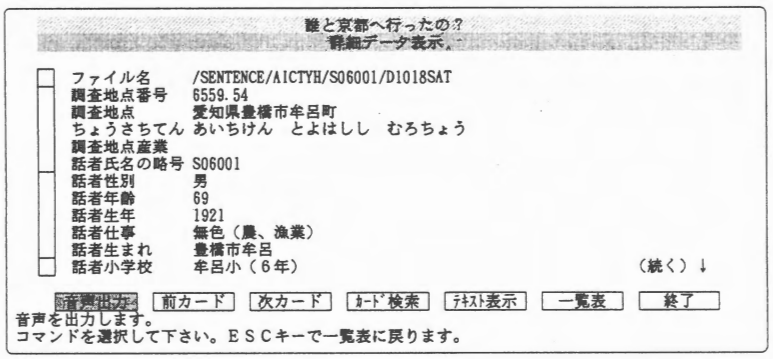
← 選択した結果一覧

ここでは６、７０歳代の男性を選んでみました。
この中からさらに選択し、音声を出力します。
ここでは愛知県の男性を選びます。

テキスト表示画面 →

発話したテキストを表示します。
長い文章だと、このボックスいっぱいに表示されます。
テキストを見ながら試聴することもできます。





← 詳細データ表示画面

話者のさらに詳しい情報、その町の産業、録音した時の状況などを見ることができます。