HOKKAIDO UNIVERSITY

Title	インターネットの音声データベース : その現状と将来
Author(s)	小野, 芳彦
Citation	北海道大学文学研究科紀要, 101, 151-168
Issue Date	2000-09-14
Doc URL	http://hdl.handle.net/2115/33973
Type	bulletin (article)
File Information	101_PL151-168.pdf

Instructions for use

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

北大文学研究科紀要 101 (2000)

インターネットの音声データベース
── その現状と将来 ──

北海道大学大学院文学研究科
言語情報学講座　小　野　芳　彦

1　はじめに

データベースといえば文字によるものであった時代を画して，今は，音声・画像・動画を含むようなマルチメディアのデータベースの時代となった。さらに，マルチメディア情報の流通手段としてインターネットは世の中に欠かせないものとなりつつある。この変革の時代に，音声言語研究のための音声データベースがどのような変貌をとげようとしているのかを概観する。

音声は，即時的に空間を越えて言語を伝達するメディアであり，文字は，時間を隔てて言語を伝達するメディアである。科学技術の発展によって，音声が時間を越えることや文字が空間を越えることが可能になり，両者の特質は共通したものになりつつある。人間の思考を言語として表現し，記録蓄積しながら共用することは，文献という形でもっぱら文字によっていたのであるが，近年の電子技術が非常に手軽に音声の大量記録蓄積を可能にした。本稿で言う「音声データベース」とは，そのような，（言語研究のための）音声の収集・蓄積・再利用を行うために，電子技術的な装置を使っているもののことを言うことにする。

ところで，実音声はビデオ映像と並ぶ大容量の記憶スペースを必要とするメディアであるため，それを大量に保持することには高いコストが必要であった。そのため，大容量でも安価なメモリーがパーソナルコンピュータ用に普及するここ5年くらいより以前では，音声データベースは，磁気テープ・

CD・CD-ROM などにおさめて，利用者にオフラインで利用させるという方法が一般的であった。通信速度が音声の再生速度より十分速くなってきたここ3～4年，オンラインの音声データというものが現れてきた。したがって，インターネット上に音声データベースが隆盛になっているわけではない。大部分は，CD-ROM などで配布される音声データベースの宣伝や試聴の段階である。この先インターネットで，音声(あるいはもっと進んでビデオ映像)の流通が当たり前になってくれば，オンラインの音声データベースがありふれたものになる可能性は高い。

2　音声データベースのモデル

「音声データベース」と「文字画像データベース」はいろいろな意味でパラレルな関係にある。

データの量が大きくなるのは，

- 音声は時間方向に広がる空気の振動であり，振幅と時間の2次元の広がりを持っている，
- 文字画像は平面の縦横に広がる白黒（あるいはグレー）の画像として捉えられる，

からである。ただし，その大きさにはある程度の限界があり，文字画像なら縦または横の次元について，音声ならば時間の次元について，パターンの繰り返しがおこることで,無限の情報を作り出すような仕組みになっている。文字画像も音声も文字数や記録時間に比例する記憶容量が必要になるということであるから，コストも同様に文字数や記録時間に比例する。

計算機はディジタルな記録方式であるから，両方の次元について詳細を記録する必要がある場合は，単位記憶容量が大きくなる。たとえば，文字画像を濃淡付のグレー画像とすれば白黒に比して8倍の記憶容量が必要であるし，音声をステレオにすればモノラルの2倍の記憶容量が必要である。文字画像の縦横の点の密度を倍にすれば，記憶容量は4倍必要だし，音声のサンプル比率をラジオ品質の8ビット8kHzからCD品質の12ビット22kHz

にすれば記憶容量は約4倍必要になる。

言語とは，これらの2次元のパターンの全てを異なるものとして弁別するようなものではなく，人間の認知の過程で，情報を担うコードとして識別され(あるいは，意味のないもの Noise としてすてられ)，相互に弁別されるものである。

- 文字（特に古文書などで）のノイズには，例えば，墨のにじみやかすれのように，文字自身に付随するものと，紙の汚れ（最近ではコピー機やカメラについた埃など)や虫食いのように背景に存在するものがある。
- 音声のノイズには，咳や言いよどみのような，発話者の出すノイズと，周辺の雑音[1]がある。

人間の雑音処理と同程度の自動認識は，文字・音声ともに現時点の技術ではかなり難しい。特に，背景のノイズが広範囲にわたる場合，自動化は絶望的である。大量のデータを蓄積するデータベースにとって，背景雑音を極力押さえたデータソースが求められるのは当然であるが，ノイズがあっても利用せざるを得ない貴重なデータソースが言語データにはある。これらは，当面，人間の認知能力を駆使して解釈をくわえた（annotated）データベースとなる。

認知は言語の構成素に基づいて，トップダウンあるいはボトムアップに行われていると考えられ，概念的に階層に分けられると考えられる。言語資料として音声・文字を扱う研究者にとっては，それらのどの層までを対象とし，収集するかということが問題になる。

- 文字画像については4層くらいが考えられる。筆画，部首，単独文字，句である。
- 音声について想定される認知の階層は，音響素性・音韻素性・音素・フレーズ・会話単位などである。

文字画像で認知の低いレベルを問題にするケースはあまりなく，字体に関する研究が該当するくらいであろう。崩し字字典などでは，最初のほうの筆画が問題にされている。単独文字についての解釈が校註という形で加えられるのがもっともありふれたものである。そして，データベース化されるのは，

この単独文字レベルの認知結果を標準化されたコードに変換したもの（このような形態を全文データベースと言い習わしている）であり，通常，もとの文字画像自体を蓄積するといったことはしない。このような校訂本の形が従来からの研究スタイルであり，電子化されてもそれは変化していないようである。

逆に，音声データベースと称するデータには，研究者の解釈を一切加えていない生の音声が最もよく利用されている。現在のところ，音声というものを抽象化したもの自体はすでに音声でないと認識されているからである。そして，生の音声や認知の中間の階層の音声で検索するという作業が必要な場合[ii]をほとんど想定できないからでもある。文字画像であっても音声であっても，コンピュータ上で認知結果を表現する手段は記号によるし，データの所在を検索するキーの情報を表現する手段も現在のところ同じ記号を利用する以外に方法はない。したがって，音声データのフレーズレベルの認識結果は，文字画像の認識結果と同じ文字のコード列として表現されているのが大部分である。つまり，音声は検索の対象になりえずに，認識結果が全文データベースとして運用されるという形態になっているのが現状である[iii]。

3　音声データベースの検索利用形態

このような全文データベース型のタグを持つ音声データベースに利用者が期待することは，先にのべたノイズを含む場合と同様，解釈の妥当性の確認を利用者が自分の耳で行うことであると思われる。そのために必要となるのは，検索結果からもとの音声を聴取可能にすることである。記録された音声を頭からすべて聞かせるのでは意味がなく，瞬時に目的の位置まで巻き戻しや早送りをしなければならない。このために，もとの音声に挿入されるマーカーをラベルあるいはタグという。これをつける作業が，データベース作成者によってあらかじめ行われるが，音声の場合，非常に根気と労力の要る作業である。ある科研費のプロジェクトでは，すべてを人手で行ったため，収集にかけた年月・費用と同じだけをタグ付けに要したと言う。このため，一

部では音声認識技術を応用したプログラムにより自動・半自動的にタグ挿入が行われている。

ただ，インターネットで標準的に使われている音声データのフォーマットにはタグの取り扱い規定がない。この事情は，文字画像に関してもおなじであり，どこにどの文字が書かれているかのタグに関する規定という概念もない。つまり，インターネット上では，音声や画像は原始的なレベルのメディアでしかなく，その内部の構造化は今後の研究課題である。実質的にはタグと音声データを別々にファイルに格納し，タグを扱うための専用のプログラムを用意して音声データを再生するといった方法が取られているようである[iv]。

最も素朴なタグ付け方法は，アイヌ民族博物館のホームページ[v]で試聴が可能な「上田トシさんのウエペケレ」に見られるような形態である。ローマ字を音素の代用とする表記で単語単位にタグをつけている。以下はその最初の3文で，単語に接辞がついている場合には間に二重ハイフンが区切りとしていれてある。また，単語の意味が下の欄に記述されている。

[1]	1	Mak	Iki	wa	an=an	pe	a=ne	ruwe	ka
		どう	する	して	いる(私)	もの	(私)である	こと	も
	2	a=erampewtek	No	tumun	tum	ta			
		(私)わからない	して	ほこり	の中	で			
	3	Yayesikarunka=an※1	Matkaci						
		気がつく（私）	娘						

別のページにはカタカナに特殊小文字のカタカナ[vi]を加えた文字セットで表記した単語のバージョンもある。これらの単語を本来のタグとして働かせるには，単語あるいは文レベルで音声ファイルが分割されていなければならない。そして，タグをクリックすればその音声が再生される，というようになっていれば，所期の目的は達せられる。尤も，このホームページの目的はサンプルの試聴にあるのであるから，最初の部分(ここでは10個の文で再生時間

29秒）を一度に聞かせるようにさえなっていればよく，そのように構成されている。

全体を聞かせることと，部分を検索して聞かせることはなかなか両立しない。クリントン・アメリカ大統領への告発を記録するデータベース[vii]では，発言の全文をまず検索し，その該当部分のビデオを表示するようになっている。ちなみに〝Japan〟を検索してみると，ヒットは一カ所だけあり，全文検索のリストから四角に囲った部分が検索窓に表示されていた[viii]。

〈OPTION value=391082〉 IT WAS REPEATED ACROSS THIS NATION
〈OPTION value=391082〉 MILLIONS OF TIMES.
〈OPTION value=397011〉 AND THEN HAPPILY, I SURVIVED THAT WAR,
〈OPTION selected value=397011〉 SAW NO COMBAT, WAS ON MY WAY TO JAPAN
〈OPTION value=397011〉 WHEN IT ALL ENDED.
〈OPTION value=410090〉 I'D NEVER HAD A TERRIBLE PROBLEM WITH
〈OPTION value=410090〉 DROPPING THE BOMB, THOUGH THAT'S BEEN A
〈OPTION value=410090〉 TERRIBLE MORAL DILEMMA FOR ME, BECAUSE
〈OPTION value=410090〉 THE ESTIMATES WERE THAT WE WOULD LOSE
〈OPTION value=410090〉 AS MANY AS MY MILLION MEN IN -- AS MANY
〈OPTION value=410090〉
〈OPTION value=423920〉 AS A MILLION MEN IN THAT INVASION.

ここの左に添えられたhtmlのタグが，ビデオの早送りパラメータとなっているようにできているらしい。

全体を聞かせることと，部分を検索して聞かせることを同一のページで実現しているものとしては，子供向けの読み上げ機能付き電子ブックが該当しよう。そこで，カーソルが文字列の上をなぞるだけで再生が始まる仕組みを利用して，これを模擬的に実現できるページを試作してみた。テキストをなぞるというカーソル操作が多少難しい面はあるが，タグと音声の一体感がうまく体現できるものになった。ただし，音声ファイルを瞬時に再生できるようにするために，ページのイメージのなかに全ての音声を埋め込む必要があり，ページ全体を表示するのに多くの時間がかかるという難点がある。現在

のHTMLおよびJAVAの仕様の中での実現は難しいようである。

実際の文章の検索には，日本語の場合，漢字かな混じり文でないと実用性はかなり落ちる。したがって，どのように音声を文書に書き起こすべきかは大事な標準化であるが，広く議論されている様子はない。ただ，研究者の間では全体として統一されたラベリング方法がいくつか提案されている[ix]という。日本語の音声のデータベースでは，当分，このようなかな漢字交じりの書き起こし全文データを検索するものに限られるようである。

4　音声データベースの利用目的

〝データベース〟が一般に公開されたり共用されたりするためには，蓄積された情報にある程度の網羅性が要求される。文字のためのデータベースとして典型である文献・書誌情報を集めたものがデータベースとよばれるためには，たとえば図書館のすべての所蔵文献といったように，網羅性は必須である。このような意味で，同様の網羅性をそなえた〝音声データ資源のための文献データベース〟的なものがあるかというと，そのようなものは組織的には存在しないようである。あればそれを「音声データベース」と呼べないことはないx。しかし，「文献データベース」と同様の構造・利用法であり，「音声」という修飾語をつけるだけの特徴・差異を見ることはできない。

音声自体の網羅性とは何に当たるであろうか。工学的には，音声認識のための素データとして音声を音響的に分析することが研究の中心を占めている。音響音声学の教育的目的も兼ねて音声の音響分析を代行するホームページ[xi]があるほどである。欧米では，IPA音素記号[xii]による少し大きい単位をタグとするものもある。これらは，ただし，一般に公開して広く検索の用に資するという性質のものではない。音声認識のテストデータとして収集・公開されているものの宣伝広報[xiii]といった面が強い。

音声認識を目標とする研究開発に必要なデータベースには，音声の音韻的バリエーションが数多く集められているということが重要であり，その次には話者のバリエーションが求められよう。

日本語は音素数が少なく子音連続が原則としてないため，音韻的バリエーションはかなりすくなくすむ。たとえば，VCV/CVCバランス単語（1542語）[xiv] とかVCVパターンに対して1単語ずつ（計707単語）[xv] という形でまとめられている。

話者のバリエーションという点では，電話の会話音声を数多く採取しているものがある。遠隔地のデータもその場にいかなくて採取できるという利点があるが，採取音声の品質が電話および交換機の性能にしたがって劣悪になるという欠点もある。採取データの使用目的は，電話経由の問い合わせに自動応答をしようという音声応答システムの評価にある。そのため，最も基本的な単語として，英語ではアルファベットと数字のサンプルが収集されている[xvi]。また，その上のレベルとして，代表的な地名や，注文・問い合わせ・指令のための単語などが収集され，それらを使って構築された音声認識会話システムがインターネット経由で宣伝・販売されている[xvii]。日本でも，日本語用音声認識ソフトウエアが実用化され，それを組み込んだICチップ[xviii] や電話応答装置[xix] などが，すでに市場にでている。

語彙より上のレベルで収集されたデータは，コーパスと呼ばれることがある。コーパスと普通に言う場合は，文字データであることが多い。「音声データベース」の同義語として「音声コーパス」や〝Speech Corpora〟もよく使われる。

コーパス収集の目的は，言語の用例を幅広く収集しておき，構文や語構成の理論の実証とすることにある。収集を国レベルで実施し，国語の標準などの議論に役立てて国語政策の元としているフランス・イギリスであるが，それらはまだ文字として行われている。音声レベルで国語政策的コーパスを作ろうとしても，プライバシーや表現の自由などの問題を解決できなければ無理であろうから，当分は，音声コーパスが目標とする文や表現を限定せざるを得ないであろう。しかし，失われてしまうかもしれない言語や方言を音声として保存する手段として，国策レベルの音声コーパスはぜひ必要である。

たとえば，全国規模の方言音声コーパスとして，科学研究費により「全国方言音声データベース[xx]」が作られている。日本全国13主要都市における方

言を対象とし，5世代の男女各2名・1都市あたり計20名の音声を収録している。話者の網羅性という点では，一研究プロジェクトとしては十分すぎるところであろうし，これ以上の網羅性は本当に国を挙げて行わなければ無理であろう。どんなプロジェクトでも，無限に規模を大きくできるというわけではないから，話者のバリエーションを増やすと，比較できる語彙や文を大きくできないというジレンマが生じる。つまり，全国規模となれば，語彙のレベルの網羅性を犠牲にするしかない。ちなみに，このプロジェクトで全国共通に使われている文は，天気予報の読み上げと『桃太郎』の本文であるという。

また，地域研究の一環として方言を音声付で記録しようという試みも一種の音声コーパスといえなくもない。日本の方言についてのホームページは，学問レベルのものから，自分の驚きの体験を単に感想として書いたものまで，幅広いバリエーションがみられる。インターネット上の方言のページへのリンク集としては，㈱ATR音声言語通信研究所客員研究員の山本和英氏のホームページ[xxi]が秀逸であり，全国的な方言ページのリーダー的存在となっている。そのリンクからたどって見つけられた音声を含む方言[xxii]の記録を付表1にあげる。

このなかに小学校のホームページがいくつか見られるが，これらは小学校4年生の国語の授業に関連して，生徒自らが地元の方言を調べたものなどである。少数の学校で始まったものが，連鎖的に広がっていったもので，学校教育の広域化に大いに資している。インターネット・データベースのひとつのモデルケースとみなすことができよう。一つ一つをみれば，そこに収録される語彙に対して，網羅性・一貫性など学問的なデータベースとしての資質を見出すことはできない。しかし，いくつかの地点を集めれば，地域研究の資料となる資質は備わっており，方言として何を取り上げているかという観点からの分析などを通じて，小学生の方言観を探求するなどの社会言語学的な研究が可能かもしれない[xxiii]。それよりも，インターネットにその種の音声が蓄積されていることを知って，自らが同様の音声データを記録していくという行為が日本全国で自己拡大して，ついにはデータベースと呼べるほどに

なっていくことのほうが，あたらしいタイプのマルチメディア・データベースを立ち上げていく契機になっているという点で注目に値しよう。今後は，このような自己発展的なデータベースをいかに有効に利用するかといった技術的研究や，ある程度望みのデータを収集するためにどういった啓発を行っていけばよいかといった政策論・マスコミ論的研究も待たれるところである。

文字ではあまりみかけないコーパスとして，「音声対話コーパス」[xxiv,xxv]や「課題コーパス」と呼ばれるものがある。前者は，音声認識をそなえれば自動化できると考えられる問い合わせや注文応答の具体的シーンを想定し，そのためのデータを集めたものである。これは，自然言語の自動応答システムを設計する際のテストデータなどとして有用である。

5　まとめ

本稿では，音声データベースの利用目的が研究に限られると述べたいわけではない。たとえば，方言の例のように，言語教育の教材として蓄積されている音声であってもよいし，もっと，一般的な知識を広めるための百科事典の項目のようなもの[xxvi]であってもよい。あるいは，目の見えない人のために，講演などの音声を収集・配布するような機関の蓄積音声[xxvii]も「音声データベース」と言いうるであろう。

IT 革命というキャッチフレーズのもと，インターネットの普及を後押ししいている要素として，ビジネスがもてはやされているが，元はといえば研究と娯楽がインターネット推進の原動力であった。娯楽は，ビジネスの対象ともなりうることもあってか，今もインターネットを隆盛にする動機として有効に働いているようであるが，研究が果たしていた役割はかなり小さくなってきたといわざるを得ない。音声というマルチメディアのひとつが，ビジネスあるいは娯楽を動機として蓄積公開されてくることが，これからの趨勢となることはやむをえない動向であろう。本稿で概観したようにまだまだ豊富とはいえないインターネットの音声データベースが，言語研究にむいたデータで構成されることはほとんど期待できない。しかし，量の点では，過去に

研究のために投資した金額が，あと何倍あっても足りないくらい集積されているのである。

これを，どのように検索し利用していくかということについては，過去にテキスト検索がたどって成功をおさめた道が少なくともひとつ見えている。少数ながらインターネットの検索エンジンに，対象をオーディオデータに絞ることが可能なものもある。あとは，インターネットに音声とテキストを結びつけるプロトコルがありさえすれば，現状でもインターネット全体を蓄積領域とする巨大音声データベースが形成できる状態である。

プロトコルのサンプルも画像の世界にそのモデルを見ることができる。サイズを縮小したサムネールという１段の段階でしかないが，巨大なデータの縮約を果たすことができるようになっている。楽曲のレベルでは mpeg3 というデータ圧縮技術がインターネット標準になりつつあるように，携帯電話等で実力が明らかとなった CDMA 方式の音声圧縮技術をインターネット標準とすることも容易であろう。文章を縮訳して少数のキーワードに凝縮することは，検索エンジンでは日常的に行われている。音声の自動認識も半完成の技術ではあるが，それなりのレベルに達している。これらの技術の組み合わせで，本当に近い将来に，音声レベルの yahoo のようなサイトができても不思議はない段階に達していると思われる。

ここに紹介したインターネット上の各ホームページについては，主たるテーマの音声が紙面には再現できない性質のものであるため，そのページイメージを参考に見せることはしていない。本稿の発表時点で，ページの内容やアドレスが変更されている可能性は若干あるが，直接，それらのページを参照されることをお勧めする。

謝　辞

本稿を書くにあたって，以下のホームページのリンク集など多くのページを参考にさせていただいた。まとめを作られた諸先輩・諸兄・諸学会に深く感謝します。

WWW 上の音声言語処理に関する情報（徳島大学工学部北研究室）
http://www-a2k.is.tokushima-u.ac.jp/member/kita/NLP/index-jp.html
「音声・音」研究に関する情報一覧（奈良先端科学技術大学院大学　情報科学研究科鹿野研究室）
http://isw3.aist-nara.ac.jp/IS/Shikano-lab/database/database_list.html

本稿は，平成 11 年 7 月 3 日に科学研究費「北方諸言語の音声データベース作成と言語変容に関する共同研究」の研究会での口頭発表に加筆訂正したものである。発表の機会を下さった北大文学部津曲敏郎教授に感謝します。

i 平均的な雑音を知る目的で，電子協騒音データベースが公開されている。http://www.aist-nara.ac.jp/IS/Shikano-lab/database/internet-resource/speech_db/speech-database-japan.html を参照。

ii オーストラリアの実際の会話例を集めたデータベース Australian National Database of Spoken Language (ANDOSL) http://andosl.anu.edu.au/andosl/では，方言や移民による発音の差異を音響音韻レベルでコード化したものを収集している。アノーテーションについて，一部はプログラムで自動的に行い，重要な部分は研究者の手で行われたと書かれている。

iii もっとも，この事情は文字画像データベースでも同様である。たとえば変体かなのバリエーションを検索しようとしてもそれは難しい。ただ，文字画像と蓄積された記号とに大きな認識上のギャップがないため，直接文字画像を検索したような錯覚を覚えるだけのことである。

iv たとえば，http://winnie.kuis.kyoto-u.ac.jp/taiwa-corpus/doc/node8.html では，書き起こし文書と細切れにした音声ファイルをつなぐリンクファイルから，閲覧のための Html 文書を自動的に作り出すしくみを使っていると記述されている。音声と書き起こし文書とのリンクのために，音声認識技術は使われていないが，会話音声の中の無音期間がおおよそスペースで示されるといった工夫がなされているという。

v http://www.asahi-net.or.jp/~kt9m-ysd/index.html

vi トレロなどである。これらはホームページ記述言語である HTML のフォント縮小機能を使って表現されているので，日本語の表示が可能であれば見ることが出来る。JIS の新しい日本語文字セットに収録されている。

vii http://video.altavista.com/impeach/　2000 年 1 月現在ではすでに存在しない。

viii 前後の文は，窓の横にある▼ボタンを押すと表示される。ポップアップウインドウとかプルダウンメニューと呼ばれる表示法である。

ix http://winnie.kuis.kyoto-u.ac.jp/taiwa-corpus/doc/main.html　板橋秀一（筑波大）

氏による，重点領域研究　音声対話　での模擬対話音声コーパスの解説。

x　しいてあげれば，通信カラオケのオンライン索引が，ポピュラーソングの歌詞という音声データの〝文献データベース〟になっているといえなくもない。

xi　http://sp.cis.iwate-u.ac.jp/sp/paste/indexj.html

xii　ホームページ上ではIPAフォントがユニバーサルに利用できない不便があるため，代用アルファベット記号によるものが多い。たとえば，ミュンヘン大学音声研究所のBavarian Archive for Speech Signalsというホームページhttp://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.htmlでは，ドイツ語の会話音声を収集したものをサンプル公開しているが，そこでは〝Ich mus uber Hannover nach Hamburg fahren.〟が

QIC＋mUs＋　Qy:b6＋　han'o:f6na:x＋　h'ambU6k f'a:r@n.

と音韻表記されている。MS Windowsに採用されているUnicodeではIPA文字が利用可能であるので，いずれこのような代用表現は少なくなっていくと考えられる。

xiii　http://www.atr.co.jp/results/index.htmlでは『自然な発話の認識技術実現のため旅行会話に関する日本語の自然発話の模擬会話データ』を広告している。そこには，次のような書き起こし例が載っていて，自動翻訳が可能になったあかつきのホテルの一会話を描いている。

E：Hello. Front desk.　(1)

EJ：はい，フロントでございます。　(2)

———

J：五〇九{tt}号室ですが，今テレビをつけてみたんですけれど，全然映らないんですね。　(3)

JE：Hi, this is room five o nine.　I turned on the TV right now but the picture doesn't show.　(4)

xiv　http://www.etl.go.jp:8080/etl/onsei/電子技術総合研究所音声研究室のデータ。説明はhttp://www.etl.go.jp:8080/etl/onsei/Misc/etlwd.htmlを参照。

xv　http://winnie.kuis.kyoto-u.ac.jp/京都大学大学院情報学研究科音声メディア研究室のデータ

xvi　テキサス・インスツルメンツ社が最初に収集したもので，TI-20と呼ばれる0から9までの数とその他の語合わせて20語と，TI-26と呼ばれるAからZまでの26語の発音を集めたものを，TI-46コーパスという。http://morph.ldc.upenn.edu/lol/docs/TI46.htmlを参照。

xvii　MITコンピュータサイエンス研究所では，ジュピタープロジェクトやマーキュリープロジェクトという名前で，世界の天候の問い合わせや航空券の予約問い合わせを音声で受けて応答するシステムが研究開発されている。http://www.sls.lcs.mit.edu/sls/whatwedo/applications/jupiter.htmlや同mercury.htmlを参照。

[xviii] 音声認識用のICチップは，すでに家庭用ゲーム機のゲームコントローラーにまで使われている。http://www.nintendo.co.jp/n01/n64/software/nus_p_npgj/sound/index.html などで宣伝されている。

[xix] たとえば，http://www.nec.co.jp/japanese/product/nuivoice/には，不特定の話者の5000語を8回線同時に認識できるという製品が紹介されている。

[xx] http://www.milab.is.tsukuba.ac.jp/corpus/jeida.html ＃5に紹介されている以前からの科研費プロジェクト。http://www.age.ne.jp/x/oswcjlrc/jlrc/jlwork96.htm および http://www.age.ne.jp/x/oswcjlrc/jlrc/jlwork97.htm に報告されている科研費による収集とデータベース化の説明によれば，Windows95とMacintoshの普及により，CD-ROMによる配布が可能になったとある。残念ながらインターネット経由のオンライン検索が可能にはなっていないが，そのための技術的な問題点はほとんど残っていないと思われる。

[xxi] http://www.slt.atr.co.jp/~yamamoto/dialect/ このホームページの英語版はYahoo!のディレクトリ構造の中の，Home>Social Science>Linguistics and Human Languages>Languages>Specific Languages>Japanese>Dialectsのページに唯一のWeb directoryとして紹介されている。

[xxii] 残念ながら，音声を収録していない方言のページのほうが数が多い。東京都，宮城・埼玉・茨城・神奈川・福井・三重・滋賀・奈良・愛媛（1件音声ファイルの再編成中）・長崎・熊本・宮崎県の方言ページからは収録音声を見つけることができなかった。

[xxiii] 付表にあげたホームページの音声を一通り聞いたところでの印象であるが，罵詈雑言を収録するところは比較的多く，関西（大阪府）方言では顕著であった。このような傾向は，個人のホームページで顕著であるが，このような悪い言葉に対して，それを表に出すことに心理的抵抗を覚えるかどうかという観点で，たとえば小学校のホームページの収録語彙を分析することなどが考えられる。

[xxiv] 注xiii参照。

[xxv] http://winnie.kuis.kyoto-u.ac.jp/taiwa-corpus/は，全国5大学で分担の重点領域研究「音声対話コーパス」が作成した音声対話コーパスのページである。それぞれのサンプルを見ることができるが，京都大学収録分には，テキスト音声合成タイプのスケジュール管理システムと人間との擬似会話例が収録されており，現状での機械との会話を彷彿とさせる面白い例となっている。

[xxvi] たとえば，http://babel.uoregon.edu/cas/Languages.htmlには，オレゴン大学で集められた（多分留学生たちの）"Voices of the World" が収められている。
Bulgarian, Cantonese, Danish, French, German, Italian, Indonesian, Japanese, Korean, Mandarin, Norwegian, Russian, Spanish, Thai, Vietnamese（これのみ工事中）のそれぞれの言語について，6つの短い会話文と一つの短い詩が紹介してあり，その言語の筆記（あるいはローマ活字）をみることと音声を聞くことができる。ちなみ

に，日本語の詩は芭蕉の「古池や蛙とびこむ水の音」である。

[xxvii] http://acb.org/Audio/index.html は American Council of the Blind の音声ファイルを紹介するページである。そこには ACB の毎年のコンベンションの講演など，それぞれが 2 時間近くの音声を RealAudio という形式で公開している。RealAudio は，音声データをまるごと手元のコンピュータに転送してから再生をするのではなく，先頭のある程度の分量が手元のコンピュータにたまった段階で，再生を始め，転送と再生を並行して行うようになっている。このため，手元のコンピュータの記憶容量の負担を軽くするとともに，すばやく音声を聞き始められるという利点を持っている。ちなみに，全盲の人にとって，コンピュータの操作はかなり難しいものであり，このホームページにどうやってたどり着き，どうやってこれらの講演を聞くことができるかは（小さな）疑問であるかもしれない。現在は，ホームページのブラウザに自動読み上げプログラムを組み合わせることで，テキストを合成音声で聞くことができるようになっているので，それを活用しているようである。一見，講演自身もこのようなテキストからの自動読み上げを使えばよいかのように錯覚するが，長時間の講演を聴いていて疲れない程度にまで流暢に読み上げるソフトウエアは，残念ながら，現在の技術でまだ可能になってはいない。

地　　域	ホームページ URL（特徴）	音声収録数
北海道(札幌)	//www.bekkoame.ne.jp/~jey/p20/hougen.html	音声例文12文（一部）
北海道(道南函館)	//www.hakodate.or.jp/jr/hogen/default.htm	音声例文10文
北海道 (道南福島町)	//www.hakodate.or.jp/fukushima/shoukou/hougen/default.htm	音声例文５文
青森県(津軽)	//www.venus.dti.ne.jp/~icegin/kouza.html	音声31語（多数のうち）
青森県(津軽)	//www.sphere.ad.jp/aomori/wnn/bunka/tugaru/tsugaruh.html	音声530語弱/例文13文
青森県(津軽)	//www.jomon.ne.jp/~ja7bal/tugaruben.htm 津軽方言の赤とんぼの歌 く・け・か・め・で　にぎやがだばげ　鱈ば喰う	音声例18例/単語19語/川柳５句/例文５文
秋田県(大雄)	//www.rnac.ne.jp/~taiyujhs/hougen/hougen.html	音声２語/１文
秋田県 (西仙北町)	//www.obako.or.jp/futaba/hougen.htm 小学校のホームページでクラス26人の音声	音声26例文
秋田	//www.bekkoame.ne.jp/~kashio/akitaben.htm	音声30語
秋田	//www.ita.tutkie.tut.ac.jp/anbo/omake/akita.html	音声２語
岩手県(紫波町)	//www.rnac.ne.jp/~siwakyoi/namari01.htm	挨拶４例
宮城県(仙南)	//www4.airnet.ne.jp/makuta/hogen/	音声例文４文/会話２例各４往復
山形県(三川)	//www.town.mikawa.yamagata.jp/hougen/hougen/mikawahougen.htm	会話12例
山形県(小国町)	//www2.dewa.or.jp/siranuma/hougen/hougen.html（小学校 HP）	会話６例
山形県	//www.informatics.tuad.ac.jp/net-expo/mapping/hougen/ja/index.html（地域分布図つき）	音声12語採取時の声
山形県(新庄)	//shinjo.dewa.or.jp/shinjcci/hougen.htm	いろはカルタ48枚読上げ
福島県(会津)	//www.city.aizuwakamatsu.fukushima.jp/sc/josai-es/dialect.html（小学校のホームページ）	音声例文26例
栃木県	//www3.justnet.ne.jp/~myattun/hatsuon.html /仮想ニュース読み上げ１例	会話例６例/単語使用例12例/語尾使用例７例
群馬県	//www.gchs.ac.jp/suzuki/gunma/医療人のためのと銘打っている（症状の訴え方の方言差に目覚めて）	単語用例32例/イントネーション８例、単語アクセント10例/慣用句５例ほか
千葉県(房州)	//www.awa.or.jp/home/kamosho/hougen.html (小学校 HP)	音声単語21語
新潟県 (分水町近辺)	//www4.justnet.ne.jp/~sima-1/manabi/hogen/hogen.htm（小学校 HP）	音声単語使用例60例

インターネットの音声データベース

新潟	//www.airlab.cs.ritsumei.ac.jp/~fumiaki/niigata/niigata.html	会話例１例（３往復）
新潟県（長岡）	//www.echigo.ne.jp/~byq06460/koe.htm	会話例１例（４往復）
富山県	//www.cap.or.jp/toyama_ben/	使用例200語強
富山県	//asuna.edu.toyama-u.ac.jp/~fukuno-es/class/4-3/（小学校掲示板） //asuna.edu.toyama-u.ac.jp/~fukuno-es/class/4-3/wano.htm　岐阜県の小学校からの返書	会話２例
石川県（金沢）	//www.city.kanazawa.ishikawa.jp/bunka/custom/hogen/	例文23例/昔話８話
長野県（木曽開田村）	//www.zoone.com/gengo/gengo.html	単語多数
山梨県（大泉村）	//www.vill.oizumi.yamanashi.jp/kindaich/hougen/nabi.html　住民100名の音声	単語多数
静岡県（駿東）	//www.misu.org/dialect/	語尾表現19例/会話例（３往復半）
愛知県（蒲郡市）	//www.tcp-ip.or.jp/~p-nisi/hougen/hougen.htm（小学校 HP）	音声単語使用例42例
愛知県（三河地方）	//www.selab.tutkie.tut.ac.jp/~sasaki/hougen.html	単語用例多数
岐阜県（東白川村）	//www.rd.mmtr.or.jp/~makotos/kotoba.htm	単語用例12例
岐阜県（東濃地方）	//www.eva.hi-ho.ne.jp/ichikawa/tonoben/	音声単語11語
京都	//www.momonga.com/kyoto/kyo-kotoba-j.html	音声使用例２例
京都	//www.kyoto-np.co.jp/kp/ojikoji/kotoba/kotoba.html（市田ひろみさんによる音声例）	例文20文
大阪府（河内）	//www.duonet.co.jp/YAOCCI/KWB.html	例文５例
大阪	//www.remus.dti.ne.jp/~zono/osaka/chapter1.html	例文正５誤３/漫才突っ込み４例
大阪	//www.gulf.or.jp/~walter/gengo.html	状況つき会話例９
大阪（関西）	//www2g.biglobe.ne.jp/~gomma/akotoba.html	挨拶９文/アクセント関係９例/罵詈雑言10例
大阪（関西）	//www.coara.or.jp/~ht/KANSAI/study.html（英語の解説つき）	挨拶・罵詈27例
和歌山県（海南）	//www.bekkoame.ne.jp/~knmc/win21/hougen.htm	音声34単語（対応英語）
兵庫県（太子町）	//office-deai.co.jp/index.htm	会話８例
鳥取県（米子）	//www.hal.ne.jp/tamaru/kouzaframe.htm	１例文/単語使用例200強
島根県（出雲）	//fish.miracle.ne.jp/okuno/izumo.htm	音声例６/42単語
島根県（出雲）	//www.taisya-jhs.taisha.shimane.jp/izumoben.htm	音声41単語

岡山県(倉敷)	//www1.harenet.ne.jp/~two-bow/okayamaben/index.htm	単語使用例300強
岡山県(連島)	//www1.harenet.ne.jp/~two-bow/okayamaben/tsurajimaben/tsuraben.htm	使用例11
岡山	//members.xoom.com/_XOOM/cana_h/index.html	単語14語（多数のうち）
岡山	//www1a.mesh.ne.jp/oknes/guide/okaben.htm	使用例 4
広島	//home.hiroshima-u.ac.jp/shu/hiroshima-ben.html	使用例29
広島	//www.educ.hiroshima-u.ac.jp/~netesc/data/hougen/html/（三択クイズの正解で再生）	使用例10
広島県(備後)	//www.amy.hi-ho.ne.jp/mori-yuji/dialect.htm	使用例13例/アクセント比較 9 組
山口県(美和町)	//www.ymg.urban.ne.jp/home/simohata/hougen.htm（小学校 HP）	用例10
山口県(周東町)	//www.ymg.urban.ne.jp/home/bsyutou/hougen/t_hougen.htm（地域住民のパソコン教室で作成）	使用例41
山口県(下関)	//www.intersky.ne.jp/~haku/hougen.htm	143単語/ 4 文
香川県(大野原)	//www.niji.or.jp/home/takeshi/oonohara/hougen.htm（単語リスト作成中）	6 例
徳島	//www.tokushima-jhs.tokushima.tokushima.jp/kotoba/kotoba.html	使用例94
高知県(土佐)	//www.i－kochi.or.jp/prv/kochi/hogen/	使用例50＋音声挨拶
福岡県(筑後)	//www.kttnet.co.jp/inutukasyo/hougen/hougen.htm（小学校 HP）	用例15
福岡県(博多)	//www.fjct.fit.ac.jp/JP/hakataben.html	用例10
福岡県(博多)	//cad7.nagaokaut.ac.jp/95/kanou/hakata/	用例12
佐賀県(有田町)	//member.nifty.ne.jp/zatsunen/hougen/	用例55
大分県(直入町)	//www.oec-net.or.jp/~n5gtdsty/hogen.htm（小学校 HP）	用例29
鹿児島県(奄美)	//www.minc.ne.jp/setouchi/seto4/seto42.htm	単語53語（5セット）
鹿児島	//www.osumi.or.jp/sakata/hougen/hotitle1.htm	5 例
沖縄県(首里・那覇)	//ryukyu-lang.lib.u-ryukyu.ac.jp/	26単語男女
沖縄県(今帰仁)	//ryukyu-lang.lib.u-ryukyu.ac.jp/	多数