

連載企画—音響学の温故知新—

音声コーパス事始め

—日本を取り巻く状況—*

板橋 秀一 (筑波大名誉教授)**

43.10.Eg, Mq

1. はじめに

最近では音声認識や音声合成に関連する製品やソフトウェアが実際に少しずつ使われるようになってきた。これにはスマートフォンの普及が果たした役割が大きい。2020年開催予定の東京オリンピック・パラリンピックでは音声自動翻訳装置の活躍がかなり期待されている。

音声に関する研究を進めるためには音声データが必要なことは言うまでもない。音声研究に使用する音声データを体系的に集めたものを音声コーパスと呼んでいる。音声データベースの方が分かり易いと思うが、データベースといういわゆるデータベース(検索)システムを指すことがあり、また音声研究の分野ではデータの検索システムよりもデータの集積そのものを指すことが多いため、それを表す「コーパス」がよく使われるようになった。更にテキストコーパスや辞書、専門用語集やこれらを扱う音声・言語処理ツール等を含めて「言語資源」ということもある。

最近の音声コーパスの状況については本誌2015年6月号の解説記事をご参照下さい[1]。筆者は比較的早い時期から音声コーパスの開発に携わる機会があったので、論文にはあまり書かれなかったことを含めて、これまでの経過を振り返りながら、今後の方向についても考察してみたい。

2. 黎明期

日本では電総研の音声データファイルシステムや阪大の音声データベースシステムが挙げられる。これらはいわゆる音声データベースシステムで、コーパスを扱う枠組みであり、それに格納するデー

タの蓄積はこれからというものであった。東北大では単語音声収集システムを作成し、当時としては大規模な単語音声データを集めた。この後更にATRの音声データベースシステムが開発された。世界を見ても音声コーパスに関する本格的な動きが出てきたのは1980年代に入ってからで、その点では日本の出足はそれほど遅くなかったと言えるよう[2]。

3. 電子協/音響学会の委員会

1980年代に入ると(社)日本電子工業振興協会(電子協、現(社)電子情報技術産業協会)の委員会に主査や委員長として関わることになった。これは主に企業からの委員で構成され、「電子協日本語共通音声データ」や「電子協騒音データベース」を構築した。デジタル録音としてPCM録音ができたばかりで、ビデオレコーダとPCMプロセッサを接続して使うものであったが、この組み合わせには相性のようなものがあってうまく動作しないことがあった。ビデオデッキを早大の白井克彦先生の研究室から借用したりして何とか実現できた。電子協のデータ作成では委員の貢献が大きく、収録費用は各委員(企業)が分担した。成果の一部は1986年に東京で開催されたICASSPで筆者が報告した。

1990年に本学会に音声データベース調査研究委員会が設置され、筆者が委員長を務めた(1990–92)。この委員会の発足は現岐阜大(当時電総研)の速水悟先生に負うところが大きい。「研究用連続音声データベース」作成では大学や研究所など28機関が分担して収録にあたったが、小林哲則(早大)・竹澤寿幸(ATR, 現広島市大)各先生や筑波大の大学院生などのお世話になり、費用面では(財)日本情報処理開発協会(JIPDEC)の援助を受けた。

* The dawn of speech corpus: Past and present situation surrounding Japan.

** Shuichi Itahashi (Professor Emeritus, University of Tsukuba) e-mail: itabashi.shuichi.da@alumni.tsukuba.ac.jp

4. 種々のプロジェクト

文部省科研費重点領域研究の最初のプロジェクトの一つであった藤崎博也先生の「音声言語」(1987-90)の際は、私の不慣れから班としての申請書を提出しなかったため、データベース班を作ることができなかった。総括班に入れていただき、収録までは何とかできたが、コーパスとしての整備は別途申請の試験研究(1989-91)で実現できた。

杉藤美代子先生の科研費「日本語音声」(1989-92)では文系と理系合同の総合的な研究プロジェクトで、これにデータベース班として参加した。DAT 製品が出たばかりの頃に DAT を採用するかどうかでかなりの議論があり、文系の研究者が多く、その使用に慣れていないことも心配されたが何とか採用に踏み切り、良質の音声データを収録することができた。方言音声の調査に同行して単に録音するだけではなく、いかに自然な音声の収録を心がけるかを体験する機会となった。これに参加したことで、言語学や音声学の沢山の方と知り合いになることができた。

堂下修司先生の科研費「音声対話」(1993-95)プロジェクトでは白井克彦先生の班に所属し、土屋俊先生(現在、学位授与機構)などに助けられて対話コーパスを作成した。

通産省の RWC プロジェクトは 1991 年から 10 年間行われたが、その間に各種の対話音声コーパスや音楽データベースの作成に関係した。

新聞記事読み上げ音声コーパス(JNAS, 1997)は、小林哲則先生や武田一哉(名大)・竹澤寿幸・伊藤克亘(電総研, 現法政大)・山本幹雄(筑波大)・松岡達雄(NTT)・鹿野清宏(奈良先端大)各先生のお世話になり、文の選定は情報処理学会のワーキンググループにより行われた。これも 30 余の大学や研究機関の協力の賜物である。JNAS の高齢者音声版の SJNAS は鹿野清宏先生の IPA プロジェクトにより構築された。

広瀬啓吉先生(東大)の科研費「韻律と音声処理」(2000-03)プロジェクトでは北澤茂良先生(静岡大)の班の一員として韻律コーパスの作成に加わった。

5. 関連の会議・組織

音声データベース・音声入出力システム評価国際協調委員会(COCOSDA)の活動が 1990 年代初めに始まった[3]。これは音声コーパスの構築と利用を世界的に推進しようというもので、それを基にして 1992 年に言語データコンソーシアム(LDC)がアメリカに、1995 年にはヨーロッパ言語資源協会(ELRA)が誕生した。1998 年には東アジア部会として Oriental COCOSDA の第 1 回会議を筆者が世話人となってつくばで開き、その後アジアの各地で毎年開催されている。日本ではコーパスを扱う組織として言語資源協会(GSK)が 1999 年に設立され、テキストコーパスの配布を中心に活動している。音声コーパスについても類似の組織が必要との認識から故東倉洋一氏の強い後押しにより国立情報学研究所(NII)に音声資源コンソーシアム(SRC)を 2006 年に設置して活動を開始した[4]。SRC は各種のプロジェクトで開発されたコーパスを中心に収集・管理・配布を行っており、前節までに紹介したコーパスの大半は現在でも SRC から入手可能である。2009 年には情報通信研究機構(NICT)により高度言語情報融合フォーラム(ALAGIN)が発足して音声・言語コーパス等の各種言語資源の配布を始めた。

アメリカでは LDC, ヨーロッパでは ELRA が音声・言語コーパスの構築・管理体制を確立している。中国・韓国もそれぞれ類似の組織を立ち上げた。日本では GSK, NII-SRC, ALAGIN, 国立国語研究所などがコーパス事業を進めていて、全体ではかなりの量を扱っている。世界的には音声・言語データの番号付けや音声データの自動集積などが進められている[1]。本稿では紹介できなかったコーパスもかなりあるので、これらの活動をうまく融合して更に発展させることが今後の課題である。

文 献

- [1] 大須賀智子, 山岸順一, 小野順貴, “音声言語資源の現状,” 音響学会誌, 71, 291-297 (2015).
- [2] 板橋秀一, “音声データベース,” 信学論, 70, 433-438 (1987).
- [3] 板橋秀一, “COCOSDA: 音声データベース及び音声入出力システム評価の国際協調委員会について,” 音響学会誌, 53, 949-954 (1997).
- [4] 板橋秀一, 山川仁子, 大須賀智子, “音声言語コーパスの現状と課題,” 信学会誌, 92, 676-681 (2009).