音声言語情報処理 11－7
（1996．5．27）

RWC計画における音声対話データベースの構築

田中 和世＊1 ktanaka@etl.go.jp
速水 悟＊1 hayamizu@etl.go.jp
山下 洋一＊2 yama@ei.sanken.osaka-u.ac.jp
鹿野 清宏＊3 shikano@is.aist-nara.ac.jp
板橋 秀一＊4 itahashi@milab.is.tsukuba.ac.jp
岡 隆一＊5 oka@trc.rwcp.co.jp

＊1電子技術総合研究所 ＊2大阪大学 ＊3奈良先端科学技術大学院大学
＊4筑波大学 ＊5新情報処理開発機構

あらまし：RWC計画の下で行われている音声対話データベースの構築について、その基本設計と現在までに試験的に収集・整備されたデータについて述べる。対話は人間同士1対1の対面対話で、「車の購入」、「海外旅行」を題材としたものであるが、リアルワールド性を考慮して、話者に実際の専門家（業者）を採用している。また、利便性を良くするため、発話単位と音声波形の対応を行い、その書き起しテキスト、および属性を記述したファイルを作成した。

キーワード：音声対話処理、音声データベース、音声データ収集

Design and Data Collection for a Spoken Dialogue Database in the Real World Computing(RWC) Program

Kazuyo TANAKA*1, Satoru HAYAMIZU*1, Yoichi YAMASHITA*2,
Kiyohiro SHIKANO*3, Shuichi ITAHASHI*4, and Ryuichi OKA*5

*1 Electrotechincal Labolatory, *2 Osaka University,
*3 Advanced Institute of Science and Technology - Nara,
*4 University of Tsukuba, *5 Real World Computing Partnership

Abstract: This paper presents a basic design of a spoken dialogue database constructed under the Real World Computing(RWC) Program and also its current status of data collection work. Features of the database are 1) professional agents are employed as the agent-side speakers for producing reality of the dialogues, and 2) transcription texts and corresponding information between the texts and speech waveforms are given for the usablity in analysing the data .

1．はじめに

RWC（リアルワールドコンピューティング）計画では、柔軟なヒューマンマシンインタフェースが主要な研究開発目標の1つであり、その一環として、これらのシステム開発の基盤となる画像、自然言語、音声、マルチモーダルデータなどのデータベース構築を進めている。本稿では、RWC計画の理念に適合したリアルワールド性の強い音声対話データベースの構築を目標として、試験的に収集された音声対話データベースについて紹介する。

音声認識や音声合成など音声言語処理研究において、音声データベースが重要な役割を果たすことはいまや誰もが認めるところであろう。近年、音声対話の研究が盛んになり、これに伴って、音声対話データベースの構築もすでに内外で行われている[1)～3)]。国内においても、日本音響学会連続音声データベース[4)]などに始まり、幾つかの大学、研究機関などで音声対話データの収集が行われている[5)]。

しかし、音声対話データベースが実際のシステム構築にどのように活用できるかという問題は、この数年の経過からみると必ずしも単純ではない。1つはいわゆる模擬対話形式に対する疑問である。また、人間同士と人間－機械対話の違いも問題である。いづれにしても現段階では、収集目的や仕様も様々、入手可能で整備が行き届いたデータ量が少ないことなどが、マンマシン対話システム研究に有効に活用できていない主要な原因であるといえる。

本稿で紹介する音声対話データベースは、人間同士の1対1の目的指向対話で、RWC計画における本格的なデータベース構築に先駆けた試験的なものであるが、従来のものに比べて量的に少なくはない。また、以下のような特徴をもち、利便性をできるだけ考慮している。

1）話題は、車の購入と海外旅行であるが、エージェント側話者は実際の業者（専門家）であり、顧客もできる限り目的意識をもった者としている。

2）各話者の音声をそれぞれL／Rチャネルに録音し、一部は接話型ヘッドセットマイクで収集した。また、VTRで対話状況を録画した。

3）音声データは、書き起しテキスト、そのローマ字表記、および発話単位と波形の対応付けがその属性と共に記述されている。

4）配布用の媒体としてはCD－ROMを使用。データ量は60対話、全体で約10数時間である。

次章以降では、本データベースの具体的仕様、録音条件、データファイルの形式、発話単位へのラベリングなどについて述べるとともに、サンプルの一例を示す。

2．データベース構築の概要

2．1　構築に当っての基本方針

基本的要件は、1）ヒューマンマシン対話インタフェースへの貢献、2）現実の対話としてあり得る対話であるという2点である。さらにデータベース利用に対しての利便性を重視する。今回のデータに限れば、1対1対面の目的指向対話としたが、これによってデータの収集整備上の問題点などの洗い出しを行う意味もある。

2．2　データ収集のための仕様概要

今回のデータ収録に当って決められた仕様概要を以下に示す。実際のデータがどのような条件の下に収録されたかの詳細は第3章で述べる。

（1）基本形態

人一人、1対1、対面、話題を設定した質問応答形式の自由対話とする。

（2）話題候補、場面設定、発話の自由度等

話題は自動車購入、海外旅行計画の2話題。質問者（顧客）－回答者（専門家）のペアの対話で、専門家が計算機に置き換わることが想定できるような設定が望ましい。専門家は実際にその職業についている、ないしは経験者であること、また質問者は一般人、ただし当該話題に興味をもち、相談目的達成の意思があることとした。収録場所は、会義室など静かな部屋。パンフレット、ディスプレイなどの補助用具を使用してもよい。

発話の自由度に関しては、収録者（立会者）は、補助用具使用のために指示語が多くならないこと、（専門家側には）対話時間が長くなり過ぎたり、話題がそれないようにある程度誘導するというインストラクションを与える。その他は基本的に自由とする。

（3）収録機器、環境等

収録場所は上記の通り。マイクロホンは、机上タイプ（弱い指向性）とする。また、一部は（全体の1／4）は接話型マイクでも収録する。各話者あたり1本のマイクで、DATのL／Rに収録する。同時に8ミリビデオで2方向から全体的な画像＋音声を収録する。

（4）話者（被験者）、データ量等

話者については、日本語を母国語とする成人男女性（18歳～60歳くらい）。男女のバランスは半々が望ましい。

対話長は、1対話あたり6分～20分程度。

平成7年度までの収録分は以下のとおり：

a）車の購入（話題）：24対話
　専門家2名、質問者24名（男女）

b）海外旅行計画（話題）：36対話
　専門家2名、質問者36名（男女）

（このうち、a），b）各24対話は、今年度第一四半期に整備完了の見込み。）

2．3　収録データの整備

収録データの整備についての詳細は第4章で述べる。

（1）音声波形

音声はDATに収録、その後、16ビット、16kHzに落としてで計算機ファイルにする。

（2）転記テキスト

かな漢字書き起しテキスト、および発音に忠実なローマ字表記をつける。

（3）波形とテキストの対応づけ

概ね、1発話に対応する単位ごとに波形と対応づける。ただし、発話単位の定義の曖昧さの扱いや音声、感動詞、非言語音声、ノイズなどの区別の必要性から、これらをフラグの形でファイルに記述しておく。

3．データ収集の実際

3．1　収録環境

収録は、この実験に必要な機器、設備等を配置した実験室を使用して行われた。騒音レベルは、A特性で約27．5dB、C特性で約48dB程度である。専門家と顧客は机をはさんで対面して座り、机上にはマイクパンフレットなどが置いてある。

3．2　収録機器

音声は、スタンドマイク(SONYC-38B)で集音し、DATに録音した。なお、一部は、接話型ヘッドセットマイクロホン(Senhaiser HMD410)でも同時に集音した。収録機器の結線（データの流れ）を図1に示す。

3．3　被験者（話者）

（1）専門家（エージェント側）

話題「車の購入」に関しては、外車販売会社勤務の男性1名、および同社経験7年後退職1年経過の女性1名。話題「海外旅行の計画」に関しては、旅行代理店勤務の男性1名および女性1名。なお、全員がこのような音声データ収録の経験はない。

（2）顧客（質問者）

車の購入では、成人男性12名、女性12名の計24名。海外旅行計画では、成人男性18名、女性18名の計36名。

数名の被験者を除き、このような音声データ収録の経験はない。被験者は自ら望んで応募したが、一部を除き、必ずしも現在、実際に車の購入や海外旅行の計画をしてはいない。

3．4　収録の方法

収録の手順は、おおむね以下の通りである。

1）専門家への説明や指示、
2）質問者への説明や指示（これは1度に3人まで、このときカタログ、パンフレットなどを渡す。）
3）上記の質問者のうち、状況設定など考えがまとまった人から順に収録。収録は1回のみ、収録中にスタッフが割り込むことはない。

専門家への指示としては、研究用データの収録であること、（車の購入では）外車販売のショールーム、（旅行計画では）旅行代理店のカウンタでの通常の顧客の応対を想定すること、対話の長さは、特に指定しないが、長くなりすぎる場合には、次の来客があるような想定で話を完結できるように配慮する、マイクは意識しなくてよいが、マイクまでの距離が遠くにならないようにすること、などを指示した。

質問者にも、基本的には同様な指示を与えているが、あらかじめ顧客として話題の目的を決めてから登場すること、実生活と異なる想定も認めることなどの指示を与えた。収録を終えた被験者と待機する被験者との情報交換はない。

4．音声データの整備

4．1　音声データ

音声データは、収録したＤＡＴをD AT-Link+で16kHzダウンサンプリングしたものである。16bitのサンプルが、Ｌチャネル、Ｒチャネル交互に記録され、対話開始から終了までを１ファイルとしている。

4．2　書き起しテキスト

書き起こしデータは、仮名漢字(EUC)で記述され、おおむね、音響学会連続音声データベースの場合に準じている。各発話の先頭には、専門家はB：、質問者にはA：がマークされる。また、冗長語(えー、あのーなど)は［］で囲み、発話中の相手のあいづち、あるいはオーバーラップした発声は｛｝で囲む。なお、オーバーラップしている場合には、｛｝内には直接発声内容を記述せず、番号を記述し、後述する。

この仮名漢字での記述は、発話の言語的意味をとるのを第一義とし、有意味部では、発声の曖昧さなどは記述しない（無意味語ではなるべく発音に沿った表現にしている）。また、書き起こしテキストの読みが曖昧な部分では、＜＞で囲み、ローマ字記述してある。ただし、後述するように、発音に近いローマ字表記はかな漢字テキストとは別に与えられる。

書き起しテキストのサンプルを図２に示す。

4．3　テキスト－波形対応ファイル

このファイルは、対話音声データを解析するのに便利な情報を記述しておく目的で作成した。そのファイル形式は下記の通りである。適当な発話の単位（これは厳密には定義できないが、現時点ではその必要性もない）を１行として、この発話に関する情報を付加して行くことができるようにする。

（１）ファイルの形式

現時点では、下記の７項目が記載されている。

1) 発話seq. No.:　各対話毎に発話に順序番号（１～　）
2) Flag.:　発話の属性を２桁でコード化
3) Speaker Code:　話者を区別するため。
4) Begin time in msec:　msecの単位で整数表示
5) End time in msec:　同上
6) 書き起こしテキスト（かな漢字表記）
7) 上のローマ字表記（電子協音声入出力専門委推奨方式で、発音にできるだけ忠実に）

(改行)

（２）音声波形と書き起しテキストの対応づけ

上記の１発話の単位に相当する音声区間とその書き起しテキストとの対応づけを音声波形上で行い、各区間の時間をテキストファイル中に記録する。これによって、必要とする音声区間の探索を可能とする。

（３）発話の単位と属性（Flag）

発話の単位はおおむね以下の目安でとられており、その属性をフラグを用いて記録しておく。２桁のフラグを、言語フラグと音響フラグに分けて表わす。

a）言語フラグ（１桁目）

１：文と認められるもの。
２：文の一部、単語、文節、句で、文の先頭に位置する
３：文の一部、単語、文節、句で、文の中に位置する
４：文の一部、単語、文節、句で、文の末尾に位置する
５：間投詞、冗長語
６：言い直し
７：非言語

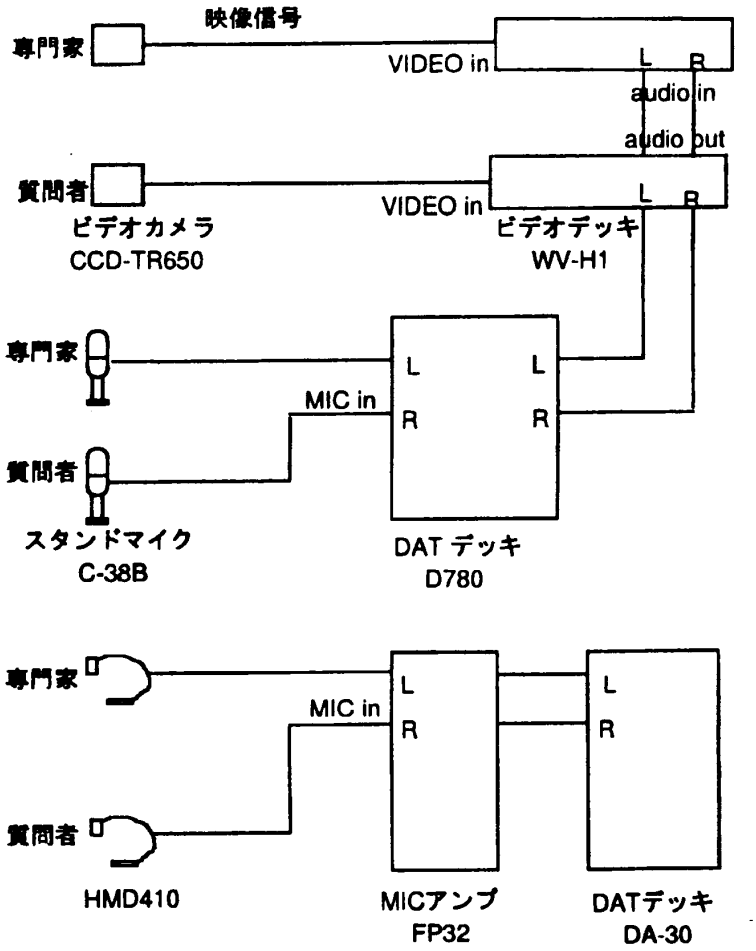

図１　音声対話データ収録機器と結線図

A： そうですね。だいたい二週間前後。
B： ［あ］かなりお長く，｛1｝取れます｛2｝ね。そうですね。二週間，ございましたら，｛3｝ぜひヨーロッパの，そうですね，奥地，［えー］例えば南仏とか。［えーと］そうですね。［ん］ニース，ま，もちろんニースは結構日本<nihon>人多いんですけれども。｛4｝ニースからず，っと南に下ったりして，［えー］そのまま，スペイン入られてアンダルシアに出るとか。その辺りは比較的日本<nihon>人会わないかもしれませんね。
A-1：はい。
A-2：［でし］A-3：はい。
A-4：はい。
A： ［えーと］そうですね。フランスってっと，ランス語ですよね。｛1｝どこに行っても日本<nihon>語は通じる，とは思うんですけれど，｛はい｝ちょっとだけ英語が喋れるんで，そこを活かしたいな，と思うんですよ。｛あー｝ええ。
B-1：基本的にはそうです。
B： ［す］英語圏ですよね。

図２　かな漢字表記書き起しテキストの例

b）音響フラグ（2桁目）

A：ほぼ、単独の発声と認められる。

B：相手の発声が重畳

C：ノイズが重畳

D：極めて曖昧な発声

E：笑い声、咳など

F：ノイズなど

（B,Cが共にある場合は支配的な方を選ぶ）

実際のファイルのサンプル例を図3に示す。

（4）CD－ROM

以上に述べた、音声波形データ、書き起しテキストファイル、テキスト－音声対応ファイルは、最終的にCD－ROMとして焼かれ、配布用媒体となる。

5．おわりに

RWC計画における音声対話データベース構築について紹介した。本文で述べたように、現在は試験的な段階であり、より広範囲の実世界音声データ収集について検討している。また、収集されたデータ自身の解析、統計データの提供についても検討している。

本研究は、新情報処理開発機構（RWCP）に設置されているデータベースワークショップにおける活動の一環として実施されている。同ワークショップでの他のデータベース（／コーパス）を含めて、配布範囲や配布方法についても検討が進められており、研究目的の場合は基本的に入手可能となる見込みである。

最後に、本データベース構築にあたりご尽力頂いた三菱電機（株）石川泰氏、ワークショップで御助力を頂くRWCP豊浦潤氏、また、データ整備についてコメントを頂いた電通大の高木一幸氏、電総研の伊藤克亘氏に深謝します。

参考文献：

1）L. Hirschman, et al, "Multi-site data collection for a spoken language corpus", Proc. ICSLP92, pp.903-906(1992).

2）V.Zue, et al, "The MIT Atis system:Preliminary development, spontaneous speech data collection and performance evaluation," Proc. Eurospeech-91, Paper 18.9 (1991).

3）Proc. ICSLP92, pp.1030-1210(1992-10). および Proc.ICSLP94 pp.1791-1834(1994-9).

4）「音声の知的処理に関する調査研究」報告書、日本情報処理開発協会刊(1992-3)。

5）例えば、（小特集）「出揃った音声データベース」日本音響学会誌４８巻１２号, pp. 876-899(1992-12).； 青野、市川他、「地図課題コーパス」、情報処理学会、音声言語情報処理 3－5（1994－10）；
文部省科学研究費重点領域研究「音声対話」研究成果報告書（平成７年３月）。

```
20/5A/B/20910/22100/えーとー/e-to-
21/2B/B/22100/24660/カマロなんですけれども，こちらにも/kamaronan' desukeredomo, kochiranimo
22/1A/A/23640/24030/はい/hai
23/5D/B/24660/24880/え/e
24/4B/B/24880/26890/お値段表があるんですけれども。/onedan' hyo-gaarun' desukeredomo.
25/1B/A/26280/27040/はい。/hai.
26/1B/B/27360/32570/カマロに四種類ありまして，やはりコンバーチブル，を御希望でいらっしゃいますか。/kamaroniyon' shuruiarimashite, yaharikon' ba-chiburu, ogokibo-deiraqshaimasuka.
27/1B/A/29000/29520/はい。/hai.
```

図３　テキスト－波形対応ファイルの例