# Voice Across Japanデータベース

工藤育男[†,☆] 中間崇夫[†]

90年代に入り音声認識の研究はフィールド実験がさかんに行われるようになり，中でも，電話回線を通したアプリケーションは大規模なシステムが開発されるにいたっている．その結果として，話者の個人差により認識率に大きなちがいがあることは分かっていたが，どのような原因がどの程度，認識率に影響を与えているのかは不明瞭であった．本論文では，電話回線を通じて大規模な話者数を要する音声データベース（VAJデータベース）を構築し，それにより話者の年齢や出身地が音声認識にどのような影響を与えるのかを分析するものである．まず，日本全国から電話回線を通して，8,866人の音声データを収録した．話者の選定には性別，年齢，出身地に関してできる限り日本の人口比に近い比率でサンプリングを行った．このデータベースを用いて，Hidden Markov Model（HMM）をトレーニングして，学習に用いていないデータで評価した結果，年齢が音声認識に最も強く影響を与えていることが明らかになった．20代，30代，40代の声は比較的よく似ているが，10代，あるいは，高齢者の声とは異なることが明らかになった．また，話者の出身地（方言）の影響も，年齢ほど強くはないが，認識率に2%から4%程度の影響を与えていることを確認した．

# The Voice Across Japan Database

Ikuo Kudo[†,☆] and Takao Nakama[†]

In 1990, a lot of speech recognition system have been tested in several fields. Especially, in a telephone application, some huge systems have been developed. As the result, it is known that speaker's age or growing area make influence into error rate of speech recognition system. But the cause was not so clear. Thus this paper reports the influence of speakers' age or growing area into speech recognition through development of a huge-speakers telephone speech database, called VAJ database, which is Japanese Polyphone database. It includes 8,866 speaker's speech data through all the Japan area as possible as portion of Japanese population about gender, age, or growing area. The database, also, has speakers' personal information about gender, age, growing area and so on. Such database did not exist in the world. Through some experiments, it was clear that speaker's age makes influence into speech recognition very strongly. The twenties, thirties, and forties are resemble, however, teenagers or senior age were very different. Also, speakers' growing area, that is, dialects makes 2 ~ 4% influence on speech recognition.

## 1. はじめに

電話回線を利用した音声認識の研究は日米欧を中心に実用化に向けて行われてきている[1)~5)]．米国では音声ダイヤリングサービス（AT&T, NYNEX, SPRINT）が実用化されており，大規模な（150万人規模）人名認識実験（Bellcore）が行われている．カナダのBNR社では株式情報案内サービスを行っている．フランスではミニテルを利用した各種音声認識を利用したサービス（CSLET）が一般顧客向けに行われている．また，デンマークでも天気予報，ニュース，星占いなどを対象とした電話音声情報案内システム（TeleDialog）が開発されている．日本でもANSERシステム（NTT）[6)]の商用化が81年に始まり，90年代には自動内線番号案内システム（KDD）[7)]，電話音声によるホームバンキング（NTTデータ）[8)]などの大規模な実験が試みられてきた．

これら電話音声認識システムには実環境下での課題に耐える頑強さ（Robustness）が要求される．たとえば，(1) 周囲雑音の問題や話者とマイクの方向や距離の問題などの録音環境の問題，(2) 入力となるマイクの種類や電話機内部のアナログ特性の違いに起因する問題，(3) 伝送の歪みや雑音に関する伝送路上で起こる問題，(4) 話者の違いによる問題（性別，年齢，音量，速度，感情），(5) アプリケーションに依存する問

† 株式会社テキサスインスツルメンツ筑波研究開発センター
Texas Instruments Tsukuba Research and Development Center Ltd.
☆ 現在，ジャストシステム
Presently with Justsystem

題（単語数，リジェクト能力），(6) 会話処理上の問題（不要語，対話制御上の問題）などへの対応が要求される．現在も個々の要素技術の改良が継続され，大規模なフィールド実験が行われてきているが，評価を行うのに十分なデータベースが揃っていない．そのため，実用上重要となる日本全国の話者にも耐えるものかどうかについての十分な報告例はない．本報告では電話回線を通じた頑強な音声認識システムを開発するために必要な基礎データを，多くの話者数を有する電話音声データベースを構築することにより探ろうとするものである．

従来の日本語の音声データベース[9)~11)]としては，組織的に大規模に開発されたATRの音声データベース[12)]や音響学会で作成されたCD-ROM[13)]などが公開されているが，収録されている話者がアナウンサーやナレーターなどの発声のプロの音声データであったり，一般の話者数が上記の目的を満たすほど十分含まれていないという問題がある．また，マイクロホンも実験室環境で用いられる高性能マイクと通常の電話のマイクでは特性が大きく異なる点や電話線における帯域の問題（300 Hz～3.4 kHz）により，電話回線を利用した音声認識システムを構築するためには実環境で用いられている機器を利用して音声データを収録する必要が出てくる．

本論文では以下の3つのことに留意して音声データの収録を行った．1つは，話者の年齢や出身地の情報を有し，後の分析に役立てること．2つ目は，男女比，年齢構成，出身地の比率をできるだけ日本の人口構成比に近づけ，統計的にも信頼性の高いものを目指すこと．3つ目は従来のデータベースより1桁以上多い話者数（約1万人）を要すること．以上のことを目標とし，データの収録に約1年（93.10～94.12），データの整理に2年あまりをかけて，8,866人のデータベースを完成させた．このような日本の人口に比例してサンプリングを行った音声データベースはいまだ公表されていない．

このデータベースを利用することにより，電話回線を通した不特定音声認識で起こる話者の違いによる問題，すなわち，話者の年齢，および，出身地が音声認識にどのような影響を与えるのか調べた．その結果，年齢が認識率に大きな影響を及ぼしていること，また出身地の影響，地域性の影響も数%あることが実験の結果判明した．

## 2. 音声データの収集

### 2.1 海外の電話音声収録動向

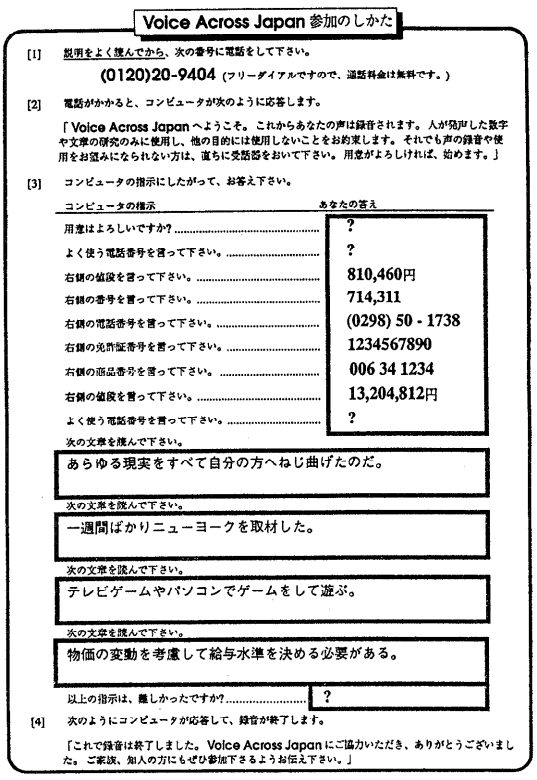
電話回線を利用して多数の話者の声を収録する国際的なプロジェクトとしてPOLYPHONEプロジェクト[14)]がある．これはCOCOSDA[11),14)]で各国別にそれぞれの言語で，3分程度の発話を5,000人程度集めようというものである．本データベース[15)~19)]も日本におけるPOLYPHONEプロジェクトとして位置づけられている．

データの収集方法は各国独自の方法が用いられている．米国ではボランティアベースで参加者を募っている．オランダ[20)]やデンマーク[21)]ではダイレクトメールの業者を利用している．スペイン[22)]では新聞紙上で景品をつけたキャンペーンを行っている．データの収集には各国の国民性が現れており，妥当な方法が選ばれたものと思われる．日本では，当初ボランティアベースで参加者を募集したが，配布資料あたりの応募数の割合は1.3%と芳しくなかった．5,000人の収録だけで3年以上かかる計算となった．そこで，テレホンカードや図書券，ビール券など500円程度のプレゼントをつけることにより参加者を募った．その結果応募の割合は14.2%前後と急増し，短期間での目標達成が可能となり，しかも，経費も半分に削減できた．

### 2.2 音声データ収録上の課題

電話を介してデータを収集する場合，2つの大きな検討課題がある．1つは話者のサンプリングの問題である．音声データの収録のためには，発声者を探さなくてはならないが，ランダムに集めればよいというものではない．性別，年齢，出身地などをできるだけ母集団の日本人口に比例してサンプリングすることができれば，統計的に信頼性の高いデータベースになる．また，性別，年齢，出身地などの話者の個人情報も音声データと一緒に集められれば，その後の分析に役立つ．欧米では個人情報の収集は，プライバシーの問題から集めにくいので，専門の言語学者が音声を聞き，話者の出身地，年齢，性別，学歴などの情報を付与することが多い．したがって話者から直接得た個人情報を持つデータベースは世界的にも珍しいものである．

もう1つの問題は，収録した音声の品質に関する問題である．電話回線を通してデータを収録する場合には，試験者が被試験者のそばにいてデータを収録する場合とは異なり，話者の発声環境や発話態度をコントロールできない．そのために収録データに背景雑音が入っていたり，笑い声などが混じっていることもある．このようなデータは音声認識の学習用データには不適



Voice Across Japan 参加のしかた

[1] 説明をよく読んでから、次の番号に電話をして下さい。

(0120)20-9404 (フリーダイアルですので、通話料金は無料です。)

[2] 電話がかかると、コンピュータが次のように応答します。

「Voice Across Japan へようこそ。これからあなたの声は録音されます。人が発声した数字や文章の研究のみに使用し、他の目的には使用しないことをお約束します。それでも声の録音や使用をお望みにならない方は、直ちに受話器をおいて下さい。用意がよろしければ、始めます。」

[3] コンピュータの指示にしたがって、お答え下さい。

| コンピュータの指示 | あなたの答え |
|---|---|
| 用意はよろしいですか? | ? |
| よく使う電話番号を言って下さい。 | ? |
| 右側の値段を言って下さい。 | 810,460円 |
| 右側の番号を言って下さい。 | 714,311 |
| 右側の電話番号を言って下さい。 | (0298) 50 - 1738 |
| 右側の免許証番号を言って下さい。 | 1234567890 |
| 右側の商品番号を言って下さい。 | 006 34 1234 |
| 右側の値段を言って下さい。 | 13,204,812円 |
| よく使う電話番号を言って下さい。 | ? |

次の文章を読んで下さい。

あらゆる現実をすべて自分の方へねじ曲げたのだ。

次の文章を読んで下さい。

一週間ばかりニューヨークを取材した。

次の文章を読んで下さい。

テレビゲームやパソコンでゲームをして遊ぶ。

次の文章を読んで下さい。

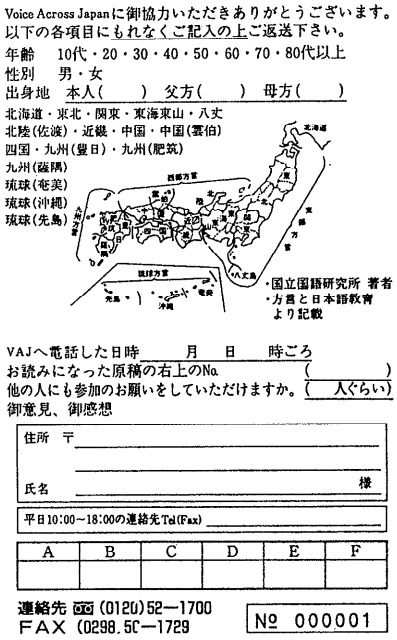
物価の変動を考慮して給与水準を決める必要がある。

以上の指示は、難しかったですか? ?

[4] 次のようにコンピュータが応答して、録音が終了します。

「これで録音は終了しました。Voice Across Japan にご協力いただき、ありがとうございました。ご家族、知人の方にもぜひ参加下さるようお伝え下さい。」

図1 発話内容（セッションシート）
Fig. 1 A VAJ session sheet.



Voice Across Japan に御協力いただきありがとうございます。
以下の各項目にもれなくご記入の上ご返送下さい。
年齢 10代・20・30・40・50・60・70・80代以上
性別 男・女
出身地 本人( ) 父方( ) 母方( )
北海道・東北・関東・東海東山・八丈
北陸(佐渡)・近畿・中国・中国(雲伯)
四国・九州(豊日)・九州(肥筑)
九州(薩隅)
琉球(奄美)
琉球(沖縄)
琉球(先島)

・国立国語研究所 著者
・方言と日本語教育 より記載

VAJへ電話した日時 月 日 時ごろ
お読みになった原稿の右上のNo. ( )
他の人にも参加のお願いをしていただけますか。( 人ぐらい)
御意見、御感想

住所 〒
氏名 様
平日10:00～18:00の連絡先Tel(Fax)

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| | | | | | |

連絡先 TEL (0120)52—1700
FAX (0298.5C—1729
№ 000001

図2 返信用はがき
Fig. 2 A reply card.

であるので，対策を講じる必要がでてくる．本論文では，2.3 節で述べるような方法をとることにより，背景雑音を含む割合を 25.6%から 14.3%までに減少させることに成功した．

### 2.3 本プロジェクトにおける音声データ収集方法

以下のような手順で収録を行った．

(1) まず，セッションシートと呼ばれる発声内容を記した用紙（図 1）と返信用のはがき（図 2）を参加者に配布する．

(2) 参加者は，24 時間受付可能なフリーダイヤルに電話して，配布されたセッションシートの内容（約 3 分程度）をガイドの声に従って読み上げる．発声された音声データは，自動的に計算機にファイリング[23]される．参加者は配布されたはがきに必要事項を書き込み返信する．

(3) 収録データの品質に問題があるかないかをチェックし，問題の多い参加者には再収録を依頼する．

(4) 返信されたはがきに書かれている話者情報をデータベースに入力し，サンプリング状況を集計しながら次の資料の配布を行う．

なお，(3) と (4) の具体的な作業については 3.1 節のデータベース化のところで述べる．

### 2.4 発声内容

特定の人の音声データを集める場合には，1 人の人間に長い時間発声してもらうことも可能であるが，不特定多数の音声を集めるためには，短時間にできるだけ効率的に音声データを集める必要がある．本論文では，図 1 に示すようなセッションシートと呼ばれる読み上げ内容を提示することにより音声データを収録する．そのために発声内容の決め方と作り方が重要な問題となる．

(1) 使用頻度の高い表現の収集：電話番号，値段，商品番号などの数字，および，「はい」「いいえ」などの応答表現などは，電話音声認識で使用頻度の高い表現であると考えられる．これらの表現を集めておくと，小語彙向きではあるが高い認識性能が期待されるワードモデルを構築することが可能となる．

(2) 音韻バランス文：少ない発声でもいろいろな音韻を集めるようにするために人工的に作られた文が音韻バランス文である．音韻バランス文はある文の集合から音韻数を減らさずに冗長な音韻を含む文を除いた文の集合体を指す．発声時間長を考えて 1 文の長さを考慮したり，意味的にある程度通じる文章を作り出したものである．Bi-phone の音韻バランスとして ATR603 文[12]が知られているが，最近の不特定音声

認識システムでは Tri-phone が用いられているようになってきているので，4,320 種類の Tri-phone を含む音韻バランス文（903 文からなる）を作成した．外来語から来る音韻について，ATR バランス文よりカバー範囲が広い特色を持つ[17]．

(3) セッションシートの管理：図 1 に示すようにセッションシートには 14 項目の質問文があるが，配布されるセッションシートはすべて異なる内容となっており，通し番号で管理されている．数字に関しては，乱数を用いて各数字の出現頻度が均一になっている．テキスト文の 4 項目は音韻バランス文，903 文のうちの 4 文がランダムに選ばれている．

(4) 発声状況のフィードバック：配布数に対する応答率は 14%ぐらいなので，配布したセッションシートのすべてが発声されるわけではない．そこで収録後の各バランス文の出現頻度が均一になるようにする必要がある．各音韻バランス文に対する発声頻度を集計し，新たに配布するセッションシートに発声頻度の低い文を選ぶようにした．

### 2.5 発声者に対する注意事項

発声者に対する注意事項としては，

(1) 参加者は 15 歳以上の日本語を母国語とする人．15 歳以下では漢字が読めないため年齢制限を設けた．

(2) 音声認識の学習用データとして使用するため，静かな部屋から電話すること．

(3) 2 秒以上のポーズを開けると，自動的に次の項目に移ることがあるが，その場合でも，継続して最後まで完了させること．

(4) セッションシートの内容は 1 人 1 人違うので，他の人の使ったシートは使わないこと．

(5) データの収録に失敗した場合には，再度収録をお願いすることがあるということである．

### 2.6 話者のサンプリング

#### 2.6.1 個人情報の収集と話者のサンプリング

返信用のはがき（図 2）には，氏名，住所，電話番号，性別，年齢，出身地，読み上げたセッションシートの番号，電話した日時，他の参加者を紹介できるかどうかの有無が記述されている．出身地の分類については，国立国語研究所の方言分類[24]を利用した．あらかじめ，1 万人を抽出することを前提として，男女別に，各地域のどの年齢層に何人集めるべきかを設定し，収録状況を管理し，人口比率に近づけた．話者数の少ない地域に重点的に資料を配布し，目標達成した地域では収録を打ち切る．このようにして，サンプリングが適切にいくよう試みた．

表 1 VAJ データベース中の話者の性別，年齢分布
Table 1 The speakers' gender and age in VAJ Database.

| 年齢 | 男性 | % | 女性 | % | 合計 | % |
|---|---|---|---|---|---|---|
| 10代 | 272 | 3.09 | 498 | 5.65 | 770 | 8.74 |
| 20代 | 948 | 10.76 | 1621 | 18.4 | 2569 | 29.16 |
| 30代 | 1115 | 12.66 | 1366 | 15.51 | 2481 | 28.16 |
| 40代 | 446 | 5.06 | 658 | 7.47 | 1104 | 12.53 |
| 50代 | 253 | 2.87 | 507 | 5.76 | 760 | 8.63 |
| 60代 | 134 | 1.52 | 166 | 1.88 | 300 | 3.41 |
| 70代 | 28 | 0.32 | 40 | 0.45 | 68 | 0.77 |
| 80以上 | 7 | 0.08 | 7 | 0.08 | 14 | 0.16 |
| 不明 | 380 | 4.31 | 363 | 4.12 | 743 | 8.43 |
| 合計 | 3583 | 40.67 | 5226 | 59.33 | 8809 | 100 |

#### 2.6.2 話者のサンプリング結果

収録した結果について述べる．

（A）性別

収録した話者について，男性が 3,583 人，女性が 5,226 人，15 歳以下の子供 57 人，合計 8,866 人である．子供を除く 8809 人を対象とすると，男性が 40.7%，女性が 59.3%ととなり，日本の人口比（男性 49%，女性 51%）[25]と比較すると，女性の比率が高いことが分かる．

（B）年齢分布

年齢の分布に関しては，15 歳以上の人口を 100 として示すと**表 1** のようになる．20 代，30 代の割合が多く，他の世代の割合が低い．特に，高齢者の割合が極端に低いことが分かる．この理由としては，高齢者が機械を相手に電話することを嫌うこと，また，配布した資料の文字が高齢者には小さすぎたことなどが考えられる．しかし，高齢者に関する音声データそのものが現在の日本のデータベースにはないことを考えると，これだけの話者数（60 歳以上 382 人）を集めたことは貴重な意味を持っているといえる．

（C）出身地の分布

出身地に関しては，収録すべき目標値と実績を日本地図上（**図 3**）に示した．図 3 の右側の値が収録すべき目標値で，左側が収録実績である．人口の密集している地域についてはかなりカバーしている．八丈，奄美，先島はもともと人口の少ないところで，必ずしも十分カバーしているとはいえないが，これら人口の少ないエリアを除いてほぼ日本全国をカバーしている．なお，話者の分類についてはすべて自己申告によるデータに基づいて付与した．1,157 人については出身地に関する情報がなかった．図 3 の右下に “Unknown” という形で表示した．その隣にある “ETC” は，出身地が想定外の地域（たとえば，満州など），子供のころ転勤が多く，出身地が複数あると申告している場合の 5 件である．外国人と思われる通話は，注意書きによりほとんどかかってこなかった．今回のデータベースには

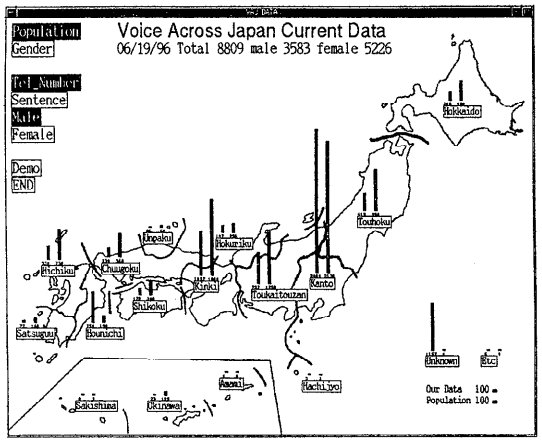



図3 出身地の分布
Fig. 3 Collected data in each area.

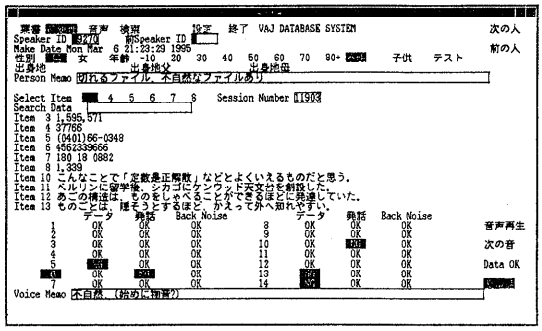



図4 前処理作業
Fig. 4 Pre-checking.

含まれていない．

## 3. データの検定

### 3.1 データベース化作業

#### 3.1.1 前処理作業

前処理作業はデータの収録と並行して行う．**図4**に示すように，収録した音声データの状態を確認する．発話されたすべてのファイルに対して，以下の項目のチエックを行う．

(1) ファイリング（図4データの項目）：自動ファイリングがうまくいったか否か？
(2) 発話態度（図4発話の項目）：発話態度（ふざける，笑う，読み方が不自然など）に問題がないかどうか？
(3) 背景雑音（図4 Back Noiseの項目）：背景雑音（他の人の声，テレビ，音楽，電話が鳴るなど）が音声と重なっていないかどうか？

以上の3つの観点から14発話についてチエックし，4発話以上問題となるファイルがある場合は再収録をお願いした．半数以上の人が再度の収録に協力してくれたので，データの品質をあげることができた．背景雑音を含むファイル数の割合を25.6%から14.3%までに減少させることができた（収録結果のフィードバック効果[17]）．

話者情報の管理として，返信はがきの情報と発話内容を対応を確認のうえ対応づける．セッションシートの番号は通し番号なので，これで管理する．また，前回の収録に失敗し再度協力を要請された話者が再度録音する可能性があるので，以前の話者IDとの関連づけを行う．なお，データベースには話者のプライバシーを考慮して，氏名，住所，電話番号，FAXなどはいっさい電子化していない．個人の識別はすべて話者IDで管理している．

その他，問題となる内容があれば，メモ（図4 Voice Memoの欄）に記入する．

#### 3.1.2 書き起こし

すべてのファイルについて内容の書き起こし作業（Transcription）を行う．書き起こしに際しては，できるだけ発声に近い書き起こしを行った．たとえば，訛っている場合は，訛ったままの音として書き起こす．たとえば，「0」なら“zero（ゼロ）”，“dero（デロ）”，“jero（ジェロ）”という具合である．

書き起こし方法としては，訓令式やIPA[26]の方法があるが，IPAの表記法は大量のデータへの作業は難しい点，訓令式での書き起こしは英米人に難読であるという理由から，ヘボン式を拡張した表記法（**表2**）を用いた．

書き起こしの際，**図5**に示すようなチエックを行う．

(1) 前後の雑音部分だけ取り除けば，音声区間は良好なデータか否か（図5のN）．
(2) 音声区間に雑音が被さっており，切り出し不能である（図5のNX）．
(3) 発話前後に言い直し区間がある（図5のR）．
(4) 発話区間に言い直しがあり，切り出し不能（図5のRX）．
(5) データが途中で切れている（図5のC）．
(6) 口ごもっていたり，発話が不明瞭で正確に書き起こしができない（図5のD）．
(7) 書き起こし困難なデータ（図5のU）．

上記の作業は判断に迷う現象も出現するが，全ファイルについて同一人物が時間をおいて少なくとも2度以上のチエックを行った．問題のあるファイルは図5のMemo欄に記述した．

前処理作業と書き起こし作業を分けた理由は，前者は収録と並行して行う必要性がある点，もう1つは前

表2 VAJで用いた表記
Table 2 The description used in VAJ.

| a | i | u | e | o | ya | yu | ye | yo | wa | wi | wu | wo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ア | イ | ウ | エ | オ | ヤ | ユ | イェ | ヨ | ワ | ウィ | ウェ | ウォ |
| ka | ki | ku | ke | ko | kya | kyu | kye | kyo | kwa | kwi | kwe | kwo |
| カ | キ | ク | ケ | コ | キャ | キュ | キェ | キョ | クァ | クィ | クェ | クォ |
| sa | shi | su | se | so | sha | shu | she | sho | swa | swi | swe | swo |
| サ | シ | ス | セ | ソ | シャ | シュ | シェ | ショ | スァ | スゥィ | スェ | スォ |
| ta | chi | tsu | te | to | cha | chu | che | cho | tsa | tsi | tse | tso |
| タ | チ | ツ | テ | ト | チャ | チュ | チェ | チョ | ツァ | ツィ | ツェ | ツォ |
| na | ni | nu | ne | no | nya | nyu | nye | nyo | nwa | nwi | new | now |
| ナ | ニ | ヌ | ネ | ノ | ニャ | ニュ | ニェ | ニョ | ヌァ | ヌィ | ヌェ | ヌォ |
| ha | hi | fu | he | ho | hya | hyu | hye | hyo | fa | fi | fe | fo |
| ハ | ヒ | フ | ヘ | ホ | ヒャ | ヒュ | ヒェ | ヒョ | ファ | フィ | フェ | フォ |
| ma | mi | mu | me | mo | mya | my | mye | myo | mwa | mwi | mwe | mwo |
| マ | ミ | ム | メ | モ | ミャ | ミュ | ミェ | ミョ | ムァ | ムィ | ムェ | ムォ |
| ra | ri | ru | re | ro | rya | ryu | rye | ryo | rwa | rwi | rwe | rwo |
| ラ | リ | ル | レ | ロ | リャ | リュ | リェ | リョ | ルァ | ルィ | ルェ | ルォ |
| ga | gi | gu | ge | go | gya | gyu | gye | gyo | gwa | gwi | gwe | gwo |
| ガ | ギ | グ | ゲ | ゴ | ギャ | ギュ | ギェ | ギョ | グァ | グィ | グェ | グォ |
| za | ji | zu | ze | zo | ja | ju | je | jo | zwa | zwi | zwe | zwo |
| ザ | ジ | ズ | ゼ | ゾ | ジャ | ジュ | ジェ | ジョ | ズァ | ズゥィ | ズェ | ズォ |
| da | di | du | de | do | dya | dyu | dye | dyo | dwa | dwi | dwe | dwo |
| ダ | ディ | ドゥ | デ | ド | デャ | デュ | デェ | デョ | ドゥァ | ドゥィ | ドゥェ | ドゥォ |
| ba | bi | bu | be | bo | bya | byu | bye | byo | bwa | bwi | bwe | bwo |
| バ | ビ | ブ | ベ | ボ | ビャ | ビュ | ビェ | ビョ | ブァ | ブィ | ブェ | ブォ |
| pa | pi | pu | pe | po | pya | pyu | pye | pyo | pwa | pwi | pwe | pwo |
| パ | ピ | プ | ペ | ポ | ピャ | ピュ | ピェ | ピョ | プァ | プィ | プェ | プォ |
| | ti | tu | | | tya | tyu | tye | tyo | twa | twi | twe | two |
| | ティ | トゥ | | | テャ | テュ | テェ | テョ | トゥァ | トゥィ | トゥェ | トゥォ |
| va | vi | vu | ve | vo | vya | vyu | vye | vyo | vwa | vwi | vwe | vwo |
| ヴァ | ヴィ | ヴ | ヴェ | ヴォ | ヴャ | ヴュ | ヴェ | ヴョ | ヴゥァ | ヴゥィ | ヴゥェ | ヴゥォ |
| zi | dji | dzu | ng | oo | fya | fyu | fye | fyo | | | | |
| ズィ | ヂ | ヅ | ン | ヲ | フャ | フュ | フィェ | フョ | | | | |
| | | | | | q | | | | | | | |
| | | | | | ッ | | | | | | | |

（注1）@は長音化を意味する （注2） _ はポーズを意味する

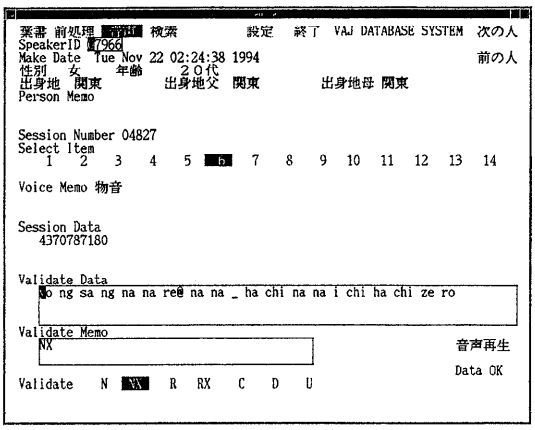



図5 バリデーション作業
Fig. 5 Validation.

処理作業は背景音に注意する作業であるのに対し，書き起こしは発話内容に注意する作業であるため，これら作業を同時に行うのは作業上困難なためである．

### 3.1.3 セグメンテーション

音声区間の切り出し作業（Segmentation）を行った．シードモデルの作成や発声時間（duration）の計測に役立つ．数字データはワード単位で，音韻バランス文は音韻単位でセグメンテーションを行った．セグメンテーションされた例を**図6**に示す．

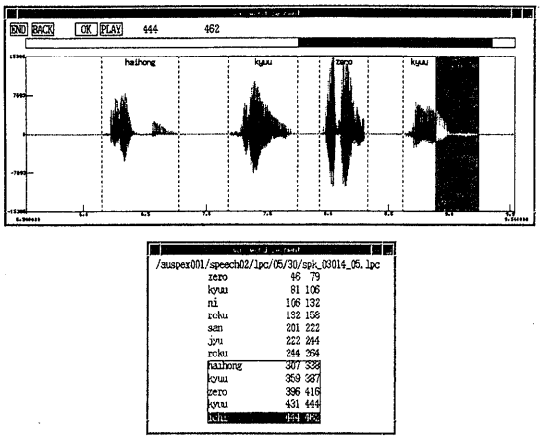



図6 セグメンテンション作業
Fig. 6 Segmentation.

## 3.2 データの品質について

### 3.2.1 人手によるファイルチエックの結果

学習に利用可能なファイルは全体の74.5%あった．残りの学習に利用できないファイルの原因は以下のようなものであった．

(1) 背景雑音（11.2%）：他人の声（1.3%），テレビの音，機械音，電子機器のファン，ゲーム，自動車の音，ベル，フルート，ピアノなどの声以外の背景雑音（9.9%）．

(2) 通信ノイズ（0.2%）

(3) 発声者の態度（0.7%）．ふざけていたり，笑っていたり，咳払いをしているなど不自然な読み方をしているものである．

(4) 収録システムのファイリングエラー（13.4%）．無音区間が2秒以上続くように設定しているため，これ以上長いポーズがあるとファイルが途中で切れてしまうために起こるエラーである．

### 3.2.2 S/N比

S/N比とは，音声区間と音声の前後にある無音区間のエネルギー平均を対比したものである．S/N比を測定してみると，**図7**に示すように，八丈，奄美の，収録データの少ない一部地域を除いては，35 dBを中央値として分布している．各地域別の平均値，標準偏差，測定に用いた標本数を**表3**に示す．ノイズの主な原因はチャネルノイズではなく，背景雑音が大きな原因であるため，このような分布になったと推測される．

性別に関して調査した結果を**表4**に示す．性別には依存していないことが分かる．しかし，年齢別に調べてみると，50代以上でS/N比の平均値が増加していること（**表5**），**図8**のグラフが右側にシフトしていることが分かる．高齢者の方が静かな環境で電話を利

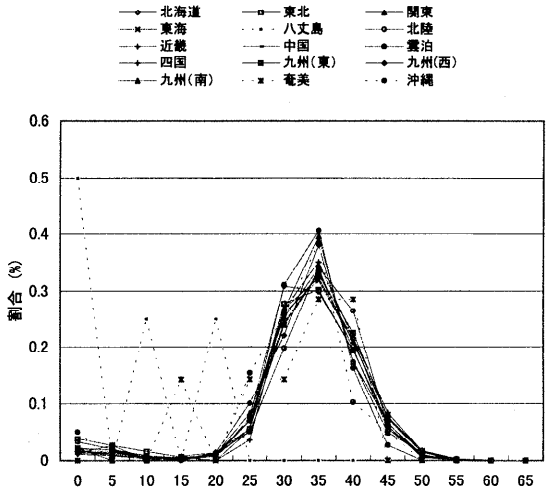



図7 各地域ごとのS/N比
Fig. 7 S/N Ratio in each area.

表3 各地域ごとのS/N比の平均と標準偏差
Table 3 S/N Ratio in each area.

| 地域 | 平均値(dB) | 標準偏差 | 標本数 |
|---|---|---|---|
| 北海道 | 33.18 | 8.05 | 408 |
| 東北 | 32.02 | 10.15 | 817 |
| 関東 | 32.83 | 8.55 | 6578 |
| 東海 | 33.98 | 7.96 | 1418 |
| 八丈島 | 10.00 | 8.66 | 4 |
| 北陸 | 33.47 | 9.43 | 389 |
| 近畿 | 33.98 | 8.17 | 2077 |
| 中国 | 33.35 | 8.18 | 480 |
| 雲伯 | 32.63 | 7.45 | 74 |
| 四国 | 34.67 | 7.29 | 356 |
| 九州(東) | 33.70 | 8.77 | 1500 |
| 九州(西) | 33.25 | 7.88 | 612 |
| 九州(南) | 33.67 | 7.45 | 139 |
| 奄美 | 31.42 | 8.32 | 7 |
| 沖縄 | 31.67 | 8.91 | 39 |
| 先島 | – | – | 0 |

表4 S/N比(男女別)
Table 4 S/N Ratio (male, female).

| | 平均値(dB) | 標準偏差 | 標本数 |
|---|---|---|---|
| 男性 | 33.16 | 8.71 | 6256 |
| 女性 | 33.26 | 8.45 | 10022 |
| 合計 | 33.22 | 8.55 | 16278 |

用していることが分かる.

米国での研究例[23]では, S/N比20dB以上では音声認識率にあまり大きな影響が与えないことが報告されている. 今回のデータではS/N比が20dB以上のファイルが92.4%あった. 20dB以下のファイルは工場内から電話してきたような背景雑音を含むものであった.

### 3.2.3 不要語, 言い直し, 言い淀みの出現頻度

不要語, 言い直し, 言い淀みについては, 現実のアプリケーションシステムを構築するうえで問題となる

表5 S/N比(年代別)
Table 5 S/N Ratio in each generation.

| | 平均値(dB) | 標準偏差 | 標本数 |
|---|---|---|---|
| 10代 | 32.82 | 9.76 | 1476 |
| 20代 | 32.25 | 8.52 | 4801 |
| 30代 | 32.90 | 8.35 | 4634 |
| 40代 | 33.03 | 8.69 | 2234 |
| 50代 | 35.91 | 7.43 | 1573 |
| 60代 | 36.68 | 6.31 | 658 |
| 70代以上 | 38.41 | 9.03 | 170 |

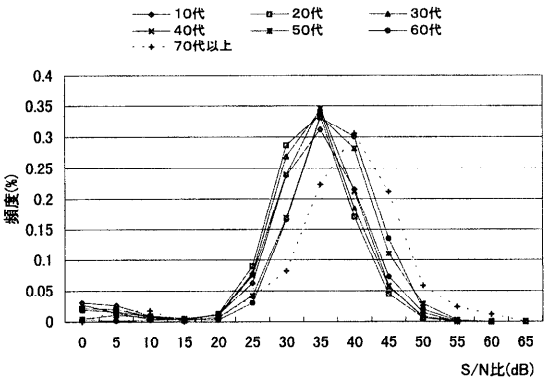



図8 S/N比(年代別)
Fig. 8 S/N Ratio in each generation.

が, 今回のデータには, "aa", "uh", "ee" という不要語が56.0%, 言い直しが3.9%, 言いよどみが0.4%存在していた.

## 4. データベースの分析

データベースから得られる知見を報告する.

### 4.1 発声時間に関する分析

発声時間について調べてみると,

(1) 米国には "The Southern drawl" という発声時間が非常に長い(つまり, ゆっくりしゃべる)方言が南部地方に存在するが, 日本にはそのような方言は存在しない.
(2) 年齢の増加とともに発声時間が長くなる.
(3) 男女間では男性の方が発声時間が短い.
(4) 年齢の影響は男女間の影響よりも大きい

ことが分かる.

データベースの連続数字に対して個々の数字150,558個に1つ1つに発声時間(duration)がついている. 各単語ごとの発声時間の平均値と標準偏差, および, 実験に用いたデータ数を**表6**に示す. 各地域ごとに横軸を時間(ms), 縦軸を頻度として発声時間の頻度をとったものが**図9**である. サンプル数の少ない一部地域を除くと, ほぼ1つの曲線状の形状になり, 米国の南部方言のような発声時間の長い方言は観測されなかった. 表6より求めた各数字の平均値に, 各地域ご

表7 各数字の発声された割合
Table 7 The ratio of each words pronounced in the test data.

| | ichi | ni | san | yon | go | roku | nana | hachi | kyuu | zero | rei | maru | no | 合計 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 北海道 | 10.42% | 10.26% | 9.77% | 8.93% | 9.39% | 8.65% | 8.76% | 9.12% | 7.46% | 11.62% | 0.60% | 0.52% | 4.52% | 100.00% |
| 東北 | 9.75% | 12.34% | 10.48% | 10.51% | 10.16% | 6.90% | 8.70% | 7.87% | 7.15% | 10.30% | 0.45% | 0.43% | 4.97% | 100.00% |
| 関東 | 8.28% | 10.88% | 11.06% | 10.69% | 9.04% | 7.85% | 8.75% | 8.76% | 8.17% | 11.10% | 0.36% | 0.45% | 4.61% | 100.00% |
| 東海 | 8.38% | 10.31% | 10.24% | 9.25% | 11.45% | 7.95% | 8.87% | 8.32% | 7.03% | 11.23% | 0.52% | 0.37% | 6.09% | 100.00% |
| 八丈島 | 18.52% | 11.11% | 3.70% | 7.41% | 7.41% | 7.41% | 7.41% | 14.81% | 7.41% | 14.81% | 0.00% | 0.00% | 0.00% | 100.00% |
| 北陸 | 8.82% | 11.50% | 8.53% | 9.22% | 9.80% | 10.40% | 10.46% | 8.01% | 6.54% | 10.17% | 0.32% | 0.69% | 5.53% | 100.00% |
| 近畿 | 8.55% | 10.59% | 9.33% | 9.60% | 8.87% | 8.32% | 10.69% | 8.12% | 8.55% | 11.17% | 0.61% | 0.41% | 5.19% | 100.00% |
| 中国 | 8.54% | 9.72% | 9.68% | 9.72% | 9.93% | 7.75% | 9.13% | 9.97% | 7.95% | 10.17% | 0.93% | 0.52% | 5.98% | 100.00% |
| 雲伯 | 7.63% | 12.90% | 8.18% | 11.65% | 9.71% | 6.93% | 6.93% | 9.15% | 7.77% | 10.96% | 0.55% | 0.28% | 7.35% | 100.00% |
| 四国 | 7.66% | 9.59% | 10.03% | 8.79% | 10.09% | 8.61% | 8.96% | 9.97% | 7.06% | 11.78% | 0.42% | 0.24% | 6.80% | 100.00% |
| 九州(東) | 8.62% | 11.40% | 9.21% | 7.88% | 8.63% | 8.39% | 11.66% | 7.08% | 8.71% | 8.17% | 1.64% | 0.78% | 7.83% | 100.00% |
| 九州(西) | 7.86% | 10.36% | 10.01% | 9.66% | 9.66% | 8.70% | 9.10% | 8.77% | 9.03% | 9.80% | 1.03% | 0.51% | 5.54% | 100.00% |
| 九州(南) | 7.96% | 11.72% | 9.49% | 7.73% | 10.26% | 7.04% | 10.41% | 6.51% | 9.49% | 10.41% | 0.23% | 0.15% | 8.58% | 100.00% |
| 奄美 | 6.25% | 1.56% | 1.56% | 3.13% | 9.38% | 3.13% | 12.50% | 23.44% | 9.38% | 23.44% | 0.00% | 0.00% | 6.25% | 100.00% |
| 沖縄 | 6.34% | 8.07% | 8.93% | 8.36% | 10.37% | 4.90% | 8.93% | 14.12% | 13.83% | 8.36% | 0.00% | 1.73% | 6.05% | 100.00% |
| 先島 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |

表6 各単語の発声時間
Table 6 Durations of the words.

| 単語 | 平均値 | 標準偏差 | 標本数 |
|---|---|---|---|
| ichi | 349(ms) | 83.9 | 12795 |
| ni | 316.5 | 85.9 | 16355 |
| san | 369.2 | 85.7 | 15479 |
| yon | 326.8 | 73.8 | 14888 |
| go | 311.6 | 81.2 | 14180 |
| roku | 357.4 | 81.0 | 12062 |
| nana | 353.8 | 72.9 | 14158 |
| hachi | 397.9 | 87.1 | 12839 |
| kyuu | 334.0 | 77.4 | 12317 |
| zero | 351.2 | 79.6 | 16143 |
| rei | 318.2 | 82.7 | 856 |
| maru | 326.8 | 73.7 | 720 |
| no | 235.0 | 58.8 | 7766 |

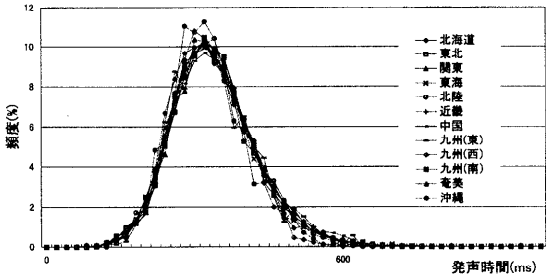



図9 発声時間の分布（地域別）
Fig. 9 The distribution of word durations in each area.

とに各数字が発声された割合（**表7**）を掛けて平均化した数値を**表8**の理論値の欄に示した．八丈，奄美の平均値と理論値がずれているが，これはサンプル数が極端に少ないために発話者のサンプリングに偏りが生じたため起こったものと考えられる．表8に示すように，全国平均35.0歳（標準偏差12.8）に対して，八丈では50.0歳，奄美では29.3歳である．また，男性と女性の比率も全国平均で38.4%に対し，八丈，奄美とも男性が0%であり，話者の分布に偏りが存在しているため，発声時間に影響したものと思われる．

表8 各地域ごとの発声時間の分布
Table 8 Word durations by each area.

| 地域 | 平均値 | 標準偏差 | 標本数 | 理論値 | 平均年齢 | 男性比率 |
|---|---|---|---|---|---|---|
| 北海道 | 337.5(ms) | 86.2 | 3675 | 341.4 | 36.2(才) | 40.1(%) |
| 東北 | 341.2 | 86.4 | 7291 | 339.1 | 36.7 | 37.0 |
| 関東 | 341.2 | 85.6 | 61791 | 340.9 | 34.4 | 36.4 |
| 東海 | 337.4 | 86.8 | 13325 | 338.5 | 36.1 | 38.2 |
| 八丈 | 472.6 | 97.6 | 27 | 349.1 | 50.0 | 0.0 |
| 北陸 | 338.9 | 84.8 | 3470 | 339.3 | 38.7 | 46.3 |
| 近畿 | 341.8 | 87.5 | 19334 | 340.0 | 34.6 | 38.3 |
| 中国 | 343.9 | 91.0 | 4413 | 339.7 | 39.1 | 37.7 |
| 雲伯 | 339.7 | 84.6 | 721 | 336.1 | 42.0 | 36.5 |
| 四国 | 337.2 | 89.4 | 3369 | 339.5 | 35.5 | 36.8 |
| 九州(東) | 336.1 | 87.9 | 13545 | 336.2 | 35.7 | 38.4 |
| 九州(西) | 341.9 | 87.3 | 5726 | 339.5 | 38.9 | 44.3 |
| 九州(南) | 327.1 | 78.2 | 1306 | 334.8 | 34.4 | 36.7 |
| 奄美 | 330.9 | 70.1 | 64 | 348.9 | 29.3 | 0.0 |
| 沖縄 | 336.7 | 83.9 | 347 | 341.0 | 27.1 | 24.3 |
| 先島 | — | — | 0 | — | - | - |

次に，年齢と発声時間の関係について，年齢の分かっている146,356個の数字について，同様の調査を行った．結果を**図10**と**表9**に示す．年齢とともに曲線が右側に移動しており，発声時間が長くなっていることが分かる．10代と60代を比較すると平均値で56.3 ms（17.5%）も発声時間が長くなっている．すべての年代において理論値が340前後であることから，読み上げた数字のばらつきが原因で起きたものではないことは明らかである．

男女間の発声時間についても同様の調査を行うと，**図11**に示すように女性の方がグラフが右にあり，男性より発声時間が長いことが分かる．年齢や標本数の偏りが原因ではない．**表10**に示すように年齢的には男女とも平均的にはほぼ等しい．年齢が影響しないように各年代ごとの男女について比較を行っても，同様の傾向はどの年代でも起きていることを観測した．

なお，実験に用いたデータは読み上げ数字および自由発声（たとえば，よく使う電話番号）が混在している．一般に女性の方が話し好きと思われており，発声

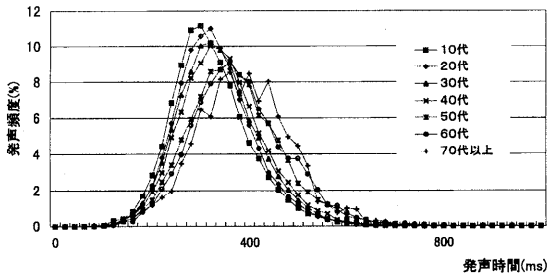



図 10 発声時間の分布（年齢別）
Fig. 10 The word durations in each generation.

表 9 各世代ごとの発声時間の分布
Table 9 Word durations by each generation.

| 世代 | 平均値 | 標準偏差 | 標本数 | 理論値 | 男性比率 |
|---|---|---|---|---|---|
| 10代 | 322.6(ms) | 84.3 | 15786 | 340.1 | 35.6(%) |
| 20代 | 330.9 | 81.2 | 45336 | 339.7 | 35.9 |
| 30代 | 336.7 | 83.9 | 42487 | 340.7 | 43.9 |
| 40代 | 345.0 | 86.1 | 20093 | 340.1 | 36.7 |
| 50代 | 368.9 | 95.9 | 14773 | 340.1 | 31.6 |
| 60代 | 378.9 | 97.0 | 6273 | 340.0 | 42.4 |
| 70代以上 | 389.4 | 98.8 | 1608 | 340.3 | 37.6 |

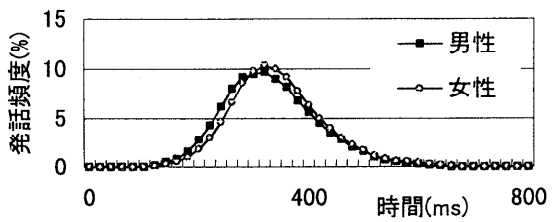



図 11 発声時間の分布（男女別）
Fig. 11 The word duration (male vs. female).

表 10 性別ごとの発声時間の分布
Table 10 Word durations by each gender.

| 性別 | 平均値 | 標準偏差 | 標本数 | 理論値 | 平均年齢 |
|---|---|---|---|---|---|
| 男性 | 333(ms) | 89.4 | 55200 | 338.5 | 35.3(才) |
| 女性 | 344.4 | 84.9 | 87603 | 340.4 | 34.7 |
| 合計 | 339.9 | 86.9 | 142803 | 339.7 | 35.0 |

時間も男性より短いのではないかと推測されるが，今回の実験では逆の結果となった．その理由であるが，年齢，性別ともに運動機能が高いほど発声時間が短いという関係が存在することから，口を動かす筋力の差が発声時間の長短に影響を与えているのではないかと考えられる．

### 4.2 方言について

数字について調査してみると，「0（“zero”，“rei”，“maru”）」を “dero”，“jero”，“zeru”，「1（“ichi”）」を “iji（イジ）”，「6（“roku”，“roq”）」を “rogu（ろぐ）” という発音している例が見られた．これらについて，性別，年齢，出身地について調べたのが，**表 11**，**12**，**13** である．男性より女性の方が出現する割合が

表 11 方言と性別の関係の例
Table 11 The number of Japanese dialects by gender.

| 性別 | dero | jero | zeru | iji | rogu |
|---|---|---|---|---|---|
| 男性 | 13 | 4 | 0 | 0 | 0 |
| 女性 | 52 | 17 | 1 | 2 | 2 |
| 合計 | 65 | 21 | 1 | 2 | 2 |

表 12 方言と年代の関係の例
Table 12 The number of Japanese dialects by each generation.

| 年代 | dero | jero | zeru | iji | rogu |
|---|---|---|---|---|---|
| 10代 | 0 | 0 | 0 | 0 | 0 |
| 20代 | 6 | 1 | 0 | 0 | 0 |
| 30代 | 10 | 0 | 0 | 0 | 0 |
| 40代 | 15 | 1 | 0 | 0 | 0 |
| 50代 | 17 | 4 | 1 | 0 | 0 |
| 60代 | 10 | 4 | 0 | 0 | 0 |
| 70代 | 2 | 11 | 0 | 0 | 0 |
| 80代 | 2 | 0 | 0 | 0 | 0 |
| 不明 | 3 | 0 | 0 | 2 | 2 |
| 合計 | 65 | 21 | 1 | 2 | 2 |

表 13 方言と出身地の関係の例
Table 13 The number of Japanese dialects by each area.

| 出身地 | dero | jero | zeru | iji | rogu |
|---|---|---|---|---|---|
| 北海道 | 0 | 0 | 0 | 0 | 0 |
| 東北 | 0 | 2 | 0 | 0 | 0 |
| 関東 | 1 | 0 | 1 | 2 | 2 |
| 東海 | 1 | 0 | 0 | 0 | 0 |
| 八丈島 | 0 | 0 | 0 | 0 | 0 |
| 北陸 | 0 | 0 | 0 | 0 | 0 |
| 近畿 | 55 | 2 | 0 | 0 | 0 |
| 中国 | 1 | 2 | 0 | 0 | 0 |
| 雲泊 | 0 | 0 | 0 | 0 | 0 |
| 四国 | 0 | 2 | 0 | 0 | 0 |
| 九州（東 | 7 | 10 | 0 | 0 | 0 |
| 九州（西 | 0 | 0 | 0 | 0 | 0 |
| 九州（南 | 0 | 0 | 0 | 0 | 0 |
| 奄美 | 0 | 0 | 0 | 0 | 0 |
| 沖縄 | 0 | 0 | 0 | 0 | 0 |
| 先島 | 0 | 0 | 0 | 0 | 0 |
| 不明 | 0 | 3 | 0 | 0 | 0 |
| 合計 | 65 | 21 | 1 | 2 | 2 |

多く見られる．年齢については，10代を除くすべての年代で出現している．また，地域については，方言学者が「方言地図をつくるのは容易ではない」[24]といっているように，地域を特定することは非常に難しい．たとえば，“dero（デロ）” は近畿地方で高く現れるが，九州（東側），関東，東海，中国などにも現れる．

アクセントについてはデータ収録する際の応答音声が標準語による音声であったためか，標準語と変わらない発声も多いが，その地方の特徴を反映されたデータも少なからず存在している．文献 19) の CD-ROM 版では関東と関西のアクセントの差を聞くことができる．

## 5. 音声認識への応用

この章では，話者の年齢や出身地に関する情報が音声認識にどのような影響を与えているのかについて報

表 14 音声認識の学習データ（12グループ）
Table 14 12 groups of training data for speech recognition.

| 性別 | 年齢 | 学習データの地域 | | |
|---|---|---|---|---|
| | | 全国 | 関東 | 西日本 |
| 男性 | 人口比 | 500人(A1) | 500人(A2) | 500人(A3) |
| | 20代 | 500人(B1) | 500人(B2) | 500人(B3) |
| 女性 | 人口比 | 500人(C1) | 500人(C2) | 500人(C3) |
| | 20代 | 500人(D1) | 500人(D2) | 500人(D3) |

告する．

### 5.1 年齢に関する実験

#### 5.1.1 実験方法

実験に用いた音声認識システムは，LPCベースの単語認識システムである．8 kHzのサンプリングを行い，高域強調（$1-0.96\,Z^{-1}$）を行い，30 msのHamming Windowを用い，20 msのフレーム周期で10次のLPCを計算する．各LPCパラメータは各フレームごとに34次元の特徴ベクトル（14次元のmel-scaleのfilter bank values，その差分，Speech level，voicingなどからなる）に変換し，次に，線形変換を行い16次元の特徴ベクトルに縮退する．

認識システムはHidden Markov Model（HMM）を用いて1単語を1語とするワードモデルを作成する．モデルの状態数は発声時間に応じて12状態から16状態である．各モデルは，遷移状態がRecursive Transition Networkで記述され，ノード部分が連続ガウス分布でモデル化されている．各モデルはViterbi searchを用いて最尤度のスコアのパスを求める．入力データと各モデルとの評価にはcoveriance matrixが同じであると仮定（pooled coveriance）してスコアを求める．

文法規則は，「(silence) → 各数字モデル → (silence)．ただし，(silence)は省略可能」という単純なものを用い，全単語モデルについて評価を行い，最上位の解のみを正解とした．

実験には，**表14**に示すように12種類のモデルを作成した．まず，性別に関して男性と女性の2つに分類し，次に，年齢に関して2つのグループ（全年齢から人口比で抽出したデータと20代のみのデータ），そして，地域に関して3つのグループ〔全国各地域の人口比に応じて抽出したデータ（All），関東地方のみからのデータ（Kanto），関西以西から九州地方までのデータ（Nishi）〕である．学習データには各条件にかなう話者500人の録音状態が良好な電話番号に関する1,500ファイル分（1人あたり10桁の数字3ファイル）を用いて，ワードモデル（数字12種類ichi, ni, san, yong, go, roku, nana, hachi, kyuu, zero, maru, reiおよびsilenceモデル）を作成した．

表 15 実験5.1の結果
Table 15 The results of experimentations in 5.1.

| 年代 | 男性 | | | 女性 | | |
|---|---|---|---|---|---|---|
| | 評価データ1 | 平均値 | 差分 | 評価データ2 | 平均値 | 差分 |
| 10代 | 958 | 12.57(%) | −1.23(%) | 2437 | 12.83(%) | 0.23(%) |
| 20代 | 5859 | 5.60 | −0.10 | 11341 | 9.81 | −0.11 |
| 30代 | 6396 | 6.97 | 0.00 | 7750 | 9.03 | 0.39 |
| 40代 | 1127 | 6.10 | −0.17 | 1994 | 10.14 | 1.03 |
| 50代 | 420 | 9.60 | 0.53 | 1759 | 10.87 | 1.48 |
| 60代 | 379 | 13.07 | 1.14 | 637 | 11.64 | 1.68 |
| 70代 | 30 | 17.97 | 2.64 | 116 | 15.40 | 2.52 |
| 80代 | 30 | 15.80 | 1.70 | 10 | 20.00 | 0.00 |
| 合計 | 15199 | | | 26044 | | |

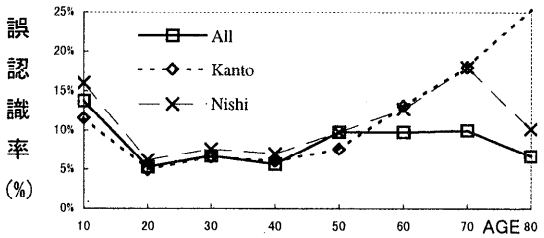



図 12 年齢と誤認識率の関係（男性1）
（学習データは男性，人口比A1, A2, A3）
Fig. 12 The relationship between speaker's age and error ratio in the speech recognition test (1). (Training data is male, all the age, A1, A2, and A3).

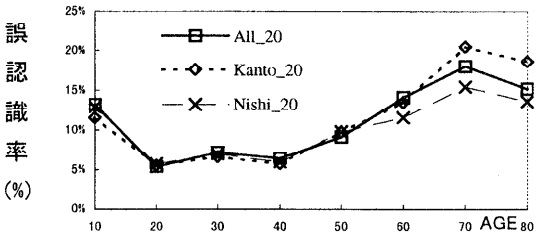



図 13 年齢と誤認識率の関係（男性2）
（学習データは20代の男性，B1, B2, B3）
Fig. 13 The relationship between speaker's age and error ratio in the speech recognition test (2). (Training data is male, twenties, B1, B2, and B3).

評価には，**表15**に示すように男性モデルに対しては評価データ1，女性のモデルに対しては評価データ2を用いた．評価データには学習データは含まれていない．

#### 5.1.2 実験結果

エラーの割合を図12から図15に示す．男性に関する結果を**図12**と**図13**，女性に関する結果を**図14**と**図15**に示す．どのグラフとも共通している点は，グラフの形状が20代から40代までは低く，10代（15歳から19歳まで）や高齢者とは異なる点である．特に，高齢者層はエラー率が高い点が注目される．また，20代から40代までは，学習データが20代であろうが，全人口比のデータの場合も同じような値を示して

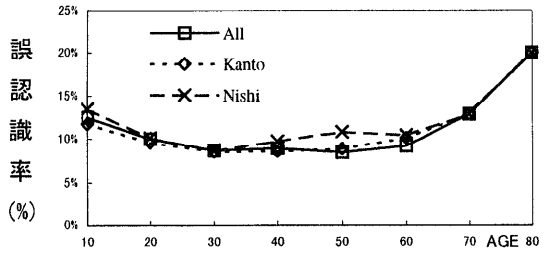



図 14 年齢と誤認識率の関係（女性 1）
（学習データは女性，人口比 C1，C2，C3）

Fig. 14 The relationship between speaker's age and error ratio in the speech recognition test (3). (Training data is female, all the age, C1, C2, and C3).

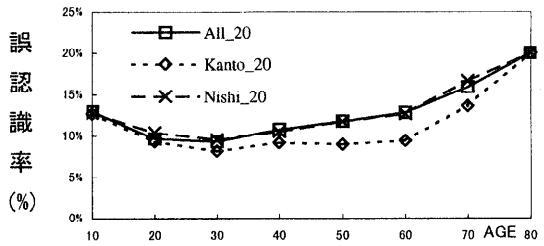



図 15 年齢と誤認識率の関係（女性 2）
（学習データは 20 代女性 D1，D2，D3）

Fig. 15 The relationship between speaker's age and error ratio in the speech recognition test (4). (Training data is female, twenties, D1, D2, and D3).

いる．このことは，20 代から 40 代までの声が，他の世代と比較して似ていることを意味している．

次に，モデルを比較する．20 代のみを学習データとしたモデル（男性 B1～B3，女性 D1～D3）の平均値および，年齢分布を人口比で構成したモデル（男性 A1～A3，女性 C1～C3）の平均との差分を表 15 の平均値，差分の欄に示す．統計的な有意差を検定[27),28)]すると，男性の 10 代の差に 25%，女性の 30 代に 25%，40 代，60 代に 10%，50 代に 2.5%の有意差が認められた．

### 5.1.3 考　　察

年齢がちがう場合，なぜ認識率に大きな影響を与えるかに関しては，1 つは，高齢になるにつれて歯がぬけるなど音響的に異なってくる点が考えられる．HMM を用いているのである程度の時間軸方向の伸縮変動は吸収していると思われるが，図 10 に示したように年齢とともに発話速度が遅くなる傾向があるので，この発話速度の影響も受けている可能性がないとはいえない．

## 5.2 地域差に関する実験

### 5.2.1 実 験 方 法

16 の地域のうち八丈，奄美，先島については十分なデータが揃っていないので，残りの 13 地域について調査した．HMM モデルには 5.1 節の実験で用いたモデル（A と C）を用いた．

テストデータには，5.1 節の実験で年齢が大きく影響を与えていることが分かっているので，20 代，30 代，40 代のみのデータで，学習に用いていないデータを使用した．**表 16** に示すように男性のモデルに対しては評価データ 3 を，女性のモデルには評価データ 4 を用いた．評価データは録音状態良好な通常 10 桁（最低でも 6 桁以上）の数字ファイルを用いた．

表 16 評価用データに用いたファイル数
Table 16 The file numbers of test data sets.

| 地域 | 評価データ3 男性 | 評価データ4 女性 |
|---|---|---|
| 北海道 | 388 | 410 |
| 東北 | 388 | 763 |
| 関東 | 5906 | 10189 |
| 東海 | 1085 | 1901 |
| 北陸 | 298 | 346 |
| 近畿 | 1623 | 2545 |
| 中国 | 326 | 474 |
| 雲泊 | 110 | 72 |
| 四国 | 167 | 341 |
| 九州(東) | 1496 | 2060 |
| 九州(西) | 396 | 415 |
| 九州(南) | 139 | 157 |
| 沖縄 | 10 | 70 |
| 合計 | 12332 | 19743 |

表 17 実験 5.2 における各モデルの平均値
Table 17 The averages in the experimentation 5.2.

| | モデル | 単純平均 | 標準偏差 | 加重平均 |
|---|---|---|---|---|
| 男性 | 全国モデル | 6.37(%) | 1.40 | 5.84(%) |
| | 関東モデル | 6.45 | 2.11 | 5.48 |
| | 西日本モデル | 6.47 | 1.34 | 6.52 |
| 女性 | 全国モデル | 8.82 | 1.11 | 9.37 |
| | 関東モデル | 8.89 | 1.53 | 9.04 |
| | 西日本モデル | 9.33 | 1.52 | 9.52 |

### 5.2.2 実 験 結 果

**表 17** に各モデルの平均値を示す．各地域ごとの値を単純に平均した単純平均，その標準偏差，および，各地域のデータ数を重みとして掛けて求めた加重平均を示した．単純平均では，男女ともに全国モデル（A1，C1），関東モデル（A2，C2），西日本モデル（A3，C3）の順に成績が良いが，加重平均でみると，関東モデル，全国モデルの順序が入れ替わる．各地域ごとの平均からのばらつきは数%（おおむね 2～4%の範囲）にある．標準偏差をみると，全国モデルが男女ともに散らばりが小さく安定していることが分かる．

**図 16** と**図 17** に各モデルの加重平均を基準に各地域の値がどれだけずれているかを示した．男女共通の傾向として，中国地方の誤認識率が高く，雲伯地方が低い．t 検定を行うと，各モデルの全国平均と比較して，中国地方は男性全国モデルで 5%，西日本モデルで 25%，女性では関東モデルで 5%，西日本モデルで

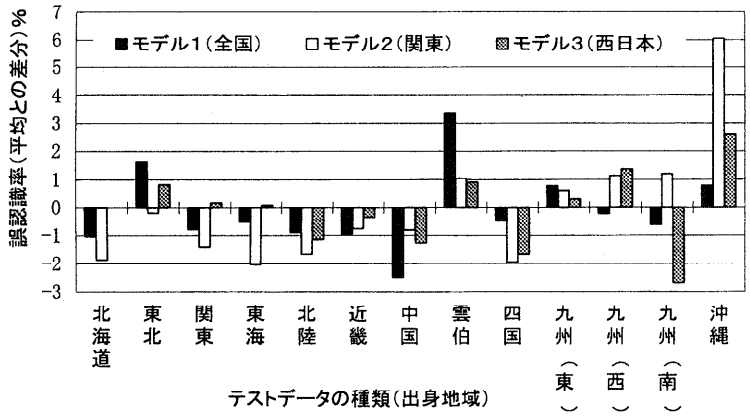



図16 各出身地の音声認識への影響（男性）

Fig. 16 The relationship between speaker's growing area and error ratio in the speech recognition test (in the case of male).

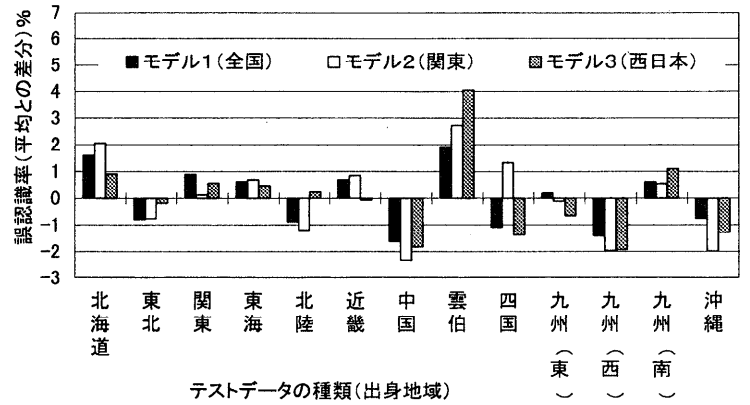



図17 各出身地の音声認識への影響（女性）

Fig. 17 The relationship between speaker's growing area and error ratio in the speech recognition test (in the case of female).

10%，全国モデルで25%で有意差がある．また，雲伯地方については，男性全国モデルで10%，女性関東モデル，西日本モデルで25%で有意差があることが示されている．

各モデルの値を比較すると，関東地方については，全国モデルと関東モデルでは，男女ともに0.52～0.71%の改善が見られた．また，西日本（近畿，中国，雲伯，四国，九州3地域）については，男性では西日本モデルが7地域平均値5.99%，全国モデルが6.27%，関東モデルが6.51%であった．女性では，西日本モデルの平均値が他のモデルの全国平均と著しくずれているので，単純に比較できないが，全国モデル，および，西日本モデルの全体の平均値（単純）と西日本7地域の平均との差が−0.10%であるのに対し，関東モデルでは0.15%と悪くなっている．学習データと評価データの間には年齢差ほどではないが，地域差が存在していると考えられる．

### 5.2.3 考　　察

ここでいう地域差の影響という意味は，方言および通信チャネルによる歪みの影響を含んでいること前提にしている．このデータベースでは発声源でのデータを収録していないため通信チャネルによる歪みやチャネルノイズの影響の度合いを計ることはできない．

今回の実験の結果として重要な点は，通信チャネルの影響なのか方言そのものの影響なのかはその原因は不明瞭であるが，少なくとも地域差による影響がみられるということである．今回の実験ではその影響は，2～4%程度であった．

電話回線の通信路における歪みについてはCCITT（現在ITU）の勧告があり，入力に対する出力の歪みが7dB以内[29]であることが求められている．歪みを測定した研究[30]によると周波数に対し一様ではなく

歪み方が4種類以上存在しているという．パスが変われば歪みも変わる可能性もあり，歪みを推定し補償する作業はかなり難しい作業になることが予想される．

### 5.3 他の研究結果との比較

研究の環境，規模，目的は異なるが，方言データベース（13都市，各22人，合計286人）で行った実験[31]がある．その報告と比較すると，年齢が認識率に影響を与えていることはこの実験でも裏づけることができた．方言に関しては，本実験では通信路の影響も含むものであるのでそのまま比較することは妥当でないが，従来の研究では地域差がほとんどなかったと報告されているのに対し，本実験においては，数%程度の影響が出ている点が異なる．

## 6. むすび

日本のPolyphoneプロジェクトについて目標を達成したことを報告する．電話回線を通して8,866人の音声データベースを作成した．このデータベースを調査することにより，電話回線を用いたアプリケーションを作るときの基礎データを得ることができる．たとえば，通信ノイズや背景雑音の頻度や影響，発話速度に関する分析，方言の分析，不要語などの言語的分析，話者（性別，年齢や出身地）の音声認識への影響の度合いなどである．今回の調査では，年齢が音声認識に無視できないほど大きな影響を与えていること，および，従来あまり報告されていないかった地域差の影響（方言，および，通信歪みの影響を含めて）も数%程度の認識率に影響を与えていることが明らかになった．

今回の収録で分かったことは，若年者と高齢者のデータの収録には大きな配慮が必要であるということである．年齢15歳以下の子供の場合は漢字の読みが障害となる．また，高齢者の場合は文字の大きさや機械操作への抵抗を取り除くような配慮が必要である．また，通信チャネルの歪みの問題については，参加者の何人かに協力を願って，代表的な地点で追加調査などを行えば，その影響を計ることが可能であるが，これについては今後の課題である．

なお，データベースの公開の有無については決まっていないが，外部機関にデータを提供した実績がある．

**謝辞** 本研究をご支援していただいた（株）テキサスインスツルメンツ筑波研究開発センター Dr. Don Show（前所長），生駒俊明所長（現日本TI社長，博士），Dr. Donald Shaver（DVD前室長），Dr. Joseph Picone（Senior Researcher）および，音声データベースの作成にご協力いただいた新井希望嬢，藤村奈保子嬢，渡辺公子嬢，亀山玲子嬢，福居恭子嬢に感謝いたします．

## 参考文献

1) 田中和世ほか：音声の知的処理に関する調査報告，日本情報処理開発協会 (1992).
2) Wilpon, J.G. and Roe, D.: Applications of Speech Recognition Technology in Telecommuications, *Proc. ICSLP'94*, pp.667–670 (1994).
3) *Proc. 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications* (1992).
4) 菅村 昇：電話網における音声認識技術の応用，信学技法，SP93-13 (May 1993).
5) 松岡達雄：IVTTA94 会議報告，信学技法，SP94-56 (Nov. 1994).
6) 中津ほか：電話音声の認識方法，信学論，Vol.J66-D, No.4, pp.377–384 (1983).
7) 黒岩，武田，井ノ上，野垣戸，山本，庄境，尾和，長濱：電話音声の連続音声認識に基づく内線電話受付装置の試作と評価，信学論A，Vol.J77-A, No.2, pp.223–231 (1994).
8) 森島，磯部，吉谷，小泉：ホームバンキングを想定した電話音声認識の評価，信学技報，SP-95-121, pp.27–33 (1996).
9) 板橋秀一：特集音声データベース，日本音響学会誌，Vol.48, No.12, pp.876–905 (1992).
10) 板橋秀一（監修）：音声データベース・音声コーパス，人文学と情報処理12号 (1996).
11) 板橋秀一：音声コーパス，情報処理，Vol.38, No.11, pp.1013–1018 (1997)
12) 阿部，匂坂，梅田：音声データベースユーザズマニュアル，ATR Technical Report 116, ATR自動翻訳研究所 (1990).
13) （財）日本情報処理開発協会：研究用連続音声データベース，音響学会CD-ROM (1992–93).
14) *Report on the Cocosda Workshop*, pp.23–25, Berlin, Germany (Sep. 1993).
15) Staples, T., Picone, J. and Arai, N.: The Voice Across Japan Database – The Japanese Language Contribution to Polyphone, *Proc. ICASSP'94*, Australia, pp.89–92 (1994).
16) Kudo, I., Nakama, T., Arai, N. and Fujimura, N.: The data collection of Voice Across Japan (VAJ) project, *Proc. ICSLP'94*, Yokohama, pp.1799–1802 (1994).
17) 中間，藤村，新井，工藤：Voice Across Japan (VAJ) data collection projectについて，信学技報，SP94-58 (1994-11).
18) Kudo, I., Nakama, T. and Watanabe, T.: An Estimation of Speaker Sampling in Voice Across Japan Database, *Proc. ICASSP'96*, Atlanta, Vol.2, pp.825–828 (1996).
19) Kudo, I., Nakama, T., Watanabe, T. and

Kamcyama, R.: Data Collection of Japanese Dialects and Its Influence into Speech Recognition, *Proc. ICSLP'96*, Pholadelphia, pp.2021–2024 (1996).
20) Damhuris, M., Boogarrt, T., et al.: Creation and analysis of the Dutch Polyphone corpus, *Proc. ICSLP'94*, Yokohama, pp.1803–1806 (1994).
21) Rosenbeck, P., Baungaard, B., et al.: The design and efficeint recognition of a 3000 speaker Scandinavian Telephone Speech database: RAFAEL.0, *Proc. ICSLP'94*, Yokohama, pp. 1807–1810 (1994).
22) Tapias, D., Acero, A., et al.: The VESTEL Telephone Speech Database, *Proc. ICSLP'94*, Yokohama, pp.1811–1814 (1994).
23) Wheatley, B. and Picone, J.: Voice Across America: Toward Robust Speaker-Independent Speech Recognition for Telecommunications Applications, *Journal of Digital Signal Processing*, pp.45–63 (1991).
24) 国立国語研究所：方言と日本語教育，大蔵省印刷局 (1993).
25) 矢野一郎（監修）：日本国勢図会地域統計版－94，国勢社 (1994).
26) 壇辻正剛：音声データベースの音声表記法，人文学と情報処理，No.12, pp.12–20, 勉誠社 (1996).
27) 蓑谷千凰彦：推測統計のはなし，東京図書 (1997).
28) 永田 靖：統計的方法のしくみ，日科技連 (1997).
29) 前田光治（監修）：有線伝送工学，電子情報通信学会 (1982).
30) 高橋，飛田，長島，菅村：電話音声認識における伝送特性模擬の効果とその分析，信学技法，SP93-106, pp.17 24 (1993).
31) 前原，謝，呂，林：方言データベースを用いた不特定話者単語認識，日本音響学会誌，Vol.44, No.7, pp.479–487 (1988).

(平成10年10月23日受付)
(平成11年 7月 1日採録)

**工藤 育男**（正会員）

1983年早稲田大学理工学部電気工学科卒業．1985年同大学院博士前期課程修了．同年CSK総合研究所入社，機械翻訳を用いた知的CAIの研究開発に従事．1987年ATR自動翻訳電話研究所出向．対話コーパス，対話翻訳システムの研究開発に従事．1991年CSK復帰，音声対話の研究に従事．1993年テキサスインスツルメンツ筑波研究開発センターに移籍，音声データベース，音声認識システムの研究開発に従事．現在，ジャストシステム所属．博士（工学）．電子情報通信学会，日本音響学会，言語処理学会，人工知能学会，教育システム情報学会，ACL各会員．

**中間 崇夫**

1989年，東京電気大学理工学部経営工学科卒業．日本アポロコンピュータ入社．1990年，ヒューマンシステム入社．1993年，テキサスインスツルメンツ筑波研究開発センター入社，音声データベース，および，音声認識合成の研究開発に従事．