社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－SLP－74 (39)
2008 / 12 / 10

話速管理機能を持った原稿提示収録システム (ReCoK5) と話速バリエーション型音声データベース (SRV-DB) の公開について

高橋 弘太† 蔦木 圭悟† 吉原 亨†

† 電気通信大学情報通信工学科 〒182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: †{kota,tsutaki,yosihara}@ice.uec.ac.jp

あらまし 高齢者に確実に音声を聞き取らせるには，どうしたらよいか，また健常者に対しては，短い時間でより多くの音声を聞き取らせるには，どうしたらよいか．これらの研究を行うためには，正確な時間管理のもと，同じ音声を同じ話者が話速を変えて発声した音声のデータベースが必要である．しかし，現時点では，そのような目的で大規模に作られた音声データベースは存在しない．そこで，我々は，平成 20 年度～平成 22 年度の科研費の研究として，さまざまな話速で録音した音声データベースを構築している．今回は，このデータベースの紹介を行うとともに，会場で音声の研究者の生の意見やコメントをうかがって，今後のデータベース作りに反映させていきたいと考えている．また，我々は，音声データベース作成のために，話速を正確に管理しつつ録音できるシステム（原稿提示システム）も製作した．システムはパソコン上で動くもので，将来は，このソフトも公開して，電通大外でも音声を追加できるようにしたいと考えている．当日は，このシステムの機能と，どのような工夫がなされているかについても紹介したい．
キーワード 音声データベース，話速変換，時間軸変更，効率的視聴

Recording system for controlling speaking rate (ReCoK5) and public domain speech database with speaking rate variations (SRV-DB)

Kota TAKAHASHI†, Keigo TSUTAKI†, and Toru YOSHIHARA†

† Department of Information and Communication Engineering, The University of Electro-Communications
Chofugaoka 1-5-1, Chofu-shi, Tokyo, 182-8585 Japan
E-mail: †{kota,tsutaki,yosihara}@ice.uec.ac.jp

Abstract A specialized speech database is required for the studies of an efficient listening method for general people or a reliable listening method for aged people. We are constructing such type of speech database. This research was supported by the Grant-in-Aid for Scientific Research for the period from 2008 to 2010. In this paper, we are aiming to demonstrate the beta version of the database in order to hear valuable opinions and new ideas from many researchers in this field. A recording system for controlling speaking rate will be also demonstrated. The speech database and software of the recording system will be public, so we are planning to extend this database by the cooperation of researchers who are interested in studies in this area.
Key words speech database, speech rate conversion, time scale modification, efficient listening

1. はじめに

我々は，平成 20 年度～平成 22 年度の科研費の研究（基盤研究 C 課題番号 205011「フレキシブルな時間軸を持つ高効率音声再生法の研究と研究者用音声データベースの研究」）の一環として，さまざまな話速で録音した音声データベースを構築している．

作成している音声データベース (SRV-DB) は，研究グループ内での利用にとどまらず，音声の研究者，および音声を勉強している学生に利用して頂くことを想定している．

様々な話速を正確に管理して発声させることは，特別な原稿提示システム無しには困難である．そこで今回，我々は，音声データベースの作成に先んじて，専用の原稿提示システム (ReCoK5) も開発した．

音声データベース (SRV-DB) を公開するだけでなく，原稿提示システム (ReCoK5) も公開する．これによって，本研究グ

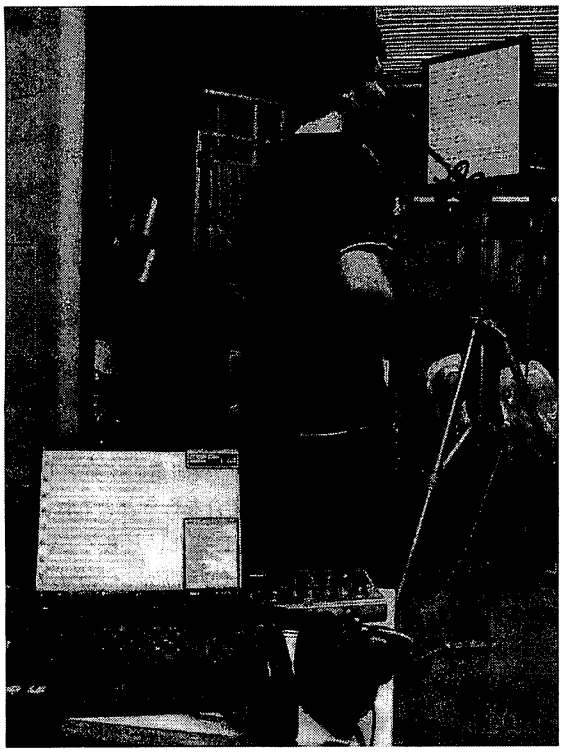

図 1　原稿提示システム ReCoK5 による録音風景.

Fig. 1　Experimental setup for recording speech using ReCoK5.

ループ外の研究グループであっても，SRV-DB にデータを追加していくことが可能である.

従来，音声データベースは，きちんと組織されたグループの中で，系統だって作られてきたと思う．今回の音声データベースは，話速のバリエーションに富んでいるという特徴の他に，参加の意図があれば誰でも音声データベースにデータをつけ加えられるという特徴を持つ．このような試みが成功するのか否かも，この研究の隠れた目標なのである.

したがって本研究は，ひとりよがりにならずに，要所要所で音声研究の専門家の意見を広く聴取し，それを取り入れていくことが重要となる．今回の発表は，その意見聴取の第一回でもある.

このような音声データベースが本当に役に立つのかどうか，役に立つとすればどんな点か，役に立たないとすれば何処を改善すれば良いのか．研究者の方々の生の感想や意見をぜひお聞きしたい．そして，将来，この音声データベースを利用してくれる研究者や，データベースへの音声追加に協力してくれる方々がひとりでも多く現れて下さることを期待している.

本稿を以下のように構成する．第 2 節で，話速バリエーション型音声データベースの構築を思い立った動機について述べる．第 3 節では，原稿提示システム (ReCoK5) について述べ，第 4 節では，音声データベース (SRV-DB) の作成の進行状況について紹介する他，現在までの作業で気が付いた点や考えた点について，いくつかの興味深い事項をひろって紹介したい．第 5 節では，現時点で公開している音声データベースとソフトウエアについて告知する.

2. SRV-DB 作成の動機

我々は，数年前から，ＨＤＤレコーダなどに記録された音声の内容を出来るだけ短時間で効率よく聞き取らせる手法について研究してきた.

例えば，[2] においては，早送り再生でユーザの指示速度のまま再生してしまうと，多くの部分で聞き落しが生じるため，ユーザの指示速度のまま再生させるのではなく，それを若干修正した速度変化で再生することを提案した.

このとき，修正速度を最適問題の解として求めるためには，聴取者が音声を聴き取れるか聴き落すかの違いが，どのような指標で記述できるかを実験的に求めなければならなかった．我々は，専用の DSP システムを作り，ユーザが音声を聞き取り可能な速度変化の限界を調べた．しかし，再生速度に関する不等式によって聞き取りの限界を定式化しようとしても，その不等式は，当然のことながら，ソース音声の話速に依存してしまうので，シンプルな定式化は困難であった.

そこで，[3] においては，ソースの話速を変化させて，この限界をさらに精密に求めることに挑戦した．この研究においては，聞き取り限界を示す式のパラメータを決定するために，話速を管理した音声信号を出来るだけ多く用意する必要があった．そのような音声ファイルは探しても [1] 存在しなかったので，我々は，それを研究室内で作成した.

そのとき痛感したのは，話速を管理して発声させることの難しさと，それを大量に作成することの，さらなる難しさであった.

「同一話者が，同一文章を，異なる話速で読み上げてくれる音声データベースが存在したら，この領域の研究は，どんなに進むことだろう！」と我々は思った.

効率的な再生法を実装するためには，話速の自動推定技術の開発は必須である．我々は，音声データを解析することで話速を推定する研究も行っている [4] が，ここでも，話速にバリエーションを持たせた音声データベースの必要性を強く感じた.

以上のような経緯で，話速バリエーション型データベースを構築することになったのである．幸い，科研費の研究として応募したところ予算が認められたので，その予算を使って 2008 年 4 月より研究を始めている.

しかし，なにぶん大学内の一研究室であるので，施設，時間，予算ともに限られている．そこで，このような特殊なデータベースを作成するにあたって，できれば同士を募り，できるだけ多くの音声を収集しようと考えた．すなわち，データベースの構築法にも特徴を持たせようと考えたのである.

具体的には，音声データベースは完全無料で公開することはもちろん，独自に開発した原稿提示システムのソフトも公開し，音声の研究に興味のある学生の方々も含めて，不特定多数で作る音声データベースという新しい方向をめざしている.

3. 原稿提示システム ReCoK5 の作成と配布

紙に原稿を印刷してアナウンサーに渡しただけでは，正確に話速を制御して読みあげてもらうことはできない．

そこで，専用の提示システムを作成した．この提示システムは，Windows パソコン上で動作する．提示機能の他，リアルタイムでの音声解析機能と録音機能も持たせてあるため，“**R**ecording system for **C**ontrolling speaking rate with **K**ind interface, version **5**” と呼べるものであり，略して，**ReCoK5** と名付けた．

3.1 ReCoK5 の概要

図1に，ReCoK5 を用いた録音の様子を示す．ここでは，ReCoK5 を操作する人を，オペレータと呼ぶことにする．ReCoK5 の典型的な利用方法においては，ReCoK5 をインストールしたパソコンに，2台のディスプレイを接続する．1台はオペレータ用，もう1台はアナウンサー用である．

図2が，ReCoK5 を立ち上げた直後の初期設定画面である．オペレータは，ここで原稿を選択し，話速を設定する．「開始」ボタンを押すと録音を行う画面へ進むが，録音を行った後でも，この画面へ戻ることができるため，複数の原稿や話速で，続けて音声収録を行うことも可能である．

図3が，録音時のオペレータ画面である．時間進行に応じて，原稿の直下に録音された音声波形と，そのスペクトログラムが表示される．原稿の文字は，発声すべきタイミングの位置に置かれているため，原稿の文字と，音声波形やスペクトログラムとを上下比較することで，正しいタイミングで発声が行われたかをリアルタイムで監視することができる．

一方，図4が，アナウンサー画面である．オペレータ画面においては，原稿の文字は読み上げるべき時間に応じた位置に置かれていたが，アナウンサーに対してこれを行うと文字の間隔がバラバラになり読みづらくなるため，文字は等間隔に配置してある．録音の開始とともに，この原稿の上を，時間進行に応じて，色付けしていく．文字を等間隔に配置したために，色付けの進行速度が一定ではなくなるという問題はある．しかし，

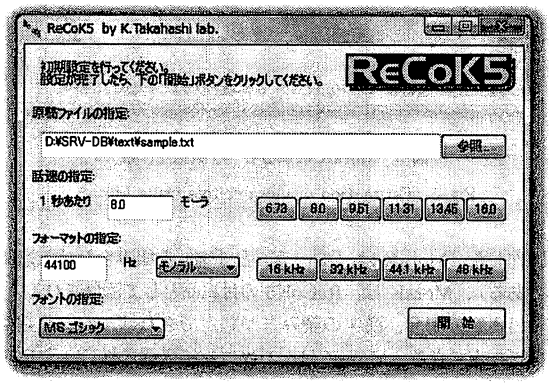

図 2 ReCoK5 の立ち上げ画面.

Fig. 2 The initial setup window of ReCoK5.

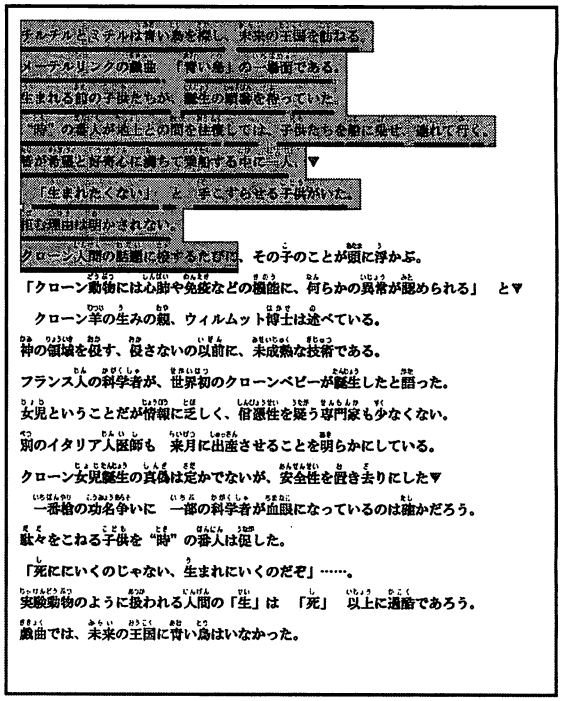

図 4 ReCoK5 のアナウンサー用画面.

Fig. 4 The announcer window of ReCoK5.

カラオケの画面と同じ表示方法であることもあってか，本システムを初めて利用する一般の人間にとってはこのほうが見易いようである．

さて，原稿だけを表示させて予備的な収録を行ってみたところ，アナウンサー役の人間から問題点が指摘された．

話速が速くなってくると，どうしても，指示された話速で発話ができず，頻繁にオペレータからＮＧが出るのであるが，なぜ自分の発話がＮＧなのか，納得することができないということであった．

これは，アナウンサーにとってストレスをためる要因になるだけでなく，どこが悪かったのかを自分で納得できなければ，再度発話しても，同じミスを繰り返すだけであるという問題につながっている．

そこで，アナウンサー用画面にも，収録音声の分析結果を表示することにした．

ただし，アナウンサー画面にスペクトログラムを表示しても，専門外の人間には理解不能であるし，意味がわかる人間であっても細かなところまでは見ることができない．このため，ReCoK5 では，アナウンサー画面には，収録された音声の強度を色の帯の濃淡に置き換えて原稿の真下に表示することにしている．

このような工夫を加えたところ，アナウンサーが納得してくれるというだけでなく，オペレータがアナウンサーに発話のどこが悪かったかを連絡する作業も効率的になるという効果があった．

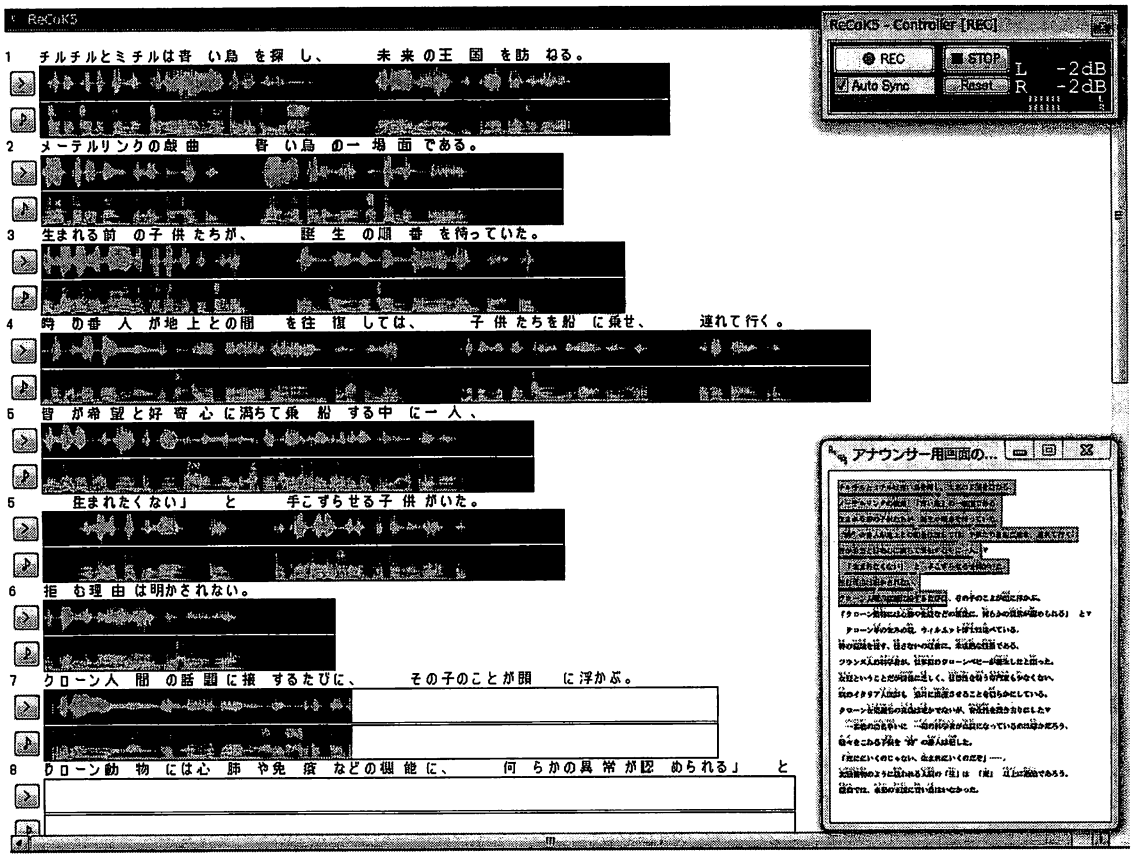

図 3 ReCoK5 のオペレータ用画面.

Fig. 3 The operator window of ReCoK5.

ところで，アナウンサーの文頭の発話開始と，画面の色付け開始は，同時でなければならない．当初は，ある文の原稿を読み終え，その文のすべての文字に色が付き終えた瞬間に，次の文頭を読み始める，という方式を採用していた．しかし，文頭の発声タイミングがなかなか合わず，アナウンサーの負荷が重かった．

そこで，アナウンサーは，自由なタイミングで読み始めてよいこととした．すなわち文頭の読み始めを入力レベルが閾値を超えることで検出し，この検出をトリガーとして自動的に色付けを開始するよう仕様を変更した．

これにより，タイミング合わせが楽になり，より発話に余裕のある，自然な音声の収録が可能となった．

3.2 MoraK の概要

一秒あたりの読み上げモーラ数を設定値になるように ReCoK5 のカーソルを動かすためには，ReCoK5 がテキスト各文字のモーラ数を知っている必要がある．

しかし，原稿のテキストファイルには，モーラ数の情報（例えば,「日本」は 3 モーラであるという情報）はないので，これを何らかの方法で取得しなければならない．

この作業を楽に行うソフトウエアとして，我々は **MoraK** を製作した（MoraK は,「モーラ」と「楽」を組み合わせた造語である）．MoraK は，ReCoK5 の付属品として添付される．

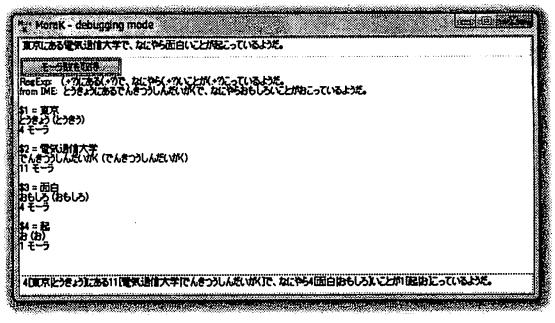

図 5 MoraK によるモーラ数推定作業.

Fig. 5 Estimation of mora counts using MoraK.

日本語の場合，漢字の読み方が判定できれば，モーラ数はほぼ推定できる．MoraK は，パソコンに搭載されている IME（かな漢字変換）機能を利用してモーラ数を推定する．図 5 に，MoraK で作業している様子を示す．なお，当然のことながら，IME で完全にモーラ数を推定することはできないので，MoraK

の出力結果に不備があれば，最後に人間の手作業で修正を加える必要がある．

4. 話速バリエーション型データベース SRV-DB

SRV-DB の作成では，研究室内のメンバーによる予備的な収録の他，専門のアナウンサーやナレータに発話を依頼することにした．

ナレータやアナウンサーと呼ばれる専門家の発話をデータベースに含めることは，(1) 明瞭な発話であるために研究価値が高い，(2) テレビやラジオなどの音声を早聴きする際には，専門家の発話を対象とすることになるので，専門家の発話もデータベースに含めておく必要がある，の２点から重要である．

それだけでなく，一般人にも馴染みのある声優やナレータなどにも発話を依頼することによって，より多くの若者に音声の研究に興味を持ってもらうきっかけとなることも狙っている．

本研究をはじめるにあたって，複数のアナウンサー事務所，ナレータ事務所に相談した．今回は，研究の意義をご理解頂いた事務所から３名のアナウンサーを派遣してもらい，電気通信大学内で収録を行った．

読み上げ原稿は，現在２種類を採用している．第一は，ATR 音素バランス 503 文の中の最初の 25 文である．音素バランス上の観点から言えば，１セットに相当する 50 文を用いるべきであるが，予備的に読み上げ実験を行ったところ，一人のアナウンサーに同一日に 50 文全てを数種類の話速で発音してもらうことは負担が重すぎると判断し，25 文となっている．以下，この原稿を「ATR25文」と呼ぶ．

ATR25文は，文ごとに独立しているので，話速を精密に制御した音声ファイルを製作するのに向いている．このような音声ファイルは基礎的研究を行うには必須である．一方，話速変換させた文章を被験者に聴かせて意味がとれるかどうかを試すためには，文と文のあいだに意味的なつながりがあり，全体としてストーリーがあるような，手頃な長さの文章も必要である．

そこで，文間に連続性のある第二の原稿として，読売新聞のコラムである編集手帳を用いることにした．過去数年分の編集手帳を調べ，あまり政治的・時事的にならず，しかも興味深い内容という観点で，３つの記事を選んだ．その後，読売新聞社の著作権関係の担当の方と相談した．幸いなことに研究の意義をご理解頂くことができ，研究利用という前提で，音声と原稿の無料配布を認めて頂いている．そのため，誰でもこのファイルをダウンロードして研究に利用することができることになっている．

現在まで，録音作業は３回行われている．第１回の録音は，専門のアナウンサー３名を招き，ATR25文を中心に録音した．第２回は，設定速度を再検討し，研究室内の人間４名を使ってATR25文を録音した．第３回の録音では，第一回の録音と同じ話者を再度招き，編集手帳の朗読を録音した．

第１回の録音の詳細について述べよう．まず，ReCoK5 により原稿提示して，それぞれの文について，5, 8, 11, 16 [モーラ/秒] で読み上げてもらった．ATR25文を４つの話速について発声してもらう時間は，録り直しを含めてひとり１時間半程度要

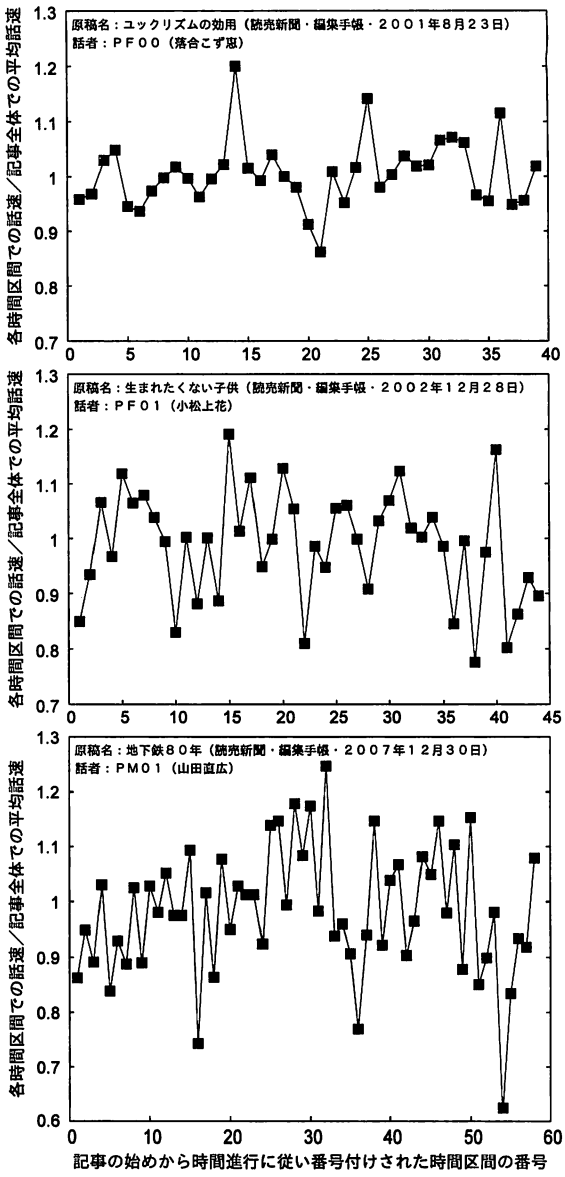

図 6　自由話速での発話における話速変動．

Fig. 6　Fluctuations of relative speaking rate without controlling speaking rate.

した．このうち，16 [モーラ/秒] については，専門のアナウンサーであっても原稿によっては発声がほとんど不可能な箇所があり，データベースとして公開するには不十分と判断し，今回は公開していない．

ATR25文の収録後に，３名の方に感想を聞いたところ，一般にアナウンサーやナレータは，ゆっくりした話速で，感情や個性を出して発声するのが得意であり，超高速の発話はほとんど依頼されることがないので発声困難であるとのことであった．

さて，第一回の録音の最後の残り時間を利用して，編集手帳の３記事を３名の方に，それぞれ読んで頂いた．このとき，あ

えて話速を制御せずに自由に発話してもらった（以下，自由話速と呼ぶ）ところ，3名の方ともに，話速を制御したときとは比較にならないほど魅力的な声質と語り口であった．

この差には，実験参加者一同が大変驚くとともに，以下の解析により，今後の収録法に関する以下の重要な示唆が得られた．

まず，解析の第一歩として，自由話速のときの話速の変動を調べた．息継ぎをせずに一気に読まれる長さを解析時区間として（これはほぼ，句読点で区切ることに等しい），各解析時区間の話速を解析した結果を紹介しよう．

図6は，横軸に区間番号をとり，縦軸にその区間での話速をとって，文章が進行するに従って話速がどのように変動するかを調べたものである．3名の話者の記事全体での平均話速は，話者 PF00 が 7.83[モーラ/秒]，話者 PF01 が 8.11[モーラ/秒]，話者 PF02 が 8.74[モーラ/秒] であった．図6では，この平均話速を 1.0 として規格化した相対話速で縦軸を目盛ってある．

図6を見ると，自然な朗読では，10%程度の話速変動は常にあり，ときに 20%を超えることもあることがわかる．

さらに，図6において大きな外れ値となった時区間をリストアップした結果，ストーリー全体の要となる重要な時区間は遅い話速で読まれていること，重要でない時区間は速い話速で読まれていることがわかった．また，「$A \rightarrow B \rightarrow A' \rightarrow B'$」の構造を持った部分では，話速も「速→遅→速→遅」のように変化させることで，構造を理解させ易くしていることもわかった．

話速はランダムに変化するわけではなく，原稿の内容や読み手の意思によって，意図的に制御されているのである．

したがって，話速を完全に制御してしまうことは，この意図を表現させないことにつながり，発話の専門家が魅力的表現をすることを妨げていたわけである．

話速にも音声としての情報が含まれている．この情報を捨てて全ての文を一定話速に制御してしまうと，発話がしにくく，文章としての自然さが失われるだけでなく，本来伝えられたはずの情報が伝わらなくなると考えられる．これは問題である．

そこで今回は，ReCoK5 に手を加え，MoraK の出力のモーラ数ファイルに図6の解析結果を書き加えると，ReCoK5 は相対速度を一定に保ったまま平均速度が設定値であるような原稿提示ができるように機能拡張を行った．

第三回の収録では，このようにして作った速度指示に基づいて発話を行ってもらった．自身の最も自然なタイミングを，均一に縮小あるいは拡大した時間軸で発話するのであるから，アナウンサーにとっても発話しやすかったようである．このため，収録した音声を試聴したところ，自然さもあり，記事の内容も楽しめる朗読になっていると思う．

図7 SRV-DB を試聴・ダウンロードする公開ページ

Fig. 7 The SRV-DB web page for hearing and downloading files.

5. 公　　開

ReCoK5 と SRV-DB については，http://www.it.ice.uec.ac.jp/SRV-DB/ で公開している．ぜひ，訪問して頂き，コメント等をお寄せ頂きたい．

ReCoK5 の標準的な使い方はデュアルディスプレイである．しかし，ノートパソコン1台で，原稿提示と録音の全てを行うこともできるので，個人で手軽に利用することも可能である．我々が想定していなような意外な利用法もあるかもしれない．

図7に，SRV-DB の暫定公開のページを示す．各文を試聴できるほか，話速ごと，話者ごとに一括ダウンロードできるようにしてある．また，話者の顔写真のサムネイルも掲載してある．これは，研究そのものには直接関係ないが，ややもすれば無味乾燥になりがちな音声解析の研究に，人間の音声の魅力を味わうという意味での潤いをもたせようという新しい試みである．

文　　献

[1] 音声資源コンソーシアム (SRC)，音声コーパスリスト，http://research.nii.ac.jp/src/list/index.html

[2] 高橋弘太，“フレキシブルな時間軸による最適な速度曲線での音声再生，” 信学技報，vol.107, no.116, pp.37-42 (2007)

[3] 吉原亨，高橋弘太，“話速適応性を有するフレキシブルな時間軸による音声再生，” 信学技報，vol.107, no.234, pp.19-24 (2007)

[4] 吉原亨，蔦木圭悟，高橋弘太，“音声の高速再生のための話速推定法と高速発話時の特性解析，” 信学技報，本件の次の発表 (2008)