論文 / 著書情報
Article / Book Information

論題(和文)	『話し言葉工学』プロジェクトのこれまでの成果と展望
Title	
著者(和文)	古井 貞熙, 前川 喜久雄, 井佐原 均
Author	SADAOKI FURUI
出典(和文)	第2回 話し言葉の科学と工学ワークショップ講演予稿集, Vol. , No. , pp. 1-6
Journal/Book name	, Vol. , No. , pp. 1-6
発行日 / Issue date	2002, 2

『話し言葉工学』プロジェクトのこれまでの成果と展望

古井 貞熙 [a,b]、前川 喜久雄 [b]、井佐原 均 [c]

[a]東京工業大学大学院情報理工学研究科計算工学専攻
〒152-8552 目黒区大岡山 2-12-1, Email: furui@cs.titech.ac.jp
[b]文化庁国立国語研究所
〒115-8620 北区西が丘 3-9-14, Email: kikuo@kokken.go.jp
[c]総務省通信総合研究所
〒619-0289 京都府相楽郡精華町光台 2-2-2, Email: isahara@crl.go.jp

あらまし　話し言葉音声の分析と認識を目指して平成 11 年度に開始した「話し言葉工学」プロジェクトのこれまでの成果を紹介し、今後の展望を述べる。プロジェクトは次の３つの目標をもっている。(1) 大規模話し言葉コーパスを構築する、(2) 話し言葉音声の言語モデルと音響モデルを構築し、音声認識理解および要約技術を構築する、(3) 話し言葉の自動要約システムのプロトタイプを作成する。これまでに、講演音声に関する収録をほぼ終了し、大規模なコーパス構築作業を、計画を上回るテンポで進めている。講演音声の認識において 70%の単語正解精度を達成するとともに、認識誤りの要因に関する種々の分析を行った。音声自動要約法と、その評価法を提案した。このように着実な進展が得られているが、解決しなければならない難しい研究課題も多数残っている。
キーワード　「話し言葉工学」プロジェクト、話し言葉コーパス、話し言葉音声認識理解、音声要約

Intermediate Results and Perspectives of the Project “Spontaneous Speech: Corpus and Processing Technology”

Sadaoki Furui, Kikuo Maekawa, Hitoshi Isahara

Tokyo Institute of Technology, Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan, Email: furui@cs.titech.ac.jp
National Language Research Institute
3-9-14 Nishigaoka, Kita-ku, Tokyo, 115-8620 Japan, Email: kikuo@kokken.go.jp
Communications Research Laboratory
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289 Japan, Email: isahara@crl.go jp

Abstract　This paper presents intermediate results and perspectives of the “Spontaneous Speech Engineering” Project started in 1999 aiming at analysis and recognition of spontaneous speech. The project has the following three major themes: 1) building a large-scale spontaneous speech corpus, 2) acoustic and linguistic modeling for spontaneous speech recognition, understanding and summarization, and 3) constructing a prototype system for spontaneous speech summarization. Recording of lectures for the corpus has been almost completed, and the corpus building process is being carried out faster than originally scheduled. 70% word accuracy has been achieved in lecture speech recognition, and several analyses of the sources of recognition errors have been performed. Automatic speech summarization technique and its evaluation method have been proposed. Although such a steady progress has been achieved, there still exist many difficult research issues that we have to solve.
Key words　“Spontaneous Speech Engineering” Project, spontaneous speech corpus, spontaneous speech recognition and understanding, speech summarization

１．はじめに

話し言葉でコンピュータシステムと対話をしたり、講演や放送などの話し言葉を文字化し、その内容を把握し、再利用や情報検索に使えるようにしたいという要望が高まっている。そのためには、言い直し、言い淀み、繰り返し、間投詞、不正確な発音などの現象を含んだ自然な話し言葉を音声認識できるようにすることが不可欠である[1]。

本プロジェクトは、このような話し言葉処理に関わる種々の重要な問題を取り上げ、話し言葉からその意味・内容・話し手の意図などを自動的に抽出する技術の基盤を確立することを目的とし、科学技術振興調整費を財源として、平成 11 年度に開始された[2]。平成 15 年度まで５年計画で進められる予定である。組織の主体は、国立国語研究所、総務省通信総合研究所、および東京工業大学であるが、国内外の大学や研究機関からの研究者が非常勤研究員として参加している。

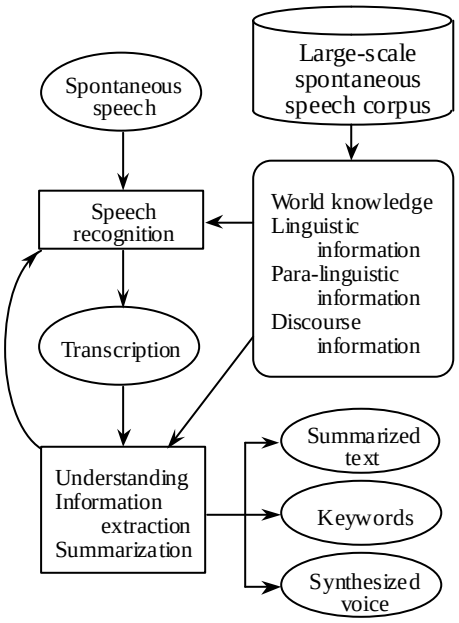

図１　プロジェクトの概要

プロジェクト全体の概要を図１に示す。本プロジェクトでは、話し言葉の大規模コーパスの構築と、それを用いた「話し言葉工学」の構築を目標としており、次の３つのサブテーマを持つ。

サブテーマ１：大規模話し言葉コーパスの構築

サブテーマ２：話し言葉を音声認識・理解・要約するための基本技術の構築

サブテーマ３：話し言葉の音声要約プロトタイプシステムの構築

本報告では、本プロジェクトの前半の成果を概観し、後半に予定している研究内容について紹介する。

２．話し言葉コーパスの構築

2.1 話し言葉コーパスの重要性

現在の音声言語処理（音声認識および音声合成）技術は、コーパスに基づく統計的手法に基礎をおいている。これによって近年の技術進歩が達成されたといっても過言ではない。しかし、この手法が機能するためには、大規模なコーパスが必要である。書き言葉に関しては、新聞記事などの大量の原稿を比較的容易に用いることができるので、これまでの音声認識のためのコーパスも、ニュース原稿を含めて、書き言葉をベースとするものが主として用いられてきた。しかしこれでは真の話し言葉をモデル化することができず、話し言葉の音声認識で高い精度を達成することはできない。話し言葉のコーパスを構築するためには、音声を録音して書き起こし、それを解析して種々の情報を付与する必要があるため、大量なコーパスの構築は容易ではない。米国、ヨーロッパなどを見ても、まだ話し言葉コーパスは十分とは言えないが、日本では特に、タスクを限定した対話音声を除くと、まだ比較的小規模なコーパスしかできていない。

2.2 コーパスの構築状況

本プロジェクトでは、タスクを限定しないモノローグを主体として、のべ約 600-700 時間、形態素にして約 700 万語（形態素）規模の話し言葉コーパス（CSJ; Corpus of Spoken Japanese）の作成を目標としている[3]。モノローグを基本とするのは、対話音声の処理に関しては別のいくつかのプロジェクトが活発に活動しており[4]、対話の流れに関する対話音声特有の重要な研究課題があるので、活動をある程度分けて分担した方がよいと考えるからである。対話音声でも、特にコンピュータシステムとの対話などでは、各文を正しく認識するのが基本と考えられるので、このプロジェクトでのモノローグを対象とする研究の成果は、対話音声処理技術の進展にも寄与できると考えている。本プロジェクトで構築するコーパスの中にも、インタビューなどは一部に含む予定である。全体のコーパスの大きさは、音声認識のための統計的言語モデルを構築するために最低限必要なデータ量に基づいて決めた。

コーパスの発声者については、あまり厳格な制限は設けず、文法と分節的特徴が共通日本語であれば、地域差や個人差に起因するアクセントやイントネーションの変動は容認している。音声のソースとしては、国内の種々の学会での講演と、国語研究所での設備を用いた模擬講演（一般の話者が簡単なメモを頼りに日常的な話題について 10 分間程度スピーチを行ったもの）を収録している。音声はヘッドセット型の接話型指向性マイクロホンを話者の頭部に装着して、48kHz、16bit で DAT に収録する。

コーパス全体について、セグメンテーション、書き起こし（仮名漢字による基本形と発音形）、形態素解析などを行うが、全体の約 10%（50 万語）の「コア」については、韻律情報などパラ言語情報まで含めたコーパスとする。話し言葉に関しては、その表記法や、形態素解析の単位などがまだ確立していない。単に文の切れ目を検出するだけでも、決して容易ではない。話し言葉音声の曖昧性のために、真剣に考えるほど解決が難しくなるような問題も多い。大規模な話し言葉コーパスを作成するためには、多くの作業者が長い年月に渡って作業することが必要で、質を揃えるためには、その作成基準を明確にすることが重要である[5]。この基準作りも、プロジェクトの重要な役割の一つと考えている。書き起こし法に関しては、間投詞、言い淀み、言い誤り、雑音などの記述法を含めて、約 300 ページにわたるマニュアルを作成した。このマニュアルは、国内外の５研究機関からの要請に応じて提供している。表記の変動を減らすためのソフトウェアツールも作成した。こうして、多数の作業者による変動を減らすとともに、プロジェクト担当者による綿密なチェックを行っている。形態素単位については、言語的分析を意識した比較的短い単位（短単位）と、音声認識を意識した比較的長い単位（長単位）を用いることとし、短単位に関しては、ほぼ確定している[6]。長単位の決定はこれからである。

これまでに約 500 時間（約 600 万語）の講演の収録と、約 400 万語の書き起こしを終了し、当初計画よりも速いペースでコーパス作成作業を進めている。コアの部分に関する、分節音ラベリングおよび J_ToBI による韻律ラベリングも始めている。今後は、インタビューや、これまでに収録した講演の書き起こしを同じ話者が朗読した音声の収録などを行う予定である。作成したコーパスは、平成 12 年 12 月に一部の公開を行い、平成 13 年 6 月に、第一回のモニター公開（86 時間、95 万語相当）を行った。130 件以上の多数の方々からの申し込みをいただいた。今後も段階を追って公開する予定である。

３．話し言葉の音声認識・理解

3.1 音響モデルと言語モデルの構築

本プロジェクトにおける「話し言葉工学」の主たる目標は、話し言葉の音声認識理解と、その自動要約である。話し言葉の音声理解のためには、話し言葉に合った音響モデルと言語モデルを構築することが必要である。このため、コーパスを構成する大量の書き起こしテキストについて、自動的に形態素解析する。このためには、話し言葉に適応した形態素解析ツールを作成することが必要である。これもプロジェクトの重要な目標の一つである。現在、ChaSen を用いる方法と、最大エントロピー法による方法について検討している。このために、コアの部分については、人手による処理を含めた精密な解析を行い、形態素解析ツールの学習に用いる。

コーパス全体から評価用コーパスを除いた部分の解析結果に基づいて、話し言葉の統計的言語モデルを作成する。上でも述べたように、話し言葉には、言い直し、言い淀み、繰り返し、間投詞、話し言葉特有の言い回しなど、書き言葉にはない多様な変動要因があるので、単に形態素を単位とする統計的言語モデル（バイグラム、トライグラムなど）を作成すればよいというわけにはいかない。話し言葉特有の変動に言語モデルとしてどう対処するか、どのようにデコーディングすればよいかについては、まだ有望な方法が見つかっていない。この解決のためには、これまでの、すべての単語（形態素）を文字化しようとする音声認識のアプローチから、不要な単語は無視し、話し手が伝えようとした内容を抽出することを目標とする音声理解への展開が不可欠である。

いくら大量のコーパスを構築しても、将来の音声認識理解の対象となるすべてのタスク（応用場面）をカバーすることはとうていできない。したがって、話し言葉に含まれるタスクに独立な現象と、タスクに依存した現象を分離したモデルを構築し、タスクに依存する部分について、新しいタスクに自動的に適応する方法を作り上げることが不可欠である。

話し言葉の音響モデルの学習も重要な課題である。現在の音声認識システムで用いられている音響モデルは、主として書き言葉を読み上げた音声から学習された環境依存型音素 HMM（hidden Markov model；隠れマルコフモデル）である。これまでの種々の実験で、

読み上げ音声と通常の話し言葉音声では、音響的特徴が大きく異なることがわかっている。話し言葉を音声認識するための最適な音響モデルの構造と、その学習法、さらにタスクや話者に自動的に適応化させる方法についても、研究を進めている。

3.2 話し言葉の音声認識の状況

種々の学会で収録された講演音声をテストセットとして、音声認識実験を行っている[7][8][9]。音響モデルや言語モデルの学習には、多数の学会で収録した音声と、その書き起こしを用いた。テストセットとの話者の重複はない。デコーダには Julius3.1 を用いた。その結果、次のようなことがわかった。

(1) 読み上げ音声から作成した音響モデルと、Web 上の講演録から作成した言語モデルを用いた場合に 55%程度あった誤認識を、CSJ から作成したモデルに置き換えることにより、35%程度にまで低下させることができた。
(2) 発声速度、言い直し頻度、および未知語率が、話者による認識率変動の主要な要因となっている。
(3) フィラーの回数も個人差が大きい。
(4) 続けて発声されている話し言葉から、文単位の音声を自動的に切り出して認識するのは極めて難しい。適切なポーズで区切れば大きな性能低下はないが、文単位に区切らずに連続してデコーディングするのが有効。
(5) まだ話し言葉のコーパスの量は少ないが、それを用いた音響モデルや言語モデルは、書き言葉やそれを読み上げ音声をもとに作成したモデルよりも極めて有効性が高い。
(6) 入力音声の話題（分野）が学習に用いられたコーパスでカバーされていないときには、その分野の教科書を用いて適応化するのが有効。
(7) 声の個人差に対処するため、教師なし話者適応が有効。
(8) 話者適応後でも単語正解精度はまだ 70%程度であり、評価法を含めて、研究課題が多い。

４．話し言葉の音声要約

4.1 音声要約法

音声を理解するとは一般的に何を指すかは、まだ必ずしも明確ではない。本プロジェクトでは、音声理解の一つの形態として、音声の自動要約技術の実現を目標としている。最終的な要約結果の出力形態には種々のパターンが考えられるが、正しく要約することができたと判断できれば、ある意味で音声が理解できたと考えることができる。音声認識結果を用いて、音声を自動要約し、要約された内容をキーワード、要約文、その合成音声などで提示する方法について研究し、プロトタイプシステムを構築する。音声要約の結果から話題を特定し、その情報を音声認識部にフィードバックすることによって、音声認識理解の精度を向上させることもできると期待される。

音声要約技術の応用としては、次のようなものが考えられる。

- 音声を含む大量のマルチメディア情報の検索のための、インデックスの自動付与
- 放送などへの字幕付与の自動化など、福祉技術の向上への寄与
- 会議・講演・講義などの速記や文字起こしの自動化の促進

これまでに、音声から自動的に抽出したキーワード（話題語、重要語）を核として、言語モデルに基づいて、各文ごと、あるいはひとつの講演全体を短縮した要約文を自動的に作成する研究を進めている [10][11]。各講演から重要な文を抽出する研究も行っているが、重要文抽出だけでは、抽出されなかった文に含まれる情報が完全に欠落して、重要文だけをつなぎ合わせたものでは意味がとれないことが少なくない。さらに、音声認識結果にはどうしてもある程度の認識誤りや不確実性が避けられないので、音声要約は、書き言葉の要約とは本質的に異なる。文の意味的構造まで考慮した方法を構築する必要がある。

最近では、DARPA を始めとして、欧米でも音声要約への関心が高まっており、新しいプロジェクトも計画されている。

4.2 音声要約の状況

音声認識結果を用いて、上で述べたように、各文ごとおよび複数文をまとめて要約する手法について検討を進めた。この方法では、音声認識結果から、与えられた要約率（圧縮率）で、要約スコアを最大にする単語（形態素）セットを自動的に選び、それを接続した文を出力する。要約スコアは、各単語の重要度（重要度スコア）、認識結果としての各単語の音響的および言語的信頼度（信頼度スコア）、選ばれた単語を接続したときの言語尤度（言語スコア）、および認識結果の文中における各単語の係り受け関係（構造）に基づく単語間遷移スコアの重みつき和からなる。係り受け関係は係り受け SDCFG (Stochastic Dependency

Context Free Grammar) によって表す。この要約スコアを最大化することにより、意味的に重要で、音声認識の信頼性が高い単語（形態素）を、文法的かつ意味的に正しい連鎖になるように選び、出力することができる。スコアを最大にする単語セットは、動的計画法によって容易に選ぶことができる。

これまでにこの要約手法を日本語および英語ニュース音声と、このプロジェクトの講演音声に適用し、評価を行ってきた。評価法として、多数の被験者が単語削除法で作成した要約結果をまとめて作成した単語ネットワークを用いる方法を提案した。このネットワークは、被験者による正解要約文の変動を表していると考えられる。このネットワークから、評価対象の要約文に最も近い単語列を抽出し、一致度を、音声認識における正解精度を測る場合と同じ方法で測る。多数の被験者がニュース音声の書き起こしから作成した要約文を正解として、種々の長さの単語連鎖の重複度によって評価する方法についても検討した。その結果、上で述べた要約手法が有効であることが確認されている。

５.むすび

「話し言葉工学」プロジェクトの前半約 2 年半を経過した状況を紹介した。これまでに着実に進展してきているが、話し言葉特有の極めて難しい種々の技術的課題（韻律ラベリング、韻律情報の利用法、形態素解析ツール、間投詞、言い直しなどのモデル化、言語モデルのタスク適応、話し言葉の音響モデル、音声要約のための構文および意味表現法など）に直面しており、これらの解決のためには、種々の新しいアプローチを開拓する必要がある。

今後、我が国で行われている種々の音声認識関連大型プロジェクトと適宜情報交換を行うなど、連絡を密にし、共有できるデータベース（コーパス）は共有していきたい。音声や言語のデータベースは、いくらあっても十分ということはなく、言葉は常に変化し新しい言葉が作られるので、データベースの構築と維持管理をプロジェクトの期間だけで終わらせず、継続できる態勢を作ることが必須である。これについては、社会的コンセンサスを得るための働き掛けをしていく必要がある。

このプロジェクトの目標達成のためには、情報科学・工学系だけでなく、音声学、言語学などに代表される人文系言語諸分野を含めた多面的かつ総合的なアプローチが必要である。このため本プロジェクトは、多様な研究者から構成されており、この面でも新しい研究環境と、一体感のある協力関係を開拓していきたい。

本プロジェクトに関連して、海外の主要関係機関とも情報交換を行っている。話し言葉の音声認識理解や要約は、世界の共通の研究課題であるので、コーパスの作成法を含めて、海外との連携も深めていく予定である。平成 15 年 4 月には、香港で開かれる ICASSP のサテライト行事および ISCA ワークショップとして、話し言葉処理に関する国際ワークショップを日本で開催する予定である。

本プロジェクトでは、この分野における国内外の 10 名のリーダーの方々に評価委員（評価委員長は京都大学の長尾真総長）をお願いしており、平成 13 年 8 月に評価委員会を行った。本プロジェクトの成果に関して高い評価をいただくとともに、今後の進め方に関する貴重なサジェスチョンや提案を多数得ることができた。本プロジェクトが、真の意味での話し言葉の認識・理解技術の構築に、役に立てれば本望である。関係各位のご協力をお願いするとともに、広くご意見やご希望をお受けしながら進めていきたい。

謝辞

コーパス作成のための音声収録にご協力いただいている、各学会および NHK 放送技術研究所に感謝する。非常勤研究員として本プロジェクトに参加し、貴重な意見やコメントを提供して下さっている大学や研究機関の方々に感謝する。

参考文献

[1] 古井：“音声ドキュメントの文字変換と情報抽出”、映像情報メディア学会誌、55、11、pp. 1394-1396 (2001)

[2] 古井、前川、井佐原：“「話し言葉工学」プロジェクトの概要と展望”、話し言葉の科学と工学ワークショップ講演予稿集、pp. 1-6 (2001)

[3] 前川：“「日本語話し言葉コーパス」の構築”、同、pp. 7-12 (2001)

[4] 小特集—音声研究の新たな方向を探る--、音学誌、56、11、pp. 746-782 (2000)

[5] 小磯：“「日本語話し言葉コーパス」の書き起こし基準”、同、pp. 13-20 (2001)

[6] 小椋：“話し言葉コーパスの単位認定基準について”、同、pp. 21-28 (2001)

[7] T. Shinozaki, C. Hori and S. Furui: “Towards automatic

transcription of spontaneous presentations", Proc. Eurospeech 2001, Aalborg, pp. 491-494 (2001)

[8] **篠崎、古井：" 話し言葉音声認識における話者間の認識率変動要因の解析 "、信学技報、**SP2001-102 (2001)

[9] T. Kawahara, H. Nanjo and S. Furui: "Automatic transcription of spontaneous lecture speech", Proc. Automatic Speech Recognition and Understanding Workshop, Madonna di Campiglio (2001)

[10] **堀、古井：" 係り受け** SCFG **に基づく音声自動要約法の改善 "、信学技報、**SP2000-127 (2000)

[11] **堀、古井：" 英語ニュース音声を対象とした音声自動要約 "、信学技報、**SP2001-109 (2001)