論文 / 著書情報
Article / Book Information

論題(和文)	「話し言葉工学」プロジェクトの概要と展望
Title	Overview and Perspectives of the Project Spontaneous Speech: Corpus and Processing Technology
著者(和文)	古井貞熙, 前川喜久雄, 井佐原均
Author	SADAOKI FURUI
出典(和文)	話し言葉の科学と工学ワークショップ予稿集, Vol. , No. , pp. 1-6
Journal/Book name	, Vol. , No. , pp. 1-6
発行日 / Issue date	2001, 3

話し言葉の科学と工学ワークショップ講演予稿集
東京工業大学、2001年2月28日～3月1日、pp. 1-6

『話し言葉工学』プロジェクトの概要と展望

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」プロジェクトの概要と、これまでの進展を紹介する。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、前者の二つの目標に関して、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 研究課題も多い。
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

Overview 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 overviews the "Spontaneous Speech Engineering" Project started in 1999 aiming at analysis and recognition of spontaneous speech, and introduces its recent progress. 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. Although steady progress has been achieved for the former two themes, there 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

不特定話者大語彙連続音声認識でも、新聞記事などのあらかじめ用意されたテキストの読み上げ音声なら、かなり高い精度で認識できるようになってきたが、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認識しようとすると、大幅に認識性能が低下してしまうのが現状である[1][2]。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が、その実現には多くの未解決の課題がある。

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[3]。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。

2．プロジェクトの概要

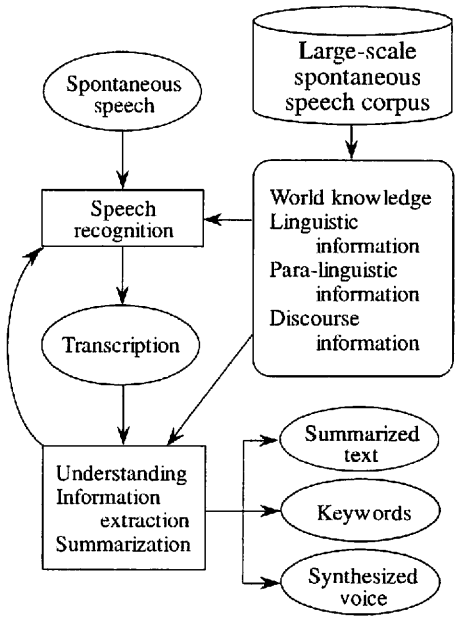

図１　プロジェクトの概要

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

以下にこれらの詳細について述べる。

3．話し言葉コーパスの構築

3.1 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。この手法には大変優れた面があり、特に音声認識に関しては、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。例えば、アナウンサーが発声した放送ニュース音声が 90%以上の単語正解精度で認識できる音声認識システムを、そのまま学会の講演や放送の対談番組に適用すると、30 ないし 40%程度の精度しか得られない。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。

3.2 構築するコーパスの概要

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 700　800 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の作成を目標としている [4][5][6]。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[7]、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、対談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する。音声のソースとしては、国内の種々の学会での講演、大学での講義、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）などから始め、徐々に対象を広げていく予定である。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、書き起こし、セグメンテーション、読みの付与、形態素解析などを行うが、全体の約 10%（50 万語）をコアとし、韻律情報などパラ言語情報まで含めたコーパスとする予定である。話し言葉に関しては、その書き起こし（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。間投詞、言い淀み、雑音などをどう表記すべきかという問題もある。話し言葉音声自身の曖昧性のために、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。この基準作りも、このプロジェクトの重要な役割の一つと考えており、すでに各種マニュアル（6 種類、計約 200 ページ）を作成し、辞書の整備を進めている。作成したコーパスは、できるだけ早く段階を追って公開する予定である。

4. 話し言葉の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。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。そのため、コーパスから種々の世界知識、言語的／パラ言語的情報を抽出・体系化し、これらを有効に組み合わせて用いる方法についても検討する。

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、研究を進めていく予定である。

5. 話し言葉の音声理解と音声要約

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する。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。

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 [8]。一方、各文ごとの要約ではなく、複数の文からなる記事から重要な文を抽出することによって要約する手法が、自然言語処理の分野で長く研究されている。しかしこれだけでは、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、という結果も報告されている。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、書き言葉の要約とは本質的に異なる。世界知識や文の意味的構造まで考慮した、全く新しい方法を構築する必要があろう。

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。

本プロジェクトが目標とする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
- 音声ワープロの高性能化への寄与

6. 本プロジェクトにおける音声認識・要約の研究状況

6.1 話し言葉の音声認識

日本音響学会と日本音声学会で収録された男性 4 名の音声に関する講演（比較的講演に慣れており、原稿を用いずに自然に話している）をテストセットとして、音声認識実験を行った[9][10]。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) 発声速度に個人差が大きい。
(2) フィラーや言い直しの回数も個人差が大きい。
(3) 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。適当なポーズで区切るか、文単位に区切らずに連続してデコーディングすることが必要。
(4) 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有効性が高い。
(5) 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。
(6) 話し言葉に整合した発音辞書を整備する必要がある。
(7) まだ単語正解精度は 55～70%であり、評価法を含めて、研究課題が多い。
(8) 発声速度が速く、フィラーや言い直しが多い話者の認識率が比較的低い。

6.2 音声要約

音声認識結果を用いて、上で述べたように、各文ごとに要約する手法について検討を進めた[8]。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）、および認識結果としての文中の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 SCFG (Stochastic 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ニュース音声に適用し、評価を行ってきた。要約結果は、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した。その結果、上で述べた種々のスコアが有効であることが確認されている。

現在この要約手法を、各文ではなく、複数の文からなる記事（発話）単位に適用する検討と、講演音声への適用を進めている。

7. むすび

「話し言葉工学」プロジェクトと概要と現状を紹介した。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。

現在我が国で複数の音声認識関連大型プロジェクトが展開しており[7]、それらは相互に関連している。本プロジェクトを進めて行くにあたっても、これら関連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、田中穂積先生の GSK 構想 [11] を軸として、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ろう。

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。

本プロジェクトでは、この分野の国内外の 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 12 年 2 月 28 日、29 日に各評価委員の講演とプロジェクトの紹介を中心とした国際シンポジウムを開催するとともに、3 月 1 日に評価委員会を行った。これらを通じて、本プロジェクト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。平成 13 年度には、2 回目の評価委員会を行う予定である。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の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。客員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る大学や研究機関の方々に感謝する。

参考文献

[1] B.-H. Juang: "Progress & challenges in automatic recognition and understanding of spoken language", Proc. International Symposium, Toward the Realization of Spontaneous Speech Engineering, Tokyo, pp. 9-13 (2000)

[2] 河原達也：“話し言葉音声認識の概観”、信学技報、SP2000-95 (2000)

[3] S. Furui: "Steps towards natural human-machine communication in the 21st century", Proc. International Symposium, Toward the Realization of Spontaneous Speech Engineering, Tokyo, pp. 1-8 (2000)

[4] K. Maekawa and H. Koiso: "Design of spontaneous speech corpus for Japanese", Proc. International Symposium, Toward the Realization of Spontaneous Speech Engineering, Tokyo, pp. 70-77 (2000)

[5] 前川、小磯、籠宮、古井、井佐原：“共通日本語話し言葉コーパスの設計”、音学春季講論、1-Q-31 (2000)

[6] 小磯、籠宮、菊池、前川、土屋、間淵、斉藤：“「日本語話し言葉コーパス」の書き起こし基準について”、信学技報、SP2000-104 (2000)

[7] 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)

[8] 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)

[9] 篠崎、齋藤、堀、古井：“話し言葉音声の認識を目指して”、信学技報、SP2000-96 (2000)

[10] 加藤、南條、河原："講演音声認識のための音響・言語モデルの検討"、信学技報、SP2000-97 (2000)
[11] 言語資源共有機構（GSK）活動報告、自然言語処理システムに関する調査報告書、5章、pp. 249-292 (2000)および http://tanaka-www.cs.titech.ac.jp/gsk/