T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	一里塚としての『日本語話し言葉コーパス』
Title(English)	
著者(和文)	古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会2006年春季講演論文集, Vol. , No. , pp. 1191-1194
Citation(English)	, Vol. , No. , pp. 1191-1194
発行日 / Pub. date	2006, 3

Powered by T2R2 (Science Tokyo Research Repository)

一里塚としての『日本語話し言葉コーパス』*

○古井　貞熙（東工大）

1　はじめに

科学技術振興調整費の開放的融合研究推進制度によるプロジェクトに、音声をテーマとする課題で申請する内容の相談を行ったのは、1998 年秋にオーストラリアのシドニーのダーリングハーバーで開かれた ICSLP98 の会期中であった。それまでの音声関係の多くのプロジェクトは、コンピュータシステムとの音声対話を研究のターゲットとしていた。しかし我々は、将来的に、人と人とのコミュニケーションにおいて、音声で発せられたメッセージ（ドキュメント）のトランスクリプション（書き起こし）をベースとする検索、すなわち SDR (Spoken Document Retrieval)が重要になると考えた。しかしそのためには、自然な話し言葉による音声、例えば講演、講義などのモノローグや、インタビューなどが十分な精度で自動的に書き起こしできることが必要である。それまでの音声認識研究は、原稿を読み上げた音声を対象とするものがほとんどで、自然な話し言葉になると、急激に認識精度が低下する問題があった。原稿を読み上げた音声と、自然な話し言葉とは、音響的にも言語的にも、大きく異なるためである。

話し言葉の音声認識を可能とするためには、話し言葉の大規模なデータベース（コーパス）が必要で、話し言葉の大規模コーパスは、1998 年当時は世界中のどこにも存在していなかった。それを構築するには、多数の人の協力と多額の資金が必要で、国家プロジェクトとして進めるべきものと考え、提案書を作成した。幸い、1999 年から国立国語研究所、通信総合研究所、および東京工業大学を中心とするグループによるプロジェクトが始まり、5 年間で約 10 億円の資金を用いて、「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」事業が推進された[1][2]。上記の他、京都大学など、種々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、極めて密度の高い研究開発が行われた。以下に、これまで何ができたか、何を学んだか、今後何が必要かについて述べる。

2　音声ドキュメントの書き起こしと検索の研究動向

音声ドキュメントの検索(SDR)に関しては、近年、米国を中心に盛んに研究が行われている[3][4]。DARPA の EARS (Effective, Affordable, Reusable Speech-to-Text) プロジェクトは、rich transcription と称して、多様なメタデータを音声ドキュメントに自動的に付与し、検索に用いることを目的に始められた。残念ながら短期間で打ち切られてしまったが、このプロジェクトは、世界的に研究を活発化させる力になった。2005 年の末にプエルトリコで開かれた IEEE の ASRU ワークショップでも、SDR に関する特別セッションが開かれ、J. Hansen (Texas 大)[5]と G. Bellegarda (Apple Computer)[6]のレビュー講演に続いて、活発な討論が行われた。Google、Yahoo!、Microsoft などの、これまでテキストの検索を対象としていたところでも、音声認識技術を用いて、音声ドキュメントを検索対象とする研究が、活発に行われている。音声データの探索は、映像データの検索より容易であると考えられている。

現在、この分野で活発に研究を行っている機関には、BBN、Cambridge 大、ICSI、SRI、LIMSI、IBM、Karlsruhe 大、CMU、MIT、HP、Washington 大、Texas 大、国立台湾大、Chinese University of Hong Kong、京大、豊橋技科大、龍谷大、NHK 技研、NTT、東工大などがある。音声ドキュメントの対象には、放送ニュース、ボイスメール、音声アーカイブ（歴史的音声、MALACH プロジェクトなど）、会議、講演、講義、国会の討論、裁判所の記録など、多様なものがある。マルチメディアアーカイブ（コンテンツ）の一部として音声を含む場合もある。

SDR の基本的構成を、図 1 に示す[6]。基本要

* "Corpus of Spontaneous Japanese" as a Milestone, by FURUI, Sadaoki (Tokyo Institute of Technology).

素技術には、書き起こし（transcription）、固有名詞抽出、話者セグメンテーション（diarization）、音響的均一区間検出、話題検出・抽出・トラッキング、インデクシング、タイトル生成、要約、翻訳、検索、質問応答（QA）システムなどがある。検索技術、要約技術の評価法なども研究されている。

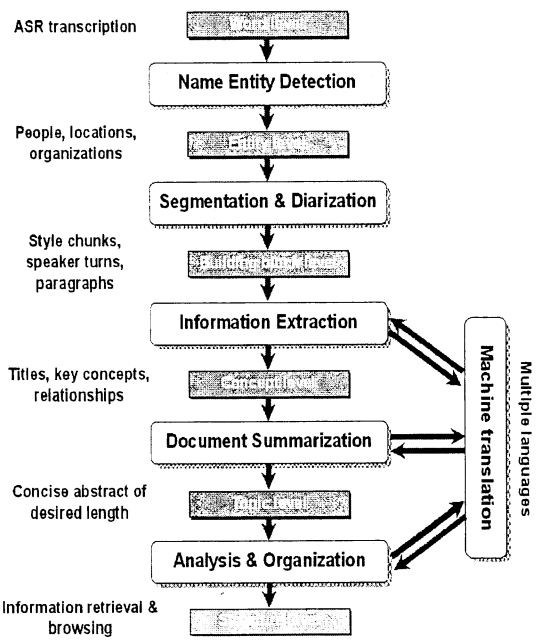

図 1 SDR(音声ドキュメントの書き起こしと検索)の基本的構成

3 『日本語話し言葉コーパス』

上記の「話し言葉工学」プロジェクトにより、タスクを限定しない独話、特に講演（学会講演と模擬講演）を主たる対象として、のべ約 650 時間、単語（形態素）にして約 700 万語規模の『話し言葉コーパス（CSJ; Corpus of Spontaneous Japanese）』を作成した[7]。独話、特に講演を対象としたのは、準備された講演音声は、音響的および言語的に、読み上げ音声と対話音声の中間に位置していると考えられ、話し言葉の特性を有しながら、対話音声よりもモデル化しやすいと考えられるためと、講演音声をコンテンツ化したり、自動的に字幕をつけたいというニーズは、極めて大きいと考えられるためである。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、最低限必要と思われるデータ量に基づいて決めた。研究の幅を広げるため、コーパスの一部に、インタビューなども含めた。

コーパス全体について、セグメンテーション、書き起こし（正書法による基本形と発音形）、形態素解析などを行ったが、全体の約 8%（50 万語）の「コア」については、人手による形態素解析結果、構文構造、要約の他、韻律情報などパラ言語情報まで含めたコーパスとした。コアの形態素解析結果を用いて、コーパス全体を自動的に形態素解析するためのツール（解析プログラム）の構築も行った。

4 『日本語話し言葉コーパス』で何ができたか

4.1 話し言葉の音響的・言語的特徴の分析

(1) 音響的特徴の分析

CSJ には、同じ話者が学会講演（AP）、模擬講演（EP）、対話音声（学会講演の内容に関するインタビューと自由対話）（D）、および読み上げ音声（自分の学会講演の書き起こしや、対話形式エッセーを読み上げた音声）（R）がデータベース化されているため、この中の男性・女性話者各 5 名による音声を用いて、話し言葉と読み上げ音声の音響的特徴の違いに関する分析を行った[8][9][10]。異なる発声スタイルで、話者が共通しているため、話者によるスペクトルの違いの影響が除去できる。

話し言葉音声においては、読み上げ音声に比べて、スペクトル空間が縮小する傾向があることが確認されたため、音素ごとにケプストラムの分布の縮小率を求めた。その結果、ほとんどすべての音素においてその傾向が見られ、対話音声において特に顕著に見られることが確認された。図 2 に、発話スタイルごとに平均した、ケプストラムの分布の縮小率を示す。

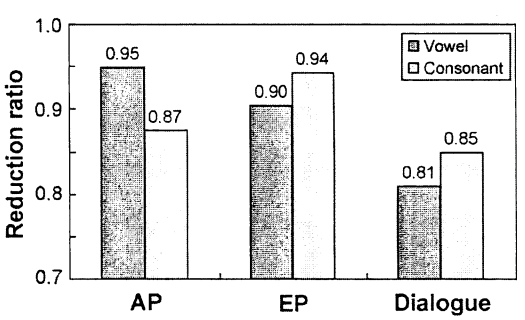

図2 ケプストラムの分布の縮小率の平均

さらに、母音・子音ともに、読み上げ音声よりも話し言葉の方が、発話速度が大きくなる傾向が確認された。

次に、音素間のマハラノビス距離と音素認識正解精度の関係を調べた。音素モデルは、CSJの学会講演と模擬講演の男女計500名の音声データを用いて作成し、各発話スタイルの音声を認識した。結果は図3に示す通りで、音素間のマハラノビス距離と音素正解精度の間には、0.97の高い相関があることがわかった。このことは、音素間のマハラノビス距離を調べれば、音素正解精度が予測できることを示している。

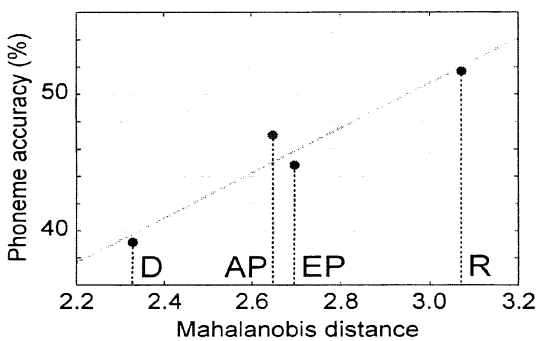

図3 音素間のマハラノビス距離と音素正解精度の関係

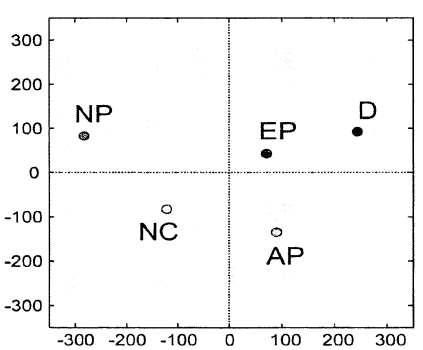

図4 言語モデル間の距離関係

(2) 言語的特徴の分析

CSJ中の種々の発話スタイルのコーパスに加えて、書き言葉およびそれに近いタスクとして、毎日新聞記事（NP）および放送のニュース解説（NC）のコーパスを用いて、それらの言語モデルの比較を行った[11]。各コーパスについて、トライグラムを作成し、コーパス相互間のテストセットパープレキシティと未知語率を調べた。コーパスによって語彙数や未知語率が異なるため、厳密な比較はできないが、テストセットと同じコーパスから言語モデルを作成した場合でも、書き言葉である新聞記事に比べて、講演、対話などの話し言葉では、パープレキシティが約5倍に大きくなることがわかった。各コーパス間のパープレキシティ行列を、距離行列化し、多次元尺度構成法により、言語モデル間の距離の可視化を行った結果を、図4に示す。音素モデルを共通にし、これらの言語モデルを用いて各タスクの音声認識実験を行った結果、パープレキシティと単語正解精度の間に、-0.98の高い相関があることが確認されている。

4.2 話し言葉の音声認識

CSJ中の講演音声をテストセットとして、すでに多数の研究者によって音声認識実験が行われ、「話し言葉の科学と工学ワークショップ」を含む、多数の学会や学会誌で成果が発表されている（例えば[12][13][14][15]）。音響モデルや言語モデルの構成や学習法、特徴パラメータ、適応化法、デコーダ、文区切り、発音変形、誤認識の要因など、多くの課題について、種々の新しい知見が得られている。CSJを学習に用い、話者適応などを行うことにより、講演音声に対して、ほぼ80%の単語正解精度が得られている。

4.3 話し言葉の音声要約

音声認識結果を用いて、その内容をキーワード、要約文、合成音声などで提示する、自動要約技術についても研究が進められている[15][16]。音声認識結果にはある程度の認識誤りや不確実性が避けられず、話し言葉の言語的構造は、書き言葉とは大きく異なるので、音声要約には、書き言葉の要約とは異なる技術が必要となる。基本的には、音声認識結果を用いて、自動的に文の区切りを推定し、重要文抽出と文圧縮の組み合わせによって要約を行う方法がとられている[8][9]。各単語の重要度スコア（情報量）、言語スコア（要約文の単語トライグラムによる言語尤度）、信頼度スコア（事後確率）、および単語間遷移スコア（係り受け構造に基づく単語間の接続可能性）が、単語や文の選択に用いられている。与えられた要約率（圧縮率）に対して、重要文抽出によって圧縮する割合と、文圧縮によって圧縮する割合を決め、それに従ってスコアを最大化する単語セットを、動的計画法を用いて選択し、要約を作成する。この要約手法は日本語および英語の、ニュース音声と講演音声に適用され、10～70%に要約する条件で、評価が行われている。多数の被験者が単語抽出によって作成した要約をターゲットとして、自動要約結果の良し悪しを、定量的に評価する方法についても研究が行われている。

5 『日本語話し言葉コーパス』は一里塚

話し言葉の言語モデルのパープレキシティや未知語率は、書き言葉の言語モデルに比べて、顕著に大きくなる。話し言葉には、言い直し、繰り返しなどが多く含まれ、それらを含めて、言語的変動が極めて大きいためである。それらをカバーするためには、大量の言語モデル構築用コーパスが必要になるが、それは極めて困難であるため、現存する話し言葉のコーパスサイズは、書き言葉コーパスに比べて極めて小さい。容易にテキストが入手できる書き言葉に比べて、話し言葉のコーパスを作成するには、実際の音声を録音して、人手で書き起こしやセグメンテーション、アノテーションなどを行わなければならない。話し言葉言語モデル学習用のコーパスを如何に効率的に収集し、活用できるようにするかが、大きな課題である。話し言葉の広がりがわかれば、それに応じて効率的にコーパスを集めることができるが、広がりがわかるためにはコーパスが必要というジレンマがある。『日本語話し言葉コーパス』は、この課題に挑戦する長い道のりの一里塚に過ぎない。

コンピュータの処理能力の向上、蓄積容量の増大により、音声ドキュメントを蓄え、処理することは、以前に比べると極めて容易になったが、話し言葉の認識処理技術は極めて不十分である。今後、認識精度を低下させている種々の要因について、統計的かつ系統的に解析し、問題点を明らかにする研究を深めるとともに、新たな特徴量やモデルについて研究する必要がある。イントネーション、アクセントなどの韻律情報の利用法、形態素解析ツール、間投詞・言い直しなどのモデル化、音声要約のための構文および意味表現法、未知語への対処法なども重要な研究課題である。

SDR を実現するためには、音声認識技術だけでなく、2 節で述べた多様な技術を高度化し、それらを適切に組み合わせることが重要である。音声から得られる知識を、Web などの多様な知識と関連付け、体系化することが、21 世紀の重要な研究課題となると予想される[17]。そのような体系化された知識が利用できるかどうかが、SDR のみならず、音声認識・理解技術自体の発展のかぎとなるであろう。音声のみでなくマルチモーダル環境でのコンテンツ作成・検索技術も、今後の重要な研究課題である。

参考文献

[1] 古井, “話し言葉の音声認識・理解を目指して,” 電子情報通信学会技術報告, SP2000-47, 2000

[2] S. Furui, “Recent progress in corpus-based spontaneous speech recognition,” IEICE Trans. Inf. & Syst., E88-D, 1, pp. 1-11, 2005

[3] L-S. Lee et al., “Spoken document understanding and organization,” IEEE Signal Processing Magazine, pp. 42-60, Sep. 2005

[4] K. Koumpis et al., “Content-based access to spoken audio,” IEEE Signal Processing Magazine, pp. 61-69, Sep. 2005

[5] J. Hansen, “Spoken document retrieval: Acoustic variability over the past 100 years,” Proc. IEEE Workshop on ASRU, 2005

[6] J. Bellegarda, “A catwalk look at text processing,” Proc. IEEE Workshop on ASRU, 2005

[7] 前川, “『日本語話し言葉コーパス』公開版の仕様,” 第３回話し言葉の科学と工学ワークショップ講演予稿集, pp. 7-14, 2004

[8] 中村他, “日本語話し言葉コーパスを用いた話し言葉音声の音響的特徴の分析,” 情報処理学会研究報告, 2004-SLP-53, pp. 7-12, 2004

[9] M. Nakamura et al, “Analysis of spectral space reduction in spontaneous speech and its effects on speech recognition performances,” Proc. Interspeech 2005, pp. 3381-3384, 2005

[10] S. Furui, “Why is the recognition of spontaneous speech so hard?” Proc. 8th Int. Conf. Text, Speech and Dialogue (TSD 2005), pp. 9-22, 2005

[11] 中村他, “読み上げ音声に対する話し言葉音声の言語的特徴の分析,” 日本音響学会秋季研究発表会講演論文集, 3-6-10, 2005

[12] 篠崎他, “日本語話し言葉コーパスを用いた講演音声認識,” 情報処理学会論文誌, 43, 7, pp. 2098-2107, 2002

[13] 古井, “話し言葉の音声認識と自動要約,” 第３回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, 2004

[14] 南條他, “大規模な日本語話し言葉データベースを用いた講演音声認識,” 電子情報通信学会論文誌 D-II, J86-D-II, 4, pp. 450-459, 2003

[15] 中村他, “音声認識システム SOLON の日本語話し言葉コーパス(公開版 Ver.1.0)による評価,” 情報処理学会研究報告, 2005-SLP-59(17), pp. 97-102, 2005

[16] 堀他, “単語抽出による音声要約文生成法とその評価,” 電子情報通信学会論文誌 D-II, J85-D-II, 2, pp. 200-209, 2002

[17] C. Hori et al., “Automatic summarization of English broadcast news speech,” Proc. Human Language Technology 2002 Workshop, San Diego, pp. 228-233, 2002

[18] 古井, “21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」,” 電子情報通信学会技術報告, SP2003-162, 2003