論文 / 著書情報
Article / Book Information

論題(和文)	音声ドキュメントの文字変換と情報抽出
Title(English)	
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	映像情報メディア学会誌, Vol. 55, No. 11, pp. 1394-1396
Citation(English)	, Vol. 55, No. 11, pp. 1394-1396
発行日 / Pub. date	2001,
権利情報 / Copyright	本著作物の著作権は映像情報メディア学会に帰属します。 Copyright (c) 2001 Institute of Image Information and Television Engineers.

小特集 人にやさしいパターン認識

4. 人にやさしい音声・文字認識技術

4-1 音声ドキュメントの文字変換と情報抽出

古井貞熙†

キーワード：音声ドキュメント，トランスクリプション，音声文字変換，大語彙連続音声認識，情報抽出

1．まえがき

統計的パターン認識理論，ケプストラムとデルタケプストラム，隠れマルコフモデル，統計的言語モデルなどを基本とする大語彙連続音声認識の技術が近年大きく進展し，音声ワープロ，放送への自動字幕付与，音声対話システム，情報検索への適用など，種々の応用が展開しつつある．音声認識技術の主たる応用は，比較的長い音声ドキュメントを対象として，音声から文字への変換を目的とするシステム（トランスクリプション，音声文字変換）と，音声によるコンピュータとの対話システムに分類できる．後者に関しては，本特集で別に解説されているので，本稿では，前者に焦点を当てる．音声文字変換の技術は，音声対話システムの技術の基礎としても重要である．

音声は，人と人との最も自然で，基本的，容易かつ効率的な情報交換手段である．多数の人を対象とした講演，講義，解説などでは，これからも音声が基本に用いられていくであろう．コンピュータへの入力手段としても，音声が使えれば極めて効率的になると期待されている．特に日本語に関しては，使いやすい優れたワープロ（ソフト）が普及しているとはいえ，仮名・漢字変換を必要とするため，英語のようにしゃべる速さに近い速度でタイプ入力するのは極めて難しい．音声認識技術を用いた日本語の音声文字変換ができれば，音声ワープロ，放送への自動字幕付与の他，会議の議事録，講演録，講義録，放送ニュースのアーカイブ化への利用など，数多くの応用が考えられる．本解説では，これまでの研究の流れと，今後の展望について述べる．音声認識技術の基本に関しては，文献1)～3)などを参照していただきたい．

2．音声文字変換研究のこれまで

2.1 読み上げ音声を対象とした研究

近年の音声文字変換研究の牽引車は，Wall Street Journalを始めとする北米の種々の新聞を読み上げた音声を対象として行われたDARPA（Defense Advanced Research Projects Agency）のプロジェクトである．このプロジェクトは1991年に開始され，文脈依存音素HMMと，単語を単位とする統計的言語モデルを基本とする，数万語規模の大語彙連続音声認識の研究が活発に行われた．メルケプストラム，状態共有HMM，言語モデルのバックオフ平滑化など，現在の音声文字変換の基本技術の多くは，このプロジェクトによって確立したと言える．それまでの人手に頼ることを前提にしていた文法を用いるよりも，認識性能が大きく向上し，65,000語の語彙を対象とした場合に，93%の単語正解精度が報告されている．

我が国においては，単語ではなく，音素，音節，仮名・漢字などを単位とする統計的言語モデルを用いた研究が，まず行われた．筆者らがNTTの研究室で，単語（形態素）を単位とする日本語の音声文字変換の研究に着手したのは，1995年の3月である．NTTの研究所で優れた形態素解析プログラムが開発されると同時に，その開発に使われた日経新聞の5年間にわたるテキストが，電子的に使えるようになったことが，研究開始のきっかけとなった．直ちに，研究用データベースの構築にとりかかり，これを用いた音声認識実験を始めた．言語モデルとして単語トライグラムを用い，英語の場合にほぼ匹敵する認識結果を得ることができた[4]．

その後我が国では，新聞の読み上げ音声を対象とした音声文字変換研究に関して，国内のボランティアグループによるデータベースとデコーダの開発が進められ，「日本語ディクテーション基本ソフトウェア」の形で成果がとりまとめられている[5]．これは，認識エンジン（Julius），日本語音響モデル，言語モデル，および形態素解析・読み付与ツールから構成され，優れた性能を示している．無償で一般に公開されており，大語彙連続音声認識研究及び開発の共通のプラットフォームとして広く使われている．

2.2 放送音声の音声文字変換

1996年早春のDARPAワークショップでは，新聞記事の読み上げ音声から，放送音声の文字変換に焦点が移った．新聞読み上げ音声との基本的違いは，対象とする音声が「実世界の音声データ」であることである．アナウンサが原稿をもとに静かな環境で発声した音声だけでなく，音楽・雑音・音声が重畳・混在している音も対象としている．米国を中心に，欧州を含む大学および研究機関がプロジェクトに参加し，語彙数は65,000語が標準的で，言語モデルの作成には，CMUで開発されたツールが広く使われている．

†東京工業大学大学院 情報理工学研究科 計算工学専攻
"Transcription and Information Extraction from Voice Documents" by Sadaoki Furui (Department of Computer Science, Tokyo Institute of Technology, Tokyo)

映像情報メディア学会誌 Vol. 55, No. 11, pp. 1394～1396 (2001)

このDARPAの動きに刺激され，日本でもNHK放送技術研究所を中心として，複数の大学を含むグループで，共同で放送ニュースの音声文字変換研究が始められた．過去5年間以上の放送ニュース原稿データベースと，実際の放送音声が用いられている．統計的言語モデルとしては，ニュース原稿に含まれる比較的頻度の高い20,000種類の単語のトライグラムを用いている．日本語では同じ単語に複数の読みがあり，正しい読みは前後関係（文脈）から決定されるため，同じ表記でも読みの違う単語は別の単位としている．さらに，話し言葉には必ず挿入される「え」「えー」などの間投詞が，文頭や読点の位置に生じやすい性質に着目して，言語テキストのこれらの位置に間投詞の発音を確率的に与えている．筆者らが先駆けて行った実験で，スタジオでアナウンサが原稿を見ながら発声した音声であれば，90%近い認識率を得ることができた[5]．

NHKでは，音声認識を用いてニュース音声に字幕を付与するシステムの運用を，2000年3月から開始した[7]．音声認識では，誤認識を完全に回避することはできないため，認識結果を人が即座にチェックして，誤認識をすばやく修正してから放送するシステムになっている．修正がすばやく実行できるレベルとして，95%以上の単語正解精度が必要で，アナウンサが原稿を読んでいる音声に関してはこれが達成されているが，それ以外のインタビューなどへの音声認識の利用は今後の課題である．ニュースでは日々新しい語彙（未知語）が生ずるため，これに対処する方法が開発されている．文末にデコーダの処理が到達するまで認識結果が確定しない従来の方法では時間遅れが大きくなってしまうので，文中でも常に一定の遅れ以内で認識結果を確定しながら処理を進めるデコーダも開発されている．

最近では，DARPAのプロジェクトを中心に，音声認識と情報検索の技術を組合せる研究が，活発に行われている．放送ニュースの音声を音声文字変換した結果から，その話題(トピックス)を表す重要語やフレーズを自動的に検出し，それを用いてインデクシングを行っている．これにより，大量のニュースを話題に応じて分類するなど，自動的にアーカイブ（データベース）化することができる．放送やビデオの中から，音声の部分だけを取出す研究や，ある特定の話題に関する部分だけを取出す研究も行われている．

2.3 ディクテーションソフト（音声ワープロ）

音声文字変換ソフト（音声ワープロ）に関しては，1990年に米国のDragon Systemsが単語区切りで発声した音声を入力とするPC用プログラムを発売し，その後，IBM，L&H，Philipsなどの会社が参入した．単語区切りでないと入力できないのはユーザへの負担が大きいので，1996年頃から単語を続けて発声できる製品が開発されるようになった．これらの製品の多くは，英語版だけでなく，ドイツ語，フランス語，イタリア語，スペイン語，中国語，日本語版などに発展している．主な利用者は，障害者，X線医師（医療所見の入力），弁護士などである．日本では，日本IBMが米国のIBMと協力して，日本語音声のディクテーションソフトの研究開発を進め，今日の製品へ結実している[8]．やや遅れてNECでも同様の製品が開発されている．

3．話し言葉の音声文字変換の実現へ

3.1 話し言葉認識の難しさとコーパスの構築

現在の音声認識技術で，スタジオでアナウンサが原稿を読んでいる音声であれば，90%あるいはそれ以上の認識率が得られるが，現場からの中継などの音声や，対談の音声に対しては大幅に認識率が低下する[6]．この原因は，発声が必ずしも明確でないこと，雑音が重畳していること，原稿を読んでいない自発的な話し言葉の音声は，言語モデルの学習に用いたニュース原稿の文とはかなり構造が異なることなどである．話し言葉特有の難しさには，助詞などの言葉の脱落・省略，間投詞などの不要語の付加，倒置，言い間違い，言いよどみ，言い直し，繰返しなどがある．

統計的言語モデルを構築するためには，認識対象に対応した大量のテキストが必須である．書き言葉のテキストは最近では比較的容易に得ることができるが，話し言葉に対しては，実際の音声を人手で書き起こすことが必要である．米国では，DARPAのプロジェクトの中で，“Switchboard”，“Call Home”などと呼ばれる，自然な会話音声を書き起こしたテキストデータベースが作られ，これを用いた音声認識研究が行われている．我が国でも1999年から，科学技術振興調整費開放的融合研究推進制度による「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」プロジェクトが開始され，5年計画で，音声認識の高度化のための，700万形態素規模の音声言語データベース（コーパス）の構築，解析，認識，要約などの研究が推進されている[9]．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．

3.2 話題語抽出と音声要約

筆者らは，ニュース音声を自動的に音声文字変換し，その結果としての単語列から話題語（重要語）を自動抽出する試みを行っている[6]．この際，ニュースの話題語はほとんどが名詞であるので，文字変換結果から名詞のみを抽出し，その中から，重要度の尺度によって話題語を抽出する方法を検討した．重要度の尺度は，大量の学習用ニュース記事に出現する各名詞の出現頻度に基づく情報量によって定義した．情報量の大きい単語を上位から順に抽出し，重要語が精度よく抽出できることを確認している．

筆者らはさらに，音声中から話題を担っている話題語をできるだけ文法的・意味的制約に従うように最適に選んで要約文を作成する方法として，各単語の重要度スコア，音声認識結果の信頼性，要約文の統計的言語スコア，単語の意味的関係をもとに，動的計画法を用いる方法を提案した[10]．統計的言語スコアは，音声よりも比較的簡潔な文体となっていると考えられる新聞（毎日新聞）の1年分の記事から計

算した単語トライグラムを用いた．認識結果の信頼性は，事後確率によって求め，意味的関係は，係り受け関係によって求めた．放送ニュース音声や学会講演の音声文字変換結果を用いて，文字数にして20～80%に要約する実験を行っている[10]．本方法は，ニュースの自動字幕作成への適用のほか，講演録・議事録作成など，種々の応用分野への展開が期待できる．講演などを，音声認識技術を用いて自動的に要約してデータベース化することができれば，情報検索に限らず種々の用途に極めて有用であろう．講演の要旨を短時間で聞くことができる，「早聞きシステム」の実現も不可能ではない．

3.3 話し言葉の音響的変動への対処

音声認識で誤認識を生ずるのは，音響モデルや言語モデルの学習に用いたデータと，実際の認識すべき音声の間に，何らかのずれがあるためである．音響的変動の原因には，音声（声質や話し方）の個人差（方言や年令によるものも含む）が極めて大きいこと，同じ人でも体調や精神状態（ストレス）などによって発声速度や話し方が変化すること，多くの場合に，部屋の反響を含む種々の雑音が音声に加わっており，しかもそれが時間的に変化すること，更にその上にマイクロホンや電話機，伝送系などの歪みが加わることなどがある[2]．

これらの変動に対処するため，音声認識システムに，少量の音声サンプルに基づいて，変動に自動的に追随する適応機能を持たせる研究が広く行われている．一般に適応化アルゴリズムには，その規則を学習するためにあらかじめ決められた言葉を発声してもらう「教師つき（Supervised）学習」と，任意の音声を用いて規則を学習する「教師なし（Unsupervised）学習」がある．前者の方が規則の獲得は比較的容易であるが，ユーザにとっては，認識装置を使う前にわざわざ特定の言葉を発声しなければならないのは煩わしい．後者の方が，認識すべき音声をそのまま使って適応化できるので，ユーザにとってはまったく意識せずに使えて便利である．

筆者らはこのような動機から，話者の交代を自動的に検出しながら，音素HMMをオンラインで，教師なしで逐次的に，入力音声に自動的に適応させる方法を提案した．話者適応には，ゆう度最大化線形回帰法（MLLR）に，MAP推定とベクトル空間平滑法（VFS）を組合せて用いた．放送ニュース音声を用いた実験の結果，本方法によって音声認識性能が向上することを確認した[11]．

3.4 話し言葉の言語的変動への対処

音声認識の応用場面が変わるごとに，言語モデルの学習のために大量の（書き起こし）テキストを集めるのは決して容易ではない．このため，共通の基本的な言語モデルを，限られた量のテキストを用いて，異なる対象に適応化させる研究も行われている．言語モデルの適応化のためには，テキストや言語モデルの中で，タスクに依存する部分と，依存しない一般的な部分を分ける必要があり，このような研究も始まっている[12]．

新しいタスクへの適応と関連して重要な課題に，新しい言葉（未知語）をいかにして言語モデルに組込むかという問題がある．最も一般的な方法は，あらかじめ単語を品詞や意味を用いてクラスに分類しておき，未知語をその中の一つのクラスに対応させてクラス言語モデルを用いる方法である．単語クラスの作り方としては，細分類された品詞，ゆう度最大化基準，潜在的意味解析（Latent Semantic Analysis），エントロピー最大化基準，相互情報量，シソーラスなどを用いる方法や，これらを組合せる方法がある．

4．むすび

音声文字変換には，極めて多様な応用が期待できるが，まだ解決しなければならない課題も多い．主たる課題は，話し言葉特有の言語的，音響的変動に対するロバストなモデルと適応アルゴリズムの構築である．その基礎となる話し言葉の大規模コーパスの構築も極めて重要である．本解説では触れなかったが，音声対話システムでも，音声認識部は，音声文字変換と同様に，大量のコーパスから作成した統計的言語モデルを用いる方法が主流になりつつある．音声文字変換との基本的違いは小さくなりつつあり，この意味でも，音声文字変換技術は音声認識の基本と言える．

（2001年5月30日受付）

〔文　献〕

1) 古井貞熙：“話し言葉処理のメカニズム－音声認識から会話意図理解へ－”，映情学誌，54，6，pp. 765-768（2000）
2) 古井貞熙：“音声情報処理”，森北出版（1998）
3) 古井貞熙：“音声トランスクリプションのこれまでと今後の展望”，信学論，J83-D-II，11，pp. 2059-2066（2000）
4) 松岡達雄，他：“新聞記事データベースを用いた大語い連続音声認識”，信学論（D-II），vol. J79-D-II，12，pp. 2125-2131（Dec. 1996）
5) T. Kawahara et al.：“Sharable Software Repository for Japanese Large Vocabulary Continuous Speech Recognition”，Proc. ICSLP，Fr4A1（Nov. 1998）
6) K. Ohtsuki et al.：“Improvements in Japanese Broadcast News Transcription”，Proc. DARPA Broadcast News Workshop，pp. 231-236（Feb. 1999）
7) 今井亨，他：“逐次2パスデコーダを用いたニュース音声認識システム”，信学技報，SP99-129（Dec. 1999）
8) 西村雅史：“日本語ディクテーションシステムの現状と今後の課題”，信学技報，SP99-94（Dec. 1999）
9) 古井貞熙，他：“科学技術振興調整費開放的融合研究推進制度－大規模コーパスに基づく『話し言葉工学』の構築－”，音響学誌，56，11，pp. 752-755（Nov. 2000）
10) 堀智織，他：“係り受けSCFGに基づく音声自動要約法の改善”，信学技報，SP2000-116（Dec. 2000）
11) S. Furui et al.：“On-line Incremental Speaker Adaptation for Broadcast News Transcription”，IEEE Workshop on Automatic Speech Recognition and Understanding”，pp. 165-168（Dec. 1999）
12) 河原達也，他：“会話スタイル依存の言語モデルを用いたキーフレーズの検出・検証”，信学技報，SP97-78（Dec. 1997）

古井　貞熙（ふるい　さだおき）　1968年，東京大学工学部計数工学科卒業．1970年，同大学院修士課程修了．同年，NTT研究所入社．以後，音声認識，話者認識，音声知覚などの研究に従事．1978年～1979年，ベル研究所客員研究員．1997年，東京工業大学大学院情報理工学研究科計算工学専攻教授．工学博士．