解説

音声ワープロ －過去・現在・未来－

西村雅史　　伊東伸泰
日本アイ・ビー・エム（株）東京基礎研究所

ディクテーションプログラム登場

最近，音声認識やディクテーションという言葉を耳にすることが多くなったのではないだろうか．SMAPの香取慎吾がテレビCMで，「先日いただいたベッタラ漬け，あなたのぬくもりがこもっていた....」などと日本語音声ワープロ（過去には入力音声を仮名文字に変換するための音声タイプライタと仮名漢字変換プログラムの組合せをこう呼んでいたが，現在の音声ワープロでは入力音声を直接仮名漢字混じり文に変換することができる．これを特にディクテーションプログラムと呼ぶことにする）を実演してみせたりしたので，背景を理解していない人にはちょっと過度に期待させてしまった面もある☆1．一方で，人工知能関連の技術の多くがそうであったように，音声認識も何年かに一度，皆が忘れた頃に再び話題となる研究テーマの1つに過ぎない，と冷めた目で見ている研究者，技術者の方もまだ多いと思う．

確かに2年前（1996年11月），日本IBMが初めて日本語のディクテーションプログラムを発表したときには，PCに付属の数あるオマケソフトの1つに過ぎなかったし，その上「先取り体験版」という名前まで付けられていた．つまり，我々研究・開発者の熱い思い入れに反し，企業としては，恐る恐る出してみたというのが実状であったようだ．しかし，その後世間で認知されるにつれ，少なくとも社内の風向きは一変したのである．当時のディクテーションプログラムはまだ単語ごとに区切って発声（離散発声）する必要があったが，その後，1997年12月にIBMから，そして1998年7月にはNECからも，自然な連続発声による入力が可能な製品が出荷された[1), 2)]．また，ジャストシステムがIBMと共同で自社のワープロソフトや仮名漢字変換プログラムを音声入力対応にする[3)]など，音声入力が急速に一般的な日本語入力手段になりつつある．ここではこのように最近注目を浴びている日本語ディクテーションプログラム（あるいは音声ワープロ）について，その技術的な背景と現状の応用分野について解説する．また，今後期待される応用分野についても紹介する．

音声ワープロ－昔と今[4)]

音声認識は人間と機械とのコミュニケーションを実現する手段として不可欠なものと考えられているが，その実現は当初予想されたほど簡単なものではなかった．実際，世界で初めて本格的なコンピュータが完成してからわずか数年後の1952年には音声認識の最初のプロトタイプが作成されたほどで，すでに50年近い研究の歴史があるが，残念ながら機械と自然で自由な対話[5)]を実現するような認識装置はいまだに完成していない．

しかし，音声認識の研究がずっと停滞していたのかというとそうではない．15年以上前なら，たとえ専用のハードウェアを用いても実時間で認識できるのはせいぜい数十単語，そして7，8年前でも1000単語程度の認識が限界だったと思われるが，今やMMX Pentium-200MHz程度のCPUを搭載した普通のパソコンで数万単語といった大語彙の連続音声認識が可能となり，音声認識に関係した者の長年の夢であった本格的な音声ワープロが実用化されるに至ったのである．

1980年代初頭に開発された日本語音声ワープロは，キーボードの代替手段としての音節認識装置と，仮名漢字変換プログラムを組み合わせた単純なもので，あらかじめ声を登録した話者（特定話者）が音節ごとに区切って発声した110個程度の音節だけを認識対象とすることで処理量を軽減していた．しかし，言語処理はすべて認識の後処理となっていたことから音声認識時には言語情報による候補の絞り込みがまったく行われず，発話者に単音節発声という負担を強いる割に認識精度は大変低いものであった．

一方，現在のディクテーションプログラムでは，統計的な音響モデルの採用によって音韻の識別能力および発話の揺らぎに対するロバストネス（頑健さ）が向上したのに加え，音響上の識別を行う際，同時に単語の前後関係に関する情報に基づいて対象単語候補を絞り込むことで高い認識精度を実現したのである．さらに，現在のPC用のプログラムでは音響的な特徴のより緻密な表現に加えてPC自体の処理能力の向上にも助けられ，連続発声の入力が可能となっている．ただし，認識対象はあくまでも書き言葉であり，読み上げかそれに近い丁寧な発声で，文法的にも正しく，言いよどみや言い誤りはほとんどないことが必要である．この理由については次節で詳しく説明する．なお，6万語を認識対象とした不特定話者の日本語ディクテーションで，96.6%の単語正解精度（文字正解

☆1 現に「ベッタラ漬け」という単語は当時の製品の辞書には入っていなかった:-p.

精度97.8%）が報告されている[6]．

統計的な音声認識手法

次にディクテーションを実現するための基本技術である，統計的な音声認識手法について詳しく説明しよう（図-1）．

ディクテーションプログラムの仕組みというと，まず音声を認識して仮名文字列に変換し，その仮名文字列（あるいは仮名文字ラティス）を辞書や文法を参照して仮名漢字混じり文に変換していると誤解している方も結構多いようである．しかし，統計的手法に基づく音声認識装置ではそのような適用順序が存在するわけではない．ここでは確率という尺度を導入することで，音響的にもそして言語的にも，両方の意味で最適な仮名漢字混じりの文字列を直接推定するのである．

現在のディクテーションプログラムで行っている音声認識は図-1に示すような情報・通信理論の問題として解釈することができる[7),8)]．人間が発声した単語列（W）が音声信号として空気中を伝わり（S），マイクロフォンで電気信号に変換される．さらにA/D変換器により量子化された後，音響処理部で周波数分析され，音声の特徴量列（X）に変換される．この部分までが音響チャネルと呼ばれ，音響チャネルは雑音のある通信路となっている．この特徴量列を言語復号部で復号化し，人間の伝えようとした単語列を復元すると考えるのである．

音響処理部の出力である特徴量列（X）が与えられたときに，単語列（W）が生じる確率$\Pr(W|X)$を最大とするようなWを選べば誤認識を最小とすることができることが分かっている．argmaxを$\Pr(W|X)$を最大にするような単語列Wを返す関数とすると，

$$\tilde{W} = \underset{W}{\operatorname{argmax}}\ \Pr(W|X)$$
$$= \underset{W}{\operatorname{argmax}}\ \Pr(X|W)\Pr(W)$$

となり，結局，確率$\Pr(X|W)\Pr(W)$を最大にするような単語列Wを求めればよいことが分かる．ここで，$\Pr(X|W)$は単語列Wが与えられたときに特徴量列Xの出現しやすさを表す確率で，一方，$\Pr(W)$は単語列Wの出現しやすさを表す確率である．

実際の認識装置では，$\Pr(X|W)$は通常HMM（Hidden Markov Model）[8)]を使って求める．このモデルを特に音響モデルと呼ぶ．HMMは統計的なモデルであり，その推定には大量の学習用音声データを必要とするが，音節認識装置などで採用されていたテンプレートマッチング手法に比べて発話の揺らぎに対するロバストネスが高く不特定話者化もしやすいという特徴がある．

一方，$\Pr(W)$は言語モデルを使って求める．言語モデルとしては文法やN-gramモデルが適用できるが，特にディクテーションではN-gramモデルが有効である．このモデルは汎用性が高く，十分な量の学習用コーパスが用意できれば一般的な日本語文章入力用のモデルが構築できる．なお，N-gramモデルとは単語列Wの出現確率を，

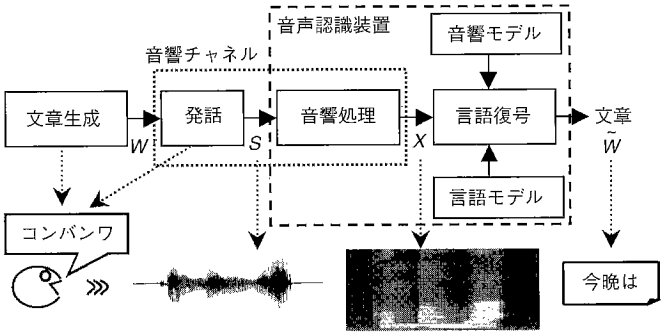

図-1　音声認識の情報・通信理論的解釈

ある単語（列）の後にはどの単語が出現しやすいかという確率の積で表現したものである（図-2）．

このように，統計的手法に基づく認識装置では，音響上の出現確率と言語上の出現確率が同時に評価され，両方の意味で最適な単語列（仮名漢字混じりの文字列）が認識結果として出力されていることに注意されたい．

なお，言語モデルの性能を評価するための尺度としてはエントロピーが用いられる[8)]．情報理論的にはエントロピーがHである情報源からは$P = 2^H$個の単語が等確率に出現すると解釈できる．つまりこの値Pは，言語モデルを適用した結果，「すべての単語が等確率に予測されるようなシステム（または言語モデルを使わないようなシステム）だと認識対象単語数が何単語程度に相当するのか」ということを表している．そしてこの値をパープレキシティ（複雑度）と呼ぶ．テスト文に対しても，その単語の出現確率が先に推定してあるN-gramモデルによって与えられるとすると同様にパープレキシティを計算できるが，これをテストセットパープレキシティと呼んでいる．

新聞記事，雑誌，電子会議室の発言などから集めたテキスト約700万文を学習データとして用い，約6万語の語彙に対してN-gramモデルを推定した場合の，種々のテスト文に対するテストセットパープレキシティを表-1に示す．このように，言語モデルを用いない場合には6万であった認識対象単語数が，1-gramを使用すると実質的には900～1400語程度に，そして3-gramまで使用すれば100～200語の音声認識と同じ程度にまで削減されていることが分かる．この程度の語彙サイズならば十分な認識精度が出ることがお分かりいただけるであろう．

ただ，N-gramモデルが有効に働くのはN-gram推定用の学習データと類似した文章を正確に読み上げた場合であることに注意されたい．現状では入手できるコーパスがどうしても新聞中心になるため，N-gramモデルも新聞記事に偏った性質のものになってしまっている．

現在の応用分野

HAL9000やドラえもんといった，人間と自由な対話ができるシステムの登場までにはまだかなり時間が必要だと思われるが，ディクテーションが実現されたことにより，音声ワープロに限らず，数多くの応用が生まれた．また，今後の改良により適用できる見通しが得られている分野もある．

IBM社製の日本語ディクテーション製品に関するアン

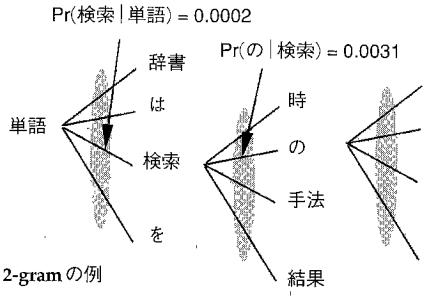

図-2　2-gramの例

データ名	Test-set Perplexity		
	1-gram	2-gram	3-gram
日経新聞	1412.3	169.1	91.2
産経新聞	1229.9	245.5	179.4
電子会議室	1269.2	266.5	201.7
ビジネストーク	903.9	149.3	112.1
小説	1060.6	245.5	190.1

表-1　テストセットパープレキシティ

ケートはがきによると☆2，日記やメール作成など，通常の音声ワープロとしての使用方法のほかに，OCRの代替手段（教師が生徒の書いた文章をディジタル化するなど），日本語学習といったものが散見される．また，身障者のコミュニケーション手段としても期待されており，健常者から聴覚障害者へのコミュニケーションの際に手話・筆談の代替手段としての使用例が報告されている．しかし，身障者用の入力手段としては今後改善すべき点も多い．

このほか，ビジネス上の代表的な応用例としては，1）放射線科の医師によるX-線写真に対する所見入力，2）法医学における検死解剖報告[9]，3）翻訳結果のテキスト入力，4）弁護士の判例入力，5）録音テープの書き起こしなどがある．特に欧米では医療関係のデータ入力への応用例が多い．

1），2）は手が離せないか離したくない場合に音声によって効率的な入力を実現する．3）も通常は頻繁な視線移動を伴う作業であるが，対象とする英文だけに集中できるので作業効率が改善される．また，1），2），4）は，ほぼ定型の大量の文章を入力する場合に音声入力が有効であることを示している．5）に関しては，現状では講演やインタビューなどを直接書き起こすことは難しいので，ユーザがテープを聞きながら，書き言葉に近い形で復唱するといった使い方がされている．それでもテープを操作しながらキーボードを打つといった従来の作業よりははるかに効率がよい．なお，連続発声による音声入力の場合，修正作業を含む場合でも，おおむねワープロ検定1級（90文字/分）以上の入力速度が得られることが分かっている[4]．

また，録音機とディクテーションプログラムを直接連携させている例もある[10]．ここでは，ディジタル録音によって認識に必要な音声品質を維持し，かつ，使用者自身の声による音声メモに限定したことで，簡単な操作による音声メモの書き起こしを実現している．

今後期待される応用分野

放送音声の書き起こし

今後の改良によって応用が期待されている分野としては，まず，ニュース音声の書き起こしがある．特に，米国ではARPA（Advanced Research Projects Agency）を中心にして精力的に研究が進められている[11), 12)]．背景雑音や音楽の重畳，インタビューにおける自由発話，電話回線でのレポートなど，ディクテーションプログラムで想定していたよりはかなり厳しい条件のデータもあるが，雑音のない環境でアナウンサがニュースを読み上げている部分については接話型マイクを用いた通常のディクテーションとそれほど大きな差はない．

すでにCNNの一部のニュース番組などでは実時間で英語の字幕が入るが，これは音声認識ではなく，ステノタイピストと呼ばれる専門家が発話内容をキーボード入力しているものである．しかし，日本語ではニュースを読み上げる速度での文字入力は，専用の速記タイプライタなどを用いても不可能であることが分かっており，また，速記タイプライタを操作できる人も非常に限られている．このため英語よりもむしろ日本語で，音声認識によるニュースへの字幕挿入に期待が集まっている．

IBMがTBSと共同で行ったラジオニュースの書き起こし実験結果を**表-2**に示す[6]．CERは文字誤り率☆3を表し，同音異義語による表記上の誤りなどもこれに含まれる．この実験では主に固有名詞（人名，地名）など，辞書に入っていなかった単語をあらかじめ登録するとともに，アナウンサの声もシステムにあらかじめ学習させてあるが，その程度のことをすれば，現状でもかなり高い精度での音声認識が可能であることが分かる．

固有名詞が頻繁に現れるスポーツや町の話題はあまり得意ではないが，特に，トップニュースになるような，事件，事故，外交問題といった話題は高い精度で認識できているといえよう．天気予報は誤りが多いが，これは言語モデル学習用のテキストに類似の文章がほとんど含まれていなかったためで，例文を学習させれば容易に改善できる．もちろん，実用化のためには次々と出現する固有名詞や新語をどのように効率的に抽出し，追加していくかという問題がある．また，音声認識で誤りを0にすることはできないので，実用化のためには，人間による確認，修正が短時間でできるようなプログラムの開発が今後必要となろう．

また，ニュースの書き起こしだけではなく，資料として残る大量の音声データやビデオテープの情報検索に音声認識を使うことも期待されている．ここでもまずディクテーションを行う．これによってテキストを作成し，この文字データを検索することで対応する部分の音声や画像を検索するというものであるが，75%以上の単語正解精度が出ていれば情報検索としては十分使えるといった報告もある[13]．さらに話者認識と併用すれば，特定の話者が，特定の発言をしていた部分を抜き出すといったことが行えると期待されている．

TV番組のナレーションなどある程度正確な台本が存在

☆2 この記述は日本アイ・ビー・エムの資料にもとづく著者の考察であり，同社の見解を代表・示唆するものではない．
☆3 文字誤り率＝（置換＋挿入＋脱落）/総入力文字数．

する場合に，字幕の挿入のタイミングを自動化する目的で音声認識を行うこともある．純粋に音響的な特徴のアライメントをとる手法だけでなく，台本とは異なる言い回しをしていた場合にも対処するため，ディクテーションを行い，テキスト上で内容の一致する部分を見つけるといった方法も検討されている．

いずれの場合も，背景雑音，発話の仕方，あるいは録音品質によって認識精度が大きく影響を受けるため，現状では課題も多いが，米国ではすでに実用化に向けた試行が始まっている．日本でも，ニュース音声の書き起こしはIBM以外にNHK技研やNTTなど[11]で，そして字幕挿入タイミングの自動化という目的では通信・放送機構等[14]で検討されている．

講演の書き起こし

現状では，講演，議会の質疑応答などの音声はディクテーションに向かない．それは発話の仕方が読み上げよりも自由発声に近くなり，文体も大幅に異なるからである．自由発声では言いよどみや不要語が避けられないし，それに至らないまでも発音が不明瞭であったり，逆に感情によって強調された発音になることも多い．一方，語彙や言語モデルの観点から見ても話し言葉に対応するために必要となる学習コーパスはきわめて少ない．また，背景雑音などの問題もある．

しかし，一般的な対話音声に比べると多くの講演では比較的丁寧な注意深い発声がなされているといえるし，原稿に基づいて話している部分も多い．筆者らは，ディクテーションから対話音声の認識に至るまでの中間的な研究目標としては，講演の書き起こしや，議会の議事録作成などが適しているのではないかと考えている．

ただし，発声内容を1字1句間違いなく書き起こせたとしても，その中には「えー」といった間投詞なども入ることになるし，一方で，句読点がない大変読みづらい文章ができてしまう．不要語の処理や句読点の自動挿入については認識結果の後処理として行う方法だけでなく，統計的手法による音声認識と同じ枠組みの中で処理する方法も提案されている[6]．さらに，読みやすさという観点からは話し言葉を書き言葉に整形し直すようなフィルタも必要となるだろう．

自然言語理解

従来の音声理解システムではキーワードスポッティングという手法によって必要なキーワードを抽出し，それを手がかりにして解析を進めるという方法をとることが多かった．しかし，この手法ではもっぱら音響的な特徴だけでキーワードを抽出するため，言語的制約が弱く，大語彙を認識対象とすることはできない．そこで最近では，入力音声をディクテーションプログラムによって書き起こし，その結果に対して言語的な解析を行う方法が主流となっている．特に，統計的な言語理解モデル[15),16)]を用いて，誤りを含む書き起こし結果から意味を理解しようとする研究が，ARPAのATIS（Air Travel Information System）タスク☆4を中心にして進められている．

☆4 航空機による旅行の案内や予約に関する問合せを音声で行うことを目的としたタスク．

話題	入力文字数	Perplexity	CER(%)
天気予報	4,770	848.5	6.7
スポーツ	16,243	743.8	7.2
町の話題	5,323	342.0	5.3
交通情報	1,975	288.0	2.9
事故	948	249.0	1.1
外交問題	4,859	186.8	1.8
事件	10,804	177.1	1.8
経済	1,086	146.9	3.2
全体	46,866	309.0	4.7

表-2　ラジオニュースの認識実験結果

発話は読み上げではなく自由発話であり，言いよどみや言い誤り，不要語なども含んでいるため音声認識および理解の精度を上げることは容易でないが，将来的にはテレバンキングやテレショッピングなどにおいてユーザフレンドリーな入力手段としての応用が期待されている．

おわりに

ディクテーションプログラムを中心にして大語彙連続音声認識の現状について説明した．また，今後期待される応用分野についていくつかの例を紹介した．音声認識の究極の目標であるところの音声対話の実現までにはまだまだ研究課題が山積しているが，一方で，音声データのテキスト化手段としてのディクテーション技術は一応の成果をあげつつあり，ここで紹介したものに限らず，さまざまな分野での応用が期待できる．

謝辞　データ使用を許可していただいた，産経新聞社，日本経済新聞社ならびに（株）ピープルワールドカンパニーに感謝します．また，評価実験用音声データの採取にご協力いただくとともに，原稿データの使用を許可してくださった（株）東京放送に感謝します．

参考文献
1) http://www.ibm.co.jp/voiceland/index.html
2) http://www.psinfo.nec.co.jp/shabette/
3) http://www.justsystem.co.jp/product/applicat/vtaro/index.html
4) 西村雅史: 音声ワープロ最新事情, 日本音響学会誌, Vol.54, No.3, pp. 229-234 (1998).
5) 中川聖一: 小特集に寄せて－音声対話システム構築, 日本音響学会誌, Vol.54, No.11, pp.783-790 (1998).
6) 西村他: ニュース音声書き起こしシステムに関する検討, 日本音響学会講演論文集, 1-R-14 (Sep. 1998).
7) Bahl, L.R. et al.: A Maximum Likelihood Approach to Continuous Speech Recognition, IEEE Trans. PAMI-5, No.2, pp.179-190 (1983).
8) 中川聖一: 確率モデルによる音声認識, 電子情報通信学会 (1988).
9) 新島他: 音声認識による剖検記録システムの開発, 日本法医学会雑誌, Vol.52, Suppl., p.62 (1998).
10) http://www.olympus.co.jp/LineUp/VTREK/vt1000rv.html
11) 古井貞熙: 大語彙連続音声認識の現状と展望, 日本音響学会講演論文集, 1-6-10 (Mar. 1998).
12) Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop (Feb. 1998). http://www.nist.gov/speech/proc/darpa98/index.htm
13) Hauptmann, A.G. et al.: Indexing and Search of Multimodal Information, Proc. ICASSP'97, Vol.1, pp.195-198 (1997).
14) 丸山他: 字幕送出タイミング検出におけるワード列ペアモデルの構成検討, 日本音響学会講演論文集, 1-1-13 (Sep. 1998).
15) Miller, S. et al.: A Fully Statistical Approach to Natural Language Interfaces, Proc. ACL, pp. 55-61 (1996).
16) Epstein, M. et al.: Statistical Natural Language Understanding Using Hidden Clumpings, Proc. ICASSP'96, Vol.1, pp.176-179 (1996).

（平成10年12月8日受付）