論文 / 著書情報
Article / Book Information

論題	音声認識技術の現状と今後の展開 ー今、何ができて、何ができないかー
著者	古井貞熙
Author	SADAOKI FURUI
掲載誌/書名	Techno-Stream TST, Vol. 24, No. 7, pp. 6-10
Journal/Book name	, Vol. 24, No. 7, pp. 6-10
発行日 / Issue date	2001, 7

特別寄稿

Special Contribution

音声認識技術の現状と今後の展開
－今、何ができて、何ができないか－

音声認識技術は近年大きく進展しています。その主たる応用は、コンピュータシステムへの音声コマンドの入力と、種々の音声ドキュメントの文字化です。本稿では、これらの技術について解説するとともに、今後の展望について述べます。特に、自由に発声した話し言葉を音声認識するために必要な話し言葉の大規模コーパスの作成、種々の変動を含む話し言葉に対応できる適応的言語モデルと音響モデルの構築、音声から話題語を抽出し、話し手の意図を理解し、要約する機能の実現、ハンズフリーコミュニケーションのための音声認識技術の重要性などに焦点を当てます。

古井 貞熙 教授（東京工業大学大学院 情報理工学研究科 計算工学専攻）

Prof. Sadaoki Furui

音声認識技術の最近の進歩

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す。人とコンピュータとの情報交換においても、コンピュータから情報を受け取るには一般にディスプレイが最も効率的ですが、電話による情報サービスや、車を運転中のように両手がふさがっているときなどは、コンピュータへの入力手段として音声が使えれば極めて効率的になります。

電話を用いて、音声でコンピュータシステムと対話し、コンピュータシステムから種々の情報を得るサービスが、米国を中心に急速に広がっています。今後、PDA（Personal Digital Assistants）のような携帯端末によるインタフェースが主流になってくると、キーボードを装備するわけにいかないので、音声入力の役割が一層大きくなってくると予想されます。

音声認識技術によって音声を自動的に文字化することができれば、音声ワープロとしての実現形態だけでなく、会議の議事録や講演録の作成、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ます。日常生活の中で、会議、討論、講演、講義など、音声で伝達される情報は極めて多く、その重要性はマルチメディアの世の中でも大きく変わることはないでしょう。

これらの音声ドキュメントを録音して蓄えることは容易ですが、そのままでは聞き直すしかないので、その利用は極めて困難です。これらを音声認識技術によって、テキスト、要約などの利用しやすい形に変換することができれば、検索も容易になり、その有用性は極めて大きいと期待されています。

音声認識の研究は、1950年代から現在まで半世紀にわたって、世界中の研究者によって進められてきました。近年、統計的パターン認識理論に基づく技術の発展と、コンピュータの高速化に支えられて、音声認識技術は大きく進歩しています。図1に、音声認識技術の進歩を、認識対象語彙数と発声スタイルの二次元平面を用いて示します。

図の左下が比較的容易な領域で、右上への対角線に沿って難しさが増していきます。図には、1980年、1990年、そして2000年の実用化レベルが示してあります。線よりも下が、その年代でほぼ達成できている応用技術です。3本の線が平行でないのは、自由に発声した自発的な話し言葉では、語彙数が少なくても実用化のレベルに達していないこと、すなわち話し言葉の音声認識の難しさを示しています。

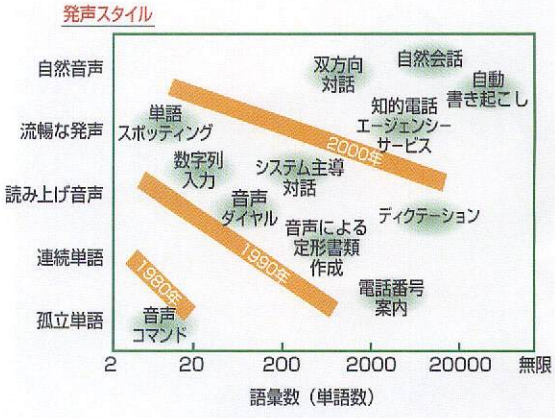

図1　音声認識技術の進歩

音声認識の基本的原理

人が音声を生成するプロセスを、通信理論のアプローチからとらえると、図2のように表すことができます[1]。メッセージ（意図）情報源で意図したメッセージMが生成され、言語チャンネルによって単語列Wに変換されます。同じメッセージを表すにも、言語や語彙の違い、個人による表現法の違いなどにより、種々の方法があります。このため言語チャンネルは、各言語について代表的な単語列を中心として確率的に機能します。

単語列Wは、調音チャンネルによって、音の系列Sとして実現されます。調音器官は話者によって異なり、同じ話者でも全く同じ音を繰り返すことはできませんので、調音チャンネルも確率的に機能します。音の系列Sは話者の口から放射され、部屋の音響特性が畳み込まれた後、音響入力信号Aとしてマイクロホンに到達します。この過程をここでは音響チャンネルと呼んでいます。音響入力信号Aは、マイクロホンによって電気信号に変換され、何らかの歪みを受けて、音声認識系にXとして入力されます。この最後の過程を、ここでは伝達チャンネルと呼んでいます。

以上の各チャンネルには、不確実性あるいは変動があり、P(W|M)、P(S|W)、P(A|S)、P(X|A)の各確率分布で表現されます。P(M)はメッセージの事前分布です。

入力された電気信号Xから、元のメッセージMを推定するのが、音声認識の目的です。残念ながら、まだ、メッセージ（意図）情報源や、メッセージを単語列に変換する言語チャンネルを表現する効果的な方法がわかっていません。このため、現在の音声認識では、単語列Wを情報源として、それを推定します。

ベイズ決定理論に基づくパターン認識理論によれば、$X=x_1, x_2, \cdots, x_T$が与えられたときの音声認識誤りを最小にするには、事後確率$P(W|X)$を最大にする単語列$X=w_1, w_2, \cdots, w_k$を選べばよいことがわかっています[2]。通常、事後確率を直接計算することはできないので、ベイズの定理に従って、$P(X|W)=P(X|W)P(W)/P(X)$のように展開します。このうち、P(X)はWとは無関係なので、事後確率を最大化するWを選ぶためには、$P(X|W)P(W)$を最大化するWを選べばよいことがわかります。またP(X|W)は、単語列Wが与えられたときの観測信号Xの確率分布で、通常HMM（隠れマルコフモデル、hidden Markov model）で表します。一方P(W)は、単語列Wの事前分布で、通常、統計的言語モデルで表します。

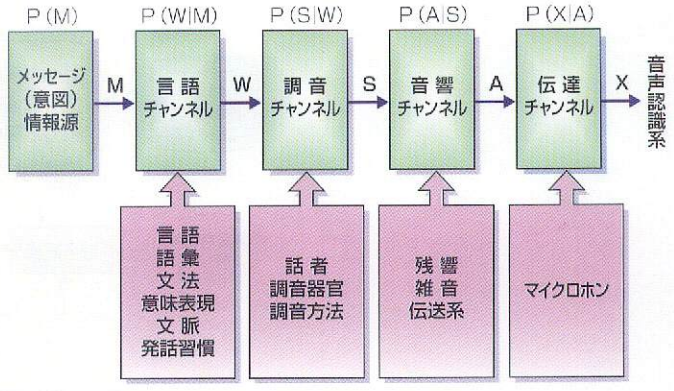

図2　通信理論のアプローチからの音声生成モデル

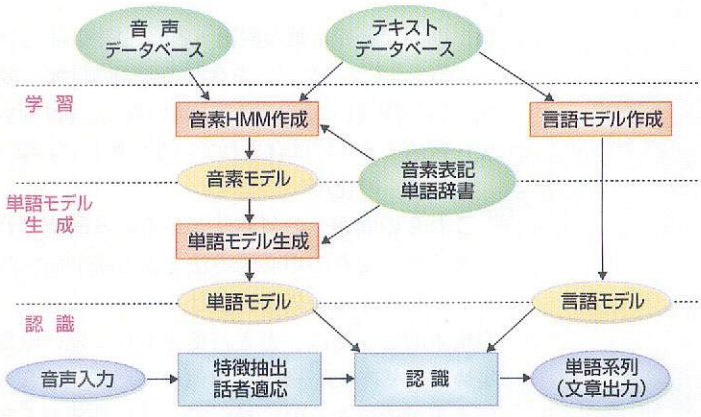

図3　音声認識系の基本的構成

HMMは、多数の話者が発声した音声を集めた音声データベースを用いて自動的に学習することができます。実際には、各音素のHMMを学習し、それを単語辞書中の音素表記に従って接続して、単語や単語列のHMMを作ります。統計的言語モデルは、大量のテキストを集めたテキストデータベースを用いて学習します。従って、現在の音声認識系の基本的構成は、図3のようになっています。

話し言葉の音声認識の難しさ

現在の音声認識技術において、過去のニュース原稿で学習した言語モデルを用いて、放送音声を音声認識すると、スタジオでアナウンサーが原稿を見ながら発声した音声であれば、95%以上の認識率が得られます。しかし、解説、現場からの中継、対談、講演などの音声に対しては、大幅に認識率が低下してしまうのが現状です。

この原因は、普通の人の発声では個々の音の発音が必ずしも明確でないこと、実環境では常に何らかの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言語的構造は、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異なることなどにあります。

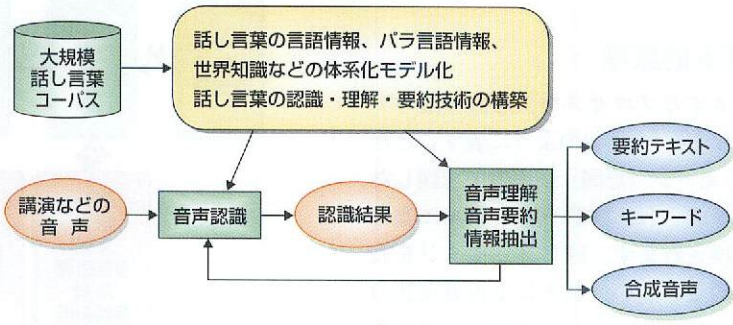

図4　話し言葉工学プロジェクトの概要

自発的な話し言葉には、次のような特徴があります。

(a) 助詞などの言葉の脱落・省略、間投詞（「あのう」「えー」など）を含む不要語の付加、倒置
(b) 言い間違い、言い淀み、言い直し、繰り返し
(c) 言語モデルには含まれていない新しい言葉（未知語）の発声

これらの問題への対処法がいろいろ研究されていますが、現在の中心的方法である統計的言語モデルと整合する有効な方法はまだありません。その根本的な原因は、書き言葉やそれに近い原稿を集めたテキストデータベース（コーパス）なら大量に入手可能であるのに対して、話し言葉の言語モデルを学習するためのテキストデータベースが十分にないことです。

このため、1999年度から、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もとづく『話し言葉工学』の構築」プロジェクト[3]が、5年計画で開始されました。

話し言葉工学プロジェクト

このプロジェクトは、話し言葉の大規模コーパスを構築し、話し言葉音声認識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、すなわち「話し言葉工学」の構築を目的としています。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すが、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ます。

プロジェクト全体の概要を図4に示します。本プロジェクトは、次のようなサブテーマを持っています。

大規模話し言葉コーパスの構築

本プロジェクトでは、種々の話題に関する講演を主体とし、のべ700～800時間、形態素（単語）にして約700万語規模の話し言葉コーパスの作成を目標としています。録音した音声を日本語として正しい表記と実際の発音の両方に書き起こすとともに、間投詞、言い直し、言い間違い、雑音などの個所を記述します。さらに形態素単位に分割し、品詞の情報などを加えるとともに、一部の音声に関しては、アクセント、イントネーション、構文解析などの結果を加える予定です。このプロジェクトをきっかけとして、放送音声などを含む多様な話し言葉のデータベースが、関係する研究者共通の財産として、継続的に集積され管理、活用されていくことが期待されています。

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

作成する大規模コーパスに基づいて、自発的な話し言葉を対象とした音声認識技術を構築します。この場合、従来の基本的方法に加えて、話し言葉特有の性質に対処できる新しい方法を作り出すことが必要です。基本的には、すべての言葉を文字にしようとするこれまでのアプローチを越えた、新しいアプローチが必要になります。

これまでのように$P(W|X)$を最大化するのではなく、図2のモデルに示すように、$P(M|X)$を最大化する発声者のメッセージMを抽出することを目標とすべきです。これにより、単純に$P(W|X)$を最大化するよりも、Wの正解精度を上げることが可能になります。さらに、メッセージ理解の結果を音声認識部にフィードバックして、メッセージ理解をやり直し、この結果を再び音声認識部にフィードバックすることにより、MとWの両方の精度を上げることも可能になります[1]。

発声者のメッセージMの表現形態、すなわち音声理解の一つの形態として、音声の要約作成があります。音声中から、話題を担っている重要語（話題語）をできるだけ文法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、我々は、各単語の重要度スコア、要約文の統計的言語スコ

ア、音声認識結果の信頼性、さらに単語間の係り受け関係をもとに、動的計画法を用いる方法を試みています[4]。

重要度スコアには、情報検索の分野で用いられているtf-idf尺度を変形した尺度を用いています。統計的言語スコアには、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の記事から計算した単語トライグラムを用いています。音声認識結果の信頼性は、認識結果として得られる単語グラフから計算される各単語の事後確率を用いています。単語の係り受け関係は、文脈自由文法にもとづいて表現しています。放送ニュース音声や、学会講演の音声認識結果を用いて、文字数にして20～70%に要約する条件で実験を行っています。

この方法は、ニュースの自動字幕作成への適用の他、会議の議事録や講演録作成など、今後の種々の応用分野への展開が期待されています。ニュース音声、講演、講義などを、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と思われます。

ハンズフリーコミュニケーションのための音声認識

音声によるコンピュータシステムへの入力が特に有用なのは、電話を通じて行う場合と、車を運転しているときのように、両手がふさがっている場合です。後者のいわゆるハンズフリーコミュニケーションに関しては、図5に示すような問題があります。すなわち、マイクロホンが発声者からある程度遠いところにあるために、スピーカからの音や種々の雑音、部屋の反響などが、音声に重畳して入力してしまうこと、発声者の位置が変動することによって、音声の特性が変わってしまうことなどにより、認識性能が低下してしまうといった問題です。

このような問題に対処するため、様々な研究がされていますが、最近、音声と唇の動きを組み合わせたバイモーダルの音声認識が盛んに研究されています。我々は、図6に示すような方法を試みています[5]。唇の動きに関しては、オプティカルフロー法を用い、その出力の統計的特徴を音声特徴に組み合わせて、音声認識を行っています。

オプティカルフロー法には、唇の輪郭などを正確に抽出できなくても、唇の動きの情報を取り出すことができる

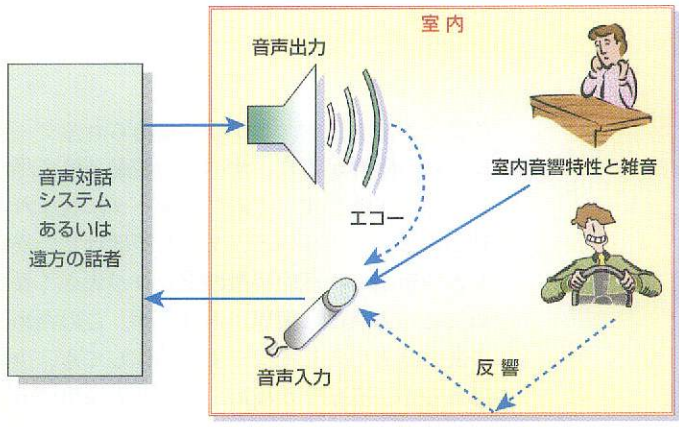

図5　ハンズフリーコミュニケーション

という特徴があります。このバイモーダル認識により、特に周囲の雑音が大きいときに、音声認識性能が向上することが確認されています。

音声の音響的・言語的変動への対処

音響モデルの適応化

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（mismatch）があるためです。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などがあります。多くの場合、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、さら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もあります[6]。

これらの種々の音声変動をカバーするような大量の音声データを用いて、統計的方法によりモデルを作成すれば、ある程度高い認識精度を達成す

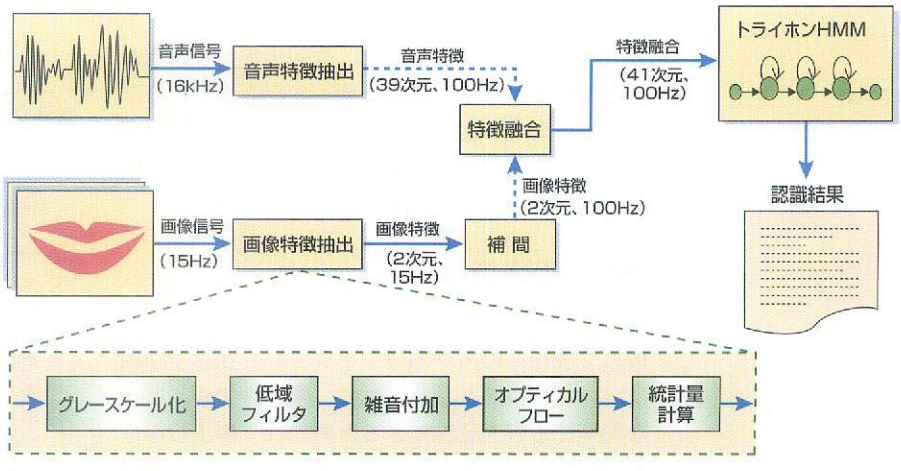

図6　オプティカルフローを用いたバイモーダル音声認識

ることができます。しかし、集められるデータ量には自ら限界があり、すべての変動条件を尽くすことは不可能です。さらに、データに含まれる変動があまりに大きいと、モデルがぼけて、異なる音素や単語の物理的特徴の重なりが大きくなるため、この方法には限界があります。このため、大多数の話者に対してはうまく動作しても、少数ではあるが認識精度が極めて低い話者を生ずることがあります。

これらの種々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もとづいて、変動に自動追随するような適応機能をもたせる研究が広く行なわれています。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「教師あり学習」と、任意の音声を用いて規則を学習する「教師なし学習」があります。

前者では、規則の獲得は比較的容易ですが、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で、煩わしくなります。後者では、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、全く意識せずに使えて便利です。極めて個人差や歪みの大きい音声に対しても、この機能を実現する方法を追及していくべきでしょう。さらに加えて、認識装置を使っているうちに自動的に性能が向上していく、オンライン逐次型適応が望ましいといえます。

言語モデルの適応化

音声認識の対象が変わるごとに、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ありません。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なわれています。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、対象に依存する部分と、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ます。

将来の音声認識

これまで述べたように、音声認識システムには、極めて多様な応用が期待できますが、任意の話題について、自然な話し言葉で、誰の声でも、どんな環境でも音声認識できるためには、解決しなければならない課題が多く残っています。主たる課題は、話し言葉特有の言語的・音響的変動、周囲雑音や電話音声の歪み、音声の個人差などに対するロバストなモデルとアルゴリズムの構築です。その基礎となる話し言葉の大規模コーパスや、実環境での音声を大量に収録したデータベースの構築も極めて重要です。今後、音声認識の研究は、単に音声を単語列に変換するだけでなく、発声者の意図の理解、話題語（フレーズ）抽出、内容の自動要約へと展開し、さらに幅広い応用が開けてくるものと期待されます。

コンピュータの小型化、高性能化、低価格化により、ユビキタス（遍在）計算とウェアラブル計算環境の時代が来るといわれています[6]。このような環境下では、キーボード入力はなじまないので、音声認識がヒューマンインタフェースの基本的手段の一つとして広く使われるようになると予想されます。その際、音声認識の多くは、個人が身につけ、個人に特化したマイクロホンおよびコンピュータで行なわれるようになり、そのシステム構成は、これまでとは大きく異なった形になるでしょう。

雑音などの環境情報は常時モニタでき、タスクに依存した知識は個人のコンピュータへ即座にダウンロードできるので、タスクへの適応性とロバスト性の向上が期待できます。ただし、それらの実現のためには、解決しなければならない課題も多く、今後の多角的な研究の展開が必要です。

参考文献

(1) B.-H. Juang and S. Furui : Automatic recognition and understanding of spoken language-A first step towards natural human-machine communication, Proc. IEEE, 88, 8, pp.1142-1165 (2000)
(2) 古井貞熙：“音声情報処理”、森北出版 (1998)
(3) 古井貞熙、前川喜久雄、井佐原均：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-、日本音響学会誌、56、1、pp.752-755 (2000)
(4) 堀智織、古井貞熙：係り受けSCFGに基づく音声自動要約法の改善、電子情報通信学会技術報告、SP2000-116 (2000)
(5) K. Iwano, S. Tamura and S. Furui: Bimodal speech recognition using lip movement measured by optical-flow analysis, Proc. Int. Workshop on Hands-Free Speech Communication (HSC 2001), Kyoto (2001)
(6) 古井貞熙：Ubiquitous／Wearable Computing環境における音声認識の展望、電子情報通信学会技術報告、SP99-114 (1999)