論文 / 著書情報
Article / Book Information

論題(和文)	話し言葉の音声認識と自動要約 ―話し言葉コーパスの必要性―
Title	
著者(和文)	古井貞熙
Author	SADAOKI FURUI
出典(和文)	国立国語研究所公開研究発表会 「話し言葉のデータベース ―『日本語話し言葉コーパス』―」 講演予稿集, Vol. , No. , pp. 21-26
Journal/Book name	, Vol. , No. , pp. 21-26
発行日 / Issue date	2003, 12

話し言葉の音声認識と自動要約
－話し言葉コーパスの必要性－

古井　貞熙

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 東京都目黒区大岡山 2-12-1

E-mail: furui@cs.titech.ac.jp

あらまし　書き言葉を読み上げたような音声だけでなく、自然な話し言葉を音声認識し、発声者の意図を自動的に抽出するための、統計的枠組みの構築が求められている。そのため、話し言葉に対する音声認識精度を向上させるための、話し言葉コーパス（CSJ）を用いた音素モデルと言語モデルの構築、さらにそれらの話者への適応化法に関する研究を行った。次に、音声を文字に変換することを目的としてきた従来の音声認識の枠組みを超えて、音声認識結果から発声者の意図を抽出する枠組みの研究を行った。話題を担う重要文や重要語を自動的に抽出して、発声者の意図を伝える要約文を自動的に生成する統計的方法を提案した。本方法を講演音声などに適用し、人が作成した要約文との比較により、その有効性を示した。最後に、話し言葉の音声変動をモデル化する方法として、ベイジアンネットの利用について述べ、話し言葉コーパスの構築や利用を含む、今後の音声認識研究の課題について述べる。

キーワード　話し言葉、コーパス、音声認識、音声要約、適応化

Spontaneous Speech Recognition and Summarization
－Effectiveness of Spontaneous Speech Corpus－

Sadaoki FURUI

Tokyo Institute of Technology, Department of Computer Science, 2-12-1 Ookayama, Meguro, Tokyo, 152-8552 Japan

E-mail: furui@cs.titech.ac.jp

Abstract　Although read speech and similar types of speech can be recognized with high accuracy, it is crucial to build a statistical framework for recognizing spontaneous speech and automatically extracting the intensions of the speakers for broadening the application of speech recognition. From this point of view, we have been investigating methods for constructing acoustic and language models using the spontaneous speech corpus (CSJ) and adapting the models to each speaker. We have also investigated the framework of extracting speakers' intensions using speech recognition results, instead of simply converting speech into text. Important sentences and words are automatically extracted and concatenated to produce a summary which conveys intensions of the speaker based on a statistical method. The proposed method has been applied to various speech including lectures and evaluated in comparison with manually created summaries. Effectiveness of the method has been confirmed. For modeling acoustic variations of spontaneous speech, the Bayesian network method has been investigated. The paper finally presents future research topics of speech recognition research, including construction and application of spontaneous speech corpus.

Keyword　Spontaneous speech, corpus, speech recognition, speech summarization, adaptation

1. はじめに

コンピュータによる音声認識の技術は、最近の 10 年ないし 20 年の間に、大きく進歩している。図 1 に、種々の話し方と認識できる単語の種類を基準にした技術的進歩の様子と、代表的な応用技術を示す[1]。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す。図には、1980 年、1990 年、そして 2000 年の実用レベルが示してある。線よりも下側が、その年代でほぼ達成できている応用技術である。この図からわかるように、単語で区切って発声した音声や、書き言葉に近い極めて丁寧に発声された音声の認識は、かなり高い精度でできるようになってきた。

しかし、図中の 3 本の線が平行でないことで示されているように、自然な話し言葉音声に対しては、語彙数が少なくても、まだ実用には程遠い状況にある。その主たる原因の一つは、後述するように、現在の音声認識が統計的な音素モデルと言語モデルに基づいており、対象とする音声の性質を表現する統計モデルの構築に、大規模なデータベース（コーパス）を必要とすることにある。新聞などの書き

言葉の大量なテキストを手に入れるのは、現在では難しくないので、それを用いて作成した言語モデルや、それらを読み上げた音声（読み上げ音声）を用いて作成した音素モデルを用いれば、書き言葉に近い音声の認識で比較的高い精度を得るのは難しくない。ところが、話し言葉に関しては、それを文字化するのに大きな労力を要するため、話し言葉の大規模なコーパスを得るのが難しく、それが話し言葉のモデル構築の大きな障害となっている。

コミュニケーションにおける音声の主要な役割は、自然に話しながら意図を伝達することである。自然な話し言葉には、「あのー」「えーと」などの不要語、言い直し、言いよどみ、言葉の省略、繰り返しなどが必ず含まれる。これまでの音声認識では、発声されたすべての言葉を正しく文字に変換することを目標にして来たが、話し言葉を対象とする場合は、不要な言葉や言い直しなどはむしろ無視し、発声者が伝えようとする内容を表す文を生成することが必要である。これまでの「正しい単語列」を最終目標とする理論ではなく、「正しい意味内容」を最終目標とする音声理解理論を構築することが求められている。

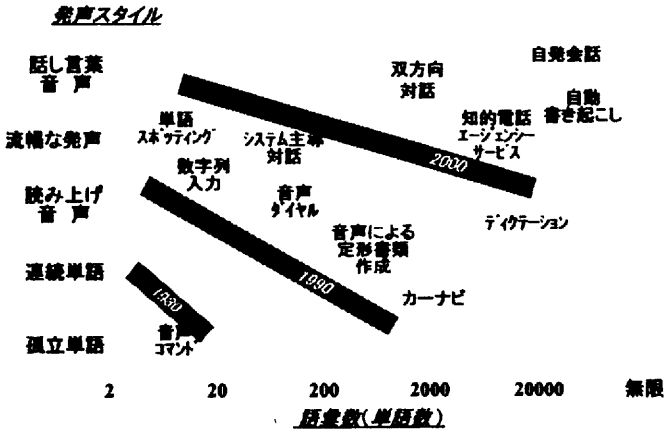

図1　音声認識技術の進歩

2. 現在の音声認識の原理

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る[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 に変換される。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、種々の方法がある。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する。単語列 W は、調音チャンネルによって、音の系列 S として実現される。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ないので、調音チャンネルも確率的に機能する。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する。この過程をここでは音響チャンネルと呼んでいる。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 として入力される。この最後の過程を、ここでは伝達チャンネルと呼んでいる。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る。$P(M)$はメッセージの事前分布である。

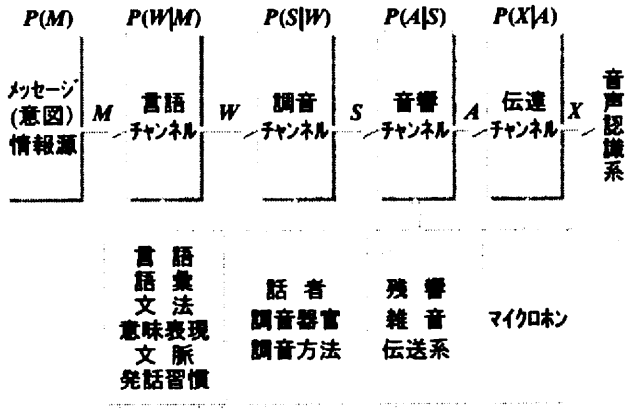

図2　通信理論のアプローチからの音声生成モデル

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、本来の音声認識の目的である。残念ながら、まだ、メッセージ（意図）情報源 $P(M)$や、メッセージを単語列に変換する言語チャンネル $P(W|M)$を表現する効果的な方法がわかっていない。このため、現在の音声認識では、単語列 W を情報源として、それを推定する[2]。ベイズ決定理論に基づくパターン認識理論によれば、$X=x_1, x_2, \ldots, x_T$ 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 $W=w_1, w_2, \ldots, w_k$ が選べばよいことがわかっている。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、$P(W|X) = P(X|W)P(W)/P(X)$ のように展開する。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る。このうち $P(X|W)$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）で表す。HMM は、各単語や音素を確率状態遷移機械（マルコフモデル）で表現したもので、音声の変動を確率モデルとして表現するのに適している。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す。実際には、単語の2連鎖（バイグラム）ないし3連鎖（トライグラム）の確率で表現する。

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る。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 を作るのが一

般的である。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する。従って、現在の音声認識系の基本的構成は、図３のようになっている。

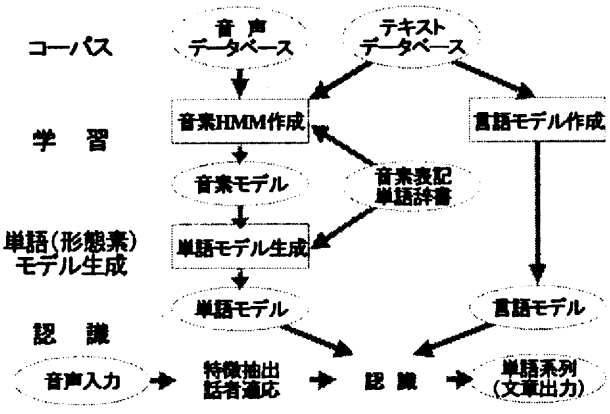

図３　音声認識系の基本的構成

3. 話し言葉を音声認識するための基本技術の構築

科学技術振興調整費開放的融合研究推進制度による「話し言葉工学」プロジェクト[3]で、学会講演音声のコーパス（CSJ: Corpus of Spoken Japanese）が構築されるようになったため、それを用いて、話し言葉を音声認識するための音素モデルおよび言語モデルの構築と、それを用いた評価実験を行った。その結果、次のような成果が得られた[4]。

(1) 読み上げ音声から作成した音素モデルと、Web 上の講演録（講演音声を読みやすいように書き言葉に近い形で整形されているもの）から作成した言語モデルを用いた場合に 55%程度あった誤認識を、CSJ から作成した音素および言語モデルに置き換えることにより、35%程度にまで低下させることができた。図４に、種々のモデルを用いたときの、認識実験の結果を示す[5]。音素モデル、言語モデルのいずれに関しても、実際の自然な話し言葉音声から作成することが必須であることがわかる。

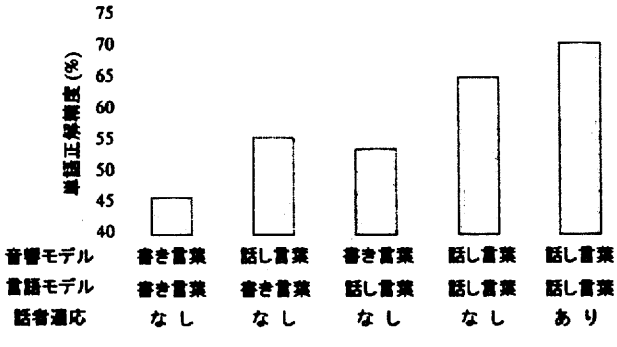

図４　自然な講演音声の音声認識結果

(2) 発声速度、言い直し頻度、および未知語率が、話者による認識率の違いの主要な要因となっている[6]。すなわち、概して、早口の音声、言い直しの多い音声、および音声認識用辞書に登録されていない語彙を多く発声している音声の誤認識が大きい。

(3) 単語の連鎖確率に基づく言語モデルを用いているため、1%の言い直し頻度と未知語率の増加は、それぞれ、3.3%と 2.3%の単語正解精度の低下に拡大される[6]。

(4) フィラー（「えー」「あー」のような言葉）の回数も個人差が大きい。

(5) 話し言葉では、文の区切りが必ずしも明確でなく、極端に長い文が頻繁に発声される。このため、書き言葉を読み上げた音声を認識する場合と異なり、文単位に区切らずに連続して音声認識（デコーディング）する方がよい[7]。

(6) 話し言葉のコーパスは極めて重要で、今後さらに拡張することにより、話し言葉の音声認識に、より適した音素モデルや言語モデルを構築できると予想される。

(7) 発声された音声の話題（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、その分野の教科書などを用いて適応化するのが有効[5]。

(8) 図４の「話者適応あり」と「なし」の比較からわかるように、声の個人差に対処するため、音素モデルを自動的に話者適応することによって、認識精度を向上させることができる[5]。

4. 音素モデルと言語モデルの話者適応化

上述の(8)で述べたように、音素モデルを話者に自動的に適応化することにより、音声認識性能が向上する。その際、音声認識結果には誤認識が含まれることがあるため、認識結果をそのまま用いて音素モデルを適応化すると、不適切なモデルが作られる可能性がある。そこで、音素を複数のクラスに分類し、各音素クラス内での適応化を、尤度最大化基準を満たす線形変換に拘束し、それを一つの講演全体に対して繰り返すことにより、適応化を行う方法を提案した。これにより、誤認識があっても、認識性能を向上させることができることが確認された[5]。図４における話者適応は、そのような原理で行ったものである。

また、言語モデルに関しても、話者や話している内容が変わるごとに、それに応じて適応化できることが望ましいが、そのために大量の（書き起こし）テキストを集めるのは現実的ではない。このため、単語を自動的にクラス化し、クラス言語モデルに基づいて、共通の基本的な言語モデルを、一つの講演全体の認識結果を用いて、適応化させる方法を提案した[8]。これにより、音素モデルの適応化法と同様に、誤認識があっても、認識精度が向上することが確認された。さらに、音素モデルと言語モデルの適応化を組み合わせることにより、一層の性能向上が達成できることがわかった。

5. 音声理解・要約の基本技術の構築

話し言葉音声を認識・理解するためには、従来の基本的認識方法に加えて、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という視点から、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超えた、新しいアプローチについて研究を進めた。図2に示したモデルで、これまでのように $P(W|X)$ を最大化するのではなく、$P(M|X)$ 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した。$P(W|M)$ をモデル化するためのコーパスを作成するのは極めて困難であるので、メッセージ M に依存した特有の単語の共起関係があると考え、そのモデルで近似した。この定式化により、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できることが確認された[9]。

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、音声の自動要約法について研究を進めた。発声者の意図が過不足なく表現された要約文が生成できれば、音声の意味表現ができた、すなわち音声が理解できたと考えられるからである。そこでまず、音声認識結果として得られる単語列から、話題を担っている重要語（話題語）を自動的に抽出する研究を行った。重要語を選ぶ種々の尺度について検討した結果、情報検索の分野で用いられている tf-idf 尺度を変形した情報量尺度が最も優れていることがわかった[9]。次に、重要語を中心とする単語セットを、できるだけ文法的制約に従い、かつ意味が変わらないように最適に選び、それらを接続して要約文を作成する方法について研究を進めた。図5に、音声自動要約システムの構成を示す[10]。

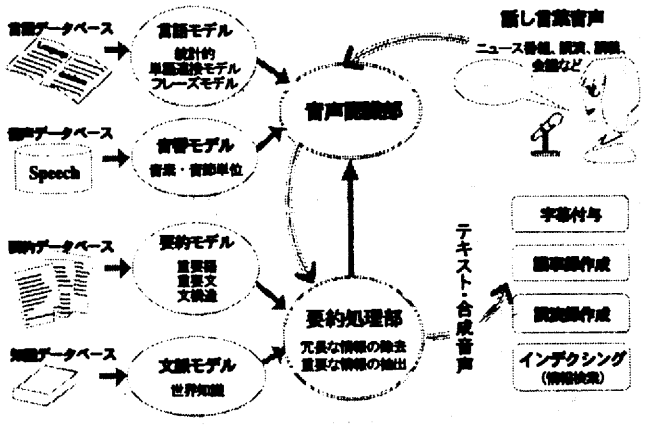

図5　音声自動要約システムの構成

無数にある単語の組み合わせの中から、上記の目的を達成する単語セットを効率的に抽出するため、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、これらのスコアの総和が最大となる単語セットを、動的計画法を用いて自動的に選択する方法を提案した[11]。重要度スコアには、上記と同じ情報量尺度を用いている。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る。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る。単語の係り受け関係は、文脈自由文法に基づいて表現している。これら4種類のスコアは、それぞれ、話題を担う単語の重視、読みやすい文の生成、音声認識誤りが要約文に含まれることの防止、要約によって文の意味が変わってしまうことの防止に寄与する。

さらに、特に圧縮率が高い場合に効果的な方法として、上記の種々のスコアをもとに、多数の文の中からまず、意味的に重要で、音声認識結果としての信頼性の高い文を抽出し、抽出された各重要文に対して、上に示した方法でさらに圧縮（要約）する方法を提案した[12]。これらの方法に関して、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い、人が要約した結果に近い結果が得られることが確認された。

提案した要約法は、音声・言語コーパスさえあれば、言語に依存せずに適用することができるため、英語の放送ニュース音声にも適用して評価実験を行った。その結果、日本語に対する場合と同様に、英語の音声に関しても効果的であることが確認できた[13]。

今後、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、種々の応用分野への展開が期待されている。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であろう。

6. 音声から音声への要約技術

音声の要約結果をテキストで提示する場合、音声認識誤りが避けられないため、読みづらくなるだけでなく、誤った情報をユーザに伝える可能性がある。また、韻律によって伝えられる話者の意図や感情情報を伝えることが難しい問題がある。元の音声をそのまま用いて提示すれば、それらの問題を回避することができる。

そこで、上で述べた音声要約手法を用いて、要約結果を音声で提示する「速聞きシステム」について研究を行った[14]。図6に、システムの構成を示す。要約音声は、テキストで出力された音声要約結果に対応する音声を、元の音声から切り出し接続して作成する。切り出した音声どうしをそのまま接続したのでは、音声の振幅や波形の差から、雑音が生ずる可能性があるので、音声端の振幅を徐々に減衰させ、前後の言語的関係に依存した長さを有するポーズ（無音区間）を挿入してから接続する。

要約音声を作成するのにふさわしい切り出し単位を調べるため、単語単位、文単位、文をさらにフィラーで区切った単位、の3つの単位の選択により作成した要約音声について、「聞きやすさ」と「要約音声としてのふさわしさ」の2項目で、聴取実験による評価を行った。講演音声を用いた実験の結果、要約率50%では、文単位選択による要約が最も有効であることが確認された。

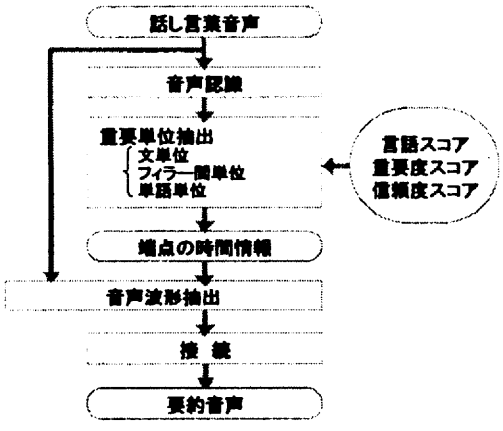

図6　速聞きシステムの構成

7. ベイジアンネットによる音声のモデル化

3章で述べたように、話し言葉音声の認識誤りを生ずる主たる要因の一つに、発声速度の変動がある。発声速度は、話者や文単位で変動するだけでなく、文の中、さらに単語の中でも音素ごとに変動する。このため、発声速度の時間的な変動に適応する方法として、発声速度変動を隠れ変数を用いてモデル化するベイジアンネットを用いた音響モデルについて研究を進めた[15]。動的ベイジアンネット（DBN; Dynamic Bayesian Network）は、HMMを含む様々な確率モデルを記述できる柔軟性があり、一連の学習・推論アルゴリズムが開発されている[16]。

提案したモデルにおける入力の1フレームに対応するベイジアンネットを図7に示す。ベイジアンネットでは、ノードは確率変数に対応し、ノード間の有向枝は、確率的な依存関係を示す。このモデルは、従来のHMMのパラメータ（音素と、音素HMM内の状態位置）に加えて、発声速度の状態を表す隠れ変数を持ち、その値に応じて音響特徴量の出力確率密度および遷移確率の制御が行われる。モデルの学習・評価（認識）時には、図7のネットワークは、与えられた音声の音響特徴量ベクトル列の長さに合わせて展開される。実験の結果、CSJおよび英語の会議音声の認識で、従来モデルよりも高い認識率が得られることが確認された。

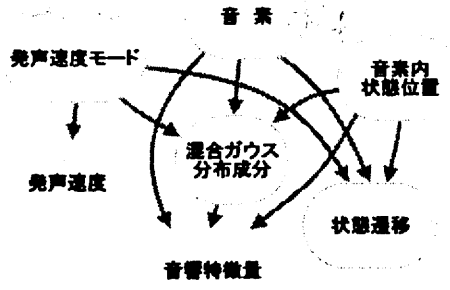

図7　発声速度変動を隠れ変数を用いてモデル化するベイジアンネット

8. これからの音声認識・理解

音声認識システムには、極めて多様な応用が期待できるが、任意の話題について、自然な話し言葉に対して、誰の声でも、どんな環境でも正しく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る。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ある。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ある。

これまでの音声認識では、主としてスペクトル包絡情報とパワー情報が用いられ、人が音声を認識するときに重要な役割を果たしている基本周波数などの韻律情報は、ほとんど用いられていない。限られたタスクの音声認識では、基本周波数の有効性が示されているが[17]、大語彙の話し言葉音声認識では、まだほとんど成功していない。CSJのコア部分には、韻律情報が付与されているので、今後これらの解析や、音声認識、要約などへの利用技術の研究が進むと期待される。

テレビのニュース、学会や一般の講演、大学の講義、インタビューなど、毎日大量に生まれているこれらの音声ドキュメントを、録音しておいただけでは利用が難しい。このため、音声認識して文字化し、自動的に内容のインデックスをつけることが期待されている。今後の音声認識の研究は、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、話題語あるいはキーフレーズの抽出、内容の自動要約へと展開し、これによって、誰が何を言ったかといった情報を含め、いろいろな人が持っている知識を即座に流通させ、活用することができるようになると期待される。さらに、それらの知識を、Webなどの多様な知識と関連付け、体系化することができるかどうかが、21世紀の重要な研究課題となると予想される。そのような体系化された知識が利用できるかどうかが、逆に、音声認識・理解技術の発展のかぎとなるであろう[18]。

コンピュータの小型化、高性能化、低価格化により、ユ

ピキタス（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る。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る。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と予想される。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る。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ある[19]。

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, 東京, 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, vol.56, no.1, pp. 752-755, 2000

[4] 古井貞熙 "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

[5] T. Shinozaki and S. Furui "Towards automatic transcription of spontaneous presentations" Proc. Eurospeech 2001, pp. 491-494, 2001

[6] T. Shinozaki and S. Furui "Analysis on individual differences in automatic transcription of spontaneous presentations" Proc. 2002 IEEE Int. Conf. Acoust., Speech, Signal Process., SP-P11.07, pp. I-729-732 , 2002

[7] T. Kawahara, H. Nanjo and S. Furui "Automatic transcription of spontaneous lecture speech" Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU2001), Madonna di Campiglio, Italy, 2001

[8] 横山忠介, 篠崎隆宏, 岩野公司, 古井貞熙 "言語モデルのバッチ型教師なし適応化法" 電子情報通信学会技術報告, SP2002-151, 2002

[9] K. Ohtsuki, S. Furui, A. Iwasaki and N. Sakurai "Message-driven speech recognition and topic-word extraction" Proc. 1999 IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 525-628, 1999

[10] 堀智織, 古井貞熙 "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, J85-D-II, 2, pp. 200-209, 2000

[11] C. Hori and S. Furui "Advances in automatic speech summarization" Proc. Eurospeech 2001, pp. 1771-1774, 2001

[12] 菊池智紀, 古井貞熙, 堀智織 "重要文抽出と文圧縮による音声自動要約" 電子情報通信学会技術報告, SP2002-158, 2002

[13] C. Hori, S. Furui, R. Malkin, H. Yu and A. Waibel "Automatic speech summarization applied to English broadcast news speech" Proc. 2002 IEEE Int. Conf. Acoust., Speech, Signal Process., SP-L01.03, pp. I-9-12. 2002

[14] 新中庸介, 菊池智紀, 岩野公司, 古井貞熙 "音声自動要約を利用した講演速聞きシステムの検討" 情報処理学会研究報告, 2003-SLP-46(3), 2003

[15] 篠崎隆宏, 古井貞熙 "発話変動を考慮した隠れモードHMM による音声のモデル化 - 音声認識におけるベイジアンネットの応用 -" 電子情報通信学会技術報告, SP2003-41, 2003

[16] J. Bilmes and G. Zweig "The Graphical Models Toolkit: An open source system for speech and time-series processing" Proc. 2002 IEEE Int. Conf. Acoust., Speech, Signal Process., ITT-P02.07, pp. IV-3916-3919 , 2002

[17] 岩野公司, 関高浩, 古井貞熙 "雑音に頑健な音声認識のための韻律情報の利用" 情報処理学会研究報告, 2003-SLP-46(10), 2003

[18] 古井貞熙 "音声認識と知識の体系化" 日本音響学会誌, vol. 59, no. 10, pp. 589-590, 2003

[19] 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