論文 / 著書情報
Article / Book Information

論題(和文)	統計的枠組みによる話し言葉の音声理解
Title(English)	
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	ＳＣＡＴ Technical Journal, Vol. , No. 38, pp. 43-49
Citation(English)	, Vol. , No. 38, pp. 43-49
発行日 / Pub. date	2003,
Note	本研究は平成１０年に研究費助成に採用され、平成１１年度から１３年度にかけて財団法人テレコム先端技術研究支援センターの研究費助成を受けたものです。

統計的枠組みによる話し言葉の音声理解

Spontaneous Speech Understanding Based on Statistical Frameworks

古井貞熙
Sadaoki FURUI, Dr. Eng., Department of Computer Science, Professor, Tokyo Institute of Technology

東京工業大学大学院情報理工学研究科教授
1970年東京大学大学院工学系研究科修士課程修了
電子情報通信学会フェロー　米国音響学会フェロー　IEEEフェロー　情報処理学会　計算言語学会　日本音響学会　国際音声通信学会（ISCA）　会員
受賞：1975年電子情報通信学会学術奨励賞　1988年、1993年電子情報通信学会論文賞　1985年、1987年日本音響学会論文賞　1989年科学技術庁長官賞研究功労者表彰　1990年電子情報通信学会著述賞　1993年IEEE ASSP Society論文賞　他
著書：「ディジタル音声処理」東海大学出版会　1985年
"Digital Speech Processing, Synthesis and Recognition", Marcel Dekker, (1989)　"Advacces in Speech Signal Processing", (編著), Marcel Dekker, (1992)　「音響・音声工学」近代科学社 1992年　「音声情報処理」森北出版 1998年　他
研究専門分野：音声情報処理　マルチモーダルインタフェース

あらまし　従来の、音声を文字に変換することを目的としてきた音声認識の枠組みを超えて、自然な話し言葉から発声者の意図を自動的に抽出するための、統計的枠組みを構築することを目的に本研究を進めた。

まず、話し言葉に対する音声認識精度を向上させるための音素モデルと言語モデル、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、音声認識結果から発声者の意図（メッセージ）を抽出する枠組みの研究を行い、話題を担う重要語を自動的に抽出する方法について検討した。さらに、自動的に選択された重要文と重要語を中心に単語のセットを抽出し、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、日本語のニュース音声と講演音声、さらに、英語のニュース音声に適用し、人が作成した要約文との比較により、提案法の有効性を示した。

1. はじめに

コンピューター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。この図から分かるように、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。NHKのアナウンサーが原稿を読んでいる音声を音声認識し、自動的に字幕を作成するシステムが、そのよい例である。

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。たとえば、NHKのニュース音声にこれまでの方法を適用すると、原稿を読んでいるアナウンサーの音声に対しては95％以上の単語認識率が得られるが、それ以外のインタビューや現場からのリポートの音声に対しては、80％程度の認識率しか得られない。

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。現在の音声認識は、書き言葉のテキストデータベースに基づく言語モデルに基づいており、書き言葉を発声した音声を用いて音素モデルを作成しているため、話し言葉音声が持つ種々の問題に対応することが難しい。

このため本研究では、話し言葉を理解するための基本的な理論や、モデルを構築することを目的とした。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

を目指した。音素モデルや言語モデルに関しても、話し言葉のデータベースを用いて、話し言葉における変動の統計的性質を分析し、これに基づいたモデルを構築し、認識性能の向上を試みた。

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捉えると、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、$P(W \quad M)$、$P(S \mid W)$、$P(A \quad S)$、$P(X \quad A)$ 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。

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 $P(W \mid M)$ を表現する効果的な方法が分かっていない。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 $P(W \mid X)$ を最大にする単語列 $W = w_1, w_2, \ldots, w_k$ を選べばよいことが分かっている。通常、事後確率を直接計算することはできないので、

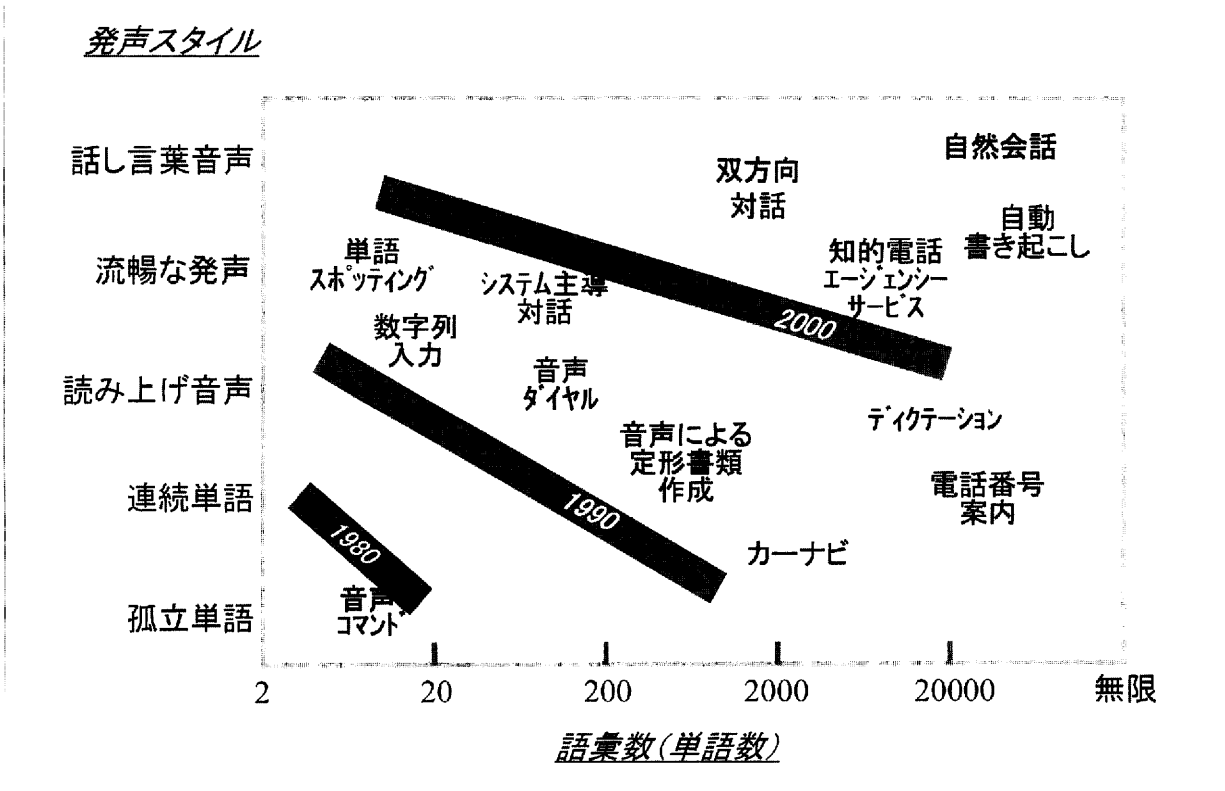

図1 音声認識技術の進歩

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分かる。

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。したがって、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。

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]で構築されている、学会講演音声のデータベース（CSJ : Corpus of Spoken Japanese）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分かる。

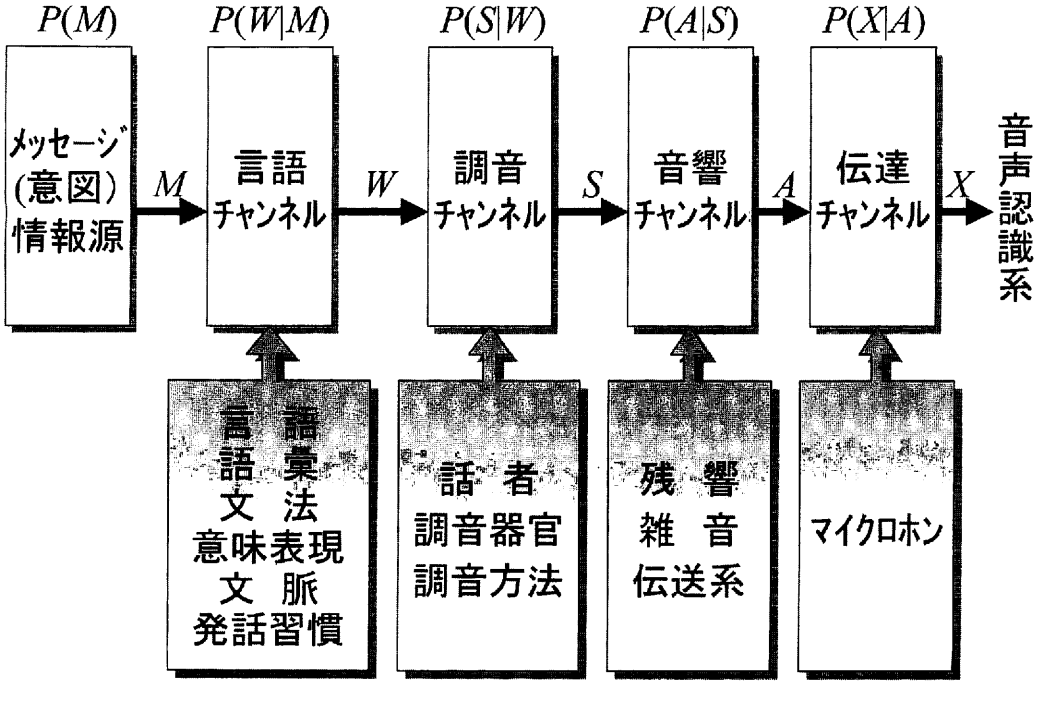

図 2 通信理論のアプローチからの音声生成モデル

(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) 話し言葉のデータベースは極めて重要で、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である[5]。

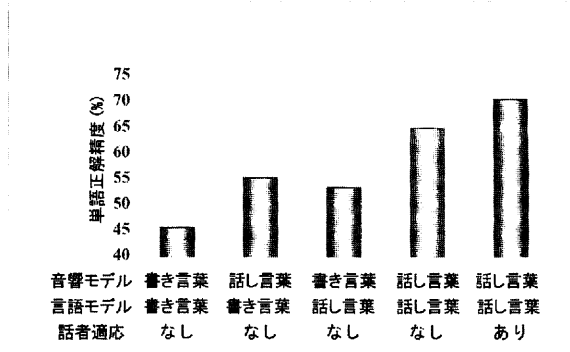

図4 自然な講演音声の音声認識結果

(8) 図4の「話者適応あり」と「なし」の比較から分かるように、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)項で述べたように、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

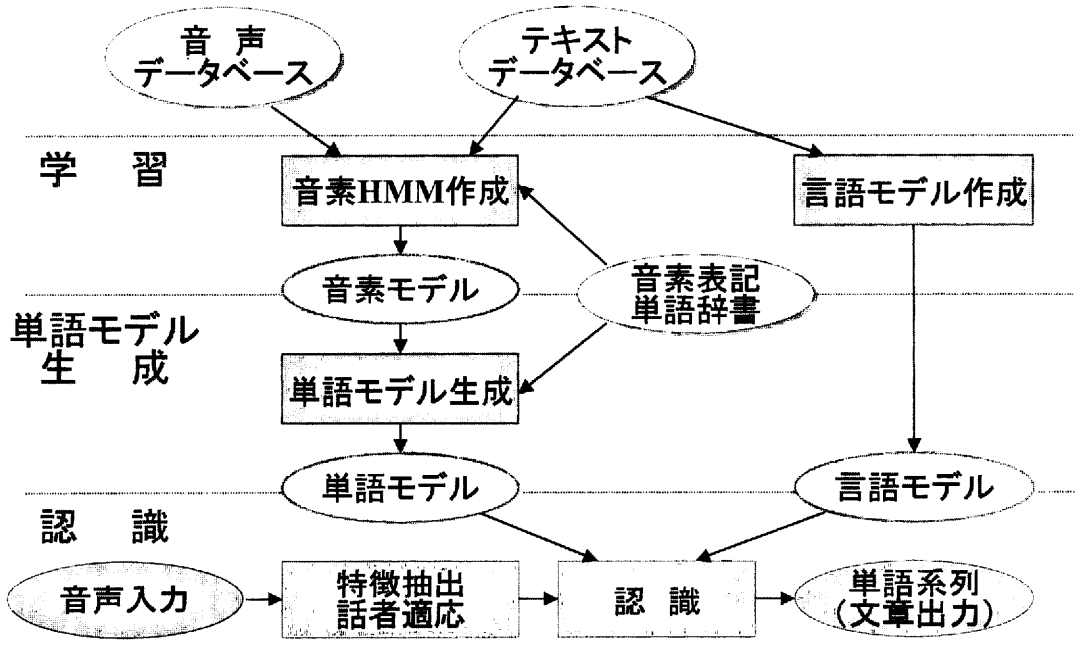

図3 音声認識系の基本的構成

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、限られた量のテキスト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分かった。

5. 話し言葉を音声理解・要約するための基本技術の構築

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]。

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、その結果に基づい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という方法で、M と W の両方の精度を上げることができる可能性もある。

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分かった[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]。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。また、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・言語データベースさえあれば、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. 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データベースや、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。

今後、音声認識の研究は、本研究がターゲットとしたように、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、これによってさらに幅広い応用が開けてくると期待される。

コンピューターの小型化、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る[14]。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ーで行われるようになり、そのシステム構成は、これまでとは大きく異なった形

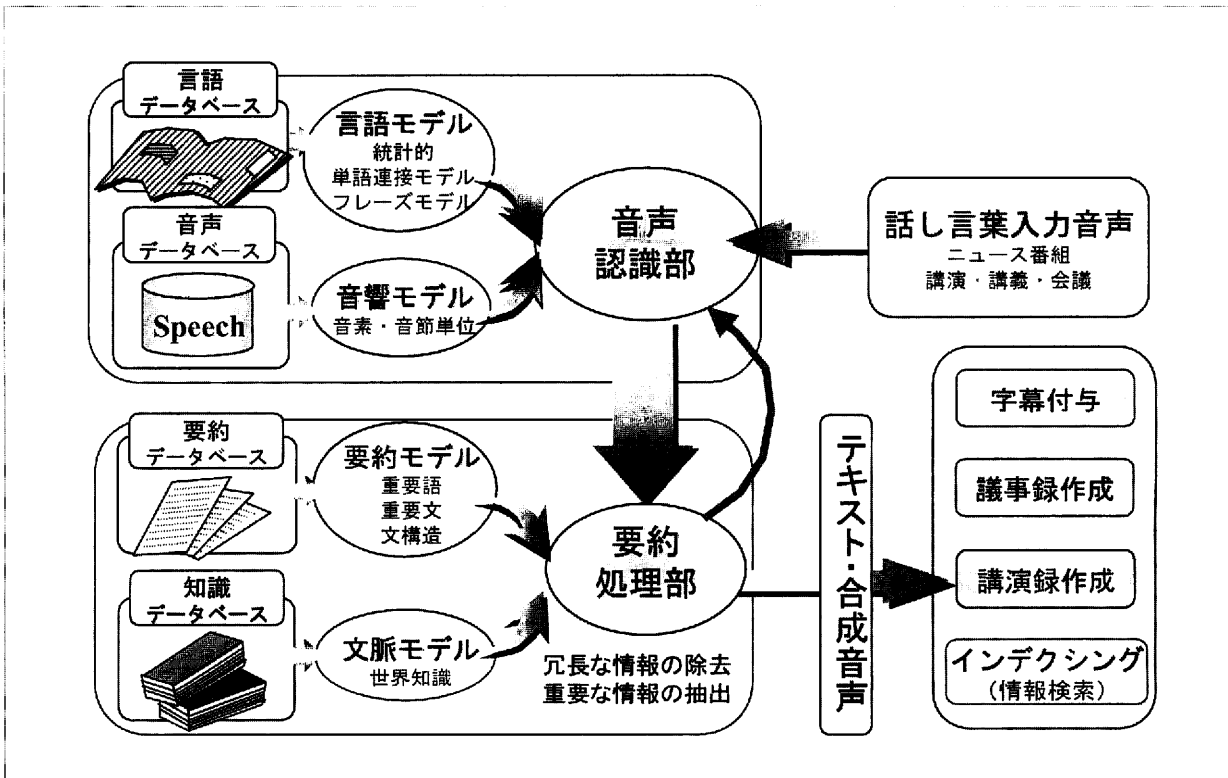

図5 音声自動要約システムの構成

になると予想される。

雑音などの環境情報は常時モニターでき、タスクに依存した知識は個人のコンピューター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。

参考文献

(1) Juang, B.-H. and Furui, S. : Automatic recognition and understanding of spoken language-A first step towards natural human-machine communication, *Proc. IEEE*, **88**. **8**, (2000), pp. 1142-1165.

(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「大規模コーパスに基づく『話し言葉工学』の構築」(科学技術振興調整費開放的融合研究推進制度), 日本音響学会誌 **56**, **1**, (2000), pp. 752-755.

(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) Shinozaki, T. and Furui, S. : Towards automatic transcription of spontaneous presentations, *Proc. Eurospeech 2001*, (2001), pp. 491-494.

(6) Shinozaki, T. and Furui, A. : Analysis on individual differences in automatic transcription of spontaneous presentations, *Proc. 2002 IEEE Int. Conf. Acoust., Speech, Signal Process.*, **SP-P11.07**, (2002), pp. I-729-732.

(7) Kawahara, T., Nanjo, H. and Furui, S. : 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、古井貞熙「講演音声を対象とした言語モデルの話者適応化」日本音響学会秋季研究発表会, **3-9-6**, (2002).

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

(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**, (2000), pp. 200-209.

(11) Hori, C. and Furui, S. : Advances in automatic speech summarization, *Proc. Eurospeech 2001*, (2001), pp. 1771-1774.

(12) 菊池智紀、古井貞熙、堀　智織「音声自動要約における重要文抽出の応用」日本音響学会秋季研究発表会, **2-9-18**, (2002).

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

(14) 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).

この研究は、平成10年度SCAT研究助成の対象として採用され、平成11年度～13年度に実施されたものです。