話し言葉の音声理解へ
― 存在感のある研究を期待して ―

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
Tel/Fax: 03-5734-3480, Email: furui@cs.titech.ac.jp

あらまし　最近大きく進展しつつある大語彙連続音声認識の研究の流れをさらに進めるには、話し言葉の音声理解を実現する研究が重要であることを述べ、特にどのような視点からの研究が必要かを述べる。具体的には、言葉の省略、不要語の付加、倒置、言い直し、未知語などの話し言葉特有の問題に対処するため、意図駆動音声認識の枠組み、detection-basedアプローチ、音声要約、種々の音声の変動にロバストな音声認識などの研究が重要であることを述べ、ニュース音声からの話題語（キーワード）抽出実験についても触れる。最後に、若い研究者に対して、国際的に存在感のある研究への期待を述べる。

キーワード　話し言葉、音声認識、音声理解、音声要約、意図駆動音声認識、ロバスト音声認識

Towards Spoken Language Understanding
— In Anticipation of Significant Research —

Sadaoki Furui

Tokyo Institute of Technology, Department of Computer Science

2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8852 Japan
Tel/Fax: +81-3-5734-3480, Email: furui@cs.titech.ac.jp

Abstract This paper argues the importance of research on spoken language understanding in advancing the recent remarkable progress in large-vocabulary continuous-speech recognition, and describes crucial research viewpoints. Specifically, it emphasizes the importance of the message-driven speech recognition framework, the detection-based approach, speech summarization, and robust speech recognition technology in solving spoken language specific problems, such as word omission, addition of extraneous words, inverted sentences, repairs and out-of-vocabulary words. This paper also describes experimental results on topic word (keyword) extraction from broadcast news speech. Finally the author encourages young Japanese researchers to accomplish significant research results and make their presence felt in the international community.

Keywords Spoken language, Speech recognition, Speech understanding, Speech summarization, Message-driven speech recognition, Robust speech recognition

1．まえがき

大語彙連続音声認識とそのシステムの研究開発が、最近大きく進展している[1]。この流れを着実に進め、その世界の技術競争の中で存在感を示していくことが求められている。そのためには、タイムリーな研究テーマとアプローチを選択することが必要である。

音声認識技術は、一人の人間がしゃべっている音声を対象とする場合と、二人の人間、あるいは一人の人間と擬人化されたコンピュータシステムとの対話音声を対象とする場合に分けられる。個々の発声単位（文）の認識を重視する立場と、対話の流れを重視する立場の違いということもできる。音声対話システムの意義と課題に関しては、最近の日本音響学会誌の特集号[2]を含めて、たびたび論じられているが、どこに重点をおいて研究をするべきかに関しては、いろいろな考え方がある。人間生活における音声の使われ方は、人と人とが対話する形が多いため、コンピュータによる音声認識でも、人間どうしの対話のような形でないと用途が広がらないと考える研究者もいるようだが、必ずしもそうとは言えないように思われる。その理由は次の通りである。

(1) 人間とコンピュータシステムとのインタフェースによって、情報検索、予約などを行う場合について考えると、ディスプレイ（タッチスクリーン）が使えるなら、その画面を入出力に用いる方が、音声を入出力に用いるよりも、能率がよい場合がほとんどであると考えられる。（音声より画面の方が能率的）

(2) 人間がコンピュータシステムと情報をやりとりする場合は、遊びは別として、効率を考えれば、人間相手と同じやり方を期待することはあまりないと考えられる。（機械は機械らしく）

(3) 上記 (2) の結果として、ユーザがどう応答あるいは質問すればよいかが明確なシステム主導の対話形式が、最も受け入れられやすいと考えられる。（システム主導）

(4) 上記 (2) の結果として、人間どうしの電話対話の中間にコンピュータシステムが挿入された音声翻訳システムのような場合でも、多くの利用者は、直接相手に話すというよりも、機械としてのコンピュータを意識した話し方になるであろう。（機械を意識した発声）

もちろん、研究が進んで、本当に人間相手と同じように動作するコンピュータシステムができれば、面白いに違いない。そのような擬人化されたコンピュータとの音声対話システムが使われる時代がいずれは来るであろう。さらに、音声対話は人間の思考のメカニズムに密接した重要な人間行動であるから、その研究は大いに進めるべきであろう。しかし、基本はあくまでも、発声者が「きちんと」発声したつもりの個々の話し言葉を、単独に正しく認識（理解）すること、すなわち発声者の意図を正しく理解することであろう。文脈情報の重要性や、対話制御の重要性を無視するわけではないが、研究の順序としては、対話としての側面を強調しすぎない方がよいのではないかと思う。ここでは、話者が頻繁に代わることはあっても、直接的には一人の人間がしゃべった音声を対象として認識（理解）する場合に限って、研究課題を考えていくことにしよう。

最後に、若い研究者へ、国際的に存在感のある研究への期待を述べたい。

2．話し言葉の音声理解

当面の音声認識（理解）研究の対象として、真に役に立つものが工学的に実現できるとすれば、どのようなものであろうか？　我々が現在利用することができる技術と科学を土台にしてさらに発展させ、真に使われるシステムを5年ないし10年で作るとすれば、次のような対象ではないかと思われる。

(1) 他人に聞かれる（聞いてもらう）ことを前提として、あるいは、コンピュータを意識して、ある程度丁寧に発声された文音声の認識（理解）システム。具体的には、放送音声の字幕化、講義や講演のディクテーション（自動書き起こし）、これらの要約など。

(2) 単語あるいはフレーズ（文節）程度に区切られた音声を、大語彙（例えば人名や地名）を対象として、誰が発声しても、雑音などがあっても極めて高い精度で認識するシステム。情報検索などで音声が使わ

れるとすれば、このような技術が最も重要であろう。

このいずれの場合も、言うまでもなく、次のような話し言葉特有の難しさがある。

(a) 助詞などの言葉の脱落・省略、間投詞などの不要語の付加、倒置への対処

(b) 言い間違い、言い淀み、言い直し、重複への対処

(c) 意味のある未知語の発声への対処

ディクテーションと言われるタスクの場合でも、これらの問題を避けて通ることはできない。逆に言えば、このような問題がないほどきちんと発声されるのは、そのような文書がすでに書かれていて読み上げている場合と考えられ、その場合には元々音声を認識する必要はないと考えられる。

これらの話し言葉特有の問題への対処法としては、

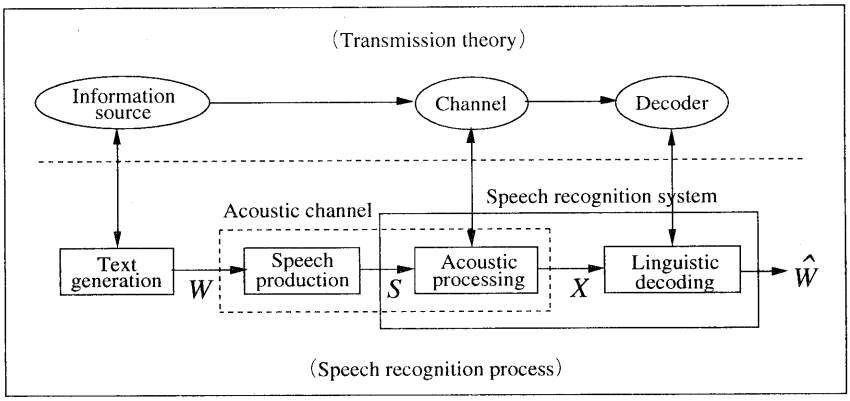
(i) 単語やフレーズを単位としたスポッティング (spotting) あるいは検出 (detection) と、発声確認 (utterance verification) のアプローチ

(ii) 余計な言葉をスキップしながら構文解析する方法

(iii) 未知語検出アルゴリズム

などが研究されている。ただし、現在の中心的方法である統計的言語モデルとマッチする効率的な方法はまだない。

今後の重要なアプリケーションの一つとして、上でも触れたが、音声の要約がある。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。音声メールを自動的に要約して文字化できれば、内容を一覧するために役に立つであろう。これらの要約技術は、音声理解技術と密接に関連している。



$$\hat{W} = \underset{W}{\operatorname{argmax}}\ P(W|X) = \underset{W}{\operatorname{argmax}}\ \frac{P(X|W)\,P(W)}{P(X)}$$

図1　これまでの音声認識の枠組み

3．意図駆動音声認識

これまでの音声認識の枠組みでは、音声生成のステップを図1のように定式化し、$P(W|X)$ (ここでWは単語列、Xは音響パラメータベクトル時系列) を最大化する規準が用いられる。読み上げ音声をディクテーションする目的の場合はこれでもよいかもしれないが、実際に人が音声を発声する場合は、まず相手 (それがコンピュータであったとしても) に伝えたいメッセージあるいは意図を頭の中で構成し、それを、何の言語か、語彙の種類、文法、意味論、文脈、発話習慣、などに従って単語の列に変換する。音声認識の究極の目的は、音声から、その発声者の意図を抽出することにある。従って、図2に示すようなモデルで考えることが必要である [3]。ここでMは、発声者が伝えたいメッセージ (意図) を表す。このモデルに基づいて考えると、音声認識のプロセスは、$P(M|X)$を最大化するMを選ぶことになる。我々は最近、このための新しい定式化を提案し、実験を進めている [4]。

この定式化は、これまでに試みられてきた種々の方法を包含し、かつ新しい方法を示唆する。すなわち、条件付き確率$P(W|M)$の表現法には、これまでに研究された話題別の言語モデル [5] [6]、cacheモデル [7] など種々の方法が考えられる。我々は、Mとして連続量を扱うことができ、しかもそれを明示的に表現する必要

がない方法として、Mは単語の共起によって表される[8]と考え、これによる定式化を行っている。さらにこの方法を放送ニュース音声に適用し、これによって結果的に単語列の認識精度が向上することを示している[4]。

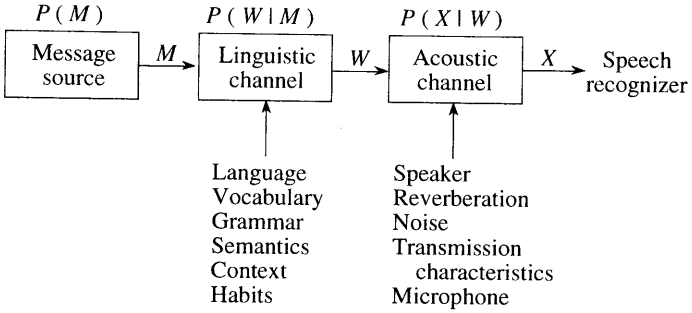


図2　意図駆動音声認識の枠組み

4．音声理解のためのDetection-Basedアプローチ

不要語などを多数含む話し言葉から、意味のある部分だけを抽出（意味抽出、意味検出）する方法として、単語およびフレーズの検出と発声確認を組み合わせたアプローチが研究されている[9][10]。不要語を吸収する音響モデルや、その単語やフレーズの受理を判定するための尺度を適切に構成することによって、よい性能を得ることができることが実験的に示されている。ただしこれまでの実験は、中規模程度の語彙数のシステムに対してであり、大語彙のシステムにどのように適用し、適度な処理量で、これまでの方法をどれだけ上回れるかは今後の課題である。

図3に示すように、flexibilityは少ないが効率のよいdictation (transcription) orientedのアプローチと、flexibilityは大きいが必ずしも効率の良くないdetection (spotting) basedのアプローチを、どのレベルで組み合わせる（融合させる）かがこれからの鍵となろう。この研究は、後述する音声認識結果からの話題語（キーワード）抽出とともに、音声要約のための一つのアプローチとなろう。

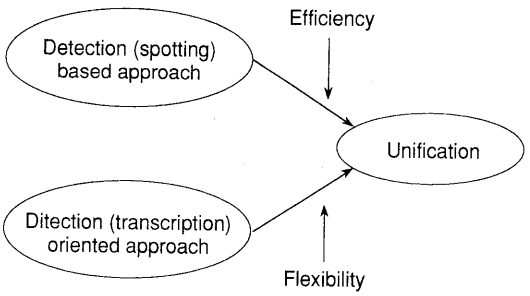


図3　話し言葉を認識（理解）する新しいアプローチ

5・ロバストな音声認識への取組み

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）があるためである。その原因には、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、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がある。図4に、音声の種々の変動要因を示す[11]。

これらの種々の音声の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ることができる。しかし、集められるデータ量には自ずから限界があり、すべての変動条件を尽くすことは不可能である。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る。このため、「Sheep and goats 現象」と言われるように、大多数の話者に対してはうまく動作しても、少数では

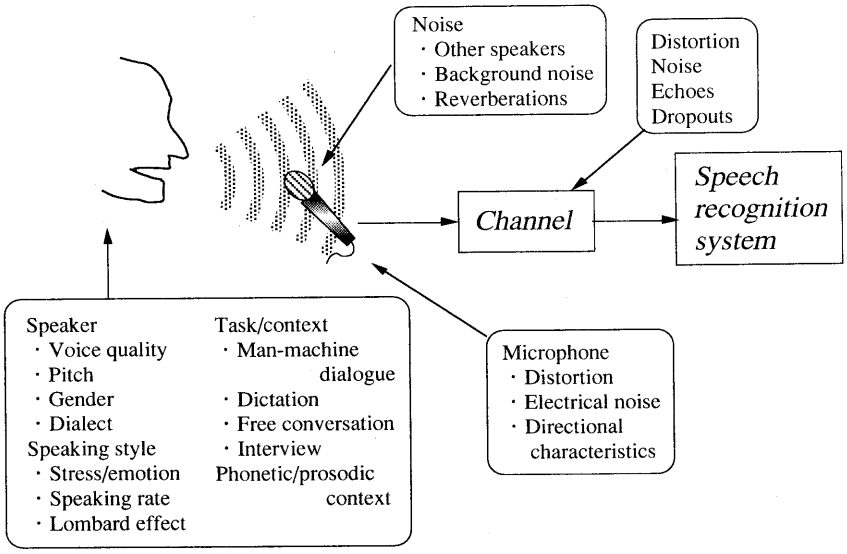


図4　音声の種々の変動要因

あるが認識精度が極めて悪い話者や条件を生ずることになることが多い。

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基づいて、変動に自動的に追随してくれるような、適応機能を持たせる研究が広く行なわれている。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」による方法がある。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。後者の場合は、認識すべき音声をそのまま使って、適応化（instantaneous adaptation）できるので、ユーザにとっては全く意識せずに、あたかもどんな声にも動作する装置であるように使えて便利であるが、極めて個人差や歪みの大きい音声に対してこの機能を実現するのは難しい。しかし、適応化が真に必要なのは、このように認識率が特に低い話者や条件に対してであり、このときに効果がない適応化法は意味がない。今後は、教師なしで動作するアルゴリズムを追及していくべきであろう。さらに加えて、認識装置を使っているうちに、自動的に適応化が進んでいくこと（on-line incremental adaptation）が望ましい。

音声は通常、スペクトルあるいは対数スペクトル領域のパラメータで表現されるが、上で述べた種々の音声の変動は、これらの領域での線形あるいは非線形の変化として表される。さらに、その変化が、元の音声パラメータに対して、加算的な場合と乗算的な場合がある。また、その影響が種々の音素に共通に現れるものと、音素に依存して異なるものとがある。雑音やマイクロホンなどによる歪みは音素に共通であるが、個人差は音素によって異なる。この違いによって、適応化技術も異なってくる。また、これらの種々の変動は、単独にではなく複合して音声に加わるのが普通である。従って、これらに同時に対応できる適応化法を構築することが必要である。

適応化法はさらに、音声の波形や特徴パラメータの領域で適用する方法と、音響モデルの領域で適用する方法がある。前者は発声内容に無関係に適用できるので容易であるが、音素に依存した適応化ができない限界がある。両者の特徴を生かして、うまく組み合わせることが必要であろう。

これまでの話者適応化では、不特定話者の音素モデルを話者に適応させるケースが多い。ある特定の話者のモデルを適応化させるよりは、この方がよい結果が得られるが、ぼけたモデルの適応化では自ずから性能に限界があると思われる。適応化前の初期モデルをどのように用意するかも、これからの重要な課題の一つとなるであろう。さらに基本的には、音声の個人差とは何なのかを少しでも明らかにし、そのモデルに基づいた適応化法を構築すべきであろう。いつまでも

MLLR(maximum likelihood linear regression) [12] でもあるまい。

6．ニュース音声の認識と話題語（キーワード）抽出

音声から発声者の意図を抽出したり、音声の内容を要約したりすることは極めて重要であり、そのためには、話題語（キーワード）が正しく抽出できることが必要である。3章の図2のモデルにおいて、発声者の意図Mが話題語の集合によって表されると考えることに相当するとも言える。我々は、ニュース音声を自動的にディクテーションし、その認識結果としての単語列から話題語を自動抽出する試みを行っている [13]。

この際、ニュースの話題語はほとんどが名詞であるので、ディクテーション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。すなわち、話題語を抽出すべきニュース音声（1ないし複数の文から構成される）の認識結果から、情報量の大きい単語を上位から順に抽出した。人がニュース文を見て抽出した話題語を正解として評価した結果、図5に示すような結果が得られた。図には参考のため、音声認識誤りがなかった場合、すなわち正しいテキストを入力した場合の結果も示してある。音声認識誤りのために、話題語の抽出精度（precision）が5～10%程度低下する。音声認識の結果から話題語を上位5個（recall：13.0%）抽出すると、84.1%のprecisionが得られる。図6に、ニュース音声の原音声、

原音声

阪神大震災で兵庫県淡路島の地表に現れた野島断層は地下ではどのぐらいの長さまで続いている活断層なのかこれまでよくわかっていませんでしたが日米の研究者の合同調査で全長２０キロにも達している可能性が高いことがわかりました。

え調査をした人はこの周辺は大きな被害を受けており活断層の位置を把握して建築物を建てる必要を改めて確認させる結果となったと話しています。

認識結果 (trigram)

阪神大震災で兵庫県淡路島の地表に現れた野島断層は地下ではどの問題の長さまで続いている活断層は七日頃までよくわかっていませんでしたが道での研究者の合同調査で全長二十キロにも達している貨物が高いことがわかりました

調査をした人は空港の周辺は大きな被害を受けており活断層の位置を把握して建築物をたてる必要を改めて確認させる結果となったと話しています

話題語抽出結果 (上位20位)

活断層　野島断層　調査　合同調査　阪神大震災　地表　頃　兵庫県淡路島
全長　長さ　した　把握　位置　地下　貨物　被害　建築　確認　必要　周辺

図6　ニュース音声の原音声、ディクテーション結果、およびそこから自動抽出した話題語の例

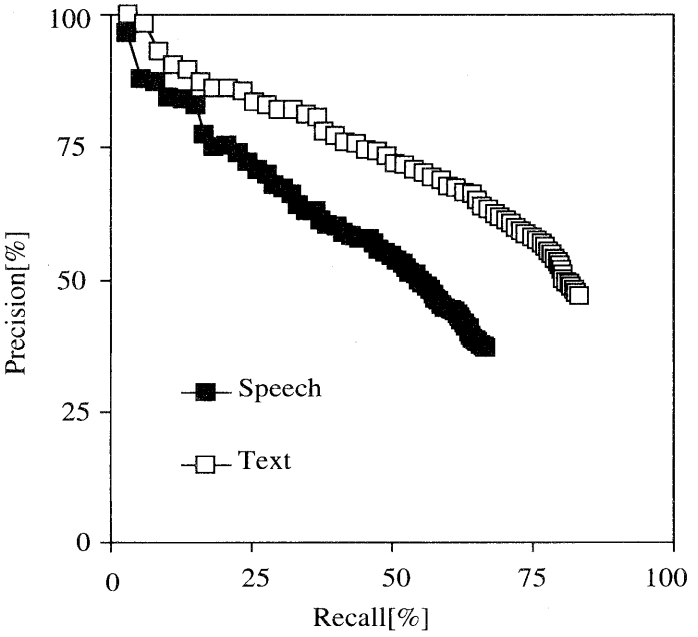


図5　ニュースからの話題語抽出の結果（recallとprecision）

ディクテーション結果、およびそこから自動抽出した話題語の例を示す。

今後、さらに単語の意味関係などを用いて、話題語の抽出精度を上げて行きたい。

7．国際的に存在感のある研究を

これまでの音声認識研究の歴史の中で、日本で提案・発明された技術が果たしてきた役割は大きい。しかし、現在、世界を股にかけて活躍している日本人は、その実力の割に少ない。工学的な意味での音声認識研究をリードしている米国には、中国人で活躍している研究者は多いが、日本人は極めて少ない。これは国民性にもよるものであるが、言葉の壁を克服して、世界の場で活躍する研究者が、今の日本の若い研究者や学生の中から沢山でてきてくれることを期待したい。そのためには、世界の研究の流れをきちんと把握し、価値ある方向を見極め、それを踏まえた上で、新しい方向を示し主張していくことが必要である。日本の中でしか通用しない研究をやっていても、存在感を示すことはできない。日本における学会や、その学会誌が、このような意味で役に立ってほしいと思う。

8．むすび

最近、音声認識が新聞などのマスコミに取り上げられ、一般的な関心が高まっている。過去40年間の音声認識研究の歴史の中で、このような一般的関心の高まりを迎えたことが何度かあったが、そのたびに技術の不十分さに直面して、関心が失われるということを繰り返してきた。今回の高まりのきっかけには、PC上でのディクテーションソフトの性能向上、米国における電話サービスでの実用化などがあるが、技術的に未完成な部分を依然として多く抱えることも事実であろう。現在の社会的関心を一時的なものとせず、着実に社会に定着させるためには、しっかりした戦略にもとづいた研究の展開が必要である。

本稿では触れなかったが、音声認識の今後の展望を長期的な視野から考えると、人間の脳における音声の知覚や生成のメカニズムの解明に挑戦する研究も、並行して進めることが不可欠である[14]。大きなブレークスルーを得るためには、工学的アプローチをいろいろ取り換えて有効性を評価する方法から、困難ではあっても、少しでも人間の音声情報処理の本質に近づく方向へのシフトが必要であろう。

本稿は、音声研究専門委員会からの依頼を受けて行う講演のために書いたものであるが、教科書ではないので、音声認識全般をカバーするものではない。私見をもとに、当面の音声認識研究の方向について考えていることを書かせていただいた。日本人の若手研究者が、これから世界で活躍するきっかけの一つとなれば幸いである。

謝辞

本年８月の１カ月のベル研究所滞在中に、いろいろ考えるきっかけを与えて下さり、種々の討論に応じて下さったB.-H. Juang博士に感謝する。日頃討論や実験に協力して下さるNTT研究所の大附克年氏をはじめ、研究をサポートして下さるNHK研究所およびNTT研究所の方々に感謝する。

参考文献

[1] 古井貞熙：大語彙連続音声認識の現状と展望、春季音学講論、1-6-10 (1998)

[2] 小特集—音声対話システムの実力と課題—、音学誌、54、11、pp. 783-822 (1998)

[3] B.-H. Juang: Automatic speech recognition: Problems, progress & prospects, IEEE Workshop on Neural Networks for Signal Processing (1996)

[4] 大附、古井、桜井、岩崎、張：ニュース音声認識のための言語モデルと音響モデルの検討、信学技法、SP98- (1998)

[5] R. Pierraccini, et al.: A speech understanding system based on statistical representation of semantics, Proc. ICASSP 92, pp. I-193-196 (1992)

[6] S. Miller, et al.: Statistical language processing using hidden understanding models, Proc. DARPA Human Language Technology Workshop, pp. 278-282 (1994)

[7] R. Kuhn and R. De Mori: A cache-based natural language model for speech recognition, IEEE Trans. PAMI-12, 6, pp. 570-583 (1990)

[8] Z. S. Harris: Co-occurrence and transformation in linguistic structure, Language, 33, pp. 283-340 (1957)

[9] T. Kawahara, et al.: Combining key-phrase detection and subword-based verification for flexible speech understanding, Proc. ICASSP 97, pp. 1159-1162 (1997)

[10] B.-H. Juang: From speech recognition to understanding: Shifting paradigm to achieve natural human-machine communication, Proc. 16th ICA and 135th Meeting ASA, pp. 617-618 (1998)

[11] 古井貞熙：音声情報処理、森北出版 (1998)

[12] C. J. Leggetter, et al.: "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language, 9, pp. 171-185 (1995)

[13] 高木、桜井、岩崎、古井：ニュース音声を対象とした言語モデルと話題抽出の検討、信学技法、SP98-33 (1998)

[14] S. Furui: Future Directions in Speech Information Processing, Proc. 16th ICA and 135th Meeting ASA, pp. 1-4 (1998)