論文 / 著書情報
Article / Book Information

論題(和文)	音声認識技術の実用化への取り組み
Title(English)	
著者(和文)	古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	情報処理, Vol. 51, No. 11, pp. 1387-1393
Citation(English)	, Vol. 51, No. 11, pp. 1387-1393
発行日 / Pub. date	2010, 11
権利情報 / Copyright	

音声認識技術の実用化への取り組み

古井 貞熙
東京工業大学 大学院情報理工学研究科

音声認識技術が，ユーザフレンドリーなインタフェースとして，また音声を文字化して記録や検索の目的に用いる手段として，種々のシステムで使われるようになってきた．今後，用途をさらに拡大し，実用化を大きく進めるためには，話者による声の違い，周囲の雑音などに対する頑健性の向上，インタフェースとしての透明性の向上，新たなアプリケーションを開発する際の開発者からの手離れを良くする技術の向上などが必要である．将来的に人間並みの音声認識を実現するためには，統計的枠組みの中で，多様であいまいな知識を適切に組み合わせて用いる方法の構築など，解決しなければならない基本的研究課題が存在している．

音声認識技術実用化を目指して

音声認識技術は，多様な機能に対し簡易なインタフェースを実現する手段として，あるいは音声のテキスト化によって便利さを増し，マルチメディアコンテンツの検索を可能にする手段として，発展が期待されている．これまでの50年以上にわたる研究開発の成果として，利用場面あるいは利用者を限った場合には，高い精度で認識することができ，種々の応用システムが利用されている．これまでの技術開発において，音声の時間軸を非線形に伸縮してマッチングする方法への動的計画法の適用，音声のケプストラムの動的特徴の利用，種々のモデル適応アルゴリズムの提案など，日本の研究開発者の果たしてきた役割はきわめて大きい．これまでに我が国の研究開発者が生み出してきたこれらの基本技術や概念は，国際的に広く認知されている．

しかし，現在の技術では，静かな環境で，想定した範囲の声の話者が，想定した範囲の方法で利用する場合には十分な性能を与えるものの，それらの条件が異なる場合には著しく性能が低下することがある．誤認識を生じた場合に，その原因がユーザに分からず，使いこなすのが難しいという問題もある．そもそも使い方がよく分からない，という場合もある．また，対象とするアプリケーションや言語などを変更しようとする際の，開発者からの手離れの悪さも，実用化の足かせとなっている．これらの問題の解決のためには，実環境での利用に耐える高精度な音声認識技術の開発に加え，ユーザと開発者の連携によって想定と実際のギャップを埋める技術，音声インタフェースの手離れを良くするための技術などの開発が重要である．この観点から，音声認識実用化技術を対象とした初めての国プロとして，2006年度からの3年間（先立って行われた先導研究の期間を含めると4年間）にわたって，経済産業省のプロジェクト「音声認識基盤技術の開発」が実施され，種々の成果が得られた[1]．本特集では，このプロジェクトの成果報告を含み，実用化を目指した最近の研究開発の状況と今後の展望が述べられている．

音声認識の基本技術

●基本原理

音声認識の基本技術に関しては，2004年の情報処理学会誌で特集「音声情報処理技術の最先端」（編集：古井貞熙，田中穂積）[2]が組まれた．その「編集にあたって」で，音声情報処理，主に音声認識技術の動向についてまとめて解説した．その後5年以上が経過したが，基本的な部分はほとんど変わっていないので，興味のある方はご参照いただきたい．なお，最近の研究動向に関しては，日本音響学会誌の小特集「自動音声認識研究の動向と展望」[3]に詳しく紹介されている．

音声認識の原理を，図-1に示す．音響処理には，

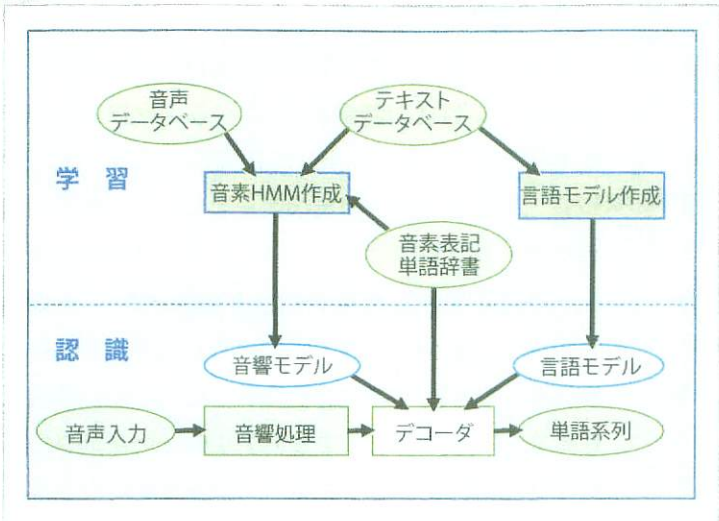

図-1　音声認識の原理[4]

音声波からのスペクトル特徴抽出，音声区間の検出などが含まれる．音声認識に用いられる基本的なモデルに，音響モデルと言語モデルがあり，これらのモデルをいかに適切に学習し，さらにアプリケーション，話者，雑音などの変化に適切に適応させるかが，認識性能を左右する．現在のほとんどのシステムは，統計的パターン認識の理論に基づいて作られており，音響モデルにはHMM（Hidden Markov Model；隠れマルコフモデル），言語モデルにはN-gramが使われている．デコーダは，音声認識のエンジンで，音声の特徴パラメータとモデルを用いて，音声入力に対し，最も可能性の高い文を単語(日本語の場合は形態素）系列として出力する．人とコンピュータが対話するシステムでは，音声認識結果から対話の進行状況を判断し，音声合成などの出力手段を制御してユーザに情報を出力し，音声による円滑な対話を実現する対話処理技術が，重要な役割を果たす[4]．

近年，モデルの構築法やデコーダの機能が高度化しており，個々の研究開発者が，そのすべてを自らプログラミングし，メンテナンスするのはきわめて困難になっている．また，音声と書き起こしテキストのデータベース(コーパス)を，音声の多様な変化がカバーされるように，いかに大量に用意し，効率的に使うことができるかが，音声認識の性能を決めるが，個々の研究開発者が大規模なコーパスを構築するのは難しい．さらに，各種の手法を比較・評価するためには，共通に使えるコーパスが必要である．そのような背景をもとに，近年，誰でも入手して使うことができるソフトウェアツールやデータベースが，多数作成されており，それが世界的なレベルでの技術の発展に大きく貢献している．

●ソフトウェアツール

ツールキット[5]としては，ケンブリッジ大学（英国）で開発されたHTK（Hidden Markov Model ToolKit）が最も広く使われている．種々の最新の音響処理や音響モデルの構築法をカバーしており，最近のバージョンにはデコーダも含まれている．言語モデルの構築ツールとしては，CMU（米国のカーネギーメロン大学）とケンブリッジ大学で作られたツールキットが1990年代から広く使われてきたが，1999年に開発が終了してしまったため，やや時代遅れとなった．その後，SRI（米国のスタンフォード研究所）でツールキットが開発された．多様な機能を持っていて，よくメンテナンスされているため，デファクトスタンダードとなっている．HTKにも言語モデル作成ツールが含まれているが，広くは使われていない．やや特殊な言語モデルツールキットとしては，MIT（米国マサチューセッツ工科大学）で開発されたMITLMや，マイクロソフト（米国）で開発されたMSRLMがある．

デコーダに関しては，従来型のデコーダで，よくメンテナンスされ，広く使われているものに，京都大学で作られたJuliusと，CMUで作られたSphinx 3およびSphinx 4がある．最近では，AT&T研究所（米国）で最初に提案されたWFST（Weighted Finite State Transducer）を用いたデコーダが，種々のアルゴリズムを容易に取り込むことができる数学的に簡明な枠組み，フレキシビリティ，および高い性能のために，世界的に主流になりつつある．AT&Tのデコーダと関連ツールは多くの研究者に使われてきたが，用途に制限があり，ソ

ースコードは公開されていない．Google（米国）のツールキットOpenFSTは，AT&Tから移った研究者によって作られ，ソースコードも公開されているので，広く使われている．MITのツールキットも，広く使われている．これに対して，上記の「音声認識基盤技術の開発」プロジェクトにおいて，我が国の音声認識実用化を含む研究開発で使われることを目的に，WFSTに基づく高性能の「T^3デコーダ（T-cubed Decoder）」が開発された．このデコーダは，次節で紹介するALAGINフォーラムから，広く公開される予定になっている．技術の詳細は，本特集の解説の1つで紹介されている．

●データベース

国内外における音声・言語データベース（コーパス）に関する活動の動向については，電子情報通信学会誌の解説[6]がある．米国のDARPAプロジェクトで作られた多数かつ大規模のコーパスが，言語データコンソーシアム（LDC）から公開されている．カバーされている主な言語は，英語，中国語，アラビア語である．ヨーロッパではヨーロッパ言語資源協会（ELRA）が，日本では，言語資源協会（GSK），音声資源コンソーシアム（NII-SRC），および高度言語情報融合（ALAGIN）フォーラムがコーパスの保存・利用を促進するための活動を進めている．コーパスの中には研究目的の利用に限られているものもあるが，個人情報が流出しない方法での商用利用を認めているものもある．企業で作られたものは，公開されない場合が多い．

我が国では，これまでに種々のプロジェクトや学会，研究グループによって，多様なコーパスが作成され，利用されているが，話し言葉音声のコーパスは少ない．その中で最も有用なコーパスは，日本語話し言葉コーパス（CSJ）[7]である．1999年から5年間行われた「話し言葉工学」プロジェクトで構築されたもので，講演音声を主たる対象としているが，一般の人による模擬講演，インタビューなども含み，700時間，700万形態素に及ぶ話し言葉音声が，きわめて精密にデータベース化されており，研究目的のみならず商用にも用いることができる．CSJは，多様な応用分野の音声認識のモデル作成に共通に用いることができる[8]．CSJはきわめて多様な研究開発に用いられており，CSJを用いた研究論文がこれまでに1000件近く発表されている．

音声認識の応用分野

●分類

音声認識の主たる応用分野を表-1にまとめて示す．これらは，目的によって，

- 使いやすいインタフェースを実現しようとするもの
- 音声の文字化を実現しようとするもの

に分類できる．表中，基本的に，コマンド制御から教育までが前者に，口述筆記（ディクテーション）から自動翻訳までが後者に対応する．また，実装の形態によって

- サーバ型（多数の組込み機器（携帯電話，カーナビゲーションなど）で収録された音声をネットワーク上のサーバに送り，サーバ上の音声認識装置で認識する）
- 組込み型（組込み機器で収録した音声を，組込み機器上の音声認識ソフトウェアで認識する）

に分類でき，サーバ型はさらに

- オンライン型（発声された音声を即座に認識する）
- オフライン型（発声された音声をいったん保存し，後で認識する）

に分類できる．

●インタフェース応用

インタフェースとして，車の中では目や手が運転操作に使われていて，キー入力やスクリーンをタッチする入力が困難なため，音声認識入力を備えたカーナビが広く普及している．ただし，実際に使われている割合はまだ少ない．

コールセンタでの音声認識の利用は，ユーザからの問合せや予約などのオペレータ業務を自動化し，

産業分野	用途・市場
コマンド制御	組込み機器（カーナビゲーション，情報家電，ロボットなど）への音声による指示．手や目が塞がった状況（ハンズビジー，アイズビジー）で，機器を制御するニーズから使われる．
データ入力	業務系機器（PDA など）への音声によるデータ入力．
介護／福祉	介護／福祉機器への音声による指示．ハンズビジーな状況や高齢者，身体障碍者への支援などに利用．
コールセンタ	ユーザからの問合せや予約などのオペレータ業務を自動化し，人件費を削減する．
音声ポータル	音声認識，音声合成を利用したインターネットコンテンツ（ニュース，天気，スポーツ，株価など）アクセスサービス．
音声ブラウザ	音声認識，音声合成を利用した音声によるインターネットコンテンツアクセス機能を有するマルチメディアブラウザ．
教育	語学教育における発音チェックや e-Learning における音声利用．
口述筆記（ディクテーション）	音声からテキストへの自動変換．PC ソフトが製品化．医療分野での電子カルテの作成や，スマートフォンやカーナビゲーションでのメール作成などに利用．
書き起こし	講演音声，会議音声などのテキスト書き起こし．会議録などの作成．
放送	聴覚障碍者のためのニュースなどのクローズドキャプション．
索引付け	TV プログラム，ビデオカメラ，IC レコーダの音声部を利用した索引付けによる検索自動化．コールセンタへの問合せ音声の自動分類．
自動翻訳	会話を認識し，他の言語に翻訳し，テキスト表示または音声合成で出力．

表-1　音声認識の主たる応用分野（文献 9）中の表に手を加えた）

人件費を削減する目的で開発された．オペレータや顧客の発話内容をテキストに変換することにより，サービスの向上，問題点の発見と解決，コンプライアンスの強化などに用いている．応対の所要時間短縮，コスト削減に貢献している．米国では，人間による電話受付の人件費 5.5 ドルを 10 分の 1 以下に削減する効果が確認されている[10]．

世界で最初の電話音声を用いた音声認識システムは，1981 年にバンキングサービスのために NTT が開発した，ANSER（Automatic answer Network System for Electrical Request；アンサー）システムであった．当時の技術レベルとして，16 種類の単語発声音声(数字とコマンド)しか認識できなかったが，日本全国の銀行で使われた．

使いやすいインタフェースとして最近注目されているものに，Google の音声検索などがある．このシステムでは，iPhone や Android スマートフォンを用いて音声で Web 検索ができ，検索結果は通常の場合と同様に，画面に表示される．日々新しい語彙の学習，音響モデルの適応化が行われているため，高い検索性能が実現されている．

音声翻訳は，音声認識・機械翻訳・音声合成の組合せによって実現され，特に日本で，1980 年代から永年にわたり精力的に研究が行われている．旅行会話のように限定された領域では実用レベルに近づいているが，依然として難しい研究課題である．

●音声の文字化応用

電子カルテの入力や，調剤薬局での服薬指導の記録など，医療現場での種々の文書作成に音声認識が広く使われている[11), 12)]．複雑な専門用語をキーボードで入力するのは難しいが，辞書に登録さえしておけば，音声で容易に入力できる．放射線画像診断や病理診断では，画像や顕微鏡から目を離さずに入力できるメリットがある．

国会，地方議会，株主総会，セミナなどの議事録作成，裁判所の記録作成などへの，音声認識の利用も進んでいる．会議等の議事録作成では，音声認識誤りの修正・編集ツールが重要であるが，修正方法を含む技術進歩により，記録作成のコスト削減や迅速化に貢献している．裁判員裁判向け音声認識システムでは，公判の音声・映像を記録すると同時に音声認識を行い，音声認識結果をその音声の時刻情報とともに記録しておく．評議の場面でキーワードを入力して自動的に検索を行い，音声・映像の頭出し再生を行う．裁判員裁判では，一般の裁判員を長期間にわたって拘束しないようにするため，短期間で次々に評議を行う．人手で公判記録を作成している

時間がないため，音声認識結果を用いて，必要な個所の音声・映像の検索・再生ができるようになっている．検索目的であるので，必ずしも認識精度が高くなくても，用いることができる．

NHK や BBC（英国）によるニュースの字幕（クローズドキャプション）作成にも，音声認識技術が使われており，聴覚障碍者や高齢者に対する有用な情報保障手段となっている[13]．NHK 総合（デジタル）では，総放送時間の約半分に字幕がついている．アナウンサーが原稿を読み上げている音声に対しては，音声認識で完璧に近い文字化ができるが，スポーツ中継など音声認識が困難な音声に対しては，リスピーク方式が使われている．これは，字幕専用アナウンサーが，ヘッドホンで番組音声を聞きながら，音声認識しやすいように言い換え，要約し，状況説明を加えて復唱するものである．直接音声認識する方式と組み合わせて用いるハイブリッド方式も使われている．

リスピーク方式は，CapTel という電話対話の自動キャプションシステム（米国）でも使われている[12]．話すことはできるが聴覚に障碍のある人が電話サービスを使うためのもので，訓練を受けたオペレータが，対話の相手の言葉を復唱して音声認識装置に入力して，音声を文字に変換し，特殊な電話についている小型ディスプレイに表示して，電話をかけた人に伝えるシステムである．音声認識誤りはオペレータによって修正されるが，対話をするのに十分な速度が確保できている．WebCapTel システムでは，電話をかける人は普通の電話を用い，テキスト化された相手の音声は，コンピュータディスプレイに表示される．

音声認識の実用化を促進するための今後の課題

●音声認識の普及を妨げている要因

これまで述べたように，技術的進歩やコンピュータの進歩に支えられて，多様な音声認識応用システムが開発され，使われている．特に米国では，広い国土に人が分散して住んでいて，多くのサービスが電話を通じて行われているため，莫大なオペレータのコストを削減する目的で，音声認識が種々のアプリケーションで利用されている．一般の人のほとんどが，何らかのサービスで音声認識を使った経験を有している．ソフトウェアツールやデータベースなどの開発環境も整備されつつある．しかし，まだ広く使われていると言える状況ではない．

日本では，さらに実際の利用は限られており，多数の音声認識技術を生み出してきたにもかかわらず，多くの人の期待ほどには使われていないというのが実態である．「なぜ音声認識は使われないか」という話題は，20 年前から何度も議論されてきた．その主たる原因は，冒頭でも述べたように，技術の頑健性の不足，インタフェースとしての不完全性である．米国では，サービスとして他に選択肢がない上，不完全でも使えれば使ってみるという風土があるが，日本の場合は，技術の完成度への要求が高い一方，人前で電話で話すことへの心理的抵抗があり，うまく使えなかったら恥ずかしいという意識も普及を妨げている．

●音声認識したい音声とそれ以外の音の分離

技術の頑健性を向上させるために解決しなければならない課題の 1 つは，認識対象としての音声とそれ以外の音との分離である．この両者は同時に重なっていることもあるし，時間的につながっていることもある．認識対象以外の音には，周囲のいわゆる騒音や雑音もあるが，他の人の音声であることもある．さらに，家庭内での家電製品の制御などの場合では，音声認識を使うたびにスイッチをオンオフするのは煩わしいので，常にオンしておくとすると，同じ人の音声でも，音声認識させる目的でしゃべった音声と，他の人との雑談など認識対象以外の音声を区別する必要が生ずる．

同時に重なっている音を分離する方法としては，人が両耳で音の方向を検知して音を区別するように，複数のマイクロホン（マイクロホンアレイ）を用いて指向性を制御し分離する方法や，雑音のスペクトル

を引き算する方法，重なっている音の時間・スペクトル的特徴の違いを用いて分離する方法などがある．効果を発揮させるためには，音が来る方向が大きく変化しないとか，雑音のスペクトルが大きく変化しないとか，複数のマイクロホンをある程度離して配置できるなどの条件が必要である場合が多い．音として分離することが困難な場合は，マイクロホンに入ってくる音をすべて音声認識してみて，アプリケーションに合った認識結果が得られた場合だけ動作させるようにするという方法も研究されている．ただし，後述するように，「合っているかどうか」という判断は難しい場合が多い．

●話し言葉音声のモデル学習

書き言葉(を読み上げた)音声と話し言葉音声とは，音響的にも言語的にも大きく異なるので，あらかじめ用意した原稿を読み上げた音声コーパスから作成したモデルは，音声認識にはあまり役に立たない．一方，自由に発声した話し言葉音声を録音して，人手で書き起こしてコーパス化するのはコストがかかりすぎる．したがって，話し言葉音声コーパスの効率的構築法と，書き言葉に整形された不完全な書き起こししかない音声コーパスをモデルの学習に効果的に利用する方法の開拓が重要である[3), 10), 14)]．

●音響および言語モデルの適応化

音声認識のための音響モデルを，いかにして発声者，周囲の雑音，発声スタイルなどによる変化に対応して適応させるかも重要な課題である．言語モデルを作成するためには，音響モデルのためのコーパスよりも格段に大きなデータベースが必要なので，アプリケーションが変わるたびに，データベースを集めるのは難しい．このため，一般的な言語モデルを，新たなアプリケーションに効果的に適応させる方法を開拓する必要がある．

●未知語への対応

音声認識で最も深刻な問題の１つは，未知語である．これはパターン認識全般，さらに人工知能全般に言えることであるが，入力音声が，認識装置の辞書にある単語かそうでないかを正しく判定することができない．本質的に辞書にない言葉，つまり未知語は音声認識できないが，入力音声は常に何らかの変動を伴っているので，辞書にある言葉の音声でも，蓄えられているテンプレートやモデルにぴったり合うことはない．したがって，未知語を含めて，どのような音声でも，言語モデルによる制約を考慮した上で，辞書にある最も近い単語と判定される．その近さに関して，事後確率のような形式で，認識結果の信頼度を計算することはできるが，その信頼性が必ずしも高くない．このため，できるだけ未知語を減らして誤認識を防ぐために，Web を検索したりして，新しい語彙を辞書に追加することが行われるが，完全に網羅することは難しい．アプリケーションによっては，認識できる語彙や言い回しをある範囲に限定しておくことも可能であるが，それをユーザにいかに伝えるかが難しい．種類が少なければ画面に表示できるが，それならそれをタッチしたりして選べばよく，音声認識はいらない．種類が多い場合は表示が難しい．

●分かりやすいインタフェース

実用化を進めるためには，未知語の課題を含めて，分かりやすいインタフェースをいかに作るかが，音声認識技術そのものと同じくらい重要である．このためには，ユーザから見たインタフェースの透過性の実現が重要である．たとえば，音声認識誤りを生じたときに，なぜ認識できなかったかがユーザに分かるシステムであれば，ユーザが躊躇なく使い，ユーザも認識しやすいように発声できるようになる．また，認識誤りが生じても，それをすばやく容易に修正することができれば，キーボードを使うときにはタイプミスを恐れないように，認識誤りを恐れる必要はなくなる．音声認識の利用に関する習熟度に応じて，ショートカットのように，インタフェースを変える仕組みも必要である．

また，最近のスマートフォンなどでは，画面やカメラを用いたインタフェースが容易に構築できるの

で，音声単独ではなく，マルチメディア・マルチモーダル環境で，画像・映像・テキストなどの情報と組み合わせて音声認識技術を利用する方法に関しても，研究開発を進める必要がある．

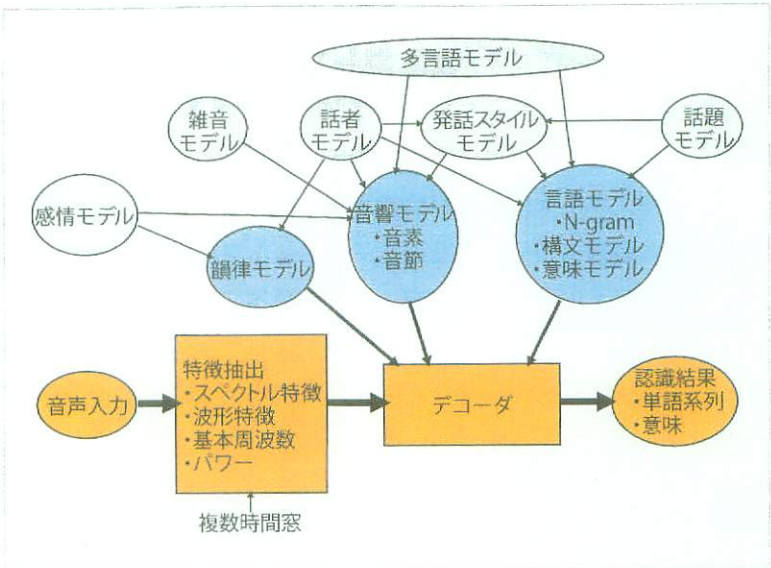

図-2　多様な知識を総合的に用いる音声認識

基礎研究と実用化研究

音声認識に関する国内および国際会議は毎年多数開かれており，学術誌による論文の発表も減ることはない．人間並みの音声認識が実現できるまでには，まだ10～20年はかかると思われ[4]，そのためには，現在のような比較的単純な音響モデルと言語モデルだけではなく，図-2に示すように，人間が用いているさらに高度で多様な知識，たとえば，話者の個人差，感情，発話スタイル，方言，周囲の雑音の影響，話題の流れなどによる音声の音響的・言語的変動，言葉の意味関係，文節の係り受け関係などがモデル化できることが必要である．それらを，状況に応じて適切に組み合わせて用いる枠組みと，それに基づくデコーダを研究開発することが求められている[14]．

現在学会などで発表される研究の中身は，長期的な視野からの基礎研究でもなく，実用化に必要な課題ともかけ離れているものが少なくない．実用化の促進のためには何がまず必要なのかを強く意識して，研究開発ターゲットを明確にし，課題を解決していく必要がある．上で述べたように，近年種々の課題が明らかになるとともに，大量のデータベースや便利なツールが整備されつつあるので，着実な課題解決により，役に立つ，使いやすい音声認識システムが，これから続々登場することを期待したい．

参考文献
1) 古井貞熙，小林哲則，矢頭　隆，大淵康成，河村聡典，三木清一，庄境　誠：総合報告：音声認識技術の展開，電子情報通信学会誌，Vol.93, No.8, pp.725-740 (2010).
2) 古井貞熙，田中穂積(編集)：特集「音声情報処理技術の最先端」，情報処理，Vol.45, No.10, pp.1001-1049 (Oct. 2004).
3) 大川茂樹，伊勢友彦(編集)：小特集「自動音声認識研究の動向と展望」，日本音響学会誌，Vol.66, No.1, pp.12-40 (2010).
4) 古井貞熙：人と対話するコンピュータを創っています—音声認識の最前線—，角川学芸出版，東京 (2009).
5) Nguyen, P.：TechWare：Speech Recognition Software and Resources on the Web, IEEE Signal Processing Magazine, pp.102-105 (May 2009).
6) 板橋秀一，山川仁子，大須賀智子：音声言語コーパスの現状と課題，電子情報通信学会誌，Vol.92, No.8, pp.676-681 (2009).
7) 前川喜久雄：『日本語話し言葉コーパス』公開版の仕様，第3回話し言葉の科学と工学ワークショップ講演予稿集，pp.7-14 (2004).
8) 西井俊介，篠崎隆宏，古井貞熙：日本語話し言葉コーパスを用いた異なるタスクに対する音声認識，日本音響学会春季研究発表会，2-7-1 (2010).
9) 平成14年度 特許出願技術動向調査報告書—音声認識技術—，特許庁総務部技術調査課 (2003).
10) Wilpon, J., Gilbert M. E. and Cohen, J.：The Business of Speech Technology, Springer Handbook of Speech Processing (Benesty, Sondhi and Huang, Eds.), Springer, pp.681-703 (2008).
11) 藤田泰彦：音声認識 実用化事例の紹介とその課題，情報処理学会研究報告，2009-SLP-78, Vol.6 (2009).
12) Zhao, Y.：Speech-recognition Technology in Health Care and Special-needs Assistance, IEEE Signal Processing Magazine, pp.87-90 (May 2009).
13) 今井　亨：リアルタイム字幕放送のための音声認識，電子情報通信学会技術研究報告，SP2009-52 (2009).
14) 古井貞熙：招待講演：何かが欠けている音声認識研究，信学技報，SP2009-80 (2009).

(平成22年7月13日受付)

古井 貞熙（正会員）furui@cs.titech.ac.jp
1970年東京大学大学院計数工学専攻修士課程修了，工博．NTTヒューマンインタフェース研究所音声情報研究部長，古井特別研究室長などを経て，1997年より東京工業大学教授．附属図書館長．音声情報処理の研究に従事．紫綬褒章，文部科学大臣表彰など．