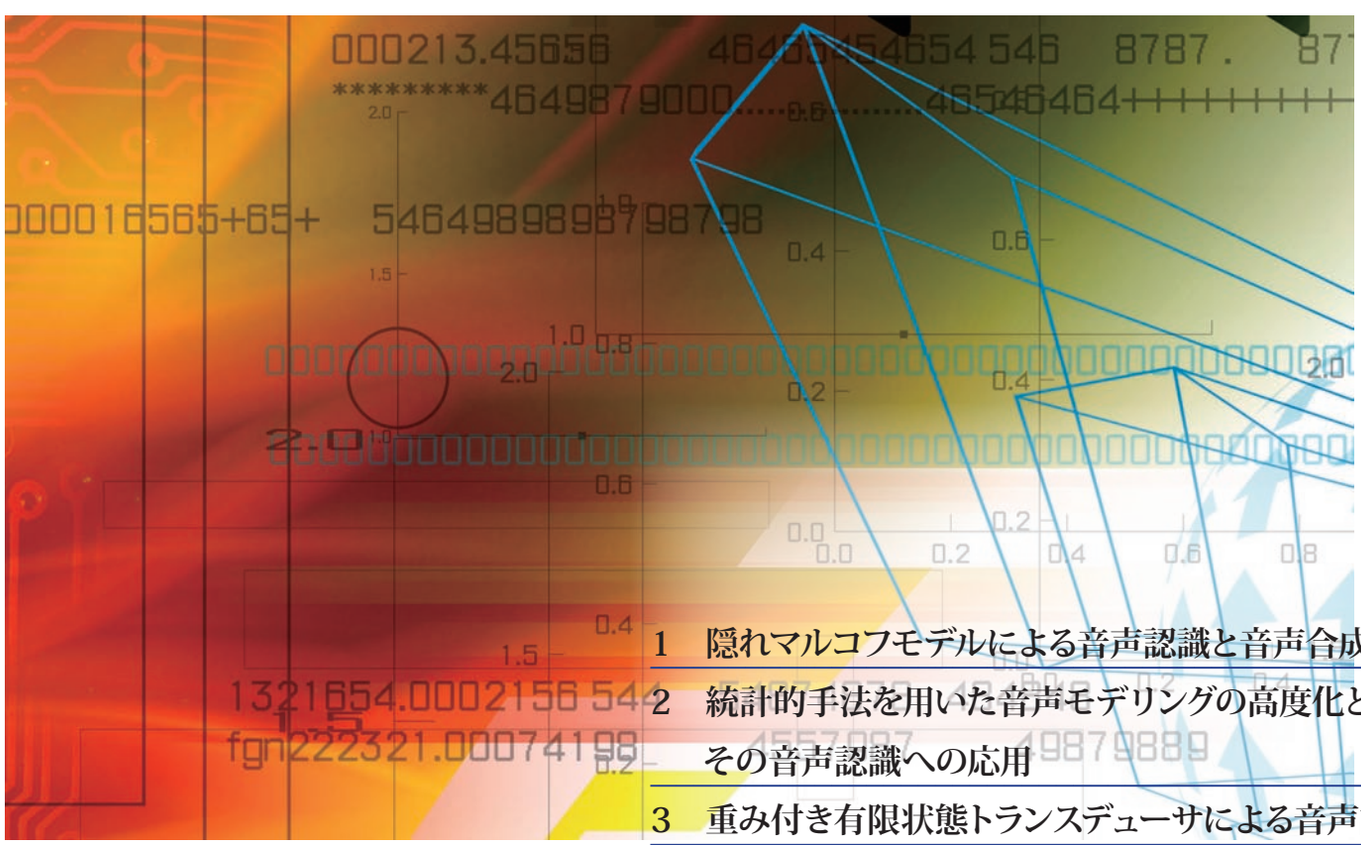
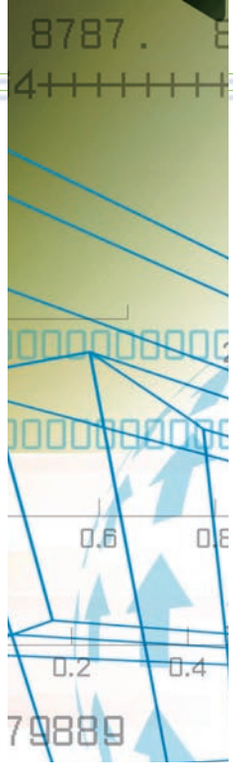
特集

音声情報処理技術の最先端





特集 音声情報処理技術の最先端

編集にあたって

古井 貞熙
東京工業大学大学院情報理工学研究科
furui@cs.titech.ac.jp

田中 穂積
東京工業大学大学院情報理工学研究科
tanaka@cs.titech.ac.jp

音声情報処理技術は，この10～20年の間に大きな進歩を遂げ，種々の実用システムが使われるようになってきた．20年前の，いわば手作りの音声認識・音声合成システムに比べ，現在は大規模なデータベース（コーパス）と統計的理論に基づいたシステムが主流となっている．20年前の音声認識技術では，十数種類の単語を認識するシステムの実用化がやっとであったが，最近では，丁寧に発声された音声であれば，数万単語の語彙を対象とした連続音声認識で90％以上の精度が得られるようになっている．

現在の大語彙連続音声認識システムの典型的な構成を図-1に示す．音声波形は，まず10ms程度の細かい時間ごとに，ケプストラム（対数スペクトルのフーリエ逆変換）に変換され，さらにその動的特徴であるデルタケプストラムと合わせて，特徴ベクトルが構成される．特徴ベクトルの時系列Xに対して，音響モデルと言語モデルを用いたデコーダによって，事後確率$P(W|X)$が最大となる単語列Wを選ぶのが音声認識の過程である．

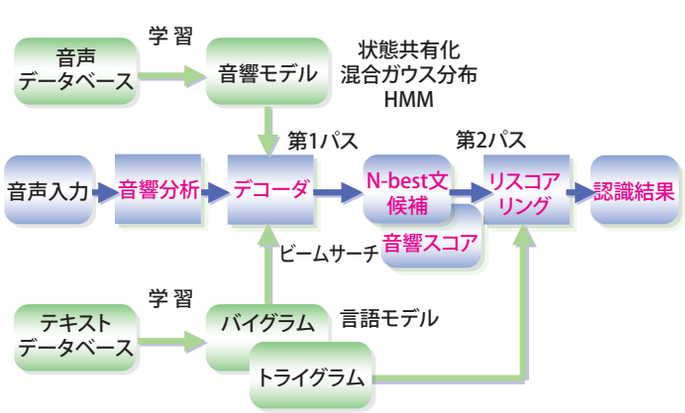


図-1　現在の典型的な大語彙連続音声認識システムの構成

事後確率を直接的に最大化することは困難なので，ベイズの定理によって，音響モデルから計算される尤度$P(X|W)$と，言語モデルから計算される言語確率$P(W)$の積に変換し，その積を最大化する．音響モデルとしては，音素を単位とし，前後の音素の影響，個人差，時間的伸び縮みなどを考慮した統計的モデルであるHMM（Hidden Markov Model；隠れマルコフモデル）が用いられる．言語モデルとしては，単語のバイグラム（2つ組確率）およびトライグラム（3つ組確率）に代表される統計的言語モデルが用いられる．音響および言語モデルは，音声コーパスおよびその書き起こしであるテキストコーパスを用いた学習によって作成されるが，モデルの推定精度を上げるため，平滑化をするなど種々の工夫がされている．

音声認識（デコーディング）の過程では，莫大な数の可能な文（単語連鎖）仮説の中から，事後確率最大の仮説を効率よく探索するため，動的計画法に基づいた処理が行われる．いきなりトライグラムを用いると処理量が莫大になってしまうので，図にあるように，バイグラムとトライグラムを用いた処理を2段階に分けて行うのが普通である．これらの認識処理を可能とした背景には，コンピュータやハードウェア技術の進歩，デコーダなどのソフトウェア技術の進歩などがあり，これらの基本技術は，学習コーパスさえあれば，どのような言語の音声にも適用できる．バイグラムやトライグラムでは，近接した単語の連鎖確率しか考慮できないので，文脈自由文法などを組み合わせる方法も検討されている．

音声認識のタスクは，表-1に示すように，対象音声に関する2つの規準，すなわち，(1) 人がコンピュータに対して発声している音声か，人に対して発声している音声か，(2) 対話か独話かによって，4つのカテゴリー

	対話	独話
人対人	(カテゴリー I) 電話による対話音声の文字化 インタビューの文字化 会議録作成	(カテゴリー II) 放送ニュースの字幕化 講演録，講義録の作成 ボイスメールの文字化
人対コンピュータ	(カテゴリー III) 情報案内システム 予約システム コールセンターの自動化	(カテゴリー IV) ディクテーション

表 -1　音声認識タスクの分類と応用例

に分類することができる．

カテゴリー I と II は，人と人との対話あるいは独話を対象とするもので，いたるところに存在する音声ドキュメントをアーカイブ化（コンテンツ化）し，検索・活用できるようにする技術として，重要度が増している．この音声ドキュメントの処理では，音声をそのまま文字化するだけでなく，さらに要約・インデキシングなどによるメタデータ化によって，その後の処理を容易にすることが必要である．講演や講義のように，比較的長くしかも冗長な表現を含む話し言葉音声を，自動的に要約する研究も行われている．講演や講義のような独話（カテゴリー II）は，それだけを人が聞いて理解できるように発声されるが，会議や対談のような対話（カテゴリー I）では，省略や，「それ」が何を指すかなどといった，照応を含む相互のやりとり（文脈）が情報の伝達に重要な役割を果たし，各発声が断片的になりがちなので，音声認識はより一層難しくなる．

カテゴリー III は，人とコンピュータシステムとの対話を対象とするもので，情報検索，予約などを行う実用システムが，米国を中心にすでに多数利用されている．あらかじめ明確に定義された応用タスクを前提としてシステムを設計するのが普通で，この点で他のカテゴリーとアプローチが異なる．人がコンピュータと対話するときには，相手が人の場合と異なり，コンピュータを意識して比較的単純な発声がされるのが普通であるが，ユーザにとっては，アイコンで表示される GUI と異なって，何をどうしゃべったらよいのか分からないといった難しさがある．入力音声と望まれる動作との対応付け（意味理解）に関しては，種々の方法が研究されている．多くの場合，音声認識結果としての単語列あるいはその集合から，内容を表す単語を抽出し，意味あるいは対話のゴールへ変換する処理が行われる．

カテゴリー IV は，人がコンピュータに独話で話している音声のディクテーションであるが，この場合も音声認識誤りを避けることはできないので，その修正を含むシステムは，通常，キーボード入力を含む対話形システムの構成をとる．カテゴリー III では，システムへの入力手段として音声が用いられるのに対し，他のカテゴリー I，II，IV では，音声そのものがドキュメント，言い換えると情報コンテンツとして扱われるところが異なる．

これまでの技術的進歩にもかかわらず，音声認識の実用化の拡大には多くの課題が残っている．その第 1 は，対象の発話スタイルによって，認識性能が大きく異なることである．孤立単語，読み上げ音声などの認識は，数千語あるいは数万語の大語彙を対象としても，すでにほぼ実用レベルに達しているが，自然な話し言葉に対しては，残念ながらまだ限られた性能しか得られていない．その原因は，話し言葉音声には，言い直し，言い淀み，繰り返し，間投詞，不正確な発音などが含まれ，音響的にも言語的にも，書き言葉を読み上げた音声と大きく異なること，そのためにまだモデル化がほとんどできていないことにある．さらに，それを学習するための大規模な話し言葉コーパスがないことも大きな障害になっている．システムに登録されていない語彙（未知語）をユーザが発声したときの対処，話者による認識性能の違い，雑音や部屋の残響が加わった音声に対する認識性能の低下に対する対処なども，重要な研究課題である．

音声合成技術に関しても，最近では，コーパスベースと呼ばれる方式の研究が盛んに行われ，大量のデータに基づく自動学習や音声単位選択法によって，高品質で自然性の高い合成音ができるようになった．しかしこの方法では，任意の話者性や感情・発話スタイルの制御など，多様で表情豊かな音声を合成しようとすると，途方もなく膨大なコーパスを構築することが必要となり，非現実的となってしまう問題がある．

このような課題にチャレンジするため，最近，音声処理に直接関係のある次のような多数の大型研究プロジェクトが進められ，今後の新たな展開への萌芽が得られつつある．

(1) 科学技術振興調整費「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」プロジェクト（責任者：古井貞熙）：1999 年度より 2003 年度までの 5 年間，次の 3 つのサブテーマを中心とする活動が行われた．
  (a) 大規模話し言葉コーパスの構築（主として講演音声を対象として，延べ 661 時間，752 万形態素からなる，質量ともに世界最高レベルの話し言葉コーパスが構築され，公開されている）．
  (b) 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．
  (c) 話し言葉の音声要約プロトタイプシステムの構築．
(2) 学術創成研究「言語理解と行動制御」（責任者：田中穂積）：行動制御という観点から言語理解の仕組みを明らかにする研究が，2001 年度から 5 年計画で行われ

ている．新しい学術の創成を目指し，言語・行為・認知に関する基礎理論，言語処理，音声処理，ロボティクス・コンピュータグラフィクス（CG）の4つの分野に分けて研究を行っている．話し言葉音声による対話理解に関する分野横断的な実証研究として，対話理解の結果を，仮想空間内にシミュレートしたソフトウェアロボットの行動の映像化で表現するプロトタイプシステムを構築している．

(3)特定領域研究「韻律と音声処理」（責任者:広瀬啓吉）:文字言語にはない音声言語特有の特徴である，韻律の基礎から応用までを統合して発展させることを目的として，2000年度から2003年度まで行われた．韻律のモデル化，韻律の多様性の分析，韻律コーパスの作成，韻律の観点からの音声合成・音声認識の性能向上の研究などが行われた．

(4) 名古屋大学COE「多元音響信号の統合的理解」（責任者：板倉文忠）：1999年度から2003年度までの5年間，空間物理，信号構造，情報変換，言語論，認知論の5つの視座から音響信号を捉える多面的な研究が行われた．車内音声収集用実験車を用いて，実走行車内での対話音声コーパスの構築が行われた．その規模の大きさ，複数センサによる収集，実環境下での収集，音声対話システムとの対話の収集など，多くの特徴を持つコーパスである．

(5) IPA「擬人化音声対話エージェント基本ソフトウェア開発プロジェクト」（責任者：嵯峨山茂樹）：2000年度から2002年度までの3年間，音声認識･音声合成･顔画像合成を主たる機能として持つ，擬人化音声対話エージェントのツールキット「Galatea」が構築された．研究のプラットフォームとして利用されることを想定して，カスタマイズ可能性が重視されており，顔画像が容易に交換可能，対話制御の記述変更が容易などの特徴がある．

(6) ATR「多言語音声翻訳技術と評価プロジェクト」（責任者：山本誠一）：2003年度から3年間の計画で，携帯型端末を入力端末とした日本語と，英語・中国語・韓国語間の音声翻訳技術の研究開発プロジェクトを推進している．

本特集では，音声情報処理技術の最先端を研究している方々に，上記のプロジェクトの成果の一部を含み，最新技術の内容と今後の展望を，できるだけやさしく解説していただいた．

第1編「隠れマルコフモデルによる音声認識と音声合成」では，HMMの定義および関連するアルゴリズムについて概説した上で，音声認識および音声合成におけるHMMの利用について述べ，HMMの限界を指摘した上で，次世代音声モデルとして期待される手法について述べている．

第2編「統計的手法を用いた音声モデリングの高度化とその音声認識への応用」では，HMMによる音声のモデル化に，より一層の柔軟性を持たせ，統計的モデルとしての高度化を目指す新しいアプローチの中から，特に3つのトピックス:モデル選択，話者適応化，ダイナミックベイジアンネットワークを用いたモデリングについて解説し，今後の展望を述べている．

第3編「重み付き有限状態トランスデューサによる音声認識」では，複雑化している音声認識アルゴリズムの問題点を解消し，新しい機能が容易に組み込める方法として，有限状態トランスデューサに基づく音声認識の利用について解説している．まず有限状態トランスデューサの基礎から応用までを概説し，従来の音声認識手法との違い，現在注目されるに至った経緯，今後の展望などを述べている．

第4編「話し言葉による音声対話システム」では，人間と自然な話し言葉音声を介して対話を行うシステムを実現するための音声認識，音声理解の方法論について述べた上で，典型的な対話システムの構成法について解説している．

第5編「話し言葉における言い直しの処理」では，音声対話システムの実現において重要な課題の1つとして，言い誤りとそれに伴う言い直しに関する研究の必要性と歴史を概観している．言い直しの生成メカニズムのモデルを説明した後で，言い直しを検出・処理するための技術を音声情報処理と自然言語処理の2つの側面から解説している．

第6編「自動車の中での音声認識」では，音声認識の重要な実用化ドメインと考えられている，走行自動車内の情報インタフェースについて，現状，高度化に向けた技術的課題，および，要素技術の研究動向について解説している．

第7編「擬人化音声対話エージェント」では，今後の知的な対話システムの実現において重要と考えられる，人間的な外面とインタフェースを持つシステムについて解説している．特に，ユーザと音声で対話をし，表情豊かな顔の動画像を持つ擬人化音声対話エージェントの技術：音声認識，音声合成，顔画像合成，対話制御などについて述べている．

本特集が，読者にとって，音声情報処理の最先端技術に関する理解を深め，今後の展望を考える上で役に立つことを願っている．最後に，ご多忙の中，快くご執筆いただいた著者の方々に厚くお礼申し上げる．

（平成16年7月13日）