# 音声認識技術

State of the Speech Recognition Technology

河原達也

Abstract

音声認識技術は近年急速な進歩を遂げており，音声翻訳を含め，音声書き起こしや音声検索・音声対話など多くの実用化がされている．本稿では，音声認識の歴史と基本的な方法論を概観した上で，最近の技術革新の基盤となったビッグデータパラダイムとニューラルネットワークによるモデルについて述べる．その上で，現在実用化が進められている音声認識アプリケーションと音声翻訳への展開について紹介する．

キーワード：音声認識，ニューラルネットワーク，ビッグデータ

## 1. 音声認識の歴史

音声認識や音声翻訳は長い間 SF の範ちゅうであった反面，なかなか実用レベルに到達しない技術であった．しかし 21 世紀に入って，機械学習の方法論と計算機・情報通信技術（ICT）の進歩に伴って，飛躍的な性能改善を遂げ，様々な実用化が行われた．今では，スマートフォンに搭載されている音声検索やアシスタントアプリは多くの人に認知され，音声翻訳アプリも複数リリースされている．また，テレビ放送の字幕付与や国会の会議録作成に音声認識技術が導入されるに至っている．更にこの数年の間で，ニューラルネットワークに基づくモデルにより一層の性能の向上が実現されている．本稿ではまず，歴史的経緯を簡単に振り返り，最近の技術革新の主な要因について述べる．

音声認識の研究が開始されたのは今から 50 年以上も前に遡る．京都大学では 1960 年頃に単音節単位の認識を行う「音声タイプ」が構築されている[1]．その後，音声認識に有効な音響特徴量と，DP マッチングに代表される動的パターンのマッチング手法に関する基礎的な研究が世界中で行われた．これは，パターン認識の観点からはテンプレートベースの方法と言える．特定話者の音声認識は何とか動作しても，多数話者のバリエーションをモデル化するには不十分であった．これに対して，確率的なモデルを導入することにより解決が図られた．DP マッチングを拡張した形で隠れマルコフモデル（HMM）が導入され，その改良が 20 年以上にわたって行われた．まず，HMM の各状態の音響特徴量のパターンを連続分布でモデル化する混合ガウス分布（GMM）が導入された．そして，これを最ゆう推定する代わりに，識別誤りが最小化されるように学習（識別学習）するための様々な方法が提案された．2000 年代に実用化された音声認識システムは，基本的に GMM-HMM の識別学習に基づくものである．一方，言語モデルについては，単語の連接規則（文法）をオートマトンで記述したものから，その遷移を確率的なものにし，その確率をコーパスから最ゆう推定する $n$-gram モデルに移行していった．以上の変遷をまとめたのが表 1 である．世代の定義は古井[2]に従ったものであるが，第 4 世代は筆者が追加したものである．この第 4 世代が，ニューラルネットワークに基づくモデルである．音響モデルについては，GMM による確率計算をディープニューラルネットワーク（DNN）に置き換えた DNN-HMM が，言語モ

表 1　音声認識の方法論の変遷

| | | |
|---|---|---|
| 第 1 世代 | 1950～1960 年代 | ヒューリスティック |
| 第 2 世代 | 1960～1980 年代 | テンプレート…DP マッチング，オートマトン |
| 第 3 世代 | 1980～1990 年代 | 統計モデル…GMM-HMM，$n$-gram |
| 3.5 世代 | 1990～2000 年代 | 統計モデルの識別学習 |
| 第 4 世代 | 2010 年代 | ニューラルネット…DNN-HMM，RNN |

河原達也　正員　京都大学学術情報メディアセンター
E-mail kawahara@i.kyoto-u.ac.jp
Tatsuya KAWAHARA, Member (Academic Center for Computing and Media Studies, Kyoto University, Kyoto-shi, 606-8501 Japan).
電子情報通信学会誌　Vol.98 No.8 pp.710-717 2015 年 8 月
©電子情報通信学会 2015

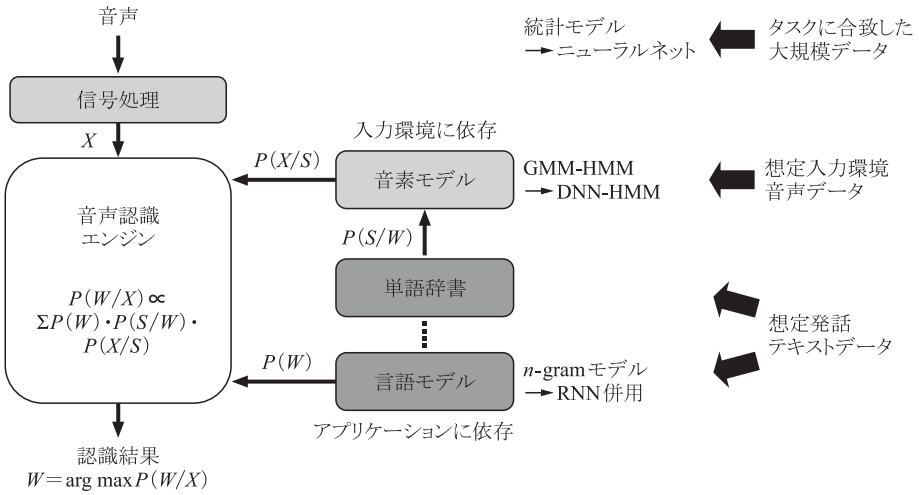


**図1　音声認識システムの構成**

デルについては，リカレントニューラルネットワーク（RNN）を *n*-gram と併用するモデルが一般的になっている．この展開について次章で述べる．

## 2. 音声認識の原理とシステム構築法

音声認識は，音声 $X$ が与えられたときにその単語列 $W$ を同定する問題である．これは，以下の式（1）のように，$p(W|X)$ をベイズ則で書き換えて得られる二つの項の積が最大となる $W$ を同定する問題として定式化される．

$$\arg\max p(W|X)=\arg\max p(W)p(X|W) \qquad (1)$$

実際には，単語は音素などのサブワード単位 $S$ でモデル化され，単語と音素の関係は辞書で決定的に与えられる（$p(S|W)=\{1,0\}$）ので，右辺の中身は以下のようになる．

$$\begin{aligned} p(W)p(X|W)&=\sum_{s} p(W)p(S|W)p(X|S) \\ &\approx \max p(W)p(X|S) \end{aligned} \qquad (2)$$

これは，単語列 $W$ の言葉が音声という雑音のある通信路を伝わってきたのを情報理論に基づいて復号するモデルである．$p(W)$ は（その言語あるいは状況において）単語列 $W$ が生成される先験的な確率であり，$p(X|W)$ は単語列 $W$（音素列 $S$）から音声（音響特徴量）$X$ が生成される確率である．

これは，音声認識が二つの確率モデルを推定する問題に分割され，各々が生成モデルとして定式化できることを意味する．具体的に，$p(W)$ を計算するモデルは言語モデルと呼ばれ，時系列（left-to-right）に探索するという制約・相性から単語 *n*-gram モデルが主に採用されてきた．これは，テキストデータを収集して単語連鎖（二つ組・三つ組）の出現頻度を計数すれば最ゆう推定できる．ただし，実際にはスムージングを要する．一方，$p(X|S)$ を計算するモデルは音響モデルと呼ばれ，音素の状態ごとに音声の音響特徴量の分布を GMM でモデル化する HMM が採用され，EM アルゴリズムによる最ゆう推定がベースラインの手法となった[(3)]．

以上の原理に基づく音声認識システムの構成を図1に示す．この枠組みは 1990 年頃に確立され，以降四半世紀以上にわたって，世界中（あらゆる言語）において普遍的に用いられてきた．しかしながら，（言語を特定しても）あらゆる用途に用いることができる普遍的・万能な音声認識システムが存在するわけではない．図1に記しているように，音響モデルは，音声認識システムが使われるアプリケーションの入力環境，具体的には音響条件・話者層・発話スタイルに合致するように，データを収集して学習する必要がある．言語モデルと単語辞書は，アプリケーションのタスクドメインに合致するように，想定発話のデータを収集して学習する必要がある．なお，音声認識エンジンは普遍的になっているが，技術的に高度・複雑になっているので，世界中でも筆者らが開発してきた Julius[(注1)] を含めて少数になっている．

要するに，音声認識の原理や音声認識エンジンは普遍的でも，万能な音声認識システムが世の中に存在するわけではない．アプリケーションごとに合致したモデルを構築する必要があり，このモデルの善し悪しが認識性能を左右する．モデルの善し悪しは，最先端（といってもかなり標準的）の技術を用いたとすると，学習データ

（注1）　http://julius.sourceforge.jp

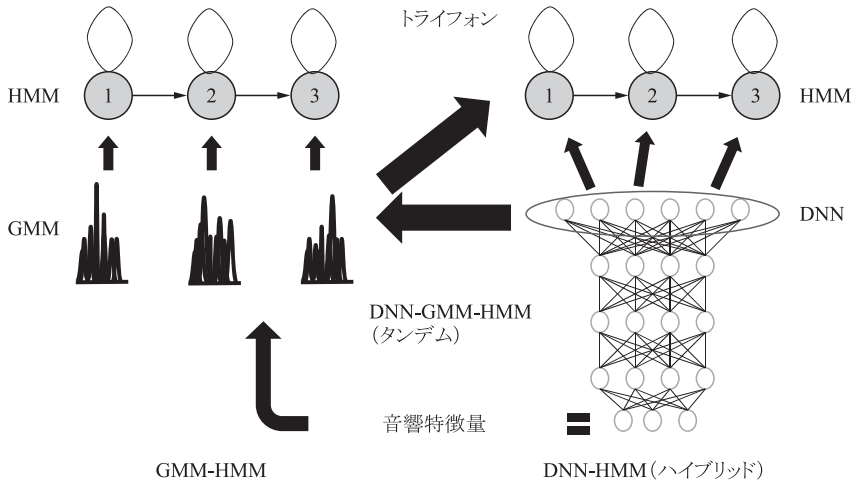


図2　GMM-HMM と DNN-HMM

ベースの規模が最も重要になる．したがって，音声認識システムの開発は，①アプリケーション設計，②データ収集，③モデル学習という流れにより構成される．

このように音声認識システムの構成論は 21 世紀初めに確立されたように思われたが，この 5 年ほどの間に更なる技術革新が行われた．一つは，ニューラルネットワークの導入であり，もう一つはビッグデータパラダイムである．筆者を含めて多くの人にとって認識性能が本当に良くなったと実感できるようになったのは，これら二つによるものである．以下に各々について説明する．

## 3. ニューラルネットワークの“逆襲”

音声認識においてニューラルネットワークを用いることは，1990 年頃にも盛んに研究が行われ，音素識別などでは良い性能が報告されたものの，式(1)，(2)に示した確率的な枠組みに基づく連続音声認識システムでは統計的なモデルである GMM-HMM が標準的になった．その後，GMM-HMM を改善するための様々な方法が研究され，最ゆう推定に代わって，識別誤りを最小化する学習法（識別学習）が導入された．これは，生成モデルのパラメータを，競合する仮説を用いて学習するものである．

これに対して，近年再びニューラルネットワークが注目されるようになり，数年のうちに主流になった．1990 年頃と比べて，入力特徴量（セグメント：数百次元），出力カテゴリー（トライフォン状態（後述）：数千クラス），中間層の層・ノード数共に，巨大化したのが最大の特徴であり，ディープニューラルネットワーク（DNN）と呼ばれる．多層のネットワークを逐次的に事前学習した上で，全体をバックプロパケーション学習するディープラーニングにより，このような大規模なネットワークの学習が可能になった[4]．

### 3.1　DNN-HMM による音響モデルの構成

DNN を音声認識の音響モデルに用いる直接的な方法は，従来の GMM-HMM における各状態の GMM による確率計算を DNN に置き換えるものであり，DNN-HMM ハイブリッドシステムと呼ばれる．別の方法として，DNN の出力若しくは中間層の値を特徴量として用いて，GMM を学習する DNN-GMM-HMM タンデムシステムもある．これらを図 2 に示す．

DNN-HMM ハイブリッドシステムでは，図 2 の右側に示すように，音響特徴量を DNN に入力する．DNN では出力層まで順次各層の計算を行う．出力層のノードは，HMM の各状態に対応付けられるが，一般的な音声認識ではトライフォンモデルの共有状態となる．これは，先行音素と後続音素の文脈を考慮したもので，クラスタリングを行っても数千個のオーダになる．音声認識で必要となるのは，式(1)の $p(X|W)$（より正確には式(2)の $p(X|S)$）であるが，DNN で計算される出力確率は通常 Softmax 関数を経た事後確率の形となるので，事前確率 $p(S)$ で除して，HMM に渡す．事前確率は，学習データベース中の各状態の頻度から推定する．GMM と比べて DNN の方が計算量が大きいが，単純な行列計算の組合せであるので，GPU による高速化が容易であり，リアルタイムの認識も十分に可能である．

DNN-HMM の学習手順の概要は以下のとおりである．

① GMM-HMM を学習する．これは，トライフォン状態のクラスタリングを含む．
② 学習データベースをトライフォン HMM の状態でアライメントする．
③ アライメントされたデータ（音響特徴量と HMM の状態ラベルの対）を用いて，DNN を学習する．これは通常，各層ごとの制約付ボルツマンマシン

（RBM）の教師なし事前学習とネットワーク全体の教師ありバックプロパゲーション学習から成る．

HMM の状態遷移確率は①で推定されたものを用い，出力確率のみ DNN で計算する．前述の事前確率の推定を含めて，学習データのアライメントを行うために，高い精度の GMM-HMM を必要とする．DNN の学習は，GMM と比べてはるかに時間を要するが，GPU などにより高速化が可能である．（GPU なしでは現実的に不可能と言ってよい．）

各種の音声認識タスクにおいて，DNN-HMM が従来の GMM-HMM をしのぐ認識精度を得られることが示されている．種々のベンチマークの結果を表 2 に示す．音素認識から大語彙連続音声認識までの様々なタスクにおいて，誤り率をおおむね 20〜30% 削減している．単一の方法により認識精度がこれほど大幅に改善したことは，筆者の知る限りほとんどなく，非常に画期的なことであった．

DNN が GMM に比べて優れている理由については様々な説明がされているが，最大の理由は，識別器に特徴抽出を統合して最適化しているためであろう．従来は，当該フレームのメル周波数ケプストラム係数（MFCC）やその回帰係数（ΔMFCC）などが主な特徴量として用いられてきたが，DNN を用いる際には，比較的広い範囲（前後 11 フレーム程度）のフィルタバンク出力をそのまま用いるのが最も効果的とされている．“生”の周波数特徴量を与えて，特徴抽出もニューラルネットワークの学習に委ねるブラックボックス化の発想と言える．これに対して，GMM-HMM を学習するには，統計的推定の信頼性の点から特徴量の次元を余り大きくできず，しかも無相関にすることが望ましいとされていた．そのため，MFCC や ΔMFCC に変換していたのであるが，このような単純な特徴抽出が性能のボトルネックになっていたことを示唆している．

### 3.2 RNN による言語モデル

音響モデルだけでなく，言語モデルにおいてもニューラルネットワークの研究・導入が進められている．言語モデルでは余りディープなネットワークは用いられず，中間層の出力を次の入力にフィードバックさせるリカレントなニューラルネットワークが一般的である．ただし，入力単語を少ないノードの数値データに射影する層を別途用意する．中間層はこれと履歴を符号化したものと捉えられ，*n*-gram モデルと比べて非常に長い履歴を考慮することができる．ただし，*n*-gram モデルの方が低頻度語のスムージングが効果的に行えることもあり，*n*-gram モデルと併用・線形補間する場合が多い．リアルタイムの認識に組み込むのは容易でないが，従来の *n*-gram モデルで生成した *n*-best 候補に対して，リスコアリング（別の高精度なモデルでゆう度を再評価）する枠組みでおおむね 5〜10% 程度誤り率の改善が得られることが報告されている．表 2 の CSJ 日本語講演音声認識で筆者らがベンチマークした結果では，絶対値で 1.7% 改善された．

## 4. データベース構築の限界 ——ビッグデータパラダイム——

音声認識システムの構成において，タスクに合致した学習データ量が鍵であることは先に述べた．そのため，研究コミュニティを挙げて，大規模な音声・テキストデータベースの構築が進められた．

図 3 に，代表的な音声データベースの構築時期とデータ量（時間数）をプロットしたものを示す．時代とともに，対象が読上げ音声から話し言葉音声に推移し，それに伴ってデータサイズが大規模化していることが分かる．更に，図 4 に筆者らが開発している国会審議の音声認識システムの音響モデルの学習音声データ量と認識精度の関係を示す[5]．線形ではないが，単調に改善していることが分かる．言語モデルの学習テキストデータ量についても，またほかのシステムでも同様の報告がされている[6]．

それではどのようにして，これだけ大規模なデータを集めるのであろうか．音声に限らず，文字や画像などのパターン認識の研究においては，単独の研究機関でデータベースを構築するのが困難なため，研究コミュニティで協力してデータを収集することがよく行われてきた．実際にこの「協調と競争」パラダイムは，1990 年代に

表 2　GMM-HMM と DNN-HMM の比較

| | 学習データ量 | GMM-HMM 単語誤り率 | DNN-HMM 単語誤り率 |
|---|---|---|---|
| TIMIT 音素認識 | 10 時間 | 27.3% | 22.4% |
| Switchboard 電話音声認識 | 300 時間 | 23.6% | 17.1% |
| Google 音声検索（英語） | 5,870 時間 | 16.0% | 12.3% |
| JNAS 日本語新聞記事読上げ | 85 時間 | 6.8% | 3.8% |
| CSJ 日本語講演音声認識 | 257 時間 | 20.0% | 17.5% |

上段の三つは文献（4）から引用，下段の二つは筆者らによる．

世界的に成功を収めた．

しかし最近では，この「データを頑張って集める」という発想自体が限界になってきている．実際に，そうやって頑張って集められるのはせいぜい数十～数百時間が限界である．また，被験者を集めて収集したデータが，実際のユーザが発話するものと適合するかも不明である．したがって，リアルなデータを自然に集積できる枠組みを構築することが考えられた．このようなビッグデータパラダイムが，音声認識の最近の成功の鍵となっている．

以下にその二つの典型的な事例について述べる．

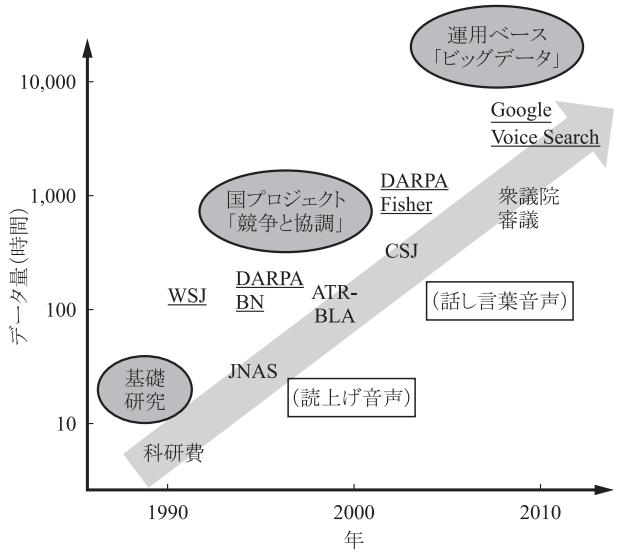


図3　代表的な音声データベースの構築時期とデータ量

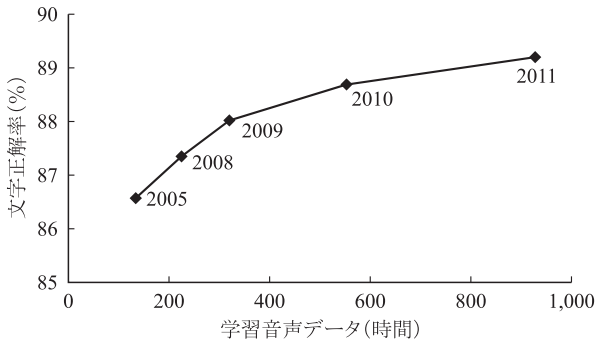


図4　国会審議音声認識における学習データ量と認識率の関係

### 4.1　携帯端末用クラウドサーバ型システム

携帯端末，特にスマートフォンのアプリケーションでは，クラウドサーバ型の音声認識が用いられている．これは，携帯端末に入力された音声をパケット化し，ネットワーク経由でサーバに送信して，認識処理を行うものである．これにより，端末の処理能力・記憶容量を気にせずに，大規模なモデルを用いた高精度な音声認識が可能になった．更に重要な点は，ユーザの発した音声データをサーバ側に蓄積できることである．サービスは無償のものが多く，利用者は数百万人にも達する[6]ので，リアルなデータが巨大な規模で蓄積されている．Googleの英語の音声検索では，発話データが1万時間規模になっている[4]．この枠組みを図5に示す．

### 4.2　会議音声と会議録の活用

会議や講演などの話し言葉の音声認識システムを構築するには，そのような音声とその忠実な書き起こしテキストを用意する必要がある．会議や講演は毎日のように行われるので，その音声を収録すること自体は容易である．しかし，これらには通常書き起こしがない．議会の場合は逐語的な会議録が作成されるが，忠実な書き起こしではなく，そのままでは音声認識のモデル学習には使えない．そこで筆者らは，会議録のテキストから実際の発言内容を確率的に予測する枠組みを考案した．例えば，「あのー」などのフィラーがどこに入りやすいかも予測することができる．この枠組みによって，会議録から話し言葉の統計的言語モデルを推定するとともに，会議録と音声から発言内容を復元し，1,000時間規模の会

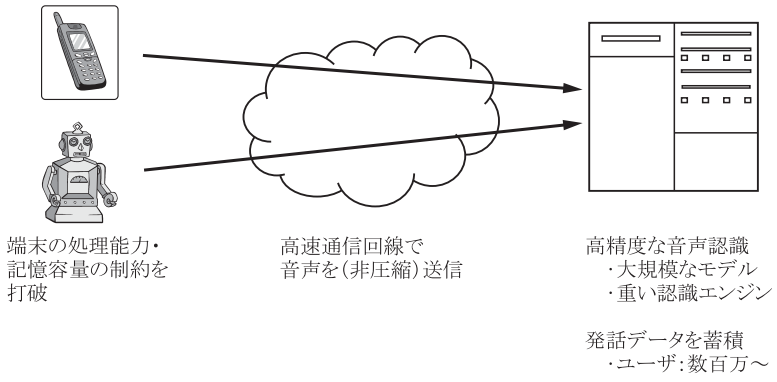


図5　携帯端末用クラウドサーバ型音声認識

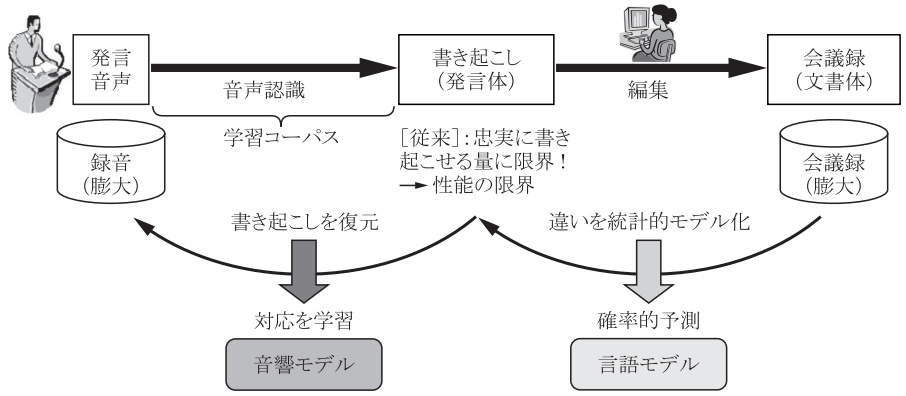


図6　会議音声と会議録テキストからのモデル学習

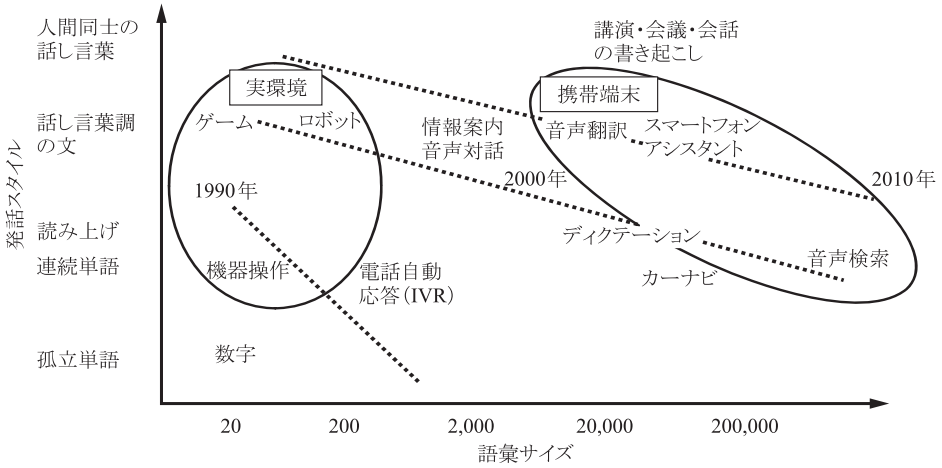


図7　音声認識のアプリケーション

議音声からほぼ自動的に音響モデルの学習が可能になった[5]．この枠組みの概要を図6に示す．この効果が図4に示されている．

## 5. 音声認識のアプリケーション

前章までは技術（シーズ）の観点から音声認識について述べてきたが，本章からは応用（ニーズ）の観点から述べる．これまでに紹介した以外のものも含めて，音声認識技術の代表的なアプリケーションを，使用環境・発話スタイル・語彙サイズの観点からプロットしたものを図7に示す．実環境では小語彙のアプリケーションが多く，携帯端末では大語彙で話し言葉調の文に対応していることが分かる．以下に，代表的なものを分類し，各々について説明する．

### 5.1　音声によるテキスト入力（ディクテーション）

音声タイプ・音声入力ワープロは，音声認識の長年の目標の一つであった．パソコンのディクテーションソフトとして1990年代に商品化され，現在ではWindowsにも標準搭載されているが，実際には余り利用されていない．大半の世代がキーボード入力に習熟し，音声入力の方がかえって疲れる，周囲に迷惑をかける，といった理由からであろう．携帯端末で簡単なメッセージを作成するなどの用途はあるが，音声のテキスト化という点では後述の書き起こしの方が主なニーズになると考えられる．

### 5.2　音声によるコマンド入力（カーナビ・ゲーム機など）

キーボードなどの入力装置が使えないハンズフリーの状況で，カーナビやゲーム機・家電機器を操作するのに音声は適している．ただし，未知の騒音・残響下で頑健に動作させることは容易でない．これらの機器では，計算資源・メモリが限られるので，十分な性能が得られないことが多かった．ただし，これらの機器もネットワー

クに接続されるようになると，携帯端末と同様にクラウドサーバ型の音声認識が導入できるようになる．そうなると，単なるコマンド入力ではなく，後述の情報アクセスへの展開も期待される．例えば，カーナビで近隣の情報案内を行うとか，家電機器で操作方法の説明を行うなどが考えられる．

### 5.3 音声による情報アクセス（電話応答装置・携帯端末）

電話や携帯端末で情報アクセスや予約などを行う際に，手順や選択肢が複雑だと，音声入力の方が便利である．米国では，2000 年頃から多くのコールセンターで音声認識を用いた電話音声自動応答（IVR）システムが導入されている．ただし，日本ではそれほど普及していない．これは，丁重なサービスに対する要求が高い反面，単純なことは早くから携帯電話のネットサービスで提供されていたためと考えられる．

しかしながら，スマートフォンの登場・普及により状況が一変した．パソコンと同様の複雑なことができるにもかかわらず，キーボードがない状況（音声入力のニーズ）が現れたのである．シーズ面からも，クラウドサーバ型の音声認識により性能が格段に改善した．その典型的なアプリは，音声検索（＝音声による情報検索）と Siri や「しゃべってコンシェル」[6] などのアシスタントソフトである．音声検索は，超大語彙にもかかわらず，効率良く Web や地図の検索が行えるので，特に携帯端末で重宝する．アシスタントソフトは，携帯端末の操作（コマンド入力）と情報アクセスを組み合わせたアプリで，情報アクセスについては音声検索のように検索結果の候補を表示するのではなく，天気や乗換案内などの情報をより直接的に応答するようになっている．

この項目の詳細は文献（7）を参照．

### 5.4 音声による会話（人間形ロボット・エージェント）

人間形ロボットや仮想エージェントの研究開発が進んでおり，音声による会話の機能も求められている．ただし現状では，挨拶程度の極めて小語彙の会話しか実現されていない．自由な話し言葉や実環境への対応が技術的に困難なためである．また，利用者として子供やお年寄りが想定されていることも技術的に困難な要因になっている．人間形ロボットの場合，ロボット自体が動き，非定常な雑音源となるため，ロボットに搭載したマイクで音声認識を行うのは非常に困難である．ただし現実には，利用者がロボットやエージェントとの会話に実質的に多くを求めているわけではないので，いわゆる“ゆるキャラ”に則した能力でよいとも考えられる．このように，音声認識だけでなく，音声合成を含めてキャラクタを設計することが重要である．

### 5.5 音声の書き起こし（会議録・講演録・字幕付与）

音声をテキスト化するという点では，ディクテーションと同じであるが，ユーザが機械に向かって話す設定ではなく，会議や講演などの人間同士の自然な音声コミュニケーションを対象とする場合はるかに困難になる．音声認識システムに向かって話す場合，必然的に発話が丁寧・明瞭になる上に，認識がうまくいかないとすぐにフィードバックされる．これに対して，会議や講演などでは考えながら発話がされるため，区切りが明確でなく，個々の発声も明瞭とは限らない．したがって，このような話し言葉に特化したモデル化が必要になる．現状ではニュース番組や議会・学会講演などの公の場で話される音声（＝パブリックスピーキング）に関して，個別のモデル化を行うことで実用的なレベルに達しつつある．例えば，テレビ番組への字幕付与や議会の会議録作成において音声認識システムが実用化されている．

この項目の詳細は文献（8）を参照．

### 5.6 音声の検索・マイニング

音声を人間が読む形式でテキスト化しなくても，長時間・大規模の音声データを音声認識することにより，検索やマイニングが可能になる場合がある．例えば裁判所では，公判を録音・録画しているが，検索を容易にするために音声認識を導入している．またコールセンターでも，顧客とオペレータの会話を収録しているが，音声認識により，どのようなトラブルが増えているか，適切な対応がされているかなどのマイニング・分析を行うことができる．また，米国で大規模に行われた電話会話音声認識の研究プロジェクトは公安目的が想定されている．これらに限らず，音声コンテンツを検索するために音声認識は有用と考えられ，今後の展開が期待される．

### 5.7 語学学習支援

外国語（特に日本人が英語）の発音や会話スキルを習得する動機と手間が大きいため，それを支援するシステムのニーズは大きい．これには，与えられた単語や文章の発音をチェックするものと，ショッピング等の特定の場面での会話を模擬するものがある．日本人が誤りやすいパターンをモデル化することで，より的確な誤り検出とフィードバックを行うことができる．ただし，十分な精度を得るには，非母語話者である日本人が発声した英語のデータに基づいて音声認識のモデルを構築する必要がある．また，対象（小児・大学生・社会人など）や目的（試験・旅行・ビジネスなど）に応じた適切なコンテンツの作成や指導を行うためには，語学教室・教師との連携が必要である．

この項目の詳細は文献（9）を参照．

## 6. 音声翻訳における音声認識

音声翻訳は，音声認識を行った結果に対して機械翻訳を適用するものであるが，幾つかのアプリケーションが考えられる．最も典型的なのは，外国語話者とコミュニケーションを行う際に支援を行う場合で，双方向・リアルタイムなシステムが要求される．例えば，日英音声翻訳では，日本語と英語の音声認識がリアルタイムで動作する必要がある．数十年前は夢の技術であったが，現在ではスマートフォンアプリとしてリリースされている．ドメインを限定しないものから，旅行や医療などドメインを絞ったものがある．後者の方が，音声認識や翻訳の精度も高くなることが期待される．

単方向でリアルタイムな音声翻訳として，講演やテレビ番組の同時通訳がある．音声認識としては，5.5 の音声の書き起こし（会議録・講演録・字幕付与）に該当する．また，単方向でオフラインの音声翻訳として，外国語の音声アーカイブの検索がある．これは，5.6 の音声の検索・マイニングに該当する．

以下に音声翻訳のための音声認識において，留意する点を述べる．

### 6.1 ドメインの設定とドメイン外発話の検出

旅行会話，更にはショッピングやレストラン，ホテルなどのドメインを設定した方が，発話パターンが限定されるので音声認識・翻訳は容易になる．しかし，ユーザがドメインを必ずしも明確に認識しなかったり，別のドメインに自然に遷移したりする場合もある．例えば，買い物の話をしていても，移動のための交通の話になったりする．また，会話の途中で個人的な経験や時事的な話題の話をする場合もある．ユーザが明確に分かるようなドメインでなければ，オープンドメインの汎用的な音声認識システムを構築する方がよい．

### 6.2 固有名詞の検出と翻字

旅行会話などでは，氏名や地名などの固有名詞が重要であるが，音声認識辞書でこれらを全てカバー・モデル化するのは容易でない．氏名や地名の部分のみを別モジュールにすることも考えられる．

### 6.3 文単位の検出

音声認識で用いる言語モデルは通常，単語の連鎖を扱っており，文や節などの言語的な単位を処理するわけではない．しかし機械翻訳は，文や節の単位を想定していることが一般的であるので，音声認識結果を区分化する処理が必要となる．日本語の場合は，句読点を言語モデルに組み込むことも考えられるが，SVM や CRF などの識別モデルを用いてチャンキングを行う方が一般的に性能がよい．講演などの同時通訳を行う場合は，リアルタイムに適した単位に区切る必要がある．

## 7. 人間並みの音声認識の実現に向けて

音声認識はかなり高度になり，実用レベルになったとはいえ，基本的には（能力の高い）外国語話者の域を出ない．一般人の話し言葉にはほとんど対応できないし，騒音下では途端に性能が低下する．母語話者のようなリスニング能力が実現されるのは想像できないくらい先のことのように思われ，それにはまだまだ素朴なブレークスルーが必要と思われる．

例えば，現在の音声認識システムでは，周波数特徴量に関する音響モデルと局所的な単語連鎖に基づく言語モデルのゆう度のみしか用いていないが，韻律に関するモデルや，意味や話題を考慮した高次・大局的な言語モデルを組み合わせることが期待される．そのためには，伝統的な式（1）の定式化から脱却し，一般的な情報統合の枠組みを構成する必要がある．

### 文　　献


（1） T. Sakai and S. Doshita, “The phonetic typewriter,” Proc. IFIP Congress, pp. 445-450, 1962.

（2） S. Furui, “Selected topics from 40 years of research on speech and speaker recognition,” Proc. InterSpeech, pp. 1-8, 2009.

（3） 鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄，音声認識システム，オーム社，東京，2001.

（4） G. Hinton, L. Deng, Y. Dong, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82-97, 2012.

（5） 河原達也，“議会の会議録作成のための音声認識―衆議院のシステムの概要―，”情処学音声言語情報処理研報，no. SLP-93-5, 2012.

（6） 辻野孝輔，栄藤　稔，磯田佳徳，飯塚真也，“実サービスにおける音声認識と自然言語インタフェース技術，”人工知能誌，vol. 28, no. 1, pp. 75-81, 2013.

（7） 河原達也，“音声対話システムの進化と淘汰―歴史と最近の技術動向―，”人工知能誌，vol. 28, no. 1, pp. 45-51, 2013.

（8） 河原達也，“話し言葉の音声認識の進展―議会の会議録作成から講演・講義の字幕付与へ―，”メディア教育研究，vol. 9, no. 1, pp. 1-8, 2012.

（9） 河原達也，峯松信明，“音声情報処理技術を用いた外国語学習支援，”信学論（D），vol. J96-D, no. 7, pp. 1549-1565, July 2013.


（平成 27 年 3 月 3 日受付　平成 27 年 3 月 31 日最終受付）


河原（かわはら）　達也（たつや）（正員）

京大学術情報メディアセンター／大学院情報学研究科教授．音声言語処理，特に音声認識及び対話システムに関する研究に従事．主著に，「音声認識システム」，「音声対話システム」（いずれもオーム社）．IEEE，情報処理学会，日本音響学会，人工知能学会，言語処理学会各会員．