Vol.2014-SLP-100 No.3
2014/1/31

音声認識の方法論に関する考察 —世代交代に向けて—

河原達也[†1]

音声認識技術の歴史的変遷を概観し、今後の展望について述べる。特に、音声認識の統計モデルの方法論に関して、従来“常識”と考えられてきたことが徐々に変遷していることを説明する。まず、学習コーパスを人手で編纂するという方法論は限界に達し、自然に超大規模に集積するビッグデータパラダイムが近年の実用システム成功の鍵であることを述べる。次に、HMM や N-gram などの生成モデルの最尤推定に代わって、最近研究コミュニティで主流になっている識別学習及び識別モデル、特に DNN について概観する。その上で、従来の通信路モデル（情報理論）に基づく定式化が、より一般的な枠組みに置き換えられるべきであることを指摘する。

1. はじめに

音声認識は「なかなか使いモノにならない」と長らく言われ続けたが、最近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システムは一般の多くの方に認知されるようになった。また、放送番組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など、話し言葉への対応も一定の範囲では実用的な水準に達している。実際に、これらのシステムの性能は、我々研究者が見ても（ひいき目で？）相当高く感じられる。

現代の音声認識システムの原型ができたのは 1980 年代と考えられる。これは、通信路モデル（情報理論）に基づく音響モデルと言語モデルの確率的な定式化と、それらの統計的モデル化・機械学習を基盤としている。具体的には、隠れマルコフモデル(HMM)や N-gram モデルに代表される生成モデルの統計量を学習データに基づいて最尤推定するという方法論がベースラインとなっている[1][2][3]。この枠組みはその後四半世紀（これは著者の研究キャリアと符合する）にわたって、音声認識の普遍的な原理として、世界中で用いられてきた。

このように基本的な方法論が変わっていないにも関わらず、音声認識システムの性能はその間に飛躍的に進歩した。これは、統計モデルの洗練と学習データの大規模化によるものである。その間の計算機の処理能力の大きな向上によるところもある。特にここ最近実用化されたシステムは、データの大規模化が量的なレベルから質的なレベルに転じてきた。すなわち、学習コーパスを人手で編纂する（“頑張って集める”）という限界を超え、超大規模に集積する（“自然に集まる”）データを活用しようという考え方になっている。いわゆるビッグデータパラダイムといえるが、これは、従来のパターン認識の基本的な教師付き学習がそのまま適用できず、準教師付き学習の枠組みが鍵になることを意味する。

また、HMM や N-gram モデルの学習方法自体も、識別学習や適応・正規化などの様々な洗練（変形？）が加わることで、ベースラインとはかなり異なった複雑なものになっている。さらに近年になって、ディープニューラルネットワーク(DNN)といった直接的な識別モデルの検討が進められ、HMM を凌ぐ認識精度が続々と報告されている[4]。

本稿では、このような動向を概観しながら、音声認識の方法論に関する考察を行う。

2. 音声認識の“常識的な”定式化

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２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。

$$\arg\max p(W \mid X) = \arg\max p(W)p(X \mid W) \quad (1)$$

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から音声（の特徴量）Xが生成される確率である a。

これは、音声認識が２つの確率モデルを推定する問題 b 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いる。これは、テキストデータを収集して単語連鎖（２つ組・３つ組）の出現頻度を計数すれば最尤推定できる c。一方、$p(X|W)$を計算するモデルは音響モデルと呼ばれ、音素毎に音声の特徴量の分布をモデル化するHMMが採用され、EMアルゴリズムによる最尤推定が行われる。

---

a 実際には、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、以下のようになる。

$$p(W)p(X \mid W) = \sum_S p(W)p(S \mid W)p(X \mid S) \approx \max p(W)p(X \mid S) \quad (2)$$

b 大語彙連続音声認識では、２つを組み合わせて最尤の仮説を探索する問題もある。

c 実際にはスムージングを要する。

†1 京都大学
Kyoto University

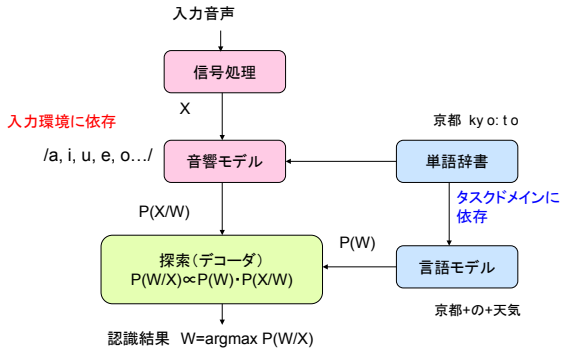

図 1 音声認識の原理

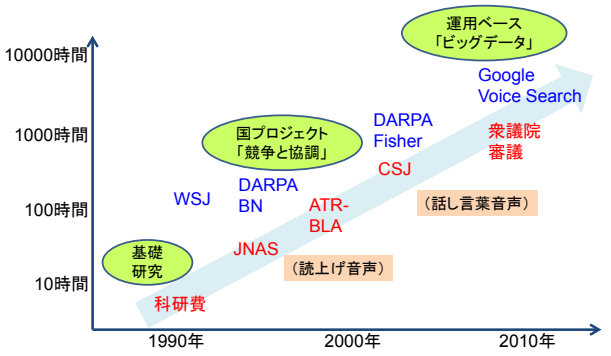

図 2 代表的な音声データベースの構築時期とデータ量

以上述べた音声認識の原理を図 1 に示す。この原理は、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。実際に、世界中のあらゆるテキストに記述されているし、世界中の“実用的な”音声認識システムのほとんどすべてが、HMM と N-gram モデル（もしくは生成文法規則）に基づいて構成されていると思われる。

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Juliusを含めて少数になっている d。

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ない。アプリケーション毎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から構成されるエンジニアリングとして確立されてきている。

表 1 に典型的な音声認識システムの構成例を示す。

表 1 典型的な音声認識システムの構成

	音響モデル	言語モデル
ディクテーション(Julius)	260 時間	2.7G 語
Google Voice Search	5000 時間	#検索数
国会審議の書き起こし	2000 時間	200M 語

3. データベース構築の限界—ビッグデータパラダイム

図 2 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わかる。さらに、図 3 に著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わかる。言語モデルの学習テキストデータ量についても、また他のシステムでも同様の報告がされている[7]。

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は限界があるため、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２つの典型的な事例について述べる。

d このように音声認識エンジンと音響モデル・言語モデルを完全に分離して、様々な研究機関が様々なシステムを構成できるようにしたのが Julius の（オープンソースであることに加えて）最大の特長である。

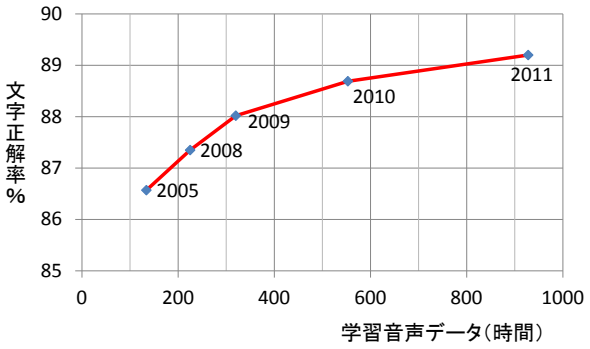

図 3　国会審議音声認識における学習データ量と認識率の関係

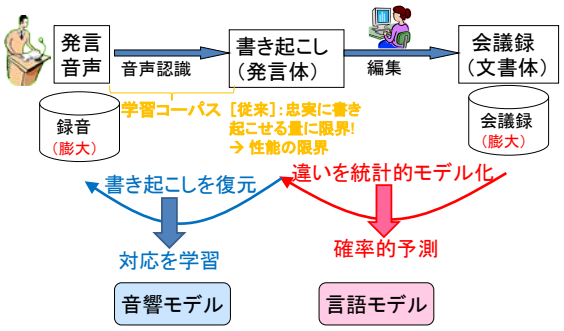

図 4　会議音声と会議録テキストからのモデル学習

3.1 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。さらに重要な点は、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[7]ので、リアルなデータが巨大な規模で蓄積されている。Google では、英語の音声検索の発話データが 5000 時間規模になっている[4]。

3.2 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。そこで著者らは、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、千時間規模の会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。この枠組みの概要を図 4 に示す[5][6]。この効果が図 3 に示されている。

4. 生成モデル・最尤推定の限界—識別モデルの導入

4.1 モデル推定手法の変遷

HMM ベースの音声認識システムが本格的に研究開発され始めた 1990 年代は、基本的な EM アルゴリズムに基づく最尤推定を行うのが主流であった。著者らが IPA のプロジェクトで開発した「日本ディクテーションツールキット」[e]、それに基づいて執筆したテキスト「音声認識システム」（2001 年刊行）[1]、そして毎年夏に開催している「音声認識技術講習会」はこの基本的なベースラインに沿っている。

しかし、音響モデルの技術はその間様々な研究開発を経て、VTLN・fMLLR などの正規化技術、MLLR などの適応技術、そして MPE・MMI などの識別学習が、現代の state-of-the-art システムには必須となっている。VTLN や MLLR は、話者や環境毎に特徴量やモデルパラメータの変換パラメータを最尤推定するもので、最尤推定の延長ともいえる。しかし、MPE などの識別学習は、誤り率最小化を学習規範とするもので、競合他クラスのサンプルもモデル推定に必要とする点で、最尤推定と根本的に異なるものである。ただし、あくまで生成モデルのパラメータを識別的に学習するというアプローチである。

これらの手法はタスクやデータベースによって効果にばらつきが大きく、それらの間に一見冗長性があるにも関わらず相乗効果もあり、これらを組み合わせて構成される state-of-the-art システムは結果として複雑怪奇なものになっている。

4.2 識別モデル・DNNの導入

これに対して近年、より直接的に識別モデルを導入しようという動きが大きくなっている[8]。識別モデルとは、条件付き確率$p(X|W)$ではなく、事後確率$p(W|X)$を推定するモデルである[f]。そのためには、競合他クラスのモデルと同時に最適化する必要がある。音声認識は本来識別タスクであるので、識別モデルの方が自然であり、HMMのような生成モデルは音声認識より音声合成に用いるのが自然かもしれない。

具体的な識別モデルとして、最大エントロピマルコフモデル[9]やセグメント単位の CRF[10]なども研究されてきた

e 「音声認識システム」[1]の付録 CD-ROM に梱包されている。
f ただし、式(1)(2)の枠組みで言語モデルと組み合わせてデコーディングするために、$p(S|X)$を先験確率 $p(S)$で除して $p(X|S)$に変換することが多い。

表 2　JNAS 及び CSJ における GMM-HMM と DNN-HMM の比較
（単語認識精度）

	JNAS	CSJ
HMM（3000 状態 x16 混合）	93.2%	80.0%
DNN（隠れ層 5x 各層 2048 ノード）	96.2%	82.5%

が、最近特に注目を集めているのがディープニューラルネットワーク(DNN)である[11][4]。基本的なハイブリッド型のシステムは、入力音声（の特徴量）X に対する HMM の各状態 S の確率 $p(X|S)$を GMM ではなく、ニューラルネットワークにより計算するものである。音声認識にニューラルネットワークを用いるのは、1990 年代前半にも盛んに行われていたが、入力特徴量（セグメント：数百次元）、出力カテゴリ（トライフォン状態：数千クラス）、中間層の層・ノード数ともに、巨大化したのが最大の特徴である。

この１～２年の間で各種の音声認識タスクにおいて、DNN が従来の GMM に基づく HMM を凌ぐ認識精度を得られることが世界中で報告されている[4]。JNAS 及び CSJ 学会講演の評価セットでベンチマークを行った結果を表 2 に示す。DNN では MLLR のような（教師なしの）話者適応を行うのが容易でないが、HMM に話者適応を行った場合と比べても認識精度が高い。認識の際の処理速度も速い。

4.3 DNNはなぜよいのか？

DNN が GMM に比べて高い認識性能を実現する理由については様々な説明がされているが、（著者が考える）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 MFCC や Δ MFCC 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いえる。ちなみに、特徴抽出を行う事前学習の過程で一般に用いられている制約付きボルツマンマシン(RBM)は生成モデルであり、最尤推定によって学習されている。

5. 通信路モデルの限界(?)—統計的機械翻訳の教訓

識別モデルの隆盛により、根源的に式(1)の通信路モデルが妥当であるかということも検討する段階に入ってきた。同様の事例として統計的機械翻訳がある。統計的機械翻訳は当初、式(1)と同様の通信路モデルで定式化されたが、現在の state-of-the-art システムは、様々な知識源・統計モデルから計算される尤度を統合する対数線形モデルの枠組みとなっている。すなわち式(3)のようになる。

$$\arg\max p(W \mid X) = \arg\max \frac{1}{Z} \sum \lambda_i * f_i(W, X) \qquad (3)$$

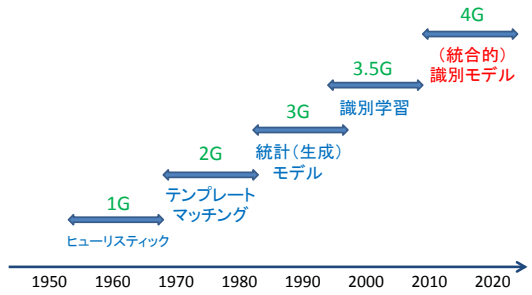

図 5　音声認識の方法論の変遷　（[12]から引用・一部改変）

ここで、$f(W,X)$は$p(W)$, $p(X|W)$, $p(W|X)$, $p(W,X)$などを計算する様々なモデル gによる尤度であり、λはその重みである。

音声認識もこのような枠組みにするのは必然的な流れではないだろうか。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 N-gram 言語モデルの尤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。

これは前節で述べたように、人為的なモジュール分割や特徴抽出を行わずに、できるだけ広範囲の“生”の情報をニューラルネットワークに与えて、学習・最適化に委ねるブラックボックス的な考え方にも通じるところがある h。

図 5 に音声認識の方法論の世代変遷を示す。文献[12]の時点（約５年前）は「第３.５世代(3.5G)」であって、「第４世代(4G)」に関する明確なイメージは示されていない。しかし、おそらくニューラルネットワークを発展させた識別モデル、あるいは式(3)のような統合的な識別モデルが主流になる可能性が高いと考えられる。

6. おわりに

音声認識システムの研究開発は、データもモデルも大規模になるのに従って、敷居が高くなり、学生等が行うのは容易でなくなってきた。著者が大学院生だった 1990 年頃は、学生の素朴なアイデアを実装し、自前のデータで評価しただけで、ICASSP などのトップカンファレンスに論文が採択されていた。今ではほとんど考えられないことである。

それをもって「音声認識研究は終わった」と言う向きもある。しかし現在の音声認識はかなり高度になったとはいえ、しょせん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、騒音下ではとたんに性能が低下する。

g 事後確率 $p(W|X)$を計算するためには、条件付き確率が１つは必要。
h ただし、音声と書き起こしを与えて、音響モデルも言語モデルも統合的に学習するという方法論は有望でない。なぜなら、テキストのみのデータが膨大に存在するためである。

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。DNN が音声認識を専門にしていない大学の研究室から提案された点も特筆すべきであろう。

謝辞： 図 3 と表 2 のベンチマークは各々秋田祐哉助教と三村正人研究員によるものである。

参考文献

[1] 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, 2001.
[2] F.Jelinek. Continuous speech recognition by statistical methods. Proc. IEEE, Vol.64, pp.532—556, 1976
[3] S.E.Levinson, L.R.Rabiner and M.M.Sondhi. An Introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. Bell Syst. Tech. J., Vol.62, No 4, pp.1035—1074, 1983.
[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 Processing Magazine, 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－. 情報処理学会研究報告, SLP-93-5, 2012.
[6] T.Kawahara. Transcription system using automatic speech recognition for the Japanese Parliament (Diet). In Proc. AAAI/IAAI, pp.2224--2228, 2012.
[7] 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. 人工知能学会誌, Vol.28, No.1, pp.75-81, 2013.
[8] M.Gales, S.Watanabe and E.Fosler-Lussier. Structured Discriminative Models for Speech Recognition. Signal Processing Magazine, Vol.29, No.6, pp.70—81, 2012.
[9] H.-K.J.Kuo and Y.Gao. Maximum Entropy Direct Models for Speech Recognition. IEEE Trans. Audio, Speech & Language Process. Vol.14, No.3, pp.873—881, 2006.
[10] G.Zweig and P.Nguyen. A Segmental CRF Approach to Large Vocabulary Continuous Speech Recognition. Proc. IEEE-ASRU, 2009.
[11] A.Mohamed, G.E.Dahl and G.Hinton. Acoustic Modeling Using Deep Belief Networks. IEEE Trans. Audio, Speech & Language Process. Vol.20, No.1, pp.14—22, 2012.
[12] S.Furui. Selected Topics from 40 Years of Research on Speech and Speaker Recognition. Proc. InterSpeech, pp.1-8, 2009.