自動音声認識における尤度についての直観

井原 健紘
電気通信大学 情報工学専攻

An intuition about likelihood of automatic speech recognition

Takehiro IHARA
Department of Computer Science, the University of Electro-Communications

概 要

パターン認識は入力を出力クラスに分類する問題として定式化されている．自動音声認識では出力クラスに分類するにあたって，各クラスへの連続値の出力確率を求めそれらを比較する．しかしながら，直観的には人間が音声認識をおこなう場合に出力確率に該当するような連続的な値を算出しているようには感じられない．また，出力に連続値を用いることは，同じクラスに分類されるべき入力にそのクラスへの帰属度合いを付与することを意味していると見なせるが，果たして人間は同一発話内容の複数の入力に対して序列を与えることができるだろうか．計算器は連続値しか出力することができず，人間は連続値を処理しているようには感じられないという隔たりについて本稿では考察をする．

1 はじめに

近年，自動音声認識技術が徐々に実用化されつつあるが，まだ性能は充分ではないとの見方もある．例えば，カーナビゲーションシステムは認識性能が低すぎて運転に影響が出るのではないかと危惧されている [1]．また，従来の自動音声認識技術を英語の発音矯正に使うことを疑問視する声もある [2]．実用化された音声認識器の性能は 90%程度であったという報告もあり [3]，これは人間の聴覚の性能には遠く及ばない．

[4] では「HMM (Hidden Markov Model)」「トライグラム」の改良によって自動音声認識性能は充分に向上するだろうと予想されている．この要素技術の分け方は，入力がそのクラスに属する確率が音響モデルによる尤度と言語モデルによる確率の積を計算することによって求まるという考え方に由来している．この考え方を導入することにより自動音声認識の性能は飛躍的に向上したのであるが，著者はこの考え方に違和感を覚える．人間による音声認識でも出力は確率で求められているだろうか？

著者の主観では，人間は,「あらゆる現実」というフレーズが背景雑音の極めて小さな環境で発せられた場合に，その発話が「あらゆる現実」である確率を考えることなく認識しているように思える．また，パターン認識一般を考えると，鳥が飛んでいる様子を見て，鳥である確率を考えているようには思えない．

しかしながら，計算器が音声認識をする場合には，どうしても出力が連続値になってしまう．これは避けられないことである．人間が「非確率的」に認識しているように感じ，計算器が「確率的」に出力を計算せざるを得ないというギャップについて，本稿では考察をする．なお，検証実験は一切おこなっていない．

2 「確率的な出力」の意味

まず，出力が確率であるとどのような問題が起きるのかということを述べる．現在の自動音声認識器は入力 x に対して，確率

$$P(W|x) \tag{1}$$

を与えている．ここで W は単語クラスである．このような確率を与えるということは，物体の認識に喩えれば，次のような処理が可能であることを意味する．

例えば，入力に対して「それが鳥である確率」を計算することのできるモデルが作られたとする．このモデルを用いて様々な種類の鳥に対して「鳥」としての確率を算出する．その結果として，スズメは 0.8 の確率で「鳥」，鷹は 0.7 の確率で「鳥」，ペ

ンギンは 0.2，カラスは 0.9，ダチョウは 0.5 などと「鳥」らしさの確率が算出される．

1 つ目の問題は，この鳥のモデルが入力の鳥に対して図 1 のようなヒエラルキーを作ってしまっているということである．カラスは非常に鳥らしく，ダチョウはさほど鳥らしくない．パターン認識は未知の入力に対して出力クラスを割り当てる問題であるとされることが多いが [5]，その実，出力クラスではなく出力確率を与えているに過ぎない[1]．比喩的にいえば，名詞を出力すべきであるのに形容詞を出力しているようなものである．なお，対象が「鳥」であると切迫感がないが,「人間」であると途端に切迫感が増す．自分がどれくらい人間らしいかを測りたがる人はおそらくいないだろう．また，人間らしさを他人と積極的に比べたがる人もいないだろう．そもそも，人間を認識するにあたって人間らしさを考える必要はないようにさえ感じられる．

話を音声に戻せば，1 つ目の問題というのは，複数の異なる「あ」という発音に対してヒエラルキーを作るのは妥当かということになる.「あめあがり」という単語を発話した際に，最初の「あ」と次の「あ」で/a/らしさが（明示的でなくとも）比較できてしまうのは奇妙な話である．

2 つ目の問題は，ヒエラルキーの頂点の存在である．鳥の例でいえば，図 1 の星の位置に相当する鳥が存在するということである．どのような確率分布を用いるにしろ，最も高い確率を与える入力が存在する．その最大確率の鳥を人間は果たして「それ以上ないほど鳥らしい鳥」であると感じるだろうか．また，人間がそのようなモデルをもし心の中に持っているとすれば,（個人個人で異なるにせよ）「それ以上ないほど鳥らしい鳥」を想像できるはずであるが，想像できないのはなぜだろうか．想像できないのは著者だけなのだろうか．

再び話を音声に戻す．2 つ目の問題は，/a/のモデルが与えられれば，最大の確率を与える「あ」を生成することができるということである．/a/やその他の音響モデルによって人間が音韻を認知しているならば,「それ以上ないほどあらしいあ」が想像できないのはなぜだろうか．想像できないのは著者だけなのだろうか．

以上，出力が確率であるというのは，これら 2 つの意味を持つ．必ずしもこれらが深刻な問題かどうかについては議論の余地があるが，少なくとも著者にとっては違和感がある．なお，認知科学の分野では事物の典型例（＝プロトタイプ）の存在が仮定されており，様々な事例はプロトタイプの派生例として説明されている [6]．認知科学の分野においてはこれら 2 つは全く問題とはならないことだろう．

[1]なお，線型識別器では連続的な確率を与えることはないが，入力が領域の中心であるか周辺であるかという概念は存在する．これは本稿の論旨に沿えば，確率を与えていることと同義である．

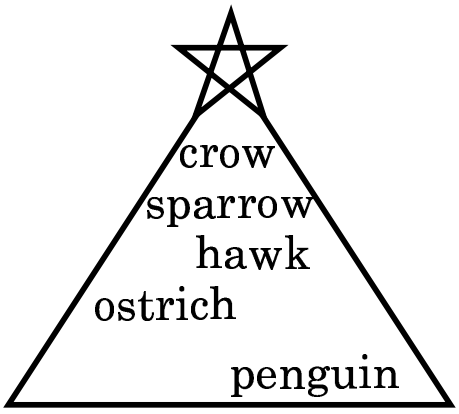

図 1: 「鳥」のヒエラルキー．星は「それ以上ないほど鳥らしい鳥」を示す．

3 それが問題ではないとする観点

冒頭にも書いたように，本稿で扱う問題は「自動音声認識の出力が確率であることの是非」である．そして，出力が確率であることは，入力に序列を与えることを可能にし，最大確率を与える入力の想定を可能にする．しかしながら，人間にはそのどちらもおそらく不可能である．この計算器と人間とのギャップに対しては，大きく分けて 2 つの立場をとることができる．これが問題であるとする立場と，これは問題ではないとする立場である．まずは，これが問題ではないとする観点について述べる．

問題の所在は，計算器が連続確率を出力し，人間が連続的でない出力を生成すると思われるところにあるので，とり得る立場は次の 2 種類になる．

3.1 自動音声認識の出力は確率ではない

もし，自動音声認識の出力が確率ではないと仮定すれば，計算器と人間との整合性がとれる．実際，途中までは自動音声認識器は式 (1) のような確率を計算しているものの,「それが何か」を識別した結果は離散的な単語となっている．

ただし，ここで問題になるのは，単語モデル間で比較する前の確率が，何を意味しているのかということである．確率は自動音声認識の途中経過の一部であり，何を意味しているのかを考えてはいけないという考え方も持ち得るかもしれないが，意味を持たない変数が存在するのは気分の悪いものがある．

3.2　人間の音声認識の出力は確率である

もし，人間の音声認識の出力が連続的な確率であると仮定すれば，計算器と人間との整合性がとれる．人間がどのように音声を認識しているのかは今のところ調べようがないので，立証することも反証することもできない．人間の認知機構は，異なる話者の同じ発話内容の発話に対して，異なる確率を出している可能性もある．

3.3　人間と計算器は異なる

本稿の趣旨とは外れるが，当然のことながら，人間と計算器の計算方法は異なっていてもよいとする考え方も存在する．この考え方は著者も否定しない．ただ，もしも計算方針が大きく異なるのだとしたら，それが認識器の性能の向上の妨げになっている可能性もある．

4　それが問題であるとする観点

直観的には，人間はある発話がどれだけその発話らしいかということを数値化することができない．例えば，異なる 2 人が「あめあがり」と発話したときに，それぞれの「あめあがり」がどれだけ標準的な「あめあがり」であるかを数値化することもできなければ，どちらがより標準的な「あめあがり」であるかを比較することもできないように思われる．ところが計算器には数値化や比較が可能である．このミスマッチが認識器の性能の向上の妨げになっているのかもしれない．

聴覚ではなく視覚の話であるが,「それが何であるか」と「それがどういった色や動きをしているか」を認知する脳の部位は異なると述べている著作物が存在している [7]．一般大衆向けの著作物であり，話を易しくしている可能性は考えられるが，話の本質をねじ曲げているということは考えられないだろう．音声における認知に関しては [7] では触れられていないが,「何を言っているか」と「どのように言っているか」を認知する方法が異なる可能性は否定できない．もし,「何を」と「どのように」の認知方法が異なるのだとしたら，現在の自動音声認識の手法にはねじれが存在することになる．式 (1) で表される確率は「発話入力がどれくらいその単語の標準的な発音に近いか」を表しており要するに「どのように」を表しているのであるが，その「どのように」を表す数値を用いて現在の自動音声認識は「何を言っているのか」を導き出そうとしている．本稿の表面的な問題は「自動音声認識の出力が確率であることの是非」であるが，より深く突き詰めると「どのように」と「何を」を区別すべきかどうかの問題となる．

なお,「何を」と「どのように」が区別されていないと，物体認識において「鳥らしくない鳥」を認識するのは困難となる．「入力物体が標準的な鳥から遠く」且つ「入力物体が標準的な鳥から離れすぎていない」という範囲を決める必要があるからである．同様に，ロボットに「きみらしくないね」と発話させるのも困難である．音声においても，非母語話者の発音に対して「標準的な『あめあがり』らしくない」と判定するのは難しい．やはり,「何を」と「どのように」は切り離して考える必要があるように思える．このような問題はこれまで「頑健性」という言葉で片づけられてきたが，自動音声認識が頑健になってもこの問題は解決しないと考えられる．なぜなら，頑健になるというのは認識におけるグレーゾーンが狭まることを意味し,「入力が標準から遠く，且つ，離れすぎていない」という範囲も狭まるからである．

「発話入力がどれくらいその単語の標準的な発音に近いか」を示す連続値を求めることはこれまでの自動音声認識ですでにおこなわれている．問題はそれらの連続値の比較で「どの単語を示しているのか」を識別することが妥当であるかどうかというところにある．ただし，この問題はともすると水掛け論になってしまうおそれがあるので，今はまだ提起するにとどめる．

5　領域に依存する存在密度

これまで，自動音声認識とは「何を言っているか」を推定する問題であるという前提の下に議論してきた．これは，パターン認識が入力をクラスに分類する問題として定式化されているためである．もしも，この自動音声認識の目指している問題自体が不適切だとしたらほかにどのような考え方ができるのかということについて本節では述べる．

計算器と人間との出力のミスマッチが本稿の主題である．計算器と人間とのミスマッチを解消する 1 つの方法として，両方とも連続値で出力できるような問題にしてしまうという方針が挙げられる．どのような問題でも計算器は連続値で出力すると考えられるので，人間が連続値を出力し，且つ，音声認識の趣旨を損なわない問題を考える必要がある．一例として，次のような案が考えられる．

説明の都合上，まずは鳥の認識に喩える．このとき一般的には物体に鳥である確率を付与することを考えるが，そうではなく，ある特定の領域に鳥がどれくらいの密度で存在するかを計算することを考え

る．言い換えれば「それがどれくらい標準的な鳥であるか」を計算するのではなく「そこに鳥と鳥以外のものはどれくらいの割合で存在するか」を計算する．例えば，日本全土を認識対象の領域とした場合，鳥の存在密度はかなり低くなることだろう．また，鳥の近傍を認識対象の領域とした場合，鳥の存在密度はかなり高くなることだろう.「そこにどれくらいの密度で鳥がいるか」というＳＮ比は，人間も計算器も同様に連続値で出力をするのが自然である．

話を音声認識に戻す．ある発話内容を認識するときに，本節の案では，まず時間領域を入力としてその領域内にその発話内容がどれくらいの密度で存在するかを出力とする.「あめあがり」という単語を認識することを考えると，1 時間の発話内に「あめあがり」という発話内容が存在する密度は低くなることと思われ，また，最初の「あ」から最後の「り」までを発話領域とすると存在密度はかなり高くなるはずである．また，雑音の音圧が相対的に大きいほど存在密度は低くなる．

前述の序列の問題は，背景雑音の問題となる．領域内には大なり小なり背景雑音が存在するので，その雑音が耳障りであるときには存在密度は低くなり，雑音が小さな場合には存在密度が高くなる．所望の信号そのものには序列がつかない．また，どのような時間領域にも異なる雑音が存在し，その雑音に応じて最大の存在密度を与える発話が変化するため，同じモデルを使用しても最大の存在密度を与える発話が一意に定まることはない．

この存在密度の考え方は，信号ではなく領域に確率を与えているという点で，ワードスポッティング [8] の考え方に近い．ただし，この存在密度の考え方が音声認識として本質的かどうかについては議論の余地が大いにある．

6　おわりに

本稿では，自動音声認識の出力が確率であることに対する違和感を述べた．自動音声認識の出力が連続確率であると，同一発話内容に序列を作ることができてしまい，また，最大確率を与える入力を想定することができてしまう．これら 2 つの作業を人間がおこなえるかどうかは疑問である．

また，自動音声認識は，発話入力がどれくらいその単語の標準的な発音に近いかという合致度を測っているが，これはその単語を「どのように言っているか」を表しているに過ぎないと解釈できると述べた．その単語を「どのように言っているか」を表す連続値を用いて，その発話が「何を言っているのか」を識別することは妥当であるのかという問題を提起した．

さらに，自動音声認識の問題の枠組みが何を言っているのかを推定することとなっていることに対しての疑念を述べた．そして，信号ではなく時間領域を入力としてその領域内にその発話内容がどれくらいの密度で存在するかを出力とする考え方を示した．この考え方が本質的かどうかは別として，最初に述べた序列の問題と最大確率の問題を回避することはできる．

今後の課題として，どのように工学的な研究を始めるかを考える必要がある．またその前に,「どのように言っているか」と「何を言っているか」の差異を見極める必要がある．

参考文献

[1] 神沼 光伸, “自動車用音声インタフェースへの期待,” 情報処理学会研究報告, 2006-SLP-63 (9-2), 2006.

[2] 峯松 信明, “音声に内在する音響的普遍構造とそれに基づく語学学習者モデリング,” 電子情報通信学会技術研究報告, SP2003-179, 2004.

[3] 鹿野 清宏, Cincarek Tobias, 川波 弘道, 西村 竜一, 李 晃伸, “音声情報案内システム「たけまるくん」および「キタちゃん」の開発,” 情報処理学会研究報告, 2006-SLP-63 (7), 2006.

[4] 中川 聖一, “音声言語処理研究の進展と今後,” 電子情報通信学会技術研究報告, SP2006-89, 2006.

[5] 上坂 吉則, “パターンの構造を考慮に入れた学習理論について,” 電子情報通信学会論文誌 D-II, Vol.J82-D-II, No.4, pp.575–582, 1999.

[6] 山梨 正明, 認知言語学原理, くろしお出版, 2000.

[7] 池谷 裕二, 進化しすぎた脳, pp.55, ブルーバックス, 2007.

[8] 山下 洋一, 岩橋 大輔, 溝口 理一郎, “基本周波数パターンを利用したキーワードスポッティング,” 電子情報通信学会論文誌 D-II, Vol.J81-D-II, No.6, pp.1065–1073, 1998.