音声言語情報処理 29－1
(1999. 12. 20)

[特別招待論文]

音声認識研究の課題

中川 聖一

豊橋技術科学大学 情報工学系
〒441-8580 豊橋市天伯町字雲雀ヶ丘1-1
Tel. (0532)44-6759
E-mail: nakagawa@slp.tutics.tut.ac.jp

あらまし 始めに、現在の機械による音声認識能力がまだ人間の能力に及ばないことを述べ、特に音響モデルの改善が必要なことを論じる。情報理論の観点から音響モデルと言語モデルとのバランスについて述べる。次に音響モデルの中心技術となっている隠れマルコフモデル (HMM) の原理と課題を述べる。最後に言語モデルの音声認識における役割は、発声されえない認識候補の除外、言い換えればエントロピーを小さくするモデルが好ましいという観点から統計的な言語モデルとその課題を述べる。

キーワード 音声認識、音響モデル、HMM、言語モデル、パープレキシティ、N-gram

Some Problems on Automatic Speech Recognition

Seiichi Nakagawa

Department of Information and Computer Sciences, Toyohashi University of Technology,
Tenpaku-cho, Toyohashi, 441-8580, Japan
Tel. (0532)44-6759
E-mail: nakagawa@slp.tutics.tut.ac.jp

Abstract

In this paper, we discuss some problems on automatic speech recognition. First, we state that the current art of speech recognition technology is inferior to the ability of human beings, especially, on phoneme recognition (perception) performance. Therefore, we point out the importance of acoustic models. Next, we describe a blance a between an acoustic model and a language model from a view point of information theory. Finally, we focus on research trends on statistics-based language model and we describe the limits and future works.

key words speech recognition, acoustic model, HMM, language model, N-gram

1 はじめに

表 1 は現在の機械による音声認識能力と人間の能力を比較したものである [1][2]。表 1 から、分析合成音の特徴パラメータでも音声認識には有効な特徴は十分保存していると言える。しかし、LPC 係数と LPC ケプストラムは相互に変換可能だが、認識パラメータとしては、前者は悪く、後者は比較的良い。このことからも、認識系と整合性の良い (認識時に非線形処理の必要のない) 特徴パラメータの抽出・発見が必要であることがわかる。また、LPC メルケプストラムよりも MFCC の方が良い。これは、MFCC の方が大まかな特徴をとらえているため不特定話者に頑健なためだと思われる。表 2 は音響モデルと言語モデルの能力を示している (機械の言語モデルは認識対象のドメインが限定されている場合)。

表 1: 人間と機械の音声認識（知覚）能力の比較

タスク	人間の誤聴率	機械の誤認識率
アルファベット	1.6% (連続発声)	5% (孤立発声)
連続数字列 (ストリング認識)	0.009% 0.105%(分析合成音)	0.72%
1000 語彙の 自然言語文 (文法未使用時の 単語認識)	2%	17%
5000 語彙の 自然言語文 (朗読音声)	S/N 10dB 1.1% S/N 16dB 1.0% S/N 22dB 0.9% クリーン音声 0.9%	12.8% 10.0% 8.4% 7.2%
スイッチコーパス (自然発話)	4%	30～43%
20000 語彙の 自然言語文 (朗読音声)	2.6% 7.4% (非母語話者)	12.6%

表 2: 音響レベルと言語レベルの性能

	音響レベル (音韻認識率)	言語レベル (単語パープレキシティ)
人間	90%～95%	100
機械	70%～80%	70～200

図 1 は音韻（音素）認識率と単語単位のパープレキシティ (2 のエントロピー乗、認識対象カテゴリ数に対応)、単語認識率 (1 単語は 6 音節からなっていると仮定)、文認識率 (≒ 単語認識率の 8 乗、1 文は 8 単語からなっていると仮定) の関係を示している [3][4][6]。また、これより、10～20%のパープレキシティの減少と 1～2%の音韻認識率の向上は、同等の効果のあることがわかる。これらの結果により、言語モデルの改良により 20～30%のパープレキシティの減少を図ることは効果が大きいが、現在の主流である trigram 言語モデルとそのタスク (トピック) による適応化によりほぼ限界に達していることがわかる。一方、音韻認識率は人間の能力にまだ遠く及ばない。特に雑音環境下やマイクロフォンなどの通信回線の違いによる機械の能力は人間と比べて極端に悪い。

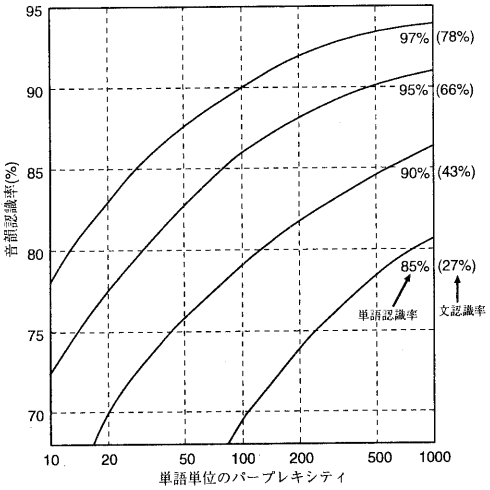

図 1: 音韻認識率と単語・文認識率の相互関係
（1 単語は平均 6 音韻、1 文は平均 8 単語から構成されていると仮定）

言語モデルで用いられる n-gram は (n-1) 重マルコフモデルのことで、単純マルコフモデルに変換でき、隠れマルコフモデル (HMM) に包含される。また、線形予測分析 (LPC) は音声を定常ガウス過程と仮定しており、p 次の分析は p 重マルコフ過程であり、これも隠れマルコフモデルに包含される。しかし、包含している抽象度の高いモデルの方が良いとは必ずしも言えず、対象に特化したモデルを探求する必要がある。例えば、トライグラムや可変グラムの言語モデルを隠れマルコフモデルで表現するためには、膨大な状態数が必要であり、パラメータの推定は非常に難しくなる。

会話文のような自然な発話 (spontaneous speech) に対しては音声認識率は極端に劣化する。これは、自然な発話の言語モデルの構築が難しいこともさながら、発声がいあいまいになることが主因である [7]。例えば、表 3 に示すように自然発話音声の音韻間距離 (HMM のモデル間の Bhattacharyya 距離)、特に母音間距離が小さくなり、発話速度の変動も大きくなる。ましてや、雑音環境下や伝達特性の異なる環境下での音声認識は難しく、このことが音声認識の実用化の妨げになっている。自然な発話に関しては発音辞書の研究も重要となろう。

表 3: 発話スタイルの違いによる音響的特徴の相違

発話スタイル	母音間距離 (分散)	子音間距離 (分散)	フレーム数／音韻 (分散)
孤立単語発声	5.32 (5.75)	5.09 (5.80)	16.7 (42.5)
朗読	3.63 (2.38)	3.71 (4.01)	14.1 (29.1)
自然発話	2.62 (0.43)	4.21 (4.99)	17.3 (125.0)

2 音響モデルと言語モデルのバランス

2.1 音声認識の確率モデル

今、$Y = y_1, y_2, \cdots, y_T$ を音声時系列パターンとしよう。ここで、y_i は第 i 時間区分 (第 i フレームという) の音声の特徴を表わす特徴ベクトル (通常はスペクトル包絡を表現するパラメータ集合) である。このとき、Y を観測して、単語列 $W = w_1, w_2, \cdots, w_n$ (音韻列、音節列と考えてもよい) を見い出す問題を考える。このとき $P(W|Y)$ を最大にする W を見い出すのが妥当であろう [5]。

$$P(W|Y) = P(Y|W) \cdot P(W)/P(Y) \qquad (1)$$

であるから、$P(Y|W), P(W), P(Y)$ が求まればよい。ここで、$P(Y)$ は最適化しようとしている W とは無関係であるから考慮しなくてよい。$P(Y|W)$ は音響・音声モデルと呼ばれ、通常 HMM でモデル化される。$P(W)$ は W の事前生起確率であり、認識対象の言語モデル (文法など) から計算できる。実際のインプリメントに於いては、音響モデルの尤度と言語モデルの尤度のバランスを考慮して、言語重み λ を導入して、

$$\hat{W} = \underset{W}{\operatorname{argmax}}\{\log P(Y|W) + \lambda \log P(W)\} \qquad (2)$$

とする。また、挿入誤り率や脱落誤り率を制御するためにペナルティ δ を導入した次式がよく使われる。

$$\hat{W} = \underset{W}{\operatorname{argmax}}\{\log P(Y|W) + \lambda \log P(W) + n\delta\} \qquad (3)$$

ここで、n は W を構成する言語モデルの単位数である。

2.2 相互情報量と言語重み

言語重み (とペナルティ値) は値を種々変えて、最適な認識率が得られる値を使う場合が多い。

音響モデルの尤度は多次元生起分布の混合で表現する確率密度値であり、n グラムを用いる言語モデルの尤度は離散確率分布の値であり、大幅にダイナミックレンジが異なる。確率密度値を多次元の単位体積を乗じて確率値に変換してもその対数値は一定の値が加算されるだけであり、両者の尤度のバランスには無関係である [22]。まず、音響モデル (単語単位) の相互情報量を求めよう。

$$\begin{aligned}
I(V;Y) &= H(V) - H(V|Y) \\
&= H(Y) - H(Y|V) && (4) \\
&\fallingdotseq H(V) - \sum_i H(C_i|Y_i) + \alpha && (5) \\
&\fallingdotseq H(Y) - \sum_i H(Y_i|C_i) && (6) \\
I(C;Y) &= H(C) - H(C|Y) = H(Y) - H(Y|C) && (7) \\
I(V;L) &= H(V) - H(V|L) && (8)
\end{aligned}$$

ここで、$W = V_1 V_2 \cdots V_n$, $V_i = C_{i1} C_{i2}, \cdots, C_{im}$。$C$ は音韻、V は単語を表わす。語彙サイズを 20000 語とすると、言語情報を用いない場合は $H(V) = log 20000 = 14.3$ ビット／単語となる。また、日本語の音韻数を 30～40 とすると $H(C) = 5$ ビット／単語となる。もし、音韻の認識率を 70% とすると、音韻単位のパープレキシティは $1/0.7 = 1.4$ となる。これは $H(C|Y) = log 1.4 = 0.5$ ビットに対応する。厳密に言えば、等確率にすべての競合カテゴリに誤るか、偏った誤り方をするかによって $H(C|Y)$ の値が異なり、$0 < H(C|Y) < 0.88$ である[3]。これから、$I(V;Y) = 14.3 - 0.5 \times 6 = 11.3$ ビット／単語となる。もし、音響レベルで単語辞書を用いないなら (通常の大語彙連続音声認識システムでは、単語辞書の情報は音響レベルに組み込まれている)、$I(C_i;Y_i) = 5 - 0.5 = 4.5$ ビット／音韻となるから、$I(V;Y) = 4.5 \times 6 = 27$ ビット／単語となる。一方、言語モデルとして trigram を用いる場合、パープレキシティは約 100 とすると、$H(V|L) = 6.5$ ビット／単語となる。両者の相互情報量の比較から、音響モデルによって得られる情報量は言語モデルによって得られる情報量よりも相当多いことがわかる。

もともと、(1) 式には言語重み λ は入っていない。確率論から考えれば言語重みは 1 である。例えば、ケプストラム係数と△ケプストラム係数を用いて音響尤度を求める時、それぞれの尤度の値を等しく扱う場合が通常最も良い結果が得られる。言語モデルの精度が高い程、言語重みを大きくすれば良さそうに思われるが、そうする必要はない。例えば、言語モデルの究極としてパープレキシティが 1 となると、言語重みに関係なく、予測される単語以外の出現確率は 0 となるので、必ず正しく認識できる。逆に、ランダムに単語を予測する言語モデルを用いるパープレキシティは語彙サイズに等しく、言語モデルによって認識システムは何等影響を受けない (但し、ペナルティ項の制御は必要)。

言語重みを導入する必要があるのは、次の 2 点の理由による。(1) フレームシフト幅、(2) 音響パラメータの冗長性。(1) 式はある音韻の認識は観測ベクトル y_i で得られるとしている。実際には、数フレームから十数フレームが一つの音韻に対応しており、その分だけ尤度が加算されていることになる。例えば、単純にフレーム幅を半分にすると音響尤度は 2 倍になる。また、LPC メルケプストラム係数と MFCC を同時に使用すると、認識精度はそれ程向上しないとは思われるが、尤度はほぼ 2 倍になると考えられる。このように、特徴パラメータの集合には、冗長な情報が独立に用いられている。これらの 2 点を正規化するために言語重みが必要となる。その値は次式によって推定される。

$$\lambda = (\text{音韻当りの平均フレーム数}) \times \frac{\text{音響モデルによる相互情報量}}{\text{真の音響モデルの相互情報量}} \qquad (9)$$

通常、フレームシフト幅は 10ms 前後なので第 1 項は約 5～10 程度、第 2 項はケプストラム、△ケプストラム、△△ケプストラムなどを用いると約 2 ぐらいであると推定され、$\lambda = 15$ 前後が適当であろうと予想される。事実、我々の認識システムの音響モデルでは $H(y) = 723.5$ ビット／状態 (ガーベッジモデル)、$H(y_i|C_i) = 714.7$ ビットであり、$I(y_i;C_i) = 8.7$ ビット／音韻となるから、理論的予想値 4.5 ビット／音韻の約 2 倍であった。

発声単語列を W_o、誤認識結果の単語列を W_e としよう。もし、

$$P(A|W_o) > P(A|W_e) \text{ かつ } P(W_o) < P(W_e) \quad (10)$$

なら、誤りの原因が言語モデルの不備をとらえ、このような誤り例が

$$P(A|W_o) < P(A|W_e) \text{ かつ } P(W_o) > P(W_e) \quad (11)$$

となる場合よりも多ければ、音響モデルよりも言語モデルの改良が重要であると解釈するのは正しくない [23]。正しく認識された事例で、

$$P(A|W_o) > P(A|W_e) \text{ かつ } P(W_o) < P(W_e) \quad (12)$$

となる方が、

$$P(A|W_o) < P(A|W_e) \text{ かつ } P(W_o) > P(W_e) \quad (13)$$

となる場合よりも圧倒的に多い (はず) だからである。(10) 式の場合でも $P(A|W_o) >> P(A|W_e)$ となるように音響モデルを改善するのが得策かも知れない。あくまでも音響モデルと言語モデルの改良には限界があるのである。どちらかがより多くの改良の余地があるかを見極めるのが大切である。

3 音響モデル

— HMM の原理と改良 —

$P(Y|W)$ を算出する音響モデルは以下のようにモデル化される（W は省略する）。

$$\begin{aligned}
&P(y_1 \cdots y_t) \\
&= \sum_x P(y_1 y_2 \cdots y_t, x_1 x_2 \cdots x_t) \\
&= \sum_x P(y_1 y_2 \cdots y_t | x_1 x_2 \cdots x_t) P(x_1 x_2 \cdots x_t) \\
&= \sum_x \prod_i P(y_i | y_1 y_2 \cdots y_{i-2} y_{i-1} x_1 x_2 \cdots x_{i-1} x_i) \\
&\quad \times P(x_i | x_1 x_2 \cdots x_{i-1}) \qquad (14) \\
&\simeq \sum_x \prod_i P(y_i | y_{i-3} y_{i-2} y_{i-1}, x_{i-1} x_i) P(x_i | x_{i-1}) \\
&= \sum_x \prod_i \frac{P(y_{i-3} y_{i-2} y_{i-1} y_i | x_{i-1} x_i)}{P(y_{i-3} y_{i-2} y_{i-1} | x_{i-1} x_i)} P(x_i | x_{i-1}) \\
&\qquad (15) \\
&\simeq \sum_x \prod_i P(y_i | x_{i-1} x_i) P(x_i | x_{i-1}) \qquad (16)
\end{aligned}$$

(16) 式が隠れマルコフモデルと呼ばれるもので、第一項が出力確率、第二項が状態遷移確率である。式の変形から明らかなように隠れマルコフモデルは原式からみれば第 1 次程度の近似でありよくないことがわかる。それにもかかわらず、パラメータ数が比較的少なく学習が容易であることから、広く一般に用いられてきた。第一項の近似の欠点を補うために動的特徴パラメータが付加され、第二項の近似の欠点を補うために継続時間制御モデルが導入されている。

我々の研究では、より近似の少ない条件付き出力確率 $P(y_t | y_{t-3} y_{t-2} y_{t-1} x_{t-1} x_t w_k)$ よりも $P(y_{t-3} y_{t-2} y_{t-1} y_t | x_{t-1} x_t w_k)$ を用いるのが従来のどの方法よりもよいことを見出している [8]。つまり、4 フレームのセグメントの統計量を用いるのが有効である。なお、小林らは (9) 式を次式のように変形し、状態遷移確率を観測ベクトルで条件付けると効果が大きいことを示している [24]。

$$P(y_t | y_{t-1} x_t) \cdot P(x_t | x_{t-1} x_{t-2}) \qquad (17)$$

Ostendorf らのセグメントモデルによる音声認識は次のように定式化される [9]。

$$\begin{aligned}
&P(y_1 y_2 \cdots y_T | w_1 w_2 \cdots w_N) \\
&= \sum_{l,N} P(y_1^T, l_1^N | w_1^N) \\
&= \sum_{l,N} P(y_1^T | l_1^N, w_1^N) P(l_1^N | w_1^N) \qquad (18)
\end{aligned}$$

$$P(y_1^T, l_1^N | w_1^N) = \prod_{i=1}^{N} P(y_{t(i-1)+1}^{t(i)} | l_i, w_i) \qquad (19)$$

$$P(l_1^N | w_1^N) = \prod_{i=1}^{N} P(l_i | w_i, l_{i-1}, w_{i-1}) \qquad (20)$$

ここで、l_i は単語（音韻）w_i の時間長である。l_i をさらにいくつかの区間に分けてモデル化すると、この区間は状態に対応し、単純化した場合は HMM と同様のモデルになる。(20) 式はセグメンテーションの確率を表しており、認識に重要であると言われている [10]。HMM の状態 i に割り付けられる観測ベクトル $y_1 y_2 \cdots y_T$ のフレーム間の相関を混合多項式回帰モデルでモデル化する Deng らの方法は次式で与えられる [11]。

$$y_t = \sum_{r=0}^{R} B_{i,m}(r)(t - \tau_i)^r + N_t(0, \Sigma_i) \qquad (21)$$

ここでは、i は状態、m は混合分布のインデックス、τ_i は状態に割り付けられた最初の時間を表わす。通常の HMM は R=0 に相当する。R=0 よりも R=1 の方が、R=1 より R=2 の方が良い結果が得られている。このモデルにおいても 1 次回帰係数の併用は効果がある。但し、コンテキストに独立な音韻モデルでは効果があるが、コンテキスト依存の音韻モデル（トライフォン）などではパラメータが多くなり過ぎ、その効果は小さい [12]。

HMM やセグメントモデルを拡張した次の確率線形動的システムの適用が試みられている [13]。

$$\begin{aligned}
x_{t+1} &= F_t x_t + G_t w_t \qquad \text{（状態方程式）} \\
y_t &= H_t x_t + v_t \qquad \text{（観測方程式）} \quad (22)
\end{aligned}$$

ここで、x_t は非観測の状態ベクトル（連続値）、y_t は観測ベクトル、w_t, v_t は x_t, y_t と無相関な平均値 0 のガウス確率変数ベクトル、F_t は状態遷移行列、G_t は駆動行列、H_T は観測行列である。一般に $\{x_t\}$ はガウス・マルコフ過程で $\{y_t\}$ はガウス過程であるがマルコフ過程でない。通常の HMM や AR モデルは上記の特殊例（時不変システム）としてモデル化される。

本モデルは、統計的セグメントモデルのセグメント内

の相関を考慮した方法の近似モデルになっており、△ケプストラムよりも効果は小さいが、その併用効果はある [13]。なお、最近のこの手法に関連する研究発表はなく、モデルの精密化（複雑化）の割には効果が少ないということによると思われる。

この他、HMM を包含するモデルとして、連続確率文脈自由文法、マルコフランダム場、動的ベイジアンネットワークなどが試みられているが、HMM を有意に上回るには至っていない。

継続時間制御のためには、各状態ごとの継続時間分布をガウス分布・ガンマ分布、ポアソン分布、多項分布（離散分布）でモデル化する方法が試みられてきた。より厳密には、各状態毎と各モデル毎の 2 種類の分布、およびモデル間の分布の相関等も用いる必要がある。なお、状態遷移確率を状態の滞在時間長によって可変にするモデルが提案され状態継続時間分布を用いるよりもよい結果が報告されている。また、出力確率を状態継続時間長に依存する方法も提案されている。

4 言語モデル

現在の音声認識技術では、音声認識結果のあいまいさは避けられず、音響レベルだけでは候補となりうる可能な認識結果は無限と考えてよい。探索空間を小さくするということは、よい言語モデルを構築することであり、よい言語モデルとは認識対象の文集合のエントロピー (パープレキシティ) を小さくするモデルである (カバーレイジが大きいことが必要なことは言うまでもない。但し、通常カバーレイジが小さいとエントロピーは大きくなるが、受理できない文はエントロピーの計算に使わない場合もあるので注意を要する)。このことから、認識対象の文集合に存在する統計的性質を利用するのが自然である。これが言語の確率モデルである。音声ワープロのように書かれた文を音声で入力する場合は、文法を中心とした文字言語の言語モデルで充分と思われるが、話し言葉である音声言語では、言い淀み、言い直し、助詞落ち、倒置、間投詞の挿入など、文字言語では非文扱いとなる文が頻繁に生じるため、言語モデルの構築は極めて難しい。例えば、「観光案内」という限定されたタスクにおいてさえ、これに関して発声される文集合を受理する文脈自由文法を作成することは非常に難しく、カバーレイジは約 40%～80%程度であり、パープレキシティも 100 前後と大きくなる。この場合でも、数百文の学習データから単語単位の (クラス) バイグラムを用いるとカバーレイジも大きくパープレキシティも小さくなる [14]。

形態素や単語の定義によりパープレキシティの値が異なるが、3gram と 4gram の差はあまりないこと、3gram を改良して定型表現や 4gram を導入してもパープレキシティの減少は約 10%であり、 1 節で述べたように、この程度の改善は単語認識率の向上にあまり貢献していないと考えられる。

一方、言語モデルの改良として、トピックや直前のテキストによるモデルの適応化 (ドモインモデルの推定) が試みられており、これらにより、パープレキシティを 10%～ 20%減少させることができる。また、言語モデル構築用のテキストデータ量の増大も効果が大きい [15]。

アメリカの経済雑誌の 3900 万語 (語彙サイズは 2 万語) で言語モデルの学習をした場合、テスト文に対して、パープレキシティは単語単位のバイグラムで 163、トライグラムで 97 (400 万語での学習ではそれぞれ 205, 153) 、2000 個のクラス (意味的、構文的に同じ単語の集合を自動的にクラスタリングしたもの) を用いたバイグラムで 187、トライグラムで 114、単語トライグラムとクラストライグラムの内挿で 93 である [16]。さらに、30 個のトピックに分類し、各トピック別のトライグラムの内挿 (内挿率は直前の 500 単語で決定) で適応化するとパープレキシティは 77 に減少し、直前の文集合を用いて適応化 (キャッシュ) すると 83 に、両者を併用すると 71 に減少する。これにトライグラムの代わりにより長い 4～5 グラムを一部用いると (varigram) さらに 67 に減少する [16]。今後は新しいドメイン・タスクに対する言語モデルの構築法が重要な研究となろう。

日本語の言語モデルはテキストを形態素分析で形態素に分解されてから統計処理を行なう。この形態素解析が 100%でないことや読みの付与が完全でないこと、形態素単位がアルゴリズムによってまちまちで、一律な比較が困難なことなどの問題点が挙げられる。読みに関しては、複数の読みの確率を辞書に与えたり [17]、形態素と読みのペアを単位とする [18] 対処法がある。また、句読点や息つぎを 1 単位として言語モデルを構築するのも効果がある [19]。これは、話し言葉では文の切れ目が音声データからでは検出が難しいためである。また、新聞やニュースでは、話題が時間的に推移するので時期に依存したモデルの適応化も効果がある。

先に述べたように、句構造文法のような規則に基づく方法では、話し言葉のような多様な言語現象をカバーし非文を排除することは極めて難しい。当然、規則に確率を付与する確率文法が考えられるが、元となる規則を作成する困難さは変わらない。そこで、規則自体を学習する試みがなされている。

自然言語の比較的良いモデルと言われている文脈自由文法（CFG）に対しては、大量データからの確率文脈自由文法の学習法は Inside-Outside アルゴリズムとして知られている（確率文脈自由文法よりも記述能力の低い確率正規文法に対しては、ほぼ HMM と等価と考えて良い [5]）。しかし、この学習は繰り返し計算量が多いこと、局所的最適解に収束することなどにより難しい。CFG の規則の確率化よりも、導出途中の部分木の出現確率を用いる方が解析精度は向上するが、学習データとして大量の解析木を必要とし、音声認識への適用化は難しい [20]。

係り受け構造の制約は文脈自由文法で記述できる [5]。例えば、チョムスキーの標準形である $A \rightarrow BA$ の規則は B の最語尾文節が A の最語尾文節に係かり、全体として A の係り受けのための「受け」の性質を保存すると解釈する。このように規則の形を制限して Inside-Outside アルゴリズムによって確率文脈自由文法を学習する。EDR のテキストコーパスに対して 3gram にはおよばないが、3gram に近いパープレキシティが得られている [21]。

言語モデルは音響モデルと比べれば、モデル化が進んでいると言える。例えば、国会中継の答弁の音声認識は非常に難しいが、言語モデルは議事録から 3gram を作成し、答弁と同じ内容を改めて読み上げた朗読音声に対しては、ニュース音声のアナウンサーの音声部分の

認識と同程度の認識率が得られている（答弁の TV 音声に対しては単語認識率は 46%、朗読音声に対しては 80%）[19]。このことは、音響モデルの精度が不十分であることを如実に示している。認識率を向上させる最も単純な方法は学習データ量を増やすことである [15]。実用システムの構築のためには数百人の音声データを用いる必要がある。

5 むすび

静かな部屋で比較的ていねいに発声すれば高精度に音声認識ができるようになってきた。しかし、誰もが動きながら自由に発声した音声を認識するためには、まだまだ基礎研究が必要である。これには、音声分析や特徴パラメータ及び言語モデルの研究よりも音響モデル (動的特徴パラメータの扱い) と認識アルゴリズムの研究が重要だと考えられる。これには、人間の知覚過程の解明が手がかりを与えてくれるであろう。なぜなら、現在主に用いられる特徴パラメータから音声を再生すれば、我々はほぼ正しく認識理解できるし、ある程度タスク (ドメイン) が限られていれば、人間の単語予測能力と n-gram の適応化による単語予測能力にはそれ程大きな差がないからである。視覚のパターン認識能力は優れてはいるが、スペクトラムリーディングは聴覚の能力よりかなり悪い。視覚はパターン認識技術と比較的直結している。このことからスペクトラム時系列の非線形変換による特徴抽出が重要であると予想される。

なお、紙数の関係で、実環境下における音声認識の問題点や参考文献をすべて挙げることはできなかった。筆者のサーベイ論文 [25] を参照されたい。

参考文献

[1] S.J. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.L. Gauvain, D.J. Kershaw, L. Lamel and D.A. Leeuwen : "Multilingual large vocabulary speech recognition in the European SQALE project, ", Computer Speech and Language. 11, 73-89 (1997)

[2] R.P. Lippmann : "Speech recognition by machine and humans,", Speech Communcation. 22, 1-15 (1997)

[3] 中川聖一："情報理論の基礎と応用", 近代科学社 (1992)

[4] 中川聖一："統計的言言語モデルの可能性と限界", 日本音響学会講演論文集, 1-6-11 (1998)

[5] 中川聖一："確率モデルによる音声認識", 電子情報通信学会, (1988)

[6] 中川聖一："パターン情報処理", 丸善, (1999)

[7] 山本一公, 岩井直美, 中川聖一："発話スタイルの違いが音声認識に及ぼす影響についての検討", 電子情報通信学会, 音声技報, SP99- (1999)

[8] 中川聖一, 山本一公："セグメント統計量を用いた隠れマルコフモデルによる音声認識", 電子情報通信学会論文誌, Vol.J79-DII. No.12, pp.2032-2038 (1996)

[9] M. Ostendonf, V.V. Digalakis and O.A. Kimball : "From HMMs to segment models : a unified view of stochastic modeling for speech recognition,", IEEE Trans. Speech and Audio Process, Vol.4, No.5, pp.360-378 (1996)

[10] J. Verhasselt, I. Illina, J.P. Martens, Y. Gong, J-P. Haton : "Assessing the importance of the segumentation probability in segment-based speech recognition,", Speech Communication, Vol.24, pp.51-72 (1998)

[11] L. Deng and M. Aksmanovic : "Speaker-independent phonetic classification using hidden Markov models with mixtures of terend functions,", IEEE Trans, Speech and Audio Process, Vol.5, No.4, pp.319-324 (1997)

[12] R. Chengalvara and L. Deng : "Speech trajectory descrimination using the mimimum classfication error learning,", IEEE. Trans. Speech and Audio Process. Vol.6, No.6, pp.505-515 (1998)

[13] V.Digalakis, J.R. Rohlicek and M. Ostendorf : "ML estimation of a stochastic linear system with the EM algolithm and its application to speech recognition,", IEEE Trans. Speech Audio Process. 1, pp.431-442 (1993)

[14] 中川聖一, 大谷耕嗣："bigram の使用による話し言葉用確率文脈自由文法の自動学習", 情報処理学会論文誌, Vol.39, No.3, pp.575-584 (1998)

[15] V. Steinbiss et al : "Continuous speech dictation-from theory to practice", Speech Communication, Vol.17, pp.19-38 (1995)

[16] S. Martin, J. Liermann and H. Ney : "Adaptive topic-dependent language modeling using word-based varigrams,", Proc. Eurospeech, pp. 1447-1450 (1997)

[17] 高木幸一, 古井貞："形態素の読みの確率を考慮したニュースの音声のディクテーション", 日本音響学会春季大会, 1-6-5 (1998)

[18] 桜井直之, 古井貞, 大附克年："ニュースの音声認識における言語モデルの改良", 日本音響学会春季大会, 2-1-3 (1999)

[19] 今井亨, 斉藤洋平, 安藤彰男, 古井貞："連続発話認識のための言語モデル", 日本音響学会春季大会, 2-1-6 (1999)

[20] R. Bod : "Spoken dialogue interpretation with the DOP model,", Proc, ACL, pp.138–144 (1998)

[21] 堀智織, 加藤正治, 伊藤彰則, 好田正紀："確率文脈自由文法を用いた言語モデルの構築と音声認識実験による評価", 電子情報通信学会, 音声技報, SP99-37 (1999.6)

[22] 武田一哉："音声確率と言語確率の統合について", 情報処理学会, 音声言語情報処理, SLP 22-14 (1998.7)

[23] 南条, 李, 河原："大語彙連続音声認識における認識誤り原因の自動同定", 情報処理学会, 音声言語情報処理, SLP 27-6 (1999.7)

[24] 古山純子, 小林哲則："平滑化部分隠れマルコフモデルによる音声認識", 情報処理学会, 音声言語情報処理, SLP 28-5 (1999.10)

[25] 中川聖一："音声認識研究の動向", 電子情報通信学会論文誌, Vol.83-DII, No.2 (2000)