音声確率と言語確率の統合について

武田一哉
名古屋大学 大学院工学研究科
email: takeda@nuee.nagoya-u.ac.jp

1. はじめに

HMMとNグラム言語モデルを利用する，確率的な枠組みによる大語彙連続音声認識技術は，PC上で動作するディクテーションシステムを実現するに至った。より難度が高いと考えられている音声対話システムにおいては，様々な知識源から得られる情報を合理的に統合することが不可欠であり，当面は確率に基づく情報の統合が最も見通しの良い方法の一つであることは間違いない。そこで本稿では，異なる知識源から得られる情報の統合という観点から，現在の連続音声認識における音声確率と言語確率の統合方法の問題点をいくつか指摘する。

2. 音響モデルと言語モデルの統合

統計的な連続音声認識の基本原理は，音響特徴ベクトル系列 $\mathbf{A}=\mathbf{a}_1\mathbf{a}_2\cdots\mathbf{a}_m$ が観測されたという条件の下で，単語系列 $\mathbf{W}=w_1w_2\cdots w_n$ が観測される条件つき確率

$$P(\mathbf{W}|\mathbf{A})=\frac{P(\mathbf{A}|\mathbf{W})\cdot P(\mathbf{W})}{P(\mathbf{A})} \quad (1)$$

を最大化する $\mathbf{W}$ を求めることである。

実際のシステムでは，$\mathbf{W}$ の最大化に関与しない分母を無視した対数確率値，

$$\ln P(\mathbf{A},\mathbf{W})=\ln P(\mathbf{A}|\mathbf{W})+\ln P(\mathbf{W}) \quad (2)$$

に換えて，以下の対数確率（密度）値の重みづけ和を $\mathbf{W}$ の「スコア」とすることが一般的である。

$$\ln \tilde{p}(\mathbf{A},\mathbf{W})=\ln p(\mathbf{A}|\mathbf{W})+\alpha\{\ln P(\mathbf{W})+nq\} \quad (3)$$

α は言語重み，q は単語挿入ペナルティー，と呼ばれるパラメータであり，それぞれ実験的に最適な値が定めることで，HMMの対数出力確率と，言語モデルにより計算される対数言語確率とを，単に加えた場合に比べて非常に高い認識精度を得ることができる。

このことは，それぞれ最尤基準に基づいて一貫して学習された確率モデルである，HMMとNグラムモデルとが，それらを「統合する」という観点からは必ずしも最適に学習されていない，ということを意味している。以下ではその原因として考えられる要素を3点指摘する。

3. 連続分布と離散分布

通常のHMMでは，出力確率の密度関数として混合正規分布が用いられる。観測された音声特徴ベクトル $\mathbf{a}$ が出現する「確率値」は，$\mathbf{a}$ を取り囲む微小領域における確率「密度関数」の積分値として与えられる。しかし，通常の音声認識システムでは，音響「確率」として $f(\mathbf{x})$ を直接利用することが一般に行われている。すなわち，

$$\int_{\mathbf{a}} f(\mathbf{x})d\mathbf{x}\approx f(\mathbf{a})\Delta\mathbf{x}$$

のように書き換えて考えると，確率計算において $\Delta\mathbf{x}$ が考慮されていないことが分かる。今 K をベクトルの次元数，L を音響特徴ベクトル系列の長さとして，音響確率の計算結果

$$\begin{aligned}\ln P(\mathbf{A}|\mathbf{W})&\approx\ln\left[f(\mathbf{A}|\mathbf{W})\cdot\Delta\mathbf{x}\right]\\&=K\cdot L\ln\Delta x+\ln f(\mathbf{A}|\mathbf{W})\end{aligned}$$

と，式（3）とを見比べると，確率分布を連続表現することに起因する確率値のレンジの違いは，言語重みによって補正することはできず，音響パラメータの次元数と，特徴ベクトル系列の長さに応じて定まる定数を，スコアに加えることにより統合されたスコアの精度を向上可能なことが示唆される。

4. シンボル数の異なり

HMM，Nグラムはマルコフ過程をモデル化の基本としており，音声，言語確率値は，それぞれの系列の長さに対して単調に減少する。言語重みを導入する必要性を，両者の処理単位であるシンボルの数の違いに求めることも可能である。

音響確率がフレーム単位で計算されるのに対して，言語確率は単語単位に計算される。フレーム毎の平均音響確率値と，単語当たりの平均単語確率値がそれぞれ，p，qである場合，Lフレームの音声が，N単語の単語列から構成される場合の対数確率値は

$$\ln\left\{p^L \cdot q^N\right\} = L\ln p + N\ln q$$

により与えられる。$\ln p$と$\ln q$の値をおおよそ同じオーダーにすることが望ましいと仮定すれば，

$$\ln q' = \frac{N}{L}\ln q$$

なる，補正を施した単語確率値q'を用いれば良い。式（3）と見比べることにより言語重みは，単語列を構成する単語数とフレーム長の比（単語の継続時間）により正規化するために導入されることになる。

この考え方に基づいて，後述する単語挿入ペナルティーと同様な挿入ペナルティーを，音素毎に作用させる方法などが報告されているが，2つの確率値のレンジを同じオーダーにそろえることの意味は，必ずしも明確ではない。

5. 言語確率の事象空間

3節では，音響確率がK次元空間上で連続に「密度関数」の形で定義されていることに起因する問題を議論した。本節では同様の問題がNグラム言語モデルにも存在し，単語挿入ペナルティーに関連していることを示す。

Nグラム言語モデルを用いた場合，等しい数の単語で構成される全ての単語系列の集合を全事象と考えており，この集合全体に対して確率値の合計が1となる。単語系列の実効的な種類数は，系列を構成する単語数nと実効的な語彙サイズである，単語パープレキシティーPPにより，PP^nと求められる。したがって，任意の長さの単語系列全体をを，全事象とする確率モデルとして，

$$P(\mathbf{W}) = \frac{P_{NG}(\mathbf{W})}{\frac{1}{PP^n}}$$

を定義することは，合理的と考えられる。ここで，nは単語列$\mathbf{W}$に含まれる単語数，PPは単語列を生成した言語Λのパープレキシティーを，それぞれあらわす。これは，言語長Λから生成される長さnの typical な単語系列の出現が等確率である場合の系列の出現確率を基準に，Nグラム確率を正規化した確率である。

両辺の対数を取って（3）式と比較することで，単語挿入誤りqが，言語のパープレキシティーPPに対応することが分かる。この，単語挿入ペナルティーと，言語のパープレキシティーとの関係は，「Nグラムにより与えられる言語確率を，真の言語確率に補正する単語挿入ペナルティーが最も高い認識率を与える。」ということを意味しており，この仮説の妥当性は小規模な認識実験により確認されている。

6. まとめ

音声確率と言語確率は，一貫して最尤基準により学習されているにも関わらず，それらを統合する際にはいくつか注意すべき点があることを指摘した。

参考文献

(1) L.R.Bahl el al. "Language-model/acoustic channel balance mechanism", IBM Technical Disclosure Bulletin, 23(7B), pp.3464-3465, Dec.1980

(2) 磯健一:「音素記号と特徴ベクトルの同時出力確率を用いた音声認識」音講論集 2-Q-7, 平成10年3月

(3) A.J.Rubio et al. "On the influence of Frame Asynchronous grammar scoring in a scr system", Proc. of ICASSP 97, vol.1, pp.895-898, April 1997

(4) 小川厚徳他「連続音声認識結果からの言語エントロピーの推定」, 信学技報 SP98－31, pp. 61-66, June 1998