E-043

文音声認識における最適言語重みの予測式
Prediction Formula of Optimal Language Weight on Sentence Recognition

大槻 恭士†
Takashi Otsuki

1．はじめに

現在の文音声認識では，音響モデルに基づく尤度（音響尤度）と言語モデルに基づく尤度（言語尤度）を統合して文仮説の尤度を求めている．その際，両尤度間のダイナミックレンジの差を補正するために，言語尤度に重み（言語重み）を付けることが広く行われている．その導入理由からも明らかなように，最適な言語重みの値が各尤度の分布に依存することは予想に難くない．それでもなお，言語重みの値の最適化手法については，認識実験による探索の域を出ていないのが現状であり，最適言語重みに対する何らかの指針を導くことは，有意義と考えられる．

そこで本稿では，文音声認識の確率モデルを解析し，最適言語重みの予測式を導出する．さらに，実測した尤度差の分布に基づいて予測した最適言語重みの値と，認識実験により求めた最適言語重みの値を比較することで導出した予測式の評価を行う．

2．最適言語重みの予測式

2.1 記号の定義

$X=x_1x_2\cdots x_L$ ：長さ L の正解文

x_i $(i=1,2,\cdots,L)$ ：文 X を構成する単語

$W=w_1w_2\cdots w_L$ ：長さ L の不正解文

w_i $(i=1,2,\cdots,L)$ ：文 W を構成する単語

$S_A(x_i), S_A(w_i)$ ：単語 x_i, w_i の音響尤度

$D_A(x_i,w_i)=S_A(x_i)-S_A(w_i)$ ：単語 x_i, w_i 間の音響尤度差

μ_A, σ_A^2 ： $D_A(x_i,w_i)$ の平均と分散

$S_A(X)=\sum_{i=1}^{L} S_A(x_i)$ ：文 X の音響尤度

$S_A(W)=\sum_{i=1}^{L} S_A(w_i)$ ：文 W の音響尤度

$S_L(X)$, $S_L(W)$ ：文 X, W の言語尤度

$D_L(X,W)=S_L(X)-S_L(W)$ ：文 X, W 間の言語尤度差

μ_L, σ_L^2 ： $D_L(X,W)$ の平均と分散

$S(X)=S_A(X)+\gamma S_L(X)$ ：文 X の尤度

$S(W)=S_A(W)+\gamma S_L(W)$ ：文 W の尤度

γ ：言語重み

2.2 予測式の導出

尤度差の分布に関して以下の仮定を置く．

仮定１ $D_A(x_i,w_i)\sim N(\mu_A,\sigma_A^2)$ で，$(x,w)\neq(x',w')$ のとき，$D_A(x,w)$ と $D_A(x',w')$ は独立である．

仮定２ $D_L(X,W)\sim N(\mu_L,\sigma_L^2)$ で，$(X,W)\neq(X',W')$ のとき，$D_L(X,W)$ と $D_L(X',W')$ は独立である．

また，正解文 X と不正解文 W の間のハミング距離を 1 に限定する．すなわち X と W は 1 つの単語のみが異なる対であるとする．この限定は実際の文認識からかけ離れているが，非常に類似する競合文との識別において最適な言語重みが通常の文認識においても有効であると考えるならば，妥当な限定であるといえる．

X が W に誤る確率は，

$$P[S(X)<S(W)]=P[S_A(X)-S_A(W)+\gamma D_L(X,W)<0]$$

と書ける．ここで，X と W のハミング距離が 1 であることと仮定 1 より，

$$S_A(X)-S_A(W)\sim N(\mu_A,\sigma_A^2)$$

であり，仮定 2 と正規分布の再生性より，

$$S_A(X)-S_A(W)+\gamma D_L(X,W)\sim N(\mu_A+\gamma\mu_L,\sigma_A^2+\gamma^2\sigma_L^2)$$

となる．したがって X が W に誤る確率は，

$$P[S(X)<S(W)]=\Phi\left(-\frac{\mu_A+\gamma\mu_L}{\sqrt{\sigma_A^2+\gamma^2\sigma_L^2}}\right)$$

で与えられる．ここで $\Phi(x)$ は標準正規分布関数

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt$$

である．$\Phi(x)$ は単調増加関数であるから，$\frac{\mu_A+\gamma\mu_L}{\sqrt{\sigma_A^2+\gamma^2\sigma_L^2}}$ を最大にする $\hat{\gamma}$ が誤り確率を最小にする最適言語重みである．そこで，$f(\gamma)=\frac{\mu_A+\gamma\mu_L}{\sqrt{\sigma_A^2+\gamma^2\sigma_L^2}}$ を γ で微分して 0 とおき，γ について解くことで，最適言語重みの予測式

$$\hat{\gamma}=\frac{\mu_L/\sigma_L^2}{\mu_A/\sigma_A^2}$$

を得る．つまり，最適言語重みの予測値は，正解単語とそれ以外の単語との間の音響尤度差 $D_A(x_i,w_i)$ の分布パラメータ μ_A, σ_A^2 と，正解文とそれからハミング距離が 1 離れた文との間の言語尤度差 $D_L(X,W)$ の分布パラメータ μ_L, σ_L^2 より求めることができる．

3．予測式の評価

3.1 認識対象と認識システム

ASJ-JNAS[1]中の IPA-98-TestSet を認識対象とした場合の最適言語重みの予測を行った．認識エンジンは Julius Version 4.1[2]を用い，音響モデルは IPA の日本語ディクテーション基本ソフトウェア最終版[3]付属の男性話者トライフォンモデル（2000 状態，16 混合分布）を用いた．

言語モデルは同ソフトウェア付属の 2-gram モデルから 2 つ（LM1, LM2）と，新たに作成した 2-gram モデル（LM3）1 つの計 3 種類を用いた（表 1）．語彙数はすべて約 2 万語である．学習データ量が少なく細かなチューニングを省いた LM3 は，テストセットパープレキシティが非常に大きく，他に比べて性能の低い言語モデルである．

† 山形大学大学院理工学研究科, Graduate School of Science and Engineering, Yamagata University

表 1:使用言語モデル

名称	学習データ量	Test Set Perplexity
LM1	毎日新聞 75 ヶ月分	122.6
LM2	同　45 ヶ月分	126.8
LM3	同　12 ヶ月分	380.7

表 2:言語尤度差分布のパラメータと予測最適言語重み

名称	μ_L	σ_L^2	$\hat{\gamma}$
LM1	6.03	8.31	11.30
LM2	5.86	7.98	11.44
LM3	4.40	8.45	8.11

3.2　音響尤度差分布の推定

以下に述べる孤立単語音声認識実験により，単語間の音響尤度差の平均と分散である μ_A, σ_A^2 を推定した．

(1) 正解単語系列を用いたビタビアライメント（HTK[4]の HVite コマンドを使用）を行い，IPA-98-TestSet の音声データをラベリングする．

(2) ラベルに従い音声データを単語単位に分割し，前後に無音区間を付加した擬似単語発声データを作成する．

(3) 擬似単語発声データに対し Julius Version 4 から新たに導入された機能である孤立単語認識を行う．このとき，出力候補数を語彙数とすることで，正解単語とそれ以外の単語の間の尤度差を求めることが可能となる．

音響尤度差のヒストグラムの概形を図 1 に示す．仮定 1 に反して，正規分布型の左右対称形ではなく，偏った分布となっており，この分布の平均と分散を用いて最適言語重みを予測するのは問題があると考えられる．

そこで，尤度差 0 の近傍（図 1 の左側）における分布を正規分布で近似することを考える．これは，実際の認識に影響するのは尤度差の小さい領域の分布の形であるという考えに基づいている．本来ならばフィッティングを行うべきところではあるが，ここでは目視により μ_A=130, σ_A^2=45^2 と推定した．図 2 に実測値の累積相対頻度分布曲線と $N(130, 45^2)$ の比較を示す．

3.3　言語尤度差分布の推定

認識対象文と 1 単語のみ異なる文をすべて生成し，認識対象文との間の言語尤度差の分布を推定した．各言語モデルによって与えられる言語尤度は CMU-Cambridge SLM toolkit[5]ライブラリを用いて計算した．

表 2 に 3 つの言語モデルについて推定した，言語尤度差分布のパラメータ μ_L, σ_L^2 を示す．LM3 を用いた場合の言語尤度差の平均 μ_L が他のモデルと比較して小さいことからも，LM3 の性能が他よりも低いことがうかがえる．一方 σ_L^2 についてはモデル間に大きな差異は認められない．

3.4　認識実験との比較

得られた尤度差分布パラメータより求めた，各言語モデルにおける予測最適言語重み $\hat{\gamma}$ の値を表 2 に示す．これらの値を評価するため，IPA-98-TestSet を対象とした文認識実験を行った．認識エンジンと使用モデルに前述のものを用い，挿入ペナルティを 7，ビーム幅を 1500 に固定し，言語重みを 1 刻みで変えて単語認識精度の変化を調べた結果を図 3 に示す．最も単語認識精度が高い言語重みは，LM1 と LM2 の場合 12，LM3 の場合 8 と，表 2 の $\hat{\gamma}$ の値に近い値であり，予測式の妥当性を示唆している．

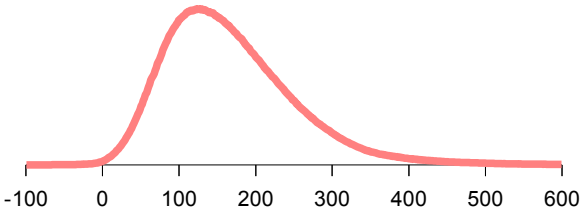

図 1:音響尤度差の分布

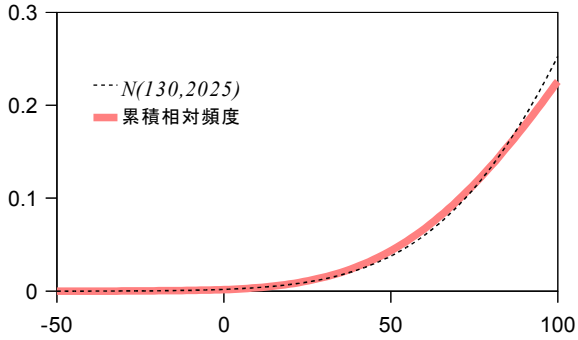

図 2:ゼロ近傍における正規分布近似

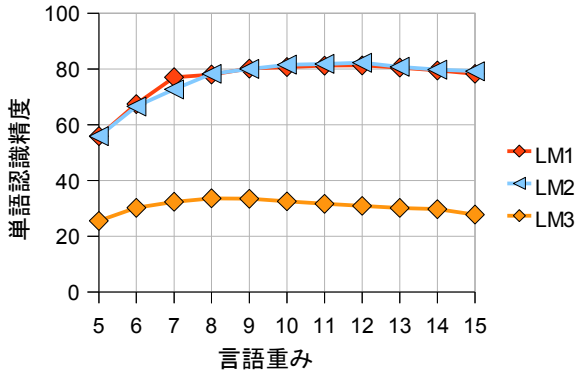

図 3:言語重みと単語認識精度の関係

4. おわりに

文音声認識の確率モデルを解析し，最適な言語重みの予測式を導出した．この予測式は尤度差分布の正規性を拠り所としているため，正規性が成立しない場合への対処法，例えば何らかの基準による比較対象単語の限定や正規分布関数によるフィッティングなどを今後検討する必要がある．また，さらに多くの認識実験との対照を行い，予測式の妥当性を確認する予定である．

参考文献

1) Itou et.al.:Journal of ASJ, **20**, 3, pp.199-206 (1999)
2) http://julius.sourceforge.jp/
3) 鹿野他:音声認識システム, オーム社 (2001)
4) http://htk.eng.cam.ac.uk/
5) http://www.speech.cs.cmu.edu/SLM/toolkit.html