尤度差に基づく n-gram 言語モデル評価のための指標

伊藤彰則[†,☆] 好田正紀[†]

N-gram をはじめとする統計的言語モデルの評価尺度として，パープレキシティやクロスエントロピーがこれまで広く用いられてきた．しかし，ドメイン外テキストを併用する言語モデルや混合言語モデルなどの複雑な言語モデルに関しては，認識システムの単語正解精度とこれらの評価尺度との相関が悪いという結果が近年報告されている．本稿では，パープレキシティに代わりうる評価尺度 LEA について検討した結果を報告する．パープレキシティやクロスエントロピーが評価テキストの単語の出現確率のみを用いるのに対して，ここで提案する指標は，評価テキストに出現する単語の言語尤度と，その単語が出現した文脈における最大言語尤度との差に基づいている．この尤度差から各単語の認識確率を推定し，その平均を算出する．音声認識シミュレーション実験および実音声認識実験の結果と，ここで提案した指標との相関を調べてみたところ，クロスエントロピーに比べて高い相関を示すことが確認された．

A Metric Based on Likelihood Difference for n-gram Language Model Evaluation

Akinori Ito[†,☆] and Masaki Kohda[†]

Perplexity and cross entropy have been widely used as an evaluation metric of stochastic language model. Recently, several papers reported that correlation between these metrics and word error rate was poor when complicated language models were used, such as mixture model. In this paper, a new metric called LEA for n-gram language model is proposed, that is intended to substitute perplexity. The major difference of the proposed metric from perplexity is that, while perplexity utilizes probabilities of word occurences in the evaluation text, the proposed metric accumulates differences of linguistic scores between a word in the evaluation text and the maximum score available in that context. The word recognition probability is estimated using the score difference. Then the average of the probabilities is calculated. Correlation between the proposed metric and word accuracy was investigated against two recognition results through a speech recognition simulator and a real speech recognizer. The result proved that the proposed metric had higher correlation with word accuracy than cross entropy.

1. はじめに

現在の音声認識システムには，n-gram に代表される統計的言語モデルが広く使われている．統計的言語モデルの開発に重要なのがその評価尺度であり，現在広く使われている評価尺度として，パープレキシティと単語正解精度（または単語エラー率）がある．パープレキシティ（テストセットパープレキシティ）は，言語モデルによって付与された評価テキストの単語の出現確率の逆数の幾何平均である．パープレキシティは，数学的な意味が明解であることと，計算が容易であることから，これまで言語モデルの評価に広く使われてきた．しかし，パープレキシティは単語間の音響的な類似性を無視しているため，パープレキシティの改善が必ずしも認識性能の向上に結びつかない例も見られた．一方単語正解精度は，音声認識実験によって実際に得られる性能であり，その言語モデルの実際の性能を測る指標として使われてきた．しかし，一般に音声認識実験には時間がかかるため，多数の言語モデルを比較するような実験において単語正解精度を評価尺度とすることは難しい．そのため，これまではパープレキシティと単語正解精度を併用する形で言語モデルの評価が行われてきた．

パープレキシティと単語正解精度とは，数%の変動はあるにしても，相関が高いものと思われてきた．実

† 山形大学工学部
Faculty of Engineering, Yamagata University
☆ 現在，東北大学大学院工学研究科
Presently with Graduate School of Engineering, Tohoku University

際，新聞記事を対象とした認識においては，パープレキシティと単語正解精度は比較的高い相関を持つ[1]．しかし，タスク適応モデルや混合言語モデル，最大エントロピーモデルなど，比較的複雑な構造を持った言語モデルの場合，パープレキシティと単語正解精度の相関が低くなることが近年指摘されている[2)~4)]．これに対して，パープレキシティに代わる指標がいくつか提案されてきた．

本稿では，そのような尺度の 1 つとして，単語ごとの尤度差に基づく評価尺度を提案する．ここで提案する指標は，評価テキストに出現する単語の言語尤度と，その単語が出現した文脈における最大言語尤度との差に基づいている．この尤度差から単語認識率の推定値を算出し，その単語ごとの平均を算出する．

本稿では，まずパープレキシティについて述べ，またパープレキシティに代わる尺度としてどのようなものが提案されてきたかについて概観する．次に，本稿で提案する尺度の定義を与え，その意味について考察する．最後に，音声認識シミュレーション実験の結果と，実音声認識のリスコアリング実験の結果に基づき，提案した尺度と単語正解精度の相関について調べる．

2. パープレキシティとその問題点

クロスエントロピーは，1 単語が生起する対数確率の相加平均の符号を反転したものであり，次の式で与えられる．

$$H_0 = -\frac{1}{n}\log_2 P(w_1 \dots w_n) \tag{1}$$

ここで，$w_1 \dots w_n$ は評価用のテキストであり，一般には学習用のテキストと独立に与えられる．

パープレキシティ（テストセットパープレキシティ）とは，1 単語の生起確率の幾何平均の逆数であり，次のように与えられる．

$$PP = 2^{H_0} \tag{2}$$

パープレキシティと音声認識における認識率の関係の研究としては，中川らによる研究[5)] がある．この研究では，パープレキシティと文の認識率の関係をシミュレーションによって求めている．しかし，中川らが使っているパープレキシティは，上記のテストセットパープレキシティとは違うことに注意しなければならない．中川らの用いているパープレキシティは

$$H(L) = -\sum_{\mathrm{w}_1^k \in L} \frac{1}{k} P(\mathrm{w}_1^k)\log_2 P(\mathrm{w}_1^k) \tag{3}$$

$$PP = 2^{H(L)} \tag{4}$$

という定義である．ここで，L はモデルの表現する言語，w_1^k は長さ k の単語列である．式 (3) は言語 L に属するすべての単語列について対数確率の平均をとっているのに対し，式 (1) では評価用テキストに出現した単語だけについての平均をとっている．評価用のテキストが非常に大きく，実質的に L の近似と見なすことができるとすれば，式 (1) は式 (3) の近似として使うことができる．しかし，実際には評価テキストの量は言語全体を近似できるほど大きいことは稀である．特に，大語彙連続音声認識においては，言語 L はすべての単語の任意の組合せであるのに対して，評価に用いるテキストは非常に小さいことが多い．評価に用いるテキストの量がどの程度であれば十分かについての明確な基準はないが，多くの場合は評価よりもモデルの学習に多くのテキストをあてるため，評価用のテキストは学習用のテキスト量に比べて数分の 1 から数十分の 1 程度であろう．このような場合，式 (1) と式 (3) とを同一視することはできないので，文献 5) の結論をそのままテストセットパープレキシティに適用することはできないと思われる．

テストセットパープレキシティと単語正解精度の間の相関が悪くなる原因として，いくつかの要因を考えることができる．

(1) パープレキシティは，単語間の音響的な類似性を無視している．
(2) 評価用のテキストの量が十分ではない．
(3) 単語出現確率の増加が，単語正解精度の向上に寄与しない場合がある．

(1) は，言語モデルのみに着目した評価尺度全般に共通した問題点である．音響的な類似性は一般には音響モデルに依存するので，言語モデル単独の評価をする場合には厳密な意味での音響的類似性を導入するのは困難であろう．音素レベルでの類似性を用いる方法[6)] は可能であろうが，評価尺度がかなり複雑になると思われる．

(2) は重大な問題であるが，一般に学習・評価用に利用できるテキストの量は限られており，そのうち多くを学習に回してしまうため，評価のためのテキストを多くするのは難しいであろう．

(3) は，たとえばすでに正解として認識されているような単語の出現確率が改善された場合，パープレキシティは低下するが，認識結果はそれ以上改善されない．同様に，尤度が非常に低くて枝刈りされてしまうような単語の場合，出現確率が多少改善されてもその単語が正解になることは難しいであろう．N-best のリスコアリングを行う場合には，n-best 候補からその単語が落ちていれば，言語モデルでの出現確率をいくら改善し

ても，その単語の正解率を改善することはできない．

3. これまで提案された尺度

パープレキシティに先行する言語モデルの評価尺度として静的分岐数などがあるが[7]，ここでは最近提案されている尺度とその問題点について簡単に述べる．

3.1 SMR-perplexity

中川らは，パープレキシティに代わる尺度として SMR-perplexity を提案している[8]．パープレキシティが単語の出現確率の幾何平均であるのに対して，SMR-perplexity は確率の 2 乗平均である．この尺度が用いる情報はパープレキシティと同一であるが，単語正解精度に対する線型性がパープレキシティに比べて優れている．問題点として，パープレキシティと同じ情報を用いているため，パープレキシティと単語正解精度との相関が非常に低い場合，SMR-perplexity を用いてもそれほど相関が改善されないということがあげられる．

3.2 決定木に基づく指標

Iyer らは，数種類の適応言語モデルについてパープレキシティと単語エラー率の関係を調べ，単語エラー率と相関が高いのは trigram のカバー率であるとしている[2]．しかし，カバー率は言語モデルの構造を反映しないため，Iyer らは単語エラーの改善率を推定するモデルを提案している．このモデルは決定木に基づいており，unigram〜trigram の言語尤度，単語の音素数，付近の単語との尤度差，品詞 n-gram の尤度などの情報を基にして決定木を学習する．この決定木によって計算された尺度は，単語エラー率の差と高い相関を持つ．問題点としては，学習によって決定木を構成しているため，構成された決定木の一般性に問題があるかもしれないという点があげられる．

3.3 M-ref と AWER

Chen らは，言語モデルの良さを評価するための手法を 2 つ提案している[3]．1 つは M-ref と呼ばれる尺度を用いる方法で，もう 1 つは音声認識のシミュレーションを用いる手法である．M-ref は，単語の出現確率と，その単語が正しく認識される確率との関係に着目した尺度である．この 2 つの間に線型な関係があれば，パープレキシティと単語エラー率は高い相関を持つことになるが，実際にはこの 2 つの確率の間には非線型な関係がある．そこで，出現確率から正解確率への非線型な対応を実際のデータから求め，それを使って単語エラー率を推定したものが M-ref である．この手法の問題点は，非線型な対応関係をデータから求めなければならない点であるが，いくつかの言語モデルについて調べた結果では，言語モデル間でそれほど大きな差はない．また，絶対的な出現確率ではなく，正解単語の出現確率との比を用いる方法も提案している．Chen らの提案したもう 1 つの方法は，擬似的に単語ラティスを生成し，そこからの認識シミュレーションを行うことによって，擬似的な単語エラー率（AWER）を推定するものである．いずれの手法によって得られた値も，タスク外のテキストを併用する言語モデルの場合には，実際の実験結果に対してパープレキシティよりも高い相関を持つ．

しかし，同じ文献 3) の中で，Chen らは実際の単語エラー率以外を用いた言語モデルの評価について悲観的な結論を述べている．上記の評価尺度は単語エラー率との間にパープレキシティよりも高い相関を持っているけれども，その評価尺度と単語エラー率との間には，単語エラー率で 1 ポイント以上の幅がある．認識システムを改善する研究では，0.5 ポイントの改善であっても有為な差と見なされることも多く，そのような状況で 1 ポイントも幅のある評価尺度は結局使いものにならないという主張である．ある意味でこれは正論であるが，筆者らはそれに完全に同意するわけではなく，言語モデル独自の評価尺度も重要だという立場である．言語モデルを独自に評価する評価尺度があれば，言語モデルの開発が容易になるだけでなく，mixture model[9] のように決定すべきパラメータを含む言語モデルにおいて，パラメータを決定するための基準としてその尺度を用いることができる．

3.4 複数の指標の組みあわせ

Clarkson らは，パープレキシティとその他いくつかの指標を比較し，それらを組み合わせることで単語エラー率との相関を改善できると報告している[10]．この文献では，各コンテキストでの (1) 単語の確率，(2) 確率が大きい順に全単語をソートした場合の当該単語の順位，(3) そのコンテキストでの単語出現確率のエントロピー，(4) ある閾値よりも確率の大きい単語の個数の 4 つの指標について検討している．ここで，(3) のエントロピーは，

$$H(w_1^{i-1}) = -\sum_{w\in \mathrm{V}} P(w|w_1^{i-1}) \log P(w|w_1^{i-1})$$

で表される．ただし，V は語彙の集合である．これらの指標のうち，(1) と (2) は類似の情報であるため，(1) と (3) を線型結合した指標

$$C_{\log}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \Big\{-\lambda H(w_1^{i-1}) + (1-\lambda)\log P(w_i|w_1^{i-1})\Big\}$$

を用い，$\lambda = 0.1$ の場合にパープレキシティよりも高

い相関が得られることを示した．また，この指標を最大にするように混合言語モデルの混合率を決定し，最尤推定によって混合率を決定した場合よりも低い単語エラー率が得られることを示している．

4. 尤度差を用いる評価尺度

従来の手法の特徴と問題点についてまとめると，次のようになる．

(1) パープレキシティや SMR-Perplexity は理論的に明確で計算も容易であるが，実際の単語エラー率との相関が低い場合がある．
(2) 単語の生起確率と，その単語が正解になる確率の間には非線型な関係がある．
(3) 正解以外の単語の確率を考慮する（Clarkson らの用いたエントロピーなど）ことによって，単語エラー率との相関を高めることができる．

これらの点を考慮し，単語間の尤度差を用いる新しい評価尺度 LEA を提案する．この尺度は，不正解単語のうち正解単語と競合しそうな単語（具体的には，出現確率が高い単語）と正解単語との尤度差を算出し，そこから単語の認識確率を推定し，その平均を評価値として用いる．この認識確率の算出には，音響スコアの分布と関係したパラメータ μ と σ が用いられる．これは，言語モデルの評価において，それと組み合わせて使う音響モデルの性質を考慮に入れるということを意味している．このように，単語の音素列間の類似度のような音響類似性でなく，音響モデルのマクロな性質を言語モデルの評価に利用するという考え方はこれまでなかった．

ここで提案する評価尺度の定義は次のようなものである．まず，評価に用いるテキストを $w_1, w_2, \ldots, w_N$ とする．また，単語 w_i の出現確率を算出するための単語履歴を h_i とする．Bigram 言語モデルの場合，

$$P(w_i|h_i) = P(w_i|w_{i-1}) \tag{5}$$

trigram 言語モデルの場合

$$P(w_i|h_i) = P(w_i|w_{i-2}w_{i-1}) \tag{6}$$

である．以下の議論は，n-gram モデルに限らず，現在の単語より前の単語履歴から現在の単語の出現確率を決定するタイプのモデルに適用可能である．

次に，$w_{c(i)}$ を次のように定義する．ただし，V は語彙の集合である．

$$w_{c(i)} = \underset{\substack{w\in \mathrm{V} \\ w\neq w_i}}{\operatorname{argmax}} P(w|h_i) \tag{7}$$

このとき，$w_{c(i)}$ は，単語履歴 h_i において最大の出現確率を与える単語である．ただし，$P(w_i|h_i)$ が最大であった場合には，$w_{c(i)}$ は 2 番目に高い出現確率を与える単語となる．ここで提案する評価尺度では，この $w_{c(i)}$ を，正解単語と競合する候補と見なす．

連続音声認識において，単語のセグメンテーション誤りがないと仮定し，単語 w_i に対応する特徴ベクトル系列を x_i とする．このとき，w_i の音響確率は $P(x_i|w_i)$ となる．さて，言語重みを α とするとき，単語 w_i のスコアは

$$L(w_i) = \log P(x_i|w_i) + \alpha \log P(w_i|h_i) \tag{8}$$

となる．同じく，$w_{c(i)}$ のスコアは

$$L(w_{c(i)}) = \log P(x_i|w_{c(i)}) + \alpha \log P(w_{c(i)}|h_i) \tag{9}$$

となる．ここで，w_i と競合する単語が $w_{c(i)}$ のみだと仮定すると，認識において正解単語 w_i が選ばれる条件は

$$L(w_i) - L(w_{c(i)}) > 0 \tag{10}$$

である．言語尤度差を

$$d(w_i|h_i) = \log P(w_i|h_i) - \log P(w_{c(i)}|h_i) \tag{11}$$

とすると，式 (10) の左辺を α で割ったものは

$$\frac{L(w_i) - L(w_{c(i)})}{\alpha} \tag{12}$$

$$= d(w_i|h_i) + \frac{\log P(x_i|w_i) - \log P(x_i|w_{c(i)})}{\alpha} \tag{13}$$

となる．

したがって，競合単語に関する上記の仮定の下では，言語尤度差 $d(w_i|h_i)$ と音響スコア $\log P(x_i|w_i)$, $\log P(x_i|w_{c(i)})$ が分かれば，その単語が正解になるかどうかが分かる．しかし，ここでは言語モデルを単独で評価するのが目的なので，評価の時点では音響スコアを知ることはできない．そこで，音響スコアの値がある分布に従うと仮定し，単語 w_i が正解になるかどうかを確率的に推定することにする．ある単語の音響スコアは，各フレームでの尤度の和で表されるため，ここではそれが正規分布で近似できるものと仮定する．$\log P(x_i|w_i)$, $\log P(x_i|w_{c(i)})$ の従う分布をそれぞれ $N(\mu_C, \sigma_C^2)$ と $N(\mu_D, \sigma_D^2)$ とする．すると，式 (13) の従う分布は

$$N\left(d(w_i|h_i) + \frac{\mu_C - \mu_D}{\alpha}, \frac{\sigma_C^2 + \sigma_D^2}{\alpha^2}\right) \tag{14}$$

である．簡単のために

$$\mu = \frac{\mu_C - \mu_D}{\alpha} \tag{15}$$

$$\sigma^2 = \frac{\sigma_C^2 + \sigma_D^2}{\alpha^2} \tag{16}$$

$$g = d(w_i|h_i) \tag{17}$$

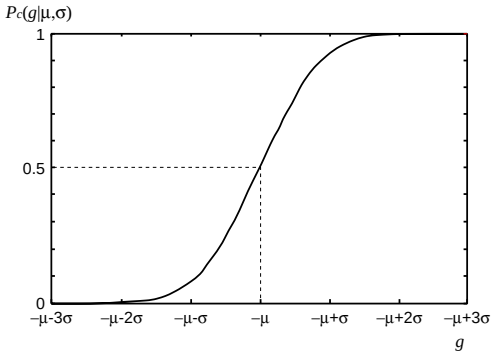

図 1 $P_c(g|\mu,\sigma)$ の概形
Fig. 1 Overview of $P_c(g|\mu,\sigma)$.

とおくと，式 (10) が成り立つ確率は

$$P_c(g|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^{\infty} e^{-\frac{(t-g-\mu)^2}{2\sigma^2}} dt \quad (18)$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{g+\mu}{\sigma}} e^{-\frac{t}{2}^2} dt$$

となって，P_c は言語尤度差 g にシグモイド状の関数をかけたものになる．この関数の概形を**図 1** に示す．最後に，言語モデルの評価値を

$$\mathcal{L}(\mu,\sigma) = \frac{1}{N} \sum_{i=1}^{N} P_c(d(w_i|h_i)|\mu,\sigma) \quad (19)$$

とする．$\mathcal{L}$ は評価テキストの各単語が正解となる確率の平均値となる．この指標をここでは「言語的推定正解率」（Linguistic Estimated Accuracy，LEA）と呼ぶことにする．

この評価値を求めるためには，パラメータ μ と σ が必要である．これは正解の音響スコアと $w_{c(i)}$ の音響スコアの分布を測定することによって推定できる．$w_{c(i)}$ は w_i と音響的な関係がないことから，$w_{c(i)}$ の音響スコアの分布は全不正解単語の音響スコア分布で代用できると考えられる．

最後に，LEA 導出のために置いた仮定についてまとめておく．

(1) 単語のセグメンテーション誤りがない．
(2) 単語 w_i と実質的に競合する単語は $w_{c(i)}$ のみである．
(3) w_i と $w_{c(i)}$ の音響スコアの差は，平均 μ・分散 σ^2 の正規分布に従う．

5. 評価実験

ここで提案した LEA とパープレキシティとで，どの程度単語正解精度との相関があるかを実験によって調べた．ただし，パープレキシティそのものよりも，その対数であるクロスエントロピーの方が単語正解精度との相関が高いため[8]，パープレキシティの代わりにクロスエントロピーを用いた．

表 1 実験 1 に用いたコーパス
Table 1 Corpora used in the experiment 1.

文セット	文数	単語数	単語種
タスク外テキスト	3,000	96,779	3,593
適応テキスト（京都 2，3，4）	341	3976	569
評価テキスト（京都 1）	117	1542	328

ここでは，3 種類の実験を行った．実験 1 は，音声認識のシミュレーションを用いた非常に小規模なタスク適応実験[4] である．実験 2 は，Switchboard コーパスからの n-best の認識結果に対してリスコアリングを行う実験である．実験 3 は，新聞記事読み上げ音声の認識であり，一般的な音声認識タスクでの性能比較である．

5.1 実 験 1

実験 1 に用いたコーパスは，日本音響学会連続音声データベースの対話書き起こしテキストである．その中から京都観光案内をタスクとして選び，「京都観光案内 2，3，4」を適応用に，「京都観光案内 1」を評価用に選んだ．タスク外テキストとして，京都観光案内を除くテキストの中から 3,000 文を選んで用いた．これらのテキストの文数，単語数，単語の種類を**表 1** に示す．

実験に用いた言語モデルは，適応テキストとタスク外テキストを重み付きで混合してから n-gram を作成するタイプの適応 trigram モデルである[4),11)]．このモデルは，タスク外テキストの語彙制限の閾値 T_o，適応テキストの語彙制限の閾値 T_i，混合重み W の 3 つのパラメータを持つ．この 3 つのパラメータを変えることにより，47 種類の言語モデルを作成した．確率の平滑化には Witten-Bell discounting による back-off 平滑化を用いている．これらすべての言語モデルについて，学習に用いたテキストは同じである．これらの言語モデルの作成条件を**表 2** に示す．表中，T_o が – になっている部分では，適応テキストのみを用いてモデルを作成した．重み W が 1 でないモデルの場合，適応テキスト中の単語の出現回数に単純に重みを掛けてモデルを作成した．平滑化に Witten-Bell discounting を用いているため，単語の出現回数が多くなるとディスカウント量が少なくなるという効果がある．

認識実験のシミュレーションとして，乱数によって約 8%の誤りを与えた音素系列からの連続単語認識を行った．音響モデルの代わりに，confusion matrix に基づ

表 2 実験 1 に用いた言語モデル
Table 2 Language model conditions for the experiment 1.

T_o	T_i	W
4,8,12	1	1
		2
		4
		8
		16
		32
		64
		128
18	1	1
		2
		4
		12
		16
		32
		64
		128

T_o	T_i	W
18	2	1
		2
		4
	4	1
		2
	8	1
		2
	10	1
		2
		12
		64
–	1	1
		2
		4
		8

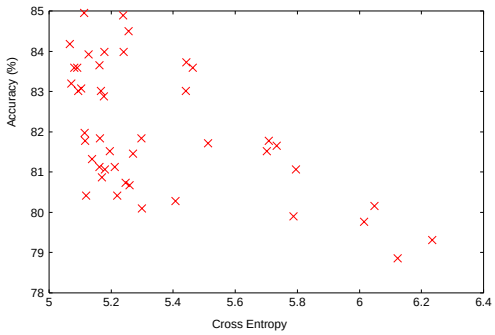

図 2 クロスエントロピーと単語正解精度
Fig. 2 Cross entropy and word accuracy.

く音素間の置換・挿入・脱落確率（音素環境独立なもの）を用いている．認識に用いたデコーダは onepass DP に基づくもので，言語モデルとして trigram を使用し，第 1 候補のみを出力する．言語重みは，1.4 から 3.8 の間で変化させ，最も性能の高かった重みである 2.8 を用いた．挿入ペナルティは用いていない．認識に使用した語彙サイズは 800 である．

それぞれの言語モデルによって認識を行った場合の，言語モデルのクロスエントロピーと単語正解精度の関係を**図 2** に示す[4]．クロスエントロピーと単語正解精度の相関係数は -0.57 であり，相関が低いことが分かる．

これに対して，評価尺度に LEA を用いた場合の相関係数を調べた．音響スコアの分布 $N(\mu_C, \sigma_C^2)$ および $N(\mu_D, \sigma_D^2)$ を直接推定することが困難であったので，ここではさまざまな μ と σ について LEA を算出し，単語正解精度との相関を調査した．

最初に，$-3 \leq \mu \leq 20$，$0.25 \leq \sigma \leq 20$ の範囲で μ と σ を変化させて LEA を測定したときの単語正解精

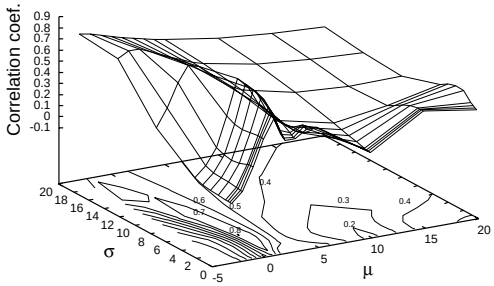

図 3 μ, σ と相関係数の関係
Fig. 3 Relationship between μ, σ and correlation coefficient.

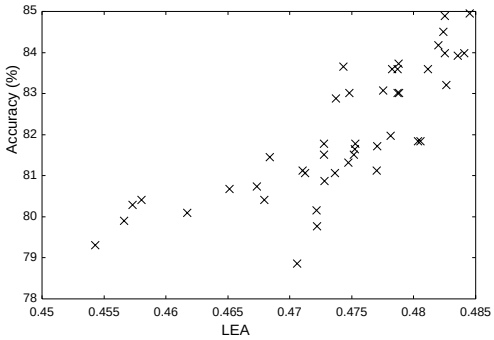

図 4 LEA と単語正解精度
Fig. 4 LEA and word accuracy.

度との相関を**図 3** に示す．$\mu = 1$ 付近に相関の高い領域があり，今回試した条件の中では $\mu = 1$，$\sigma = 5$ の場合に最も高い相関が得られた．相関係数は 0.80 である．このときの LEA と単語正解精度との関係を**図 4** に示す．

また，参考のため，式 (11) の $d(w_i|h_i)$ の平均値を指標として単語正解精度との相関を計算したところ，相関係数は 0.57 であった．

5.2 実 験 2

実験 2 として，Switchboard corpus からの n-best の認識結果をリスコアリングする実験を行った．言語モデル作成に用いたテキストは 2 種類で，適応テキストとして Switchboard corpus の対話書き起こし約 16 万 5 千文，タスク外テキストとして CNN の放送ニュース原稿約 220 万文を用いた．評価には，Switchboard corpus の中から適応テキストと独立な 684 文を用いた．これらのテキストの文数と単語数を**表 3** に示す．

言語モデルとして，タスク外テキストと適応テキストからそれぞれ trigram モデルを作成し，n-gram レベルで混合したもの[9] を用いた．適応テキストから作成した trigram による確率を $P_I(w_i|w_{i-2}w_{i-1})$，タスク外テキストからのものを $P_O(w_i|w_{i-2}w_{i-1})$ とす

表 3 実験 2 に用いたコーパス
Table 3 Corpora used in the experiment 2.

文セット	文数	単語数
タスク外テキスト（CNN）	2,189,437	35,986,571
適応テキスト（Switchboard）	165,003	2,276,027
評価テキスト（Switchboard）	684	7,744

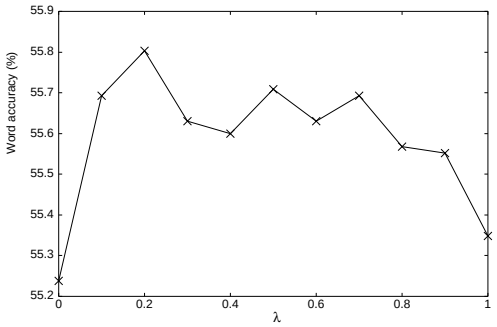

図 5 λ と単語正解精度
Fig. 5 The interpolation factor λ and word accuracy.

るとき，単語の生起確率は

$$P(w_i|w_{i-2}w_{i-1}) = \lambda P_I(w_i|w_{i-2}w_{i-1}) + (1-\lambda)P_O(w_i|w_{i-2}w_{i-1}) \quad (20)$$

によって求められる．結合係数 λ は通常は held-out 法などによって決められるが，ここでは λ を 0〜1 まで 0.1 ずつ変えることによって 11 個の言語モデルを作成し，比較した．平滑化には Witten-Bell discounting を用い，語彙サイズは約 35000 とした．

リスコアリングの元になるデータとして，BBN の Byblos システムによって得られた，最大 100-best の認識候補を用いた．この認識には bigram 言語モデルが用いられており，1 位候補の単語正解精度は 53.9%であった．このデータに対して，上記の各言語モデルを用いたリスコアリングを行い，単語正解精度と言語モデルの評価尺度とを比較した．リスコアリングの際には，言語重みを 1 から 50 まで 1 刻み，挿入ペナルティを −50 から 50 まで 5 刻みに変化させ，そのすべての組合せについてリスコアリング結果を計算した．この結果，言語重み 17，挿入ペナルティ −15，$\lambda = 0.2$ の場合の認識結果が最も良かった（55.8%）ので，この言語重みと挿入ペナルティでの結果を利用した．

まず，各言語モデルでリスコアリングした場合の，1 位候補の単語正解精度と λ との関係を**図 5** に示す．上記のとおり，$\lambda = 0.2$ の場合に単語正解精度が最大となった．この場合のクロスエントロピーと単語正解精度の関係を**図 6** に示す．両者の相関係数は −0.66 であった．

次に，μ，σ を変化させながら，LEA と単語正解精度の相関を調べてみた．これを**図 7** に示す．この実験では広い範囲で相関が高くなったが，最も相関が高い条件は $\mu = 15$，$\sigma = 1$ で，相関係数は 0.90 であった．このときの LEA と単語正解精度の関係を**図 8** に示す．また，参考のため，式 (11) の $d(w_i|h_i)$ の平均値を指標として単語正解精度との相関を計算したところ，相関係数は 0.77 であった．

実験 2 における最適な μ の値は，実験 1 における値と大きく異なっている．これは，2 つの実験における音響モデルの差に起因すると思われる．実験 1 におけ

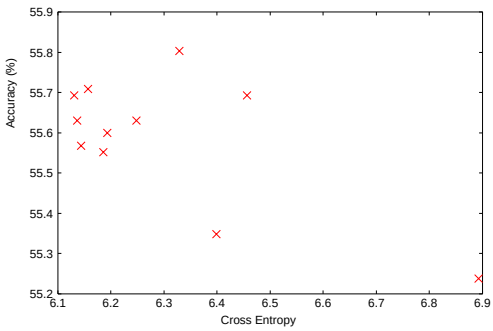

図 6 クロスエントロピーと単語正解精度
Fig. 6 Cross entropy and word accuracy.

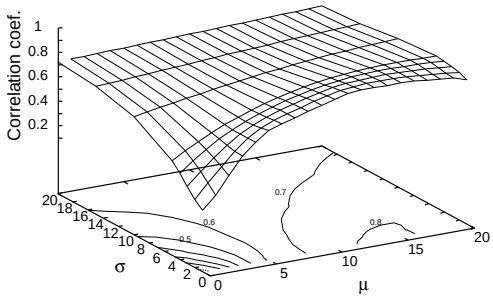

図 7 μ，σ と相関係数
Fig. 7 Relationship between μ, σ and correlation coefficient.

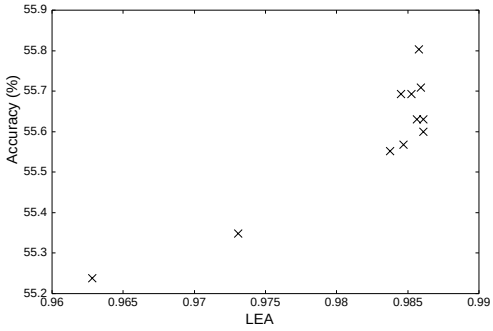

図 8 LEA と単語正解精度
Fig. 8 LEA and word accuracy.

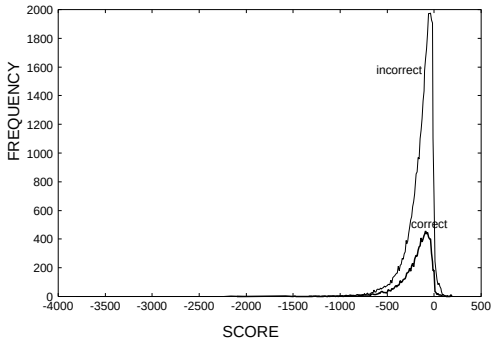

図 9 音響スコアの分布
Fig. 9 Acoustic score distributions for correct and incorrent words.

る音響スコアは音素の混同確率の対数であり，その大きさは言語スコアと同程度である．これに対して，実験 2 の音響スコアは HMM の出力確率密度であり，その対数は言語スコアと比較してかなり小さな値になる．これが原因で μ の値が変化した可能性がある．そこで，実験 2 における音響スコアの分布を調査してみた．

実験 2 において，100best 候補中の正解単語と不正解単語の音響スコアの分布を**図 9** に示す．正解単語の分布は $\mu_C = -199.28$，$\sigma_C = 177.02$ で，不正解単語の分布は $\mu_D = -154.49$，$\sigma_D = 151.77$ であった．ここから μ と σ を計算すると $\mu = -2.6$，$\sigma = 13.7$ となるが，これは最適値である $\mu = 15$，$\sigma = 1$ とは大きく異なっている．この食い違いの原因として，ここでの不正解単語の分布が 100best 候補のみから求めたものだという点が考えられる．μ，σ の推定に用いるべき不正解単語のスコア分布はすべての不正解単語のスコアであるため，上位の候補だけから分布を推定すると，μ が小さい方向に分布が偏る．このため，μ，σ の最適値と推定値に食い違いが出たと思われる．

5.3 実　験　3

実験 3 では，新聞記事読み上げ音声の認識実験を行った．前述のとおり，新聞記事読み上げタスクのように大規模なコーパスから比較的単純な言語モデルを作成する場合には，パープレキシティと単語正解精度との相関は比較的高い．ここでは，そのような場合に LEA とクロスエントロピーの性能の比較を行う．実験 3 における音響分析条件と音響モデルの作成条件を**表 4** に示す．言語モデル作成のための学習データとして，毎日新聞 91 年〜96 年の 6 年分の記事を用いた．ただし 94 年は 1〜9 月分である．これは，評価データが毎日新聞 94 年の 10〜12 月の記事であるためである．各年の記事のみを用いた場合と，6 年分の記事すべてを用いた場合で 7 通りの言語モデルを作

表 4 分析条件と音響モデル作成条件
Table 4 Analysis conditions and acoustic model conditions.

標本化周波数	16 kHz
量子化ビット数	16 bit
分析フレーム長	32 msec
分析周期	8 msec
分析窓	ハミング窓
高域強調	$1 - z^{-1}$
特徴ベクトル	1〜12 次の LPC メルケプストラムと対数パワー，および 1 次と 2 次の回帰係数（計 39 次元）
正規化	発話ごとのケプストラム平均正規化
音響モデル	音素環境依存 HMNet，2000 状態 16 混合
音響モデル学習データ	日本音響学会新聞記事読み上げデータベース（ASJ-JNAS）男性話者 102 名，15,732 文

成する．また，それぞれについて bigram と trigram を作成し，またカットオフ頻度を 0〜5 まで 6 通りに設定する．ただし，trigram のカットオフについては，trigram 頻度と bigram 頻度について同じカットオフを適用するものとする．以上の条件により，計 84 種類の言語モデルを作成した．

評価データは ASJ-JNAS の男性 10 名（学習データに含まれないもの）が発声した計 100 文である．ここで用いているデコーダは 2 段階で認識を行う[12]．1 段目では，音響モデル・bigram 言語モデル・辞書を用いて入力音声から単語グラフを生成する．2 段目では，単語グラフ内の候補に対して bigram または trigram 言語モデルを適用し，リスコアリングによって最終候補を生成する．ここでは，1 段目の bigram 言語モデルとして，毎日新聞 91〜96 年度で学習されたカットオフ 0 の bigram モデルを使用した．認識における言語重みは 20，挿入ペナルティは -5 である．2 段目のリスコアリングで種々の言語モデルを用いることにより，異なる認識結果を生成している．

単語グラフを生成するにあたり，第 1 段目でのビーム幅を変えて 2 種類の単語グラフを生成した．1 つ目は単語内ビーム幅 400，単語間ビーム幅 300 であり，2 つ目は単語内ビーム幅 150，単語間ビーム幅 100 である．それぞれの単語グラフにおけるグラフエラー率は 0.6%および 23.4%であった．これらの単語グラフをリスコアリングする際に，言語重み α と挿入ペナルティの p の設定として，$(20, -20)$ と $(50, 0)$ の 2 つを用いて比較した．前者は最適な設定に近いもので☆，

☆ 最適な (α, p) の値は言語モデルごとに異なる．ここでは，最も多くの言語モデルで最適だった値を用いた．

表 5 各種の指標と単語正解精度との相関係数
Table 5 Correlation coefficients between the word accuracy and various metrics.

ビーム幅	α	p	相関係数 entropy	相関係数 LEA	μ	σ
400/300	20	−20	−0.94	0.95	10	9.5
	50	0	−0.95	0.97	6	5.5
150/100	20	−20	−0.85	0.92	0.5	5
	50	0	−0.67	0.84	−3.0	5

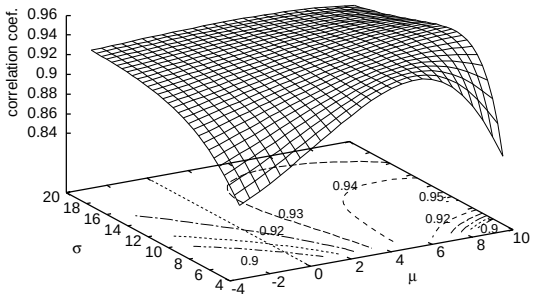

図 10 μ, σ と相関係数
Fig. 10 Relationship between μ, σ and correlation coefficient.

後者は最適値から大きく外れた設定である．このような比較をする理由は，言語重みと挿入ペナルティの設定の最適性が，言語モデル評価基準と単語正解精度との相関に影響することを確かめるためである．

クロスエントロピーと LEA で相関を比較した結果を**表 5** に示す．表中の μ, σ は，相関が最も高かったものを示している．また，単語内/単語間ビーム幅が 400/300，$\alpha = 20$，$p = -20$ のときの μ, σ と相関係数の関係を**図 10** に示す．表 5 より，ビーム幅が 400/300 の場合は，クロスエントロピーと LEA はいずれも単語正解精度と高い相関を持つことが分かる．また，式 (15) より μ と α は反比例の関係にあるが，ビーム幅 400/300 における $\alpha = 20$ と $\alpha = 50$ における μ の最適値の相違はこの関係を反映していると考えられる．ただ，それぞれの条件での μ の最適値は完全な反比例関係にあるわけではないので，LEA 導出時の仮定（セグメンテーション誤りがない，対立候補が唯一，など）と実際の認識処理とのずれによって最適値が変化した可能性がある．これについては今後さらに検討する必要がある．また，ビーム幅を絞って正解候補が多く脱落している状況では，クロスエントロピーと単語正解精度の相関が低くなることが分かった．この場合でも LEA は単語正解精度と高い相関を示すが，このときの μ の最適値はビーム幅が広い場合と比べて小さくなる．この現象は，定性的には次のように考えることができる．すなわち，単語グラフにおいて「正解が脱落した」という状況は，言語モデルからみれば「正解単語に対して非常に低い音響スコアが与えられた」という状況と等価である．その影響によって，実質的な音響尤度差 μ の値が小さくなるという現象が起きたと考えられる．しかし，ビーム幅 150/100 における $\alpha = 20$ と $\alpha = 50$ の結果については μ に反比例関係があるとはいえず，まだ不明な点が多い．

図 10 の分布は，図 7 の分布に類似しており，広い範囲で高い相関があることを示している．これらの結果から，音声認識タスクにおいては，μ, σ はこれらと類似した分布を示すことが予想される．

6. むすび

言語モデル評価のための新しい指標について述べた．本稿で提案した指標は，あるコンテキストでの単語の言語スコアと，同コンテキストでの最大の言語スコアとの差をとり，そこから単語の認識確率を推定して平均したものである．言語スコアの差をとることにより，他の単語との相対的な良さを評価することができる．シミュレーション実験および実音声の認識結果のリスコアリング実験の結果から，ここで提案した指標は，単語正解精度に対してパープレキシティよりも高い相関を持つことが示された．

残る問題点は，μ と σ をどうやって事前に推定するかである．これを事後に決めたのでは，あらかじめ言語モデルを評価しておくことができないので，事前にこれらの値が決定できるようにしなければならない．理論的には，これらは音響尤度差の分布のパラメータである．したがって原理的には求めることができるが，実際には正解単語と不正解単語の音響スコアの分布をそれぞれ推定することが困難なことも多く，また仮定と実際の認識プロセスとの違いによって μ, σ の最適値にずれが生じることも考えられる．これらの値の簡易な推定方法，理論値と実際の最適値のずれについては，今後さらに検討する必要がある．また，今後は，より多くの実験を通して本手法の有効性を検証していきたいと考えている．

謝辞 研究の機会を与えてくださった Prof. Mari Ostendorf（当時米国ボストン大学，現米国ワシントン大学）に感謝いたします．また，実験を担当していただいた穐本由紀子氏に感謝いたします．

参考文献

1) 亀山誠裕，加藤正治，伊藤彰則，好田正紀：新聞記事コーパスから作成した各種 N-gram 言語モデルの音声認識実験による評価，日本音響学会秋

季講演論文集，2-1-18, pp.73–74 (1998).
2) Iyer, R., Ostendorf, M. and Meteer, M.: Analyzing and Predicting Language Model Improvements, *Proc. IEEE Workshop on Speech Recognition and Understanding*, pp.254–261 (1997).
3) Chen, S., Beeferman, D. and Rosenfeld, R.: Evaluation metrics for language models, *Proc. DARPA broadcast news transcription and understanding workshop* (1998).
4) 伊藤彰則，好田正紀：N-gram 出現回数の混合によるタスク適応の性能解析，電子情報通信学会論文誌（D-II），Vol.J83-D-II, No.11, pp.2418–2427 (2000).
5) 中川聖一，大黒慶久，村瀬　功：連続音声認識システムの評価法—タスクの複雑性と文認識率の関係，電子情報通信学会論文誌（D-II），Vol.J73-D-II, No.5, pp.683–693 (1990).
6) 大槻恭士，伊藤彰則，牧野正三：音素・文字間の遷移情報を用いた単語認識の性能予測，電子情報通信学会論文誌（D-II），Vol.J76-D-II, No.6, pp.1090–1096 (1993).
7) 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).
8) 中川聖一，伊田政樹：連続音声認識のタスクの複雑さを表わす新しい尺度，電子情報通信学会論文誌（D-II），Vol.J81-D-II, No.7, pp.1491–1500 (1998).
9) Iyer, R., Ostendorf, M. and Rohlicek, J.R.: Language modeling with sentence-level mixtures, *Proc. ARPA Human Language Technology Workshop*, pp.82–87 (1994).
10) Clarkson, P. and Robinson, T.: Towards improved language model evaluation measures, *Proc. Eurospeech '99*, pp.1927–1930 (1999).
11) Ito, A., Saitoh, H., Katoh, M. and Kohda, M.: N-gram language model adaptation using small corpus for spoken dialog recognition, *Proc. Eurospeech '97*, pp.2735–2738 (1997).
12) 堀　貴明，岡　直生，加藤正治，伊藤彰則，好田正紀：大語彙連続音声認識のための音素グラフに基づく仮説制限法の検討，情報処理学会論文誌，Vol.40, No.4, pp.1365–1393 (1999).

(平成 13 年 11 月 12 日受付)
(平成 14 年 4 月 16 日採録)

伊藤　彰則（正会員）

昭和 61 年東北大学工学部通信工学科卒業．平成 3 年同大学大学院博士課程修了．同年同大学応用情報学研究センター助手．平成 10〜11 年米国ボストン大学客員研究員．現在，山形大学工学部情報工学科助教授．音声言語情報処理の研究に従事．工学博士．日本音響学会，電子情報通信学会，ISCA 各会員．

好田　正紀（正会員）

昭和 40 年名古屋大学工学部電子工学科卒業．昭和 42 年同大学大学院修士課程修了．同年日本電信電話公社電気通信研究所入社．昭和 62 年山形大学工学部情報工学科教授．音声認識を主とする音声情報処理の研究に従事．工学博士．日本音響学会，人工知能学会，言語処理学会各会員．