音声言語情報処理　31－10
（2000．6．2）

音響モデル尤度と言語モデル尤度のバランスの理論的・実験的検討

堀部 千寿*　峯松 信明**　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
**東京大学 大学院 工学系研究科 電子情報工学専攻

現在の多くの音声認識システムでは、単語仮説や文仮説などを求めるために、音響モデルと言語モデルが使われる。しかし、音響モデルと言語モデルの尤度のバランスを取るために言語モデルの値に重み付けをしたり定数値のペナルティやボーナスを加えたものが用いられる。この重みつき言語尤度と音響尤度を合わせ仮説全体の尤度として扱う。ここで用いられる言語重みや定数ペナルティは経験的に決定されることが多い。しかし、音響モデルの情報量などを考慮すると言語重みは音韻あたりの平均フレーム数と音響モデルの相互情報量に依存すると考えられる。これを理論的・実験的に検証し、音響モデルの相互情報量が言語重みに依存することを示した。

Theoretical / Experimental Investigation of Balance between Acoustic Model Likelihood and Language Model Likelihood

Yukihisa HORIBE* , Nobuaki MINEMATSU** and Seiichi NAKAGAWA*

*Department of Information and Computer Sciences
Toyohashi University of Technology, Tenpaku, Toyohashi, 441-8580, Japan
**Department of Information and Communication Engineering, School of Engineering,
University of Tokyo

Many speech recognition systems integrate an acoustic model with a language model to generate word hypotheses or sentence hyoptheses. In order to balance the likelihood between acoustic and language models we weight a language model and add a constant penalty or bonus to compensate the difference. The hypotheses are generated by the integration of weighted language model likelihood with acoustic model likelihood. Language model weight and constant penarty are often decided experimentally. From a view point of information theory, a language model weight depends on acoustic model entropy and the average number of frames of every phoneme. In this paper, we investigate this hypothesis and show that the language model weight depends on the mutual information of an acoustic model.

1　はじめに

現在、数多くの音声認識システムが商品化、実用化されている [1, 2, 3, 4] 。

このような中で、ほとんどの音声認識システムは単語仮説や文仮説などを求めるために、音響モデルと言語モデルの２つのモデルが使われる。しかし、この音響モデルと言語モデルの間には各モデルが出力する尤度のレンジに差がある。そこで、その差を補うために言語モデルの値に重み付けをしたり定数値のペナルティやボーナスを加えたものが用いられる。この重みつき言語尤度と音響尤度を合わせ仮説全体の尤度として推定を行う。ここで用いられる言語重みや定数ペナルティは経験的に決定されることが多い。この言語重みや定数ペナルティを理論的に

推定する研究も行われている [5] 。[5] では言語モデルに対するペナルティ値を一般化ベルヌーイ試行に基づいて決定する方法で効果が得られている。

本稿では、言語重みは音響モデルの相互情報量に依存すると考え、2節でその理論的な検討を行ない、4節において実験的に検証を行った。

2 確率モデルによる音声認識 - 音響・言語尤度のバランス -

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 $P(W \mid Y)$ を最大にする W を見い出すのが妥当であろう [6]。

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

であるから、$P(Y \mid W), P(W), P(Y)$ 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。$P(Y \mid W)$ 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 [7] 。

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

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 [8] 。[8] では同じ音響特徴パラメータでも HMM の構造 (状態総数や混合分布数) によって最適な言語重みとペナルティの値が異なること、1パス目 (bigram) と2パス目 (trigram) の最適値がほぼ同じであることを実験的に示している。しかし、テスト文集合によって結果が異なる可能性もある。

音響モデルの尤度は多次元正規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 [9]。ただ、時間方向の積分値に影響を与えるのはフレームレートであり、言語重みはフレーム周期の逆数に比例することがわかる。

まず、音響モデル (単語単位) と言語モデルの相互情報量を求めよう。

$$\begin{aligned}
I(V;Y) &= H(V) - H(V|Y) \\
&= H(Y) - H(Y|V) && (4) \\
&\fallingdotseq H(V) - \sum_i H(C_i|Y_i) + \alpha && (5) \\
&\fallingdotseq H(Y) - \sum_i H(Y_i|C_i) && (6) \\
I(C;Y) &= H(C) - H(C|Y) && (7) \\
&= H(Y) - H(Y|C) && (8) \\
I(V;L) &= H(V) - H(V|L) && (9)
\end{aligned}$$

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。(5) 式の α は音韻系列の制約のための補正項である。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 $H(C) \fallingdotseq 5$ ビット／音韻となる。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 [10]。これから 1 単語が平均 6 音韻から構成されるとすると、$I(V;Y) = 14.3 - 0.5 \times 6 = 11.3$ 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 64 とすると、$H(V|L) = 6.0$ ビット／単語となり、$I(V;L) = 14.3 - 6.0 = 8.3$ 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。ベイズ確率論から考えれば、音響モデルと言語モデルが独立なら、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 [11]。一方、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ので以下で考察する。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 (等確率) 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) 。それでは、この中間に対応する一般の場合はどうであろうか。複数の入力モーダルからの統合法に関する場合を考える [13]。例えば、ケプストラム次数の i と j が独立である場合 (対角共分散行列を使用する場合) は、それぞれの正規分布からの確率密度の積で統合される [14]。一方、各情報源から得られる結果の信頼性を相互情報量で表現し、重み付けして統合する方法も提案されている [15]。

言語重みを導入する必要があるのは、次の 3 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、(3) 各情報源の信頼度、である。(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の総和は 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。

$$\lambda_1 = (\text{音韻当りの平均フレーム数}) \times \frac{\text{音響モデルによる相互情報量}}{\text{真の音響モデルの相互情報量}} \quad (10)$$

さらに、(3) の信頼度を相互情報量で表現すると、

$$\lambda_2 = \lambda_1 \frac{\text{言語モデルによる相互情報量/単語}}{\text{真の音響モデルによる相互情報量/単語}} \quad (11)$$

通常、フレームシフト幅は 10ms 前後なので (10) 式の第 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。以上の考察は、1 単語あたりの音韻数を一定と仮定しているが、実際は異なっているので、言語重みも単語長に依存する可能性がある [12]。一方、(11) 式の第 2 項は unigram で 0.45、bigram で 0.66 、trigram で 0.73 程度であるので、λ_2 は 6 ～ 10 前後となる。

3 大語彙連続音声認識システム

まず、本稿における基準となる大語彙連続音声認識システム (baseline system) について簡単に述べる。ベースラインシステムにおける音声認識デコーダは 2 パスで構成される。1 パス目において多少粗い音響モデルと言語モデルを使用して入力音声に対する N-best 仮説を求め、2 パス目で 1 パス目より精度の良い音響モデルや言語モデルを使って最終的な認識結果を出力する。

1 パス目における連続音声認識のアルゴリズムには One pass Viterbi 法を用いている。これは、各フレームを各単語境界と仮定し、言語モデルによる確率の対数値を音響モデルの累積スコアに加えることを繰り返すことによって次の (12) 式を満たす最尤の単語列候補を求める。

$$P(w^*|y_1^T) = \underset{\{w_1^N\}\{t_1^N\}}{\text{argmax}} \left\{ \sum_{n=1}^{N} \log(P_a(y_{t_{n-1}+1}^{t_n}|w_n)) + \sum_{n=1}^{N} \{W_L \log(P_l(w_n|h_{n-1})) + P_c\} \right\} \quad (12)$$

ここで $P_a(y_{t_{n-1}+1}^{t_n}|w_n)$ は観測パターン系列 $y_{t_{n-1}+1} y_{t_{n-1}+2} \cdots y_{t_n}$ における単語 w_n の音響的な尤度、$P_l(w_n|h_{n-1})$ は単語系列 h_{n-1} の次に単語 w_n が接続する言語確率である。音響的尤度と言語尤度の間には尤度の値のレンジが異なるため、結合の際に言語重み W_L と、単語の接続数に対する定数ペナル

ティ(またはボーナス) P_c を使用する。本稿のベースラインシステムで標準的に使用されている音響モデルの仕様を表1に示す。本研究において、1フレームあたりの音響尤度の計算に用いられるケプストラムは10次元である(この特徴パラメータを以降、frameと呼ぶ)。この10次元のケプストラムを4フレーム分まとめてベクトル化し(40次元)、これに対してK-L展開を用いて20次元に圧縮したものを最終的な特徴ベクトルとして用いている(この特徴パラメータを以降、segmentと呼ぶ)。

なお、(12)式における音響尤度 $P_a(y_{t_{n-1}+1}^{t_n}|w_n)$ は、$t_s = t_{n-1}+1$ とすると、特徴パラメータにsegmentを用いる場合は(14)式の近似式を用いる。特徴パラメータにframeを用いる場合は(15)式を用いる。

また、言語尤度 $P_l(w_n|h_{n-1})$ は以下の(16)式で定義される。

$$P_a(y_{t_{n-1}+1}^{t_n}|w_n)$$
$$= \sum_x \prod_{i=t_s}^{t_n} \frac{P(y_{i-3}y_{i-2}y_{i-1}y_i|x_{i-1}x_i)}{P(y_{i-3}y_{i-2}y_{i-1}|x_{i-1}x_i)} P(x_i|x_{i-1}) \quad (13)$$
$$\approx \sum_x \prod_{i=t_s}^{t_n} P(y_{i-3}y_{i-2}y_{i-1}y_i|x_{i-1}x_i) P(x_i|x_{i-1}) \quad (14)$$
$$P_a(y_{t_{n-1}+1}^{t_n}|w_n)$$
$$= \sum_x \prod_{i=t_s}^{t_n} P(y_i|x_{i-1}x_i) P(x_i|x_{i-1}) \quad (15)$$

$$P_l(w_n|h_{n-1}) = \begin{cases} P(w_n) \\ \quad (\text{unigram の場合}) \\ P(w_n|w_{n-1}) \\ \quad (\text{bigram の場合}) \\ P(w_n|w_{n-2}, w_{n-1}) \\ \quad (\text{trigram の場合}) \end{cases} \quad (16)$$

4 実験結果と考察

4.1 音響モデルと言語モデル

比較結果を述べる前に、比較実験を行うために作成した音響モデルの作成条件について述べる。作成した音響モデル全てにおいて固定とした実験条件を表1に、個々の音響モデルに依存する条件を表2に示す。表2において、パラメータの欄のLPCMCはLPCメルケプストラム(1フレーム10次元)を表し、LPCMC-KLはLPCメルケプストラムで計算した特長パラメータを4フレーム分を1セグメントとして40次元の特徴パラメータとし、これに対してKL展開による次元圧縮(20次元)を行ったものである。MFCCでは、サンプリング周波数が12 kHzの場合は1フレーム10次元、16 kHzの場合は1フレーム12次元の特徴パラメータを用いる。MFCC-KLは4フレーム分を1セグメントとしてKL展開による次元圧縮を行うが、圧縮後の次元数はサンプリング周波数が12 kHzの場合は20次元、16 kHzの場合は24次元とする。また、Δはケプストラムと Δ パワーの使用を、ΔΔはΔΔケプストラムと ΔΔ パワーの使用をそれぞれ表す。表2のSFの欄はサンプリング周波数を表し、dimsは音響モデル全体のパラメータの次元数を表す。

表1: 全音響モデル共通の実験条件

窓関数	21.33 ms ハミング窓 (256 points)
frame 周期	8 ms (96 points)
モデル単位	音節 (正確にはモーラ単位)
音節種類数	114
HMM	連続出力分布型 HMM (5 状態 4 出力分布、全共分散行列、 4 混合ガウス分布)
音響モデル 学習データ	ASJ Database(男性 30 名 4518 文) 新聞読上げ音声 (男性 125 名 12703 文)

表2: 実験に使用する音響モデル

モデル名	パラメータ	SF	dims
LPCMC-f-12	LPCMC	12k	10
LPCMC-f-12-d	LPCMC,Δ	12k	21
LPCMC-f-12-dd	LPCMC,Δ,ΔΔ	12k	32
LPCMC-s-12-d	LPCMC-KL,Δ	12k	31
LPCMC-s-12-dd	LPCMC-KL,Δ,ΔΔ	12k	42
MFCC-f-12-dd	MFCC,Δ,ΔΔ	12k	32
MFCC-f-16-dd	MFCC,Δ,ΔΔ	16k	38
MFCC-s-12-dd	LPCMC-KL,Δ,ΔΔ	12k	42
MFCC-s-16-dd	LPCMC-KL,Δ,ΔΔ	16k	50

表3に各音響モデルの情報量を示す。表から、(10)式の第2項の値はおよそ2～3程度であることがわかる(LPCMC-s-12-dだけなぜか大きい)。

次に、比較実験に使用した言語モデルの性能を表4に示す。表4におけるPPは未知語をスキップしたテスト文パープレキシティ、PP'は未知語を含んだテスト文パープレキシティ、APPは未知語の種類数を考慮した補正パープレキシティである。なお、語彙数は20,000語である。言語モデルの学習はJNASの新聞記事読み上げコーパスの1991年1月～1994年9月分、約330万文を使用し、評価は

表 3: 各音響モデルの情報量の値

モデル名	$H(Y\|C)$	$H(Y)$	$I(Y;C)$
LPCMC-f-12	182.98	190.01	7.03
LPCMC-f-12-d	399.23	406.88	7.65
LPCMC-f-12-dd	520.80	529.71	8.91
LPCMC-s-12-d	594.20	611.51	17.31
LPCMC-s-12-dd	713.74	—	—
MFCC-f-12-dd	832.15	843.30	11.15
MFCC-f-16-dd	514.54	525.25	10.71
MFCC-s-12-dd	586.44	598.49	12.05
MFCC-s-16-dd	697.22	712.58	15.36

新聞読み上げコーパスより 100 文 (話者 10 人分、IPA'98 の評価セットと同様のもの) を使用した。

表 4: 実験に使用する言語モデルとパープレキシティ

モデル名	PP	PP'	APP
unigram	746.3	731.1	781.6
bigram	109.9	109.0	116.5
trigram	69.3	69.2	74.0

4.2 言語モデルの違いによる比較 - unigram , bigram , trigram -

各言語モデル、各音響モデルを用いて認識実験を行った結果、認識精度が最大となった言語重みを表 5 に示す。ここでは言語重みと定数ペナルティは認識精度が最大になるように各モデルごとに最適化した。表 5 において表記が " 12-18(12) " となっている場合は、認識精度が最大となる言語重みは 12 で、最大精度からの認識精度の差が 1 % 以内 (言語モデルに unigram を使用する場合は認識性能が低いため、2 % 以内とする) である言語重みが 12 ~ 18 であることを示す。まず、言語モデルの違いによる言語重みについて比較する。音響モデル LPCMC-s-12-dd における認識精度が最大になる言語重みは、unigram < bigram ≒ trigram となっており、bigram と trigram の言語重みは同程度の値と見ることができる。1%の誤差範囲でも同様に unigram < bigram ≒ trigram という傾向が見られる。音響モデル MFCC-f-12-dd においてもほぼ同様の傾向が見られる。それ以外の音響モデルについては言語モデルの違いによる明確な差はない。

表 5: 認識精度が最大となる言語重み

(括弧内は最適値)

モデル名	unigram	bigram	trigram
LPCMC-f-12-d	12-15(15)	12-18(18)	12-18(12)
LPCMC-f-12-dd	12-18(12)	12-18(15)	12-18(12)
LPCMC-s-12-d	12-18(15)	18(18)	12-18(15)
LPCMC-s-12-dd	9-18(15)	15-21(18)	18-21(18)
MFCC-f-12-dd	9-15(12)	15-18(15)	15-18(18)
MFCC-f-16-dd	12-15(15)	12(12)	12-24(15)
MFCC-s-12-dd	12-18(15)	18-21(18)	15-24(18)
MFCC-s-16-dd	15-18(15)	15-21(18)	15-21(18)

4.3 特徴パラメータの違いによる比較 - Δ, ΔΔ, セグメント -

特徴パラメータの種類の違いによる言語重みの値の差について検討する。まず、表 5 の LMCMC-f-12-dd と MFCC-f-12-dd の比較をすると、有意な差は見られなかった。これはセグメントレベルの音響モデルの場合でも同じ傾向が見られる。以上から LPCMC と MFCC の特徴パラメータによる言語重みの差はないということができる。(後述の triphone 単位の音響モデルを使った場合とは異なる傾向)

次に、ΔΔ ケプストラムの有無による言語重みの差の比較をする。表 5 の LPCMC-f-12-d と LPCMC-f-12-dd の比較をすると、この 2 つのモデルの間には有意な差は見られなかった。これは LPCMC-s-12-d のモデルの場合でも同じ傾向が見られた。以上のことから ΔΔ ケプストラムの有無による言語重みの差は見られなかった。(後述の triphone 単位の音響モデルを使った場合とは異なる傾向)

セグメント単位とフレーム単位のモデルの間での言語重みの差の比較をする。表 5 の LPCMC-f-12-dd と LPCMC-s-12-dd の比較をすると、セグメント単位の音響モデルのほうがフレーム単位の言語重みより大きくなっている。この傾向は他の音響モデルでも同じである。

これはセグメント単位の音響モデルのみかけ上の相互情報量が大きくなったためと考えられる。

4.4 Triphone 単位の音響モデルを用いた場合

音響モデルに triphone 単位モデルを用いた場合の比較を行うために、IPA の Julius 2.0 による認識実験を行った。なお、フレーム周期は 10 ms であり、多次元正規分布には、対角共分散行列を用いている。結果を表 6 に示す。表 6 の結果より、言語

表 6: triphone を用いた場合の認識精度が最大となる言語重み

(括弧内は最適値)

モデル名	unigram	bigram	trigram
LPCMC-f-12-dd	—	15-16,18(16)	15-19(16)
LPCMC-s-12-dd	—	22,24(24)	19-26(22)
MFCC-f-12-d	11-14(13)	10-12(12)	10-14(11)
MFCC-f-12-dd	13-17(17)	17,19(17)	14-16(14)
MFCC-f-16-d	12(12)	11(11)	9-12(9)
MFCC-f-16-dd	—	13-15,17(15)	11-16(14)

モデルの違いによる言語重みの値の差については、全体的に unigram ≧ bigram ≧ trigram というような傾向が見られた。特徴パラメータの種類の違いによる言語重みの値の差については、LPCMC と MFCC では、LPCMC の方が言語重みが大きいことがわかる。また、ΔΔ ケプストラムの有無による言語重みの差は、ΔΔ ケプストラムを持つモデルの方が言語重みが大きくなっている。セグメント単位とフレーム単位のモデルの間での言語重みの差は音節モデルの場合と同様にセグメント単位の音響モデルの方が言語重みの値が大きくなっている。

以上の結果より、triphone モデルを用いた場合ではそれぞれの比較で何らかの差が見られる。これは言語重みを振らす値の幅が音節モデルの場合 (言語重みの値を 3 ずつ変動させている) より細かくなっているため、より厳密な言語重みを得ることができたため、その差を厳密に求めることができたためと思われる。

Triphone と音節モデルの比較では、フレーム周期が 10 ms と、音節モデルに比べフレーム数が 0.8 倍になったにもかかわらず、LPCMC のモデルでは triphone の方が言語重みが大きくなっており、MFCC のモデルではほとんど同じような言語重みとなっている。これは、triphone モデルは対角共分散行列を使っており、そのため各ケプストラム係数が独立なものとして扱われるために、見かけ上相互情報量が大きくなっているためと考えられる。

5 むすび

音声認識で使用される音響モデルと言語モデルの間にはモデルから得られる尤度に、大幅なレンジの差がある。そのため、通常は言語モデルに重み付けを行い、尤度のレンジの調整を行っている。この際の言語重みは経験的に決定されることが多いが、本稿ではモデルの相互情報量に着目し、言語重みが言語モデルには依存せず、音響モデルの相互情報量に依存することを理論的、実験的に検討した。

参考文献

[1] Proceedings of the Broadcast News Transcription and Understanding Workshop (1998)

[2] Proceedings of the DARPA Broadcast News Workshop (1999)

[3] 西村雅史 , 伊東伸泰 , 山崎一孝 : "単語を認識単位とした日本語の大語彙連続音声認識" , 情報処理学会論文誌 Vol.40 No.4 ,pp.1395-1403 (1999 年 4 月)

[4] 小林彰夫 , 今井亨, 安藤彰男 , 中林克巳 : "ニュース音声認識のための時期依存言語モデル" , 情報処理学会論文誌 Vol.40 No.4 , pp.1421-1429 (1999 年 4 月)

[5] 小川厚徳, 武田一哉, 板倉文忠 : "一般化ベルヌーイ試行に基づく言語確率の補正方法" , 電子情報通信学会論文誌, D-II Vol.J81-D-II No.12 pp.2703-2711 (1998 年 12 月)

[6] 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)

[7] Kristie Seymore, Stanley Chen, Maxine Eskenazi, and Ronald Rosenfeld : "Language and pronunciation modeling in the CMU 1996 Hub4 evaluation," Proc. of Speech Recognition Workshop, pp.141-146 (Feb. 1997)

[8] 斎院俊典, 岡直生, 加藤正治, 伊藤彰則, 好田正紀 : "単語グラフ生成の言語重み・挿入ペナルティ最適化の検討", 音響学会講論集, 2-8-10 (2000 年 3 月)

[9] 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)

[10] 中川聖一 : "情報理論の基礎と応用", 近代科学社, (1992)

[11] Seiichi Nakagawa, Li Zhao and Hideyuki Suzuki : "A Comparative Study of Output Probability Functions in HMMs," IEICE TRANS. INF. & SYST., VOL. E78-D, NO. 6, pp.669-675 (JUNE 1995)

[12] 甲斐充彦, 廣瀬良文, 中川聖一 : "N-gram 言語モデルと効率的探索法を用いた大語彙連続音声認識システムの検討", 電子情報通信学会 信学技報, SP97-99 (1998)

[13] 松永浩之, 浦浜喜一 : "マルチモーダルパターン識別器の教師なし学習", 電子情報通信学会論文誌, D-II Vol J81-D-II No.3 pp.573-582 (1998 年 3 月)

[14] I. Bloch : "Information Combination operators for data fusion : A comparative review with classification," IEEE Trans. Syst. Man and Cybermetics, Part A, vol.26, No 1, pp.52-67 (1996)

[15] Hakan Altincay, Mubeccel Demirekler : "An information theoretic framework for weight estimation in the combination of probabilistic classifiers for speaker indentification," Speech Communication 30, pp 255-272 (2000)