音声言語情報処理 19－10
(1997. 12. 12)

音素連接HMMを用いた尤度正規化に基づくワードスポッティングの検討

加藤 正治　堀 貴明　伊藤 彰則　好田 正紀
山形大学工学部

〒992 山形県米沢市城南4丁目3-16
Tel:0238-26-3367
Email:katoh@ei5sun.yz.yamagata-u.ac.jp

あらまし　近年の音声認識では，確率モデルに基づく隠れマルコフモデル(HMM)が広く利用されている．連続音声認識では，仮説の長さや位置が異なるためにHMMの尤度スコアを直接用いることには問題がある．本研究では，確率論的な考えに基づいて音素HMMの尤度を正規化する手法について検討する．具体的には，日本語の任意の音素並びを表現できる音素連接HMMを用いて，HMMの尤度を正規化する．音素連接HMMに基づく手法は，特別なモデルを作成する必要がなく，認識システムの枠組にとり込める，といった特徴がある．本研究で提案する手法をワードスポッティングに適用しその効果を評価する．

キーワード　尤度正規化，ワードスポッティング，スコア関数，音素連接HMM

A Study on Word Spotting based on Likelihood Normalization Using Phoneme HMMs

Masaharu KATOH　Takaaki HORI　Akinori ITO　Masaki KOHDA
Yamagata University

Jounan 4-3-16, Yonezawa-shi Yamagata, 992 Japan
Tel:0238-26-3367
E-mail:katoh@ei5sun.yz.yamagata-u.ac.jp

Abstract　In recent speech recognition, hidden Markov model (HMM) has been useful. We consider likelihood score of HMMs from a point of theory of probability. In continuous speech recognition, each hypothesis will have different length and position of speech segment. It affects the system performance by comparing the HMMs' scores directly. In this paper, we describe normalization of likelihood based on Bayes' theorem. To normalize likelihood, we use connected phoneme HMMs that allow Japanese syllable rule. In this method, we need no additional calculation to get scores, and we need no models except phoneme HMMs to the system. We apply it to the word-spotting, and obtain significant improvement of system performance.

key words　normalization of likelihood, word spotting, score function, connected phoneme HMMs

1 はじめに

近年の音声認識では，確率モデルに基づく隠れマルコフモデル (HMM) が広く利用されている．認識では，HMM の出力する尤度に基づいて仮説のスコアを評価する．しかし，連続音声中の単語を対象とする場合において，仮説に対応する音声区間が異なる場合は，尤度を直接比較することは問題がある．文献 [1] では，相互情報量に基づいて尤度を正規化する手法について述べられている．この文献では，エルゴディック HMM に基づいて正規化尤度が提案されていた．また，文献 [2] では，エルゴディック HMM の他に，音素 HMM，継続時間フレームなどから正規化尤度を定義する手法が提案されている．

本研究では，音素 HMM から音声の正規化尤度を推定する手法について検討する．具体的には，任意の音素並びを表現できる音素連接 HMM の状態スコアを用いて，正規化尤度を推定する手法について検討する．音素連接 HMM に基づく手法は，

- 特別なモデルを作成する必要がない．
- 認識システムの枠組みに取り込める．

といった特徴がある．

本研究で提案する手法をワードスポッティングに適用しその効果を評価する [3].

2 相互情報量に基づく HMM の尤度正規化

確率論的な考え方に基づく連続音声認識において，入力音声データ Oに対して，事後確率 $\mathrm{P}(W|O)$ を最大とする単語あるいは単語列 Wを求めることが目標である．$\mathrm{P}(W|O)$ はベイズの定理により次式で表すことができる．

$$\mathrm{P}(W|O) = \frac{\mathrm{P}(O|W)\mathrm{P}(W)}{\mathrm{P}(O)} \tag{1}$$

ここで対数を取ると

$$\log \mathrm{P}(W|O) = \log \mathrm{P}(O|W) + \log \mathrm{P}(W) - \log \mathrm{P}(O) \tag{2}$$

右辺第一項の $\mathrm{P}(O|W)$ は Wが与えられたときの観測系列 Oの条件付き確率である．これは Wを表す音響 HMM の尤度で求めることができる．

右辺第二項の $\mathrm{P}(W)$ は Wの現れる先験確率を表す．通常，言語モデルによって求めることができるが，これは観測系列 Oには無関係な量である．

右辺第三項の $\mathrm{P}(O)$ は音声の尤度を表し，相互情報量を求めるのに重要な要因となる．孤立単語認識においては，全単語仮説に対して観測系列 Oが共通であることから，$\mathrm{P}(O)$ はすべての候補で等しく，考慮する必要はない．しかし，連続音声中の単語を検出するような場合には，仮説毎に観測系列 (区間) が異なる．$\mathrm{P}(O)$ を無視することはできない．

相互情報量 $\mathrm{I}(O;W)$ は式 (2) の右辺第一項と第三項により次式で定義される．

$$\mathrm{I}(O;W) = \log \mathrm{P}(O|W) - \log \mathrm{P}(O) \tag{3}$$

相互情報量に基づいて，$\mathrm{P}(O)$ を推定し音響 HMM の尤度を正規化する．

3 ワードスポッティング

3.1 システムの概要

HMM に基づくワードスポッティングでは検出単語の前後の区間に対する尤度の推定が重要である．ここでは，日本語の音素列の規則に基づいて音素 HMM を連結した音素連接 HMM を用いて尤度の推定をする．音素連接 HMM は経路の組み合せによって，任意の単語系列を表現できる．つまり，検出する単語の直前までの音声の系列を音素連接 HMM によって表現することが可能である．単語 wの HMM の最終状態 $s_w(J_w)$ での経路スコアがしきい値 θ との間に，次式を満たす時刻 t について検出する．

$$f(t, s_w(J_w)) > \theta \tag{4}$$

単語の開始時刻は，経路をトレースバックすることで求める．終了時刻 t に対応する開始時刻$\tau(t)$ とおくと，各τ毎に，最大となるフレームを求める．

$$\hat{t}(\tau) = \max_{t:\tau(t)=\tau} f(t, s_w(J_w)) \tag{5}$$

3.2 音素連接 HMM

音素連接 HMM は日本語の音韻の規則（たとえば，子音は連続しない．母音または撥音で終了する等の条件）を基に音素 HMM を連結したものである．音素連接 HMM の説明図を図 1に示す．

音素連接 HMM の経路の組み合わせによって，あらゆる単語を表現できる．すなわち，最適経路と同じ音節列の未知語が存在すると解釈できる．ワードスポッティングには単語の直前に音素連接 HMM を連続したモデルを用いる．

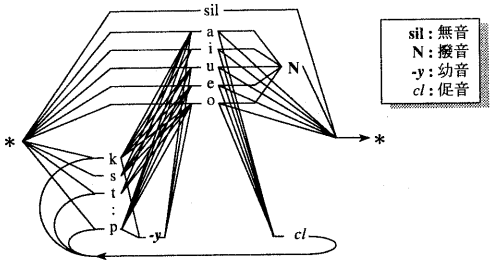

図 1: 音素連接 HMM

3.3 スコア関数

3.3.1 基本スコア関数

相互情報量に基づいて，P(O) を推定して音響HMMの尤度を正規化する．HMM の状態 s での，入力音声の時刻 t までの累積スコア (対数尤度) を次式で定義する．

$$f(t,s) = g(t,s) - h(t) \tag{6}$$

ここで，$g(t,s)$ は HMM の状態 s での累積経路スコアをあらわす．$h(t)$ は正規化尤度である．

3.3.2 継続時間によるスコア関数の正規化

経験的に仮説の長さに応じてスコアを正規化することは有効であるとされている．ここでは，継続時間によりスコア関数を正規化することを考える．

スコア関数を次のようにおく．

$$f_d(t,s) = \frac{f(t,s)}{t - \tau(t)} \tag{7}$$

ここで$\tau(t)$ は終了時刻 t に対する開始時刻で，分母は継続時間に相当する．

3.3.3 音素数によるスコア関数の正規化

仮説の長さについては，音素数に応じたスコアも考えられる．この場合，単語 wの音素数を $p(w)$ とおくと，

$$f_p(t, s_w(J_w)) = \frac{f(t, s_w(J_w))}{p(w)} \tag{8}$$

3.4 正規化尤度

3.4.1 音素連接 HMM の出力による正規化尤度

音素連接 HMM の出力するスコアに基づいて正規化尤度を求める．音素連接 HMM の最終状態を $s_p(K)$ における HMM の尤度スコアを

$$h(t) = g(t, s_p(K)) \tag{9}$$

で求める．これは，各時刻で任意の音素列が生成されるときのスコアを意味する．

3.4.2 全状態による正規化尤度

音素連接 HMM 内の最大尤度スコアを利用して正規化尤度の推定を行う．音素連接 HMM の状態を $s_p(k)$ とする．正規化尤度は

$$h(t) = \max_k g(t, s_p(k)) \tag{10}$$

で求める．つまり，正規化尤度は音素 HMM のネットワークの全ての状態から求める．

4 ワードスポッティングの実験

4.1 特定話者

特定話者のワードスポッティングの実験をする．標本化周波数 12kHz，32msec のハミング窓を分析周期 8msec で用いる．特徴ベクトルとして，対数パワーと 1〜12 次の LPC メルケプストラム，およびそれらの 1 次と 2 次の回帰係数 (39 次) を用いる．

HMM は，音素環境に独立な 28 種類を作成する．各音素 3 状態 8 混合．学習データは ATR の男性話者 1 名 (MHT)，重要語偶数番目 2620 単語．検出の対象単語は ATR キーボード対話（国際会議問合せ) から 3 音節以上の自立語を抽出し利用する．語彙 (V) は 305 単語．評価資料に，ATR 国際会議問合せ文 (SA セット）を利用する．継続時間は 0.427 時間で，検出対象単語の出現回数 (W) は 385 回．評価データに対する音素正解率は，78.2%である．

検出率 (A/W) と単位時間当たりの湧きだし誤り率 (FA/V/H) の関係を調べる．FA/V/H が 10 について表 1 に示す．

音素連接 HMM による正規化について，全音素 HMM と HMM の全状態の最大値に基づく正規化尤度を用いる場合の結果を図 2 に示す．全音素 HMM と HMM の全状態との結果の違いは，主に，検出位置ずれによるものである．検出の許容範囲を±80[$msec$] から±160[$msec$] まで伸ばすと検出率はそれぞれ AS: 94.1%，AP: 88.6%になる．pp 全状態の最大値に基づく正規化尤度を用いる場合について音素数，継続時間で正規化した場合の結果を図 3に示す．音素数，継続時間で正規化の効果はほとんど見られない．

表 1: 検出率の評価：ATR 音声データ

Normalize	Model	Base	Phoneme	Duration
AS	SD	89.5	90.9	90.5
AP	SD	71.1	—	—

FA/V/H=10.0
AS:All State, AP:All Phoneme, SD:Speaker Dependent.

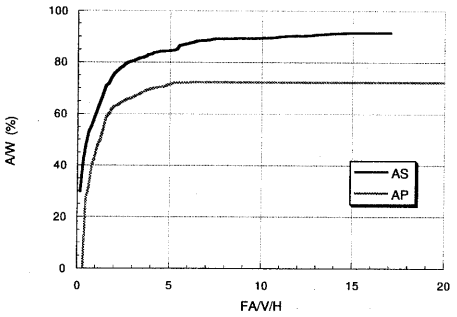

図 2: 正規化尤度の比較
AS:All State, AP:All Phoneme.

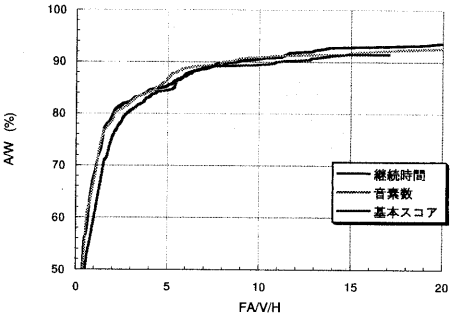

図 3: スコア関数の正規化の比較

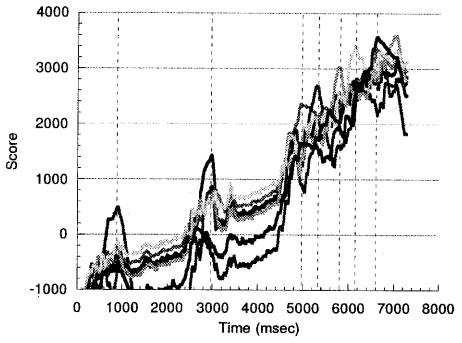

図 4: エルゴディック HMM による尤度正規化

入力音声「もしもし，こちら通訳電話国際会議事務局です」について評価スコアのグラフを示す．単語 HMM は {もしもし，こちら，通訳，電話，国際，会議，事務局} である．図 4 はエルゴディック HMM による正規化尤度を示す．音素連接 HMM による正規化スコアについて，図 5 に全音素 HMM による正規化スコア，図 6 に HMM の全状態による正規化スコアを示す．図中の縦の点線は単語境界を表し，ここで最大値をとる関数が良い．

エルゴディック HMM の場合は仮説の長さにしたがってスコア関数が上昇していることがわかる．これに対して，音素 HMM に基づく場合はスコアが安定して得られている．また，最適なモデル系列が正解系列と一致する場合はスコア関数の値が 0 となることが保証されている．

全音素 HMM の出力に基づいて正規化したスコアは 0 となる区間が長くできてしまい正確な検出位置を定めることができない．一方，HMM の全状態の出力に基づくいてスコアを正規化すると，関数の最大値をとる位置がはっきりして検出区間を定めることが可能であり，検出性能も良い．

4.2 不特定話者

不特定話者のワードスポッティングの実験をする．標本化周波数 16kHz．ここでは，ケプストラム平均正規化する．

HMM は，音素環境に独立な 28 種類を作成する．各音素 3 状態 8 混合．不特定話者モデル (SI) の学習データは ASJ の音素バランス文で男性 20 名，3000 文を用いる．話者適応モデル (SA) の学習には A-set, 50 文を用いる．

評価資料は学習テキストに独立な，話者 6 名の音素バランスの J-set を利用する．継続時間は 0.335 時間．検出対象として文中より 3 音節以上の名詞を抽出

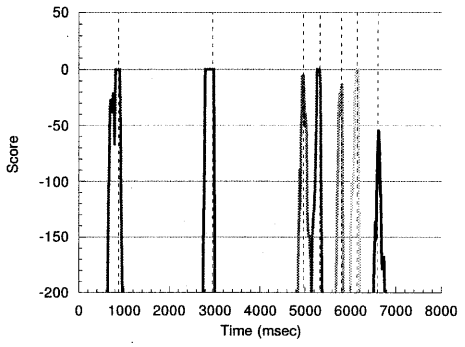

図 5: 全音素 HMM による尤度正規化 (AP)

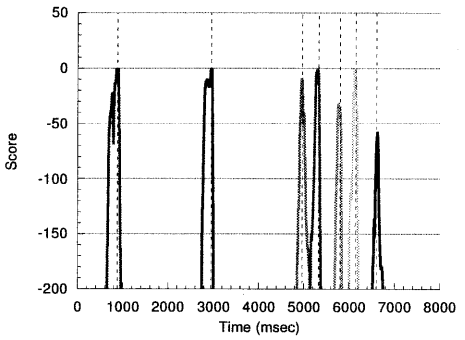

図 6: HMM の全状態による尤度正規化 (AS)

した．検出対象の語彙は 100 単語．出現回数は，107 回である．

評価データに対する音素正解率は，SI:67.1%，SA:77.8%である．

FA/V/H=10 について検出率を表 2 に示す．

音素 HMM の全状態で正規化するときの A/W と FA/V/H の関係について，不特定話者モデルを図 7，話者適応モデルを図 8 に示す．全音素 HMM の出力で正規化するときの A/W と FA/V/H の関係を図 9 に示す．

スコア関数は音素 HMM の全状態で正規化するものが良い．また，仮説の長さに対する正規化は継続時間に基づく場合が良い．

表 2: 検出率の評価：ASJ 音声データ

Normalize	Model	Base	Phoneme	Duration
AS	SI	57.1	65.6	75.4
	SA	77.8	84.7	89.9
AP	SA	69.2	76.4	83.4

FA/V/H=10.0
AS:All State, AP:All Phoneme, SI:Speaker Independent, SA:Speaker Adapted.

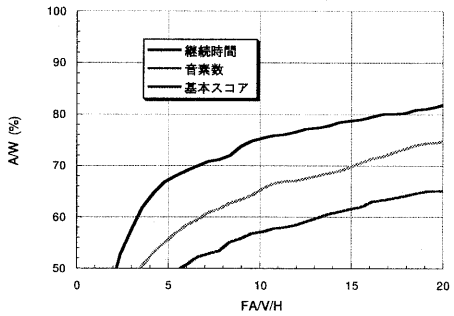

図 7: 単語検出率の評価 (SI:AS)

5 考察

HMM の尤度正規化に基づくスコア関数を用いたワードスポッティングの実験を行なった．実験結果から以下のことがわかる．

- 正規化尤度は，音素連接 HMM の全状態に基づく手法が良い．
- スコア関数は，継続時間で正規化すると良い．
- 検出性能として，特定話者で FA/V/H=10 のとき A/W=90.5%．不特定話者で FA/V/H=10 のとき A/W=89.9%．

6 まとめ

HMM の尤度の正規化について，音素連接 HMM の最大尤度に基づく手法を検討した．ワードスポッティングの実験に適用した場合について評価した．

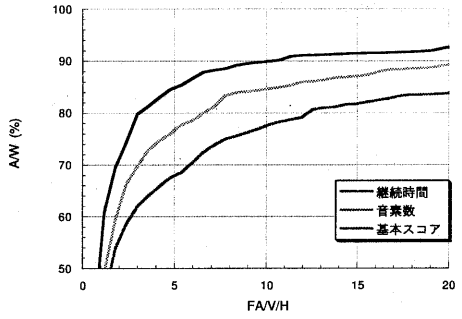

図 8: 単語検出率の評価 (SA:AS)

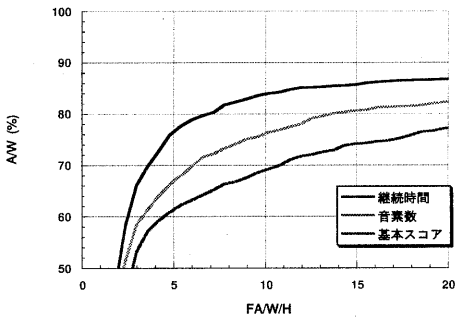

図 9: 単語検出率の評価 (SA:AP)

本手法は以下の特徴がある．音素連接 HMM に基づく手法は，特別なモデルを作成する必要がない．音素連接 HMM を未知の単語の推定に用いる場合の枠組みの場合，新たな計算を必要としない．

今回の実験では，継続時間や音素数といった単純な正規化しか行なっていない．音素の種類や識別率などに基づくより高度なスコア関数の検討も考えられる．さらに，HMnet に代表される環境依存モデルへの応用．連続音声認識システムへの応用などを検討したい．

参考文献

[1] K.Ozeki, "The Mutual Information as a Scoring Function of Speech Recognition", 信学技報 SP95-101, pp.53-60, 1995-12

[2] 小黒玲, 近藤法夫, 尾関和彦: "日本語連続音声認識におけるスコア関数の比較", 信学技報 SP96-9, pp.61-67, 1996-5.

[3] 加藤正治, 堀貴明, 伊藤彰則, 好田正紀 "音素連接 HMM に基づく尤度正規化を用いるワードスポッティングの検討", 音講論集 2-1-16, pp.79-80, 1997-9.