発声制約の緩和を可能にするサーチの検討

小高 俊之　天野 明雄　畑岡 信夫
（株）日立製作所　中央研究所

単語音声認識システムの利用者は，例えば「すずき」ではなく「す」「ず」「き」のように音節単位で発声をする場合がある。しかし，単語単位での発声入力を前提としている従来の単語音声認識システムでは，音節単位での発声を許容しないために，認識性能に著しい劣化が生じることになる。そこで本報告では，発声形態の制約を緩和し，様々な発声形態による入力音声に対して認識性能を回復できるサーチ手法として，累積尤度方式を提案する。累積尤度方式は，照合処理において，音節単位のような部分的な発声の終端が検出される毎に，照合データの一部を次の音声入力まで保持することにより，ポーズで分割された音節単位発声の単語音声や，一息では言いづらい長い単語の音声に対応する。人名100語の認識タスクでの評価の結果，従来の単語単位の発声形態を前提とした方式では全く正解が得られない入力音声に対して，71%の認識率が得られ，提案方式の効果が実証できた。

A Study of Search Algorithm to Accept More Flexible Utterance Input

Toshiyuki Odaka, Akio Amano, and Nobuo Hataoka
HITACHI, Ltd. Central Research Laboratory

Some users may speak a word by isolated syllables on word recognition systems. But ordinary word recognition systems which accept only isolated word utterances can't achieve good performance for isolated syllable utterances. In this report, we propose an improved search algorithm, named accumulated likelihood method, which can reduce utterance restriction and which can recover the recognition performance. This accumulated likelihood method keeps intermediate matching scores about search strategy after detecting an end point of one isolated utterance until the next utterance. Therefore, the isolated word recognition system can also deal with input words uttered by an isolated syllable style and by putting one or two pauses in long words. As the result of preliminary experiments, the proposed method achieved the recognition rate of 71% for isolated syllable utterance input speech which could not be recognized by the ordinary method.

1．はじめに

従来の単語音声認識システムにおいては，予め登録された単語が一塊りの音声として入力されることを前提としており，利用者にその前提に沿った発声をするような制約を課し

てしまっている。ところが，不慣れな利用者はそのような制約は知らずに，一音節ずつ発声する場合がある。特に，一度誤認識された場合に，より丁寧に発声するつもりで，音節単位でしっかりと区切って発声してしまう利用者も少なくない（「ひたち」ではなく，「ひ」「た」「ち」等）。こうなると，先の前提が崩れてしまい，音声認識システムは全く正解を出さなくなってしまう。このような状況でも，できる限り正しい結果を出せるようにしたいと考えたのが，本研究の動機である。

本稿では，利用者に対する発声の制約を緩めることが可能なサーチ手法として，累積尤度方式を提案し，音節単位発声の単語認識実験の結果を報告する。

2．従来の単語音声認識

2．1　音声検出による単語音声認識

音声検出を用いる一般的な音声認識システムでは，取り込まれた音声は，ある一定区間（フレーム）毎に音響分析され，ある特徴量に変換される。この特徴量に基づき，音声区間内か音声区間外かを判定し，予め登録されていた単語に対応した標準パターンに対して音声区間内の特徴量を用いた照合処理を行う。音声区間の終端が検出されると，照合結果を用いて，それまで入力された一連の音声に対する認識結果が求められる。

この方法では，音声検出の精度が認識率に影響し易く，音声検出に失敗すると認識率が低下する。ただし，照合処理は比較的処理量が多いため，音声検出処理を入れることで，全体的に処理量を抑えられるという利点がある。

2．2　ワードスポッティングによる単語音声認識

ワードスポッティングでは，音声が入力されているかどうかは関係なく常に照合処理をし続け，入力された音声中に含まれる単語を検出する手法である。ワードスポッティングでは，照合と結果判定を毎フレーム密接に連携させて処理を進める点が特徴と言える。音声検出を行っていないため，実際の発声とはずれた位置に結果がでるような湧き出しという現象が起きやすい。

2．3　従来方式の問題点

音声検出，およびワードスポッティングによる音声認識方式のいずれも，認識処理の単位はほぼ一塊りで発声された音声である。すなわち，単語認識システムとして使う場合は，単語単位の発声を前提としている。したがって，たとえ単語音声であっても，「す」「ず」「き」のように音節単位で間にポーズが挿入されたような音声が入力された場合は，正しい認識結果を出力することはほぼ不可能となる。「す」「ず」「き」のそれぞれの音節発声を１つの単語と見なして，それぞれに対応して最大３つの認識結果を出そうとしてしまう。こうした点が従来の音声認識システムを使いにくくしている一つの要因であり，従来方式の問題点の１つと言える。「ずずき」でも「す」「ず」「き」でも正しく認識できるロバストな音声認識システムが望ましい。

3．累積尤度方式によるロバスト音声認識

前記の問題点を解決するために提案する累積尤度方式についてHMMの場合を例に説明する。処理量の観点から，音声検出は行うこととする。また，照合部内で保持している標準パターンも従来と同じ単語認識用のＨＭＭをそのまま使う。

累積尤度方式は，音節単位のような部分的な発声の終端が検出される毎に，照合結果を

基に受理する単語を判定すると同時に，部分的にマッチしている単語を保留すべき単語として判定し，その単語の照合用の内部データの一部を次の音声入力まで保持する。その結果，ポーズで分割された音節単位発声の単語音声や，一息では言いづらい長い単語の音声認識が可能となる。

3．1　基本アルゴリズム

図１に提案方式の処理フローを，図２に処理イメージ図を示す。図１，図２に沿って基本アルゴリズムの詳細について説明する。

説明に用いる照合用の内部データを示す。

$S(w)$…単語ｗの中のＨＭＭ状態数

$L(w,i,t)$…単語ｗ中のｉ番目の状態のｔフレームにおける尤度

$Ls(w,i)$…単語ｗ中のｉ番目の状態の累積尤度

$Fs(w,i)$…単語ｗ中のｉ番目の状態の累積フレーム数

（１）尤度初期設定（図２の(1)に対応）

認識開始時に，単語ＨＭＭ状態の各内部データに初期値を設定しておく。例えば，認識対象を表す先頭の状態に尤度として確率値１，途中の状態に確率値０を設定する。また，全ての状態に対する累積尤度と累積フレーム数には０を設定しておく。すなわち，

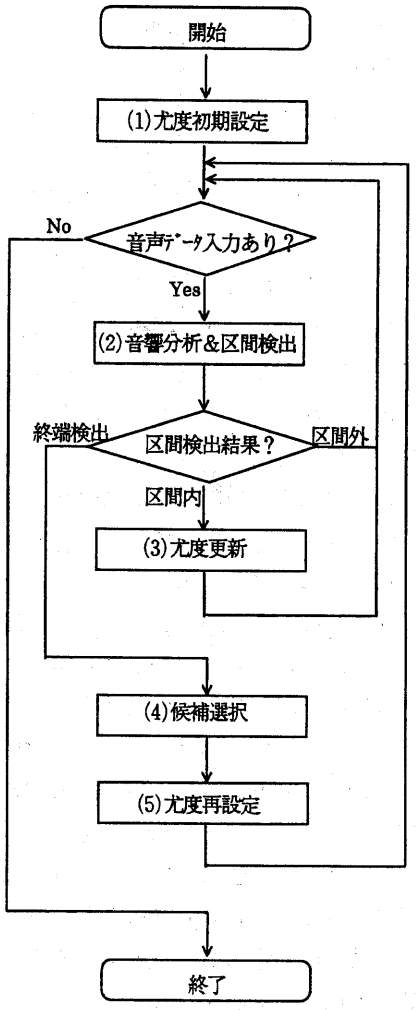

図１　累積尤度方式の処理フロー

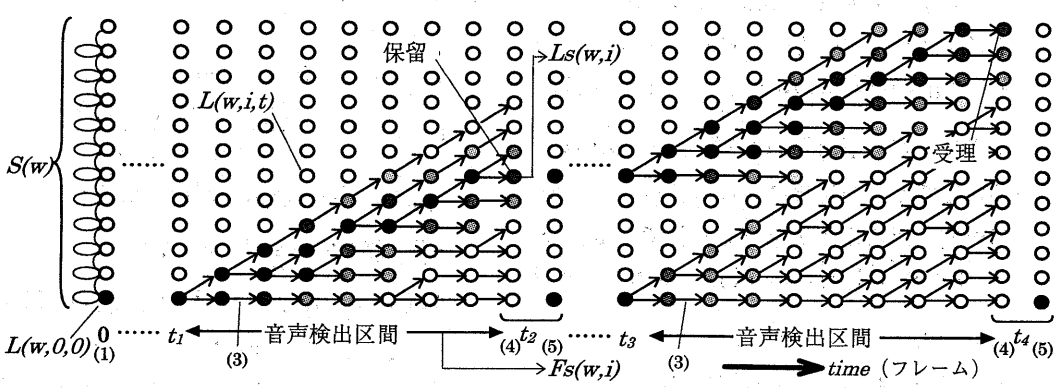

図２　累積尤度方式の処理イメージ（単語ｗのトレリス，○が状態，濃淡が尤度の大きさを表す）

$\forall w \in W$ について，

$L(w,1,0) \leftarrow 1$

$L(w,i,0) \leftarrow 0$，　$(2 \leqq i \leqq S(w))$

$Ls(w,i) \leftarrow 0$，　$Fs(w,i) \leftarrow 0$

$(1 \leqq i \leqq S(w))$

ここで，Wは認識対象単語の集合である。

（2）音響分析＆区間検出

音声データが入力され始めると，入力されたデータは，ある時間（フレーム）間隔毎に音響分析される。その結果の一部から音声区間かどうかを判定し，基本的には，音声区間内であれば認識対象に対する尤度計算を実行してから次フレームの音響分析へ進み，音声区間外であれば何もせずに次フレームの音響分析へと進む。ここで，音声区間の終端を検出すると，その時点でのHMMネットワーク内の状態の尤度に基づく候補選択へと進む。

（3）尤度更新（図2の(3)の→に対応）

尤度更新においては，従来法と同様にビタビアルゴリズム［1］により各状態の尤度更新を行う。本手法においては，さらに，ビタビアルゴリズムにより選択された遷移に沿って累積尤度と累積フレーム数をそのまま次状態へ伝搬させる。すなわち，

$\forall w \in W$，$1 \leqq i \leqq S(w)$ について，

$$L(w,i,t) \leftarrow \max_j (L(w,j,t-1) * P_{trans}(i,j,V))$$

$$\left(\forall j \in \{\text{状態 i への遷移元の状態}\}\right)$$

$$Ls(w,i) \leftarrow Ls(w,k),\ Fs(w,i) \leftarrow Fs(w,k)$$

$$\left(k = \arg\max_j (L(w,j,t-1) * P_{trans}(i,j,V))\right)$$

ここで，$P_{trans}(i,j,V)$ は状態 j から状態 i への遷移確率とその遷移における特徴ベクトル V に対する出現確率を乗算した値である。なお，$\arg\max_j(X(j))$ は，$X(j)$ を最大にする j の値とする。

（4）候補選択（図2の(4)に対応）

候補選択では，照合パラメータから算出される評価値に基づいて，第1の候補群と第2の候補群を選択する。第1の候補群とは，この時点までに入力された音声データが，認識対象（単語，文節，文等）として非常に尤もらしく，認識結果として受理すべき上位 N_1 個の候補の集まりである。具体的には，その認識対象を表現しているHMMネットワーク上の各状態の尤度を観察したときに，状態系列間で尤度最大の状態の位置（尤度ピーク位置）が状態系列の終端付近にある場合とする。一方，第2の候補群とは，この時点までに入力された音声データが，認識対象（単語，文節，文等）の先頭の一部分らしく，認識結果としての受理も棄却（リジェクション）も保留すべき上位 N_2 個の候補の集まりである。具体的には，その認識対象を表現しているHMMネットワーク上の各状態の尤度を観察したときに，状態系列間で尤度ピーク位置が状態系列の終端付近に到達せず，途中にある場合とする。

N_1 および N_2 は，尤度の大きい順に選択する際の個数を定数として決めても良いし，ネットワーク全体の状態の中で最大の尤度からある一定の尤度差内にある候補数と見なしても良い。後者は以前に提案した簡易リジェクション［2］の応用となり，ここでは後者を採用することとする。

すなわち，

$\forall w \in W$，$1 \leqq i \leqq S(w)$ について，

$$L'(w,i,T) \leftarrow \frac{L(w,i,T) + Ls(w,i)}{T + Fs(w,i)}$$

$$S_{\max}(w,T) \leftarrow \arg\max_i \left(L'(w,i,T)\right)$$

$$L'_{g\max}(T) \leftarrow \max_{w,i} L'(w,i,T)$$

$$\Delta L(w) \leftarrow L'_{g\max}(T) - L'(w, S_{\max}(w,T), T)$$

のとき，以下の判定を行う。

if $S_{\max}(w,T) > \theta_p$ and $\Delta L(w) < \Delta L_{th}^{Accept}$

then　wは受理（第１の候補群）

else if

$S_{\max}(w,T) \leqq \theta_p$ and $\Delta L(w) < \Delta L_{th}^{Continue}$

then　wは保留（第２の候補群）

else

wは棄却（リジェクション）

ここで，$L'(w,i,T)$は尤度$L(w,i,t)$を音声検出フレーム数($=T$)で正規化した値であり，累積尤度および累積フレーム数も含めて正規化する。$S_{\max}(w,T)$は終端検出されたフレームTにおいて単語wの状態列の中で尤度が最大となる状態の位置，θ_pはその状態位置に対する閾値である。θ_pは実験的に値を決める。また，$L'(w,i,T)$はその絶対値が話者や音声の発声環境に対して安定した評価値とはならないため，HMMネットワーク全体の中での最大尤度である$L'_{g\max}(T)$により補正した値$\Delta L(w)$を評価値として求める［2］。$L'_{g\max}(T)$の代わりに，日本電気の渡辺らの提案している任意の音節列の尤度を用いることも可能である［3］。さらにまた，ΔL_{th}^{Accept}と$\Delta L_{th}^{Continue}$は$\Delta L(w)$に対する閾値であり，それぞれ認識結果の候補としての受理判定，保留判定に使う。いずれも，実験的に値を求める。但し，$\Delta L_{th}^{Accept} \leqq \Delta L_{th}^{Continue}$である。

（５）尤度再設定（図２の(5)に対応）

候補選択が終わると，次の音声入力に備えた尤度の再設定を行う。ここでの尤度は先頭の状態に確率１を与えるとともに，第２の候補群として選択された認識対象の単語の中で，最大尤度が得られていた状態系列の途中の状態$S_{\max}(w_c,T)$も，確率値１を与える。さらに，同じ$S_{\max}(w_c,T)$に対応する累積尤度に，その$S_{\max}(w_c,T)$での尤度を加算し，それ以外の状態及び第２の候補群に属さない各単語内の全ての累積尤度は０に再設定する。また，累積フレーム数についても同様な再設定する。

すなわち，

$\forall w \in W$について，

$L(w,1,0) \leftarrow 1$

$\forall w_c \in$｛Wのうち保留する単語｝について，

$L(w_c,S_{\max}(w_c,T),0) \leftarrow 1$

$L(w_c,i,0) \leftarrow 0$,

$(2 \leqq i \leqq S(w_c) \wedge i \neq S_{\max}(w_c,T))$

$L(\overline{w}_c,i,0) \leftarrow 0$,　$(2 \leqq i \leqq S(\overline{w}_c))$

$Ls(w_c,S_{\max}(w_c,T)) \leftarrow Ls(w_c,S_{\max}(w_c,T)) + L(w_c,S_{\max}(w_c,T),T) \cdots (a)$

$Fs(w_c,S_{\max}(w,T)) \leftarrow Fs(w_c,S_{\max}(w,T)) + F$

$Ls(w,i) \leftarrow 0$,　$Fs(w,i) \leftarrow 0$

$(w \neq w_c \vee i \neq S(w_c))$

こうした尤度の再設定により，再度音声区間を検出した際に，第２の候補群に対して途中から照合を再開することができる。

以降，（２）から（５）を繰り返しながら，（４）で受理された認識結果は随時出力する。

３．２　音節単位発声単語音声認識への適用

本節では，前節で示した累積尤度方式の基本アルゴリズムを音節単位発声の音声認識に適用する場合の問題点を挙げ，その対策案を説明する。

累積尤度方式では，単音節のような認識対象単語の部分的な発声を，単語モデルの標準パターンで無理矢理に照合しようとするため，認識単位モデル（音節モデル，半音節モデル，音素片モデル等）によっては照合で不整合が起きるため，認識率が悪くなる。

我々は認識単位モデルとして音素片モデルを採用しており，本モデルでもその不具合が起きる。音素片モデルは，音声を，子音定常部（C モデル），母音定常部（V モデル），遷移部（CV モデル，VC モデル，VV モデル）に分けて，モデル化している。音素片モデルで音節発声の認識をするのであれば，本来以下のようなモデルを標準パターンとして用意

すべきである。“すずき”という単語を例に挙げると，

「す」には，［ ＊s s su u u＊ ］

「ず」には，［ ＊z z zu u u＊ ］

「き」には，［ ＊k k ki i i＊ ］

しかし，単語発声を認識するための次のような標準パターンを使っている。

「すずき」には，

［ ＊s s su u uz z zu u uk k ki i i＊ ］

“＊”部分は無音部との境界を示しているが，前に示した音節単位の場合と比較してわかるように，波線を付けた部分が異なっており，この部分で照合に不整合が生じ，尤度が抑えられてしまう。

（改良１）継続時の定数ボーナス尤度加算による尤度補正

前記の問題点を解消するために，保留候補に対して，累積尤度 Ls を更新する時点で，定数ボーナス尤度（L_{bonus}）を加算することを考える。すなわち，前記（a）式を，

$$Ls(w,S_{\max}(w,T)) \leftarrow Ls(w,S_{\max}(w,T)) + L(w,S_{\max}(w,T),T) + L_{bonus}$$

とする。これにより，照合不具合による尤度の減少をある程度補正でき，認識率の改善が期待できる。

（改良２）継続時の累積フレーム数で重み付けしたボーナス尤度加算による尤度補正

結果判定部において，入力フレーム長がより長い単語を優先させることを考える。そのため，累積尤度 Ls の更新時に，累積フレーム数 Fs に比例したボーナス尤度を加算することを考える。すなわち，前記（a）式を

$$Ls(w,S_{\max}(w,T)) \leftarrow Ls(w,S_{\max}(w,T)) + L(w,S_{\max}(w,T),T) + Fs(w,S_{\max}(w,T),T) \times L_{bonus}$$

とする。これにより，過去の入力フレーム長が長い単語が優先的に候補に挙がるようになり，全体的に認識率の向上が期待できる。

4．音節単位発声による単語音声認識実験

4．1　実験条件

人名 100 語を認識タスクとして，提案した累積尤度方式の効果を調べた。使用した音声資料および分析条件は，表１および表２の通りである。また，使用した HMM は分布を 256 で量子化した半連続型 HMM で，2 状態 3 混合の音素片モデル 387 種類からなる。

表１　音声資料

学習データ	216 音韻バランス単語×男性 30 名×1 回
評価データ	人名 100 語中先頭 10 語×男性 10 名×1 回 （但し，電子協音声データベースの音節データを組み合わせて作成した音節単位発声）

表２　分析条件

サンプリング	12kHz，16bit
分析窓長／周期	20ms／10ms
特徴量	14 次 LPC ケプストラムとその回帰係数 14 次

4．2　実験結果（1）

ボーナス尤度（L_{bonus}）の加算による尤度補正（改良1）を含んだ累積尤度方式による評価結果を図3に示す。図3では，ボーナス尤度をパラメータとした結果である。θ_p，ΔL_{th}^{Accept}，$\Delta L_{th}^{Continue}$も認識性能に影響するパラメータであるが，今回は最適に近くなるような値として，θ_p＝単語の先頭から70%の位置，ΔL_{th}^{Accept}＝10,000，$\Delta L_{th}^{Continue}$＝20,000に固定した。詳細は別途検討したい。また，本実験において，評価単語途中の段階で誤った単語を受理してしまった場合を受理誤りとし，途中誤受理率を以下の通り定義する。

$$\text{途中誤受理率} = \frac{\text{評価単語途中での単語誤受理総数}}{\text{評価単語途中の音節総数}}$$

図3において，ボーナス尤度が0の場合が尤度補正なし（改良前）に相当するが，従来認識できなかったデータ（単語認識率0％）に対して認識率29%が得られた。さらに，ボーナス尤度（L_{bonus}）の導入による提案方式の効果があることも明らかである。ボーナス尤度の値を90,000付近に設定すると最も高い認識率（67%）が得られている。受理誤りに関しては，例えば「わ」「た」「な」「べ」という発声の途中で認識結果として，「わだ」が検出されてしまったり，あるいは「し」だけで「いしい」として検出されてしまうというような現象があったが，ボーナス尤度（L_{bonus}）の誤検出率に対する影響はほとんどないと言える。

4．3　実験結果（2）

図4に，累積フレーム数で重み付けしたボーナス尤度（$Fs(w,S_{\max}(w,T),T)\times L_{bonus}$）の加算による尤度補正（改良2）の実験結果を示す。

改良1の場合と比較すると，湧出率はほとんど変わらないが，認識率は改善されている（最高で，71%）。人名100語の認識タスクの場合，単語単位の発声であれば少なくとも90%程度の認識率は得られ，その場合との差がまだある。単語単位の標準パターンを用いているため完全に差を埋めるのは困難であるが，少しでも認識率を改善できるように，音響モデル自身の構造も含めた他のパラメータに関して再検討が必要と考える。

5．まとめ

5．1　結　論

従来の単語発声を対象とした音声認識システムをベースにして，音節発声でも対応可能なサーチ手法として累積尤度方式を提案し，人名100語のタスクで評価した。この結果，次のことが判明した。

（1）人名100単語のタスクで，内10単語×男性10名の音節単位発声の音声データを評価した結果，従来の方法では全く正解が得られない（単語認識率0％）データに対して，71％の単語認識率が得られ，その効果が実証できた。

（2）通常の単語単位発声の音声データの場合の認識率との差が20%程度残った。

5．2　今後の課題

今回の評価実験等を通して，今後の課題として以下の3つが挙げられる。

（1）単語検出誤りの抑制（「わたなべ」で「わだ」を抑制，等）

（2）促音への対応（「はっとり」と「はとり」の区別，等）

（3）音響モデルの改善

（1）と（2）に関しては，時間情報を利用した対応を行いたいと考えている。（3）に関しては，提案方式にとっては特に音節単位のモデルの精度が重要であり，様々な音響モデリングを検討したい。

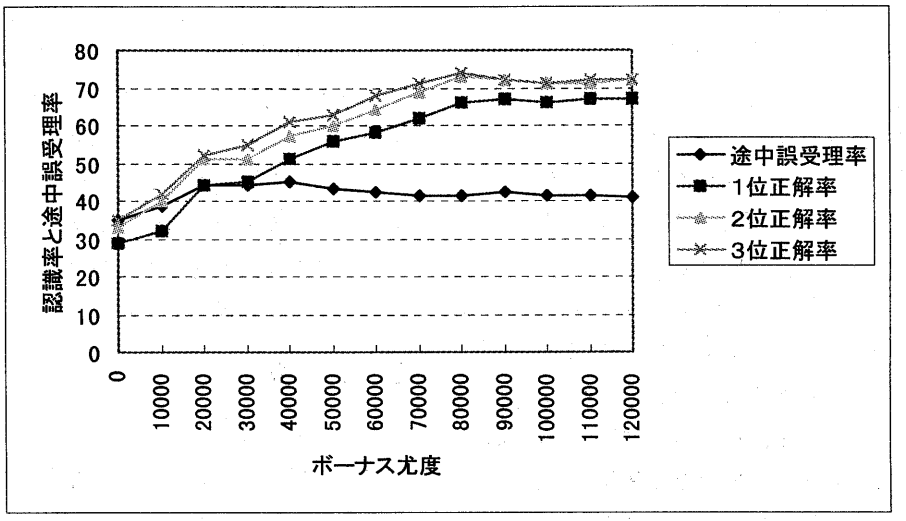

図3　ボーナス尤度加算利用の累積尤度方式による音節発声単語認識結果

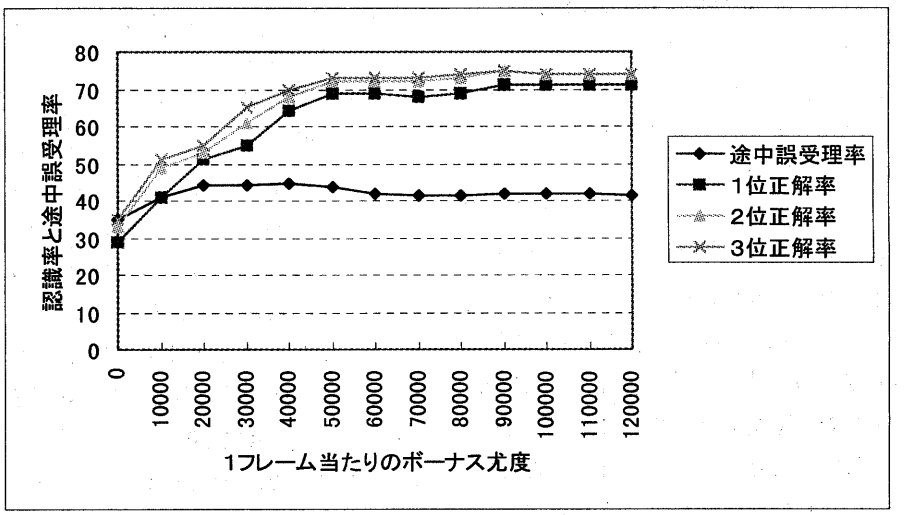

図4　継続長重み付きボーナス尤度加算利用の累積尤度方式による音節発声単語認識結果

参考文献

[1] X.D.Huang, 他3名：HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION, Edinburgh, 1990.

[2] 小高，天野：音声認識におけるリジェクト方式の検討，日本音響学会講演論文集 3-2-5, 1995.8

[3] 渡辺，塚田：音節認識を用いたゆう度補正による未知発話のリジェクション，電子情報通信学会論文誌 D-II, Vol.75-D-II, No.12, 1992.12