TR-SLT-0008

単語を構成する音素の連接特性を考慮した音声タイプライタの検討

Studies on a speech typewrite using phonetic word modeling

佐藤 隆　　大西茂彦
Takashi SATO　　Shigehiko ONISHI

匂坂芳典　　山本博史
Yoshinori SAGISAKA　　Hirofumi YAMAMOTO

2002.3.15

個別的タスクに依存しない音声タイプライタの認識精度が向上するような言語モデルを考察するにあたり，近接する音素を拘束する音素結合と単語レベルの音素列を律束するような制約を検討し，制約導入による音素正解率の変化を調べる認識実験を行った．今回は，単語の音素列の両端の音素によってクラスを定義し，階層化言語モデルを適用した．まず，音素 0-gram（全等確率）に結合音素を加えていったところ，最大約 15%上昇し，局所的な音素列に対する制約が性能向上に高い効果を示すことが確認できた．また，上層に始終端クラス 2-gram，下層に Word-structure モデルによるサブワードモデルを持つ階層化言語モデルを適用した場合，音素 0-gram を常に上回る結果となり，性能向上の効果が認められた．このとき上層のクラスを分けず，下層の Word-structure モデルだけを適用した場合でも階層化言語モデルに近い結果が得られ，単語レベルの音素列のモデル化を行っている Word-structure モデルが，認識性能を大きく向上させる制約として働くことが予想された．さらに上層の始終端クラスのような学習を伴わない大まかなクラス 2-gram でも，ある程度の制約力があり，音声タイプライタの制約として有効に機能するものと予想された．

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288 京都府相楽郡精華町光台二丁目 2 番地 2　TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-0288,Japan
Telephone:+81-774-95-1301
Fax　　:+81-774-95-1308

1. はじめに

一般に，入力音声に対し音素列を出力するものを音声タイプライタと言い，従来，音素N-gramやシラブルN-gramによって駆動させることが多い．これは，大語彙連続音声認識のように厳密な言語モデルを用いることなく音素列が得られ，個別的なタスクに依存せず，文法的に制約の緩やかな「話し言葉」をタスクとする場合にも適していると考えられる．しかし音素N-gramは，言語モデルとして制約が弱く，音声タイプライタの性能はあまり高くない．性能を上げるには大語彙コーパスに基づいて厳密に学習を行えばよいが，反面，学習時に使用した特定のタスクへの依存性が強まってしまう．

今回の実習では，音声タイプライタの性能を向上させる方法として，高頻度で出現する音素列を1つの結合音素として扱い，音素N-gramモデルを構築する際の学習データ中に加えることを試みる．

また，単語レベルの音素列を律束する制約の導入を検討する．今回は，データ自体の特徴による指標によってクラスタリングを行い，音素N-gramを構築することを考える．本実習では，単語間の連接部位である各単語の始端と終端の音素を指標とする始終端クラスを導入することにする．

さらに，言語モデルを構築する際，始終端クラスN-gramモデルを上層に，下層に各クラスに対応するサブワードモデルを持つ階層化言語モデル[1]を適用することを試みる．このとき，下層のサブワードモデルには，高い精度で音素列を推定することができるWord-structureモデル[1]を適用し，音声タイプライタの制約について検証を行う．

2. 先行研究による知見

音声タイプライタは，従来，音素 N-gram モデルやシラブル N-gram モデルといった言語モデルによって駆動させるが，音素やシラブルのような微細な単位の履歴のみによる制約では言語モデルとしての拘束力が弱く，認識精度はあまり高くならない．認識精度を上げるには，単語モデルのように大規模コーパスに最適化して学習を行ったモデルを導入すればよいが，厳密に学習を行えば行うほど，学習に使用した特定のコーパスに依存するようなモデルとなってしまう．

今回の実習では，タスクに依存しない音声タイプライタの認識精度を向上させる方法として，高頻度で出現する音素列を 1 つの結合音素として音素と同等に取り扱い，音素 N-gram を生成する際に学習データとして加えることを試みる．結合音素によって近接する音素列を拘束することで，局所的に信頼性の高い確率値が得られることが期待される．

また，より範囲の大きい制約として，言語モデルを構築する際に，単語レベルの音素列の連接を律束するような制約を導入することを検討する．今回は，モデルの学習に用いる語彙単語に対し，事前にトップダウン的に与える品詞情報や学習の結果から得られる情報ではなく，データ自体の特徴からボトムアップ的に得られる指標によってクラスタリングを行い，音素 N-gram を構築することを考える．本実習では，単語間の連接部位である各単語の始端と終端の音素によって単語間の連接に偏りが存在すると考え，単語の両端の音素の組み合わせを指標とする始終端クラスを導入する．

さらに，制約を言語モデルとして形成するにあたり，始終端クラス N-gram モデルを上層に，下層に各クラスに対応する単語の音素列を導出するサブワードモデルを持つ階層化言語モデル[1]を適用することを試みる．下層のサブワードモデルの構築には，高い精度で単語レベルの音素列を推定することができる Word-structure モデル[1]を適用する．また，クラスモデルやサブワードモデルを構築する際には，接続の方向性を考慮して効率よくモデル化を行うことができる多重クラス N-gram モデル[2]を適用することにする．

なお，タスク依存性の検証を行うには，単語を構成する音素 N-gram の学習に，クラス N-gram とは別のタスクのコーパスを用いる必要があるが，今回は実験に使用できるデータの都合により，同一タスクのデータにおいてのみの検証とし，タスク依存性についての検証は行わない．

3. クラスの定義方法

各単語の音素列の始端と終端の組み合わせによって，単語のクラスタリングを行う．

（例）

[予約] j o j a k u　　→「j__u」クラス（j で始まり u で終わる単語の集合）
[会議] k a i g i　　→「k__i」クラス（k で始まり i で終わる単語の集合）
⋮

このようにして定義したクラスを「始終端クラス」と呼ぶ．

4. 言語モデルのモデリング手法

統計的言語モデルとして広く使われている単語 N-gram モデルにおいて，単語 N-gram モデルとクラスモデルを補間等により組み合わせたモデルは，それぞれのモデルを単独で使う場合よりも高い精度を示すことが多い[3]．今回，始終端クラスを言語モデルに導入するにあたって，上層に始終端クラス 2-gram，下層に各クラスに対応したサブワードモデルを持つ階層化言語モデル[1]を適用する．このモデルは，上下層のそれぞれが独立して統計的制約を持つため，上層では単語間の連接特性を，下層ではそれぞれのクラスに含まれる単語を構成する音素の連接特性を表現することが出来る．

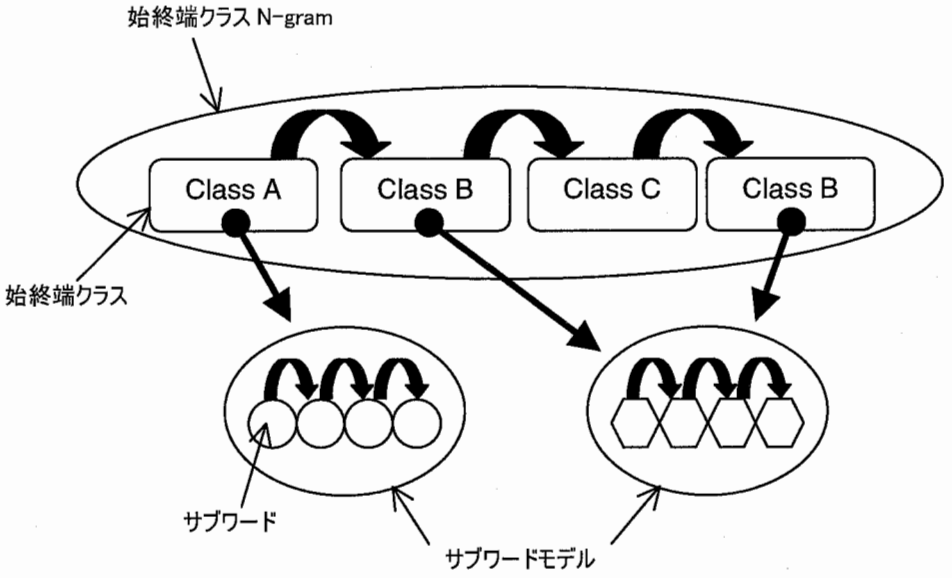
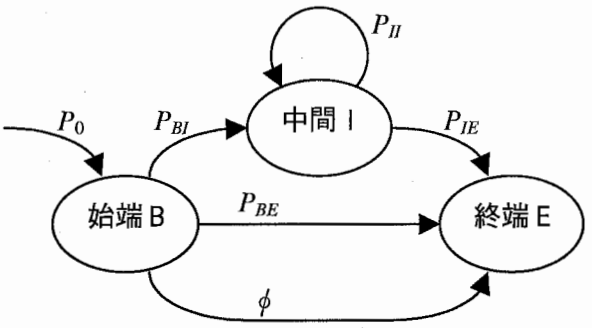


図 1 上層に始終端クラス N-gram，下層にサブワードモデルを持つ階層化言語モデル

5. サブワードモデルのモデリング手法

階層化言語モデルにおいて，上層の始終端クラスに対応する下層のサブワードモデルを導出する方法として，サブワード列を始端，中間，終端に分けてモデリングを行う Word-structure モデル[1]を適用する．このモデルは，サブワード列の両端にクラス特有の特徴が存在すると仮定したモデルであり，始終端クラスでは始端と終端の組み合わせが同じである単語集合の音素列を扱うので，今回のサブワードモデルのモデリングに適している．



（但し，ϕは NULL 遷移）

図2 サブワードモデルの構築に用いる Word-structure モデル

6. 音素の始終端で単語のクラスタリングを行う際の注意点

今回，クラスタリングの対象とする単語は，ATRSPREC（後述）の単語辞書（Lexicon）を用いた．この形式には1つの単語に複数の読み（音素列）が既定されている．

（例）25343　[相当]　s o {o|u} t o {o|u} → 4通り
↑
単語ID

このため，音素表記の括弧部を展開し，別々の音素列として扱う必要がある．単語の音素列の展開方法を下に示す．

- 全ての場合の音素列を列挙し，それぞれ別の単語として扱う
- 展開したそれぞれの音素列の出現頻度は，音素表記のバリエーション数で等分する
- 単語の音素列末の {l-} は無視する

25343_0　s o o t o o
25343_1　s o u t o o
25343_2　s o o t o u
25343_3　s o u t o u
} 元の出現回数N ⇒ 展開後それぞれ4/N

- 単語の音素列途中の {l-} は以下の2通りに分ける．この場合の音素列の出現頻度はそれぞれ1/2とする．

（例）21416　[やーあの]　j a a {l-} a n o

1. {l-}前後の音素列を結合する
2. {l-}前後の音素列を別の単語の音素列とみなし，その遷移をカウントする

21416_0　j a a a n o
21416_1_0　j a a
21416_1_1　a n o
} 元の出現回数N ⇒ 展開後それぞれ2/N

＜参考資料＞ 複数の音素表記を持つ単語

- 3,778 語（単語辞書 LEX.ALL(27,498 語)中の 13.79%）
- 音素表記のバリエーション数：2～32 通り

表記数	単語数
32 通り	2 個
16 通り	2 個
12 通り	1 個
8 通り	41 個
6 通り	2 個
4 通り	514 個
3 通り	14 個
2 通り	3202 個

（例）32 通りの音素列のバリエーションを持つ単語

[奈良交通定期観光営業所]

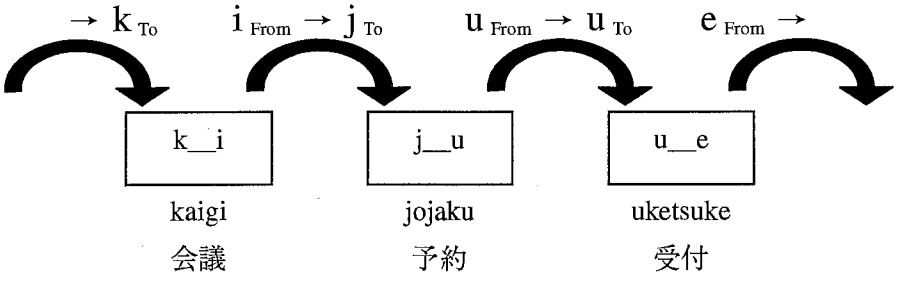
n a r a k o {o|u} ts u u t e {e|i} k i k a ng k o {o|u} e {e|i} g j o {o|u} sh j o {|-}

この処理の結果，展開前に 27,398 語だった語彙（音素列）数は，展開後に 36,348 語となった．

7. モデルの構築

7.1. 多重クラス N-gram モデルの適用

モデルの構築には，接続の方向性を考慮した多重クラス N-gram モデル[2]を適用する．このモデルはクラス N-gram において，単語の前方への接続性と後方への接続性を別々にとらえ，各単語に対し，前方と後方の 2 種類のクラスを割り当てるもので，効率的かつ信頼性の高いクラスタリングが行えることが報告されている[2].



※例えば，クラス j__u は，前方属性 j To と後方属性 u From の 2 つを持つことになる

7.2. 多重クラスによるクラスタリングとクラス 2-gram 構築

単語の前方属性である始端音素で To クラスを，後方属性である終端音素で From クラスをそれぞれ定義してクラスタリングを行う．クラスタリングの結果を以下に示す.

To class　: 26 クラス

From class : 6 クラス（指標である終端音素は母音となった）

同様に，要素をシラブル単位とし，単語の始端シラブルで To クラスを，終端シラブルで From クラスをそれぞれ定義してクラスタリングを行ったところ，以下のようになった.

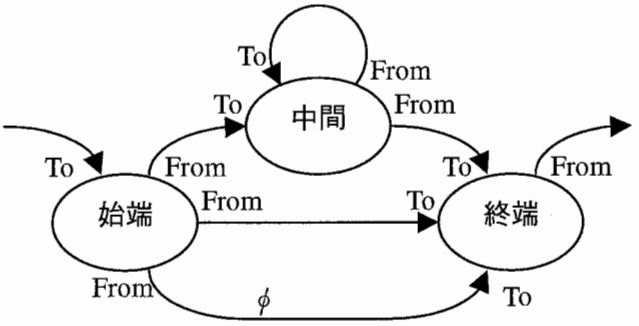
To class　: 120 クラス

From class : 96 クラス

この始終端クラスを用いて，階層化言語モデルの上層部にあたる多重クラス 2-gram を構築する.

7.3. 多重クラスによるサブワードモデル構築

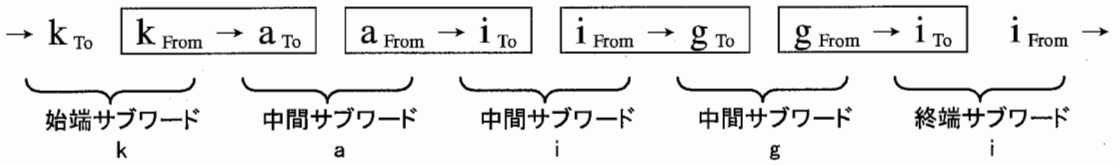
次に，Word-structure モデル[1]を適用し，始終端クラスに対応したサブワードモデルを導出するが，この際にも多重クラスの考えを導入する．つまり，単語を構成するサブワード（音素や結合音素）にも前方属性と後方属性があるものとして扱う.



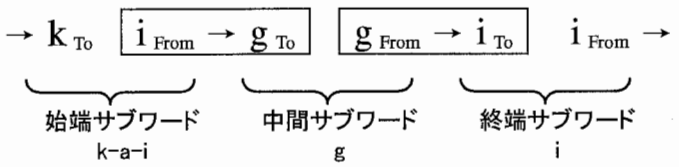
（但し，ϕは NULL 遷移）

図 3 サブワードに多重クラスを適用した Word-structure モデル

例えば，上層の始終端クラス k__i に対応するサブワードモデルを構築する場合，「会議」は



という遷移をカウントすることになり，またこのとき，結合音素 k-a-i が存在する場合には



と遷移する可能性もある．

このような形で，サブワード間の遷移をカウントし，サブワード 2-gram を構築していく．

今回の実習では，階層化言語モデルの下層であるサブワードモデルの構築に用いる学習用データに，上層で用いたものと同じデータを用いたが，本来は，上層のクラス N-gram とは独立した制約力を持たせるために，一般的な単語データベースを別に用意し，学習に用いる必要がある．そのため，今回はタスク依存性に関する検証は行わない．

以上のデータを用いて，上層に始終端クラス 2-gram，下層にサブワードモデルを持つ階層化言語モデル[1]を構築した．

8. 音声認識実験

8.1. 実験条件

認識器：ATRSPREC r08r01
実験マシン搭載メモリ：2GB
結合音素の導出，始終端クラス 2-gram およびサブワードモデルの学習セット
　ATR 旅行対話 V8 データ
　　総単語数　　1,606,951
　　異なり単語数　　16,355
評価セット
　訓練セットに含まれない 42 片対話 6,326 語（オープン）[S1S2S4]

8.2. 実験項目

各制約を表現するモデルをそれぞれ用いて音素正解率を求め，比較を行うことで，各制約の影響や効果を調べる．

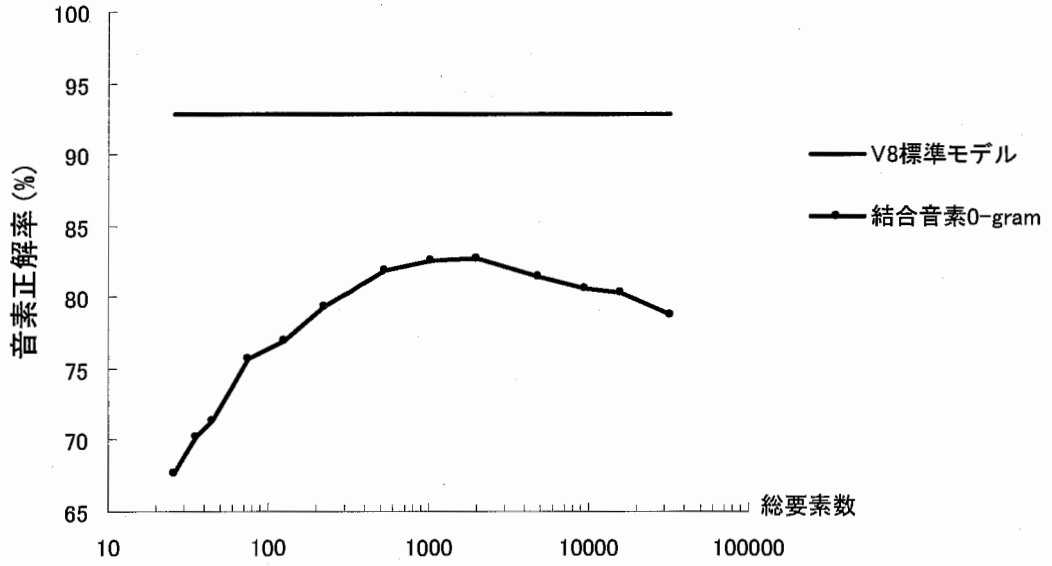
1. 音素 0-gram に結合音素を導入
2. 単語登録方式で始終端クラスを導入
3. 始終端クラス 2-gram とサブワードモデルによる階層化言語モデルを適用
4. 単一クラスのサブワードモデルのみ適用

8.3. 結合音素の性能実験

まず，なるべく学習を行っていない音声タイプライタの状態から確認する意味で，今回は，音声タイプライタの音素 N-gram として，N=0（全て音素の出現確率が等確率）の場合について実験を行った．これに結合音素を加えていくことにより，音素正解率の変化を測定し，結合音素による制約力を調べる．

8.3.1. 結合音素導入結果

実験結果を以下に示す．



結合音素を加えない状態（総要素数 26）での音素正解率は 67.7%で，総要素数 1,989 個の時に最高値 82.75%となった．以降，結合音素を加えても音素正解率は上がらず，かえって悪くなった．

8.3.2. 結合音素導入の考察

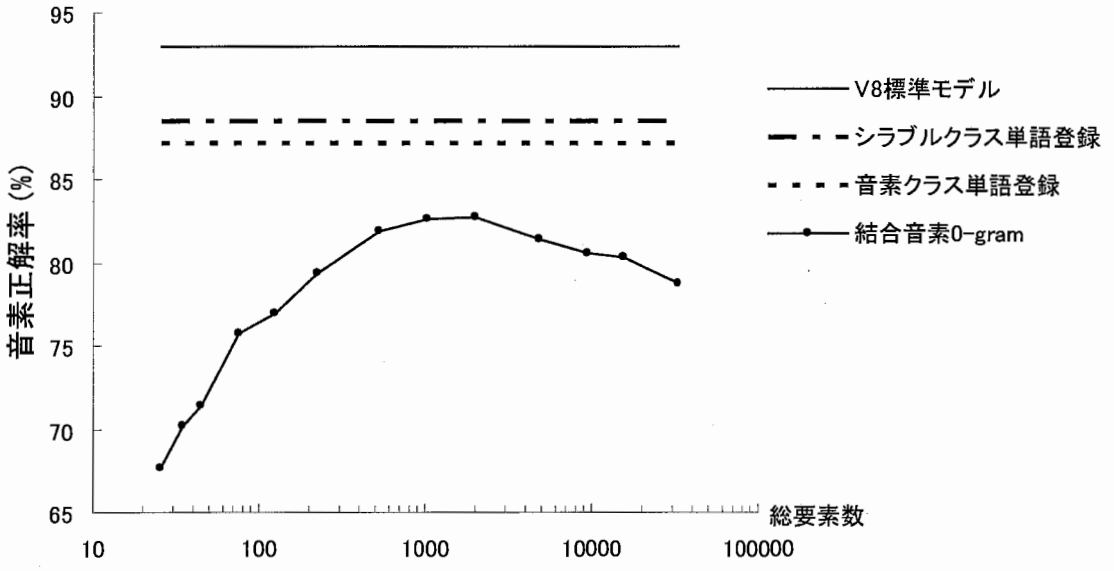
音素 0-gram に結合音素を加えていくことで，音素正解率が最大で約 15%改善した．しかし，その後も結合音素を加えていくと，逆に正解率は低下し始めた．これは，出現頻度の高い順に結合音素を追加していく過程で，徐々に出現頻度の低い音素列まで含まれるようになり，逆に正解に対して妨げとなる候補が増えていったためと考えられる．

8.4 始終端クラスによる性能実験

階層化言語モデル導入による性能を調べる前に，単語を始終端クラスで分類した単語登録方式での認識実験を行い，始終端クラス 2-gram 導入による制約力を確かめる．始終端は音素とシラブルの両方の場合についてそれぞれ調べた．

8.4.1. 始終端クラス導入結果

実験結果を以下に示す．



コーパスに最適化してクラスタリングを行った V8 標準モデルの音素正解率は 92.9%であった．始終端クラス導入による音素正解率は，両端音素によるクラス分けで 87.19%，両端シラブルによるクラス分けで 88.45%となった．また音素に比べ，シラブルで分けたクラスの方が音素正解率が高かった．

8.4.2. 始終端クラス導入の考察

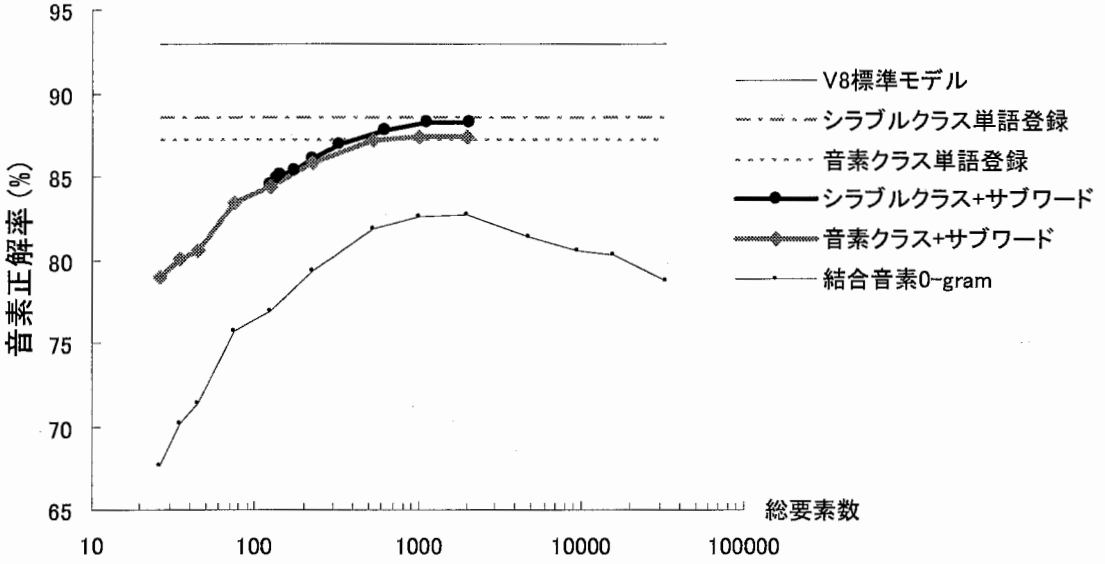
始終端クラス導入による音素正解率は，V8 標準モデルと結合音素 0-gram の中間に位置する形となった．コーパスに最適化した V8 標準モデルには及ばないものの，始終端クラスのような比較的緩やかなクラスタリングでも，ある程度の制約力が期待できることが分かった．また，クラスタリングを行わず 1 つのモデルだけで動かすよりも性能向上に効果があると予想できる．また，シラブルで分けたクラスの方が音素で分けたクラスよりも音素正解率が高かったが，この結果は，日本語の音声単位がシラブルベースであるという一般的な知見に反しないものであり，妥当であると思われる．

8.5. 階層化言語モデルによる性能実験

始終端クラスとそのサブワードモデルによる階層化言語モデルの性能を調べる．音素の場合とシラブルの場合の両方について，それぞれ調べた．

8.5.1. 階層化言語モデル導入結果

実験結果を以下に示す．



階層化言語モデルによる音素認識率は，結合音素 0-gram に比べ，平均して 6～7%程度高かった．また，今回実施した実験の範囲では，結合音素が増えるにつれ，始終端クラスに単語を登録して行った認識実験の性能に近づく傾向が見られた．本モデルでは，シラブル単位の正解率の方が音素単位よりも若干上回る傾向が見られた（平均差 1%未満）．

8.5.2. 階層化言語モデル導入の考察

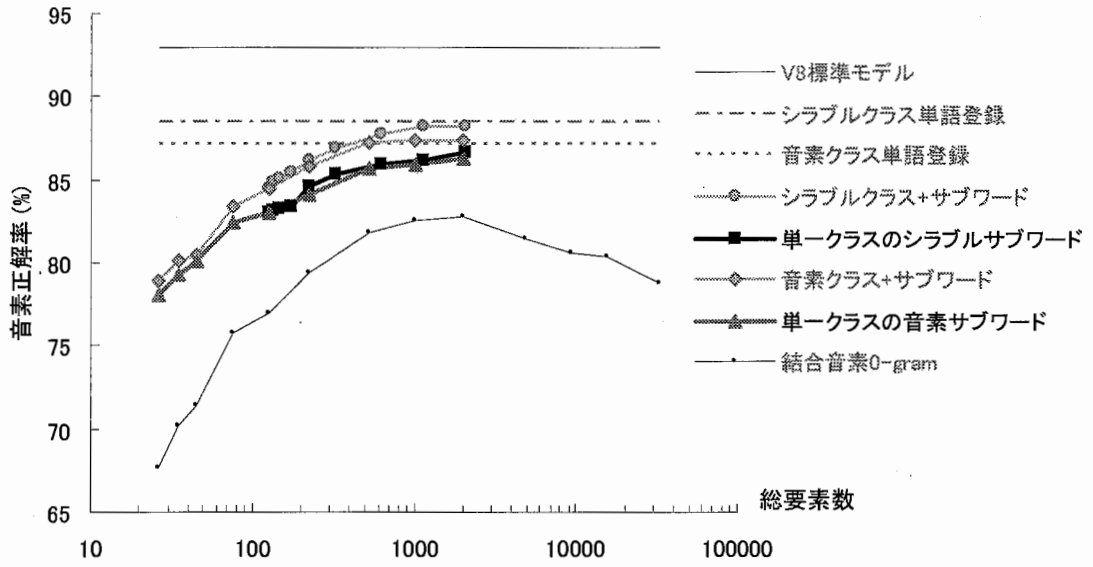
階層化言語モデルを適用した本モデルでは，音素 0-gram の認識性能を常に上回る形となり，性能向上の効果が認められた．また，サブワードに結合音素を加えることにより，始終端クラスに単語を登録したモデルの性能に近づいたが，これは，十分な学習データによるサブワードモデルを持つ階層化言語モデルは，単語登録方式と同等の性能を持つ，という知見[1][4]に反しない結果である．

8.6. 単一クラスのサブワードモデル導入による性能実験

前節で，階層化言語モデルと音素 0-gram を比較しただけでは，性能向上が上層の始終端クラス 2-gram によるものなのか，下層のサブワードモデルによるものなのかが確認できない．そのため，クラス分けを行わず単一のクラスで前節と同様の認識実験を行う．すなわち，全ての単語の音素列を対象に Word-structure モデルを適用する．

8.6.1. 単一クラスサブワードモデル導入結果

実験結果を以下に示す．



単一クラスのサブワードモデルによる音素認識率は，結合音素 0-gram に比べ，平均して 4～6%程度高かった．また，前節の階層化言語モデルに比べ，平均して 1～2%程度低かった．本モデルでは，音素とシラブルではあまり差は見られなかった（平均差±0.5%未満）．

8.6.2. 単一クラスサブワードモデル導入の考察

クラスタリングを行わず全ての単語の音素列を対象として，Word-structure モデルを適用した場合，音素 0-gram の認識性能を常に上回る形となり，前述の階層化言語モデルに近い性能の向上が認められた．このことから，単語レベルの音素列のモデル化である Word-structure モデルが，認識性能を大きく向上させる制約として働くことが予想される．また，階層化言語モデルとの比較で，上層の始終端クラス 2-gram モデルがあった方が 1～2%性能が改善されており，このような学習を伴わない大まかなクラスタリングによっても，制約として有効に機能するものと予想される．

9. まとめ

個別的タスクに依存しない音声タイプライタの認識精度が向上するような言語モデルを考察するにあたり，近接する音素を拘束する音素結合と単語レベルの音素列を律束するような制約を検討し，制約導入による音素正解率の変化を調べる認識実験を行った．

単語レベルの音素列の連接特性を直接的に表現するものとして，単語を構成する音素列の始端と終端の音素を指標とする始終端クラスを定義した．さらに，言語モデルを構築する際，上層に始終端クラス 2-gram を，下層に単語レベルの音素列を推定することができる Word-structure モデル[1]によるサブワードモデルを持つ階層化言語モデル[1]を適用した．

まず，結合音素導入による認識精度の変化を調べるために，音素 0-gram に結合音素を加えていったところ，音素正解率が最大で約 15%上昇し，局所的な音素列に対する制約が性能向上に高い効果を示すことが確認できた．

次に，始終端クラスに単語を登録して認識実験を行ったところ，音素正解率は約 87%となり，タスクに最適化してクラスタリングを行った V8 標準モデルの 92.9%には及ばないものの，始終端クラスのような大まかなクラスタリングでも，ある程度の制約力が期待できることが分かった．

また，上層に始終端クラス 2-gram モデル，下層にサブワードモデルを持つ階層化言語モデルでは，音素 0-gram の認識性能を約 6～7%程度常に上回る形となり，性能向上の効果が認められた．但し，この性能向上は，上層の始終端クラス 2-gram による効果なのか，下層のサブワードモデルによる効果なのかが確認できないため，階層化言語モデルの上層でクラスタリングを行わず全ての単語の音素列を対象として，Word-structure モデルのみを導入して認識実験を行ったところ，音素 0-gram の認識性能を平均して 4～6%程度常に上回る形となり，前述の階層化言語モデルに近い性能向上が認められた．

このことから，単語レベルの音素列のモデル化を行っている Word-structure モデルが，認識性能を大きく向上させる制約として働くことが予想された．さらに，階層化言語モデルの結果から，上層の始終端クラスのような学習を伴わない大まかなクラスタリングによるクラス 2-gram であっても，制約として有効に機能するものと予想された．

10. 今後の課題

今後，次のような認識実験を行って，性能向上の要因を検証していく必要がある．

・音声タイプライタに一般的に用いられている音素 N-gram（N>0）の性能
・始終端クラスに自動クラスタリングを適用した場合の性能
・別の指標によるクラスの検討
　－ 品詞情報によるクラスタリング
　－ 学習データの尤度最大化による自動クラスタリング 等

また，パープレキシティによってモデルの評価を行い，情報量尺度の観点から精度改善の確認を行う．

今回は，実験で使用可能な学習データや時間的な制約から，タスク依存性を確かめる実験は実施できなかったが，別コーパスから学習データを用意して同様の認識実験を行うことで，タスク依存性についての検証を行うことが可能である．

11. 参考文献

[1] 大西茂彦,小窪浩明,山本博史,匂坂芳典,“大語彙連続音声認識における未知語の sub-word モデリング手法”,信学技報,EA2001-5,SP2001-5,pp.33-39.Apr.,2001.

[2] 山本博史,匂坂芳典,“接続の方向性を考慮した多重クラス複合 N-gram モデル”,日本音響学界,平成 10 年秋季研究発表会論文集

[3] 北研二,”言語と計算 4 確率的言語モデル”, 東京大学出版会, 1999

[4] 谷垣宏一,山本博史,匂坂芳典,“クラスに依存した語彙の確率的既述に基づく階層型言語モデル”,信学技報,Vol99,No.526,pp.49-54,1999