音声言語情報処理 44－9
(2002. 12. 19)

日本語話し言葉音声認識のための音節に基づく高精度な音響モデルの検討

緒方 淳† 有木 康雄†

† 龍谷大学理工学部 〒520-2194 滋賀県大津市瀬田大江町横谷 1-5
E-mail: †ogata@arikilab.elec.ryukoku.ac.jp, ††ariki@rins.ryukoku.ac.jp

あらまし 日本語話し言葉音声認識のための音節に基づく音響モデリング手法について検討している．従来,「モーラ」単位に基づくサブワード音響モデルが検討され，読み上げ音声認識においてその効果が確認されている．それに対し本報告では,「音節」と「モーラ」をその定義から明確に区別し，話し言葉音声認識において,「音節」が「モーラ」より音響モデルの単位として適していることを示す．具体的には，話し言葉音声特有の現象で，頻繁に発生する「長母音化」に注目し，これを明確に考慮した音節モデル，並びにその高精度なモデリング手法を提案する．学会講演音声を対象とした認識実験の結果，提案モデルによって従来の triphone モデル，モーラモデルを上回る認識性能を得ることができた．

キーワード 話し言葉音声，音節，モーラ，長母音化，HMM

Syllable-Based Acoustical Modeling for Japanese Spontaneous Speech Recognition

Jun OGATA† and Yasuo ARIKI†

† Faculty of Science and Technology, Ryukoku University
1-5, Yokotani, Oecho, Seta, Otsu, Shiga, 520-2194 Japan
E-mail: †ogata@arikilab.elec.ryukoku.ac.jp, ††ariki@rins.ryukoku.ac.jp

Abstract We study on a syllable-based acoustical modeling method for Japanese spontaneous speech recognition. Traditionally, mora-based acoustic models have been adopted for Japanese read speech recognition system. In this paper, syllable-based unit and mora-based unit are clealy distinguished in their definition, and syllables are shown to more suitable as an acoustic model in Japanese spontaneous speech recognition. In spontaneous speech, a *vowel lengthening* occurs frequently, and recognition accuracy is greatly affected by this phenomena. In this view point, we propose an acoustical modeling technique that emplicitly incorporates the *vowel lengthening* in syllable-based HMMs. Experimental results showed that the proposed model could exceed the performance of conventionally used cross-word triphone model and mora-based model in Japanese spontaneous speech recognition task.

Key words spontaneous speech, syllable, mora, vowel lengthning, HMM

1. まえがき

近年，大語彙連続音声認識技術が格段に進歩し，パソコン上で実時間動作可能なディクテーションソフトウェアやフリーの共通プラットフォームソフトウェア [1] などが登場し，読み上げ音声の認識においては一定の成功を収めているといえる．しかし，人間同士のコミュニケーションにおいて,「読み上げ音声」は極めて不自然であり，これからは，より自然ないわゆる「話し言葉」の認識，理解が必要である．

本研究では，話し言葉音声認識における高精度な音響モデルの構築を目的としている．話し言葉音声においては，とりわけ調音結合や異音化の影響が大きく，また話し言葉特有の現象なども頻繁に発生することから，音響モデルの精密化が必要不可欠である．現在の音声認識システムのほとんどが隠れマルコフモデル (HMM) に基づく音響モデルを採用しており，音響モデル自体の単位としては，音素が一般的に用いられている．音素単位の HMM を用いたシステムは，もともと海外の研究機関で提案され，現在，国内の多くの研究機関においても一般的に

用いられている．しかしながら，音響モデルの単位として必ずしも音素が最適であるとは限らず，最近では，英語音声に対して音節を基本とした音響モデリングを行い，音素を用いた場合と比べて高い認識性能が得られるといった報告がなされている [2]．ただし，文献 [2] では，音節を基本として，triphone や単音節単語を組み合わせてモデリングするハイブリッドシステムについて報告されており，英語音声においては，音節単独を用いて音響モデリングを行うことは難しい．この理由としては，英語音声においては音節の種類が膨大に存在し (10000 以上)，それらを限られた学習データで十分に学習することが困難であることが挙げられる．

一方，日本語は英語などに比べて音節数が極端に少なく，音響モデルの単位として音節を用いることの有効性がこれまでに報告されている [3]- [6]．いずれの報告においても，音節モデルは triphone モデルと比較して，ほぼ同等の認識精度が得られることが実験的に示されている．ただし，これらの報告では全て，新聞記事読み上げ音声タスクにおける音響モデリング，認識実験が行われており，日本語話し言葉・自由発話のための音節ベースの音響モデリング，評価等の報告はなされていない．また，従来研究における音節モデルは，基本的にモーラ単位に準じており，厳密な音節に基づいたモデルではないと考えられる．

本報告では，「音節」と「モーラ」をその定義から明確に区別し，話し言葉音声認識において，「音節」が「モーラ」より音響モデルの単位として適していることを示す．具体的には，話し言葉音声特有の現象で，頻繁に発生する「長母音化」に注目し，これを明確に考慮した音節モデル，並びにその高精度なモデリング手法を提案する．学会講演音声を対象とした認識実験を通して，認識精度，認識処理コストの両面から提案モデルの評価を行う．また，従来より用いられている triphone，モーラの両モデルとの比較を行う．

2. 日本語音声認識のための音響モデル単位

現在の音声認識においては，大語彙，不特定話者の条件下では，サブワード単位の音響モデルが用いられることが多い．サブワード音響モデル (ほとんどの場合は HMM) の高精度化に関しては，これまでに多くの研究報告がなされている．

高精度なサブワード音響モデルを作成する上で重要な要素の一つとしては，「コンテキスト依存性」が挙げられる．すなわち，先行または後続の環境 (コンテキスト) に依存する形で当該モデルを学習する．これにより，コンテキストに依存した調音結合や異音化をモデル化することができ，より精密な音響モデルを作成することができる．また，音響モデリングにおいてはコンテキスト依存性とは別に，サブワードの各々のユニットの「モデリング区間長」が重要となる．一つのユニットのモデリング区間長が長いほど，スペクトルの時間的な依存性をモデル化することが可能であると考えられる [2] [6]．同時に，より長いモデリング区間長のサブワードを用いることによって，コンテキスト依存性を一つのユニットの中に直接考慮することもできる．

ただし，これら 2 点を考慮した音響モデリングは，より精密で高精度なモデル化が可能である反面，モデル数，モデルパラメータが膨大に増加してしまう．したがって，限られた学習データの中で，より精密に，かつデータの不足を起こすことなく効率的に学習可能なサブワードモデルが望まれる．このような観点に基づき，以下ではこれまでに報告されている主なサブワードモデルについて考察する．

a) 音　　素

現在，音声認識のための音響モデル単位として，最も広く用いられている．特に大語彙音声認識システムにおいては，音素モデルは通常，音素環境によってコンテキスト依存させ，コンテキスト依存モデルとして用いられる．コンテキスト依存モデルとしては，前後の音素環境を考慮した triphone モデルが主流となっている．triphone モデルを用いることにより，あらゆる音素の前後パターンの渡り部分をモデル化できる．しかし，triphone を含めたコンテキスト依存モデルでは，モデル数が膨大になることやモデルの構造が複雑になるため計算量が増加する (詳細は後述)．また，モデリング区間長に関しては，現在主に検討されているサブワード音響モデルの中では最も短い．

b) 音　　節

音節モデルは，日本語音声認識においては音素モデルに次いでよく用いられている．これまでにサブワード音響モデルに音節を用いた認識システムに関する報告がなされている [3]- [6]．ただし，これら先行研究における「音節」とは，基本的に「モーラ」単位を指しており，後述するように本研究では「音節」と「モーラ」を区別して取り扱う．音節は「母音を中心とする音のまとまり」であり [8]，日本語音声の知覚の単位とも報告されている [6] [9]．日本語における音節は，大部分が子音と母音 (CV) からなり，CV のコンテキスト依存性に関しては一つの単位の中で考慮されている．また，モデリング区間長に関しては，大体が音素 2 つ分の長さでモデル化されるため，その分スペクトルの時間依存性を考慮できる．これに関連して，音素レベルで発生する発音変形 (母音の無声化など) を一つのモデルに含めて学習するので，音素レベルの発音モデリングを必要としない [2]．ただし，コンテキスト依存度は triphone より低く (triphone は前後の環境を併せて 3 音素分に対して，音節は CV の 2 音素分)，母音から子音への遷移に関してはモデル化されない．

c) 半 音 節

半音節単位は，triphone などに比べて，比較的少ない種類であらゆる音素遷移パターンを表現できる単位であり，その基本的な定義は「音節をその母音中心で分割したもの」とされる [7]．半音節は，音節を精密にしたものと考えられ，音節では捉えることのできない母音から子音への遷移も表現できる．コンテキスト依存性に関しては，音節とほぼ同等であると考えられるが，音節をより細かく分割する分，モデリング区間長は小さくなる．また，連続音声認識の場合，triphone と同様，単語間のコンテキスト依存性 (後述) も扱う必要がある．

本研究では，話し言葉音声認識のための音節に基づくサブワード音響モデリングについて検討している．以下では，上述した 3 種類のサブワード単位のうち，主に音素，音節に焦点を

当て，それらのモデリング法並びに比較について述べる．なお，半音節との比較は今後の課題とした．

3. 話し言葉音声認識のための音節に基づく音響モデリング

3.1 音節単位とモーラ単位

音節単位サブワード音響モデルに関するこれまでの先行研究においては，いずれもモデリングの単位として厳密な音節ではなく，モーラの単位に基づいていた (以下ではこれら従来手法をモーラモデルと呼ぶ) [3]- [6]．本研究では，モーラと音節をその特徴によって厳密に区別して扱い，それぞれに対する音響モデリングを検討する．まず，文献 [8] に基づき，音節，モーラの定義並びにその特徴について以下に述べる．

一般的に，音節は「母音を中心とする音のまとまり」として定義され，日本語音声の場合，大体は 1-3 の音素のまとまりとして構成される．その構成パターンとしては，V，VV，CV，CVV，CVC が主に挙げられる．一方，モーラは音節を部分的に分解した単位であり，例えば CVC であれば CV と V に分解し，長母音や二重母音を含む CVV であれば CV と V に分解する．日本語においては，モーラは，ひらがなあるいはカタカナとほぼ 1 対 1 の対応をなす．撥音「ん」や促音「っ」，長母音に対応する「ー」に対しても独立した単位(注 1)を割り当てる．ここで，表 1に音節，モーラの構成例を示す．表中，「・」は音節あるいはモーラの境界を表している．

表 1　音節，モーラの構成例

例	音節	モーラ
「セーター」	せー・たー	せ・ー・た・ー
「言語」	げん・ご	げ・ん・ご
「学会」	がっ・かい	が・っ・か・い

以上の定義，並びにそれぞれの単位の構成例より，音節単位とモーラ単位の特徴的な違いとして次の 3 点が挙げられる(注 2)．本研究ではこれらの特徴をそれぞれ，音節のモーラに対する「長母音化」，「促音化」，「撥音化」と呼ぶことにする．

- 長母音化: 「せ・ー」 ⇒ 「せー」
- 促音化: 「が・っ」 ⇒ 「がっ」
- 撥音化: 「げ・ん」 ⇒ 「げん」

3.2 提案モデル

本研究では，話し言葉を対象とした高精度な音声認識を目的としている．話し言葉音声や実際の会話音声において，読み上げ音声にはない特徴の一つとして，長母音化が頻繁に発生することが挙げられる．長母音化は，話し言葉特有の現象である言い淀みやフィラーなどに連動して起こる場合や，母音の連続により発生する場合 (例えば，「チョウ」⇒「チョー」) などがある．従来検討されてきたモーラモデルにおいては，長母音を単母音の連続としてモデル化するため，長母音化を厳密にモデル化することができない．このことは話し言葉音声を認識するうえで，認識性能を劣化させる原因となる可能性が十分にある．したがって，3.1節での各単位の定義からもわかるように，従来のモーラモデルより音節モデルの方が，話し言葉において発生する現象の厳密なモデル化が可能であると考えられる．

ここで実際に，CSJ コーパスモニター版中の男性約 200 名の講演書き起こしテキストを用いて，長母音化，促音化，撥音化が発生している音節の割合を調べた．表 2にその結果を示す．表からもわかるように，3 種類の現象のうち特に長母音化の発生率は高く，全出現音節の約 11 %を占める．これは，長母音化した音節を単独でモデル化するのに十分な学習データ量であるといえる．

表 2　長母音化，促音化，撥音化の発生率

	発生率
長母音化	11.4%
促音化	2.9%
撥音化	6.2%

以上のような考察から，本報告では，話し言葉音声認識のためのサブワード音響モデルとして，長母音化を考慮した音節モデル，並びにそのモデリング手法を提案する．提案モデルは，長母音化をその直前の音節に含めて一つのモデルとして学習するというものである．ただし，前述の促音化，撥音化に関しては，今回提案するモデルにおいては，直前の音節に依存することなく，促音，撥音それぞれ単独のモデルとして，従来のモーラモデルと同様に扱った．この理由としては，表 2の結果から，促音化，撥音化した音節の発生率は低く，モデル化するための学習データが十分に確保できないことと，話し言葉音声特有の現象ではないため，長母音化のように実際の認識率に直接影響するものではないと考えられるからである．

本報告では，提案する音節モデル，従来のモーラモデル並びに音素モデル (triphone)，それぞれについての比較を行う．ここで，本研究で用いた音素，モーラ各モデルの種類数を表 3に示す．モーラモデルには，外来音節も含めた 124 個のモデルを用いている．音素モデルは，日本語音声認識において標準的に使用されている音韻体系とほぼ同じものを用いている(注 3) [1]．また，提案する音節モデル，モーラモデル，音素モデルそれぞれの音韻の表記例を表 4に示す．すなわち，音素モデルでは長母音を単独でモデル化し，従来のモーラモデルでは長母音を母音の連続としてモデル化しているのに対し，提案するモデルでは長母音を直前の音節ごとモデル化している．提案モデルは，母音，CV，促音，撥音，無音のモデル (124 個) に，CV，母音それぞれに対して長母音化したモデル (120 個) を加えた合計 244 個のモデルで構成される．

3.3 音節モデルの状態共有化

提案する音節モデルでは，通常のモーラモデルに比べると，長母音化が考慮されてモデル化される反面，モデル数の増加に伴う学習データの不足が発生することで，性能劣化が生じる可能性がある．その対処として，本研究では各 HMM の状態共有

(注 1)：実際には長母音は母音の連続として表現する．

(注 2)：ただし，音節「CVV」における長母音以外の二重母音 (例:かい ⇒ か・い) の相違点に関しては，ここでは取り扱わないこととした．

(注 3)：相違点は，無音モデルの数だけである．

表3　各モデルの種類数

	音素	モーラ
母音	5	5
長母音	5	-
促音	1	1
撥音	1	1
子音	27	-
CV	-	115
無音	2	2
総モデル数	41	124

表4　それぞれのモデルの音韻表記例

	表記例「チョー」
音素モデル	ch o:
モーラモデル	cho o
提案モデル	cho:

化を導入する．提案モデルにおいては，長母音化したCVと単独のCVが別々のモデルとして学習されるが，両者は発声の前半部分(子音部分，あるいは子音直後の母音の一部)に関しては同一であるとみなすことができる．したがって，長母音化したCVと単独のCVのモデルの状態の前半部分を共有化することにより，学習データの不足を防ぐことができると考えられる．本研究では，各CVのモデル5状態のうち，前半の3状態を共有化しモデル化を行う(図1)．このように，状態を共有することによって，モデル全体の総状態数を減らすことができ，認識処理コストも減らすことができる．

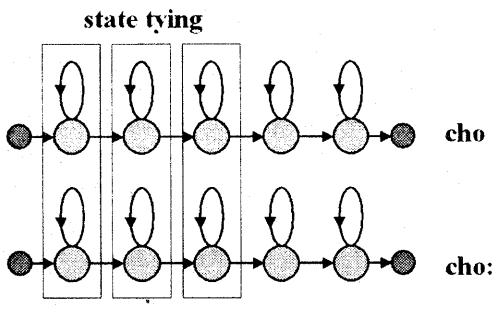

図1　長母音化音節の状態共有化

3.4　コンテキスト依存性と認識処理効率

ここでは，コンテキスト依存性によるモデルの精密化と，実際にそのモデルを認識に用いたときの処理コストについて述べる．前述したように，音素モデルにおいては，通常，前後の音素環境に依存したコンテキスト依存モデル，特に前後の1音素に依存したtriphoneモデルが用いられる．triphoneモデルは，単語間の渡り部分に関してもコンテキスト依存を行った「単語間依存型triphone (cross-word triphone)」，単語間の渡りに関してはコンテキスト依存をせず，単語内のコンテキスト依存のみを考慮する「単語内依存型triphone (word-internal triphone)」に大別される．単語間依存型triphoneモデルにおいては，デコーディングの際，単語の始端，終端部分で，コンテキストに依存してその後の候補を展開するので，全体のネットワーク展開数が劇的に増加する．その結果，認識処理時間や必要メモリー量が著しく増加してしまう[10]．単語内依存型triphoneにおいては，単語の始端，終端部分はbiphone，あるいはmonophoneでモデル化される．すなわち，単語内依存型triphoneは，単語間依存型triphoneの近似モデルであり，認識性能は劣化するが，処理効率は良い．本研究では，認識精度を優先し，音素モデルとしては，単語間依存型triphoneモデルを用いる．

一方，モーラモデルや提案する音節モデルにおいては，モデル間のコンテキスト依存は考慮しない，いわゆるmono-syllableモデルであるため，処理効率に関しては，単語間依存型triphoneよりも遥かに良いと考えられる．

4.　実験と考察

4.1　実験条件

本報告では，学習，認識を全て日本語話し言葉コーパス(C-SJ:Corpus of Spontaneous Japanese)モニター版を用いて行った[11]．CSJは，話し言葉，自由発話研究を目的とした大規模なコーパスであり，国内の学会講演音声や模擬講演音声が収録されている．各音響モデルの学習には，CSJコーパスモニター版のうち，男性話者200名の講演音声を用いた．音響特徴量は39次元のMFCC(12次元MFCCとパワー，およびそれぞれのΔ，$\Delta\Delta$)である．言語モデルには，CSJモニター版に同梱されている「講演音声認識用言語モデル (2001-06；京都大学)」のbigramモデルを用いた．612の講演(1480834単語)から学習されており，言語モデルの語彙数は19Kである．評価用データとしては，CSJモニター版中の，学習に用いたデータ以外の男性話者4名の学会講演(A01M007，A01M0035，A01M0074，A05M0031)のうち，冒頭の100発話ずつ，合計400発話を用いた．各発話へのセグメンテーションは，300msecの無音を基準に自動的に行った．

4.2　triphoneによる実験結果

まずtriphoneを用いた実験を行った．3.4節で述べたように，triphoneには単語間コンテキスト依存型のモデルを用いている．triphoneの状態の共有化は，音素決定木に基づくクラスタリング法を用いた[12]．図2に各状態数(800，1500，2500)における認識結果を示す．実験結果より，状態数が1500(正確には1517)のときに最も高い認識率が得られた．1状態あたりの混合数は40のときに最も高い認識率を示している．これより，以下では，triphoneの実験結果として状態数1500のモデルを用いることにする．なお，状態数2500のモデルでは，最適な混合数は32となっているが，これは推定すべきパラメータが比較的多いため，学習時にデータ不足が生じていると考えられる．

4.3　提案モデルによる実験結果

図3に各混合数に対する単語誤り率(WER:Word Error Rate)を，モーラモデル，triphoneモデルの結果も併せて示す．図中，「mora」はモーラモデルを，「syllable-m244」は提案する音節モデル(以下，これを単に音節モデルと呼ぶ)を，「syllable-m244-tied」は状態共有化を行った提案モデル(以下，これを状態共有

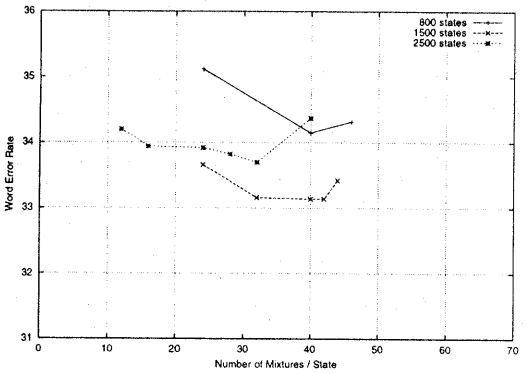

図2　triphoneによる実験結果

表5　各モデルのモデル数と総状態数

	logical モデル数	physical モデル数	総状態数
triphone	27659	2357	1517
モーラ	124	124	607
音節	244	244	1192
状態共有化音節	244	244	**847**

表6　最終的な認識結果

	最適混合数	WER(%)
triphone	40	33.16
モーラ	60	34.72
音節	56	33.35
状態共有化音節	60	31.36

化音節モデルと呼ぶ) をそれぞれ表している．また，表5に各モデルのモデル数と総状態数を示す．

実験結果より，提案する音節モデルを用いることによって，triphone モデルとほぼ同等の精度が得られていることがわかる．従来のモーラモデルの性能を大きく上回っていることから，話し言葉音声認識においては長母音化を考慮することが重要であるといえる．また，状態共有化音節モデルを用いることで，さらに精度が向上し，triphone の性能を上回っていることがわかる．状態共有化を行うことで精度が向上していることから，状態共有化を行わない音節モデルでは，CV のモデルの前半部分，特に子音部分に対する学習データの不足が生じていると考えられる．したがって，状態共有化音節モデルは長母音化が考慮され，かつ学習データの不足も少ない頑健なモデルであるといえる．

最後に，各モデルの最適混合数における単語誤り率を表6に示す．状態共有化音節モデルは，従来のモーラモデルと比較して3.3%，triphone モデルと比較して1.8%の WER を削減しており，話し言葉音声認識において有効であることがわかる．このとき，状態共有化音節モデルの総分布数は50820，triphone の総分布数は60680であり，提案モデルの方が高精度かつ僅かではあるがよりコンパクトなモデルといえる．

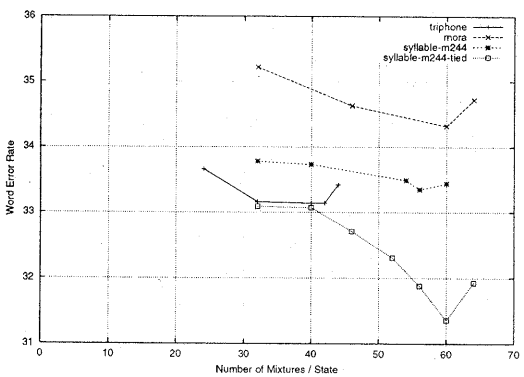

図3　提案モデルによる実験結果

4.4　各モデルの認識処理時間

ここでは，認識の際の処理時間について考察する．比較対象のモデルは，triphone と提案モデルである状態共有化音節である．実際の認識処理時間に特に強く影響するファクターは，3.4節で述べた単語間のコンテキスト依存性と，各モデルの1状態あたりの混合数であると考えられる．

単語間のコンテキスト依存性に関する処理コストは，認識に用いるデコーディング手法に強く依存する．一般的に，音声認識におけるデコーディング手法は，探索空間の構成に関して2つに大別される[10]．1つは静的ネットワーク展開型[13][14]，もう一方は動的ネットワーク展開型[15][16][17]である．単語間コンテキスト依存性に関しては，前者の方では認識前にあらかじめ展開するため，認識段階における取り扱いが簡便になり，単語間コンテキスト依存性による処理コストの増加は後者の方が著しいと考えられる．しかしながら，N-gram を用いたデコーディング，特に大語彙が対象となると，静的ネットワーク展開型は必要メモリー量などの問題で不向きであり[10]，動的ネットワーク展開型に基づく手法が一般的であると考えられる(注4)．

ここでは，動的ネットワーク展開型デコーダを用いて，各モデルごとの RTF(Real Time Factor) を算出した．実験結果を図4に示す．結果より，提案モデルは，triphone を用いた場合よりも大幅に処理コストを抑えることができている．triphone において処理時間に最も強く影響しているファクターは，単語間コンテキスト依存性である．提案モデルやモーラモデルは，mono-syllable であり基本的にモデル間のコンテキスト依存，単語間のコンテキスト依存を考慮しないため，triphone との処理効率の差は歴然としている．提案モデルに関しては，mono-syllable でありながら，単語間のコンテキスト依存性を考慮した triphone を上回る認識性能を示していることから，話し言葉音声認識において非常に有効なモデルであるといえる．

5. むすび

本報告では，話し言葉のための音節に基づく音響モデル，並

(注4)：ただし，近年，有限状態トランスデューサの適用により，静的ネットワーク展開型を用いた効率的なデコーディング手法が検討されている[14]．

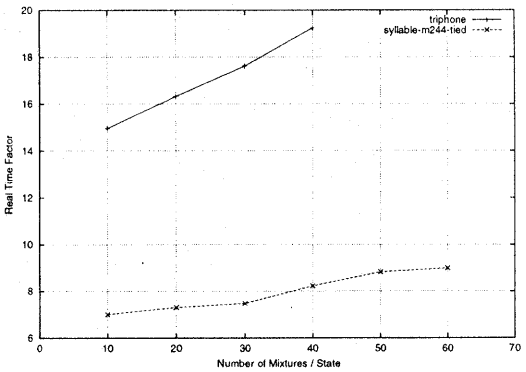

図 4　認識処理時間の比較

びにモデリング手法について検討した．提案手法は，話し言葉音声特有の現象である長母音化を直接考慮した，音節単位のサブワードモデルである．学会講演タスクを対象とした実験を行った結果，提案した音節モデルを用いることで，単語間のコンテキストまでも依存した triphone とほぼ同等の認識性能を得ることができた．また，提案モデルに状態共有化を導入することにより，triphone を上回る認識性能が得られた．提案モデルは，実際の認識時の処理コスト面においても，triphone を大幅に上回っており．話し言葉音声認識において、従来のモデルより高精度かつコンパクトなモデルであるといえる。

今後の課題としては，コンテキスト依存音節モデル，半音節モデルとの比較等が挙げられる．

文　献

[1] 河原達也 他：“日本語ディクテーション基本ソフトウェア (99 年度版) の性能評価”, 情処研報, **SLP2000**-38-2, (2000).

[2] A.Ganapathiraju, J.Hamaker, J.Picone, M.Ordowski and G.Doddington : “Syllable-Based Large Vocabulary Continuous Speech Recognition”, IEEE Trans. on *Speech and Audio Processing*,Vol.9, no.4, pp.358-366(2001).

[3] 中川聖一，花井建豪，山本一公，峯松信明：“HMM に基づく音声認識のための音節モデルと triphone モデルの比較”, 信学論,Vol.J83-D II, No.6, pp.1412-1421 (2000)

[4] 高橋信寿，中川聖一：“コンテキスト依存音節 HMM の評価”, 音講論集, pp.97-98 (2001).

[5] 諸戸正憲，山本一公，中川聖一：“大語彙連続音声認識における音節モデルの改良”, 音講論集, pp.95-96 (2001).

[6] 山本一公，中川聖一：“音声知覚実験による音声認識モデル単位の検討”, 信学技法，**SP99**-43, pp.23-30 (1999).

[7] 渡辺隆夫，磯谷亮輔，塚田聡：“半音節を単位とする HMM を用いた不特定話者音声認識”, 信学論,Vol.J75-D-II, No.8, pp.1281-1289 (1992).

[8] 窪薗晴夫，本間猛：“英語学モノグラフシリーズ 15　音節とモーラ”, 研究社 (2002).

[9] 大竹孝司：“日本語音声のセグメンテーションユニット”, 信学技法，**SP90**-108, pp.41-46 (1991).

[10] X.Aubert : “A Brief Overview of Decoding Techniques for Large Vocabulary Continioous Speech Recognition”, ISCA ITRW ASR 2000, pp91-96 (2000).

[11] 古井貞熙，前川喜久雄，井佐原均：“『話し言葉工学』プロジェクトのこれまでの成果と展望”, 第 2 回話し言葉の科学と工学ワークショップ, pp.1-6 (2002-2).

[12] J.J.Odell : “The Use of Context in Large Vocabulary Speech Recognition”, PhD thesis, Cambridge University (1995).

[13] .G.Antoniol et al : “Language Model Representations for Beam-Search Decoding”, Proc. ICASSP'95, pp.588-591, (1995).

[14] M.Mohri et al : “Network Optimizations for Large Vocabulary Speech Recognition”, *Speech Communication*, Vol.28, Nr.1, pp.1-12, (1999).

[15] S.Ortmanns, H.Ney and X.Aubert: “A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition”, *Computer Speech and Language*, Vol.11, No.1, pp.43-72, 1997.

[16] 緒方 淳，有木 康雄：“大語彙連続音声認識における最ゆう単語 back-off 接続を用いた効率的な N-best 探索法”, 信学論, Vol.84-D-II, No.12, pp.2489-2500, (2001).

[17] 李晃伸，河原達也，堂下修司：“単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識”, 信学論, J82-D-II, No.1, pp.1-9, (1999).