社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－SLP－54（33）
2004／12／21

音節継続長比モデルを用いた音声認識の検討

蟻生 政秀† 益子 貴史† 田中 信一† 河村 聡典†

†東芝 研究開発センター 〒212-8582 川崎市幸区小向東芝町 1

E-mail: †{masahide.ariu,takashi.masuko,shinichi.tanaka,akinori.kawamura}@toshiba.co.jp

あらまし 本稿では話速に対してロバストな継続時間情報のモデル化手法について述べる．本手法では，日本語のモーラ等時性から，各音節の相対的な継続時間は話速によらず一定であると仮定し，隣接する音節の継続長の比をモデル化する．これを音節継続長比モデルと呼び，隠れマルコフモデル（HMM）と組み合わせて用いることにより，HMM のみでは不十分であった継続時間情報を考慮した音声認識を行う．本手法の効果を確認するために，単語認識実験及び連続数字認識実験を行った．その結果，認識誤りは単語認識実験では最大で 18%，連続数字認識実験では最大で 28%減少した．また，音節単位の継続長モデルを用いた場合に比べ，学習データと話速が異なるデータに対しても効果のあることを確認した．

キーワード 音節， 継続長， 話速， N-Best

A Study on Speech Recognition Using Syllable Duration Ratio Model

Masahide ARIU† Takashi MASUKO† Shinichi TANAKA† and Akinori KAWAMURA†

† Corporate Research & Development Center, Toshiba Corporation

1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, 212-8582, Japan

E-mail: †{masahide.ariu,takashi.masuko,shinichi.tanaka,akinori.kawamura}@toshiba.co.jp

Abstract This paper describes a new duration model for speech recognition. Syllable duration ratio model is defined as a probability function for a duration ratio of neighboring syllables. The syllable duration ratio model can add duration information to speech recognition system, and is less subject to speech rate. The modeling technique is evaluated in two tasks of speech recognition. In the experiments, the proposed model improves the speech recognition, although the speech rates of test data are different from that of training data.

Keyword Syllable, Duration, Speech rate, N-Best

1. はじめに

従来の標準的な音声認識手法は，特徴系列を出力する状態の遷移をモデル化した隠れマルコフモデル（HMM）によって，多少の時間的な揺らぎに対してはロバストに音声現象のモデル化を行えることを利用している．しかし HMM では時間構造の表現力について問題となる場合があることが指摘されている．

その中でも，状態や音素などの継続長に関する問題は古くから議論されている．例えば標準的な HMM では，状態遷移確率によって陰に継続長をモデル化することができる．しかし状態遷移確率による継続長のモデル化では，実際の現象を表現するのに適切ではないとして，明示的な継続長分布を導入することによって，より高精度なモデル化を行う手法が提案されている[1][2]．さらには，単独の状態だけではなく前後関係を考慮して継続長分布を導入した手法も考えられている[3]．この文献[3]では，条件付分布によって状態間の継続長の関係のモデル化を行っている．

前後関係を考慮して継続時間情報を扱う際に，状態よりも長い単位である音節を対象として，既に得られた音節の継続長から次の音節の継続長を予測する手法も提案されている[4]．この手法においては，複数の要因ごとに音節の平均時間を持ち，それらを利用して既知の音節の継続長から次の音節の継続長を予測してマッチング区間の制御を行う．要因としては，当該音節の種類や先行音節の種類などが検討されている．当該音節だけではなく先行音節の情報も利用可能であることや，平均の継続長との関係を利用することにより，話速の影響を受けないようにして継続長を予測することができる手法として提案されている．また，確率的な制御を導入して，先行音節から後続音節の継続長を予測する手法も提案されている[5]．この手法では，先行音節の継続長に係数を掛けることで，後続音節の継続長の予測を行う．学習データから，その係数と，予測した時の予測残差の分布を求め，認識時には予測残差分布も利用している．この手法では，先行音節の継

続長に係数を掛ける形であるので，話速の変化に対応でき，また予測残差分布によって確率的に継続長の評価を行うことができる．

本稿で提案する手法も，これらの手法と同様に，前後の関係を考慮して相対的な継続時間のモデル化を行うことで，従来よりも高精度に，かつ話速にロバストな音声認識を行うことを目的とする．前述の文献[5]においては，先行音節の継続長に係数を掛ける形にすることで話速の影響を受けないようにして音節間の関係を表現している．しかし，予測残差自体のとる値は学習データの話速の影響を受けるため，確率的な制御を行うための予測残差分布も話速の影響を受けることが考えられる．そこで本稿ではより直接的に，隣接する音節の継続長の比をモデル化することを考える．これは，日本語のモーラ等時性から，隣り合う音節の継続長の比は話速によらず一定の分布に従っているという仮定に基づくモデル化である．また音節間ごとに分布を考慮することによって，音節同士の関係をより適切に表現するためのモデル化を行うことにする．

本稿で提案している，隣接する音節の継続長の比のモデルを音節継続長比モデルと呼ぶことにする．次節ではその基本的な考え方と，認識時の用い方について述べる．3節では単語認識実験及び連続数字認識実験を行うことでその特性を調べ，4節で本稿のまとめを述べることにする．

2. 音節継続長比モデル

2.1. 基本的な考え方

ここでは，本稿で提案する音節継続長比モデルの基本的な考え方について説明する．まず次のような仮定を考える．

- 日本語のモーラ等時性より，隣接する音節の継続長の比は一定の分布に従い，話速によって変わらない
- 継続長の比の分布は隣接する音節のコンテキストのみに依存する

このような仮定のもとに，隣接する音節の継続長の比のモデルを音節継続長比モデルと呼ぶことにする．本稿では比の取り方は後続音節の長さを先行音節の長さで割ったものとする．

また，本稿では音節として CV 音節のみ扱うことにする．そして，促音については後続音とまとめて一つの音節とみなし，長音についてはそのまま一音節とする．このような単位にすることにより，長音の有無や促音の有無といった，通常の HMM を用いた音声認識では弁別が難しい状況においても，音節継続長比によって差が出ることを期待できる．以上のことを説明するための概念図が図 1である．

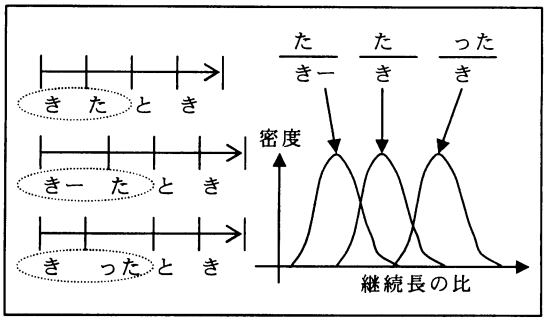

図 1　音節継続長比モデルの概念図

図 1では，「きたとき」「きーたとき」「きったとき」を例に挙げている．本稿では隣接する音節の継続長の比をモデル化するため，二音節毎に注目することになる．図 1中の語の最初の二音節に注目した場合，音節「き」と「た」の継続長の比（図中や，以降では“き＼た”で表現する）と，“きー＼た”，“き＼った”は，それぞれある分布に従うとするのが音節継続長比モデルである．そして認識対象内の全ての隣接音節間で音節継続長比モデルは評価される．つまり図 1の「きたとき」ならば，“き＼た”・“た＼と”・“と＼き”が評価されることになる．

2.2. 音節継続長比モデルの学習

音節継続長比モデルの学習は，EM アルゴリズムによって，音響モデルの HMM と同時に行うことが可能である．しかし本稿では簡単のために，既存の音響モデルによる発声内容既知のデータのアライメント結果から，音節継続長比のモデルを求めた．音節継続長比のモデルについては，継続長の比の対数値が単一の正規分布に従うものとし，前述のアライメント結果から，各音節の組み合わせ毎にそれぞれの分布のパラメータを求めた．

また，認識に必要な音節の組み合わせの中には，充分なデータが存在しない組み合わせが存在する．そこで，MDL 基準を用いた二分木のコンテキストクラスタリング[6]を用い，学習データの少ない音節の組み合わせについてはクラスタリング結果のリーフに対応する分布で代用することとした．コンテキストクラスタリングの質問は，先行・後続音節のそれぞれについて「破裂音か」，「長音か」などの，調音的な属性に対する質問群によって構成した．

2.3. N-Best 候補のリスコアリングによる認識

音節継続長比モデルを用いた認識の方法について述べる．簡便な認識方法によって音節継続長比モデルの効果を確認するために，本稿では N-Best 候補を用いたリスコアリングによる認識を行った．N-Best の各候補のアライメント結果から，音節継続長比モデルに対

する対数尤度S_Rを求めて，音響モデルの対数尤度S_Aと重み付け和したものを新しいスコアS_{NEW}とする．

$$S_{NEW}=w\times S_R+(1-w)\times S_A \quad (0\leqq w\leqq 1) \quad (1)$$

本稿の後述の認識実験では，これらS_RとS_Aの対数尤度はそれぞれ音節数，フレーム数による正規化を行った．この重み付け和でリスコアリングされた候補の内で，スコアの最大値を取るものを認識結果とする．

3. 認識実験

音節継続長比モデルの効果を確認するために，単語認識実験及び連続数字認識実験を行った．また比較対象として，隣接する音節の継続長の比を用いずに，音節の継続長の対数値をそのままモデル化したモデルを作成した．これを音節継続長モデルと呼ぶことにする．そして音節の扱い方は音節継続長比モデルと同一とした．

3.1. 実験条件

音響分析条件については，サンプリング周波数11.025KHz，フレーム長256点（約23ms），フレーム周期88点（約8ms）とした．特徴量はMFCCの12次元とパワー，それぞれのΔとΔΔを追加した39次元の特徴量とする．

音響モデルは約75時間分の短文読み上げデータから，Left-to-Right型で1音素に3状態を持つTriphoneモデルを学習した．状態数は約千であり，各状態は対角共分散行列で混合数6の混合正規分布からなる．音響モデルを構成する音素には長音やショートポーズは含めなかった（長音は母音の連結で表現する）．また，実使用環境で想定される雑音に対するロバスト性を確保するために，音響モデルの学習時において，0dBから18dBまでのSN比で定常な自動車走行雑音の重畳を，一部の学習データに対して行った．

音節継続長，音節継続長比モデルについては，雑音を重畳していない音声データ(約75時間分)を用いて，上記の音響モデルで発声内容をもとにアライメントした結果から学習した．また，前述の二分木のコンテキストクラスタリングを行い，学習データの少ないパラメータに対する対応も行った．

認識については，まずN-Best（候補数は10とした）を求めてから，2.3節で述べたリスコアリングを行った．フレーム数で正規化した音響モデルのスコアについては，音節継続長（比）モデルのスコアと近いレンジにするために一定の規則で線形変換を行っている．また，評価データを4群にわけ，1群を重み付け和の重みの調整用として残りの3群を評価するということを，対象を変えて4回行った．以降の実験結果ではその4回分の平均の値を挙げることにする．また，音声区間は既知としていずれの認識実験も行った．

3.2. 単語認識実験

まず比較的容易なタスクで音節継続長比モデルの効果を調べることにする．評価には，電子協日本語音声DB（100都市名発声）の男女計40名分を用いた．また定常な自動車走行雑音（学習データに重畳したものとは異なるもの）を重畳した評価によって雑音に対するロバスト性も調べた．各環境下での正解率を表1に示す．

ここで音響モデルのみを使用した場合の認識率C_Aを基準とし，継続長に関するモデルを併用した場合の認識率C_Rとしたときの，誤り改善率R_Cを次のように定義する．

$$R_C=100\times\frac{C_R-C_A}{100-C_A} \quad (\%) \quad (2)$$

この誤り改善率R_Cは，継続長に関するモデルを併用することにより，従来の音響モデルのみを用いた場合の認識誤りがどの程度減少したのかについて表す．表1の結果に対して，この誤り改善率を図にしたものが図2である．

表1と図2から，本稿で提案している音節継続長比モデルを用いることで，高SN比環境下では約18%の誤り改善率が得られることがわかる．一方で，継続長モデルでは特に効果は見られなかった．

表1　単語認識結果の正解率

SN比	音響モデルのみ	音響モデル＋継続長	音響モデル＋継続長比
0dB	74.8%	75.3%	76.6%
5dB	87.7%	87.8%	89.2%
15dB	94.7%	94.7%	95.6%
25dB	96.5%	96.5%	97.2%
clean	96.9%	96.9%	97.4%

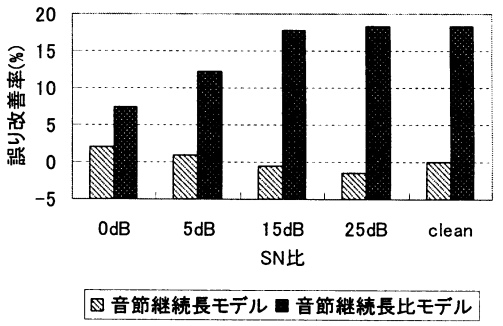

図2　単語認識における誤り改善率

音節継続長比モデルに注目すると，高 SN 比環境下の方が効果的であった．これは，N-Best 認識後のアライメント結果を用いているため，低 SN 比環境下では雑音によって，信用できないアライメント結果となってしまったことによるものと考えられる．

また高 SN 比環境下での認識結果の詳細に注目すると，「ほんど」を「ほんじょう」に，「つる」を「つるが」に誤認識する回数が大きく減少しており，これらの単語対では音節継続長比モデルが効果的に働いたものと考えられる．一方で，「つる」を「すず」に，「ちりゅう」を「きりゅう」に誤認識する場合はほとんど改善が見られなかった．音節継続長比モデルではその定義から，「ちりゅう」と「きりゅう」のような同様の音節パターンの識別は困難であることは止むを得ないと思われる．このように効果が語彙に依存する可能性はあるが，そのような場合においても，継続長やリズム的に不自然な認識誤りを減らす効果はあるものと考えられる．

3.3. 連続数字認識実験

次に連続数字認識実験の結果を挙げる．評価には，男女計 20 名による 3,4 桁の連続数字発声を用いた．各話者は 50 発声をしている．雑音の重畳については単語認識実験と同様に行った．数字の挿入誤りを減少させるために，数字 1 桁ごとに挿入ペナルティを課して認識を行った．挿入ペナルティの値は，予備実験から全ての SN 比でバランスよく認識できる値を求めて，その値による固定値とした．また，各数字の読みは「ぜろ，いち，にー，さん，よん，ごー，ろく，なな，はち，きゅー」とした．

ここでの音節継続長比モデルの定義については，N-Best 結果において無音が挿入された区間は無視して考慮しないとした．つまり，「２６」ならば“にー＼ろ”，“ろ＼く”の音節継続長比モデルを考慮するが，「２（無音）６」ならば，無音区間を除いた“にー＼ろ”，“ろ＼く”がそれぞれ評価されることになる．

単語認識と同様に，雑音を重畳させたときの文認識率（全桁一致したときを正解とする）を表 2に示す．また，音響モデルのみを使用した場合の文認識率を基準とした，各 SN 比での誤り改善率を図 3に示す．

表 2　連続数字認識結果の文認識率

SN 比	音響モデルのみ	音響モデル＋継続長	音響モデル＋継続長比
0dB	44.5%	53.4%	50.5%
5dB	78.2%	82.7%	81.0%
15dB	94.3%	95.1%	95.9%
25dB	95.7%	96.2%	96.6%
clean	95.3%	95.5%	96.2%

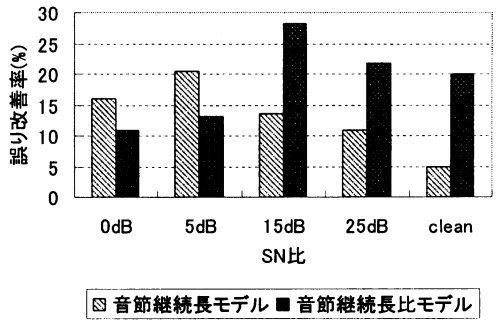

図 3　連続数字認識における誤り改善率

単語認識実験の場合とは異なり，低 SN 比環境下では音節継続長モデルの方がより効果的であった．しかし SN 比が 15dB 以上では音節継続長比モデルの方が高い認識率を示した．

音節継続長比モデルの誤り改善率が低 SN 比環境下で低下するのは，単語認識と同様にアライメントの精度が低下するためだと考えられる．一方で音節継続長モデルについては，モデルが音節の種類数分しかなくて粒度が粗いために，低 SN 比の下ではアライメント誤りの影響を受けにくいが，高 SN 比環境下での効果は限定されたことが考えられる．また，アライメントが正確になるほど学習データと評価データとの間の継続長の分布の差が現れやすくなったことも考えられる．後者については後述する．

以上から，単語認識実験及び連続数字認識実験において，実環境のように雑音が多少重畳されていても，音節継続長比モデルを用いた認識が，破綻することなく効果のあることが確認できた．N-Best 認識からの後処理でなく，デコード中に音節継続長比モデルの情報を用いることができれば，低 SN 比環境下でも継続長として不自然な候補を抑制することにより，認識精度向上に寄与できるものと考えられる．

また，比較対象の音節継続長モデルについて，コンテキスト依存（先行音節に対して，後続音節の継続長をモデル化）の実験も行ったが，特に効果は見られなかった．

3.4. 話速との関係

先の実験では，音節継続長比モデルが単語認識実験及び連続数字認識実験のどちらのタスクにおいても効果があり，かつその効果は高 SN 比では顕著であったことが示された．しかし音節継続長モデルではその性質がタスクによって大きく異なっていた．ここではその原因について調べることにする．

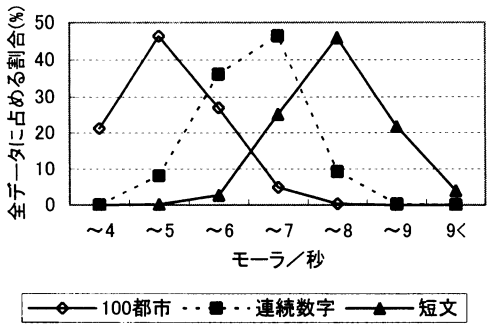

図 4　各データの話速の割合

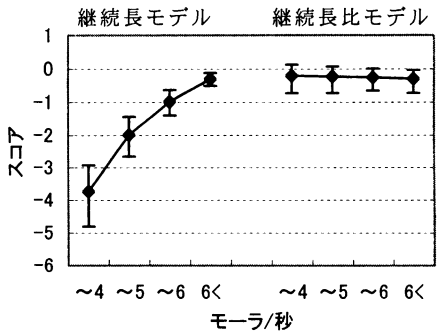

図 5　話速とスコアとの関係（100 都市）

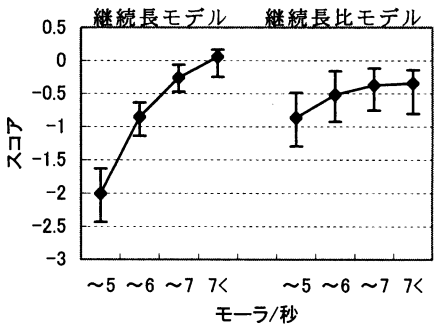

図 6　話速とスコアとの関係（連続数字）

前述のように，継続長をモデル化する場合の問題の一つとして話速の問題が挙げられる．そこで今回用いたデータについて，話速に対応する値として単位時間あたりのモーラ数を求めた．この値が大きい値になるほど話速が速かったことになる．各データをモーラ／秒でクラス分けしたときのデータの割合を図 4に示す．

図 4の横軸については，例えば“～5”の表記は「モーラ／秒が前のクラスである 4 より大きく，5 以下」であるクラスに対応する．そして”9<”は「モーラ／秒が 9 より大きい」クラスを表す．この図からわかるように，音節継続長（比）モデルを学習した短文データでは，平均的なモーラ／秒が 7～8 であったのに対し，連続数字では 6～7，100 都市の単語発声データでは 4～5 となり，特に単語発声データは学習データと大きく話速が異なっていたことがわかる．

さらに，雑音を重畳していない評価用のデータに注目して，発声内容をアライメントして得られた音節継続長（比）モデルのスコア（対数尤度）とモーラ／秒の関係を調べた．話速に対して音節継続長（比）モデルのスコアの性質が変われば，学習データと評価データとの話速の違いによる影響を受けたものと考えられる．

まず 100 都市の単語発声データについて，モーラ／秒でデータを分けたときの，各クラスの音節継続長（比）モデルのスコアの中央値を示したのが次の図 5 である．図の左側が音節継続長モデルで，右側が音節継続長比モデルに対応する．誤差表記は 4 分位点を表している．そして，連続数字発声データについても同様の結果を次の図 6に示す．

音節継続長モデルと音節継続長比モデルのスコア自体は比較できないが，それぞれのモデルのスコアについて見ると，図 5の単語発声データでは音節継続長モデルのスコアの分布が話速によって大きく変化しているのに対して，音節継続長比モデルのスコアはほとんど変化していない．同様に図 6の連続数字発声についても，音節継続長モデルでは話速によってスコアの分布が変化している．その中では，モーラ／秒が 6 を超えるクラス間の分布は近くなっている．これらのクラスでは学習に用いた短文発声データと話速が近く，評価音声とモデルとが合致しているためと考えられる．一方で音節継続長比モデルでは，モーラ／秒が 5 以下のクラスのみがややずれた分布となっているが，全体的にどの分布も重なっており，話速に対してロバストな様子が確かめられた．このように話速に変化しにくい特性から，学習データと評価データの話速が大きく異なった単語認識実験のような状況下であっても，音節継続長比モデルは効果があったものと考えられる．

以上から，単語認識実験及び連続数字認識実験で音節継続長モデルの特性が大きく異なるのは，学習デー

タと評価データの話速が大きく異なっていることが主要因と考えられる．そして，そのような状況においても音節継続長比モデルは特性が大きく変わることなく，認識精度を高められたものと考えられる．

4. まとめ

本稿では標準的な HMM では表現力の乏しい継続時間情報のモデル化について，話速にロバストな形でモデル化を行う手法について述べた．本稿で提案している音節継続長比モデルは，隣接する音節の継続長の比をモデル化することによって，学習データと評価データの話速が異なるような状況でも，相対的な継続時間を認識に使うことができることを特徴としている．

音節継続長比モデルを音声認識に用いる方法としては，本稿では従来通りの N-Best 認識を行ってから，N-Best 候補の音節継続長比モデルと音響モデルの対数尤度の重み付け和を行うというリスコアリング手法を用いた．音節継続長比モデルの効果を確認するために，短文発声データから学習された音節継続長比モデルを用いて，単語認識実験及び連続数字認識実験を行った．その結果，音節継続長比モデルによって，実験環境下において単語認識実験では最大で 18%，連続数字認識実験では最大で 28%の誤りを削減することができた．これらの認識実験は雑音重畳下についても行い，音節継続長比モデルの効果は低 SN 比環境下でも破綻することなく，15dB 以上ならば顕著に見られることを確認できた．また，単語認識実験及び連続数字認識実験に用いた評価データは，学習データと話速の分布が異なっていたことから，音節継続長比モデルは話速に対してロバストに相対的な継続時間を表現できることが確かめられた．

文　献

[1] L. Rabiner and BH Juang, 音声認識の基礎(下), 古井貞熙(監訳), pp.147-152, NTT アドバンステクノロジ, 1995.

[2] J. Pylkkönen and M. Kurimo, "Duration Modeling Techniques for Continuous Speech Recognition," Proc. ICSLP-2004, pp.385-388, 2004.

[3] Z. Qingwei, W. Zuoying and L. Dajin, "A Study of Duration in Continuous Speech Recognition Based on DDBHMM," Proc. EUROSPEECH-99, pp.1511-1514, 1999.

[4] 滝沢由実, 坪香英一, "音節継続時間予測法を用いた不特定話者連続音声認識," 電子情報通信学会論文誌, Vol. J77-A, No. 2, pp.173-181,1994.

[5] 實廣貴敏, 嵯峨山茂樹, "音節継続時間制御の語彙制約なし認識形での検討," 日本音響学会講演論文集, 2-2-8, pp.53-54, September, 1995.

[6] 篠田浩一, 渡辺隆夫, "情報量基準を用いた状態クラスタリングによる音響モデルの作成," 電子情報通信学会技術報告, NLC96-48, SP96-79, pp.9-15, Dec. 1996.