# 単語を認識単位とした日本語の大語彙連続音声認識

西村雅史† 伊東伸泰† 山崎一孝†

我々は先に，日本人が感覚的にとらえている単語単位を，既存の形態素解析プログラムの出力である形態素単位との統計的対応関係から自動推定する方法を提案し，それを認識および発声の単位とする離散単語発声の日本語ディクテーションシステムを構築した．この人間の考える単語単位を連続音声認識の認識単位としても利用することを試み，特に，他の大語彙連続音声認識システムで用いられることの多い形態素単位と比較してその有効性について調査した．また，認識単位の定義が一意に決まらない現状をふまえて，日本語の連続音声認識システムの評価方法を提案するとともに，不特定話者の大語彙音声認識実験結果について報告する．男女各10名に対する認識実験の結果，文字誤り率3.0%，単語誤り率4.3%が得られた．さらに，句読点の自動挿入方法や，未知語モデルと単語N-gramモデルによる単語単位の自動分割方法などについても述べる．

# Word-based Approach to Large-vocabulary Continuous Speech Recognition for Japanese

MASAFUMI NISHIMURA,† NOBUYASU ITOH† and KAZUTAKA YAMASAKI†

In this paper, we discuss a word-based continuous dictation system for Japanese. We previously proposed a statistical method for segmenting a text into words on the basis of human intuition, and developed an isolated-word-based Japanese dictation system. By comparing the word units used for the isolated word recognition with grammatical units, we show that the former are also very useful for continuous speech recognition. Evaluation of the performance of this continuous dictation system showed that the character error rate was 3.0%, and that the word error rate was 4.3%. We also present a method for inserting punctuation marks in spoken texts automatically, and a method for segmenting Japanese text into words by using an N-gram model, focusing on the handling of unknown words.

## 1. はじめに

近年欧米では，統計的言語モデルを用いたディクテーションシステムが実用化され，現在では医療所見の入力といった特定分野から，徐々にではあるが個人ユーザが日常的な文章を入力する手段へと市場が広まってきている．そして研究の対象も，次第に読み上げ文からニュース音声の書き起こしなど，さらに自然な発話へと移行しつつある[1),2)]．

一方，日本語については，単音節発声用の日本語音声ワープロが検討されて以来，長い間，音声による日本語文の入力システムは実用化されなかった．その直接の原因としては，欧米に比べて音声および言語データベースの整備が遅れたことが大きいが，技術的には，日本語の単位に関する定義があいまいで，言語的な単位への自動分割が困難であったこと，基本的にN-gramなどの単純な統計モデルでは日本語を正確に表現することは難しいと信じられていたこと，また，欧米語で行われていたような離散単語発声が困難または，不適切であると考えられていたこと，などが理由としてあげられる．

その後，日本語でも新聞記事などを中心にコーパスが整備されたが，それにともない，認識時の処理が容易な離散単語発声にこだわることなく，形態素を認識単位としていくつかの大語彙連続音声認識システムが試作されるようになった[3),4)]．しかし，連続発声の認識を必要とするため計算量が多く，パソコンで実時間動作するようなレベルにはない．また，連続発声ゆえに発声変形も大きく，認識精度のうえからも実用化までには課題が残っている．

一方，我々は形態素解析プログラムの出力である形態素と，日本人が感覚的にとらえている“単語単位”との対応関係を統計的なモデルで表現し，それによっ

† 日本アイ・ビー・エム株式会社東京基礎研究所
IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd.

て日本語の単語単位を自動推定する方法を提案した．そして，この単位を使えば日本語でも離散単語発声によるディクテーションシステムが，欧米語と同様にN-gram言語モデルと音素HMMによって構成される認識システムとして実現できることを示した[5]．また，この手法を用い，Intel社製のCPU，Pentium-133 MHz程度の処理能力で実時間動作が可能なパソコン用の日本語ディクテーションソフトウェアを実用化した[6]．

本研究では，このシステムの自然な拡張として，離散発声用の単語単位を連続音声認識の認識単位としても使用することを試みる．この単位は人間が感覚的にとらえている単語単位に近いことから，認識の結果得られる単語列の分割の仕方に違和感が生じにくく，単語ごとの修正がしやすいことや，離散発声可能な発声の最小単位でもあることから，文中に自由にポーズを置くことができるなどの利点がある．一方，既存の形態素単位などに比べて分割の揺らぎを単位に含むために，語彙サイズやパープレキシティの増大が懸念されるため，ここでは，実験によってこれらの認識単位の基本的な性質を比較するとともに，単語単位が音響的にはむしろ識別しやすい単位となっていることを示す．

次に，認識単位が異なる認識システムの性能評価方法を提案した後，不特定話者大語彙連続音声認識実験の結果について報告する．連続発声中の発声変形に対処しうる十分な量の学習音声データや精密な音響モデルを用意することで，高い認識精度が得られることを示す．また，句読点の自動挿入方法や，未知語モデルと単語N-gramモデルによってテキストを単語単位へ自動分割する方法など，ディクテーションシステムの実用化を進めるうえで重要となる技術についても触れる．

## 2. 認識単位の定義

日本語のように単語の概念が明確でない言語においてディクテーションシステムを実現しようとする場合，どのような単位を認識単位とするかがシステムの性能を決めるうえで重要な問題となる．連続発声に対処できるアルゴリズムを適用する限り，原理的には1音素以上のどのような単位を使用することも可能ではあるが，音響モデルとしても，言語モデルとしても，ある程度まとまった長さの単位が効率良く定義されていることが望ましい．短すぎる認識単位は音響的にも言語的にも識別が難しいし，逆に個々の単位を長くすると必要となる単位の種類数が増え，これも認識精度や処理速度を下げる要因となるからである．

日本語処理研究の成果として，形態素解析プログラムが開発され，機械翻訳やテキスト音声合成などで広く利用されている．大語彙連続音声認識においても，この中で定義された解析的な最小単位☆を認識の単位としてそのまま利用することが多い．しかし，この単位はあくまでも解析的な観点から定義されたものであり，人間が感覚的にとらえている単語単位や発声の単位とは一致しない．たとえば，「行った」という単語は「行」，「っ」，「た」という3つの形態素で構成されている．また，解析を精度良く進めるために，長単位の複合語を1つの形態素として出力することも多く，離散発声には明らかに向かない．

一方，我々は，日本語においても，離散発声入力が可能との条件を満たすような単語単位を音声認識における認識単位とした[5]．そして，そのような単語単位を得るため，統計的な単語分割手法を導入した．その方法について簡単に触れておく．まず，多数の被験者に対し，その単位で発声できることを前提に，日本語の文章を「単語単位に分割する」作業を指示し，たとえば，名詞の後に「人」という名詞が続く場合，「者」が続く場合よりもそこで分割が起こりやすいといった傾向を調査した．その結果を既存の形態素解析プログラムの出力と対比させると，ある品詞の連鎖においては分割が起こり難いのに対し，別の品詞連鎖では分割が起こりやすいといった傾向が観察される．このような分割の起こりやすさ（確率）を品詞連鎖と関連付けて統計的に推定した．我々はこれを単語分割モデルと呼んでいる．

この単位は日本語の最小発声単位をほぼ包含しているから，連続発声に適用した場合には，発話者が，文中に自由にポーズを置くことができるという利点がある．当然，これまでどおり離散単語発声による入力も可能である．また，離散発声は，認識結果の修正作業の際など，単語分割位置に関する誤りを回避したい場合には特に有効である．

## 3. 連続音声認識における認識単位の比較

離散単語発声のために定義した単語単位が，連続音声認識の認識単位として，どの程度有効であるかについて，他の単位との比較を行う．具体的には，形態素解析プログラム[7]の出力の複合名詞部分を短単位に分割した，より本来の形態素に近い単位（以降，これを形態素と呼ぶことにする）とカバレージ，パープレキシティ，単位あたりのモーラ長および読みの異なり数

☆ この単位を単に形態素と呼ぶことが多いが，文法学者が呼ぶところの形態素とはかなり異なっている．

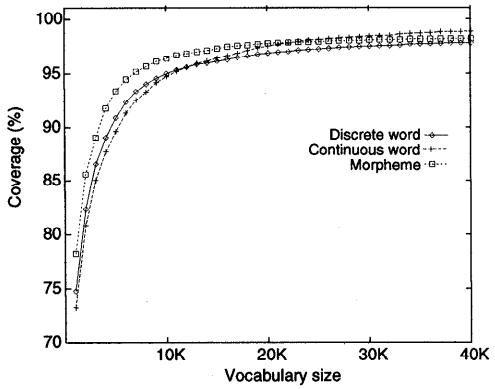



図 1 認識単位ごとの語彙サイズとカバレージの関係
Fig. 1 Relation between vocabulary size of each recognition unit and coverage.

の観点から比較する．なお，本研究では各単位の表記のみを辞書のエントリーとして使用している．

### 3.1 カバレージの比較

日経新聞 3 カ月分のデータを，形態素単位，単語単位にそれぞれ自動分割し，各単位を頻度順にならべた辞書を用意した．また，このデータとは時期が異なる 600 文のテストデータを用意し，それぞれ，(1) 人手で単語単位に分割したもの（Discrete word），(1') 単語の Tri-gram モデルを用い，得られる単語列 ($W = w_1, w_2, \ldots, w_n$) の尤度 ($Pr(W) = \sum_{i=1}^{n} Pr(w_i|w_{i-1}, w_{i-2})$) が最大になるような単語単位に自動分割したもの（Continuous word）☆，(2) 形態素解析プログラムを用いて解析後，複合名詞を自動分割し，さらに人手で修正して形態素単位としたもの（Morpheme）の 3 種類のデータを作成し，それぞれの単位に対応する辞書を用いて，語彙サイズとテスト文のカバレージの関係を調べた．結果を図 1 に示す．なお，単語単位に関し，(1) は離散単語発声の音声認識に相当する．一方，(1') は同じ単語単位を使っているが，音響モデルおよび未知語部分の影響を無視すれば，言語モデルを用いて連続音声認識を行った場合に相当すると考えられる☆☆．

我々が用いている単語単位は，単語分割の揺らぎを反映した単位となっているが，離散単語発声 (1) の場合にも，揺らぎを表現するために必要とされる単位の数はそれほど多くはなく，不利と考えられたカバレージに関しても，形態素単位に近い効率を持つことが分かる．また，連続音声認識を想定した (1') の場合には言語モデルを参照して尤度が最大となるような単語列が選ばれることから，カバレージはさらに改善されることとなり，4 万語では約 99%に達する．なお，この結果を見る限りでは 2 万語以上の語彙を用意する場合には，形態素単位よりも単語単位の方がカバレージは高いが，これには各単位の自動分割プログラムの性能差も影響している．特に形態素において複合語の分割精度が若干低かったが，このために語彙サイズの増大がカバレージの改善に直接結び付いていない面がある．

### 3.2 言語モデルの能力の比較

4.2 節に示す単語単位の Tri-gram モデルと同じ学習データから推定した形態素単位の Tri-gram モデルの能力を比較した．学習用データの分割は，いずれの場合もすべて自動処理である．一方，テスト文としては学習領域内のタスクとして日経新聞からとった 600 文を，また，学習領域外のタスクとしてはビジネストークからとった 679 文を使用した．なお，テスト文は，先の実験と同様に 3.1 節における (1)，(2) の基準で分割した．

未知語の有無がパープレキシティの値に大きな影響を与えるため，いずれの場合もパープレキシティの算出時には未知語の予測を計算から除外している．また，単語単位に付いてはすでに用意してあった約 4 万語の辞書を用いたが，形態素単位に対してはそれぞれのテストデータに対して単語単位と同等のカバレージ（ビジネストーク：97.5%，日経新聞：98.0%）が得られるように高頻度のものから順次単位を追加し，新たに辞書を作成したうえで言語モデルを推定した．

これらの単位はそれぞれ長さが異なり，文を構成する単位数が異なるため，何らかの基準で正規化を行わないと比較ができない．他の研究機関では形態素のパープレキシティを評価基準として用いることが多いので，ここでも単語単位については 1 文あたりのエントロピー ($H_{snt}$) を推定した後，次式で形態素あたりのパープレキシティ ($PP_{mor}$) に正規化して比較を行った．ここで，$N$ は文あたりの形態素の単位数である．結果を表 1 に示す．

$$H_{mor} = H_{snt}/N$$
$$PP_{mor} = 2^{H_{mor}}$$

この結果から分かるように，いずれのタスクに対しても形態素の単位あたりに換算して求めたパープレキシティの差は小さく，言語モデルの予測能力としてはどちらの認識単位を用いても本質的な違いはないといえる．

☆ この自動分割手法を適用するためには未知語部分も統計モデルとして表現する必要があるが，このモデルについては 5.3 節で述べる．

☆☆ たとえば，“例文集” という音声が入力され，“例文” および “集” が辞書にあり，“例文集” は辞書にない場合，離散単語認識ではこの単語は未知語となるが，連続音声認識の場合には未知語ではない．

表 1 認識単位による言語モデルの能力の比較
Table 1 Comparison of the LM capabilities of each recognition unit.

| タスク | 評価尺度 | 認識単位 | |
|---|---|---|---|
| | | 単語 | 形態素 |
| 日経新聞 | 文あたりの平均単位数 | 30.6 | 34.8 |
| | $PP_{mor}$ | 51.9 | 53.3 |
| ビジネストーク | 文あたりの平均単位数 | 24.2 | 30.2 |
| | $PP_{mor}$ | 44.5 | 42.7 |

表 2 認識単位あたりの平均モーラ長および読みの異なり数
Table 2 Average number of morae and pronunciations in each recognition unit.

| タスク | 評価尺度 | 認識単位 | |
|---|---|---|---|
| | | 単語 | 形態素 |
| 日経新聞 | 平均モーラ長 | 2.5 | 2.2 |
| | 読みの異なり数 | 1.13 | 1.18 |
| ビジネストーク | 平均モーラ長 | 2.3 | 1.9 |
| | 読みの異なり数 | 1.05 | 1.15 |

### 3.3 平均モーラ長および読みの異なり数の比較

現在の音声認識では，各音素の前後の音素環境別にどれだけ正確なモデルを用意できるかが認識精度に大きく影響することが知られているが，一般的には認識単位間にわたる音素環境，特に右側の環境をすべて考慮しようとすると計算量が膨大になる．その意味では単位内のモーラ長が長い方が音素環境を考慮したデコーディングを広い範囲で容易に行うことができる．

また，単位が長いほど単位あたりの読みの種類数を絞り込めると期待される．たとえば,「行く」という単語の読みは「イク，ユク」しかないが，動詞の語幹の「行」だけに分割されると「イ，ユ，コウ，ギョウ，オコ，オコナ，イキ，ユキ」など，このコンテキストでは不必要な多数の読みを考慮する必要が生じ，認識を困難にする 1 つの要因となる．

先ほどと同じ 2 種類のタスクに対し，それぞれの単位（単語，形態素）において，テスト文中の単位あたりの読みのモーラ長を比較した．また，表記から読みの辞書を参照し，単位あたりの読みの異なり数を推定した．結果を**表 2** に示す．単語単位の平均モーラ長は形態素に比べ 14～21%程度長く，読みについても 4～9%とわずかではあるが絞り込まれていることが分かった．

このように，人間の振舞いに基づいて推定した単語単位は離散発声が可能であるという特徴を持つ一方で，一般的に用いられることの多い形態素単位と比較してもカバレージや言語モデルの能力の面では同程度の性能を持つことが分かった．また，音響的には相対的に識別しやすい単位となっていた．

## 4. 連続音声認識システム

### 4.1 認識システムの構成[5),8)]

離散発声も可能な単語単位を認識単位として大語彙の連続音声認識システムを構築した．本認識システムは**図 2** に示すように，ランクラベラー，ファーストマッチとディーテイルマッチを併用したスタックデコーダーおよび Tri-gram モデルなどから構成される．

音響モデルは混合正規分布で表現された音素環境依存型の HMM である．各中心音素の音素環境は HMM 上の各状態ごとに判別木で表現され，葉の部分に混合正規分布が対応付けられている．ランクラベラーはこの音素 HMM のすべての状態，言い換えるとすべての混合正規分布に対して入力特徴量の出現確率を推定した後，それを順序付けするもので，順位をラベルとして出力する．一方，ラベル（つまり順位）の出力確率は，ビタービアライメントによって HMM の各状態に対応付けられた正解ラベル（音素環境が一致する混合正規分布の順位）の出現頻度から状態ごとに推定しておく．デコーディングの際にはこのラベル出力確率テーブルを参照することになる．なお，入力特徴量の出現確率を尤度計算に直接使用しないのはデコーディング時のダイナミックレンジを確保するためである．

デコーダーはスタックデコーディングアルゴリズムを用いて，入力単語列を推定する．まず，現時点でのスタックを参照して次に展開すべきパス（単語列）を決め，ファーストマッチを実行する．ファーストマッチは，我々のシステムの中では処理の高速化のための中心的な役割を果たす部分である．木構造で表現された発音辞書を参照して，候補単語を絞り込む処理を行うが，この際には音素環境を考慮しない．また，継続時間を表現するため最小滞留時間は考慮するものの同一音素の HMM 内では各状態のラベル出力確率も区別せず，最大値で代用する．具体的には，音素ごとの全判別木上の混合正規分布の中で最大の出現確率を与えた混合正規分布のラベル（つまり順位）に対応する出力確率をその音素に対する入力特徴量の出現確率として用いることになる．こうすることにより，正解候補を落とさずに高効率な絞り込みが可能になる．

さらにこの単語候補に対し Tri-gram モデルを適用してプルーニング処理を行い，候補を 100 語程度にまで絞る．この候補に対して音素環境を考慮したディーテイルマッチを行って終端の範囲と尤度を求め，さらに Tri-gram モデルを併用して結果をソートし，スタックに積む．この操作を繰り返し，入力単語列を推定する．

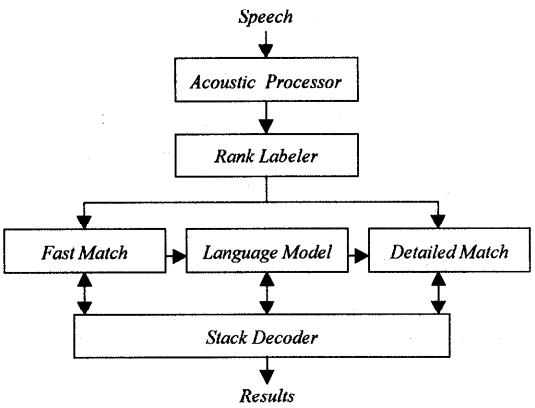



図 2 認識システムの構成
Fig. 2 Block diagram of a speech recognizer.

表 3 言語モデル用学習データの内訳
Table 3 Sample corpora and their approximate sizes.

| コーパス | サイズ（文） |
|---|---|
| 日経新聞 | 415 K |
| 産経新聞 | 1,291 K |
| EDR[9] | 169 K |
| 電子会議室の発言 | 173 K |
| 合計 | 2,048 K |

### 4.2 言語モデル

**表 3** に示す約 200 万文のテキストデータに対して，形態素解析処理を行った後，単語分割モデルを用いて単語単位に自動分割し，これを言語モデルの学習データとした．なお，学習データには句読点や括弧類など読み上げが可能と思われる記号はすべてそのまま含んでいるが，極端に記号が多い文章など，読み上げに不適切な文章はあらかじめ除いてある．

この学習データを用いて Tri-gram モデルの推定を行った．認識対象語彙は主に出現頻度に基づいて選択した 39,295 語（40 K 語彙）とし，Tri-gram の平滑化には Held-out 補間法を適用した．なお，カットオフなどによるパラメータ数の削減は行っていない．推定されたモデルのエントリー数は Tri-gram が 20,358 K，Bi-gram が 3,845 K であった．

### 4.3 音響モデル

20～60 代の男性 110 名，女性 105 名がそれぞれ 120 文（Set-25 K：計 25,800 文，355,454 単語），20～50 代の男性 312 名，女性 346 名がそれぞれ 148 文（Set-100 K：計 97,384 文，1,338,307 単語）を連続発声したデータを用いて男女共通の不特定話者用音響モデルを推定した．サンプリングレートは 11.025 KHz とし，採取は比較的静かな室内で，接話型のマイク（SHURE-SM10）を用いて行った．各発声文には新聞記事，手紙，小説など多様な文書を用意し，各人がそれぞれ異なる文章を読み上げている．

なお，各音素モデルはおおむね 3 状態からなるモデルとして表現されている．また，各状態に対し最大前後 5 音素分の音素環境を考慮した判別木が推定され，その葉の部分に混合正規分布が割り当てられている．本研究では音素環境依存モデルの総状態数を約 3,000 とした．また，特に断らない限り正規分布の平均混合数は 20，正規分布の総数は約 60,000（=60 K）である．

## 5. 評価方法

日本語を対象とする場合，文字単位のような明確な単位を用いる場合は別として，1 つの文を認識単位列に分割するやり方には曖昧さが残る．特に，ここで定義したような単語単位を用いた場合，1 つの入力文にあてはまる単語列候補は多数存在し，入力単語数ですら一意には決められない．このため，ここではシステムの性能評価（誤認識率，単語カバレージ，パープレキシティ）を次のようにして行う．

### 5.1 誤認識率

ディクテーションシステムを考えた場合には，入力後の修正作業も含めてその性能を評価できたほうがよい．その意味では認識結果の単語分割の正確さも重要な要因となりうるが，認識単位の異なるシステムを比較するには，単純に認識結果の表記上の誤り，つまり文字誤り率（$CER$）と文字正解精度（$Accuracy$）で認識性能を評価することが望ましい．

$$CER = 100 \cdot \frac{Ins. + Del. + Sub.}{N}$$
$$Accuracy = 100 - CER$$

なお，$Ins.$，$Del.$，$Sub.$，はそれぞれ文字の挿入，脱落，置換誤りを，$N$ は入力文字数を表す．

### 5.2 単語カバレージとパープレキシティ

音響上の性能も含めた場合には，上記文字誤り率が良い指標となるが，言語モデルの観点からは，単語カバレージやパープレキシティも調査する必要がある．

ここでは，3.1 節でも述べたように，認識時と同じ Tri-gram モデルを使用して言語モデルの観点から認識対象文を出力する確率が最も高い単語列を探索する．そしてその単語列を用いて単語カバレージおよび単語あたりのパープレキシティを評価する．ただし，最大化したのは文としての尤度であり，単語あたりのパープレキシティを最小化したのではない．単語誤り率を推定する必要がある場合には，最適パス上の単語列を仮に正解単語列としている．

表 4 未知語部分の抽出精度
Table 4 Accuracy of extracting unknown words.

| テキスト | 再現率 | 適合率 |
|---|---|---|
| 日経新聞 | 86.5% | 84.5% |
| 産経新聞 | 85.6% | 79.2% |
| 電子会議室 | 70.6% | 80.7% |
| 毎日新聞 | 92.7% | 79.7% |

### 5.3 未知語モデルと N-gram モデルによる単語の自動分割

上記のように，N-gram モデルを使って最適な単語系列を推定するためには未知語を何らかの方法で統計的にモデル化する必要がある．統計的な未知語モデルとしては文字 N-gram モデルを用いる方法がすでに提案されている[11),12)]．ただ，状態数が多くなり，その推定には非常に多くの訓練データが必要となることから，ここでは文字種（漢字，ひらがな，カタカナ）をクラスと見なした以下のモデルで表現することにした[13)]．ここで $c$ は未知語の文字，$g$ は文字種のクラスを表す．

$$
\begin{aligned}
&Pr(c_1 c_2 \ldots c_n) \\
&= Pr(c_1 \ldots c_n | g_1 \ldots g_n) Pr(g_1 \ldots g_n) \\
&\approx \prod_{i=1}^{n} Pr(c_i | g_i) Pr(g_1 \ldots g_n)
\end{aligned}
$$

参考のため，未知語モデルと N-gram モデルを用いて単語を自動分割し，人手による分割結果と比較して未知語部分の再現率および適合率を評価した結果を**表 4** に示す．ただし，分割位置が厳密に一致しなくとも意味上の異なりを生じていないものは正解としてカウントしている．

N-gram モデルによる単語の自動分割は，上記のような性能評価のためだけでなく，未知語の自動抽出や，既存の形態素解析プログラムに依存しない言語モデルの構築手段を提供するという意味で重要である．この手法によって単語の自動分割と未知語の抽出を行い，その結果に基づいて辞書と N-gram モデルを再構築し，それを再び単語分割に利用するという一連の作業を繰り返すことにより，既存の言語モデルとの整合性を保ちつつ精度の高い未知語抽出および言語モデルの推定が可能になると期待される．

## 6. 実験結果

### 6.1 カバレージとパープレキシティ

6 種類のタスクに対し，40K 語彙のカバレージおよびパープレキシティを，先の定義に基づいて調査した．結果を**表 5** に示す．ここで，毎日新聞，ビジネストークおよび小説は言語モデルの学習時には用いておらず，学習領域外のデータである．

表 5 各タスクに対するカバレージとパープレキシティ：離散発声と連続発声の場合の比較
Table 5 Coverage and perplexity for each task: Comparison of continuous and discrete speech recognition.

| タスク | Coverage | | Perplexity | |
|---|---|---|---|---|
| | Disc. | Cont. | Disc. | Cont. |
| 日経新聞 | 98.0 | 99.5 | 92.2 | 77.4 |
| 産経新聞 | 95.5 | 97.6 | 166.7 | 191.9 |
| 電子会議室 | 95.2 | 97.3 | 246.8 | 258.2 |
| 毎日新聞 | 96.1 | 98.0 | 138.6 | 150.6 |
| ビジネストーク | 97.5 | 98.7 | 122.4 | 118.1 |
| 小説 | 94.6 | 96.7 | 196.4 | 217.2 |

表中の Cont. は 5.2 節で述べた方法で自動推定された単語系列に対する結果を示しており，言語モデルを用いて連続音声認識を行う場合に相当すると考えられる．また，離散単語認識との比較のため，同じ文章を人間が単語単位に切り出した場合のカバレージとパープレキシティも示す（Disc. と表示）．予想どおりこの定義では，連続音声認識においてカバレージが一様に改善されることが分かる．一方，単語パープレキシティは離散に比べ若干増大する場合も見られるが，さほど大きな変化ではない．また，離散と連続でタスクごとのパープレキシティの順序関係は変化していない．

### 6.2 不特定話者連続音声認識

訓練時とは異なる 20～50 代の男女各 10 名（20 代 3 名：話者番号 1～3，30 代 4 名：話者番号 4～7，40 代 2 名：話者番号 8～9，50 代 1 名：話者番号 10）が，それぞれ 30 文ずつ読み上げた合計 600 文（入力総数=15,768 文字：8,252 単語，単語パープレキシティ=148.3，未知語率=0%）を用いて認識実験を行った．なお，この実験では全体の 4/5 の文章に対しては句読点も読み上げるように指示しており，それ以外の文章に対しても句読点の自動挿入は行っていない．内容としては，新聞，雑誌，電子会議室の発言をそれぞれ 1/3 ずつ含む．結果を，単語あたりのパープレキシティ，単語誤り率（WER）とともに**表 6** に示す．

このタスクにおける 1 単語あたりの平均文字長は 1.91 であり，平均文字誤り率から推定される単語正解精度は $0.97^{1.91} = 0.943$ であるが，誤りの分布に偏りがあるため，それよりは高い単語正解精度が得られている．20 名の話者の中では 1 名だけ，10%を超える文字誤り率を示した話者がいるが，それ以外の話者についてはおおむね良好な結果が得られており，性別，年齢あるいはパープレキシティと，誤り率の関連は見い出せない．また，この表にはないが，発声速度も人

表 6 不特定話者連続音声認識実験結果（Error Rate：%）
Table 6 Experimental results of speaker-independent speech recognition.

| 話者 | Per-plexity | Error Rate (%) | | | | |
|---|---|---|---|---|---|---|
| | | Sub. | Ins. | Del. | CER | WER |
| 男性 1 | 168.5 | 2.89 | 0.52 | 0.26 | 3.68 | 5.6 |
| 男性 2 | 125.6 | 1.34 | 0.36 | 0.48 | 2.19 | 3.44 |
| 男性 3 | 126.7 | 2.53 | 0 | 0.8 | 3.33 | 3.64 |
| 男性 4 | 124.1 | 1.95 | 0.12 | 0.24 | 2.32 | 2.91 |
| 男性 5 | 218.9 | 0.87 | 0.12 | 0.5 | 1.50 | 1.23 |
| 男性 6 | 168.5 | 1.97 | 0.52 | 0.13 | 2.63 | 4.14 |
| 男性 7 | 125.6 | 1.7 | 0.6 | 0.24 | 2.56 | 4.58 |
| 男性 8 | 126.7 | 3.46 | 0.4 | 0.13 | 4.0 | 5.58 |
| 男性 9 | 124.1 | 1.95 | 0.12 | 0.36 | 2.44 | 3.4 |
| 男性 10 | 218.9 | 1.37 | 0.25 | 0.37 | 2.0 | 2.21 |
| 男性平均 | 148.3 | 1.98 | 0.3 | 0.35 | 2.64 | 3.68 |
| 女性 1 | 168.5 | 1.3 | 0 | 0.26 | 1.58 | 2.68 |
| 女性 2 | 125.6 | 2.56 | 0.12 | 0.36 | 3.04 | 4.81 |
| 女性 3 | 126.7 | 7.73 | 2.26 | 0.13 | 10.1 | 12.6 |
| 女性 4 | 124.1 | 1.46 | 0.12 | 0.24 | 1.83 | 4.13 |
| 女性 5 | 218.9 | 0.50 | 0.12 | 0.25 | 0.87 | 0.24 |
| 女性 6 | 168.5 | 3.29 | 0.26 | 1.05 | 4.61 | 6.34 |
| 女性 7 | 125.6 | 1.7 | 0.24 | 0.24 | 2.19 | 3.89 |
| 女性 8 | 126.7 | 2.13 | 0.66 | 0.93 | 3.73 | 4.61 |
| 女性 9 | 124.1 | 1.96 | 0.36 | 0.61 | 2.93 | 4.62 |
| 女性 10 | 218.9 | 2.25 | 0.5 | 0.75 | 3.5 | 4.18 |
| 女性平均 | 148.3 | 2.45 | 0.45 | 0.48 | 3.39 | 4.81 |
| 全平均 | 148.3 | 2.22 | 0.38 | 0.42 | 3.02 | 4.25 |

によって大きな違いがあった（1.3～2.2 単語/秒）が，その影響も見られない．なお，認識率が特に悪かった話者についても，音声品質には問題がなく，ただ舌足らずな発声が特徴的な話者であった．

なお，この実験では処理速度についてはほとんど考慮しておらず，精度に最適化した結果を求めている．処理速度と認識精度の関係に関しては十分な調査を行っていないので参考値を示すにとどめるが，Intel 社製の MMX Pentium-200 MHz を使用し，ファーストマッチの出力候補単語数など，処理速度に直接影響する閾値だけを変化させて実時間処理を実現したところ，文字誤り率は 3.0%から 4.0%にまで増大した．

### 6.3 パープレキシティと文字誤り率の関係

表 6 について，単語パープレキシティと文字誤り率（CER）の関係を，パープレキシティが一定範囲内に収まる文ごとに集計した結果が**表 7** である．表 6 を見る限りではパープレキシティと CER の相関は非常に低いように思われたが，この表からだとおおむね単調な相関関係が見てとれる．ただ，実際に言語モデルの能力が認識率に影響を及ぼすのは，パープレキシティとして求まるような平均分岐数ではなく，局所的に出現確率が低下する部分であり，今後そのような局所的な統計値との対応関係も調べる必要があると考えている．

表 7 パープレキシティの範囲とその範囲に収まるテスト文に対する文字誤り率
Table 7 Character error rate (CER) of the test sentences within the range of perplexity.

| パープレキシティの範囲 | 該当文数 | CER (%) |
|---|---|---|
| 40～80 | 88 | 2.20 |
| 81～120 | 140 | 2.02 |
| 121～160 | 108 | 2.97 |
| 161～200 | 72 | 3.57 |
| 201～240 | 72 | 3.64 |
| 241～ | 120 | 4.21 |

表 8 学習データ量およびパラメータ数の文字誤り率への影響
Table 8 Effect of the training data size and parameter size on the CER.

| 学習データ | 正規分布総数 | CER (%) |
|---|---|---|
| Set-25 K | 30 K | 5.0 |
| | 60 K | 4.9 |
| Set-25 K + Set-100 K | 30 K | 3.2 |
| | 60 K | 3.0 |

### 6.4 学習データ量とモデルの複雑さの影響

次に，音響モデルの学習データ量およびモデルの複雑さが認識精度に与える影響について調査した．結果を**表 8** に示す．なお，テストデータは，表 6 と同じものである．音素環境依存モデルの総状態数は約 3,000 であるので，正規分布総数 30 K が混合数 10，60 K が混合数 20 の場合に相当する．この結果を見る限りでは混合数を増加させるメリットは学習データ量によらずあまりない．一方，モデルの学習データ量については，215 名から得られた Set-25 K（計 25,800 文，355,454 単語）では明らかに不足であり，658 名から得た Set-100 K（計 97,384 文，1,338,307 単語）を追加することで，認識率が顕著に改善されたことが分かる．

ただ，表 6 の結果に見られるように，突出して誤り率の高い話者がまだ存在することから，学習データ量についてはさらに増やす必要があると思われる．

## 7. 句読点の自動挿入[10)]

### 7.1 句読点の自動挿入方法

連続音声によるディクテーションの場合，発声をより自然なものにするため，句点（。）および読点（、）を自動的に挿入したいという要望がある．また，今後放送音声の書き起こしや会議の議事録採取などへの応用を考える場合，句読点があると認識結果の可読性を高めることができる．

ここでは息継ぎ位置と句読点位置がある程度対応することを利用して，句読点を自動挿入する．具体的に

は句点および読点に対して「マル」,「テン」といった読みを表すモデルに加え，無音のモデルを割り当てておく．この結果，ポーズが挿入された部分は音響的には無音として認識されるが，言語的には句読点または“透過単語”（この単語自体は言語モデルによって予測されるが，後続する単語の予測の際の条件としては使用されない単語）として処理され，言語的に可能性の高い方の系列が認識結果として出力されることになる．なお，透過単語の出現確率（つまり，無音の出現確率）は統計的にではなく，予備実験を通じて実験的に与えている．

### 7.2 実験結果

6.2節で用いたデータのうち，句読点を読み上げていなかった話者4名，合計120文のデータを用いて句読点の自動挿入実験を行った．なお，読み上げた文には句読点が記載されているが，句読点で息継ぎをするなどの特別な指示は与えず，あくまでも自然に読み上げさせている．

実験の結果，句点に関しては120個の指定に対し，121個が検出され，ほぼ100%正解を得ることができた．一方，読み上げ文に記載されていた読点は52か所であったが，134個の読点を自動検出した．なお，文中のポーズは232カ所で検出されている．脱落は皆無であったものの，明らかに多くの読点が挿入されていることになる．ただし，文章としては特に不自然になるような挿入ではなく，認識結果を読みやすくするという目的は十分に果たしているといえる．

## 8. おわりに

人間の考える単語単位を認識単位とし，連続音声を認識対象とした不特定話者用の日本語ディクテーションシステムの実現可能性について検討した．

認識単位間の比較に関しては，単位によってカバレージや単位長さが異なり，また自動解析精度にも差が出るため，言語モデルや音響モデルまで含めた単位間の性能比較は非常に難しい．しかしながら，ここで示した実験結果からは，言語モデルとしての性能は広く用いられている形態素単位と大差なく，カバレージにおいても遜色がないことが分かった．一方，音響的な側面からは，1文を構成する単位数が少ないために単位あたりの平均モーラ長が長く，単位あたりの読みの異なり数も少ないなどの特徴が明らかになった．

実際，この認識単位を用いた不特定話者大語彙認識実験の結果は良好なもので，約4万語からなる語彙を認識対象とし，単語パープレキシティ=148.3のテスト文に対して，文字正解精度は97%，単語正解精度も95.7%が得られた．

また，連続音声認識時の言語モデルの性能評価のために，未知語モデルとN-gramモデルによる単語の自動分割手法を導入した．この手法は既存の単語N-gramモデルを生かし，新たなコーパスを単語単位に自動分割する方法としても大変重要である．実際，我々は今後新たに得られる言語モデル学習用のコーパスをすべてこの手法によって自動分割する予定である．既存の言語モデルによる単語分割，そしてその結果に基づくモデルの再構築という一連の操作を繰り返すことにより，既存のモデルとの整合性を保ちつつ精度の高いモデルの推定が可能になる．

さらに，ポーズ位置と句読点の挿入位置にある程度対応がつくことを利用して句読点の自動挿入を試み，おおむね良好な結果が得られることを示した．今後放送音声の書き起こしなどに応用する予定である．

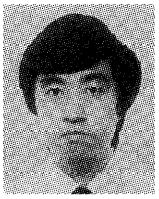
**謝辞** データ使用を許可してくださった，産経新聞社，日本経済新聞社，毎日新聞社（CD-毎日新聞94）ならびに（株）ピープルワールドカンパニーに感謝します．また，形態素解析プログラムを本研究の目的に合わせて変更ならびに整備していただいた荻野紫穂研究員に深謝します．

## 参考文献

1) Gauvain, J.L. and Lamel, L.: Large vocabulary continuous speech recognition: From large vocabulary systems towards real-world applications, 電子情報通信学会論文誌（D-II），Vol.J79-D-II, No.12, pp.2005–2021 (1996).
2) 古井貞熙：大語彙連続音声認識の現状と展望，日本音響学会平成10年度春季研究発表会，1-6-10, pp.19–22 (1998).
3) 河原達也ほか：大語彙日本語連続音声認識研究基盤の整備—評価用連続音声認識プログラムの開発，情報処理学会音声言語情報処理研究会，18-1, pp.1–6 (1997).
4) 松岡達雄，大附克年，森　岳至，古井貞熙，白井克彦：新聞記事データベースを用いた大語い連続音声認識，電子情報通信学会論文誌（D-II），Vol.J79-D-II, No.12, pp.2125–2131 (1996).
5) 西村雅史，伊東伸泰：単語を認識単位とした日本語ディクテーションシステム，電子情報通信学会論文誌（D-II），Vol.J81-D-II, No.1, pp.9–16 (1998).
6) 西村雅史：音声ワープロ最新事情，日本音響学会誌，Vol.54, No.3, pp.229–234 (1998).
7) 丸山　宏，荻野紫穂：正規文法に基づく日本語形態素解析，情報処理学会論文誌，Vol.35, No.7, pp.1293–1299 (1994).
8) Bahl, L.R., et al.: Performance of the IBM

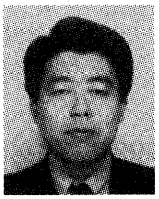
large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task, *Proc. ICASSP'95*, pp.41–44 (1995).
9) 日本電子化辞書研究所：EDR 電子化辞書仕様説明書 (1995).
10) 西村雅史，伊東伸泰，山崎一孝，荻野紫穂：単語を認識単位とした日本語大語彙連続音声認識，日本音響学会平成 9 年度秋季研究発表会，3-1-5, pp.95–96 (1997).
11) 永田昌明：単語頻度の期待値に基づく未知語の自動収集，情報処理学会自然言語処理研究会，116-3, pp.13 20 (1996).
12) 森 信介，山路 治：日本語の情報量の上限の推定，情報処理学会論文誌，Vol.38, No.11, pp.2191–2199 (1997).
13) 伊東伸泰，西村雅史：N-gram を用いた日本語テキストの単語単位への分割，情報処理学会自然言語処理研究会，122-9, pp.57–62 (1997).

(平成 10 年 10 月 5 日受付)
(平成 11 年 2 月 8 日採録)



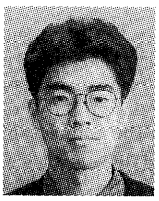
**西村 雅史**（正会員）

昭和 56 年大阪大学基礎工学部生物工学科卒業．昭和 58 年同大学大学院博士前期課程修了．同年日本アイ・ビー・エム（株）入社．東京基礎研究所にて音声認識等の音声言語処理の研究に従事．工学博士．平成 10 年本学会山下記念研究賞受賞．電子情報通信学会，日本音響学会各会員．



**伊東 伸泰**（正会員）

昭和 57 年大阪大学基礎工学部生物工学科卒業．昭和 59 年同大学大学院博士前期課程修了．同年日本アイ・ビー・エム（株）入社．東京基礎研究所にて文字認識，音声認識の研究に従事．



**山崎 一孝**（正会員）

昭和 63 年東京工業大学工学部情報工学科卒業．平成 5 年同大学大学院博士課程修了．工学博士．同年日本アイ・ビー・エム（株）入社．東京基礎研究所にて文字認識，音声認識の研究に従事．平成 6 年電子情報通信学会論文賞受賞．電子情報通信学会，日本音響学会各会員．