日本語ディクテーション基本ソフトウェア*
— 1999 年度版 —

河原達也　李晃伸 (京大)　小林哲則 (早稲田大)
武田一哉 (名大)　峯松信明 (豊橋技科大)†　嵯峨山茂樹 (北陸先端大)
伊藤克亘 (電総研)　伊藤彰則 (山形大)　山本幹雄 (筑波大)
山田篤 (ASTEM)　宇津呂武仁 (奈良先端大)‡
鹿野清宏 (奈良先端大)
http://winnie.kuis.kyoto-u.ac.jp/dictation/

2000 年 5 月 8 日

概 要

「日本語ディクテーション基本ソフトウェア」は、大語彙連続音声認識 (LVCSR) 研究・開発の共通プラットフォームとして設計・作成された。このプラットフォームは、標準的な認識エンジン・日本語音響モデル・日本語言語モデル及び日本語形態素解析・読み付与ツール等から構成される。音響モデルは、日本音響学会の音声データベースを用いて学習し、monophone から数千状態の triphone まで用意した。語彙と単語 N-gram (2-gram と 3-gram) は、毎日新聞記事データベースを用いて構築した。認識エンジン Julius は、音響モデル・言語モデルとのインタフェースを考慮して開発された。これらのモジュールを統合して、20000 語彙及び 60000 語彙の日本語ディクテーションシステムを作成し、種々の要素技術の評価を行なった。本ツールキットは、無償で一般に公開されている。[1]

1　はじめに

大語彙連続音声認識 [1][2][3] は、音声を利用した様々なアプリケーションの基盤になる技術であり、音声入力ワープロ、放送やオーディオテープの書き起こしなどの応用が考えられる一方、そこで培われる要素技術は音声対話システムや種々の音声インタフェースに利用できるであろう。

大語彙連続音声認識の実現のためには、高精度の音響モデル、高精度の言語モデル、そして効率のよい認識エンジン (デコーダ) が必要とされ、それらのバランスのよい統合化とともに、実環境においては適応化技術も要求される。このように大規模なシステムの開発と個別要素技術の研究をバランスよく推進していくためには、データベースだけでなくモデルやプログラムを含めた共通基盤を整備することが必要である。また、このような標準的なツールキットは、アプリケーション開発者や他分野の研究者による音声認識技術の利用を容易にし、技術の普及や新たな応用の展開に寄与することも期待できる。

我々は、一般全国紙の 1 つである毎日新聞の記事データを共通の言語・音声コーパス [4][5] に採用し、音声認識のためのソフトウエアツールキットを開発する 3ヶ年 (97 ~ 99 年度) のプロジェクトを推進してきた。本プロジェクトは、主として大学と公的研究機関のメンバーから構成され、情報処理振興事業協会 (IPA) の「独創的先進的情報技術に係わる研究開発」の支援の下で進められた [6][7] [8][9][10]。この成果物の「日本語ディクテーション基本ソフトウエア」は、標準的な音響モデル、言語モデル、認識エンジン、及び形態素解析・読み付与ツールから構成され、一般に無償で公開されている。これらのコーパスとソフトウエアの関連を図 1 に示す。

本稿では、このツールキットの 99 年度版に関して、各モジュールの仕様、及びこれらを統合して構成される日本語ディクテーションシステムについて述べる。さらに、各モジュールとシステム全体の性能評価についても報告する。

98 年度版との主要な変更点は以下の通りである。

---

*本ソフトウェアは、情報処理振興事業協会 (IPA) が実施した独創的先進的情報技術に係わる研究開発の成果物である

†現在 東大

‡現在 豊橋技科大

[1] 本ソフトウェアの入手方法
http://www.lang.astem.or.jp/dictation-tk/
mailto:　dictation-tk-request@astem.or.jp

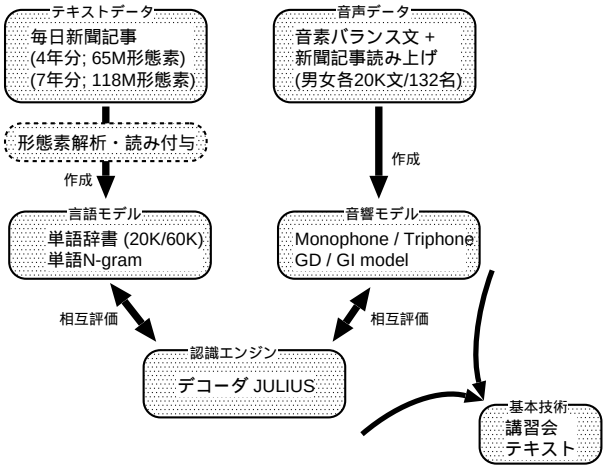

図 1: ツールキットの概要

1. デコーダ Julius における単語間 triphone の扱いを改善し、探索の高精度化を実現した。
2. 音響モデルとして、新たに Phonetic Tied-Mixture (PTM) モデルを作成し、これにより認識精度を維持しながら、処理効率を大きく改善した。
3. 言語モデルとして、60000 語のモデルを用意し、新聞記事に対しては 99%以上のカバレージを実現した。
4. これらのモデルに対応できるように、デコーダ Julius を改良した。特に PTM モデルによる認識の高速化のために、Gaussian pruning を実装した。

2 モデルとプログラムの仕様

2.1 音響モデル

音響モデル [11] は、対角共分散の混合連続分布 HMM に基づいており、HTK のフォーマット [12] で提供される。

表 1 に示すように、音素環境独立 (monophone) モデルから数千状態の triphone モデルまで、種々の日本語音響モデルを構築した。基本的に男性/女性別 (GD) に構築されているが、代表的なものについては性別非依存 (GI) モデルも用意している。また 99 年度版では、Phonetic Tied-Mixture (PTM) モデルを作成した。これは、monophone と同様にガウス分布集合を構成するが、混合分布の重みのみを triphone によって変えるものある。つまり、monophone と通常の triphone の中間的な位置づけである。[2]

本ツールキットで採用している日本語の 43 音素の

[2] monophone と状態共有 triphone は 98 年度版と同一である。

表 1: 音響モデルの一覧

model	#state	#mixture	gender
monophone	129	4, 8, 16	GD, GI
triphone 1000	1000	4, 8, 16	GD
triphone 2000	2000	4, 8, 16	GD, GI
triphone 3000	3000	4, 8, 16	GD
PTM triphone	3000/129	64	GD, GI

GD: Gender Dependent, GI: Gender Independent

表 2: 音素の一覧

```
a i u e o a: i: u: e: o: N w y
p py t k ky b by d dy g gy ts ch
m my n ny h hy f s sh z j r ry
q sp silB silE
```

一覧を表 2 に示す。この音素表記は、日本音響学会 (ASJ) の音声データベース委員会で策定されたものに基づいている。表中で、a:～o:は長母音を、q は促音を表す。ポーズに関しては、silB, silE, sp の 3 種類のモデルを用意した。これらはそれぞれ、文頭・文末・文中 (単語間) のポーズに対応している。ポーズのモデル自身はコンテキスト独立であるが、sp モデルは他の音素のコンテキストになりうる。

音響モデルの学習には、日本音響学会の音素バランス文からなる研究用連続音声データベース (ASJ-PB) の全部と、新聞記事読み上げ音声コーパス (ASJ-JNAS) のうち 100 名分を利用した。合計で男女とも、約 130 名の話者による 2 万文のデータである。

音声データは 16kHz,16bit でディジタル化され、フレーム周期 10ms で、12 次元のメル周波数ケプストラム係数 (MFCC) を計算する。その一次差分 (ΔMFCC) とパワーの一次差分 (ΔLogPow) も計算する。その結果、各フレームの特徴量ベクトルは 25(=12+12+1) 次元となる。入力チャネルのミスマッチを補正するために、発話単位でケプストラム平均による正規化 (CMN) を実行する。

各音素モデルは 3 状態 (分布を持たない初期・最終状態を除く) から構成される。状態遷移はすべて、left-to-right であり、初期状態からの遷移と最終状態への遷移は 1 つに限定している。

実際の音声認識に triphone モデルを適用するには、単語辞書中 (単語間の接続も含む) に出現可能な音素の組み合わせをすべてカバーする必要がある。そのために、言語的に可能な音素の組 (logical triphone; 約 21000 種類) と実際に用意されたモデル (physical

triphone; 約 8000 種類) の対応を指定するファイルを用意する。現実には、このすべてのモデルに対して十分な学習データはないので、決定木に基づいたクラスタリングによって状態の共有を行う。このクラスタリングのしきい値を調整することによって、種々のモデル (状態数が約 1000,2000,3000) を作成した。

PTM モデルについては、ガウス分布集合は 64 混合分布の monophone のものを、状態集合は 3000 状態の triphone のものを、それぞれ用いる。triphone の各状態は、monophone 上で対応する状態の分布集合を共有し、混合分布の重みのみを個別に持つ。その上で、分布集合も含めて再学習して最適化する。これにより、効率的に音素環境依存モデルを表現するとともに、信頼度の高いモデルを安定して学習できる [13]。

2.2 形態素解析と単語辞書

単語辞書は、{ 語彙のエントリ }-{ 表記 }-{ 音素記号列 } の集合であり、HTK のフォーマット [12] で提供される。単語辞書は、音響モデルと言語モデルの両方と整合性をとっている。すなわち、音素記号はすべて音響モデルでカバーされており、また語彙のエントリにはすべて言語モデルにより (少なくとも 1-gram により) 生起確率が定義される。なお、後述するデコーダでは、1-gram 確率が定義されていない単語は未知語 UNK カテゴリとして処理される。

日本語においては、語彙の定義が形態素解析システムに依存する。本ツールキットでは、形態素解析システムに奈良先端科学技術大学院大学で開発されている ChaSen[14] を採用した。

音声認識用の単語辞書を構成するためには、形態素に区分化するだけでなく、読みを正しく付与する必要がある [15]。そこで、単語の読みを従来の仮名遣いに基づくものから実際の読み方に基づくものに変更した。読みはすべてカタカナ表記で、その表記法は原則として、NHK 日本語発音アクセント辞典 (新版) に従った。漢語・和語・外来語を問わず、長音化して読まれる場合は長音記号「ー」で表記した。

(例) または　マタワ
綴り　ツズリ
いう　ユウ
アルミニウム　アルミニューム
東京　トーキョー

また、不規則な読みを正しく付与するための後処理プログラムを作成した。特に日本語では、多くの数詞が複数の読みを持つ上に、後続する助数詞が連濁と呼ばれる変化を生じる場合がある。このような数詞や助数詞の読み変化は、ChaSen の後処理プログラム ChaWan により対処する。

(例) 1 本，2 本，3 本
1 分，2 分，3 分

さらに、特殊な用言の読みに関して後処理を行い、数字表現を位取り標準形に正規化して、語彙エントリが決定される。

一般に、同一の表記でも品詞タグが異なると連接する形態素の傾向が異なる。また上記の例のように、読みが連接する形態素に依存する場合は、読みに応じてエントリを区別しておくことによりそのような制約を表現できる。したがって言語モデルの精度向上のために、語彙のエントリは表記だけでなく読みと品詞タグによっても区別し、{ 表記 }+{ 読み }+{ 品詞タグ } の形式で定義した。複数の読みを持つ形態素で読みが確定できない場合は、複数の読みを併記した形で 1 つの語彙エントリとなっている。

(例)	本+ホン+39	[本]	h o N
	本+ホン+33	[本]	h o N
	本+ボン+33	[本]	b o N
	本+ポン+33	[本]	p o N
	本+{ホン/モト}+2	[本]	h o N
	本+{ホン/モト}+2	[本]	m o t o

語彙は、毎日新聞の 91 年 1 月から 94 年 9 月までの 45ヶ月分の記事データ (CD-毎日新聞) において高頻度の形態素 (=単語) から構成される。また、句読点と疑問符のエントリも含めており、これらの発音はポーズに対応づけられている。種々の語彙サイズにおけるカバレージを表 3 に示す。最終的に、5000 語、20000 語と 60000 語の単語辞書を用意している。60000 語の辞書は、99%を上回るカバレージを実現している。[3]

2.3 言語モデル

設定した語彙に基づいて、N-gram 言語モデルを構築した。すなわち、単語 2-gram と 3-gram を学習した。いずれもバックオフ平滑化を行っており、バックオフ係数の推定には Witten Bell ディスカウンティングを用いている。これらは、CMU-Cambridge SLM ツールキット [16] のフォーマットで提供される。

[3] 5000 語と 20000 語の辞書は 98 年度版と同一である。60000 語の辞書の語彙は 75ヶ月分の記事データから構成している。

表 3: 語彙とカバレージ

vocabulary size	coverage
5000	88.3%
20000	96.4%
24000	97.0%
53000	99.0%
60000	99.2%
101000	99.7%
154000	99.9%

ポーズに対応づけた句読点なども通常の単語と同様に扱われており、結果として、句読点の出現確率の推定によりポーズの出現位置の推定を代用している。

ベースライン N-gram エントリのカットオフのしきい値は、2-gram、3-gram ともに 1 とした [cutoff-1-1]。また省メモリ向きのモデルを作成するために、N-gram エントリの削減を行った。従来、カットオフのしきい値を大きくすることにより言語モデルの縮小が行われており、ここでも 2-gram、3-gram ともに 4 に設定したモデル [cutoff-4-4] を用意した。これに加えて、単純に出現頻度に基づいて削減するのではなく、エントロピに基づいて削減する方法も試みた。これは、エントロピの変化 (=相対エントロピ) が最小になるように最尤推定を行いながら、エントリを逐次的に削除していく方法である [17]。これにより、3-gram のエントリのみを約 10%まで削減したモデル [compress10%] を用意した。

言語モデルの学習用のコーパスとして、毎日新聞の記事データを使用した。その際に、見出しや表などの読み上げに適さないテキストを前処理によって除去している [4]。20000 語のモデルについては、まず語彙を構成するのに用いた 45ヶ月分 (91 年 1 月 ~ 94 年 9 月; 65M 単語) で学習したが、その後 75ヶ月分 (91 年 1 月 ~ 94 年 9 月, 95 年 1 月 ~ 97 年 6 月; 118M 単語) に学習データを増やした。60000 語のモデルについては、語彙も言語モデルも 75ヶ月分で構築している。圧縮にはエントロピに基づく方法のみを適用した。用意した言語モデルの一覧を表 4 と表 5 に示す。

なお、後述するデコーダでは、2-gram の各エントリに 18 バイト、3-gram の各エントリに 6 バイトを割り当てる。forward-backward 探索を行なうデコーダのために、逆向きの 3-gram を用意した。ただし 60000 語のモデルについては、通常の前向きの 3-gram も用意している。

表 4: 20K 言語モデルの一覧

	2-gram entries	3-gram entries
45month cutoff-1-1	1,238,929	4,733,916
45month cutoff-4-4	657,759	1,593,020
45month compress10%	1,238,929	473,176
75month cutoff-1-1	1,675,803	7,445,209
75month cutoff-4-4	901,475	2,629,605
75month compress10%	1,675,803	744,438

表 5: 60K 言語モデルの一覧

	2-gram entries	3-gram entries
75month cutoff-1-1	2,420,231	8,368,507
75month compress10%	2,420,231	836,852

2.4 デコーダ

認識エンジン Julius[18] は、前述の音響モデル・言語モデルとインタフェースがとれるように開発された。種々のタイプのモデルを扱えるので、それらの評価に用いることができる。

音声波形ファイル (16bit PCM)、音響特徴量ファイル (HTK フォーマット) だけでなく、マイク入力にも対応している。Sun, SGI のワークステーション、Linux PC のマイク端子、及び DAT-LINK/netaudio 経由で音声入力が可能となっている。ただし、音声分析は前述の音響モデルで採用している特徴量のみを実装している。

Julius は 2 パス (forward-backward) 探索を行ない、第一 (forward) パスで簡易なモデル (2-gram) により単語候補をしぼった上で、第二 (backward) パスで高精度なモデル (3-gram) を用いて再探索・再評価を行う。

第一パスでは、木構造化辞書に言語モデル確率を動的に割り当てながら、フレーム同期ビーム探索を実行する。木の途中のノードには、プレフィックスを共有する単語集合の 1-gram 確率の最大値を付与しておき、木の葉 (=単語終端) に達した際に 2-gram 確率を与える。単語間の音素環境依存性の扱いについては、単語終端では可能な音素環境依存モデルの最大値で近似し、単語始端では最尤履歴から与えることにする [-iwcd1 オプション]。[4]

この探索においては、単語対近似ではなく 1-best 近

[4] 98 年度版のデコーダでは、単語間の音素環境依存性を正しく処理していなかった。

似を用いる。この粗い近似により第一パスの認識精度は低下するが、それは tree-trellis 探索を行なう第二パスで回復される。第一パスの結果である中間表現として単語トレリスインデックスを採用している。これは、各フレームにおいてビーム内に終端ノードが残った単語、そのスコア、対応する始端フレームの集合からなり、第二パスの探索で予測単語とそのスコアを効率よく逆引きできるようになっている。

第二パスにおいては、単語 3-gram に加えて、単語間の音素環境依存性の処理を正しく行なうことで、より高精度な認識を実現している。仮説終端の音素についても、遅延なく厳密に処理することもできる [-iwcd2 オプション]。スタックデコーダによる探索を行うが、単純な best-first 探索では探索に失敗して解が得られない場合があった。そこで、各単語長の仮説数に上限 (=ビーム幅) を設定して、探索が幅優先に陥った場合には強制的に前に進める方式を実装した [19]。なお、第二パスのスコアが第一パスのスコアよりもよくなる場合があるため、厳密な A*探索にはならず、最初に出力される候補が最尤解とは限らない。そこで、10 候補を出力してから最尤の解を選ぶことにする。

ビーム幅や、言語モデル重み、挿入ペナルティなどのパラメータは、各パスで調整できるようになっている。音響モデルの種類毎に、標準版と高速版のデコーディングオプションを用意した。標準版では精度を重視して、第二パスで音素環境依存性の処理を厳密に行う。高速版では、第一パスのビーム幅をしぼるとともに、第二パスで最初の候補が得られた時点で探索を打ち切る。

さらに、ガウス分布数の大きい PTM モデルの出力確率計算の高速化のために、Gaussian pruning を実装した。これは、多次元ベクトルの距離 (=確率密度関数の指数部分) の計算を行う際に、途中の次元で枝刈りを行うものである。この際のしきい値として、既に最後の次元まで計算された k-best の値を用いるのが安全であるが、効率はよくない [safe pruning]。途中の次元でもビーム幅を設定したり [beam pruning]、未計算の次元をヒューリスティックに見積ったりすることにより [heuristic pruning]、適格性は失われるが、さらに効率化される。標準版では safe pruning が、高速版では beam pruning が適用される [13]。

デコーダの概要を表 6 にまとめる。

表 6: デコーダ Julius の概要

	cross-word phone model	language model	search approx.
1st pass	approximate	2-gram	1-best
2nd pass	accurate	3-gram	N-best

3 日本語ディクテーションシステム

前章で述べた各モジュールを統合して、日本語ディクテーションシステムを設計・実装した。

システムのブロック図を図 2 に示す。デコーダの仕様に基づいて、音響モデルと言語モデルが統合されている。第一パスでは単語 2-gram を利用し、音素環境依存 (CD) モデルの処理は単語内に限り、単語間は近似している。より高精度で計算量の大きい単語 3-gram と厳密な単語間の音素環境依存性 (CD) は、しぼられた候補を再探索・再評価する第二パスで適用される。

音響モデルと言語モデルにはいくつかの種類があるので、それに伴って種々のシステム構成が考えられる。デコーダのパラメータの設定によっても、いくつかのバリエーションが考えられる。

ここでは、20000 語彙と 60000 語彙のディクテーションシステムを作成した。各モジュールは異なる研究機関で開発されたが、仕様に沿って問題なく統合することができた。

4 モジュールとシステムの評価

統合したシステムを用いて、逆に各モジュールの評価を行なうことができる。すなわち、各モジュールを交換することによって、その認識精度や処理効率に対する影響を調べる。ここでは、主に 20000 語彙のシステムを用いて評価を行った。

評価用サンプルには、日本音響学会の新聞記事読み上げ音声コーパス (ASJ-JNAS) のうち、音響モデルの学習に用いていないセット (IPA-98-Testset) [5] を用いた。これは、男女それぞれについて、23 名の話者による合計 100 文の発声からなる。

サンプル文は、94 年 10 月 ~ 12 月の記事データから抽出されており、言語モデル学習に対してもオープンとなっている。文長やパープレキシティに関しても、コーパス全体の分布を反映している。[6] サンプル中の句読点等を除いた総単語数は 1575 で、20000 語の辞

[5] www.milab.is.tsukuba.ac.jp/jnas/test-set/male/male1_LARGE.txt
www.milab.is.tsukuba.ac.jp/jnas/test-set/female/female1_LARGE.txt
[6] NORMAL:76 + LONG:24, LPP:26 + MPP:45 + HPP:29

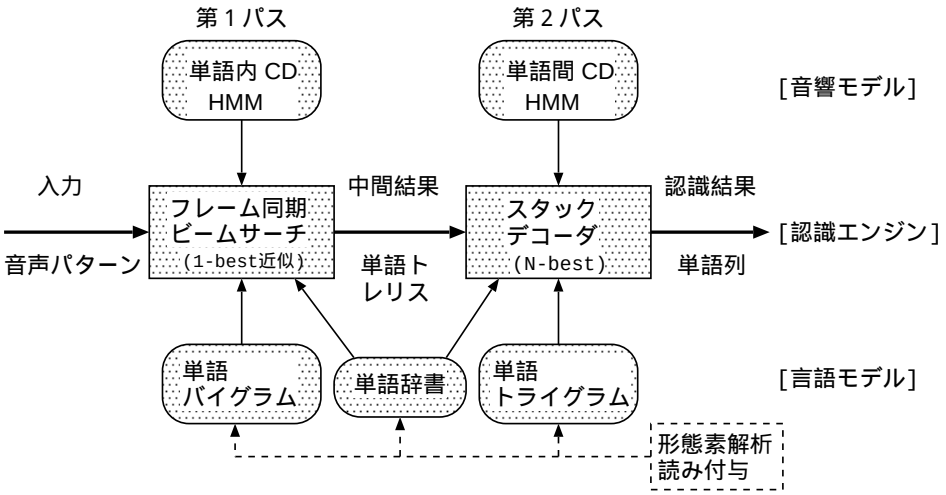

図 2: 日本語ディクテーションシステムの構成

書による未知語率は 0.44%である。

評価尺度としては単語認識精度 (word accuracy) を用いている。日本語の単語認識精度の算出にはいくつかの問題点がある。まず、単語の単位が形態素解析によって異なる問題がある。文字単位で認識率を算出する方が客観性が高いが、ここではわかりやすさのために本ツールキットで定義した単語 (=形態素) の単位に基づいている。次に、形態素解析システムを固定しても、形態素の区分化に曖昧性が生じる問題がある。例えば、「ともに」という形態素は「とも」と「に」の 2 つの形態素に区分化される場合もある。この問題に対処するために、複合語を連結する処理を施してから正解とのマッチングを行う。さらに、漢字とかなの表記の曖昧性の問題がある。例えば、「作る」は「つくる」とも表記される。ただし、すべてをかな表記に変換すると、「操作」と「捜査」のような同音異義語間の混同も正解と判定してしまうことになる。ここでは、漢字表記で算出している。これらの処理は機械的に判定しており、人間が目視で照合した場合と比べて、0.5%程度誤り率が増加する [20]。[7]

4.1 音響モデルの評価

まず、種々の音響モデルに対する評価を行なった。ここでは、75ヶ月分のデータで学習されたベースライン言語モデル [75month cutoff-1-1] と、標準版デコーディングを用いている。PTM モデルでは safe pruning を適用している。

男性話者に関する単語認識精度を表 7 に、女性話者に関する単語認識精度を表 8 に示す。

表 7: 音響モデルの評価 (男性; accuracy)

	mix.4	mix.8	mix.16
GD monophone	75.3	79.6	83.9
GI monophone	68.3	78.0	81.7
GD triphone 2000	92.0	92.6	94.3
GI triphone 2000	89.3	91.8	92.5
GD PTM 129x64 (3000)		92.4	
GI PTM 129x64 (3000)		89.5	

表 8: 音響モデルの評価 (女性; accuracy)

	mix.4	mix.8	mix.16
GD monophone	75.5	80.7	88.9
GI monophone	76.0	80.8	84.7
GD triphone 2000	92.0	94.4	95.2
GI triphone 2000	92.3	93.4	94.8
GD PTM 129x64 (3000)		94.6	
GI PTM 129x64 (3000)		94.3	

PTM モデルは少ない総分布数で、triphone モデルに近い認識精度を実現していることがわかる。なお、PTM モデルによる認識時間は、triphone の場合の半分以下である。また、性別非依存 (GI) モデルは性別依存 (GD) モデルに比べて、全般に誤り率が数%程度増加している。

4.2 言語モデルの評価

次に、言語モデルの評価を行なった。実験には、男性の triphone 2000x16 モデルと標準版デコーディングを使用した。20000 語彙と 60000 語彙について、ベースラインのモデル [cutoff-1-1] とエントロピに基づいて圧縮したモデル [compress10%] を比較した。

[7] 実験の詳細に関しては下記を参照。
http://winnie.kuis.kyoto-u.ac.jp/pub/julius/result99/

表 9: 言語モデルの評価

	accuracy	LM size
20K 75month cutoff-1-1	94.3	79MB
20K 75month compress10%	94.3	38MB
60K 75month cutoff-1-1	93.7	100MB
60K 75month compress10%	93.5	55MB

各モデルによるメモリ使用量と単語認識精度を表 9 に示す。

語彙サイズを 20000 語から 60000 語に増やしても、同一のビーム幅にもかかわらず、認識率の低下は 1%未満であった。ただし、処理時間は 30%程度増大した。また、3-gram エントリを 1/10 に削減したモデルを使用しても、ほとんど認識率が低下しないことが確認された。

4.3 デコーダの評価

デコーディングアルゴリズムの評価も、男性の性別依存で行った。ベースラインの言語モデル [75month cutoff-1-1] を用いた。

triphone 2000x16 モデルを用いる場合において、単語間 triphone の扱いの改善による高精度化に関する実験結果を表 10 に示す。第 1 パス (1st pass) と第 2 パス (final) それぞれについて、単語認識精度を示している。98 年度版に比べて、第 1 パスで近似的に単語間 triphone を扱うことにより、第 1 パスの精度が大きく向上した。さらに、第 2 パスにおける仮説終端音素の単語間 triphone をより厳密に扱うことにより、誤り率が大きく削減された。これは、探索エラーを半分以下にしたことに相当する [21]。

PTM モデルを用いる場合の Gaussian pruning による高速化に関する実験結果を表 11 に示す。単語認識精度とガウス分布距離の計算量 (=乗算回数) の相対値を示している。予備実験の結果、各 64 混合分布のうち上位 2 分布のみを計算すれば、認識率に影響しないことが明らかになった。いくつかの枝刈り手法を比較した結果、単純かつ安全な safe pruning でも距離計算の量を約 1/2 に削減でき、各次元でしきい値を設定する beam pruning により約 1/5 に削減できた。

4.4 システムの性能

日本語ディクテーションシステムの全体としての性能を、20000 語彙のシステムについて表 12 に、60000 語彙のシステムについて表 13 にまとめる。性能の指標として、単語認識精度 (Acc.:word accuracy) と単語正解率 (Corr.:word %correct)、及び実時間ファクタによる処理速度を示している。

表 10: デコーディングの評価 (triphone における高精度化)

GD triphone 2000x16	accuracy final (1st pass)
(1998 version)	92.0 (78.9)
enhanced XW-CD: 1st pass	93.0 (85.2)
enhanced XW-CD: 1st&2nd pass	94.3 (85.0)

XW-CD: Cross-Word Context Dependency handling

表 11: デコーディングの評価 (PTM における高速化)

GD PTM 129x64	accuracy	Gaussian computation
No Gaussian pruning	92.8	100
safe pruning	92.4	52
heuristic pruning	90.7	36
beam pruning	91.2	21

ここでは典型的なシステムとして、高精度版と高速版を挙げている。高精度版では、triphone モデルと標準的なデコーディングを使用することにより、約 95%の単語認識率を達成している。高速版では、PTM モデルにより 90%程度の認識率を維持した上で、高速化を図っている。これにより、(計測に使った計算機より高速なハイエンドの) パソコンでほぼ実時間に近い動作が可能となっている。また、圧縮言語モデルを用いてメモリ効率も改善している。

20000 語彙については、比較のために 98 年度版の数値も示している。高速版・高精度版ともに計算量は増大しているが、誤り率がおおむね 2/3 程度に削減されていることがわかる。

5 まとめ

本ソフトウェアの主要な特徴は、汎用性と拡張性である。各モジュールのフォーマットとインタフェースには一般性があり、また改良や置換が容易である。実際に本稿の実験は、異なる機関で開発されたモジュールを交換・統合することにより行われた。したがって本ツールキットは、個別モジュールの研究や特定の目的のシステムの開発に適している。また、形態素解析と読み付与ツールを用いることにより、種々のタスクへの適用が容易になっている。

統合して構成されるディクテーションシステムが、20000 語彙のタスクで約 95%の認識精度を達成し、また実時間に近い動作で 90%の精度を実現できることを

表 12: 20K システムの構成 (98 年度版からの改善)

	高速版		高精度版	
	98 年度版	99 年度版	98 年度版	99 年度版
acoustic model	monophone 16 (0.5MB)	PTM 129x64 (3.0MB)	triphone 2000x16 (8.6MB)	
language model	75month compress10% (38.0MB)		75month cutoff-1-1 (78.5MB)	
decoding	fast	fast	(1998 ver.)	standard
CPU time	1.1x RT	2.3x RT	8.4x RT	12.8x RT
Acc,Corr(male)	82.6, 83.5c	89.1, 91.1c	92.0, 93.2c	94.3, 95.4c
Acc,Corr(female)	85.7, 87.1c	91.8, 93.1c	93.2, 94.1c	95.2, 96.2c
Acc,Corr(GD ave)	84.2, 85.3c	90.5, 92.1c	92.6, 93.7c	94.8, 95.8c
Acc,Corr(GI)	81.5, 84.0c	89.7, 91.1c	90.3, 91.7c	93.7, 94.7c

RT (Real Time): 5.8sec./sample, CPU: Ultra SPARC 300MHz

表 13: 60K システムの構成

	高速版	高精度版
acoustic model	PTM 129x64 (3.0MB)	triphone 2000x16 (8.6MB)
language model	75 compress10% (54.5MB)	75 cutoff-1-1 (99.7MB)
decoding	fast	standard
CPU time	2.9x RT	16.9x RT
Acc,Corr(male)	89.1, 90.9c	93.7, 94.6c
Acc,Corr(female)	91.6, 92.7c	93.4, 94.9c
Acc,Corr(GD ave)	90.4, 91.8c	93.6, 94.8c
Acc,Corr(GI)	88.9, 90.5c	93.2, 94.2c

RT (Real Time): 5.8sec./sample, CPU: Ultra SPARC 300MHz

示して、本ツールキットの有用性を明らかにした。

本システム (デコーダ) は、標準的な Unix 環境 (Solaris, IRIX, PC Linux など) で動作する。今後は、Windows PC への移植を行うとともに、API などを実装していく予定である。

謝辞：本プロジェクトに対して有益なコメントや多大な協力を頂きましたアドバイザリ委員の方々や関係各位に感謝します。

参考文献

[1] S.J.Young. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing magazine*, Vol. 13, No. 5, pp. 45–57, 1996.

[2] 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, Vol. J79-DII, No. 12, pp. 2125–2131, 1996.

[3] 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, Vol. J81-DII, No. 1, pp. 10–17, 1998.

[4] 伊藤克亘, 伊藤彰則, 宇津呂武仁, 河原達也, 小林哲則, 清水徹, 田本真詞, 荒井和博, 峯松信明, 山本幹雄, 竹沢寿幸, 武田一哉, 松岡達雄, 鹿野清宏. 大語彙日本語連続音声認識研究基盤の整備–学習・評価用テキストコーパスの作成–. 情報処理学会研究報告, 97-SLP-18-2, 1997.

[5] K.Itou, M.Yamamoto, K.Takeda, T.Takezawa, T.Matsuoka, T.Kobayashi, K.Shikano, and S.Itahashi. The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus. In *Proc. ICSLP*, pp. 3261–3264, 1998.

[6] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏. 日本語ディクテーション基本ソフトウェア (97 年度版) の性能評価. 情報処理学会研究報告, 98-SLP-21-10, 1998.

[7] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏. 日本語ディクテーション基本ソフトウェア (97 年度版). 日本音響学会誌, Vol. 55, No. 3, pp. 175–180, 1999.

[8] T.Kawahara, T.Kobayashi, K.Takeda, N.Minematsu, K.Itou, M.Yamamoto, T.Utsuro, and K.Shikano. Sharable software repository for Japanese large vocabulary continuous speech recognition. In *Proc. ICSLP*, pp. 3257–3260, 1998.

[9] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 伊藤克亘, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏. 日本語ディクテーション基本ソフトウェア (98 年度版) の性能評価. 情報処理学会研究報告, 99-SLP-26-6, 1999.

[10] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 伊藤克亘, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏. 日本語ディクテーション基本ソフトウェア (98 年度版). 日本音響学会誌, Vol. 56, No. 4, pp. 255–259, 2000.

[11] 武田一哉, 伊藤彰則, 伊藤克亘, 宇津呂武仁, 河原達也, 小林哲則, 清水徹, 田本真詞, 荒井和博, 峯松信明, 山本幹雄, 竹沢寿幸, 松岡達雄, 鹿野清宏. 大語彙日本語連続音声認識研究基盤の整備 – 汎用音素モデルの作成–. 情報処理学会研究報告, 97-SLP-18-3, 1997.

[12] S.Young, J.Jansen, and J.Odell D.Ollason P.Woodland. *The HTK BOOK*, 1995.

[13] 李晃伸, 河原達也, 武田一哉, 鹿野清宏. Phonetic tied-mixture モデルを用いた大語彙連続音声認識. 電子情報通信学会技術研究報告, SP99-100, NLC99-32 (99-SLP-29-8), 1999.

[14] 松本裕治, 北内啓, 山下達雄, 平野善隆. 日本語形態素解析システム「茶筅」 version 2.0 使用説明書. Information Science Technical Report NAIST-IS-TR99008, 奈良先端科学技術大学院大学, 1999.

[15] 伊藤克亘, 山田篤, 天白成一, 山本俊一郎, 踊堂憲道, 宇津呂武仁, 山本幹雄, 鹿野清宏. 日本語ディクテーションのための言語資源・ツールの整備. 情報処理学会研究報告, 99-SLP-26-5, 1999.

[16] *The CMU-Cambridge Statistical Language Modeling Toolkit v2*, 1997.

[17] 踊堂憲道, 鹿野清宏, 中村哲. N-gram モデルのエントロピーに基づくパラメータ削減に関する検討. 情報処理学会研究報告, 99-SLP-27-18, 1999.

[18] 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. 電子情報通信学会論文誌, Vol. J82-DII, No. 1, pp. 1–9, 1999.

[19] 李晃伸, 河原達也. 大語彙連続音声認識エンジン Julius における A*探索法の改善. 情報処理学会研究報告, 99-SLP-27-5, 1999.

[20] 伊藤克亘, 山本俊一郎, 鹿野清宏, 中村哲. ディクテーションにおける日本語の特質を考慮した単語正解率判定ツール. 日本音響学会研究発表会講演論文集, 3-Q-19, 春季 1999.

[21] 河原達也, 南條浩輝, 李晃伸. 大語彙連続音声認識における認識誤り原因の自動同定. 日本音響学会研究発表会講演論文集, 2-1-17, 秋季 1999.