自然言語処理 131－6
音声言語情報処理 26－6
（1999．5．28）

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

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)
伊藤克亘 (電総研)　山本幹雄 (筑波大)
山田篤 (ASTEM)　宇津呂武仁　鹿野清宏 (奈良先端大)
http://www.itakura.nuee.nagoya-u.ac.jp/~takeda/IPA/

あらまし

「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。98 年度版では各モジュールに大幅な改良・改善がなされた。本稿ではその仕様を述べるとともに、20000 語彙のディクテーションタスクにおける要素技術の評価を報告する。本ツールキットは、無償で一般に公開されている。

Evaluation of Japanese Dictation ToolKit – 1998 version –

Tatsuya Kawahara, Akinobu Lee (Kyoto Univ.), Tetsunori Kobayashi (Waseda Univ.),
Kazuya Takeda (Nagoya Univ.), Nobuaki Minematsu (Toyohashi Univ. of Tech.),
Katsunobu Itou (ETL), Mikio Yamamoto (Tsukuba Univ.),
Atsushi Yamada (ASTEM), Takehito Utsuro (Nara Inst. of Sci. & Tech.),
Kiyohiro Shikano (Nara Inst. of Sci. & Tech.)

Abstract

A sharable software repository for Japanese LVCSR (Large Vocabulary Continuous Speech Recognition) is introduced. It has been developed under collaboration of researchers of different academic institutes in Japan. The platform consists of a standard recognition engine, Japanese phone models and Japanese statistical language models as well as Japanese morphological analysis tools. As an integrated system of these modules, we have implemented a baseline 20000-word dictation system and evaluated various components. The software repository is available to the public.

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

1 はじめに

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と考えられる。

そこで我々は、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 [1][2] に採用し、音声認識のための共有のソフトウエアリポジトリを開発する3ヶ年 (97～99年度) のプロジェクトを推進している。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」の支援を受けている [3][4][5][6]。この成果物である「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98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の構成について述べる。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。

97年度版との主要な変更点は以下の通りである。

1. 形態素解析に ChaSen を採用し、この辞書を改善した上で、読み付与プログラム ChaWan を作成した。これに基づいて20000語の語彙を用意した。この結果、単語辞書のカバレージと読みの精度が大きく改善された。
2. 7年分の記事データを利用し、また圧縮アルゴリズムを導入することにより、省メモリで高精度な言語モデルを構築した。
3. 性別非依存の音響モデルを用意した。
4. デコーダ Julius を改善し、認識率を向上した上で処理速度を2倍以上にした。その結果、高効率版システムでは、ほぼ実時間認識が可能となった。

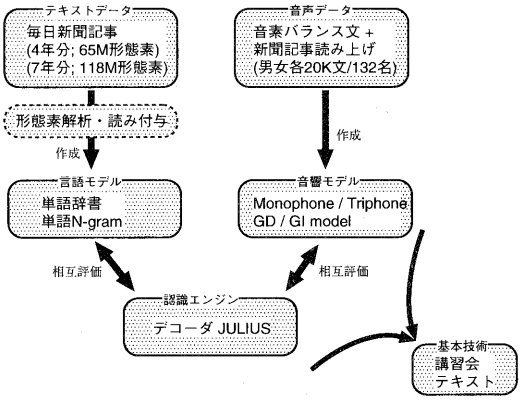

図 1: ツールキットの概要

表 1: 音響モデルの一覧

model	#states	#mixtures	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

GD: Gender Dependent, GI: Gender Independent

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 [7] は、混合連続分布 HMM(対角共分散) に基づいており、HTK のフォーマット [8] 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ており、使用目的に応じて適当なモデルを選択することができる。音響モデルは基本的に男性/女性別 (GD) に構築されているが、表1の通り、一部については性別非依存 (GI) モデルも用意している。本ツールキットで採用している日本語の43音素の一覧を表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。

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。

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							

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、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。triphone モデルに関しては、決定木に基づいたクラスタリングによって、類似した音素環境間で状態の結びを行う。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) を構築した。

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 [8] 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。

語彙 (=語彙のエントリの集合) は、毎日新聞の 91 年 1 月から 94 年 9 月までの 45 か月分の記事データ (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。

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。98 年度版では、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 を採用した。[1] ただし、本ツールキットのためにいくつかの修正を行った。まず、IPA 品詞体系に基づいて ChaSen 用の単語辞書を構成した。また、新聞記事に出現する固有名詞に対処するために、人名や地名などを中心に大量のエントリを追加した。

[1] 97 年度版では、形態素解析は新情報処理開発機構の新聞記事タグデータ (RWCP テキストデータベース) に基づいていた。

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。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

さらに、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。なお、句読点と疑問符のエントリはポーズに対応づけられている。

(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

表 3: 語彙とカバレージ

vocabulary size	coverage
5000	88.2%
6135	90.0%
20000	96.5%
22959	97.0%

種々の語彙サイズにおけるカバレージを表3に示す。98年度版では5000語と20000語の単語辞書を用意している。

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 [9] のフォーマットで提供される。

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。

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が、45ヶ月分 (91年1月～94年9月；65M単語) と75ヶ月分 (91年1月～94年9月, 95年1月～97年6月；118M単語) の2通りを比較した。なお、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 [1]。

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 [10] も試みた。これは、エントロピの変化 (元の言語モデルとのクロスエントロピ) 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。これにより3-gramのエントリのみを約1/10(=10%)に削減したモデル (compress-10%) を用意した。

表 4: 20K 言語モデルの一覧

	2-gram entries	3-gram entries
75-month cutoff-1-1	1,675,803	7,445,209
75-month cutoff-4-4	901,475	2,629,605
75-month compress-10%	1,675,803	744,438
45-month cutoff-1-1	1,238,929	4,733,916
45-month cutoff-4-4	657,759	1,593,020
45-month compress-10%	1,238,929	473,176

45ヶ月分のデータと75ヶ月分のデータの2通りに対して上記を適用したので、用意した言語モデルの一覧は表4のようになった。

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。

2.4 デコーダ

認識エンジンJulius[11]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。

98年度版では、ファイル入力 (音声波形/音響特徴量) だけでなく、マイク入力にも対応した。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。

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。最尤の単語履歴とプレフィックスを共有する単語集合に応じて2-gram確率を木のすべてのノードに分配する方式を基本とする。ただし高速化のために、木の途中のノードには1-gram確率を静的に付与しておき、木の葉 (=単語終端) に達した際に2-gram確率を与える方式も実装した (1-gram factoring)。

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 (CD) の処理を行なうことで、

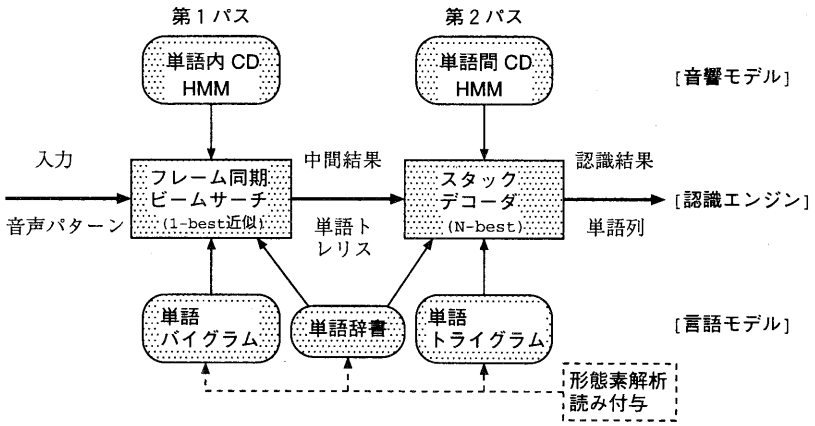

図 2: 日本語ディクテーションシステムの構成

表 5: デコーダ Julius の概要

	acoustic model	language model	search approx.
1st pass	intra-word CD	2-gram	1-best
2nd pass	inter-word CD	3-gram	N-best

CD: Context-Dependent model

より高精度の認識を実現している。スタックデコーディングサーチを実行するが、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、文長 (=単語数) 毎に仮説数の上限を設定して、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 (enveloped best-first) を実装した。

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。ただし高効率版では、第 1 候補が得られた時点で探索を打ち切る (1-best candidate)。

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。

デコーダの概要を表 5 にまとめる。

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のみに限られている。より高精度で計算量の大きい単語 3-gram と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。例えば、音素環境独立な monophone モデルを用いることにより、効率性重視のシステムが構成できる。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。

98 年度版では、20000 語彙のディクテーションシステムを開発した。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。

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) 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。サンプル中の句読点等を除いた総単語数は1575で、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 [12]。

なお、特記していない限り、ほぼ認識率が収束した十分なビーム幅を使用している。[2]

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 (75-month cutoff-1-1) と、最終的にチューンされたデコーダ (Julius 2.0) を用いている。

男性話者に関する単語認識精度を表6に、女性話者に関する単語認識精度を表7に示す。

monophoneモデルは十分な認識精度を得るのに多数の混合分布を必要とすること、及びtriphoneモデルは2000状態でほぼ認識率が収束していることがわかる。

また、性別非依存 (GI) モデルを用いると誤り率が2%程度増加している。

[2] 実験の詳細に関しては下記を参照。
http://winnie.kuis.kyoto-u.ac.jp/pub/julius/result98/

表6: 音響モデルの評価 (男性; accuracy)

	mix.4	mix.8	mix.16
monophone	75.3	79.6	83.9
triphone 1000	86.9	90.0	90.5
triphone 2000	90.5	90.9	92.0
triphone 3000	89.2	90.5	90.5
GI monophone	68.3	78.0	81.7
GI triphone 2000	87.4	88.7	90.0

表7: 音響モデルの評価 (女性; accuracy)

	mix.4	mix.8	mix.16
monophone	75.5	80.7	88.9
triphone 1000	90.3	91.6	92.6
triphone 2000	91.0	92.2	93.2
triphone 3000	90.5	90.4	91.3
GI monophone	76.0	80.8	84.7
GI triphone 2000	89.9	91.8	90.5

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 モデルを利用した。

各モデルによるメモリ使用量と単語認識精度を表8に示す。

75ヶ月分のデータで学習したモデルの方が、45ヶ月分に比べて全般に高い認識率を得ており、学習データ量を増やす効果が確認された。

省メモリ化に関しては、カットオフのしきい値を大きくする (cutoff-4-4) よりも、エントロピに基づいて圧縮したモデル (compress-10%) の方が認識精度の低下が少なくなっており、当該手法の有効性が示された。

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 triphone 2000x16 モデルとベースライン言語モデル (75-month cutoff-1-1) を用いて行なった。

表8: 言語モデルの評価

	accuracy	LM size
45-month cutoff-1-1	89.8	54MB
45-month cutoff-4-4	89.3	23MB
45-month compress-10%	89.3	28MB
75-month cutoff-1-1	92.0	79MB
75-month cutoff-4-4	90.9	34MB
75-month compress-10%	91.8	38MB

表 9: デコーダにおけるアルゴリズムの比較

	word accuracy 3-gram (2-gram)
2nd-pass best-first	91.2 (78.9)
2nd-pass enveloped best-first	92.0 (78.9)
+ 1st-pass 1-gram factoring	91.2 (73.9)
+ 2nd-pass 1-best candidate	90.8 (78.9)

表 9 に各手法による単語認識精度を、第一パスと第二パスそれぞれについて示す。

表の上半分では、単純な best-first 探索に比べて、各文長毎に仮説数の制限を設ける方式 (enveloped best-first) により探索が安定になる効果が示されている。

下半分では、高速化のために導入した手法の効果を調べている。第一パスの木構造化辞書のノード内で 2-gram でなく 1-gram factoring を使用すると、第一パスの認識結果は大きく低下するが、最終的な第二パスではわずかな低下にとどまっている。これは本デコーダが単語トレリス形式を中間表現に用いているためと考えられる。また、第二パスで第 1 候補が得られた時点で探索を打ち切ると (1-best candidate)、認識率が若干低下する。それぞれの手法により、処理速度は実時間の 1 倍程度改善される。

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表 10 にまとめる。ここでは典型的なシステムとして、高効率版と高精度版を挙げている。

高効率版では、monophone モデルを用いており、デコーディングも高速化を図っている。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。これにより、パソコンでほぼ実時間動作が可能となっている。高精度版では、triphone モデルと高精度な言語モデルを使用することにより、おおむね 93%の単語認識率を達成している。

4.5 97 年度版との比較

参考のため、本ツールキット 97 年度版によるシステムとの比較を表 11 に示す。97 年度版の仕様にあわせて、5000 語彙のタスクで、97 年度のテストセット (IPA-97-TestSet) により評価を行った。

高効率版では、同程度の認識精度を維持した上で、2 倍以上の高速化を実現し、実時間処理を可能にした。高精度版でも処理速度を約 2 倍にした上で、誤り率を改善している。

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。したがって、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だけでなく、異なった機関で開発されたモジュールの交換・統合や評価を行うことが可能である。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 90% を上回る認識精度を実現し、また実時間動作もほぼ可能であることを示して、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, Linux など) で動作する。今後、さらに高精度化と高効率化を図っていく予定である。

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く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] 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.

[2] 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.

[3] 河原達也, 李晃伸, 伊藤克亘, 伊藤彰則, 宇津呂武仁, 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–評価用連続音声認識プログラムの開発–. 情報処理学会研究報告, 97-SLP-18-1, 1997.

表 10: 20K システムの構成

	efficient version	accurate version
Acoustic Model	monophone 129x16 (0.5MB)	triphone 2000x16 (8.6MB)
Language Model	75-month compress-10% (38.0MB)	75-month cutoff-1-1 (78.5MB)
Decoding	1-gram factoring 1-best candidate	2-gram factoring 10-best candidates
CPU time	2.2x RT	8.4x RT
Male Dep.	Corr. 83.8 / Acc. 82.9	Corr. 93.2 / Acc. 92.0
Female Dep.	Corr. 87.5 / Acc. 86.2	Corr. 94.1 / Acc. 93.2
Gender Indep.	Corr. 82.7 / Acc. 81.2	Corr. 91.7 / Acc. 90.3

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

表 11: 5K システムの改善

	98 年度版		97 年度版	
	efficient	accurate	efficient	accurate
Acoustic Model	mono129x16@98 (0.5MB)	tri2000x16@98 (8.6MB)	mono129x16@97 (0.5MB)	tri2000x16@97 (8.6MB)
Lexicon	ChaSen		RWCP	
Language Model	75 compress-10 (18.5MB)	75 cutoff-1-1 (44.9MB)	45 cutoff-1-2 (23.5MB)	45 cutoff-1-2 (23.5MB)
Decoding	1-gram factor 1-best cand.	2-gram factor 10-best cand.	1-gram factor 1-best cand.	2-gram factor 10-best cand.
CPU time	1.1x RT	6.3x RT	2.8x RT	13.0x RT
Corr/Acc(male D)	87.3 / 86.5	94.5 / 93.3	87.9 / 85.4	93.6 / 92.6
Corr/Acc(female D)	90.0 / 88.9	94.4 / 93.7	89.9 / 89.0	94.1 / 93.3
Corr/Acc(GI)	86.5 / 85.7	92.0 / 91.0	NA	NA

RT (Real Time): 4.1sec./sample, CPU: Ultra SPARC 300MHz
認識率は IPA-97-TestSet による 5K タスクでの評価

[4] 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エア (97 年度版) の性能評価. 情報処理学会研究報告, 98-SLP-21-10, 1998.

[5] 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.

[6] 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.

[7] 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.

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

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

[10] 踊堂憲道, 鹿野清宏, 中村哲. 情報量に基づく trigram パラメータの逐次的削減手法. 情報処理学会研究報告, 98-SLP-22-17, 1998.

[11] 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.

[12] 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.