日本音響学会誌 57 巻 3 号 (2001), pp. 210-214

技術報告

43.72.Ne

日本語ディクテーション基本ソフトウェア (99 年度版)*

河原達也*1 李 晃伸*1 小林哲則*2 武田一哉*3 峯松信明*4
嵯峨山茂樹*5 伊藤克亘*6 伊藤彰則*7 山本幹雄*8 山田 篤*9
宇津呂武仁*10 鹿野清宏*11

[要旨] 「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。99 年度版では更なる高精度化・高速化そして大語彙化がなされた。本稿ではその仕様を述べると共に，20,000 語彙及び 60,000 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[1]。

キーワード 大語彙連続音声認識，ソフトウェア
Large vocabulary continuous speech recognition, Software

1. はじめに

1997 年度から 3 年間にわたって，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，「日本語ディクテーション基本ソフトウエア」[1-4]の開発を進めてきた。これは，音声認識研究の共通基盤を目指したソフトウェアツールキットで，一般に無償で公開されている。この全体像を図-1 に示す。

本稿では，この 99 年度版 (=IPA 最終版) に関して，各モジュールの仕様と性能評価を報告する。ここでは紙面の都合上，98 年度版[2]との変更点を中心に述べる。主要な変更点は以下のとおりである。

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，60,000 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。

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[5]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である。具体的には，ガウス分布集合は 64 混合分布の monophone のものを，状態集合は 3,000 状態の triphone のものを，それぞれ用いた。その上で，分布集合も含めて再学習して最

---

* Japanese Dictation ToolKit—1999 version—, by Tatsuya Kawahara, Akinobu Lee, Tetsunori Kobayashi, Kazuya Takeda, Nobuaki Minematsu, Shigeki Sagayama, Katsunobu Itou, Akinori Ito, Mikio Yamamoto, Atsushi Yamada, Takehito Utsuro and Kiyohiro Shikano.

*1 京都大学
*2 早稲田大学
*3 名古屋大学
*4 東京大学
*5 北陸先端科学技術大学院大学
*6 電子技術総合研究所
*7 山形大学
*8 筑波大学
*9 京都高度技術研究所
*10 豊橋技術科学大学
*11 奈良先端科学技術大学院大学
(問合先：河原達也 〒606-8501 京都市左京区吉田本町 京都大学大学院情報学研究科)

(2000 年 7 月 7 日受付，2000 年 10 月 24 日採録決定)

[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

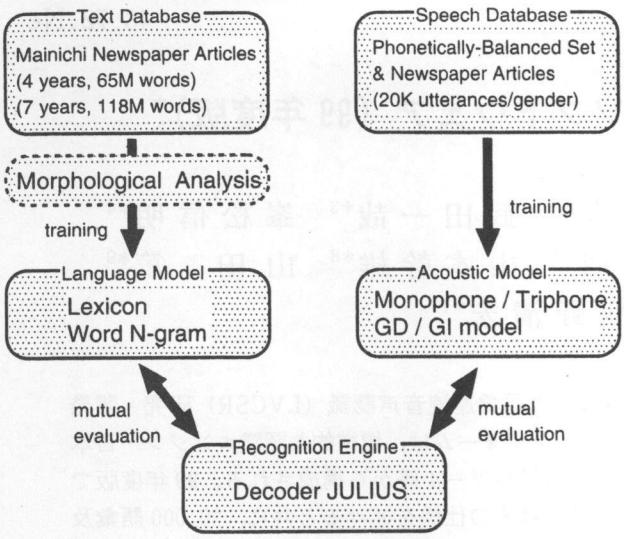

図-1　ツールキットの概要

表-1　音響モデルの一覧

model	# state	# mixture	gender
monophone	129	4, 8, 16	GD, GI
triphone 1000	1,000	4, 8, 16	GD
triphone 2000	2,000	4, 8, 16	GD, GI
triphone 3000	3,000	4, 8, 16	GD
PTM triphone	3,000/129	64	GD, GI

GD：Gender Dependent，GI：Gender Independent

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共に，信頼度の高いモデルを安定して学習できる[6]。

音素表記，音響分析条件，音響モデルの学習データに関しては，基本的に97年度[1]から同じである。また，monophoneと状態共有triphoneは，モデル自体が98年度版と同一である。

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[5]で提供される。

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ChaSen[7]を採用している。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一つの語彙エントリとなっている[8]。

語彙は，毎日新聞の91年1月から94年9月までの45ヵ月分の記事データ（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。種々の語彙サイズにおけるカバレージを**表-2**に示す。最終的に，

表-2　語彙サイズとカバレージ

Vocabulary size	Coverage
5,000	88.3%
20,000	96.4%
24,000	97.0%
53,000	99.0%
60,000	99.2%
101,000	99.7%
154,000	99.9%

5,000語，20,000語と60,000語の単語辞書を用意している。60,000語の辞書は，99%を上回るカバレージを実現している。なお，5,000語と20,000語の辞書は98年度版と同一である[2]。

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。

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。ここでは，エントロピの変化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を適用した[10]。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。

20,000語のモデルについては，98年度版と同一である。これらはまず語彙を構成するのに用いた45か月分（91年1月～94年9月；65M単語）で学習したが，その後75か月分（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。60,000語のモデルについては，語彙も言語モデルも75か月分で構築している。用意した言語モデルの一覧を**表-3**と**表-4**に示す。

前向き・後向き探索を行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

[2] 最終的な60,000語の辞書の語彙は75か月分の記事データから構成した。

表-3 20,000 語彙 言語モデルの一覧

	2-gram entries	3-gram entries
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
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

表-4 60,000 語彙 言語モデルの一覧

	2-gram entries	3-gram entries
75 month cutoff-1-1	2,420,231	8,368,507
75 month compress 10%	2,420,231	836,852

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（16 bit 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。

Julius は 2 パス探索を行い，第 1（前向き）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 2（後向き）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 1 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[-iwcd 1 オプション]。

第 2 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[-iwcd 2 オプション]。ビーム幅を設定したスタックデコーダによる探索を行う[12]。

98 年度版のデコーダでは，単語間の音素環境依存モデルは第 1 パスでは全く適用せず，第 2 パスでも仮説終端の音素については次の単語が展開されるまで処理していなかった。これが最大の変更点である[3]。

音響モデルの種類ごとに，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 2 パスで厳密な音素環境依存性の処理（-iwcd 2 オプション）を適用するが，高速版では行わない。また高速版では，第 1 パスのビーム幅をしぼるとともに，第 2 パスで最初の候補が得られた時点で探索を打ち切る。更に，状態当たりの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[6]。

3. モジュールとシステムの評価

各モジュールを統合して，20,000 語彙と 60,000 語彙の日本語ディクテーションシステムを設計・実装した。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 20,000 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）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。サンプル中の句読点等を除いた総単語数は 1,575 で，20,000 語の辞書による未知語率は 0.44%である。なお，60,000 語彙では未知語は 1 個（0.06%）であった。

単語認識精度（word accuracy）は，複合語を連結する処理を施してから，漢字表記で算出している[4]。

ここでは，今年度新たに作成されたモデルを中心に評価結果を示す。他のモデルの評価については，98 年度版の結果[2]を参考にされたい[5]。

3.1 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った。ここでは，ベースライン言語モデル[75 month cutoff-1-1]と，標準版デコーディングを用いている。男性話者に関する単語認識精度を**表-5** に，女性話者に関する単語認識精度を**表-6** に示す。

[3] 軽微な修正は多数ある。
[4] 実験の詳細に関しては下記を参照。
http ://winnie.kuis.kyoto-u.ac.jp/pub/julius/result 99/
[5] ただし昨年度と同一の音響・言語モデルについても，デコーダが改善されているために，昨年度の数値より良くなっている。

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分かる。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。

3.2 言語モデルの評価

次に，言語モデルの評価を行った。実験には，男性のtriphone 2,000×16モデルと標準版デコーディングを使用した。20,000語彙と60,000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[compress 10%]を比較した。各モデルによるメモリ使用量と単語認識精度を**表-7**に示す。

語彙サイズを20,000語から60,000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。

表-5 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 129×64 (3000)		92.4	
GI PTM 129×64 (3000)		89.5	

表-6 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 129×64 (3000)		94.6	
GI PTM 129×64 (3000)		94.3	

表-7 言語モデルの評価

	Accuracy	LM size
20 K 75 month cutoff-1-1	94.3	79 MB
20 K 75 month compress 10%	94.3	38 MB
60 K 75 month cutoff-1-1	93.7	100 MB
60 K 75 month compress 10%	93.5	55 MB

3.3 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[75 month cutoff-1-1]を用いた。

triphone 2,000×16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-8**に示す。第1パス（1 st 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[-iwcd 1オプション]により，第1パスの精度が大きく向上した。更に，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[-iwcd 2オプション]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[13]。

3.4 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，20,000語彙のシステムについて**表-9**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。

ここでは典型的なシステムとして，高精度版と高効率版を挙げている。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。高効率版では，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。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，（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。

表には，比較のために98年度版の数値も示している。高効率版・高精度版ともに計算量は増大しているが，誤り率が大きく（おおむね2/3程度に）削減されていることが分かる。

4. 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

表-8 デコーディングの評価

	Accuracy final (1 st pass)
(1998 version)	92.0 (78.9)
Enhanced IW-CD：1 st pass	93.5 (85.0)
Enhanced IW-CD：1 st & 2 nd pass	94.3 (85.0)

IW-CD：Inter-Word Context Dependency handling

表-9 20,000 語彙ディクテーションシステムの構成

	Efficient system		Accurate system	
	1998 version	1999 version	1998 version	1999 version
Acoustic model	monophone 16 (0.5 MB)	PTM 129×64 (3.0 MB)	triphone 2000×16 (8.6 MB)	
Language model	75 month compress 10% (38.0 MB)		75 month cutoff-1-1 (78.5 MB)	
Decoding	fast	fast	(1998 ver.)	standard
CPU time	1.1×RT	2.3×RT	8.4×RT	12.8×RT
Acc, Corr (male)	82.6, 83.5	89.1, 91.1	92.0, 93.2	94.3, 95.4
Acc, Corr (female)	85.7, 87.1	91.8, 93.1	93.2, 94.1	95.2, 96.2
Acc, Corr (GD ave)	84.2, 85.3	90.5, 92.1	92.6, 93.7	94.8, 95.8
Acc, Corr (GI)	81.5, 84.0	89.7, 91.1	90.3, 91.7	93.7, 94.7

RT (Real time)：5.8 s/sample, CPU：Ultra SPARC 300 MHz

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。従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。

統合して構成されるディクテーションシステムが，20,000 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 への移植を行うと共に，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いただ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。

文　献

[1] 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)," 音響学会誌 **55**, 175-180 (1999).

[2] 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)," 音響学会誌 **56**, 255-259 (2000).

[3] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro and K. Shikano, "Free software toolkit for Japanese large vocabulary continuous speech recognition," in Proc. ICSLP, Vol. 4, 476-479 (2000).

[4] 河原達也, 李 晃 伸, 小林哲則, 武田一哉, 峯松信明, 嵯峨山茂樹, 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 (99 年度版) の性能評価," 情処学会研報 SLP-31-2, NL-137-7 (2000).

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

[6] 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," 信学技報 SP 99-100, NLC 99-32 (99-SLP-29-8) (1999).

[7] 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「茶筌」version 2.0 使用説明書," Inf. Sci. Tech. Rep. NAIST-IS-TR 99008, 奈良先端科学技術大学院大学 (1999).

[8] 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," 情処学会研報 99-SLP-26-5 (1999).

[9] The CMU-Cambridge Statistical Language Modeling Toolkit v 2 (1997).

[10] 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," 情処学会研報 99-SLP-27-18 (1999).

[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," 信学論 **J82-DII**, 1-9 (1999).

[12] 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," 情処学会研報 99-SLP-27-5 (1999).

[13] 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," 音講論集 2-1-17 (1999.9).