自然言語処理 125−12
音声言語情報処理 21−10
(1998. 5. 29)

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

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)　伊藤彰則 (山形大)　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。これは、複数の大学・公的研究機関の研究者の協力プロジェクトの成果である。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、5000 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。

Evaluation of Japanese Dictation ToolKit – 1997 version –

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

Abstract

The project of developing LVCSR (Large Vocabulary Continuous Speech Recognition) platform is introduced. It is a collaboration of researchers of different academic institutes and intended to develop a sharable software repository of not only databases but also models and programs. The platform consists of a standard recognition engine, Japanese phone models and Japanese statistical language models. As an integrated system of these modules, we have implemented a baseline 5000-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 [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と個別要素の研究をバランスよく推進していくためには、共通のソフトウェアプラットフォームを整備することが必要であると考えられる。

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] に採用し、共有のソフトウエアプラットフォームを開発する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「独創的情報技術育成事業」の支援を受けている [5]。この成果である「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、及び認識エンジンから構成され、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97年度版に関して、各モジュール (音響モデル・言語モデル・認識エンジン) 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。

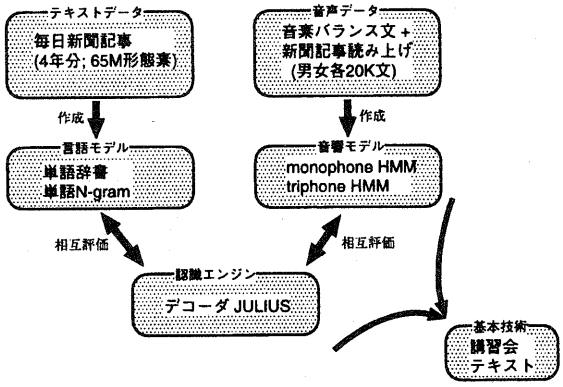

図 1: 大語彙連続音声認識研究の基盤

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 [6] 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 [7] 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 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。認識精度を優先する場合には高精度なモデルを、処理効率を優先する場合には単純なモデルを用いる。音響モデルはすべて、男性 / 女性の別に構築されている。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。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。

表 1: 音響モデルの一覧

	状態数	混合分布数
monophone	129	4, 8, 16
triphone 1000	1000	4, 8, 16
triphone 2000	2000	4, 8, 16
triphone 3000	3000	4, 8, 16

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、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、可能な音素の組(logical triphone)と実際に用意されたモデル(physical triphone)の対応を指定するファイルを用意する。現実には、すべての音素の3つ組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、類似した音素環境をまとめて学習を行なう。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1000,2000,3000)を構築した。

2.2 単語辞書

単語辞書[4]も、HTKのフォーマット[7]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は言語モデルのエントリと一致している。

語彙は、毎日新聞の1991年1月から1994年9月までの45か月分の記事データ(CD-毎日新聞91～94年版)において高頻度の単語(=形態素)から構成される。日本語においては語彙は形態素解析器によって定義されることが多いが、ここで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)に基づいている。語彙のエントリは、表記だけでなく形態素カテゴリによっても区別される。種々の語彙サイズにおけるカバレージを表3に示す。

日本語の漢字は通常、複数の読みを持つので、形態素の多くに複数の読みがふられている。記号などのいくつかのエントリはポーズに対応づけられている。

97年度版では5000語の辞書を用意している。20000語の辞書も近い将来に用意する予定である。

表3: 語彙とカバレージ

語彙サイズ	カバレージ
5000	85.8%
8129	90.0%
20047	95.7%
27634	97.0%

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。これらは、CMU-Cambridge SLMツールキット[8]のフォーマットで提供される。

単語間のポーズも統計的言語モデルの枠組みで扱われており、その出現確率はポーズに対応する記号エントリを用いて推定されている。

言語モデルの学習用のコーパス(毎日新聞91年1月～94年9月)のサイズは、前処理の結果、240万文・6500万単語(=形態素)となっている。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で1、3-gramで2とした。

5000語彙の言語モデルの仕様を表4に示す。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。また、カットオフのしきい値を大きく(4と8)することにより、省メモリ向きのモデルも用意した。なお、言語モデルのパープレキシティの値は、学習用コーパスと異なる時期(94年10月～94年12月)のテストセットの文を用いて計算した。

表4: 5K N-gramの仕様

	カットオフ	エントリ数	パープレキシティ
2-gram	1	578,653	107
3-gram	2	1,978,931	70

2.4 デコーダ

認識エンジンJULIUS[5][9]は、前述の音響モデルと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。

JULIUS は 2 パス探索を行ない、第一パスで単語 2-gram を、第二パスで単語 3-gram を用いる。

第一パスでは、木構造辞書に 2-gram 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。2-gram 確率は、最尤の単語履歴とプレフィックスを共有する単語に応じて、木のすべてのノードに分配される。

ここでは、単語対近似ではなく 1-best 近似を採用している。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。単語トレリスインデックスを中間表現に用いて、効率的なスタックデコーディング探索を実現している。単語グラフを用いる探索と比較して、かなり少ない計算量と記憶量で同等以上の認識精度を得ることが示されている。特に探索に必要なメモリ量が、標準的なパソコンで動作する程度まで大幅に削減できた [9]。

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。

第二パスにおいては、単語間の音素環境依存性 (CD) の処理を行なうことで、より高精度の認識を実現している。スタックデコーダに基づく第二パスは、文単位の N-best 候補を出力することができる。

デコーダの概要を表 5にまとめる。

表 5: デコーダ JULIUS の概要

	音響モデル	言語モデル	近似
第一パス	単語内 CD	2-gram	1-best
第二パス	単語間 CD	3-gram	N-best

CD: Context-Dependent (音素環境依存) モデル

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、音素環境依存性 (CD) 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。

97 年度版では、標準的な 5000 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を用いた。男女それぞれについて、10 名の話者による 10 文の発声である 100 サンプルを用いた。サンプル文は言語モデル学習に対してもオープンとなっている。

評価尺度としては単語認識精度 (word accuracy) を用いている。2-gram と 3-gram それぞれの制約に対して計算している。デコーダの特徴から、第一パス (2-gram) の認識結果は若干よくないが、第二パス (3-gram) の認識精度は信頼できるものである。

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、ベースライン言語モデル (カットオフ 1-2) と、最終的にチューンされたデコーダを用いている。

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 モデルにおいて状態や分布を増やしてもそれほど認識精度が向上していないことがわかる。これは、triphone モデルを十分に学習するには、さらに多くのデータ量が必要であることを示唆している。男性と女性であまり認識精度に差はみられず、これらの結果が信頼できるものであることを示している。

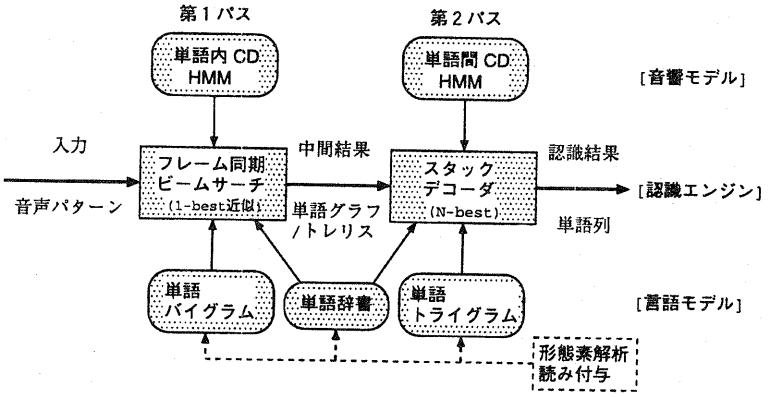

図 2: 日本語ディクテーションシステムの構成

表 6: 音響モデルの評価 (男性)

	mix.4	mix.8	mix.16
monophone	78.1 (67.2)	86.1 (74.7)	87.4 (80.4)
triphone 1000	88.2 (77.7)	91.4 (80.4)	91.7 (82.3)
2000	90.0 (78.0)	91.9 (80.1)	92.8 (82.6)
3000	90.2 (76.9)	92.4 (80.4)	92.7 (80.3)

word accuracy: with 3-gram (with 2-gram)

表 7: 音響モデルの評価 (女性)

	mix.4	mix.8	mix.16
monophone	79.1 (66.4)	84.7 (75.7)	88.6 (80.0)
triphone 1000	91.0 (79.1)	90.6 (82.9)	90.0 (82.3)
2000	91.3 (78.7)	91.8 (81.1)	93.2 (82.9)
3000	91.1 (79.3)	92.9 (80.6)	92.6 (81.7)

word accuracy: with 3-gram (with 2-gram)

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。

その結果を表8に示す。カットオフのしきい値の大きな粗いモデルを用いることにより、若干の認識精度の低下がみられた。また、このモデルはメモリ効率を改善するが、認識時間には影響しないことがわかった。

表 8: 言語モデルの評価

baseline cutoff 1-2	92.8 (82.6)
cutoff 4-8	91.2 (82.8)

word accuracy: with 3-gram (with 2-gram)

4.3 デコーダの評価

デコーディングアルゴリズムと手法の評価も、男性の triphone 2000x16 モデルとベースライン言語モデル (カットオフ 1-2) を用いて行なった。

表 9に、デコーダで導入した種々の手法の効果を示す。単語間の音素環境依存性 (CD) の処理を行なうことにより、認識精度が大きく向上している。本ツールキットでは、語彙が形態素から定義されており、これらは通常の "単語" より小さい単位であるので、単語間の調音結合を処理することが重要であると考えられる。

表 9: デコーダにおける種々の手法の効果

2-gram only	(80.2)
3-gram	86.0 (80.2)
LM weight tuned for 2-pass	86.9 (80.2)
insertion penalty used	87.5 (82.6)
inter-word CD handled (=final)	92.8 (82.6)

word accuracy: with 3-gram (with 2-gram)

4.4 システムの性能

最後に、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、典型的なシステムの構成を3つ挙げている。

処理効率を優先したシステムは、monophoneモデルを用いており、標準的なワークステーションで実時間の3倍で認識を行なうことができる。triphone 3000x8モデルを用いて、デコーディング時のビーム幅を絞ったシステムでは、実時間の6倍で単語誤り率9%程度の性能を実現している。認識精度を重視したシステムは、高精度な音響モデル(2000x16)を用いており、単語誤り率7%を達成している。

表10: 典型的なシステムの構成例

音響モデル	monophone 129x16	triphone 3000x8	triphone 2000x16
デコーディング	candidates reduced	small beam	large beam
認識時間	3x RT	6x RT	12x RT
認識精度(男性)	85.2	91.3	92.8
認識精度(女性)	87.3	91.2	93.2

CPU: Ultra SPARC 300MHz

RT (Real Time): 4.1秒/サンプル

5 まとめと今後の計画

我々が開発しているプラットフォームは、標準的かつポータブルである。また、統合して構成されるディクテーションシステムが十分な性能を実現することを示した。これにより、個々の要素技術やシステム化に関する一層の研究・開発が期待される。いくつかのモジュールは、様々な音声入力のアプリケーションに利用可能である。

本プロジェクトの今後の予定としては、(1)20000語彙のタスクに拡張すること、(2)標準的なパソコンで動作するように効率化すること、が挙げられる。

単語辞書と言語モデルに関しては、一般的な日本語で95%程度のカバレージを実現するように拡張する予定である。そのために、形態素解析と読み付与のプログラムの改善を行なう。音響モデルに関しても、より大きな語彙で十分な認識精度を得られるように改善する予定である。デコーダについても、一層の効率化を行なう。

上記以外の課題としては、音響モデル・言語モデルの種々のタスクへの適応や未知語の処理などがある。

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] 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] 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.

[6] 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.

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

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

[9] 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大語彙連続音声認識エンジンJULIUS. 電子情報通信学会技術研究報告, SP98-3, 1998.