連続音声認識コンソーシアム2000年度版ソフトウエアの概要と評価

河原達也　住吉貴志 (京大)　李 晃伸 (奈良先端大)
武田一哉 (名大)　三村正人 (ASTEM)
伊藤彰則 (山形大)　伊藤克亘 (電総研)　鹿野清宏 (奈良先端大)
http://www.lang.astem.or.jp/CSRC/

あらまし

連続音声認識コンソーシアム (CSRC) は、IPA プロジェクトで開発された「日本語ディクテーション基本ソフトウェア」の維持・発展をめざして、情報処理学会 音声言語情報処理研究会のもとで活動を行っている。本稿では、2000 年度 (2000 年 10 月-2001 年 9 月) において開発されたソフトウエアの概要を述べる。今回、大語彙連続音声認識エンジン Julius の機能拡張、大規模なデータベースを用いた音響モデルの作成、種々の音響・言語モデル及びツール群の整備を行った。本ソフトウエアは現在、有償で頒布している。

Product Software of Continuous Speech Recognition Consortium – 2000 version –

T.Kawahara, T.Sumiyoshi (Kyoto Univ.), A.Lee (NAIST)
K.Takeda (Nagoya Univ.), M.Mimura (ASTEM)
A.Ito (Yamagata Univ.), K.Itou (ETL), K.Shikano (NAIST)

Abstract

Continuous Speech Recognition Consortium (CSRC) was founded last year under IPSJ SIG-SLP for further enhancement of Japanese Dictation Toolkit that had been developed by the IPA project. An overview of the software developed in the first year (Oct. 2000 - Sep. 2001) is given in this report. We have revised the LVCSR (large vocabulary continuous speech recognition) engine Julius, and constructed new acoustic models using very large speech corpora. Moreover, a variety of acoustic and language models as well as toolkits are being set up. The software is currently available by contacting the address below.

本ソフトウェアの申込み先　http://www.lang.astem.or.jp/CSRC/
mailto: csrc@astem.or.jp

1 はじめに

日本の情報処理技術において、現在、日本語音声認識技術が注目され、実用化も視野に入れた研究・開発が活発に行われている。しかしながら、基本性能・頑健性、そしてユーザインタフェースにおいて、一層の改善を必要とするのが実情である。個別要素技術の研究とシステムの開発をバランスよく推進するためには、データベースだけでなくモデルやプログラムを含めたプラットフォームを整備することが必要である。また、これらがソースコードを含めてオープンになっていることも重要である。

そこで我々は平成 9 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」の受託事業として、「日本語ディクテーション基本ソフトウェア」[1][2] [3] の開発を進めてきた。この成果は、標準的な日本語音響モデル、言語モデル、大語彙連続音声認識エンジン Julius、及び種々のツールから構成され、フリーソフトウェアとして公開し、多数の研究機関でベースライン・リファレンスとして利用されている。[1]

平成 12 年 10 月には、本ソフトウェアの一層の拡充・発展とともに、音声認識を用いたアプリケーション開発の促進を目指して、音声認識コンソーシアムが情報処理学会 SLP 研究会のもとで発足し、50 以上の企業・大学の参加を得て、活動を行っている。

本稿では、本コンソーシアムの 2000 年度 (2000 年 10 月〜2001 年 9 月) の成果ソフトウェアの概要を紹介し、また種々のタスクでの評価を示す。

2 認識エンジン Julius

大語彙連続音声認識エンジン Julius[4] [5] について、機能強化と高速化を行うとともに、ネットワーク文法を扱えるパーザも統合した。また、これらの利便性の向上を図るために、Windows 上で Speech API の実装を行った。

2.1 音響尤度計算の高速化

音響尤度計算を高速化するために Gaussian Mixture Selection[6][7] を実装した。本手法は有望なガウス混合分布を予備選択するもので、各入力フレーム毎にまず monophone ンで評価を行い、その尤度の高い状態についてのみ対応する triphone モデルによる尤度計算を行う。これにより、ほとんど認識精度を低下することなく、音響尤度計算を大幅に削減することができ、実時間のディクテーションタスクでも 90%超の認識精度を実現した。

2.2 話し言葉音声を指向した機能強化

講演音声や人間どうしの対話のような話し言葉を認識できるようにするために、デコーダの改善を行った [8][9]。

このような音声では、発話の区切りが明示的になされないので、認識とセグメンテーションを並行して行う逐次デコーディングの方式を実装した。これは、第一パスでショートポーズの仮説が最尤となるフレームが連続した場合に、そこで区切って第二パスを実行し、その単語履歴を引き継いで、後続区間の処理を進めていくものである。なおこの機能は、configure 時にオプション (–enable-sp-segment) を指定することにより有効となる。

さらに、言語モデルで生起確率が十分に推定されていないフィラーやポーズの挿入に対処するために、単語 N-gram 評価時に特定単語をコンテキスト上でスキップする透過単語処理を実装した。透過単語については、単語辞書の第二カラムで指定する。

2.3 記述文法用認識エンジン (Julian)

IPA「日本語ディクテーション基本ソフトウェア」の Julius では、言語モデルとして単語 N-gram モデルしか扱えなかった。しかし、音声認識の比較的単純なアプリケーションでは記述文法を用いる場合が多い。そこで、京都大学で開発してきた記述文法のための認識エンジン Julian[10] を統合した。

Julian では単語カテゴリという概念を導入しており、文法ファイルでは単語カテゴリ (非終端記号) のみで BNF 記法で書き換え規則を記述し、語彙ファイルで各カテゴリに属する単語を記述する。BNF 記法では文脈自由文法を記述できるが、認識時には効率化のため決定性有限状態オートマトン (DFSA) を使用するため、文法はこれにコンパイルできるクラス (左再帰を許さない) に制限される。ただし、実際には大半のタスクに適用可能である。コンパイルや文法チェックのためのツールも用意している。

[1] 「日本語ディクテーションソフトウェア」最終版は、文献 [1] の付録 CD-ROM として収められている。

Julian と Julius はいずれも、第一パスで単語対/2-gram 制約を使用し、第二パスで高次の言語モデル (文法/3-gram) を適用する 2 パス探索で実現されている。音響モデルのインタフェースやデコーディングオプションの多くが共通である。実際に両者は、スタックデコーディングや音響確率計算を含む大半のコードを共有しており、Julius のパッケージ において configure のオプション (–enable-julian) を指定することで Julian が作成される。

2.4 Windows への移植と SAPI の実装

IPA「日本語ディクテーション基本ソフトウェア」は、基本的に音声認識の研究・開発者を対象としており、Unix でしか動作しなかった。音声認識を利用したアプリケーションの開発やマルチモーダルインタフェースやマルチメディア処理などに適用するには、標準的な API を提供することが必要である。そのため今回、マイクロソフト社が策定した Speech API (SAPI 5.1) の Julius/Julian への実装を行った [11]。

SAPI 文法フォーマット (XML) のサポート、マルチコンテクスト・マルチインスタンスの処理などが未対応で、まだ完全に要求仕様を満たしていないが、Unix 版と同様の性能で動作することを確認している。ただし、Unix 版に比べてデコーディングオプションなどは制限される。

なお、Windows SAPI 版の使用に際しては、Microsoft Speech SDK 5.1 をインストールする必要がある。[2]

3 言語モデル

IPA「日本語ディクテーションソフトウェア」では、毎日新聞記事データ (1991〜1997 年分) で学習した単語 N-gram モデルを提供していた。コンソーシアムでは、この新聞記事を用いたモデルを更新するとともに、話し言葉を指向したモデルの作成を行っている。

いずれも、形態素解析に Chasen を用いており、N-gram モデルのフォーマットは CMU-Cambridge SLM ツールキットのバイナリ形式もしくは ARPA 形式である。

3.1 新聞記事モデルの更新

毎日新聞記事データ 1991 年〜2000 年 6 月 (1994 年の後半 3ヶ月を除く) の 111ヶ月分のテキストを用いて、言語モデルを再構築した。このテキストから高頻度語を選定して、2 万語彙 (20K) と 6 万語彙 (60K) の単語辞書を作成した。そして、Julius 用に前向き 2-gram と後向き 3-gram を学習した。カットオフはいずれも 1 で、Witten Bell ディスカウンティングを適用している。また、3-gram をエントロピに基づいて 10%に圧縮したモデル [12] も作成した。

3.2 話し言葉対応モデル

対話などの話し言葉を指向したモデルとして、今回はグルメ・レシピドメインのモデルを作成した。グルメや料理に関する Web 上の掲示板から収集したテキストを元に学習しており、語彙サイズは 2 万である。レストランやレシピの検索などのアプリケーションに用いることができる。

3.3 言語モデル作成用ツール

- N-gram 学習ツールキット Palmkit[13]

  CMU-Cambridge SLM Toolkit とコマンドレベルでほぼ互換で、さらに、クラス N-gram をサポートし、また異なるタイプのモデルや、異なる長さの N-gram を組合わせて利用することもできる。

- N-gram モデル融合ツール [14]

  複数の N-gram モデルを直接融合する。相補的バックオフを適用することで、元コーパスを必要とせず、任意の融合重みや語彙数を設定することができる。

- Web からの N-gram モデル自動作成ツール [15]

  指定されたタスクやドメインのキーワードを基に、Web ページを収集し、テキスト処理や文選択を行って、N-gram モデルを作成する。

- N-gram モデル圧縮ツール [12]

  エントロピに基づいて、N-gram エントリを削減する。

[2] http://www.microsoft.com/speech/speechsdk/sdkinfo.asp

4 音響モデル

IPA「日本語ディクテーションソフトウェア」では、日本音響学会の新聞記事読み上げ音声コーパス (ASJ-JNAS)[16] で学習した音響モデルを提供していた。コンソーシアムでは、ATR の多数話者音声データベース [17] を購入・利用することにより、より高精度なモデルの構築を行う。また、この標準的なモデル以外に種々の話者層・環境における音響モデルも作成する。

いずれも、各音素 3 状態の対角共分散の混合連続分布 HMM に基づいており、HTK フォーマットで提供される。また、音素体系・表記、及び音響分析や特徴量も IPA モデル [2] と同一である。

4.1 高精度音響モデル (CSRC モデル)

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) 及び新聞記事読み上げ音声コーパス (ASJ-JNAS) に加えて、ATR の多数話者音声データベース音素バランス文セット (ATR/BLA) を用いた。

学習データの概要を表 1 に示す。このデータ量は、IPA モデルに比べて話者数で 15 倍、学習量で 3.5 倍の規模を持つ。ただし ATR/BLA コーパスでは女性のデータが男性に比べて 2 倍近く多い。

作成した音響モデルの一覧を表 2 に示す。種々の混合分布数からなる音素環境独立 (monophone) モデル、状態共有 triphone モデル、PTM triphone モデル [18] を作成した。最も高精度なものでは、30 万以上のガウス分布を持ち、IPA モデルの 6 倍以上の規模である。PTM モデルでは、混合分布のコードブックは monophone の状態 (129 個) 毎に用意し、triphone の共有状態 (3000 個) 毎に異なる重みを推定する。

なお今回作成したのは、すべて性別非依存 (GID) モデルである。

4.2 電話帯域モデル

IPA モデルで学習に用いた ASJ-JNAS コーパスに対して、電話帯域 (300〜3400Hz) に帯域制限して音響分析を行ったデータを用いることにより、電話音声向けの音響モデルを作成した。

性別非依存 (GID) の状態共有 triphone モデル (2000 状態 16 混合, 2000 状態 32 混合) である。

表 1: CSRC 音響モデルの学習データの概要

		話者数	発話数	時間
ATR/BLA	男性	1379	42057	59h
	女性	2390	70483	103h
	計	3769	112540	162h
ASJ/JNAS + ASJ/PB	男性	179	28127	47h
	女性	182	28681	51h
	計	361	56808	98h
全体	男性	1558	70184	106h
	女性	2572	99104	154h
	合計	4130	169348	260h

表 2: CSRC 音響モデルの一覧

	状態数	混合分布数
monophone	129	8, 16, 32, 64, 128
triphone 2000	2000	8, 16, 32
triphone 5000	5000	8, 16, 32, 64
PTM triphone	3000/129	64, (128)

4.3 高齢者モデル

NEDO の委託事業「シニア支援システムの開発」プロジェクトで構築されたシニア音声認識用大規模データベース [19] を用いて、高齢者向けの音響モデルを作成した。本データベースは、60〜90 歳の 301 名の被験者が各音素バランス 100 文、新聞記事文 100 文を読み上げたものである。

性別非依存 (GID) と性別依存 (GD) の両方の monophone と PTM triphone (2000 状態 64 混合) がある。

5 CSRC 音響モデルの評価

高精度音響モデルの評価を、読み上げ音声と対話音声を用いて行った。

読み上げ音声として IPA-98-TestSet を、対話音声として ATR 自然発話音声データベース (旅行会話タスク) から対面対話音声 (ATR/SDB)212 文、通訳対話音声 (ATR/SLDB)108 文を用いた。通訳対話は対面対話と読み上げの中間の性質を持つと考えられる。ATR の評価セットは男性のみを用いている。実験に用いた言語モデルは単語 3-gram で、デコーダは Julius 3.1 である [20]。

表 3: IPA モデル (ASJ で学習) による単語認識精度 (%)

		monophone	PTM	triphone
	状態数		3000	2000
	混合数	16	64	16
IPA	男性	82.07	90.55	92.45
	女性	84.03	94.79	94.79
ATR/SDB	男性	63.82	75.25	72.50
ATR/SLDB	男性	78.70	88.94	88.54
CSJ	男性			54.15

表 4: CSRC モデル (ASJ+ATR/BLA で学習) による単語認識精度 (%)

		monophone				PTM	triphone				
	状態数					3000	2000		5000		
	混合数	16	32	64	128	64	16	32	16	32	64
IPA	男性	78.58	78.58	80.77	85.16	90.17	92.46	91.95	91.63	93.60	94.61
	女性	86.25	88.76	88.76	91.43	93.97	94.22	94.36	94.80	94.54	96.13
ATR/SDB	男性	62.80	65.04	68.52	68.65	75.69	74.24	75.24	75.54	76.00	76.19
ATR/SLDB	男性	82.82	83.35	85.63	85.91	90.42	90.52	91.61	89.55	90.50	91.14
CSJ	男性						55.65				57.06

また、開放的融合研究プロジェクトで構築された日本語話し言葉コーパス (CSJ) の講演音声 4 名分 (男性のみ)[9] に対しても評価を行った。

IPA モデルの認識精度を表 3 に、CSRC モデルの認識精度を表 4 に示す。

読み上げ音声に対しては、同一の複雑さのモデル (2000 状態 16 混合) で比べると認識精度はほとんど変わらず学習量や話者数の増強の効果は見られない。しかし、IPA モデルがこのパラメータ数でほぼ飽和し、3000 状態のモデルでは認識精度が逆に低下していたのに対して、大規模な ATR/BLA コーパスを用いた CSRC モデルではパラメータ数の増加につれて認識精度も着実に上昇している。最終的に 5000 状態 64 混合のモデルで、本テストセットに対してこれまでで最高の認識精度を達成した。ただし、女性に比べて全体に男性の認識精度が低いのは、女性の方がデータ量がかなり多いためと考えられる。

対話音声に対しても、CSRC モデルが全体に高い認識精度を得て、効果が確認された。講演音声に関しては一部のモデルのみ評価を行ったが、CSRC モデルの方が高い認識精度を得ている。

6 おわりに

本ソフトウェアは、IPA「ディクテーション基本ソフトウェア」と同様に、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、またソースコードも公開されているので、汎用性と拡張性に富んでいる。認識エンジン Julius については、引き続きアプリケーションとのインタフェースの改善を進めていく予定である。

言語モデルと音響モデルについては、使用目的や環境に応じた多様なモデルの整備を進めている。言語モデルについては今後も話し言葉のモデルの構築を行う。音響モデルについては、電話音声モデルに加えて、車内音声モデルや小児音声モデルなどの作成を予定している。

また、他の種々のプロジェクトとの連携も進めていきたいと考えている。

2000 年度実行委員リスト

代表：鹿野清宏（奈良先端大）
幹事：河原達也（京都大）
　　武田一哉（名古屋大）
　　伊藤克亘（産総研）
　　山田　篤（ASTEM）
委員：伊藤彰則（山形大）
　　宇津呂武仁（豊橋技科大）
　　峯松信明（東大）
　　山本幹雄（筑波大）
　　李　晃伸（奈良先端大）
　　小林哲則（早稲田大）
　　嵯峨山茂樹（東大）
　　岩野公司（東工大）

謝辞：本コンソーシアムの設立に対して協力を頂きましたSLP研究会及び情報処理学会の関係各位、そして運営に対して支援を頂きました会員各位に深い感謝の意を表します。

参考文献

[1] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄. 音声認識システム. オーム社, 2001.

[2] 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. 情処学研報, 2000-SLP-31-2, 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, pp. 476–479, 2000.

[4] 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.

[5] A.Lee, T.Kawahara, and K.Shikano. Julius – an open source real-time large vocabulary recognition engine. In *Proc. EUROSPEECH*, pp. 1691–1694, 2001.

[6] 李晃伸, 河原達也, 鹿野清宏. モノフォンモデルを用いた状態選択に基づく音響尤度計算の高速化. 情処学研報, 2000-SLP-34-17, 2000.

[7] A.Lee, T.Kawahara, and K.Shikano. Gaussian mixture selection using context-independent HMM. In *Proc. IEEE-ICASSP*, pp. 69–72, 2001.

[8] 李晃伸, 河原達也, 鹿野清宏. 話し言葉の認識のためのデコーダ Julius の改良. 日本音響学会研究発表会講演論文集, 1-3-15, 春季 2001.

[9] 河原達也, 加藤一臣, 南條浩輝, 李晃伸. 話し言葉音声認識のための言語モデルとデコーダの改善. 情処学研報, 2001-SLP-36-3, 2001.

[10] 李晃伸, 河原達也, 堂下修司. 文法カテゴリ対制約を用いた A*探索に基づく 大語彙連続音声認識パーザ. 情報処理学会論文誌, Vol. 40, No. 4, pp. 1374–1382, 1999.

[11] 住吉貴志, 李晃伸, 河原達也. 音声認識エンジン Julius/Julian の API 実装. 情処学研報, 2001-SLP-37-16, 2001.

[12] 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. 情処学研報, 98-SLP-22-17, 1998.

[13] 伊藤彰則, 好田正紀. 単語およびクラス n-gram 作成のためのツールキット. 情処学研報, 2000-SLP-34-32, 2000.

[14] 長友健太郎, 西村竜一, 小松久美子, 黒田由香, 李晃伸, 猿渡洋, 鹿野清宏. 相補的バックオフを用いた言語モデル融合ツールの構築. 情処学研報, 2001-SLP-35-9, 2001.

[15] 西村竜一, 長友健太郎, 小松久美子, 黒田由香, 李晃伸, 猿渡洋, 鹿野清宏. Web からの音声認識用言語モデル自動生成ツールの開発. 情処学研報, 2001-SLP-35-8, 2001.

[16] 板橋秀一, 山本幹雄, 竹沢寿幸, 小林哲則. 日本音響学会新聞記事読み上げ音声コーパスの構築. 日本音響学会研究発表会講演論文集, 3-P-22, 秋季 1997.

[17] 松井知子, 内藤正樹, ハラルド シンガー, 匂坂芳典. 大規模な日本語音声データによる音響モデルの分析. 日本音響学会研究発表会講演論文集, 1-Q-28, 春季 2000.

[18] 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. 電子情報通信学会論文誌, Vol. J83-DII, No. 12, pp. 2517–2525, 2000.

[19] 馬場朗, 芳澤伸一, 山田実一, 李晃伸, 鹿野清宏. 高齢者向け音響モデルによる大語彙連続音声認識の評価. 情処学研報, 2001-SLP-35-3, 2001.

[20] 三村正人, 河原達也. ディクテーションと対話音声認識における音響モデルの差異. 日本音響学会研究発表会講演論文集, 2-8-4, 春季 2000.