2003－SLP－48 (1)
2003／10／17

連続音声認識コンソーシアム2002年度版ソフトウエアの概要

河原達也 住吉貴志 (京大) 李 晃伸 (奈良先端大) 坂野秀樹 (和歌山大)
武田一哉 (名大) 三村正人 (ASTEM)
伊藤克亘 (名大) 伊藤彰則 (東北大) 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、2002年度 (2002年10月-2003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。大語彙連続音声認識エンジン Julius の機能拡張・安定性向上と Windows SAPI 対応を行うとともに、多様な入力環境に対応した音響モデル、及び幅広いカバレージを実現する言語モデルの整備などを行った。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 – 2002 version –

T.Kawahara, T.Sumiyoshi (Kyoto U), A.Lee (NAIST), H.Banno (Wakayama U),
K.Takeda (Nagoya U), M.Mimura (ASTEM),
K.Itou (Nagoya U), A.Ito (Tohoku U), K.Shikano (NAIST)

Abstract

Continuous Speech Recognition Consortium (CSRC) was founded 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 third year (Oct. 2002 - Sep. 2003) is given in this report. The LVCSR (large vocabulary continuous speech recognition) engine Julius has been improved both in functionality and stability, and ported to Windows in compliance with SAPI (Speech API). A variety of acoustic and language models are set up to realize wider coverage of input speech. The software package 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、連続音声認識コンソーシアム (CSRC) が情報処理学会 SLP 研究会のもとで発足し、約50の企業・大学の参加を得て、活動を行ってきた [4][5]。

本稿では、この2002年度 (2002年10月～2003年9月) の成果ソフトウェアの概要を紹介する。

2 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 [6] 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 [7] を利用することにより、より高精度なモデルの構築を行っている。

また音響モデルは、話者層や入力環境が大きく変わると大幅な性能低下を引き起こすので、それらに応じて適切なものを用いる必要がある [8]。そこで、自動車内などの入力環境のためのモデルを構築した。

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 フォーマットである。また、音素体系・表記、及び音響分析や特徴量も、特記していない限り、IPA モデル [2] と同一である。コンソーシアムで提供するモデルは、原則としてすべて性別非依存 (GID) モデルである。

さらに、実環境において高い性能を得るためには、適応を行うことが有効であるので、すべて MLLR 適応が可能なモデルを用意した。

2.1 高精度成人モデル (CSRC モデル)

成人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。

昨年度に提供したモデルに加えて、混合分布数がより大きな (256 の) PTM (Phonetic Tied Mixture) triphone モデル [9] を作成した。これは、高い精度と処理効率の両立を図るものである。

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。

読み上げ音声に対する認識精度を表1に示す。評価データは新聞記事読み上げコーパス (JNAS) から選択された IPA-98-TestSet である。同一の複雑さのモデルで比べると、学習データ量や話者数の増強による認識精度の改善は見られない。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。特に今回作成した256混合の PTM モデルにより IPA モデルの精度を上回ることができた。同モデルは triphone (2000状態16混合) と比べて総ガウス分布数は同程度であるが、ガウス分布枝刈り [9] が効果的に働くので、認識時間は大幅に速い。IPA モデルが主に新聞記事読み上げ文から学習されているのに対して、ATR/BLA が音素バランス文であることを考慮すると、CSRC モデルの方が汎用性が高いと考えられる。

対話音声に対する認識精度を表2に示す。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。ここでは、男性のみのサンプルを用いている。言語モデ

[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。

表 1: JNAS 読み上げ音声に対する単語認識精度 (%)

	PTM	PTM	PTM	triphone
状態数	3000	3000	3000	2000
コードブック	129	129	129	-
混合数	64	128	256	16
IPA モデル	92.7	NA	NA	93.6
CSRC モデル	92.1	93.0	93.6	93.3

表 2: ATR 対話音声に対する単語認識精度 (%)

	PTM	PTM	PTM	triphone
状態数	3000	3000	3000	2000
コードブック	129	129	129	-
混合数	64	128	256	16
IPA モデル	82.1	NA	NA	80.5
CSRC モデル	83.1	84.5	84.8	82.4

ルは単語 3-gram である [10]。CSRC モデルが全般に高い認識精度を得ており、効果が確認された。

特に 128 混合や 256 混合の PTM モデルは、高い認識精度で実時間の大語彙連続音声認識を可能とする。

2.2 自動車内用音響モデル

カーナビなどの自動車内での音声認識インターフェースの需要が大きいが、入力環境が前述の大規模音声コーパスと大きく異なるので、専用の音響モデルが必要となる。

そこで、名古屋大学の CIAIR プロジェクト [11] で構築が進められている車内対話音声データベースを利用して、自動車内用の音響モデルを作成した。音素バランス文を中心に計 700 名による約 22000 文の音声データを使用している。アイドリング中と市街地走行中の両方の状況で、バイザー位置に設置したマイクロフォンから入力されている。SN 比の平均は 20dB である。250Hz 以上に帯域制限した分析条件で、作成したモデルは triphone (2000 状態 32 混合) である。

2.3 ささやき声モデル

携帯電話などによる音声入力においては、ささやき声で行えることが望ましい。ささやき声は通常の音声のような有声振動を伴わないなど音響的特徴が大きく異なるため、専用のモデルが必要である [12]。

そこで、名古屋大学 CIAIR プロジェクト [11] で収集されたささやき声音声データを用いてモデルを構築した。学習データは 80 名による音素バランス文発声 (計 4000 文) であり、作成したモデルは triphone (1000 状態 32 混合) である。

2.4 装着マイク用音響モデル

手軽で SN 比も比較的よいハンズフリー音声入力の形態として、講演等で使用される装着マイクが考えられる。ただし口元との距離があるため、通常の音響モデルでは十分に対処できない。

そこで、名古屋大学 CIAIR プロジェクト [11] において、種々の環境 (防音室、オフィス、車内、路上) 下で装着マイクで入力された音声データを用いてモデルを構築した。学習データは 80 名による音素バランス文発声 (計 4000 文) であり、作成したモデルは triphone (500 状態 32 混合) である。

2.5 話者・環境適応のサポート

このように、ある程度多様な音響モデルを用意したが、実際に使用する際には、適切なモデルを選択した上で、さらに個々のユーザや周囲の環境・入力チャネルに適応することが望ましい。

利用話者・使用環境への事前適応の手法として、MLLR(Maximum Likelihood Linear Regression) 法が近年最も広く使われており、HTK[13] のパッケージにも含まれている。

そこで、すべての音響モデル (HTK フォーマット) に HTK の MLLR 適応が適用できるように必要な回帰情報を埋め込んでいる。

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、この新聞記事モデルを更新するとともに、より日常的な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 Julius 用のバイナリ形式である。

3.1 新聞記事モデルの更新

毎日新聞記事データ 1991 年〜2002 年の 12 年分のテキスト (3.5 億形態素) を用いて、言語モデルを構築した。このテキストから高頻度語を選定して、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。

また今回は、新聞記事コーパスの一部から作成したクラス N-gram のサンプルも含めている。

3.2 Web 上テキストから学習したモデル

新聞記事データよりも、World Wide Web の方がより大規模なテキストを収集することができる。また、Web ページの方がより日常的な言葉や話し言葉が含まれている。

そこで今回、テキストサイズが約 27 億形態素のデータを収集し、言語モデルを構築した。語彙サイズは 6 万 (60K) である。

3.3 言語モデル作成用ツール Palmkit

統計的言語モデルを作成するためのツールである Palmkit[14] [2] の最新バージョン (1.0.28) を収めている。これは、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 を組み合わせて利用することもできる。

4 認識エンジン Julius

大語彙連続音声認識エンジン Julius[15] [16] [3] については、コンソーシアムではネットワーク文法を扱えるパーザ Julian を統合し、Windows 上への移植と SAPI (Speech API) の実装を行ってきた。今年度さらなる機能拡張を行うとともに、安定性の向上に取り組んだ。なお、Unix 版の最新バージョンは Rev.3.4 である。

[2] http://palmkit.sourceforge.net
[3] http://julius.sourceforge.jp

4.1 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、ネットワーク文法のための認識エンジン Julian[17] 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) を用いて 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、さらに SAPI の XML 形式への (半自動) 変換スクリプトもパッケージに含まれている。

4.2 クラス N-gram のサポート

統計的言語モデルにおいて学習データのスパースネスに対処するために、クラス N-gram がしばしば用いられる。特に人名や商品名などの固有名詞のモデル化において有効である。クラス N-gram のクラスを文法の単語カテゴリと対応づけることにより、単語 N-gram と記述文法の中間形として捉えることもできる。また、発音の変形のモデル化もクラス N-gram の枠組みで実装することができる。

そこで、Julius においてクラス N-gram のサポートを行った[4]。言語モデルファイルは通常の単語 N-gram と同様であるが、クラス内の単語生起確率を単語辞書で指定する。その例を図 1 に示す。ここでは人名のクラス「23」が N-gram のエントリとなっており、「岡本」のクラス内確率 (対数スケール) が 2 番目のエントリで与えられている。クラス化していない単語「屋根」については、クラス内確率は 1(対数スケールで 0) となっている。「家」のように従来の単語エントリの形式が同一ファイルに含まれていてもかまわない。Palmkit で生成されるクラス N-gram と若干フォーマットが異なるため、変換スクリプトを用意している。

[4] ただし 2003 年 9 月時点では Unix 版のみの実装である

言語モデルエントリ	@クラス内確率	クラス内単語名	[出力表記]	音素列
(例)				
23	@-2.89719	李:リ:李:23+23	[李]	r i
屋根:ヤネ:屋根:1	@0	屋根:ヤネ:屋根:1	[屋根]	y a n e
家:イエ:家:1			[家]	i e

図 1: Julius におけるクラス N-gram の形式

4.3 認識信頼度の算出

音声対話システムにおいてユーザへの確認を制御したり、音声メディアの自動書き起こしに基づいた検索や要約において取捨選択するための指標として、認識結果の単語毎に信頼度が付与されていることが望ましい。

そこで、Julius においても認識信頼度を算出する機能を追加した。これは、事後確率と同様に 0〜1 の値をとるものである。一般的に認識信頼度は、単語グラフをいったん出力してから事後的に計算することが多い。しかし、Julius で採用しているスタックデコーディングサーチでは単語グラフを求めることが容易でないため、
(1)N-best 単語列候補から求める方法 [18]
(2) スタックデコーディング途上で計算する方法 [19]
の 2 通りを実装した。(1) の方が簡便であるが、多くの候補に対して信頼度 1 になってしまう。(2) は近似計算であるが、すべての候補について実時間に計算できる。ただし、Windows 版では (1) のみを実装している。

4.4 Windows への移植と SAPI の実装

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に適用するには、標準的な API を提供することが重要である。そのため、Windows への移植を行うとともに、マイクロソフト社が策定した Speech API (SAPI-5) [5] の Julius への実装を進めてきた [20]。

Windows XP では、SAPI が標準で含まれているので、Speech SDK をインストールしなくても Julius が動作する。コントロールパネルの「音声認識」のプロパティから、エンジン自体の指定やマイクの設定、そして音響モデル・言語モデル・デコーディングオプションの指定を行うことができる。

SAPI では標準の文法が XML 形式であるので、Julian 用の外部形式・内部表現との相互変換を行う。また、MS-IME を用いて読み付与を行うこともできる。複数の文法 (1 つの N-gram を含む) を扱う場合は、各文法毎にインスタンスを生成し、並列・独立に認識処理が行われる。

現在、Speech SDK 5.1 付属のアプリケーションや MS-IME2002 の音声入力パッド、Office XP などで動作確認を行っている。また、.NET Speech SDK 1.0βに含まれている SALT(Speech Application Language Tags) [6] に対しても、動作確認ができている。

この Windows SAPI 版 Julius は、基本的に Julius Rev.3.2 をベースに作成されており、また Unix 版に比べて、オプションや機能が一部制限される。

またこれとは別に、名古屋大学で SAPI を介さない DLL 版 (Windows, MacOS X で動作) も作成されており、このパッケージも収めている。[7] こちらは、Julius Rev.3.3p4 ベースである。

5 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。今回さらに、Windows に移植し、SAPI 対応になったことにより、アプリケーション開発の利便性が向上した。また、種々の話者や環境に対する音響モデルの整備を行ったことで、やはり多様なアプリケーションへの適用の可能性が広がった。

今後も一層の充実を図るとともに、他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。

[5] http://www.microsoft.com/speech

[6] http://www.saltforum.org

[7] http://www.itakura.nuee.nagoya-u.ac.jp/people/banno/julius.html

2002 年度実行委員リスト

代表：河原達也（京大）
幹事：武田一哉（名大）
伊藤克亘（名大）
李　晃伸（奈良先端大）
山田　篤（ASTEM）
委員：鹿野清宏（奈良先端大）
伊藤彰則（東北大）
宇津呂武仁（京大）
峯松信明（東大）
山本幹雄（筑波大）
小林哲則（早大）
嵯峨山茂樹（東大）
岩野公司（東工大）
坂野秀樹（和歌山大）
北岡教英（豊橋技科大）
山田武志（筑波大）
西浦敬信（和歌山大）
西田昌史（千葉大）
三村正人（ASTEM）

謝辞：本コンソーシアムの設立・運営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、そして活動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. 情報処理学会研究報告, 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] 河原達也, 住吉貴志, 李晃伸, 武田一哉, 三村正人, 伊藤彰則, 伊藤克亘, 鹿野清宏. 連続音声認識コンソーシアム 2000 年度版ソフトウェアの概要と評価. 情報処理学会研究報告, 2001-SLP-38-6, 2001.

[5] 河原達也, 住吉貴志, 李晃伸, 坂野秀樹, 武田一哉, 三村正人, 山田武志, 西浦敬信, 伊藤克亘, 伊藤彰則, 鹿野清宏. 連続音声認識コンソーシアム 2001 年度版ソフトウエアの概要. 情報処理学会研究報告, 2002-SLP-43-3, 2002.

[6] 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. 音講論, 3-P-22, 秋季 1997.

[7] 奥田浩三, 松井知子, 内藤正樹, 匂坂芳典, 中村哲. 大規模日本語音声データベースの構築と評価. 音響誌, Vol. 58, No. 9, pp. 569–578, 2002.

[8] 河原達也. ここまできた音声認識技術. 情報処理, Vol. 41, No. 4, pp. 436–439, 2000.

[9] 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.

[10] 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.

[11] 武田一哉, 板倉文忠. 文部省 COE プログラム統合音響情報研究拠点 (CIAIR). 音響誌, Vol. 56, No. 11, pp. 748–751, 2000.

[12] 伊藤太介, 武田一哉, 板倉文忠. ささやき声の音響分析と音声認識への応用. 電子情報通信学会技術研究報告, SP2001-71, 2001.

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

[14] 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. 情報処理学会研究報告, 2000-SLP-34-32, 2000.

[15] 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.

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

[17] 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.

[18] 駒谷和範, 河原達也. 音声認識結果の信頼度を用いた効率的な確認・誘導を行う対話管理. 情報処理学会論文誌, Vol. 43, No. 10, pp. 3078–3086, 2002.

[19] 李晃伸, 鹿野清宏, 河原達也. 音声認識エンジン Julius における単語事後確率を用いた信頼度算出. 日本音響学会研究発表会講演論文集, 3-6-8, 秋季 2003.

[20] 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. 情報処理学会研究報告, 2001-SLP-37-16, 2001.