# 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

**河原達也　住吉貴志 (京大)　李 晃伸 (奈良先端大)　坂野秀樹 (名大)**
**武田一哉 (名大)　三村正人 (ASTEM)　山田武志 (筑波大)　西浦敬信 (和歌山大)**
**伊藤克亘 (産総研)　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、2001 年度 (2001 年 10 月-2002 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 Windows SAPI 対応を行うとともに、多様な話者層や入力環境に対応した音響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 – 2001 version –

T.Kawahara, T.Sumiyoshi (Kyoto U), A.Lee (NAIST), H.Banno, K.Takeda (Nagoya U)
M.Mimura (ASTEM), T.Yamada (Tsukuba U), T.Nishiura (Wakayama U)
K.Itou (AIST), 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 second year (Oct. 2001 - Sep. 2002) is given in this report. The LVCSR (large vocabulary continuous speech recognition) engine Julius is ported to Windows and compliance with SAPI (Speech API). A variety of acoustic models are set up to cover wider user generations and speech-input environments. 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、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、50 以上の企業・大学の参加を得て、活動を行っている [4]。

本稿では、この 2001 年度 (2001 年 10 月〜2002 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 [5] 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 [6] 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 [7]。そこで、高齢者や子供などの話者層、電話や自動車内などの入力環境のためのモデルを構築した。

[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。

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 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 (128 の) PTM (Phonetic Tied Mixture) triphone モデル [8] 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。IPA モデルが主に新聞記事読み上げ文から学習されたのに対して、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。言語モデルは単語 3-gram である [9]。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。

表 1: JNAS 読み上げ音声に対する単語認識精度 (%)

| | PTM | PTM | triphone | triphone |
|---|---|---|---|---|
| 状態数 | 3000 | 3000 | 2000 | 5000 |
| コードブック | 129 | 129 | - | - |
| 混合数 | 64 | 128 | 16 | 64 |
| IPA モデル | 92.7 | NA | 93.6 | NA |
| CSRC モデル | 92.1 | 92.9 | 93.3 | 95.4 |

表 2: ATR 対話音声に対する単語認識精度 (%)

| | PTM | PTM | triphone | triphone |
|---|---|---|---|---|
| 状態数 | 3000 | 3000 | 2000 | 5000 |
| コードブック | 129 | 129 | - | - |
| 混合数 | 64 | 128 | 16 | 64 |
| IPA モデル | 82.1 | NA | 80.5 | NA |
| CSRC モデル | 83.1 | 84.6 | 82.4 | 83.7 |

## 2.2 小児音声モデル

小児は声道長が短いため、成人用の音響モデルで音声認識するのは困難である。そこで、名古屋大学の CIAIR プロジェクト [10] [2] で構築が進められている子供の声データベースを用いて、小児用の音響モデルを作成した。小学生 400 名が各 100 単語程度を発話したデータを使用している。ただし、十分な学習サンプルを確保できない音素コンテクストがいくつか存在するため、成人女性モデルから MAP 適応学習を行うことにより作成した。作成したモデルは、monophone(16 混合) と PTM triphone(64 混合) である。予備的な動作実験により、成人女性モデルを大きく上回る効果を確認している。

なお 2000 年度には、高齢者音声モデルを配布している。[3]

## 2.3 電話用音響モデル

IVR などの電話を介した音声認識アプリケーションの需要が大きいが、帯域制限や回線上のノイズがあるために、電話音声用の音響モデルが必要である。そこで、京都大学で収集された電話音声データベースを使用して、音響モデルを作成した。このデータベースでは、517 名が音素バランス文各 50 文を発声しており、固定電話と携帯電話の割合は約半数である。収集に用いられた電話インタフェースボードは、Dialogic 社の D/41E-PCI である。作成したモデルは、triphone (2000 状態 16 混合) と PTM triphone (32 混合) である。

なお 2000 年度に、IPA モデルの学習に用いた JNAS コーパスに対して、電話帯域 (300〜3400Hz) で音響分析を行ったデータを用いて学習した音響モデルを配布した。予備的な評価実験により、今回作成したモデルの方が、特に携帯電話からの音声に対して高い認識精度を実現することを確認している。

## 2.4 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も大きいが、やはり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 [10] 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 105 名による 6000 文の音声データを使用している。うち、4000 文がアイドリング中、2000 文が市街地走行中の発声で、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。作成したモデルは、triphone (1000 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[11] のパッケージにも含まれている。

そこで今回、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だ。IPA モデルの代表的な monophone (16 混合)、triphone (2000 状態 16 混合)、PTM triphone (64 混合) のモデルにもこの情報を埋め込んだ。なお、以前の Julius では回帰情報を埋め込んだモデルを直接読み込めないので、認識前に回帰情報を除去するか、今年度版 (Rev.3.3) のものを使用する必要がある。

[2] http://www.ciair.coe.nagoya-u.ac.jp
[3] 各年度毎にパッケージに含まれる内容は異なる。

### 2.6 ハンズフリーツールキット

接話型マイクを用いないようなアプリケーションにおいては、マイクロフォンアレー (2 チャンネルの場合も含む) が有望な入力装置と考えられる。

そこで、マイクロフォンアレーのための信号処理ツールキットを用意した。遅延和アレーと適応型アレー (AMNOR) を実装している。また、TSP を用いたインパルス応答の測定プログラムも含まれている。

## 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〜2001 年 12 月 (1994 年の後半 3ヶ月を除く) の 129ヶ月分のテキストを用いて、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、10 万語彙 (100K) 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。

ただし、標準コンフィグレーションの Julius で扱える語彙サイズの上限は 65535 語なので、–enable-words-int のオプションで作成する必要がある。

### 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可能性が高い。

今回、産総研でテキストサイズが約 24 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 2 万で、種々のカットオフ (20, 40, 80) のモデルを作成した。

### 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[12] [4] の最新バージョン (1.0.27) 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[13] [14] [5] 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、さらに Windows 上への移植を進めている。今年度さらなる機能拡張を行うとともに、利便性の向上を図るために、Windows 上で SAPI (Speech API) の実装を行った。なお、Unix 版の最新バージョンは Rev.3.3 である。

### 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[15] 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。

今回、コンパイルや文法チェックのためのツールもパッケージに統合した。また、SAPI の XML 形式への (半自動) 変換スクリプトも用意した。

[4] http://palmkit.sourceforge.net
[5] http://julius.sourceforge.jp

### 4.2 複数の文法のサポート

文法を用いた認識においては、サブタスク毎に文法を用意しておき、文脈情報などから適当なものに絞り込んだり、切り替えたりできると、認識精度・処理効率の両面で効果的である。例えば、対話システムではプロンプト毎に、フォームフィリングでは項目毎に、ブラウザではページ毎に、受理する文法を切り替えることが考えられる。

そのため、同時に複数の文法で認識を行えるように認識エンジンを拡張し、また動作中にアクティブな文法を切り替えられるようにした。ただし、この実装は、Unix 版と Windows 版で異なる。Unix 版では、アクティブな文法はクライアントから専用の APIで指定し、(サーバによる) 認識は、単一のデコーディングでアクティブな文法ノードのみを展開する方式となっている。これに対して Windows 版では、文法の指定は SAPI に準拠しており、アクティブな文法毎に (複数の) デコーディングを行い、尤度の最も高いものを出力する。なお、(ディクテーション用に)N-gram も記述文法と同時にアクティブにできるが、読み込める N-gram は高々1 個である。

### 4.3 マルチパス音響モデルのサポート

音響モデルについても、複数のモデルを読み込んで、入力に応じて動的に選択できると、多様な話者層や発声環境の変化に対応できると考えられる。例えば、成人モデルと小児音声モデルを並列して用いたり、種々の SN 比に応じた雑音適応モデルを用意することが考えられる。

このような複数の音響モデルを、HMM のトポロジー上でマルチパスとして (事前に) 結合したモデルを扱えるようにした。ただし、この機能は Unix 版のみの実装で、また処理速度上の理由から別のパッケージになっている。

### 4.4 雑音対策

雑音下での音声認識のために、スペクトルサブトラクションを実装した。また、リアルタイムでの認識のために CMN パラメータを保存できる機能も実現した。なおデフォルトでは、前の発話のケプストラム平均を用いている。これらの機能も Unix 版のみの実装である。

### 4.5 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 5.1) [6] の Julius への実装を行った [16]。

今回ほぼ Compliance となった。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 は動作する。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 形式であり、それらと 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・独立な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 1.0$\beta$に含まれている SALT(Speech Application Language Tags) [7] 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。こちらは、Julius Rev.3.1p2 ベースであり、Julian を含まない。

## 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。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と考えられる。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、

[6] http://www.microsoft.com/speech
[7] http://www.saltforum.org

やはり多様なアプリケーションへの適用の可能性を広げたと考えられる。

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。

**2001 年度実行委員リスト**

代表：河原達也（京大）...2002 年 3 月より
幹事：武田一哉（名大）
伊藤克亘（産総研）
山田　篤（ASTEM）
李　晃伸（奈良先端大）
委員：伊藤彰則（東北大）
宇津呂武仁（豊橋技科大）
峯松信明（東大）
山本幹雄（筑波大）
小林哲則（早稲田大）
嵯峨山茂樹（東大）
岩野公司（東工大）
坂野秀樹（名大）
北岡教英（豊橋技科大）
山田武志（筑波大）
西浦敬信（和歌山大）
三村正人（ASTEM）
鹿野清宏（奈良先端大）... 前代表

**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] 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. 情処学研報, 2001-SLP-38-6, 2001.

[5] 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.

[6] 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.

[7] 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.

[8] 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. 信学論, Vol. J83-DII, No. 12, pp. 2517–2525, 2000.

[9] 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. 音講論, 2-8-4, 春季 2000.

[10] 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.

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

[12] 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.

[13] 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, Vol. J82-DII, No. 1, pp. 1–9, 1999.

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

[15] 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. 情処学論, Vol. 40, No. 4, pp. 1374–1382, 1999.

[16] 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.