日本語ディクテーション基本ソフトウエア

デモ 7

武田一哉 (名大)　河原達也 (京大)　伊藤克亘 (電総研)　鹿野清宏 (奈良先端大)

http://www.itakura.nuee.nagoya-u.ac.jp/~takeda/IPA/

1　はじめに

「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。

これらは統合することにより日本語ディクテーションシステムとして動作するが、個々のコンポーネントを用いて種々の要素技術の研究やアプリケーションの開発を行なうことが可能になっている。ここでは、この 97 年度版の紹介と音声認識システムの構成例のデモンストレーションを行う。本ツールキットは、無償で一般に公開されている。

2　モデルとプログラムの仕様

ツールキットのモジュールと使用したデータベースの関連を図 1 に示す。言語モデル及び単語辞書は、毎日新聞の記事データベースから構築した。音響モデルは、日本音響学会の音声データベースから学習した。認識エンジンは、両者とインタフェースをとっており、それらの相互評価に用いることができる。

2.1　音響モデル

音響モデルは、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。

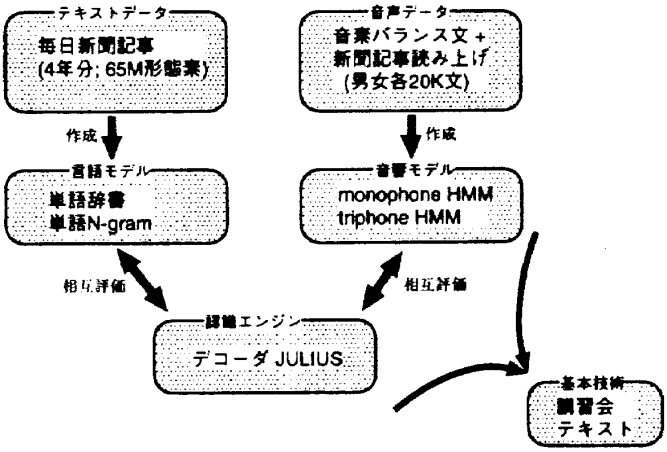

図 1: 大語彙連続音声認識のプラットフォーム

2.2　単語辞書

単語辞書も、HTK のフォーマットで提供される。

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。種々の語彙サイズにおけるカバレージを表 2 に示す。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 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: 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%

Japanese Dictation ToolKit – 1997 version –
K.Takeda, T.Kawahara, K.Itou, and K.Shikano

本ソフトウェアは、情報処理振興事業協会 (IPA) が実施した独創的情報技術育成事業の研究成果である。

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

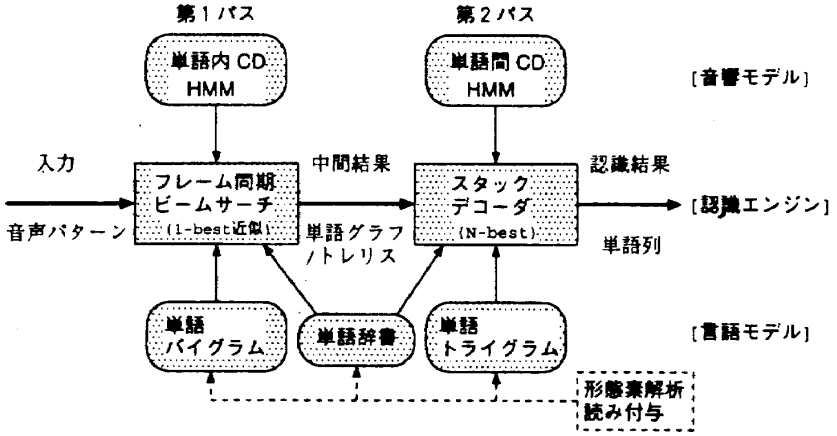

図 2: 日本語ディクテーションシステムの構成

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。

2.4 認識エンジン

デコーダ JULIUS 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。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。

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構成することができる。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。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。表 3 に、典型的なシステムの構成を 3 つ挙げる。

なお詳細な性能評価については、文献 [1] を参照されたい。

表 3: 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

4 おわりに

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。したがって、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だけでなく、異なった機関で開発されたモジュールの交換・統合や評価を行うことが可能である。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。ただし、言語モデルの読み込みを含めて、64MB 程度のメモリを要する。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。

謝辞：開発に協力頂いたプロジェクトメンバーの方々ならびに関係各位に感謝します。

参考文献

[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エア (97 年度版) の性能評価. 情報処理学会研究報告, 98-SLP-21-10, 1998.