音声言語情報処理 18－1
（1997. 10. 24）

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 －

河原達也　李晃伸 (京大)　伊藤克亘 (電総研)　小林哲則 (早稲田大)
伊藤彰則 (山形大)　宇津呂武仁 (奈良先端大)　清水徹 (KDD)　田本真詞 (NTT)
荒井和博 (NTT)　峯松信明 (豊橋技科大)　山本幹雄 (筑波大)
竹沢寿幸 (ATR)　武田一哉 (名大)　松岡達雄 (NTT)　鹿野清宏 (奈良先端大)

あらまし
大語彙連続音声認識研究の推進のためには、データベースと共に、標準 (ベースライン) となるソフトウェア (言語モデル・音響モデル・認識プログラム) が必要であり、著者らはその基盤整備を進めている。本稿では、認識プログラムについて、その仕様 (案) と基本的なアルゴリズムを説明する。

Common Platform of Japanese Large Vocabulary Continuous Speech Recognition Research
– Speech Recognizer Design –

Tatsuya Kawahara, Akinobu Lee (Kyoto Univ.), Katsunobu Itou (ETL),
Tetsunori Kobayashi (Waseda Univ.), Akinori Ito (Yamagata Univ.),
Takehito Utsuro (NAIST), Toru Shimizu (KDD), Masafumi Tamoto (NTT),
Kazuhiro Arai (NTT), Nobuaki Minematsu (Toyohashi Univ. of Tech.),
Mikio Yamamoto (Tsukuba Univ.), Toshiyuki Takezawa (ATR),
Kazuya Takeda (Nagoya Univ.), Tatsuo Matsuoka (NTT), Kiyohiro Shikano (NAIST)

Abstract
For Japanese large vocabulary continuous speech recognition (LVCSR) research, we are developing standard baseline software repository that includes language models, acoustic models and recognition engines. In this report, specifications and algorithms of the speech recognizer currently designed are described.

1 必要性

大語彙連続音声認識 (LVCSR) すなわち任意語彙音声のディクテーションは、音声認識の研究・開発者にとってはある意味で究極のテーマである。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様々な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の基盤となるであろう。

すなわち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プログラム (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。

このように広範囲な分野をすべてカバーし、大規模なシステムを構築するのは (特に大学のような) 単独の研究機関では容易ではなく、個別要素に関する研究を進めてもシステム全体としての性能を示しにくいのが実情である。また多くの機関ですべての分野に対して投資するのは効率がよいといえず、結果としていずれも不十分なシステムしかできない可能性も高い。また、音響モデル構築には大規模な音声データベース、言語モデル構築には大規模なテキストデータベースが必要であり、これらを単独の研究機関で収集していくのも非常に困難である。

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、標準的なソフトウェアを整備することが必要であると考えられる。すなわち、標準となるデータベースに加えて、さらに標準 (ベースライン) となる言語モデル・音響モデル・認識プログラムが整備されれば、個々の要素に集中して研究を進めても、正当な評価を行うことができ、さらにはシステム全体の性能の改善にもつながる。

特に大語彙連続音声認識においては、タスク設定や評価尺度がかなり明確に定義できるため、研究者間で種々の手法やシステムの比較することが容易である。

米国の DARPA[1] や欧州の SQALE[2] のプロジェクトにおいては、データベースと評価基準を共通にして、研究機関どうしで競争と協調を進めていくことで、データや知見が集積され、全体としての認識精度と研究水準の着実な向上をもたらしている。

我が国においても、1995 年に情報処理学会に大語彙連続音声認識研究用データベースワーキンググループ (主査：鹿野) が設置され、日本音響学会の連続音声データベース調査委員会との協力の下、データベースの整備が進められてきた。その結果、毎日新聞記事

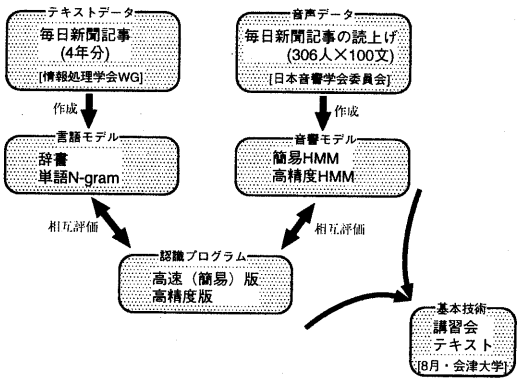

図 1: 大語彙連続音声認識のための研究基盤の整備

のテキスト [3] と読上げ音声のデータベース [4] が提供されるに至っている。

大語彙連続音声認識研究の推進のためには、引き続き標準モデルと標準プログラムが必要であるが、1997 年より情報処理振興事業協会 (IPA)「独創的情報技術育成事業」の支援を得て、著者ら (代表者：鹿野) を中心にその整備を進めている。この成果は広く一般に無償で公開される予定である。これらの一連の活動を図 1に示す。

本稿では、この中で認識プログラムに関して、その仕様 (案) と基本的なアルゴリズムを説明する。音響モデル及び言語モデルに関しては、本冊子に個別の報告があるので参照されたい。

2 認識プログラムの仕様 (案)

大語彙連続音声認識プログラムは、その仕様に関しては著者らで議論を行い、早稲田大学と電総研の協力を得ながら、京都大学で主要部分の開発を行っている。京都大学におけるコード名は JULIUS である。

入力は連続音声である。抽出する音響特徴量は音響モデルで指定する。MFCC (Mel-Frequency Cepstral Coefficients) ベクトルとパワー及びそれらの時間差分を基本とする。CMN (Cepstral Mean Normalization) を実行するか否かも音響モデルで指定する。単語間にポーズが入ってもよいが、ポーズの出現に関しては言語モデルで指定する。つまり、句読点や記号などをポーズに書き換えるようにして、それらの出現確率は N-gram の枠組みでモデル化する。ただし句点相当の長いポーズで、認識プログラムは入力を打ち切って処理する。

出力は最尤単語列である。第2パスまで実行する高精度版ではN-best候補を求める。

処理に要する時間は、ハイエンドの計算機で実時間の10倍以下とする。第1パスしか実行しない高速版では、ある程度の語彙であれば実時間に近い応答ができるように目指す。処理に要するメモリ量は、ハイエンドの計算機で標準的なメモリサイズにおさまるようにする。ただし、処理時間とメモリ量に関しては、研究の円滑な推進に必要な水準を満たすことが基本である。また、認識プログラムにおいては、語彙が小さければそれなりに処理時間とメモリ量が少なくてすむスケーラビリティも重要であると考えている。

以下に音声認識に必要な他の要素とのインタフェースについて述べる。

- 音響モデル (HTK フォーマット)

  HMMを基本とし、HTKフォーマット[5]に互換性を持たせる。

  混合連続分布HMM、及びtied-mixture HMMを扱えるようにする。ただしHTKフォーマットでは、HMMの各状態から必要な分布を参照するようになっており、両者は統一的に扱われている。

  認識プログラムでは、音素環境独立 (CI) モデルだけでなく、音素環境依存 (CD) モデルを扱えるようにするが、第1パス (高速版) では単語内の依存性のみを考慮し、単語間については第2パス (高精度版) で処理する。これもtriphoneまでとする。

- 辞書 (HTK フォーマット)

  単語と対応する音素表記の辞書は、HTKフォーマットに準拠する。

  単語のローマ字表記から音素表記への変換規則 (プログラム) は、音響モデル開発者の責任で用意する。ローマ字セットは日本音響学会の音声データベースで使用されたものを標準とする。

- 言語モデル (CMU-TK | HTK フォーマット)

  単語N-gramを基本とし、CMU-TK[6]フォーマットに互換性を持たせる。これはHTKフォーマットとほぼ等価である。高速化のため、コンパイルした形式にすることも検討している。

  第1パス (高速版) ではbigramのみを用い、第2パス (高精度版) でtrigramを適用する。

  辞書との一貫性を保証するため、単語のローマ字表記は言語モデル開発者の責任で行う。これには、格助詞「は」「へ」などの適正な処理なども含まれる。また、ポーズの出現に関しても、句読点や記号などを利用して言語モデルで表現する。

認識プログラムには高速版と高精度版の2つを用意する。高速版は実時間性を指向し、必要なメモリ量も高精度版の半分以下ですむ。高精度版は認識精度を優先するため、単語間の音素環境依存 (CD) モデルとtrigramを利用する。高精度版は2パスアルゴリズムとなっており、基本的にその第1パスが高速版に相当する。両者の比較を表1に示す。

3 認識システムの構成

認識プログラムを中心とする認識システムの構成と処理の流れを図2に示す。

認識処理は、基本的に2パスから構成される。ただし、オプションとして予備選択 (ファーストマッチ) を導入する。これは、簡易音素モデルを適用して、スコアの低いノードを照合の対象から除外 (枝刈り) するものである。第1パスでは、単語内の音素環境依存モデルと単語bigramを用いて、フレーム同期ビームサーチを行う。高速版では、この結果求められる最尤単語系列を認識結果とする。高精度版では、中間結果として、各フレーム毎に (最終的に) 残った単語候補についてその尤度と始端フレーム・先行単語のリストを、単語グラフもしくはトレリスを縮退した形式で保存する。第2パスでは、この中間結果として得られた仮説に対して、高精度なモデルを適用して再評価及び再探索を行う。すなわち、単語間の音素環境依存モデルと単語trigramを適用しながら、スタックデコーディングサーチを行う。

4 アルゴリズム

認識プログラム (JULIUS) で採用するアルゴリズムについて説明する。

4.1 フレーム同期サーチ (第1パス)

時間フレームに同期して探索を進める。HMMのトレリス上の探索と等価である。アルゴリズムの設計や実装が比較的容易である。仮説間の尤度 (評価値) もそのまま比較できる。ただし、1フレーム毎の情報

表 1: 認識プログラムの高速版と高精度版

	処理	音響モデル	言語モデル	出力	特徴
高速版	1 パス	単語内 CD-HMM	単語 bigram	1-best	実時間指向・省メモリ
高精度版	2 パス	単語間 CD-HMM	単語 trigram	N-best	認識精度優先

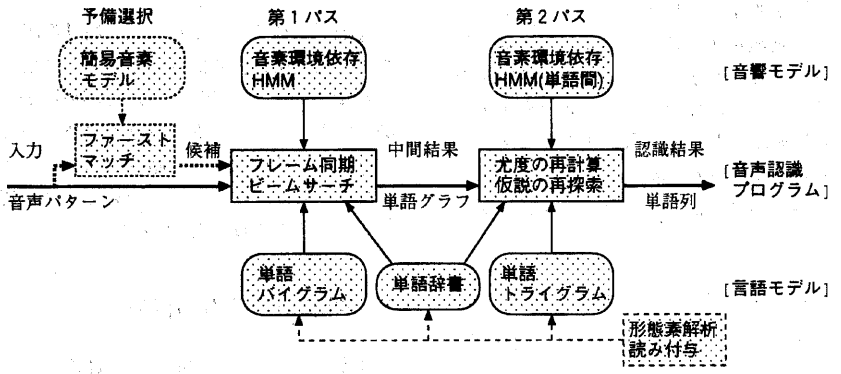

図 2: 認識システムの構成と処理の流れ

に基づいて枝刈りを行うので、安定な照合を行えず、局所的な変動の影響を受けやすい。このため、ファーストマッチによる数フレームの先読みを導入したり、マルチパスで総合的な認識を行うのが望ましい。

4.2 木構造化辞書

N-gram 言語モデルの最も単純な実装は、確率有限状態オートマトンとみなして、静的な単語ネットワークを構成する方法である。この場合、単語間の遷移に言語モデルの確率を付与する。bigram を使用する場合、ネットワークのノード数 (HMM の状態数) は、語彙サイズに対してほぼ線形に増加する。また、単語間のアーク (遷移数) は語彙サイズの 2 乗になる。[1]

単語辞書に関しては、プレフィックスを共有する木構造化を行うことにより、状態数を削減できる。この効果は、語彙が大きくなるにつれて顕著になり、大語彙連続音声認識では不可欠となる。しかし、木構造化辞書においては、単語の終端 (= 木の葉ノード) の方に達しないと単語を同定できないので、N-gram の確率を静的に埋め込むことはできない。そこで、単語の終端に達する毎に N-gram の確率を動的に加える。ただし、直前の単語 (履歴) をいちいちトレースバックすることなく同定できるように、最尤経路を与える直前の単語 (履歴) の情報を伝播 (コピー) しておく必要がある。なお、始終端を特定したい際には、単語境界のフレームもコピーする。また単語対近似 (N-best 探索・ワードグラフ作成) を導入する際には、直前の単語毎に尤度と境界をコピーする。

この木構造化辞書と N-gram の実装に関しても、あくまでも静的な木を 1 つ用意しておいて、すべてのノードにバックポインタ (単語履歴) をコピーする方法と、単語履歴毎に木を動的に生成する方法とがある [7]。ここでは前者を採用する。

4.3 bigram 確率の factoring

木構造化辞書と N-gram を単純に組合せると、単語の終端に到達するまで、言語モデルの確率が与えられないことになる。言語的制約が適用されるタイミングが遅れると、枝刈りの際の評価値に反映されないために、最適な解が単語の途中でビーム幅からあふれる可能性がある。語彙が大きくなるにつれて、この問題は顕在化する。

そこで、言語モデルの確率を分解して、木の途中のノードにも割り振るようにする。この factoring にはいくつかの方法があるが、プレフィックスを共有する単語に対する N-gram 確率の最大値を割り振り、累積値を順に精算していく方法が一般的である [7][8]。

[1] trigram を使用する場合の事態はもっと深刻である。

4.4 2パスサーチ

1パスの処理において、高精度のモデルを適用すると多くの仮説を扱うことになり、全体としてかえって処理時間が長くなりかねない。特にtrigramを適用するには、2単語履歴を考慮する必要があるので、実装が複雑になる。

そこで、効率よく高精度のモデルを適用するために、マルチパス探索[9][10]を採用する。第1パスで、ある程度の精度の音響モデル・言語モデルを用いて入力(ポーズまで)を完全に処理して、この中間結果を基に、第2パスにおける探索空間を限定すると共に、先読み情報(ヒューリスティック)として利用する。ここでは認識精度を優先して、最初から文脈依存モデルを使用する[11]。ただし、単語間の結合に関しては処理が煩雑になるため最初は考慮しない。言語モデルは、やはり処理の簡便性からbigramを最初に用いて、第2パスでtrigramを適用する。

第2パスでは、候補もかなりしぼられている上に、完全な先読みもできるので、音素・単語単位で探索を行う。これは単語の木の上の探索と等価であり、安定した照合に基づいて枝刈りを行うことができる。仮説の評価は、第1パスで計算された入力全体の尤度を反映(先読み)して行う[12]。ここでは、効率のよいbest-firstなスタックデコーディングサーチを実現する。

第1パスと第2パスの間の中間表現の形式(インタフェース)としては、以下が考えられる。

- N-best候補[13][14]

  単語列(文)の複数(N-best)候補を受け渡す。単語間の音素環境依存モデルやtrigramによる再評価の実装が容易である。しかし、入力(文長)が長い場合は、かなり多数の候補を求めても一単語のみ異なる類似候補しか得られないので、結果として効率が悪くなる。

- 単語ラティス / 単語グラフ[15][9][16]

  単語の尤度と始端・終端の集合を求める。そのグラフをたどることにより、結果として多数のN-best候補が求められるので、効率よい表現といえる。単語の尤度や境界はそれ以前の単語の影響を受けるので、単語履歴毎に異なった候補を求めるべきであるが、そうすると候補が膨大となるので、直前の単語のみに依存させる単語対近似[13]を仮定することが多い。

- トレリス[12][17]

  単語毎の尤度や区間(境界)を決定的に求めるのではなく、単語の終端の状態のトレリス(=尤度と対応する始端)を保存する。次のパスにおいてもトレリス接続の計算が必要になるが、(より高精度な音響モデルによる)仮説の正確な再評価ができる。そのままでは候補の絞り込みを直接的に行えないので、ビームに残った単語ノードを逆引きできるようにする。

ここでは、単語グラフ、あるいはそれを一般化したトレリスの表現を採用する。

4.5 ファーストマッチ

簡易な音響モデル(HMMを音素環境独立な1状態に縮退)の情報を使用して、1音素分に相当する数フレームを先読み照合する。これにより、辞書中から音響的に有望な候補を絞り込む(予備選択)ことができる[18][19]。精度も深さも限られた簡易な先読みに基づいて枝刈りを行う、高速性を指向した手法である。

4.6 言語モデルのデータ構造

大語彙のN-gramは膨大な記憶量を必要とする。参照する回数が多いので、記憶量を削減すると共に、参照の局所性を反映したコンパクトなデータ構造を採用する必要がある。具体的には、同一の単語履歴に対するtrigram(及びbigram)は連続してシーケンシャルにアロケートする、またこれらは単語履歴に対応する2単語のbigram(及び1単語のunigram)からリンクすることにより、効率よく表現する。

5 開発計画

認識プログラムの開発計画を図3に示す。上述のアルゴリズムの実装は、ファーストマッチを除いてほぼ終わっており、1997年度は語彙サイズ5000のシステムにおいて、基本性能の確認を行う。1998年度には、これをさらに大規模なシステムに拡張し、高精度化を図る。1999年度には、省メモリ化と高速化を図り、パソコン上で十分動作するように実装する。

本プログラムは、単語辞書を含む言語モデル・音響モデルなどと共に、順次無償で公開する予定である。仕様やアルゴリズムに対するコメントは歓迎する。

図 3: 認識プログラムの開発計画

謝辞

本研究は、情報処理学会音声言語情報処理研究会の大語彙連続音声認識研究用データベースWGの活動の一環として行われた。また本研究は、情報処理振興事業協会「独創的情報技術育成事業」の一環として行われている。関係各位のご支援に感謝致します。

参考文献

[1] D.B.Paul and J.M.Baker. The design for the Wall Street Journal-based CSR corpus. In *Proc. DARPA Speech & Natural Language Workshop*, pp. 357–361, 1992.

[2] H.J.M.Steeneken and V.Leeuwen. Multilingual assessment of speaker independent large vocabulary speech recognition system: the SQALE-project. In *Proc. EUROSPEECH*, 1995.

[3] 伊藤克亘, 松岡達雄, 竹沢寿幸, 武田一哉, 鹿野清宏. 大語彙連続音声認識研究のためのテキストデータ処理. 音講論, 3-3-10, 秋季 1996.

[4] 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.

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

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

[7] J.J.Odel, V.Valtchev, P.C.Woodland, and S.J.Young. A one pass decoder design for large vocabulary recognition. In *Proc. ARPA Human Language Technology Workshop*, pp. 405–410, 1994.

[8] G.Antoniol, F.Brugnara, M.Cettolo, and M.Federico. Language model representations for beam-search decoding. In *Proc. IEEE-ICASSP*, pp. 588–591, 1995.

[9] H.Murveit, J.Butzberger, V.Digalakis, and M.Weintraub. Large-vocabulary dictation using SRI's DECIPHER speech recognition system : Progressive search techniques. In *Proc. IEEE-ICASSP*, volume 2, pp. 319–322, 1993.

[10] L.Nguyen, R.Schwartz, Y.Zhao, and G.Zavaliagkos. Is N-best dead ? In *Proc. ARPA Human Language Technology Workshop*, pp. 411–414, 1994.

[11] S.J.Young. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing magazine*, Vol. 13, No. 5, pp. 45–57, 1996.

[12] F.K.Soong and E.F.Huang. A tree-trellis based fast search for finding the N best sentence hypotheses in continuous speech recognition. In *Proc. IEEE-ICASSP*, pp. 705–708, 1991.

[13] R.Schwartz and S.Austin. A comparison of several approximate algorithms for finding multiple (N-best) sentence hypotheses. In *Proc. IEEE-ICASSP*, pp. 701–704, 1991.

[14] 吉田航太郎, 松岡達雄, 大附克年, 古井貞熙. 単語 trigram を用いた大語彙連続音声認識. 情報処理学会研究報告, 96-SLP-14-14, 1996.

[15] H.Ney and X.Aubert. A word graph algorithm for large vocabulary continuous speech recognition. In *Proc. ICSLP*, pp. 1355–1358, 1994.

[16] L.Nguyen, R.Schwartz, F.Kubala, and P.Placeway. Search algorithms for software-only real-time recognition with very large vocabularies. In *Proc. ARPA Human Language Technology Workshop*, pp. 91–95, 1993.

[17] 李晃伸, 河原達也, 堂下修司. 単語 N-gram と段階的探索を用いた大語彙連続音声認識. 情報処理学会研究報告, 97-SLP-16-4, 1997.

[18] L.R.Bahl, S.V.de Gennaro, P.S.Gopalakrishman, and R.L.Mercer. A fast approximate acoustic match for large vocabulary speech recognition. *IEEE Trans. Speech & Audio Process.*, Vol. 1, No. 1, pp. 59–67, 1993.

[19] 西村雅史, 伊東伸泰, 山崎一孝, 荻野紫穂. 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. 情報処理学会研究報告, 97-SLP-15-5, 1997.