音声言語情報処理 34－14
(2000. 12. 21)

[サーベイ] 音声認識における探索

今井 亨

ＮＨＫ放送技術研究所

〒157-8510 東京都世田谷区砧1-10-11

imai@strl.nhk.or.jp

あらまし 大語彙連続音声認識技術は、統計的な音響モデルと言語モデルの高精度化、そして探索の効率化によって近年著しく進歩した。計算機処理速度の向上はより複雑な処理を可能にし、さらなる高精度化と高速化を目指して広く研究が行われている。本稿では、音声認識の探索に関する今年１年の研究をサーベイし、最近の探索方式や効率化に対する取り組みなど、その現状を概観する。

キーワード 音声認識、探索

[Survey] Search in Speech Recognition

Toru Imai

NHK (Japan Broadcasting Corp.) Sci. & Tech. Res. Labs.

1-10-11 Kinuta, Setagaya, Tokyo 157-8510 JAPAN

imai@strl.nhk.or.jp

Abstract Large-vocabulary continuous speech recognition techniques have greatly advanced in recent years due to the use of statistical acoustic and language models and efficient search methods. With improved CPU speed allowing increased processing complexity, research into both faster and more accurate search methods has been widely pursued. This paper surveys recent research on search techniques and efficient approaches in speech recognition.

key words speech recognition, search

1．まえがき

大語彙連続音声認識技術は、統計的な音響モデルと言語モデルの高精度化、そして探索の効率化によって近年著しく進歩した。計算機処理速度の向上はより複雑な処理を可能にし、さらなる高精度化と高速化を目指して広く研究が行われている。本稿では、音声認識の探索に関する今年１年の研究をサーベイし、その現状を概観する。

2．探索方式

2.1 2パス・デコーダ

処理の効率化のために、第１パスで粗い探索を行い、中間結果のN-best文や単語ラティス（あるいはグラフ）を第２パス以降で高精度に探索するマルチパス・デコーダが使われることが多い。第２パスの探索効率化では、単語ラティス上で一番短い仮説から展開してスコアによる枝刈りを行う、時間同期型に近い時間非同期型探索手法が提案されている[1]。また、第１パス実行中に逐次的に第２パスを実行することで、発話終了を待たずに認識結果を早期確定する方式も提案されている[2]。

2.2 時間同期型１パス・デコーダ

一方、始めから高精度なモデルで探索を行う時間同期型１パス・デコーダも広く検討されるようになり、２パス・デコーダとほぼ同等の処理速度が報告されている[3][4]。具体的には、バイグラム・ファクタリングを用いた１パス・トライグラム・サーチ[3][5]や、高次N-gram下での効率的な単語グラフ生成法[6]、単語終端のクロスワード用ノードを探索中に動的に生成する方法[7]などが検討されている。

2.3 時間非同期型スタック・デコーダ

時間非同期型探索であるスタック・デコーダについては、仮説の展開の仕方についての様々な検討[8][9]や、履歴の異なる仮説でDetailed Matchを共有して計算量を削減する報告[10]がある。

3．探索の効率化

3.1 ネットワーク構造

大語彙の音声認識では、処理の高速化を図るために、単語辞書を木構造の音素ネットワークに展開して探索空間を圧縮することが一般的である。単語に共通の接頭辞だけでなく接尾辞もマージするDAWG(Directed Acyclic Word Graph)構造は、人名タスクにおいてノード数削減の効果が報告されている[11]。また、単語辞書のネットワークを直前単語ごとのバイグラム接続用と全単語共通のバックオフ接続用(1-best近似併用)に分け、高速化を図る報告もある[12]。

3.2 ファクタリング

単語辞書を木構造で展開した場合、音素を共有する単語の中で最大のバイグラムを枝刈り用スコアとして使うなど、言語モデルのファクタリング（先読みともいう）が一般的に行われている。

ファクタリングのための計算量とメモリ量の削減の研究では、語頭の数音素のみ最大バイグラムをテーブル化しておき、それ以外は必要になった時に計算する方法[2]、品詞バイグラムやユニグラムを併用する方法[13][14]、単語エントリーを発音でソートして共有単語リストのメモリ量を削減する方法[15]、１状態モノフォンモデルによる音響スコアと最大バイグラムを併用して先読みスコアとする方法[5]、音響先読みと最大バイグラムによる先読みを音素内で切り替えて、最大ユニグラムのランク付けで枝刈りをする方法[16]など、多くの報告がある。

3.3 マージ

ネットワーク上での探索の効率化の１つに仮説のマージがあり、直前の単語が同じ仮説を各ノードでマージする単語対近似（あるいは単語依存N-best探索）と、直前の単語に関わらず仮説を各ノードでマージする1-best近似がある。これらの中間的な近似として音素履歴近似が提案されており、真のバイグラム適用時に先行単語を再選択することで近似誤差の解消を図る報告がある[17]。

3.4 枝刈り

仮説を絞り込むために、仮説の数やスコアによるビーム幅を設定することが一般的である。固定のビーム幅ではなく、アクティブなノードの最大スコアと平均スコアから動的にビーム幅を変更する方式[18]や、韻律境界を利用して動的にビーム幅を変更する方式[19]、状態継続時間長に基づく枝刈り[20]、音響スコアの事前分布を利用して、音声終端でスコアが低くなると予想される候補を各フレームで枝刈りする方式

[21]が提案されている。

4．音響尤度計算の高速化

大語彙連続音声認識における音響尤度の計算量は、探索処理時間全体の多くを占めることがあり、効率よく計算量を削減することが重要である。音響尤度計算法の改良による探索高速化の研究では、状態位置ごとにコードブックを共有するPTM(Phonetic Tied-Mixture)モデルにガウス分布計算の足切りを導入した方法[22]、モノフォンモデルで高い尤度を与えた状態のみトライフォンモデルで尤度を再計算する方法[23]、尤度最大のガウス分布の選択と二分木探索を組み合わせる方法[24]、ガウス分布をサブスペース（実際には各次元）でクラスタリングしておき、参照されるクラスタの組み合わせを有効に利用する方法[25]、分布をVQでクラスタリングしておいて厳密な尤度計算と近似的な計算を切り替える方法[18][26]、プロセッサのSIMD (Single Instruction Multiple Data) 命令を利用して音響尤度計算に必要なl_2ノルムを3倍高速に計算する方法[27]などがある。

数種類の音響特徴量を組み合わせてデコードする際、各フレームでスコア最大の音響特徴量を採用するランク・ベース方式が提案されている[28]。全特徴量を同等に扱った場合と同等の性能だが、別々にデコードして結果を統合するROVER[29]よりは高速で認識率がわずかに改善されることが報告されている。

5．信頼度によるスコアリング

音声認識の確率的アプローチでは、音声Xが与えられた時の単語列Wの事後確率$P(W|X)$がスコアに用いられる。一般には、これをViterbiアルゴリズムによる最適パスで近似して求め、異なるセグメンテーションや異なるコンテキストは考慮しない。一方、$P(W|X)$をW中の各単語w_nの事後確率$P(w_n|X)$の積で近似してリスコアリングに用いる方式が提案されている[30][31]。単語事後確率は、各単語w_nの前後で単語列W以外のコンテキストや異なるセグメンテーションも考慮する点が従来のスコアとは異なり、認識結果における各単語の信頼度とみなすことができる。単語事後確率は、最初の探索で得られた単語ラティス（あるいはグラフ）を同一時刻の同一単語でマージするなどして再構成しておき、forward-backwardアルゴリズムによって求められる。単語列Wへの依存性を改善した報告[32]もあるが、いずれも従来のスコアリングに比べて大きな改善には至っていない。信頼度を認識結果の改善に利用する試みは、他にも単語グラフにおける前後の単語への接続数に基づいた方法も提案されている[33]。

6．その他

音響特徴量や言語以外の情報も探索に利用する研究では、話者交替を言語モデルで表現しつつ話者認識と単語認識を同時に行う探索法[34]や、音源方向を推定しつつ複数話者の音声を同時に認識する探索法[35]などがある。

固定長の短い音声区間を連続して探索することにより、音声の端点検出を不要とした方式が提案されており[36]、特に音声切り出しが不安定な雑音環境下での有効性に期待できる。

探索時には、音響スコアに対する言語スコアの重み係数や単語挿入ペナルティを設定する必要があるが、このパラメータの最適化について検討がなされている[37]。

目新しい研究では、言語処理で広く使われるようになってきた有限状態変換器(FST; Finite-State Transducer)を適用した探索方式がある[38][39]。これは、HMMとバイグラムと発音辞書をすべて重み付きFSTの合成で表現しておき、FST上でフレーム同期Viterbiサーチを実現するものである。HTKに比べて25倍高速にデコードできることや[38]、発音辞書の自動生成に適していることが示されている[39]。

7．むすび

本稿では、今年発表された音声認識の研究報告の中から探索に関するものを取り上げ、その現状を概観した。今後、無駄な処理を効率よく削減する高速化の研究と、より複雑な処理を導入可能にする高精度化の研究が並行して発展していくと期待される。

文　献

[1] A. Ogawa, Y. Noda, and S. Matsunaga, "Novel two-pass search strategy using time-asynchronous shortest-first second-pass beam search," Proc. ICSLP, vol. IV, pp. 290-293 (2000.10).

[2] T. Imai, A. Kobayashi, S. Sato, H. Tanaka, and A. Ando, "Progressive 2-pass decoder for real-time broadcast news captioning," Proc. ICASSP, vol. III, pp. 15559-1562 (2000.6).

[3] 本間真一、今井亨、安藤彰男、“ニュース音声認識のための1パストライグラムデコーダの検討、” 音講論集、2-8-9 (2000.3).
[4] J.-L. Gauvain and L. Lamel, "Fast decoding for indexation of broadcast data," Proc. ICSLP, vol. III, pp. 794-797 (2000.10).
[5] X.L. Aubert and R. Blasig, "Combined acoustic and linguistic look-ahead for one-pass time-synchronous decoding," Proc. ICSLP, vol. III, pp. 802-805 (2000.10).
[6] C. Neukirchen, X. Aubert, and H. Dolfing, "Extending the generation of word graphs for a cross-word M-gram decoder," Proc. ICSLP, vol. IV, pp. 302-305 (2000.10).
[7] A. Sixtus, S. Molau, S. Kanthak, R. Schlüter, and H. Ney, "Recent improvements of the RWTH large vocabulary speech recognition system on spontaneous speech," Proc. ICASSP, vol. III, pp. 1671-1674 (2000.6).
[8] D. Willett, C. Neukirchen, and G. Rigoll, "DUCODER – The Duisburg university LVCSR stackdecoder," Proc. ICASSP , vol. III, pp. 1555-1558 (2000.6).
[9] 小早川健、世木寛之、今井亨、本間真一、田中英輝、安藤彰男、“連続音声認識のためのスタックデコーダの作成、” 音講論集、1-5-16 (2000.9).
[10] M. Novak and M. Picheny, "Speed improvement of the tree-based time asynchronous search," Proc. ICSLP, vol. IV, pp. 334-337 (2000.10).
[11] K. Georgila, K. Sgarbas, N. Fakotakis, and G. Kokkinakis, "Fast very large vocabulary recognition based on compact DAWG-structured language models," Proc. ICSLP, vol. II, pp. 987-990 (2000.10).
[12] J. Ogata and Y. Ariki, "An efficient lexical tree search for large vocabulary continuous speech recognition," Proc. ICSLP, vol. II, pp. 967-970 (2000.10).
[13] 小窪浩明、山本博史、“木構造ネットワークサーチにおけるbigram factoringテーブルの削減、” 音講論集、1-5-18 (2000.9).
[14] W.N. Choi, Y.W. Wong, T. Lee, and P.C. Ching, "Lexical tree decoding with a class-based language model for Chinese speech recognition," Proc. ICSLP, vol. I, pp. 174-177 (2000.10).
[15] H. Yamamoto, T. Fukada, and Y. Komori, "Effective lexical tree search for large vocabulary continuous speech recognition," Proc. ICSLP, vol. IV, pp. 322-325 (2000.10).
[16] J. Wu and F. Zheng, "Reducing time-synchronous beam search effort using stage based look-ahead and language model rank based pruning," Proc. ICSLP, vol. IV, pp. 262-265 (2000.10).
[17] 堀貴明、野田喜昭、松永昭一、“音素履歴近似とDelayed Bigramを用いた効率的探索手法の検討、” 音講論集、1-5-15 (2000.9).
[18] J. Suontausta, J. Häkkinen, and O. Viikki, "Fast decoding in large vocabulary name dialing," Proc. ICASSP, vol. III, pp. 1535-1538 (2000.6).
[19] S. Lee, K. Hirose, and N. Minematsu, "Efficient search strategy in large vocabulary continuous speech recognition using prosodic boundary information," Proc. ICSLP, vol. IV, pp. 274-277 (2000.10).
[20] Z. Song, F. Zheng, and W. Wu, "Statistical knowledge based frame synchronous search strategies in continuous speech recognition," Proc. ICASSP, vol. III, pp. 1583-1586 (2000.6).
[21] 秋葉友良、伊藤克亘、“音声認識における精度と効率の制御を可能とする確率的ビームサーチ法、” 音講論集、2-Q-3 (2000.9).
[22] A. Lee, T. Kawahara, K. Takeda, and K. Shikano, "A new phonetic tied-mixture model for efficient decoding," Proc. ICASSP, vol. III, pp. 1269-1272 (2000.6).
[23] 李晃伸、河原達也、“モノフォンモデルを用いたコードブック選択による音響尤度計算の高速化、” 音講論集、1-5-13 (2000.9).
[24] 尾本幸宏、村井則之、小林哲則、“高速計算可能な確率分布モデルSGMとその話者識別への応用、” 音講論集、1-Q-6 (2000.3).
[25] A. Aiyer, M.J.F. Gales, and M.A. Picheny, "Rapid likelihood calculation of subspace clustered Gaussian components," Proc. ICASSP, vol. III, pp. 1519-1522 (2000.6).
[26] 堀部千寿、中川聖一、“出力確率密度分布のクラスタリングを用いた大語彙連続音声認識の高速化の検討、” 音講論集、1-5-14 (2000.9).
[27] S. Kanthak, K. Schütz, H. Ney, "Using SIMD instructions for fast likelihood calculation in LVCSR," Proc. ICASSP, vol. III, pp. 1531-1534 (2000.6).
[28] B. Ramabhadran, Y. Gao, and M. Picheny, "Dynamic selection of feature spaces for robust speech recognition," Proc. ICSLP, vol. III, pp. 913-916 (2000.10).
[29] H. Schwenk and J.-L. Gauvain, "Combining multiple speech recognizers using voting and language model information," Proc. ICSLP, vol. II, pp. 915-918 (2000.10).
[30] F. Wessel, R. Schlüter, and H. Ney, "Using posterior word probabilities for improved speech recognition," Proc. ICASSP, vol. III, pp. 1587-1590 (2000.6).
[31] G. Evermann and P.C. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," Proc. ICASSP, vol. III, pp. 1655-1658 (2000.6).
[32] R. Schlüter, F. Wessel, and H. Ney, "Speech recognition using context conditional word posterior probabilities," Proc. ICSLP, vol. II, pp. 923-927 (2000.10).
[33] 緒方淳、有木康雄、“信頼度に基づく繰り返しデコーディングによる音声認識、” 音講論集、3-5-1 (2000.9).
[34] N. Murai and T. Kobayashi, "Dictation of multiparty conversation using statistical turn taking model and speaker model," Proc. ICASSP, vol. III, pp. 1575-1578 (2000.6).
[35] P. Heracleous, S. Nakamura, and K. Shikano, "A technique for likelihood normalization in the 3-D N-best search for simultaneous recognition of multiple sound sources," 音講論集、2-Q-14 (2000.9).
[36] 瀬川修、武田一哉、板倉文忠、“端点検出を行わない連続音声認識の検討、” 音講論集、2-Q-4 (2000.9).
[37] 加藤正治、斎院俊典、伊藤彰則、好田正紀、“単語グラフ生成におけるパラメータ最適化の検討、” 音講論集、1-5-17 (2000.9).
[38] G. Boulianne, J. Brousseau, P. Ouellet, and P. Dumouchel, "French large vocabulary recognition with cross-word phonology transducers," Proc. ICASSP, vol. III, pp. 1675-1678 (2000.6).
[39] K. Livescu and J. Glass, "Lexical modeling of non-native speech for automatic speech recognition," Proc. ICASSP, vol. III, pp. 1683-1686 (2000.6).