3F-6

「質より量のアプローチ」による音声認識のシミュレーションと計算機の性能評価

川辺弘之† 杉森公一† 瀬戸就一‡ 下村有子†
金城大学† 金城大学短期大学部‡

1. はじめに

本研究を含む進行中の研究プロジェクトの目的は、聴覚障害学生が大学の講義を不自由なく受講できるシステムの構築である。この研究プロジェクトでは、多数の入力ボランティアのキーボード入力によりノートテイクを実現していた。[1]「質（正確さ）より量（人数）」の概念にもとづいたノートテイクシステムである。そこで問題になったのは、多数の入力ボランティアを確保することと、キーボード入力の正確さであった。そこで、入力ボランティアを確保する問題を解決するため、キーボード入力を講師による音声入力に置き換えることを我々は構想している。さらに、音声認識率は約 80%と高くはないが、初心者によるキーボード入力より優る。したがって、音声認識に「質より量」のアプローチを適用することで上記の問題を解決できる。すなわち、多数の音声認識エンジンを同時並列実行することにより音声認識率の向上を目指すことが本研究プロジェクトの次の目標となっている。

並列実行はマイクロプロセッサにおける現在の趨勢を反映している。最近のパーソナルコンピュータは２並列ではあるが並列コンピュータとなっている。また、８から 16 個のプロセッサコアを備えたワークステーションも廉価に市販されている。この状況を考慮すると、音声認識エンジンのアルゴリズムを工夫して認識率を向上させるアプローチ以外に、多数のプロセッサコアで異なった特徴を持った音声認識エンジンを同時並列実行するアプローチも有望である。この場合、多数の音声認識結果から最終的な音声認識結果を多数決で抽出することになる。

本研究では、まず、音声認識システムの並列実行についての数学モデルとそのコンピュータシミュレーション結果を簡単に紹介する。次に、音声認識プログラムを並列実行させて、計算機への負荷の程度を調べることで、計算機ハードウェアの性能評価を行う。そして、現在の技術水準で実現可能なノートテイクシステムとその将来像について検討する。

2. 数学モデルとコンピュータシミュレーション

2.1 数学モデル

我々のシステムでは、多くの初心者が同時に講義データを入力すると想定している。したがって、講師が発した文章データを複数得ることができる。複数の入力データがあれば、正しく入力された単語もあれば、誤って入力された単語もある。このとき、入力する人数が増えれば、正しい単語も多くなることが期待できる。一方、単語の入力誤りの傾向とその発生率や発生箇所はランダムで、全く同じ入力誤りは現れないと仮定する。したがって、複数の入力単語データにおいて、二つ以上同じ入力単語データが現れたならば、それは正しい単語であると仮定する。すなわち、入力誤りの完全なランダム性を仮定する。そして、複数の入力単語データから正しい入力箇所を抽出し、つなぎ合わせることで、元の文章の再現が可能になる。

2.2 コンピュータシミュレーション

入力者数が増加すれば、正しい入力の期待値が向上すると期待できることを示す。すべての入力者が確率 0.5 で正しく入力できるという条件下で入力者数を変えた場合、６～８人の入力者で十分な精度が得られる。さらに、10 並列程度で 95%を超える認識率が得られる。これは、現在のワークステーションの能力で処理できる領域である。[2]

3. 負荷実験

音声認識プログラムの同時並列実行がコンピュータに与える負荷を調べた。音声認識プログ

Computer Simulation and Load Evaluation of Voice Recognition by “Quantity rather than Quality” Approach
† H.Kawabe, K.Sugimori and Y.Shimomura, Kinjo University
‡ S.Seto, Kinjo College

ラムとして Julius を使用した。Julius は音声認識システムの開発・研究のための高性能な汎用大語彙連続音声認識エンジンであるが、オープンソースソフトウェアなので、容易に入手でき、また、自由な改変が可能である。[3]

Julius を実行するコンピュータとして、２種類のコンピュータを用意した。ノート PC（Apple MacBook Pro、2 Core）とワークステーション（HP Z800/CT、16 Core）である。これは、現在利用可能な可搬型の PC と将来容易に入手できるであろう PC という意味を持つ。

音声認識の対象としたのはデジタル録音したファイル（23 秒）である。実際の運用ではマイクから入力することになるのだが、負荷実験では同一の音源とすることで、音源の差異による影響を排除した。

２つの観点から負荷実験を行った。１つ目は同時に実行される複数の Julius がすべて実行を終えるのに要する時間であり、２つ目は同時実行時に Julius が要求する総主記憶容量である。top コマンドで消費される主記憶容量を調べつつ、time コマンドで実行時間を測定した。また、負荷実験での Julius の並列数を、ノート PC（２コア）では１から８、ワークステーション（12 コア）では１から 36 とした。

4. 結果と考察

ノート PC では、同時実行プログラム数に比例して実行時間が延びた。OS やウィンドウシステムが 1 コアを占有し、残りの 1 コアが音声認識プログラムに割り当てられたと考えられる。また、6 並列程度で実行した場合、22 秒の音声の認識に 26 秒程度要している。認識に要する時間が大きすぎ、ノートテイクにおいて応答が緩慢になってしまう。

次に、ワークステーションでの負荷実験の結果を述べる。各コアに１つのプログラムが割り当てられている 12 並列実行までは、実行時間の上昇が低く抑えられた。それ以上の並列実行を行うと、１つのコアを２つまたは３つのプログラムが共同で利用することになるので、実行時間の増える割合が上がった。また、6 並列程度での実行時間は、22 秒の音声の認識に 6 秒程度であった。ノート PC に比べ十分に短い実行時間であり、ノートテイクにおいても快適な応答を期待できる。

Julius のメモリ消費量は、ノート PC、ワークステーションいずれにおいても、音声認識プログラム１つあたり、約 120MB であった。メモリに対す負荷はそれほどではない。これだけのメモリ消費量ならば、現代のノート PC であってでも相当数の同時実行が可能である。

これらのことから、音声認識に基づく「質より量のアプローチ」によるノートテイクシステムを実現するにあたり、現在のノート PC はメモリ搭載量では十分だが演算能力では非力で、ワークステーションはメモリ搭載量・演算能力ともに十分だと、結論できる。

5. まとめ

ノートテイクシステムのための音声認識プログラムの並列実行についての数学モデルとそのコンピュータシミュレーション結果を簡単に紹介し、6～8 程度の並列実行が求められていることを示した。また、音声認識プログラムを並列実行させて、計算機への負荷の程度を調べることで、計算機ハードウェアの性能評価を行った。現在市販のノート PC では能力不足だが、ワークステーションならば、十分な性能を持つことがわかった。

この結果から、聴覚障害学生が大学の講義を不自由なく受講できるシステムを現在の技術水準で構築するならば、現在のワークステーション級の演算能力を持つノート PC が必要である。技術水準は日々、進歩している。このようなノート PC が入手可能になる時が待ち遠しい。

謝辞

本研究は文部科学省科学研究費補助金基盤研究（C）22500519 の助成を受けたものである。ここに記して感謝の意を表す。

参考文献

[1] S.Seto et.al., Mathematical Model of Text Reproduction Based upon “Quantity Rather Than Quality” Concept, The 20th National Conference of Australian Society for Operations Research, Gold Coast, Australia (2009)

[2] H.Kawabe et.al., Mathematical Model and Computer Simulation of Text Reproduction Based upon “Quantity Rather Than Quality” Concept, The 40th International Conference on Computers and Industrial Engineering, Awaji-shima, Japan (2010)

[3] A.Lee et.al., Julius - an open source real-time large vocabulary recognition engine, Proc. European Conf. on Speech Communication and Technology, pp.1691-1694 (2001)