2F-5

音声認識エンジンの複数実行の効果

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。このような熟練を要しない音声入力による聴講支援システムの構築は，就業の場においても多くの応用が期待できる。例えば聴覚障害者の就業支援や遠隔会議システムへの字幕付加などである。一方健常者においても業務日報の音声入力などに応用が可能である。

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こと以外に，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に異なった設定を施した場合の認識結果、そして、多数決原理で抽出した結果を与え、本手法の有効性を示す。

Effectiveness of Multiple Execution of Voice Recognition Engine
†H. Kawabe, K. Sugimori, Y. Shimomura・Kinjo University
‡S. Seto・Kinjo College

２．モデルとコンピュータシミュレーション

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きた。本研究では、人力での入力をコンピュータによる音声認識に置き換えることを目指している。この際、特性の異なった音声認識プログラムが並列実行される。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。この中には，正しく認識された単語もあれば，誤って認識された単語もある。このとき，並列動作する音声認識プログラムの数が増えれば，正しい単語も多くなることが期待できる。一方，単語の認識誤りの傾向とその発生率や発生箇所はランダムで，全く同じ認識誤りは現れないと仮定する。したがって，複数の単語データにおいて，２つ以上同じ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，認識誤りの完全なランダム性を仮定する。そして，複数の認識単語データから正しい認識箇所を抽出し，つなぎ合わせることで，元の文章の再現が可能になる。

音声認識プログラムの並列実行の数が増加すれば，正しく認識する確率が向上すると期待できる。コンピュータシミュレーションにおいて、すべての音声認識プログラムが確率 0.5 で正しく認識できるという条件下で並列数を変えた場合、6～8 並列で十分な精度が得られている。さらに，10 並列程度で 95%を超える認識率となる。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．調整パラメータによる音声認識への影響

音声認識システムは認識エンジンと認識エンジンの調節パラメータとから構成される。同一の音声であってでも、異なったパラメータで調整された認識エンジンに与えると、異なった認識結果を得る。

本研究では、音声認識エンジンとして Julius[3]を用いた。Julius では、音響モデルや、言語モデル、デコーダを変更可能であり、また、

ID	モデル	状態数	性別
1	Triphone 3000	3000	Dependent
2	PTM triphone	3000	Dependent
3	PTM triphone	3000	Independent

表１：音響モデル

ID	モデル	語数	圧縮率
A	20k.1-1	20K	-
B	20k.1-1.10p	20K	10%
C	60k.1-1	60K	-
D	60k.1-1.10p	60K	10%

表２：言語モデル

	単語間 トライフォン	言語 モデル	探索法
第 1 パス	近似的	Bigram	最尤近似
第 2 パス	厳密	Trigram	N ベスト

表３：デコーダ

解析する際、無音期間の長さなどのパラメータを調整できる。我々が今回用いた音響モデル、言語モデルを表１、２に示す。

Julius に「およそ桃太郎の話を知らない人はいない」という文章を与え、表３のデコーダを用いて音声認識させた結果が表４である。表４では、認識された単語のうち原文に含まれている単語の数を認識率とした。

それぞれの認識率は 6/10(60%)から 7/10(70%)であったが、多数決原理を用いて共通部分を取り出すことで 8/9(89%)へと認識率が向上した。

４．結論

低い認識率の音声認識エンジンであってでも、種々の音声認識エンジンを数多く同時並列実行し、認識結果に対して多数決を行うことで、少々の誤認識は隠蔽され、結果的に高い認識率が得られる。このことはコンピュータシミュレーションで予想されていたが、実際に音声認識エンジンで実行して確認した。したがって、我々の手法は効果的であることが明らかになった。

今回得られた最終的な認識率はまだまだ満足できるものではない。言語モデルにおける辞書の語彙数を増やすことや、パラメータをさらに調整することで、さらに認識率を向上させたい。また、多くの話者の種々の文章を与えてでも、高い認識率が得られることを目指したい。

謝辞

本研究の一部は日本学術振興会科学研究費基盤研究（C）No. 22500519 の助成を受けたものである。

参考文献

[1] S. Seto, et.al., The 20th National Conference of Australian Society for Operations Research, Australia (2009)

[2] H. Kawabe, et.al., The 40th International Conference on Computers and Industrial Engineering, Japan (2010)

[3] A. Lee, et.al., Proc. European Conf. on Speech Communication and Technology, pp.1691-1694, 2001.

ID	文章	認識率
原文	およそ　ももたろう　の　はなし　お　しらない　ひと　わ　いない	-
1-A	およそ　の　かん　お　はなし　お　しらない　ひと　わ　いない	7/10
2-A	およそ　もだん　は　が　て　お　しらない　ひと　わ　いない	6/10
3-A	およそ　もと　の　はなし　も　しらない　ひと　わ　いない	6/9
1-B	およそ　のうせん　の　はなし　お　しらない　ひと　わ　いない	7/10
2-B	およそ　ぼん　の　はんが　しゅう　お　しらない　ひと　わ　いない	6/10
3-B	およそ　ぼん　の　はなし　も　しらない　ひと　わ　いない	6/9
1-C	およそ　の　かん　の　はなし　お　しらない　ひと　わ　いない	7/10
2-C	およそ　もと　の　はなし　も　しらない　ひと　わ　いない	6/9
3-C	およそ　もと　の　はなし　も　しらない　ひと　わ　いない	6/9
1-D	およそ　の　かん　の　はなし　お　しらない　ひと　わ　いない	7/10
2-D	およそ　もと　の　はなし　も　しらない　ひと　わ　いない	6/9
3-D	およそ　の　こぶ　の　はなし　も　しらない　ひと　わ　いない	6/10
共通	およそ　もと　の　はなし　お　しらない　ひと　わ　いない	8/9

表４：認識結果