聴覚障害者のための音声認識システム
－ 多数決原理による認識誤りの訂正 －

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‡

{kawabe, sugimori, seto, shimo}@kinjo.ac.jp

１．はじめに

聴覚障害学生数の増加は 1950 年代には本格的になった。日本聴覚障害学生高等教育支援ネットワークによると、日本では大学の約 60％に聴覚障害学生が在籍しており、修学している聴覚障害者の約60％が大学生である(Shirasawa 2005)。聴覚障害学生を支援するために、大学のサークルやボランティア学生、外部のグループは様々な方法、身振り、即時音声認識による字幕やテキスト提示、などを使用している。ノートテイクは広く使われており、手書きや PC、音声認識ソフトウェアなどでなされている(Kitabayashi 2004, Lang 2002 and Tachiiri 2003)。だが、ノートテイクによる支援がなされているのは、聴覚障害者学生の在籍する大学の 13.6％に過ぎない。なぜなら、ノートテイクには高度な技能が必要だからである。そのため、基礎的なノートテイクの技能を身につけるまでには 4 ヶ月、熟練するには約 2 年を要する。さらに、苦労してノートテイクの技能を身につけたとしても、大学生は 4 年間で卒業してしまうので、ノートテイクに熟練したボランティア学生を確保し続けるのは困難である(Hatakeyama 1994)。

このような状況を打開するために、様々なやり方での支援が試みられている。従来のノートテイクでは、複数のノートテイカーが聴覚障害者学生と同じ教室内で支援する。熟練ノートテイカーの大学内での確保が困難なことの対策として、遠隔地の熟練ノートテイカーチームがスマートホンで教室内の聴覚障害者学生を支援する試みがある(Kobayashi 2009)。また、音声認識を利用した即時字幕システムの実験もある。前者には熟練ノートテイカーの確保に、後者には現在の音声認識結果の正確さに問題がある。

これらの問題を解決するため、多くの初心者ボランティアがノートテイクする支援システムを我々は提案した(Seto 2007, Seto 2008 and Sugimori 2008)。このシステムでは、ボランティアは初心者なので、入力できるのは話し言葉の断片と仮定している。そして、IT を活用してその断片からノート全体を再現する。すなわち、一人の熟練ボランティアではなく、多くの初心者ボランティアに頼ったノートテイクシステム、「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。だが、ひとりの聴覚障害者学生を支援するために多くのノートテイクボランティアを教室内に配置するのも無理がある。また、初心者ボランティアのキーボード入力や手書きは絶望的な精度と速度である。そこで、音声認識の利用である。現在の音声認識結果が満足できるレベルでないことは初心者ボランティアと対応している。ノートテイクボランティアを音声認識で置き換え、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。

音声認識エンジンには、言語モデルや音響モデル、音声辞書、そして認識のためのパラメータを変える自由度が存在する。これらを変えることで音声認識エンジンは同一の音声であってでも異なった認識結果を返す。したがって、個性の異なった多くの初心者ボランティアを音声認識システムで実現可能である。

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個性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，多数の認識結果から最終的な認識結果を多数決で抽出することになる。

ここでは、まず、我々のノートテイクシステムの数学的モデルとコンピュータシミュレーションの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。

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，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。

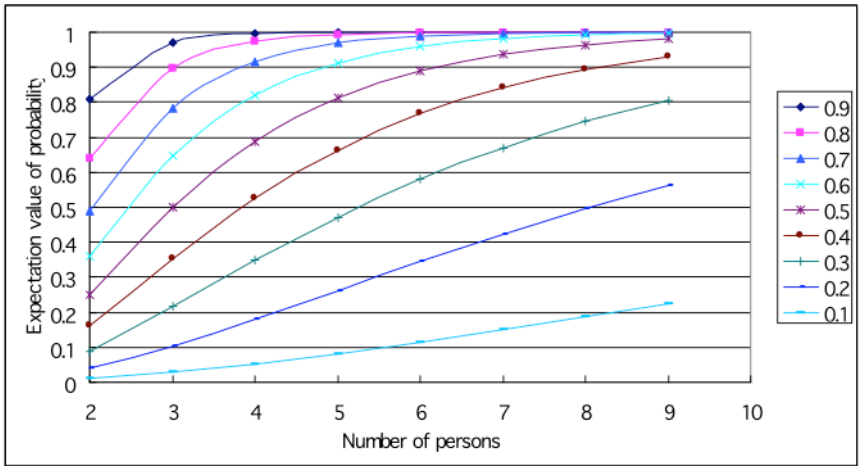

図１：入力する「人」数と正しい入力の期待値

図１に入力ボランティアの人数（音声認識エンジンの数）と多数決で正しい結果が得られる期待値の関係を示す。各曲線は入力ボランティア（音声認識エンジン）が正しく入力できる精度（確率）に対応している。音声認識エンジンの実行数が増加すれば，正しく認識する確率が向上すると期待できる。すべての入力ボランティア（音声認識エンジン）が確率 0.5 で正しく認識できるという条件下では、8〜9 人（並列）で十分な精度が得られている。さらに，10 人（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(Kawabe 2010)。

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 を用いた(Lee 1002)。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、解析する際、無音期間の長さや言語スコアの重み、ビームの幅などの音声認識のアルゴリズムでのパラメータを調整できる。

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、音響モデルや、言語モデル、音声辞書を変えて音声認識させた結果が表１である。最初の行が原文、最後の行が共通部分、その他が音声認識エンジンで認識された文である。認識された文において、上段が認識された単語、下段が音声認識のスコアである。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 6/12(0.50)から 11/12(0.92)であったが、認識率が悪い結果を含め、多数決原理により共通部分を取り出すことで最終的に 11/12(0.92)の認識率となった。

表 1：音声認識結果（モデルによる違い）

文章											認識率
およそ	ももたろう		の	はなし			お	しらない	ひと	わ	いない	-
およそ	の	かん	お	はなし			お	しらない	ひと	わ	いない	9/12 (0.75)
0.082	0.945	0.049	0.052	0.200			0.445	0.947	0.844	0.802	0.660	
およそ	もと	ろん		は	が	しゅ	お	しらない	ひと	わ	いない	6/12 (0.50)
0.110	0.112	0.188		0.107	0.253	0.119	0.529	0.954	0.920	0.740	0.675	
およそ	の	かぶ	の	はなし			も	しらない	ひと	わ	いない	9/12 (0.75)
0.088	0.772	0.011	0.229	0.207			0.619	0.872	0.882	0.675	0.627	
およそ	のーそん		の	はなし		の	お	しらない	ひと	わ	いない	10/12 (0.83)
0.105	0.044		0.173	0.293		0.294	0.459	0.901	0.838	0.782	0.515	
およそ	もと		の	はんが		しゅー	お	しらない	ひと	わ	いない	9/12 (0.75)
0.125	0.103		0.116	0.642		0.050	0.247	0.920	0.903	0.748	0.099	
およそ	の	かぶ	の	はなし			も	しらない	ひと	わ	いない	9/12 (0.75)
0.105	0.726	0.012	0.119	0.135			0.632	0.841	0.877	0.674	0.489	
およそ	のーそん		の	はなし			お	しらない	ひと	わ	いない	11/12 (0.92)
0.275	0.026		0.134	0.161			0.424	0.861	0.689	0.837	0.620	
およそ	もと	ろーん	の	はんが		しゅー	お	しらない	ひと	わ	いない	8/12 (0.67)
0.318	0.550	0.075	0.062	0.607		0.051	0.242	0.913	0.855	0.769	0.636	
およそ	もと		の	はなし			も	しらない	ひと	わ	いない	10/12 (0.83)
0.277	0.197		0.402	0.211			0.585	0.857	0.831	0.701	0.582	
およそ	のーそん		の	はなし			お	しらない	ひと	わ	いない	11/12 (0.92)
0.275	0.031		0.137	0.154			0.418	0.839	0.720	0.800	0.608	
およそ	もと	ろーん	の	はんが		しゅー	お	しらない	ひと	わ	いない	9/12 (0.75)
0.318	0.547	0.076	0.053	0.607		0.052	0.219	0.898	0.912	0.742	0.627	
およそ	もと		の	はなし			も	しらない	ひと	わ	いない	10/12 (0.83)
0.277	0.110		0.376	0.211			0.582	0.838	0.886	0.680	0.574	
およそ	もと		の	はなし			お	しらない	ひと	わ	いない	11/12 (0.92)

表 2：音声認識結果（認識パラメータによる違い）

文章											認識率
およそ	ももたろう		の	はなし			お	しらない	ひと	わ	いない	-
およそ	もと	ろん		は	が	しゅ	お	しらない	ひと	わ	いない	6/12 (0.50)
0.110	0.112	0.188		0.107	0.253	0.119	0.529	0.954	0.920	0.740	0.675	
およそ	もと	ろん		はんが		しゅー	お	しらない	ひと	わ	いない	7/12 (0.75)
0.117	0.110	0.093		0.671		0.036	0.496	0.943	0.921	0.740	0.677	
およそ	もと	ろーん	の	はなし			お	しらない	ひと	わ	いない	10/12 (0.83)
0.121	0.659	0.036	0.208	0.081			0.524	0.927	0.916	0.802	0.649	
おおよそ	もと	ろーん	の	はなし			も	しらない	ひと	わ	いない	8/12 (0.67)
0.033	0.532	0.030	0.200	0.141			0.234	0.888	0.638	0.691	0.562	
およそ	もと	ろん/ろーん	の	はなし			お	しらない	ひと	わ	いない	10/12 (0.83)
-	-	ろん	-	-			-	-	-	-	-	10/12 (0.83)

多数決原理による共通部分において、同数のため有望な単語を絞りきれない場合には、下段のスコアを考慮する。表１ではスコアを活用する必要がなかった。

表２では、音声認識のアルゴリズムのパラメータを変更した場合の認識結果を与える。認識結果の１行目は表１の認識結果の２行目と同一である。それぞれの認識率は 6/12(0.50)から 10/12(0.83)であったが、多数決原理により 10/12(0.83)の認識率となった。パラメータを変更することで、認識結果が劇的に向上することがわかる。表２の場合、多数決原理による共通部分に同数の候補が現れたので、上で述べたように、音声認識のスコア（下段の数値）により最終的な候補を決定した。

表１，２ともに、「ももたろう」を認識できていない。音声辞書に「ももたろう」が含まれていないことが原因と考えられる。もし含まれていたならば、最終的な結果はもっと良かったであろう。

４．結論

ここでは、「質（正確さ）より量（人数）」を基本概念とする我々のノートテイクシステムを紹介した。この特性から、ノートテイクボランティアを音声認識システムで置き換え可能となった。低い認識率の音声認識エンジンであってでも、個性の異なった音声認識エンジンを数多く実行し、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た。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。

聴覚障害者は発話内容を知らない。そして、音声認識の結果は玉石混交であり、ただ一つの音声認識エンジンが与えた結果が良好だという保証はない。しかしながら、複数の音声認識エンジンによる認識結果があれば、多数決原理により良好な結果を常に期待できる。

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

Hatakeyama, T., Senda, T., Furukawa, S., Ebizuka, K. and Nakagawa, Y., The Note Taking Support System for the Person Who is Hearing Impaired (in Japanese), The 10th Symposium on Human Interface, 363-368, Tokyo, 18-20, Oct. 1994.

Kano, K., Ito, K., Kawahara, T., Takeda, K., and Yamamoto, M., Voice recognition system (in Japanese), Ohmsha, Tokyo, 2001.

Kawabe, H., Sugimori, K., Shimomura, Y., and Seto, S., Mathematical Model and Computer Simulation of Text Reproduction Based upon “Quantity Rather Than Quality” Concept, Proceedings of the 40th International Conference on Computers and Industrial Engineering (CIE) 2010.

Kawabe, H., Sugimori, K., Seto, S., and Shimomura, Y., Effectiveness of Simultaneous Execution of Voice Recognition Programs for Hearing Impaired, Proceedings of the 10th Asia-Pacific Conference on Industrial Engineering and Management Conference (APIEMS) 2011.

Kitabayashi, T., Report on Remote Operation Support of Summarized Translation by PC (in Japanese), Tsukuba College of Technology Techno Report, 11, 15 -20, 2004.

Kobayashi, M., Ishihara, Y. and Miyoshi, S., Real-Time Text Presentation System with Pronunciation along with Chinese Characters Using Mobile Phones (in Japanese), Tsukuba College of Technology Techno Report, 16, 23 -25, 2009.

Lang, H. G., Higher Education for Deaf Students: Research Priorities in the New Millennium, Journal of Deaf Studies and Deaf Education, 7, 267, 2002.

Lee, A., Kawahara, T. and Shikano, K., Julius - an open source real-time large vocabulary recognition engine, Proc. European Conf. on Speech Communication and Technology, pp.1691-1694, 2001.

Nakano, S., Kuroki, H., Ino, S., Kanazawa, T., Kikuchi, M. and Ifukube, T., The Real-time

Remote Captioning System in order to Support Students with Hearing Impairment in Higher Education (in Japanese), The 42nd Conference of The Japanese Association of Special Education, Tokyo, 10-12, Sept. 2004.

Seto, S., Kawabe, H. and Shimomura, Y., A Reproduction of Time Sequential Data from a Set of Time Sequential Fragments with Random Gaps, Proceedings of the 8th Asia-Pacific Conference on Industrial Engineering and Management Conference (APIEMS) 2007.

Seto, S., Sugimori, K., Shimomura, Y., Kawabe, H. and Kimura, T., Input Device of Note Taking System of Hearing Impaired Student, Proceedings of the 9th Asia-Pacific Conference on Industrial Engineering and Management Conference (APIEMS) 2008.

Shirasawa, M., Current Situation and Problems on Deaf / Hard of Hearing Student Services in Japan: Based on the Results of A Nationwide Survey, The 2nd International Conference of Higher Education of Students with Disabilities, Tokyo, 27, Mar. 2005.

Sugimori, K., Seto, S., Kimura, T., Kawabe, H. and Shimomura, Y., A Reproduction of Time Sequential Data from a Ser of Time Sequential Fragments with Random Gaps: Improvement of Algorithms for Word Alignment, Proceedings of the 9th Asia-Pacific Conference on Industrial Engineering and Management Conference (APIEMS) 2008.

Tachiiri, H., Inoue, K. and Miyaike, Y., Support system for full-joining to the lecture using voice recognition for the students with hearing impaired (in Japanese), IEICE technical report - Natural language understanding and models of communication, 103, 43-48, 2003.