人間の第二言語運用能力との比較による音声認識性能評価

竹澤 寿幸　安田 圭志　水島 昌英　菊井 玄一郎 (ATR **音声言語コミュニケーション研究所**)

1　まえがき

機械による音声認識を人間の言語能力と比較して数量化する新しい評価手法を提案する。機械にとって人間の言語は母語ではない。人間にとって、母語であれば、相手の発話スタイルが多少変動したり、周囲の環境が多少変化したりしたところで、音声を聞き取るのはたやすい。しかしながら、母語でない場合は、発話スタイルや周囲環境の変化によって、聞き取りが難しくなる場合がある。

機械による音声認識の性能は、一般に認識率で評価される。しかし、ある特定の音声認識システムを同一の内部パラメータで利用しても、評価対象のテストセットが異なれば、性能評価結果の認識率はたいてい違う値となる。補助的な情報として、発話内容に関するパープレキシティやノイズレベルが示されることがあるが、それだけで変化の要因が説明できるわけではない。

もし、母語話者には影響を与えないが、機械による音声認識には影響を与える要因が、非母語話者の聞き取り能力に影響を与えており、かつ、機械に与える影響と非母語話者に与える影響に関係があれば、テストセットに依存せず、機械による音声認識性能を数量化することができる。

音声自動翻訳システム研究のために収集した日本語話者と英語話者の対話音声から選んだ複数のテストセットに対して、さまざまな TOEIC スコアを有する日本語ネイティブによる英語の聞き取りデータを集めた。言いよどみの含まれる割合に代表される発話スタイルの違いは、機械による音声認識結果のみならず、非母語話者である日本語ネイティブの聞き取り能力にも影響を与えることを示す。そして、英語音声認識システムの性能を TOEIC スコアに換算することを試みる。

さらに、音声自動翻訳システムを始めとする音声インタフェースでは、音声認識が単独で利用されるのではなく、何らかの応用システムと組み合わせて使われる。そこで、単に表層的に単語が一致する割合だけではなく、内容的に、応用システムに与える影響が重要となる。もし仮に人間は非母語話者であっても重要な部分を聞き取り、重要でない部分を聞き流しており、機械は単にランダムに聞いたり、聞き誤ったりしているのであれば、応用システムへ与えるダメージが異なる可能性がある。

本稿では、応用システムとして機械翻訳システムを取り上げ、非母語話者であるさまざまな TOEIC スコアを有する日本語ネイティブが聞き取ったテキストを機械翻訳した結果と、音声認識結果を機械翻訳した結果を比較し、同じくらいの聞き取り能力であれば、結果に対しても顕著な差はないことを示す。

Evaluation of speech recognition systems by measuring machine output against non-native results by Toshiyuki Takezawa, Keiji Yasuda, Masahide Mizushima, and Genichiro Kikui (ATR Spoken Language Translation Research Laboratories)

2　聞き取り能力の比較

2.1 テストセットの特徴

本稿で扱うテストセットは、音声自動翻訳システムを介して日本語話者と英語話者が課題遂行対話を行うことにより得られたデータ(MAD: Machine-Aided Dialogues) [1, 2, 3]から選んだものである。条件を変更して複数回の実験を行っており、3 回目と 4 回目に相当する MAD3 [2], MAD4 [3] のテストセットを使った。音声自動翻訳システムが現在開発途上であることから、良質な対話データ収集を目的とし、MAD3, MAD4 ともに、音声認識システムの代わりにタイピストが発話を書き起こし、機械翻訳システムに入力する形態で集めたものである。MAD3 と MAD4 では、システム利用者である話者に与える教示が異なる。MAD3 では、1 回の発話は 10 秒以内というような負担の少ない話し方を教示し、MAD4 では、短く簡潔に話すというような機械を意識した話し方を教示している。その結果として、発話スタイルが異なる。一つの目安として、日本語発話に言いよどみの含まれる割合を表 1 に示す。参考情報として、日本人同士の旅行対話(SDB/TRA) [4]と、人間の逐次通訳者を介した日英対話(SLDB) [4]の数値も示す。

表 1　テストセットの発話スタイル

	言いよどみを含む発話
日本人同士の対話(SDB/TRA)	29.4%
通訳者を介した対話(SLDB)	16.3%
機械を介した対話(MAD3)	13.8%
機械を介した対話(MAD4)	6.2%

表 1 から、日本人同士の対話に比べ、通訳者を介する状況では言いよどみを含む発話の割合が減り、MAD3 は人間の通訳者を介する状況に近く、MAD4 は大幅に減ることがわかる。英語側発話についても傾向は似ている。このような特性をもつデータから、英語の聞き取りデータを集めるテストセットを作成した。表 2 にその概要を示す。

表 2　英語テストセット

	話者数	発話数	単語数	平均発話長
MAD3	6	504	5,709	11.33 語／発話
MAD4	12	502	4,694	9.35 語／発話

日本人の英語能力を TOEIC スコア[5]で代用することとし、TOEIC スコア 300 点台から 900 点台まで、100 点台ごとに 3 名、合計 21 名の被験者に MAD3, MAD4 の英語音声の聞き取りをさせた。MAD3, MAD4 で被験者の重複はな

い。ただし、都合により、MAD3 テストセットの 400 点台は 2 名であり、MAD3 テストセットは合計 20 名である。

2.2 実験結果と考察

英語音声認識システムは、ATR で研究開発した ATRASR [6]を用いた。言語モデルは、ATR で構築したコーパス[7]で訓練したマルチクラス複合バイグラムを用いた。認識実験結果を表 3 に示す。

表 3　英語音声認識実験結果

	単語認識率	パープレキシティ	未知語率
MAD3	77.9%	55.3	0.65%
MAD4	86.4%	39.8	0.05%

パープレキシティ、未知語率、平均発話長という基本特性が異なるために、同じ英語音声認識システムを使っても、テストセットによって単語認識率が異なる。ちなみに、実験はともに屋内の打ち合わせ室で行っており、周囲環境は共通である。課題も共通である。

さまざまな TOEIC スコアを有する日本語ネイティブにこのテストセットを聞き取らせた結果を図 1, 図 2 に示す。

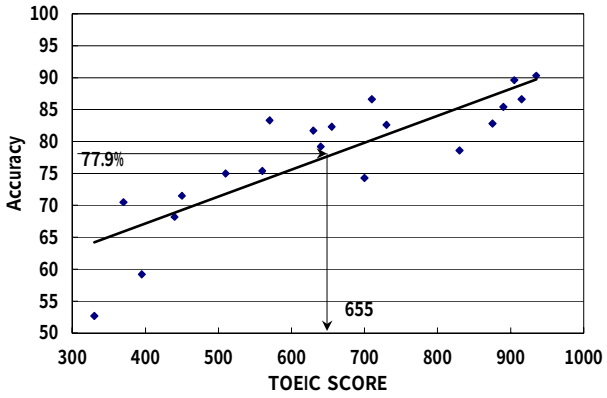

図 1　MAD3 の英語聞き取り実験結果

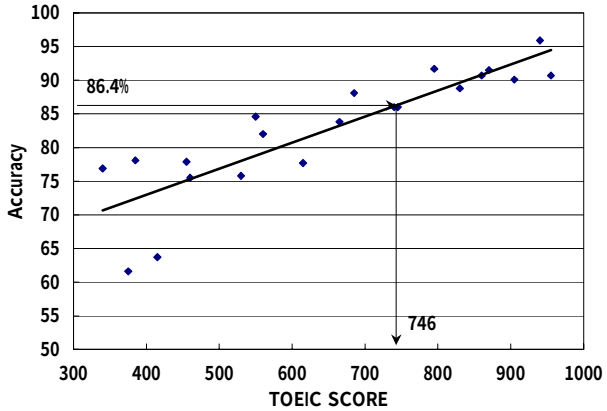

図 2　MAD4 の英語聞き取り実験結果

TOEIC 被験者データを集める際には、特に時間制限は設けず、一つの発話を 2 回まで聞くことを許し、聞き取った内容をパソコンでタイプ入力させた。辞書を引くことは認めず、スペルチェッカも使わせなかった。したがって、スペルミスが含まれ、また、たとえば数字表記について、アラビア数字を使ったり、英単語で書いたりするばらつきがある。全角記号と半角記号が混在するような形式的な不一致などについては整形を行った。その後、ATR の英語形態素解析ツールで品詞タギングを行い、音声認識結果の計算ツールで認識結果と同様に値を求めた。

図 1, 図 2 のプロットが、それぞれ被験者に対応する。回帰直線をあわせて示した。相関係数を表 4 に示す。

表 4　TOEIC スコアと聞き取り能力の相関

	相関係数
MAD3	0.848
MAD4	0.868

図 1 と図 2 を比較すると、正しく聞き取れたとみなせる割合は、MAD3 テストセットより MAD4 テストセットが高い。表 3 によれば、英語音声認識システムの認識結果も、MAD3 より MAD4 が高い。そこで、試みに、英語音声認識システムの認識率から回帰直線を介して、TOEIC スコアを推定した。推定される TOEIC スコアを表 5 に示す。

表 5　英語音声認識システムの推定 TOEIC スコア

	推定 TOEIC スコア
MAD3	655
MAD4	746

これらの結果が示唆する内容は次の通りである。

- 日本人の英語能力を TOEIC スコアで表すとすると、TOEIC スコアと英語聞き取り能力の相関は比較的高い。
- 機械による音声認識が相対的に難しいテストセットは、非母語話者の聞き取りという観点でも相対的に難しい。
- 英語音声認識システムの推定 TOEIC スコアが MAD4 より MAD3 が低いということは、短く簡潔に話すという機械を意識した話し方から、言いよどみが多く、パープレキシティも高い流暢な話し方への変化が与える影響は、非母語話者よりも機械に対して大きい可能性が高い。

誤差の分析をしたり、他の英語音声認識システムの結果を追加したりすることで、信頼性を高める議論は今後の課題とする。

3 応用システムへのダメージの評価

3.1 実験の準備

聞き取り能力の比較実験では、表層的なレベルで一致する度合いに相当する量的な議論を行った。しかし、音声自動翻訳システムを始めとする音声インタフェースでは、音声認識が単独で使われるのではなく、応用システムと組み合わせて使われる。そのため、単に量的な議論では十分ではなく、内容的に応用システムに与える影響の議論が重要となる。

本稿では、応用システムとして、機械翻訳システムを取り上げる。実験には、ATR で研究開発した統計的機械翻訳 SAT [8]を用いた。

実験するにあたり、機械翻訳システムに入力可能なレベルとなるまで、TOEIC 被験者データを整備した。具体的には、スペルチェッカでスペルミスを修正したり、数字の表記を揃えたり、形態素解析ツールで未登録語となっているものを修正したりした。まず、MAD4 テストセットのみ、そのような修正作業を行い、実験することにした。

3.2 実験結果と考察

修正を施した MAD4 テストセットに対して、まず、機械翻訳への入力側の品質を再計算した。ツールとしては、翻訳評価実験と共通のものを採用し、単語誤り率(WER)を求めた。その結果を図 3 に示す。音声認識率の計算では、表層と品詞の情報を使用したが、このツールでは表層情報のみ使用している。

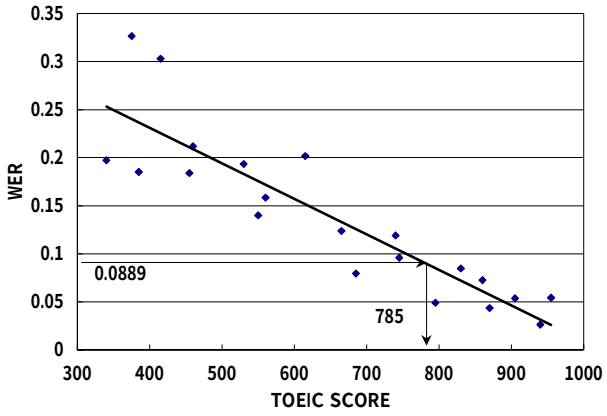

図 3　MAD4 整備データによる英語聞き取り誤り率

さらに、図 3 の整備データに対応する相関係数と、図 3 の回帰直線を介して求めた英語音声認識システムの推定 TOEIC スコアを表 6 に示す。英語音声認識システムの単語誤り率が表 3 の単語認識率と対応しないのは、表 3 ではマッチングの際に表層と品詞を使用していたのに対し、表 7 では表層のみ使用しているからである。なお、それぞれ人間と機械を比較する際には、単位や基準のずれはない。

表 6　整備データによる英語聞き取りの諸量

	MAD4
相関係数	0.890
音声認識結果(WER)	0.0889
推定 TOEIC スコア	785

表 6 によれば、整備データを用いて、表層のみのマッチングを採用すると、相関係数、推定 TOEIC スコアともに表 4, 表 5 の場合よりも高くなることがわかる。

このようにして整備した TOEIC 被験者データを SAT で機械翻訳させた。一つの翻訳に対して 15 通りの参照訳を準備し、マルチプル単語誤り率(mWER)を求めた。その結果を図 4 に示す。

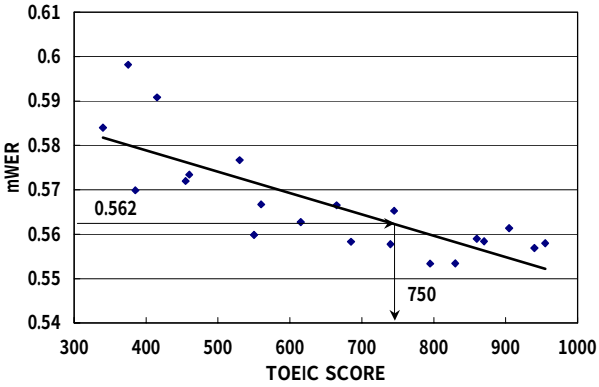

図 4　MAD4 整備データを用いた機械翻訳結果

さらに、図 4 の整備データを用いた機械翻訳結果に対応する相関係数と、図 4 の回帰直線を介して求めた英語音声認識結果の機械翻訳結果に対する推定 TOEIC スコアを表 7 に示す。

表 7　整備データを用いた機械翻訳の諸量

	MAD4
相関係数	0.798
音声認識結果の機械翻訳結果(mWER)	0.562
推定 TOEIC スコア	750

表 6 に示すように機械翻訳に対する入力側で測定した英語音声認識システムの推定 TOEIC スコアが 785 点であったのに対し、機械翻訳の出力に対する英語音声認識システムの推定 TOEIC スコアは 750 であり、大きな差はない。

4 議論と関連研究

4.1 議論

音声や言語を処理する技術に関するシステムの研究開発が盛んになっている。しかし、それらのシステムは、テストセットによって結果の評価数値が異なる。そのため、研究者や技術者の間では、共通のテストセットを用いて、技術の評価をしようという活動が盛んになっている(たとえば[9])。しかしながら、そこで得られる知見や数値は専門家でないとわからない。そこで、人間の能力にたとえることで、システムの性能を数量化できれば、非専門家にとっても有益な情報となる。仮に音声自動翻訳システムの英語音声認識システムの性能が TOEIC スコアで 700 点相当といえるならば、ユーザにとっての価値が明瞭になる。

4.2 関連研究

機械の性能を人間の能力と比較することで数量化しようとする試みに、音声翻訳システムの性能評価手法に関する一連の研究[10, 11, 12]がある。日英翻訳システムを対象に、日本人の英語能力を TOEIC スコアで表すことにし、TOEIC 被験者による英訳と機械翻訳結果を一対比較し、システムと同等とみなせる TOEIC スコアを推定する研究[10]と、その一対比較作業を自動化する研究[11, 12]である。本研究は、それを音声認識システムに適用したものとみなすこともできるが、単に聞き取り能力だけではなく、応用システムへの影響を考慮した点に特徴がある。

さらに、一般の音声インタフェースの評価[13]へ本手法を拡張する可能性について述べる。本稿では、機械翻訳システムを取り上げ，機械翻訳の評価尺度で影響を議論した。正しい内容が入力された場合のシステムの振る舞いとの差分を測定し，その差分が人間と音声認識システムの間でどのような関係にあるか調べるものが一つの拡張の可能性である。

5 むすび

機械による音声認識を人間の言語能力と比較して数量化する新しい評価手法を提案した。まず、言いよどみの含まれる割合に代表される発話スタイルの違いは、機械による音声認識のみならず、非母語話者の聞き取り能力にも影響を与えることを示した。さらに、聞き取り能力の比較のみならず、認識結果が応用システムに与える影響に着目した。機械翻訳システムを取り上げ、同じくらいの聞き取り能力であれば、翻訳結果に顕著な差がないことを示した。このように人間の能力と比較することで、テストセットに依存しないシステム性能評価が可能であることを示した。

謝辞

本研究は情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

参考文献

[1] Takezawa, T. and Kikui, G., “Collecting machine-translation-aided bilingual dialogues for corpus-based speech translation,” Proc. 8th European Conference on Speech Communication and Technology, Vol. 4, pp. 2757-2760 (2003).

[2] 竹澤寿幸, 西野敦士, 高嶌浩司, 松井孝典, 菊井玄一郎: 機械翻訳を介した対話データ収集のための実験システム, 情報科学技術フォーラム(FIT), E-036, 一般講演論文集第 2 分冊, pp. 161-162 (2003).

[3] Takezawa, T. and Kikui, G., “A Comparative Study on Human Communication Behaviors and Linguistic Characteristics for Speech-to-Speech Translation,” Proc. International Conference on Language Resources and Evaluation, pp. 1589-1592, (2004).

[4] Takezawa, T., Kikui, G., Nakamura, A., Sagisaka, Y., and Yamamoto, S., “Spoken language corpora development at ATR,” Proc. 18th International Congress on Acoustics, Vol. I, pp. 401-404 (2004).

[5] TOEIC: Test of English for International Communication, http://www.toeic.com/.

[6] 伊藤玄, 葦苅豊, 實廣貴敏, 中村哲: 音声認識統合環境 ATRASR の概要と評価報告, 日本音響学会 2004 年秋季研究発表会講演論文集 I, 1-P-30, pp. 221-222 (2004).

[7] Kikui, G., Sumita, E., Takezawa, T., and Yamamoto, S., “Creating corpora for speech-to-speech translation,” Proc. 8th European Conference on Speech Communication and Technology, Vol. 1, pp. 381-384 (2003).

[8] Watanabe, T., Imamura, K., and Sumita, E., “A statistical machine translation based on hierarchical phrase alignment,” Proc. 9th International Conference on Theoretical and Methodological Issues in Machine Translation, pp. 188-198 (2002).

[9] International Workshop on Spoken Language Translation – Evaluation Campaign on Spoken Language Translation -, IWSLT 2004 Proceedings (2004).

[10] 菅谷史昭, 竹澤寿幸, 横尾昭男, 山本誠一: 音声翻訳システムと人間の比較による音声翻訳能力評価手法の提案と比較実験, 電子情報通信学会論文誌, Vol. J84-D-II, No. 11, pp. 2362-2370 (2001).

[11] Yasuda, K., Sugaya, F., Takezawa, T., Yamamoto, S., and Yanagida, M., “Application of automatic evaluation methods to measuring a capability of speech translation system,” Proc. EACL, pp. 371-378 (2003).

[12] Yasuda, K., Sugaya, F., Takezawa, T., Kikui, G., Yamamoto, S., and Yanagida, M., “An objective method for evaluating speech translation system: Using a second language learner’s corpus,” IEICE Trans. Inf. & Syst., (to appear).

[13] 石川泰, 澤田久美子, 城戸恵美子: 音声インタフェースの評価, 日本音響学会誌, Vol. 61, No. 2, pp. 79-84 (2005).