2005－SLP－57（22）
2005／7／16

音声認識結果と非母語話者の聞き取り能力の比較による音声言語処理システムの性能評価

竹澤寿幸[†]　安田圭志[†]
水島昌英[†]　菊井玄一郎[†]

音声認識出力を非母語話者の聞き取り能力と比較して数量化する音声言語処理システムの新しい評価手法を提案する．提案手法は2種類の数量化から構成される．一つは音声認識結果と非母語話者の聞き取り結果の比較による数量化である．もう一つは音声言語処理システムの性能劣化を数量化するものであり，音声認識結果からの処理出力と非母語話者の聞き取り結果からの処理出力の比較による数量化である．応用システムとして機械翻訳を取り上げ，発話スタイルの変化が機械翻訳の性能のみならず，日本語ネイティブ話者の英語聞き取り能力にも影響を与えることを示す．

Evaluation of Spoken Language Processing Systems by Comparing Speech Recognizer's Output with Non-native Speakers' Listening Capabilities

Toshiyuki Takezawa,[†] Keiji Yasuda,[†] Masahide Mizushima[†] and Genichiro Kikui[†]

We propose a new method of evaluating spoken language processing systems by comparing a speech recognizer's output with non-native speakers' listening results. The proposed method consists of two kinds of measurements. One is a measurement of listening capabilities by comparing the speech recognizer's output with the listening results of non-native speakers. The other is an assessment of the degradation of a spoken language processing system by comparing the processing output from a speech recognition result with processing outputs from non-native speakers' listening results. We employ machine translation (MT) as an application system and show that the change in speaking style degrades not only MT but also the English listening capability of Japanese native speakers.

1. まえがき

音声自動翻訳システムで使われる機械翻訳システムは音声認識結果を入力として受け付けなければならない．現在の音声認識は誤りが避けられないため，本質的に性能劣化が生ずることになる．本稿では，その性能劣化に着目し，人間の言語能力と比較して数量化する音声言語処理システムの新しい評価手法を提案する．

そもそも，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．

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いるとすれば，機械による音声認識を非母語話者の聞き取り能力と比較して数量化することに意味がある．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合わせて使われる．そのような場合には，単に表層的に単語が一致する割合だけではなく，応用システムへ与える影響が重要となる．

そこで，本稿では，音声認識出力を非母語話者の聞き取り能力と比較して数量化する音声言語処理システムの新しい評価手法を提案する．提案手法は2種類の数量化から構成される．一つは音声認識結果と非母語話者の聞き取り結果の比較による数量化である．もう一つは音声言語処理システムの性能劣化を数量化するものであり，

† (株)国際電気通信基礎技術研究所 音声言語コミュニケーション研究所
ATR Spoken Language Communication Research Laboratories

音声認識結果からの処理出力と非母語話者の聞き取り結果からの処理出力の比較による数量化である．

まず，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，様々な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．次に，応用システムとして機械翻訳を取り上げ，発話スタイルの異なる複数のテストセットを用いて実験を実施した．そして，発話スタイルの変化が機械翻訳の性能のみならず，日本語ネイティブ話者の英語聞き取り能力にも影響を与えることを示す．さらに，システムの性能をTOEICスコアに換算することを試みる．

2.で聞き取り能力の比較について述べる．**3.**で応用システムとして機械翻訳を取り上げ，性能劣化に関する評価実験について報告する．**4.**で議論を行い，関連研究について言及する．最後に**5.**で全体をまとめる．

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)~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[3)]．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，日本語発話に言い淀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 (Speech DataBase/TRAvel)[4)]と，人間の逐次通訳者を介した日英対話SLDB (Spoken Language DataBase)[4)]の数値も示す．

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淀みを含む発話の割合が減り，MAD3では人間の通訳者を介する状況に近く，MAD4で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．MAD4の平均発話長はMAD3より短いことがわかる．なお，MAD3とMAD4の実験はいずれも同じ部屋で行われ，収録の条件は共通である．また，話者に与えた課題も同じである．

表1 発話スタイルの違い
Table 1 Evidence of changes in speaking style

言い淀みを含む発話	
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2 英語テストセット
Table 2 English test sets

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

表3 英語音声認識実験結果
Table 3 English speech recognition results

	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%

2.2 英語音声認識

英語音声認識システムは，ATRで研究開発したATRASR[5)]を用いた．言語モデルはATRで構築したコーパス[6)]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．テストセットパープレキシティや未知語率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．

2.3 実験結果

日本人の英語能力をTOEIC[7)]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．

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，例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統一等については整形を行った．その後，

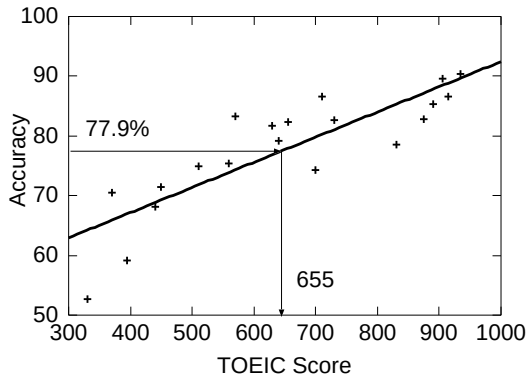

図1 MAD3の英語聞き取り実験結果
Fig. 1 MAD3 English listening experiment

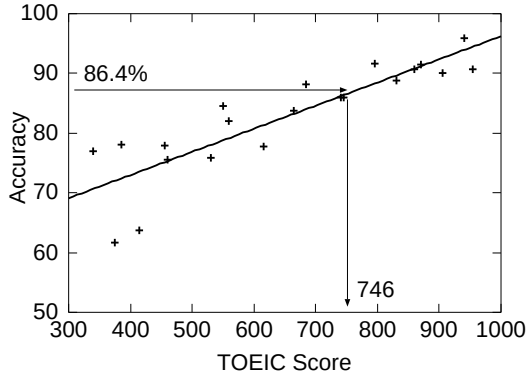

図2 MAD4の英語聞き取り実験結果
Fig. 2 MAD4 English listening experiment

表4 TOEICスコアと聞き取り能力の相関
Table 4 Correlation between TOEIC scores and accuracy

	相関係数
MAD3	0.848
MAD4	0.868

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．

MAD3テストセットとMAD4テストセットの聞き取り実験結果を図1と図2に示す．図の各プロットが，それぞれ被験者に対応する．回帰直線をあわせて示した．被験者のTOEICスコアと聞き取り能力の相関係数を表4に示す．

図1と図2を比較すると，被験者が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
Table 5 Estimated TOEIC scores of the system

	推定TOEICスコア
MAD3	655
MAD4	746

表6 推定TOEICスコアの信頼度区間
Table 6 Confidence interval of the estimated TOEIC scores

	信頼度区間
MAD3	±81
MAD4	±82

2.4 推定TOEICスコアの誤差

被験者iのTOEICスコアをX_i, その聞き取り能力をY_i, 誤差項をε_iとし，母集団が次式の母回帰方程式に従うものとする．

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (i = 1, 2, \cdots, n) \quad (1)$$

ここで，β_0, β_1は母回帰係数であり，nは被験者のサンプル数とする．さらに，誤差項ε_iは次の条件を満たすものとする．

(a) 期待値は0: $E(\varepsilon_i) = 0$
(b) 分散は一定: $V(\varepsilon_i^2) = \sigma^2 \qquad (i = 1, 2, \cdots, n)$
(c) 異なった誤差項は無相関:
$Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0 \qquad \text{if } i \neq j$

文献[8]によれば，この条件の下でのシステムの推定TOEICスコアの誤差の標準偏差σ_{TOEIC}は次式で表される．

$$\sigma_{TOEIC} = \left| \frac{\sigma}{\beta_1} \right| \sqrt{\frac{1}{n} + \frac{(C_0 - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}} \qquad (2)$$

ここで，C_0はシステムの推定TOEICスコア，$\bar{X}$は被験者のTOEICスコアの平均値である．

さらにt分布を用いれば，システムの推定TOEICスコアの信頼度区間CIは次式で表される．

$$CI = [C_0 - I, C_0 + I] \qquad (3)$$

$$I = \sigma_{TOEIC} \times t(\frac{\alpha}{2}; n - 2) \qquad (4)$$

99%の信頼度区間を求めることとし，$\alpha = 0.01$を採用した．このようにして求めた値を表6に示す．

2.5 考　　察

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 (表 4). 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 (図 1と図 2).

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，短く簡潔にという機械を意識した話し方から，言い淀みが多く，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.

3. 応用システムの性能劣化の評価

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割合に関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なく，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[9] 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になるまで， 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．MAD3, MAD4 の両者に対して修正作業を行い，実験を実施した.

3.2 実験結果

修正データに対して，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 (Word Error Rate) を求めた． MAD3 の結果を図 3に， MAD4 の結果を図 4に示す．図の各プロットが，それぞれ被験者に対応する．回帰直線をあわせて示した．被験者の TOEIC スコアと WER の相関係数，音声認識システムの WER, システムの推定 TOEIC スコア，信頼度区間を表 7に示す．音声認識システムの WER (表 7) が表 3の音声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.

この修正データを用いて，機械翻訳実験を行った．一つの翻訳に対して 15 通りの参照訳を準備し，マルチリファレンス単語誤り率 mWER (multi-reference Word Error Rate) を求めた． MAD3 の結果を図 5に，MAD4 の結果を図 6に示す．図の各プロットが，それぞれ被験者に対応する．回帰直線をあわせて示した．被験者 TOEIC スコアと被験者聞き取り結果からの翻訳出力に対する mWER の相関係数，音声認識結果からの翻訳出力の mWER, システムの推定 TOEIC スコア，信頼度区間を表 8に示す.

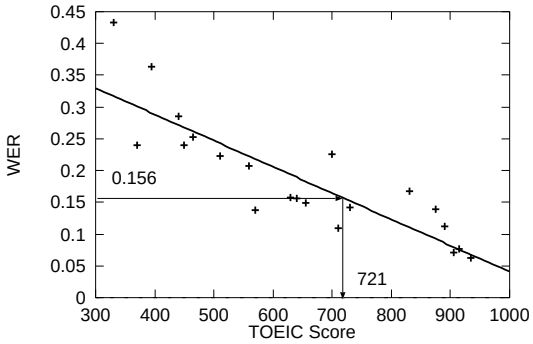

図3 MAD3 修正データによる英語聞き取り誤り率
Fig. 3 English listening experiment using MAD3 refined data

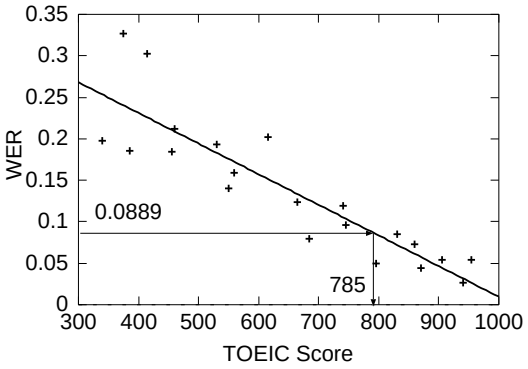

図4 MAD4 修正データによる英語聞き取り誤り率
Fig. 4 English listening experiment using MAD4 refined data

表7 修正データを用いた英語聞き取りの結果
Table 7 Results of English listening experiment using refined data

	MAD3	MAD4
相関係数	−0.858	−0.890
音声認識結果 (WER)	0.156	0.0889
推定 TOEIC スコア	721	785
信頼度区間	±81	±80

3.3 考　　察

英語聞き取り実験について比較すると， MAD3, MAD4 修正データを用いた推定 TOEIC スコアの差 (表 7) は，修正前データを用いた実験で得られた差 (表 5) より小さく，信頼度区間はいずれもほぼ同じである (表

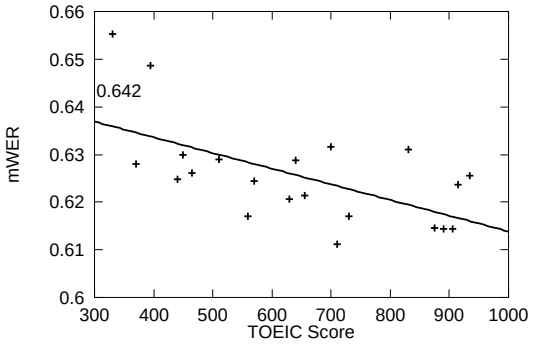

図5 MAD3修正データを用いた機械翻訳実験結果
Fig. 5 MT experiment using MAD3 refined data

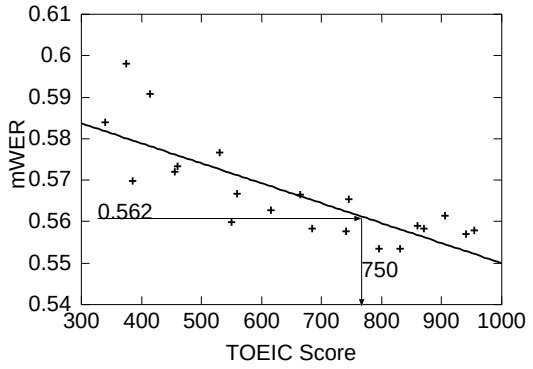

図6 MAD4修正データを用いた機械翻訳実験結果
Fig. 6 MT experiment using MAD4 refined data

表8 修正データを用いた機械翻訳の結果
Table 8 Results of MT experiment using refined data

	MAD3	MAD4
相関係数	−0.600	−0.798
機械翻訳結果 (mWER)	0.642	0.562
推定 TOEIC スコア	132	750
信頼度区間	±477	±109

6と表 7). さらに，表 7によれば，修正データを用いると，相関係数の絶対値，システムの推定 TOEIC スコアともに，以前の結果 (表 4，表 5) よりも，高くなることがわかる．

表 8の機械翻訳実験結果によれば，相関係数の絶対値，聞き取り結果からの翻訳出力に対する mWER ともに，MAD3 より MAD4 が良い．ATR の統計的機械翻訳 SAT が MAD3 より MAD4 の表現を翻訳しやすい傾向をもつことを示唆している．

表 7によれば，MAD4 聞き取り実験から得られた推定 TOEIC スコアは 785, その信頼度区間は ±80 である．表 8によれば，MAD4 機械翻訳実験から得られた推定 TOEIC スコアは 750, その信頼度区間は ±109 である．MAD4 に関するこれらの値はほぼ近い．MAD4 の音声認識率は，表層と品詞を考慮した場合に 85% 以上，表層のみを考慮する条件では 90% 以上である．これらの値は直感的にも良いもので，音声言語翻訳に十分なレベルに達している．

一方，MAD3 聞き取り実験から得られた推定 TOEIC スコアは 721, その信頼度区間は ±81 であるにも関わらず，MAD3 機械翻訳実験から得られた推定 TOEIC スコアは 132, その信頼度区間は ±477 となり，その性能劣化は著しい．MAD3 の音声認識率は，表層と品詞を考慮した場合に約 78%, 表層のみを考慮する条件でさえ約 85% である．これらの値は直感的にも良いものではなく，音声言語翻訳に十分なレベルに達していない．

TOEIC 公式ホームページ[7] によれば，TOEIC スコア 730 点以上の被験者は「どんな状況でも適切なコミュニケーションができる素地を備えている」レベルである．MAD3, MAD4 修正データを用いた英語聞き取り実験から得られたシステムの推定 TOEIC スコアは 730 点に匹敵するレベルである (表 7). そのレベルであって，かつ，MAD4 のように機械を意識した発話スタイルであれば，機械による音声認識性能も非母語話者が聞き取れる量も相対的に良く，機械翻訳の性能劣化の程度も機械と非母語話者で同程度である．しかし，MAD3 のように人間の逐次通訳者に話すような流暢な発話スタイルになり，しかもパープレキシティや平均発話長も大きくなると，機械による音声認識性能も非母語話者が聞き取れる量もあまり良くなく，認識結果が機械翻訳に与える性能劣化の程度は非母語話者に比べて著しく大きくなる．

なお，表 8に示すように，機械翻訳システムそれ自体が MAD3 より MAD4 のような表現を処理しやすい傾向をもつ点に再度注意する必要がある．

4. 議論と関連研究

4.1 議　　論

音声言語処理技術およびシステムの研究開発が盛んになっている．しかしながら，あるシステムの性能を別のシステムと比較するのは一般に容易ではない．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 (例えば国際ワークショップ IWSLT[10]). しかし，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，人間の能力に例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730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．

ただし，MAD3, MAD4機械翻訳結果によれば，音声認識率の値そのものも機械翻訳システムの性能劣化を削減するために重要である．

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[8),11)]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[8)]と，その一対比較作業を自動化する研究[11)]である．本研究は，それを音声認識システムないし英日方向の音声自動翻訳システムに適用したものとみなすこともできるが，聞き取り能力の比較と応用システムへの影響に分けて取り扱った点に特徴がある．

5. むすび

音声認識出力を非母語話者の聞き取り能力と比較して数量化する音声言語処理システムの新しい評価手法を提案した．提案手法は2種類の数量化から構成される．一つは音声認識結果と非母語話者の聞き取り結果の比較による数量化である．もう一つは音声言語処理システムの性能劣化を数量化するものであり，音声認識結果からの処理出力と非母語話者の聞き取り結果からの処理出力の比較による数量化である．応用システムとして機械翻訳を取り上げ，発話スタイルの変化が機械翻訳の性能のみならず，日本語ネイティブ話者の英語聞き取り能力にも影響を与えることを示した．さらに，日本人の英語能力をTOEICスコアで表すこととし，システムの性能をTOEICスコアに換算することを試みた．本手法は，非専門家が音声言語処理システムの性能を把握する際に有益と期待できる．

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 (EUROSPEECH)*, 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. 4th International Conference on Language Resources and Evaluation (LREC)*, 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 (ICA)*, pp. 401–404, (2004).

5) 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, 1-P-30, Vol. I, pp. 221–222, (2004).

6) 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 (EUROSPEECH)*, Vol. 1, pp. 381–384 (2003).

7) TOEIC: 国際コミュニケーション英語能力テスト (Test of English for International Communication), `http://www.toeic.or.jp/` (2003).

8) 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).

9) 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 (TMI)*, pp. 188–198, (2002).

10) IWSLT: International Workshop on Spoken Language Translation — Evaluation Campaign on Spoken Language Translation —, `http://www.slt.atr.jp/IWSLT2004/` (2004).

11) 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.*, Vol. E88-D, No. 3, pp. 569–577, (2005).