音声翻訳システムを介した対話の評価　ー誤認識及び誤訳が対話に及ぼす影響ー

○水島昌英，竹澤寿幸，菊井玄一郎（ATR 音声言語コミュニケーション研究所）

1. はじめに

ATR では，実世界で利用可能な多言語の音声対話翻訳システムの実現を目指して研究を進めている[1]。著者らは，これまで同システムを介した対話音声の収集と同システムの評価を目的とした対話実験を進めてきた[2]。また，収集された対話音声の特徴として，会話文を読み上げた音声に近い発話スタイルになること[3]や，実験前の教示が言語表現に大きな影響を与えること[4]，さらに，相手の意図を理解する際には，合成音声だけではなく，翻訳テキストや，相手話者音声などを総合的に活用していること[5]などを明らかにしてきた。

これまでの実験においては，システムそのものが開発途上であることなどから，良質な対話データ収集のために，音声認識システムの代わりにタイピストによる書き起こし方式を採用していた。音声認識率ほぼ 100%の状態を模擬していたことになる。本報告では，実際に音声認識を導入した音声翻訳システムを介した対話実験において，誤認識及び誤訳がどの程度生じ，またそれらが発話の特性や対話の進行にどのような影響を及ぼしたかについて報告する。

2. 音声翻訳システム

音声翻訳システムを介した対話の仕組み（片側）を図1に示す。話者の音声はまず，直接相手に聞こえる。そして，音声認識システムにおいてテキストに変換され，話者にその結果を表示すると共に機械翻訳システムに渡される。同システムで相手言語テキストに変換され，そのテキストが相手に表示された後，音声合成システムで合成された音声が出力される。

対話ではこの逆側のシステムも必要で，実際の実現形態としては，異なる言語話者が互いに PDA（携帯情報端末）のような小型の通信機器を持ち，相手話者の音声を翻訳した結果を PDA 画面に表示すると共にヘッドホンから合成音声で出力する方式になっている。

現行のシステムにおいては，認識された結果（テキスト）は相手には提示していない。また，翻訳されたテキスト及び合成音声は，発声者には提示されない。さらに，誤認識の訂正のしくみは導入しておらず，誤認識が生じた場合でも，自動的にそれらを含んだテキストが翻訳システムに渡されるようになっている。

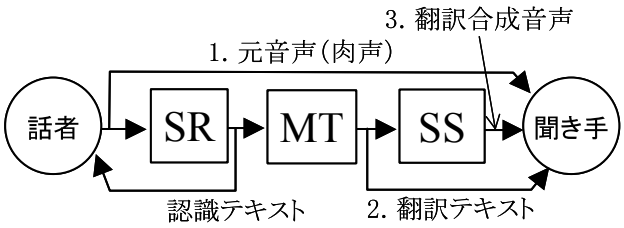

図1　音声翻訳システムの仕組み
SR:音声認識，MT：機械翻訳，SS：音声合成

音声認識には ATRASR[6]，機械翻訳には，句構造等のパターンに基づく翻訳 HPAT[7] と文を単位とする用例に基づく D3[8]に最良翻訳選択器[9] を組み合わせたシステムを使用した。日本語音声合成は XIMERA[10]，英語音声合成には，AT&T Labs' Natural Voices™を使用した。

3. 対話実験の実施

上記音声翻訳システムを使用した日英話者による対話実験を実施した。以下に実験の概略を述べる。

3.1 被験者

日本語話者6名，英語話者6名で，原則的に相手の言語は話せないことを条件としているが，日本語話者は英語教育を受けており，また英語話者は，日本在住の人であることから，ある程度は，理解できると思われる。

3.2 教示

従来の実験と同等な教示をした[4]。

1. 大き目の声で明瞭に話す。
2. 1回の発話は 10 秒以内とする。
3. 時々誤りが発生するが，確認や再発話をすることにより，対話を続ける。
4. 伝達したい情報を分割して，短く簡潔に話す。

3.3 場面と課題の設定

海外，あるいは日本における日本人と英語話者の旅行会話を想定した場面を設定した。例えば，「通りで道に迷い，道順を尋ねる」，「ミュージカルのチケットを買う」などである。なお実際の実験は実験室で実施した。各々の場面に合わせて，被験者には達成すべきいくつかの課題を与えた。例えば，ミュージカルのチケットを買う場面では，お互いにチケット種類，枚数，公演開始時間等を確認し合うことなどである。事前に，与えた課題を達成するために必要な単語や文章表現が認識及び翻訳出来るかどうかを確認している。即ち“正解（あるいは模範）発話”が存在することを確認した上で実験を実施した。無論，被験者にはそれらは伝えておらず，課題遂行に必要な最低限の情報しか与えていない。

4. 発話の特性

音声翻訳システムを介した対話実験においては，認識，機械翻訳結果に含まれる誤りに発話者は影響を受け，発話に変化が生じることが予想される。本節では我々がこれまでに収集した様々な条件下での“音声翻訳システムを介した対話”データと対比させながら，今回の実験における対話音声の特徴について述べる。

4.1 比較する対話データ

(1) **通訳者を介した対話（SLDB）**：一切のシステムを使用せず，通訳者を介して日英話者が対話した実験。

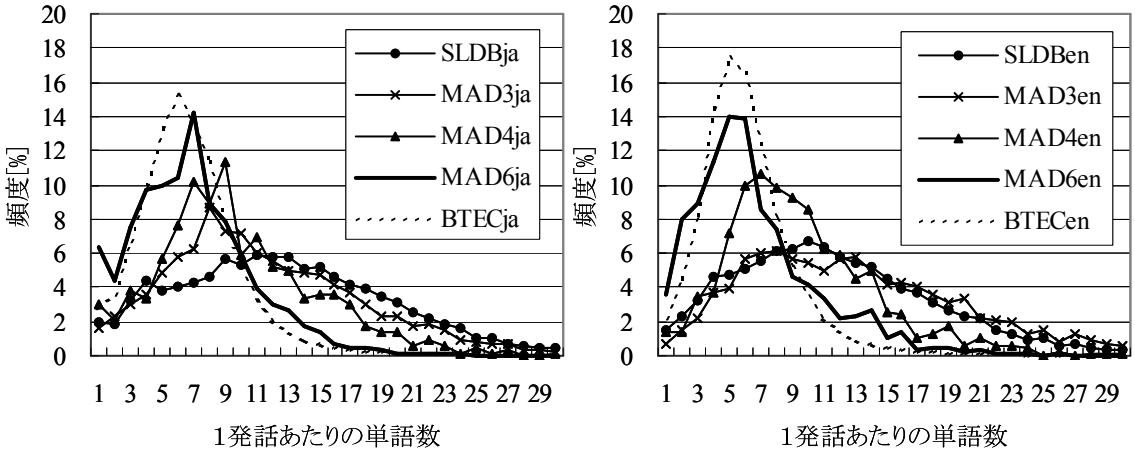

図2　1発話あたりの単語数の分布(左:日本語, 右:英語)

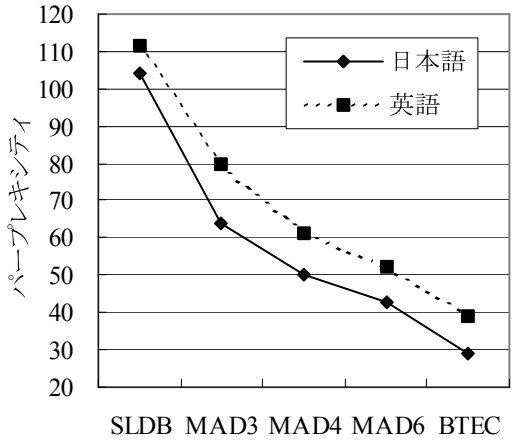

図3　旅行会話例文集とのクロスパープレキシティ

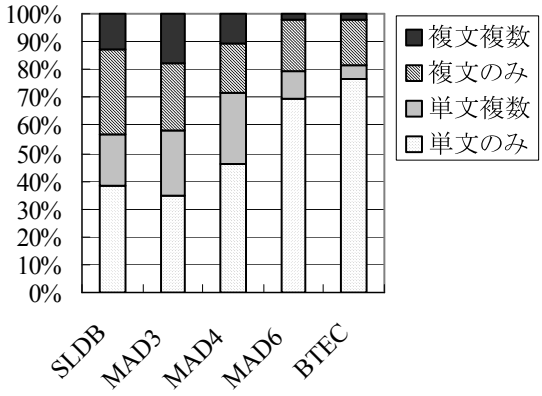

図4　発話に含まれる単文と複文の割合
(日本語側)

音声翻訳システムの理想を模擬した実験である。

(2) 音声翻訳システムにおいて, **認識システムの代わりにタイピストを導入した対話**(MAD3):音声認識率 100%を模擬した実験である。

(3) (2)において, **話者の話し方に制限を加えた対話**(MAD4):事前に「短く簡潔に話す」という教示をすることで, 翻訳結果向上を目指した実験である。

(4) (3)において, **音声認識システムを導入した対話**(MAD6):本研究が対象とした実験である。

(5) **旅行会話例文集**(BTEC):主に旅行者を対象とした対訳英会話例文集にあるような簡単な文で, これらを読み上げた音声をリファレンスとする。

4.2 発話の長さ

図2は, 各対話データにおける1発話あたりの単語数のヒストグラムである。日英でその傾向はほぼ同じになった。「短く簡潔に」という教示で確かに短く話すようになり(MAD4), 音声認識を取り入れた実験 (MAD6)では, 更に短くなり, 旅行会話例文集(BTEC)とほぼ同等の分布になった。

4.3 パープレキシティ

図3は, BTEC とのクロスパープレキシティである。比較のために, BTEC 自体のテストセットパープレキシティも入れてある。MAD6 では文章が短くなるだけでなくBTEC により近い表現を多用していることがうかがえる。

4.4 単文と複文の割合

図4は, 日本語側の発話に含まれる単文と複文の割合を示している。分析手法には丸山らの方法[11]を用いた。単文複数とは二つ以上の単文で発話が構成されている割合, 複文複数とは, 少なくとも一つの複文を含む複数の文で構成された発話の割合である。MAD4 で複文が減り, MAD6 では更に減ると共に, 1発話を一つの単文で済ませる割合が急激に増え, ここでも MAD6 は BTEC に近い分布になっている。

4.5 自然発話率(発話スタイル)

図5は日本語側発話の自然発話率を示している。自然発話率とは, 音声認識時に自然発話と朗読発話の音響モデルを発話単位に最尤選択させ, 自然発話音響モデルが選択された発話の頻度のことで, 話者の発話スタイルの目安となる[3]。音声認識を導入することで, ほぼ会話文読み上げと同等の発話スタイルになる。

5. 誤認識と誤訳の発生率

少なくとも現在の音声翻訳システムにおいては誤認識や誤訳を避けることは出来ない。また, 現行の我々のシステムは認識結果に誤認識が含まれていてもその訂

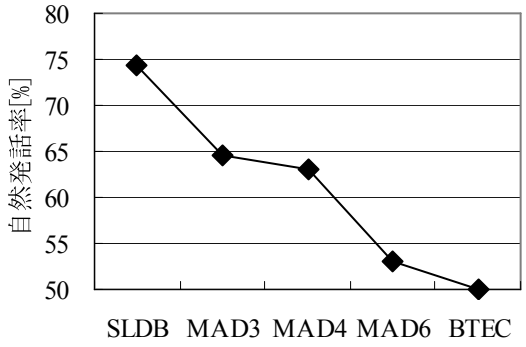

図5 自然発話率(日本語側)

正はせずに，翻訳システムに送るため，翻訳結果には，その影響も含まれる。本節では MAD6 実験時のこれらの結果について述べる。

5.1 音声認識率

表1に実験時の音声認識率を示す。なお音響モデルは，話者適応したモデルを使用した。

5.2 誤訳率

表2は実験時の誤訳率を表している。ここでいう誤訳率とは，情報の過不足を伴う誤訳が含まれる発話の頻度である[5]。一箇所(1単語)でもあれば誤訳としてカウントしている。誤訳の有無は，実際の発話に対して判断するため，誤認識が起きれば誤訳しやすくなる。また，発話長が長ければそれだけ誤訳も起こりやすいと予想されるため，それらの関係で分類した。発話の長さは，1発話の単語数が平均単語数より多いか少ないかで分けた。また，正認とは発話に一つも単語誤りが含まれない発話である。表から明らかなように，認識誤りがあると，発話長の長さに関わらず，7割程度の発話で誤訳が生じている。一方，認識誤りがない場合には，短い発話では1割前後の誤訳率だが，長い発話になると4分の1の発話で誤訳が生じたことが分かる。

6. 誤認識・誤訳に対する話者の対応

音声翻訳システムが誤認識や誤訳をすると，話者はそれに対応しようとする。本節では話者がどのように対話を進めているかを誤認識や誤訳が生じた直後の話者の発話に着目して分析を試みた。

6.1 連続再発話

発話者は誤認識の有無を確認することができるため，必要に応じて同じことを繰り返したり言い方を変えたりして再発話することが出来る。表3は，誤認識後に話者が連続再発話をした頻度を示している。日英で倍の差があるが，全体では1割強でほとんどの場合，再発話をせずに相手に発話ターンを渡していることが分かる。

6.2 誤訳に対する話者の対応

翻訳結果に誤訳が含まれているかどうかは，被験者には陽には分からないが，意味が通じない表現であったり，話の流れから外れた内容だったりすると，被験者は対話を進めるために様々な対応を取る。図6は誤訳を含む発話に対して相手話者がどういう対応をしたかを

表1 実験時の音声認識率

単語正解精度[%]		発話正解率[%]	
日	英	日	英
88.3	85.1	67.8	64.5

表2 誤認識の有無と発話長と翻訳精度の関係

	発話長短い		発話長長い		全体	
[%]	日	英	日	英	日	英
誤認誤訳率	69.8	69.0	75.2	67.4	73.4	68.0
正認誤訳率	12.4	7.3	28.4	23.6	19.6	12.3
全体誤訳率	24.8	22.6	48.0	45.9	36.9	32.1

表3 誤認識後の連続再発話

	連続再発話数	誤認識数	連続再発話率[%]
日本語	60	403	14.9
英語	29	411	7.1
全体	89	814	10.9

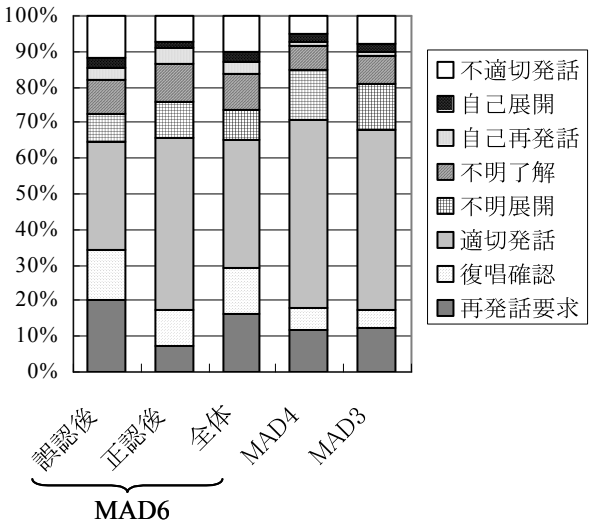

図6 誤訳後の相手話者の対応

表したグラフである。対応は以下の8種類に分類した。

適切発話:相手の発話に対して適切と思われる発話。

不適切発話:不適切と思われる発話。

再発話要求:相手の意図を汲み取れなかった場合等に再発話を要求する発話(「もう一度お願いします」等)。

復唱確認:相手の意図を汲み取れなかった場合等に，内容を確認する発話(「一万円の席を二枚ですか」等)。

不明了解:単に「分かりました」等だけの発話で，第三者からは正しく伝わったか分からない発話。

不明展開:話の流れとして不自然ではないが，第三者からは正しく伝わったかどうかが分からないままに次の話題に切り替える発話。

自己再発話:相手の発話内容に関係なく，先の自分の発話内容を繰り返す発話。

自己展開:相手の発話内容に関係なく，話の流れから不自然に話題を切り替える発話。

左側から3本が MAD6 の結果で，誤認識があった場合となかった場合とそれら全体の平均が示してある。ま

ず，タイピストを使うことで認識率がほぼ 100%であった MAD3,4 では MAD4 で不適切発話が減っていること以外，ほとんど分布に違いがない。不明了解，不明展開も含めれば7割以上は“対話が進行する発話”であり，残りの2割弱が確認発話（再発話要求と復唱確認），1割弱が不適切な発話となる。自己再発話や自己展開はその残りで数は少ない。MAD6 の認識が正しかった場合を比較してみると，大きな分布傾向は同じだが，確認発話の中で復唱確認が多いこと，不明発話のうち，不明展開がやや少ないことが特徴と思われる。また，自己再発話がやや多くなる。一方誤認識後の誤訳に対しては，不適切な対応が増えると共に確認発話が大幅に増え，対話が進む割合は5割弱となる。

7. 考察

4節から明らかなように，音声認識を導入した音声翻訳システムを介した対話における発話は，旅行会話例文集にあるような短い文章を読み上げた発話に非常に近くなった。この傾向は音声認識の代わりをタイピストがしていた時からあった[3]が，誤認識が生じる本来のシステムを導入するとその傾向がさらに強まった。認識結果は発話者に表示されるため，そのフィードバック効果が働いた上に，5,6 節で述べたように誤認識を生じると，誤訳を生じる確率が高く，その誤訳結果を見た相手から確認発話や不適切な発話が返ってくる確率も高まることから，より誤認識させないように短く丁寧に発話するようになると考えられる。

表2から明らかなように，誤認識があると誤訳が生じる確率は格段に高まるが，逆に誤認識がなければ誤訳率は全体で2割弱，かつ簡潔に発話すれば，1割前後と低くなる。しかも，誤訳が生じた場合でも，誤認識がない場合には，7割は対話が進行する対応が出来ている（図6参照）。一方，誤認識後の誤訳に対しては，再発話要求は倍以上になっており，適切な対応が大幅に減っている。音声認識は，“音響的な近さ”が大きな探索基準になることから，誤認識の際には実際の発話とは意味的にかけ離れた認識結果となることが多く，翻訳すると元の単語の推測がほとんど不可能になるのであろう。短く簡潔に話し，且つ生じた誤認識をなんらかの方法で訂正することが出来れば，よりスムーズなコミュニケーションが実現できる可能性はある。

8. まとめ

音声認識，機械翻訳，音声合成を組み合わせた音声翻訳システムを介した日英話者による対話実験を実施し，システムの性能が対話にどのような影響を与えるかについて調べた。その結果，発話の特性は，旅行会話例文集から集めたコーパスの特性とほぼ同等になった。また誤認識を伴わない誤訳に対しては，約7割は適切に対話を進行出来るものの，誤認識後の誤訳に対しては，相手に意図を聞き返す再発話要求や相手の意図を取り違うなどの不適切は発話が大幅に増え，対話の進行が困難になることが分かった。

今回の実験は，実験室の中で，しかも限られた場面と課題設定をした上で実施している。また被験者に事後インタビューをした結果，特に日本語話者の半分は相手話者の英語の大半を理解しており，それでも出来る限り翻訳結果に対して反応しようとしたと答えている。今回の結果には，これらのバイアスがかかっていることは付け加えておかなければならない。しかし実際の場面では，自分の持つ英語能力と音声翻訳機双方を活かしながら会話をするのが現実的とも考えられ，それらも考慮したより実際の利用形態に即した評価実験の設定が今後の課題の一つである。また，我々の音声翻訳システムは発展途上であり，今後システムの性能が向上した場合，それが対話にどのような変化をもたらすのか，継続して調査する予定である。

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] Yamamoto, S. “Toward Speech Communications Beyond Language Barrier - Research of Spoken Language Translation Technologies at ATR -,” Proc. of ICSLP2000, pp.406-411, 2000.

[2] Takezawa, T. and Kikui, G., “Collecting Machine-Translation-Aided Bilingual Dialogues for Corpus-Based Speech Translation,” Proc. of EUROSPEECH2003, pp. 2757-2760, 2003.

[3] 水島，竹澤，菊井，“翻訳システムを介した対話音声の発話スタイルについて-自然発話，朗読発話との関係-”，第 3 回話し言葉の科学と工学ワークショップ講演予稿集，pp.135-142, 2004.

[4] Takezawa, T. and Kikui, G., “A Comparative Study on Human Communication Behaviors and Linguistic Characteristics for Speech-to-Speech Translation,” Proc. of LREC2004, pp.1589-1592, 2004.

[5] 水島，竹澤，菊井，“翻訳システムを介した音声対話における相手話者音声と翻訳テキスト表示の影響について”，SLP-52,pp.99-106,2004.

[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] Imamura, K., “Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment for Pattern-based MT”, Proc. of TMI-2002, pp74-84, 2002.

[8]Sumita, E. “Example-based machine translation using DP-matching between word sequences, Proc. of ACL-2001, pp.1-8, 2001.

[9] Akiba, Y.,Watanabe, T., and Sumita, E., “Using Language and Translation Models to Select the Best among Outputs from Multiple MT systems”, Proc of COLING-2002, pp.8-14, 2002.

[10] Kawai, H., Toda, T., Ni, J., Tsuzaki, M. and Tokuda, K., “XIMERA: A New TTS from ATR Based on Corpus-Based Technologies”, 5th ISCA Speech Synthesis Workshop, 14th-16th June 2004, Carnegie Mellon University, 2004

[11] 丸山，柏岡，熊野，田中，“節境界自動検出ルールの作成と評価”，NLP2003，pp.517-520,2003.