「聞く」と「見る」における言語理解の違い
―話者の意図が聞き手の理解に与える影響―*

○高椋琴美，谷田泰郎（シナジーマーケティング株式会社）

1　はじめに

人はよく他人と話をするが、結構いい加減に聞いているものである。一言一句聞き漏らすまいと聞いてはみてもなかなか難しい。長い時間真剣に話した後でも、その内容を思い返すと案外概要くらいしか覚えていないものである。我々の脳は詳細より概要を捉えようとするが、逆にコンピューターでは物事を詳細に処理しがちである。しかし人間のコミュニケーションに関する事をシステムで処理する場合、果たして全てに詳細な処理が必要なのだろうか。我々人間の様ないい加減なアプローチも考えられるのではないだろうか。

我々の研究グループは、コミュニケーション研究、中でも音声対話から人の「心のモデル」のデザインを考えようとしている。そこでまず手始めに“コミュニケーションのいい加減さ”に着目した。

現在、発話内容を理解するのに自然言語処理が用いられているが、テキストから話者の意図や感情などを推定するのには限界がある。音声の持つ非言語情報にはこれらを知るヒントとなる声質や抑揚、「間」など重要な情報が含まれている。例えば[1]では話し言葉における「間」について様々な知見が述べられている。中でも前後の「間」は重要性の判断に関わっており、特に後の「間」が若干長いと重要と判断される率が高いと示されている。この様な音声に含まれる非言語情報を使って話者が重要だと伝えている情報を取得し、より話者の意図を汲み取った内容理解へと繋げたい。

我々は先行研究[2]で小説を題材に、内容を「聞く」場合と「見る」（黙読）場合とで理解する（ここでは記憶に残る）情報にどのような違いがあるのかについて調べるため実験を行い次のような結果を得た。題材の内容が叙述的であれば「見る」方が記憶量が多く，感情的であれば「聞く」方が記憶量が多かった。被験者の記述した記憶内容を見ると、「聞く」場合に言葉の置き換えがよく起こっていた。性別にみると、相対的に女性の方が男性より記述量が多く、「聞く」の場合の方がその傾向が強かった。

[2]では言語特徴からの分析を行っているが、本稿では朗読音声の音響特徴に注目して「聞く」場合と「見る」場合の違いについて分析した。また、朗読音声の音響特徴量を用いた重要箇所抽出を試みたのでその結果を報告する。

2　視聴覚実験の概要

「聞く」と「見る」における言語理解の違いを調べるため、被験者に小説の朗読を聞いたり、テキストを読んだりしてもらい、その後記憶に残ったものを書き起こしてもらうという簡単な視聴覚実験を行った[2]。

具体的には、夏目漱石の小説「こころ」(「上三」の章の前半 4 段落及び「下四十五」の章の前半 3 段落）の朗読音声[3]を「聞いた」後とテキストを「見た」後で内容を記述してもらった。約 900 文字、朗読では約 3 分半の量である。被験者にはなるべく単語ではなく文章にして他人にあらすじを説明するようなつもりで書くように指示した。「上三」は主人公が「先生」と海水浴場で話す機会を得るまでの話で風景描写などの叙述的な記述が多い場面である。「下四十五」は若い頃の「先生」が友人「K」を出し抜いて「お嬢さん」と結婚するためにその母親である「奥さん」に結婚を願い出る場面で、会話が多く感情的な記述が多い。

Table1 に示すように、全被験者 21 人（男性 11 人、女性 10 人、年齢は 19 歳から 52 歳までの平均 29 才、題材の小説を知らない、読んだことのない人が対象）を午前と午後の 2 部に分け、午前の部は「上三」をテキスト、「下四十五」を朗読、午後の部は「下四十五」をテキスト、「上三」を朗読というようにテキス

---

* Difference of understanding by reading or listening. Influence that speaker's intention gives in listener's understanding, by TAKAMUKU, Kotomi and TANIDA,Yasuo (Synergy Marketing, Inc.)

トと音声の実験素材を午前と午後で逆にすることで、同じ素材でのテキストと朗読音声の実験結果の比較ができるようにした。午前、午後ともテキスト実験、朗読実験の順に行った。テキストの見取り、朗読の聞き取り実験をそれぞれ 3 回ずつ行い、1 回ごとに記憶している内容を書き起こしてもらった。2 回目以降はそれ以前に自分が書き起こしたテキストを見ずに作業してもらった。書き起こしテキストとともに、1 回ごとに体感的な指標として、どれぐらい理解できたか、どれぐらいイメージを書き出せたか、その回の作業に対するコメント・感想を聴取した。書き起こし作業の時間は最大 20 分で設定し、全員の書き起こし作業が終わったところで切り上げた。表に示したように 1 回目が 11 分程度、2 回目が 15 分程度、3 回目が 18 分程度で終了している。また、朗読音声の長さが 3〜4 分であったため、テキストを読む時間を 3 分とした。これらの書き起こし実験回数や時間設定は予備実験を 3 人に対して行った体感値で定性的に決定した。

Table 1　実験の条件など

	被験者数（男性:女性）	書き起こし回数	テキストを「見る」		朗読を「聞く」	
			実験素材	書き起こし時間	実験素材	書き起こし時間
午前	11人（6人:5人）	1回目	上三	11.5分	下四十五	11分
		2回目		15分		15分
		3回目		18分		17分
午後	10人（5人:5人）	1回目	下四十五	11分	上三	11分
		2回目		15分		15分
		3回目		18.5分		18分

3　視聴覚実験の結果と考察

まず「聞く」と「見る」における記憶量や記憶箇所の違いについて比較を行い、次に朗読音声の音響特徴量を用いて重要箇所を抽出した場合の性能について評価した。

重要だと感じた部分は記憶に残りやすいと仮定し、「聞く」「見る」それぞれに書き起こし 1 回目の実験で記憶した部分が被験者が最も重要だと感じた部分であると考え、1 回目の実験結果のみを用いることとした。

3.1 「聞く」と「見る」の記憶の比較

テキスト及び朗読の各実験において被験者の記憶量と、記憶箇所の違いについて調べた。「上三」及び「下四十五」のテキストを意味単位（文章より短く、単文の単位に近い）に恣意的に分割し、書き起こされたテキストにその内容を表す重要なキーワードが含まれていればその意味単位を記憶しているとして被験者ごとの記憶量（＝記憶した意味単位数÷全意味単位数）を算出した。算出した記憶量を平均した結果を Table2 に示す。

Table 2　被験者ごとの記憶量の平均

	聞く	見る
上三	40%	45%
下四十五	33%	21%

「上三」では「見る」の記憶量が多く、逆に「下四十五」では「聞く」の記憶量が多かった。全体としては 20〜45%程度の内容を記憶していた。

次に 記憶度（＝ある意味単位が何%の被験者の記憶に残っていたか）を算出し、「聞く」と「見る」の記憶度の相関を調べた。PEARSON の相関係数で「上三」では 0.81、「下四十五」では 0.70 であった。また定性的ではあるが朗読を聞いて強調していると思った箇所は 1 、それ以外は 0 として「聞いた」場合の記憶度との相関を見ると、相関係数で「上三」は 0.06、「下四十五」は 0.37 で、「下四十五」の方が朗読の強調箇所と記憶度との相関が高かった。これらから「下四十五」の方が朗読の影響を受けていると考えられる。

3.2 音響特徴量を用いた重要箇所抽出

音声を解析して得られる音響特徴量を用い、人の記憶に残る箇所（＝重要箇所）の抽出が出来るかどうかについて検討した。

3.2.1. 発話区間と音響特徴量の取得

まず音声分析ソフト Praat(version 5.4.09) [4] を用い、朗読音声を発話区間と無音区間とに分類した。無音区間の条件は、最小ピッチ：10Hz／無音閾値(dB) : -25／無音時間 : 0.75 以上／音声最小時間：0.1 とした。

抽出した発話区間と 3.1 の意味単位とで区切りが違う箇所があったため、次のルールで調整を行った。①複数の意味単位と発話区間が紐付く場合は一番高い記憶度の意味単位を採用する。②複数の発話区間と意味単位が紐付く場合は一番重要な単語を含む発話区間を採用し、前後の「間」は意味単位の区切りで取得する。

発話区間と意味単位の調整後、Praat を用いて発話区間の音響特徴量（Pitch、Intensity、HNR、Jitter、Shimmer）を取得し、加えて話速（発話区間のモーラ数÷発話区間の秒数）と発話区間の前後の「間」（無音区間の秒数）を計算した。これらの特徴量とその発話区間の記憶度の相関から、今回は「Pitch の最大値・Pitch の標準偏差・Intensity の最大値・話速」の 4 つの音響特徴量を採用した。話速は記憶度と負の相関があり、遅い方が記憶度が高かった。[1]の知見から前後の「間」も用いようとしたが、題材が小説の朗読であるため、段落を区切る「間」、セリフの前後の「間」重要箇所を表す「間」など複数種類の「間」が混在していたため記憶度との相関が出なかったと思われる。簡単かつ自動的に重要箇所を抽出することを優先するため今回は「間」の採用は見送った。

3.2.2. 発話区間の得点と記憶度

音響特徴量からみて強調されている発話区間を抽出するため、発話区間毎に得点を付けた。まず 3.2.1 で採用した各音響特徴量にそれぞれ閾値を設定し、発話区間の音響特徴量毎に、閾値を超えると 1 点、それ以外は 0 点として得点を求め、それらを合算して発話区間の得点とした。音響特徴量の閾値は、各音響特徴量の分布から決定した（Table3）。

Table 3　採用した音響特徴量と閾値

採用した音響特徴量	閾値
話速	≦6
pitch_max	＞350
pitch_std	＞55
intensity_max	＞85

次に実験素材ごとの記憶度平均を Table4 に示す。ここでは誰も記憶していなかった発話区間は除外して計算している。

Table 4　素材別記憶度平均

	聞く	見る	総平均
上三	48%	54%	51%
下四十五	39%	29%	34%
全体	43%	42%	43%

全体での記憶度の総平均は 43%であった。

素材や「聞く」「見る」別にみると記憶度平均に差が見られた。特に「下四十五」の「見る」は、Table2 に示した様に記憶量が少ないにも関わらず、記憶箇所がバラついていた。「下四十五」は「見る」より「聞く」の記憶度が高く朗読の影響を受けていると考えられる。

次に発話区間の得点と朗読の場合の記憶度平均との関係を Fig.1 に示す。（得点 3 以上の発話区間が少なかったため得点 2 以上までを表示している。）

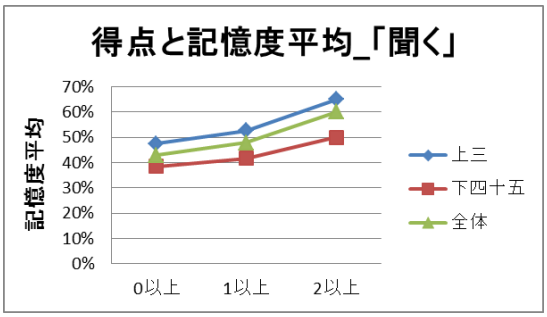

Fig. 1　発話区間の得点と記憶度平均

閾値の得点が高いほど記憶度平均も高くなり、音響特徴量を用いて人の記憶に残りやすい文章を選択できていると言える。

3.2.3. 重要箇所抽出システムの性能評価

3.2.2 を重要箇所抽出システムとしてみた場合、その性能が人間と比較してどの程度なのかを評価した。ここでは発話区間の得点の閾値を 2 以上として発話区間の抽出を行った。

被験者 11 人に重要箇所抽出システムを加えた 12 人を被験者とし、各々の被験者が記憶していた意味単位ごとに、他の 11 人の被験者の内何人が記憶していたかを一致率として算出し質的な指標とした。また全発話区間中いくつ抽出したかを抽出率とし量的な指標とした。これらを「上三」と「下四十五」それぞれで求め被験者毎にプロットしたものを Fig.2 に示す。〇で囲っているのが重要箇所抽出システムである。

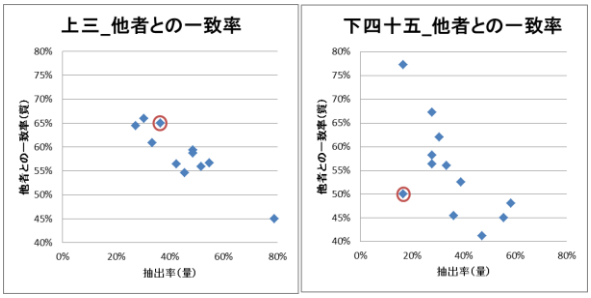

Fig. 2　抽出率（量）と他者との一致率（質）

Fig.2 から人間の能力にもバラつきがあるのが分かる。重要箇所抽出システムもそのバラつきの範囲内にプロットされており人間と比べても遜色ない性能であると言える。

4　おわりに

本研究では、「聞く」と「見る」における言語理解の違いについて実験と分析を行った。約 900 文字、朗読にして約 3 分半程度の文章では大体内容の 20～40%を記憶していた。朗読の場合はテキストより言葉の置き換えも多くみられた。割といい加減である。性別にみると、女性の方が「聞く」の影響が大きく、男性より話者の意図を汲み取って理解しているとの結果を得ている[2]。素材別にみると、3.2.2 より「上三」の叙述的な内容では、「聞く」と「見る」の記憶度平均に大きな差はなかった。また音響特徴量から抽出した重要箇所と、多くの人が記憶していた箇所が一致している場合が多かった。「上三」では内容から重要と判断して記憶に残ったのか朗読の影響なのかは分からなかった。朗読者が内容から重要だと考えた箇所を強調して読んでいたとも推察される。一方「下四十五」では「聞く」と「見る」の記憶度平均に差が見られ、「見る」より「聞く」方が記憶箇所に共通性があった。特に「下四十五」の「差し上げるなんて威張った口の利ける境遇ではありません。」と言うセリフは「見る」では誰の記憶にも残っていなかったが、「聞く」では 73%の人の記憶に残っていた。「下四十五」の方は朗読の影響を受けていると言える。音響特徴量からみると強調している箇所が少なかったせいかもしれない。また「下四十五」の中で最も重要であると思われる「お嬢さんを下さい」という結婚を申し込むセリフは抑えた表現で読まれており、音響的な特徴は出ていなかった。しかしほぼ全員の記憶に残っており、内容から重要と判断された例と言える。「下四十五」は情緒的な内容で、読み方を抑えても十分内容が伝わるものであった。あえて抑えて読むのは朗読者のテクニックなのかもしれない。

重要箇所抽出システムとしてみた場合、用いた音響特徴量は少なく、仕様としてはごく単純なものであった。実際に運用する場合には、特定話者向けのチューニングや、対話や講演など発話スタイルや場面による違い、「間」などのその他知見を入れるなど改善可能な点は数多くある。にも関わらず、人間と比較しても遜色のない結果が得られた。音声の非言語情報が重要箇所の抽出に有用であると言えるだろう。

本稿では話者の意図（ここでは話し方）が聞き手の理解に影響を与えていることを確認できた。更に音響特徴量が話者の意図理解のヒントになるとの結果も得られた。言語特徴を利用した重要箇所抽出技術に加え、音声に含まれる非言語情報からの重要箇所抽出技術を用いることで、話者の意図の理解や、要約システムの精度向上に活用できると考えている。今回は言語特徴からの重要箇所抽出技術との性能比較は行っていないが、今後比較を行いたい。また[2]の実験終了後、被験者にフリーアンサーでのアンケート（読書についてや、読むこと・聞くことに関して得意不得意があるか等）や、被験者の年齢、性別、3 種類の価値観（[5]Societas・[6]Schwartz・[7]Big5）についてのアンケートも行っているがまだ手をつけていない。異なる感覚からの情報をどのように“いい加減”に処理して理解しているのか、個人差は何に関係するのかなどの観点から分析を行い、コミュニケーションの研究に生かしていきたい。

参考文献

[1] 中村敏江,コミュニケーションにおける「間」の感性情報心理学, 音声研究 13(1), 40-52, 2009.

[2] 谷田他,いい加減な対話からの心のモデルの抽出,人工知能学会, 2015 年度人工知能学会全国大会 （第 29 回）

[3] 夏目漱石（著）岩崎さとこ（朗読）,こころ,2013,ことのは出版

[4] Praat: doing phonetics by computer http://www.fon.hum.uva.nl/praat/

[5] 谷田泰郎, 価値観マーケティングと社会知ネットワーク,人工知能 9 月号 Vol.29 No.5 P456-463

[6] Schwartz, S. H. An overview of the Schwartz theory of basic values. Online Readings in Psychology and Culture, 2(1), 2012

[7] Daniel Nettle（原著）, 竹内 和世 (翻訳) パーソナリティを科学する―特性 5 因子であなたがわかる,2009, 白揚社