

## これってヒトの声？

### —読み聞かせ場面を想定した人間の声と合成音声の印象比較—

### Is this a real person talking?

## Comparison of impressions of human voice and synthesized voice for a read-aloud situation

持舘 康太<sup>†</sup>, 友野 貴之<sup>†</sup>

Kota Mottate, Takayuki Tomono

<sup>†</sup>札幌学院大学

Sapporo Gakuin University

F220160@e.sgu.ac.jp

### 概要

本研究では、人間の声による読み聞かせと合成音声による読み聞かせの印象を比較した。その際、両音声に含まれる「人間らしさ」とは何かを具体的に記述し、合成音声における「人間らしさ」が読み聞かせ場面に適しているのかを明らかにすることを目的とした。

実験の結果、読み聞かせ場面では、合成音声よりも人間の声の方が適していると感じられることがわかった。また、合成音声の「人間らしさ」を評価する際には、「生きていると感じられるかどうか」、「感情があるかどうか」、「肯定・芸術・弾力性」といった概念が関係していることが示された。

キーワード：合成音声 (synthesized voice), 人間らしさ (human-likeness), 読み聞かせ (read-aloud situation), 印象評価 (impression evaluation)

### 1. 問題と目的

「合成音声」とは、人工的に作られた音声言語のことを指し、「機械音声」「人工音声」とも言い換えられ、幅広く世の中に普及している。

向他 (2023) は、共働き世帯の増加により、絵本の読み聞かせが難しいことから、感情を込めた合成音声の生成を試みており、「喜び」の感情を込めた合成音声が開き取りやすいことを明らかにした。高橋他 (2020) は、スマートフォンや車などに搭載された音声アシスタント等をエージェントと呼び、そのエージェントの音声に対して抵抗感を覚える人がいることに触れ、エージェントの表情や BGM によって社会的存在感を高めることで、エージェントの「人間らしさ」が高まることを示した。このように合成音声の研究は、「人間らしさ」を高めることを目的としている。

それでは、「人間らしさ」とは一体どのような概念であろうか。機械と人間を判断するテストとしてはチューリングテストが知られている。Devlin et al., (2021) は、「人間のようなナビゲーション」を評価する指標と

して、案内の成功率だけでなく、チューリングテストを行った上での振る舞いの人間らしさといった観点から評価した。Wang et al., (2025) は、評価の一貫性を保つために、音声人間か機械かを判別するといったシンプルな評価を行う指標を提案した。また、山本・片上 (2020) らは、直接人と話すことに苦手意識を持っている人と同じように、ビデオチャットでも同様の苦手意識を持つ人がいることを想定し、どのようなアバターを参加者が選択するかを調べ、結果として人間と非人間アバターの間に差はなかったものの更なる研究の必要性について言及している。

しかし、これらの研究では、そもそも「人間らしさ」とはどのように定義されうるかという部分を深く追求してはいない。当たり前になっている「人間らしさ」を問い直すことで、機械が「人間らしさ」を持っていた場合に、どのように「人間らしい」のかを考えることにつながる。また、「人間らしさ」を持っている機械が、本来人間が行っている読み聞かせ場面で使用することに適しているのかを考えるためにも、「人間らしさ」を考えることは必要なことといえる。

そのため、本研究では「人間らしさ」とはどのようなものであるのかを、人間の声と合成音声の印象比較を通じて考える。本研究の意義は、「人間らしさ」とはどのような概念かを記述することで、機械の「人間らしさ」を評価する際に具体的な指標とすることができる点と、読み聞かせ場面でどのような合成音声が適切であるのか「人間らしさ」の観点からを判断することが可能になる点である。

### 2. 方法

実験では、AquesTalkPlayer Ver.1.0.1.1 を用いて、音声を作成した。棒読み機能を起動すると、読み上げ

られた音声のアクセントすべてが無視される。合成音声構成するフォルマントにはアクセントの影響もあることから、棒読み条件も加えた。本研究では、アクセントが抑えられた棒読み条件の方が通常の合成音声よりも人間らしさを感じないと仮定する。実験の条件は、①人間の声、②合成音声、③棒読み条件の3つであった。

小林・倉片 (2023) の研究では、女性の声の方が効果は小さいものの、了解度と印象評価の結果が高かったと報告している。そのため、女性の声の高さを基準音声とした。保育士経験のある女性に音声の吹き込みを依頼した。

### 【実験方法】

A 大学の学生 11 人が実験に参加した。音声は 30 人程度を収容できる部屋においてノート PC から音声を再生した。実験に際して、静かな環境において通常の音量が聞き取れるかを事前に尋ねた。しかし、母集団を特定の聴力に限定していないため、特にその質問によって参加の可否を決定しなかった。

参加の有無によって不利益が生じないこと、倫理的な配慮についての項目と、実験中に気分が悪くなった場合には退出可能であることを事前に参加者に伝えた。

無料 e 童話の「オオカミと七匹の子やぎ」を読み上げる文章として採用した。Python のライブラリ ML-Ask を用い、感情分析を行った。生態学的妥当性を高めるために、感情価が表示された次の 3 文を用いた。

1. 「こわいオオカミがいないので、ヤギたちは安心してくらししました。」
2. 「そこへ帰ってきたお母さんやぎはびっくり。」
3. 「そして池の水を飲もうとしたとたん、ザブーン！と、お腹の重みで落ちて沈んでしまいました。」

印象評価として、質問紙による音声の評価と SD 法を用いた。質問紙の設問として、向他 (2023) のアンケートの設問 5・6・7 を使用した。これらは、子どもに対しての読み聞かせを想定した設問である。「感情表現が豊かである」、「音声が聞き取りやすい」、「子どもの読み聞かせに適している」の 3 文を 5 件法で尋ねた。また、岩宮他 (1992) の音声評価に関する SD 法の音印象表現語を 5 件法によって用いた。

実施の順序は人間条件、合成条件、棒読み条件の順であった。

### 【解析計画】

フリードマン検定を用いた。フリードマン検定は頑健性が認められているため、事前にサンプルサイズの想定を行わなかった。帰無仮説は、「合成音声棒読み条件、合成音声条件、人間の音声条件の 3 水準での印象には違いがない」である。

順位尺度の一要因分散分析（フリードマン検定）を、音声を要因として、合成音声棒読み条件、合成音声条件、人間の音声条件を 3 水準とし、それぞれの設問（SD 法を除く）で、jamovi2.3.28 により実施した。

SD 法はプロフィール分析と、合成条件の探索的因子分析を行い、音声を聴く際に重点を置く因子について検討した。機械における「人間らしさ」を考えるために合成条件のみ因子分析を行った。なお、プロフィール分析では、厳密には順位尺度ではあるものの、量的なものとみなし算術平均を用いた。因子分析には、最小残差法、オブリミン法、平行分析を用いた (jamovi 2.3.28)。

## 3. 結果

実験参加者 11 人全員に対して分析が行われた。実験は複数回に分けて実施された。音声の再生途中から部屋に入室し、音声を 2 回以上聞いた参加者もいた。

「あなたは普段の生活において、静かな場所で会話の音が聞き取れますか？」という事前質問については 11 人全員が「はい」と回答した。

音声の種類を要因とした一要因分散分析（フリードマン検定）の結果を次に示す。設問「感情表現が豊かである」における条件間の主効果は有意であった ( $X^2(2)=18.0, |p| < .001$ )。ダービン・コノバー法による多重比較の結果、人間条件と合成条件との間で有意な差が認められ (統計量 7.798,  $|p| < .001$ )、人間条件と棒読み条件との間においても有意な差が認められた (統計量 8.578,  $|p| < .001$ )。一方で、設問「音声が聞き取りやすい」における条件間の主効果は有意でなかった ( $X^2(2)=1.28, ns$ )。設問「子どもの読み聞かせに適している」における条件間の主効果は有意であった ( $X^2(2)=18.8, |p| < .001$ )。ダービン・コノバー法による多重比較の結果、人間条件と合成条件との間で有意な差が認められ (統計量 9.94,  $|p| < .001$ )、かつ人間条件と棒読み条件との間においても有意な差が認められた (統計量 8.80,  $|p| < .001$ )。各設問の多重比較の結果を以下に示す (図 1, 2, 3)。

図 1. 設問「感情表現が豊かである」の多重比較 (\*\*\*)  
 $p < .001$ )

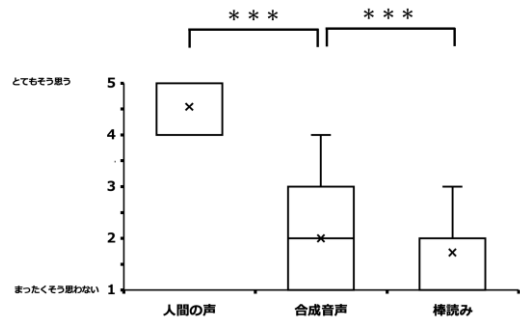


図 2. 設問「音声が聞き取りやすい」の多重比較 (\*\*\*)  
 $p < .001$ )

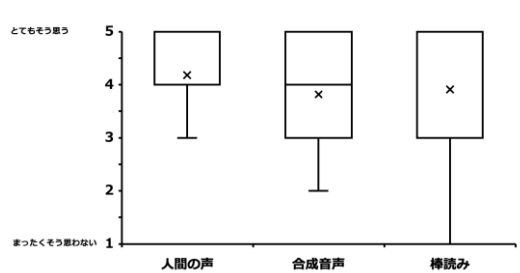
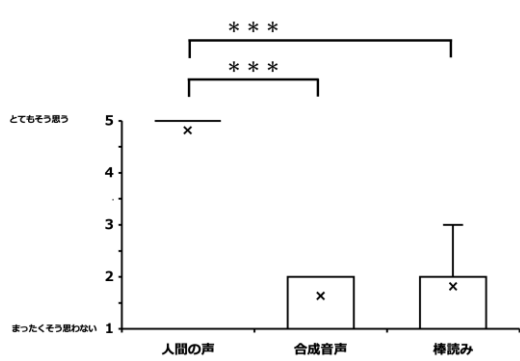
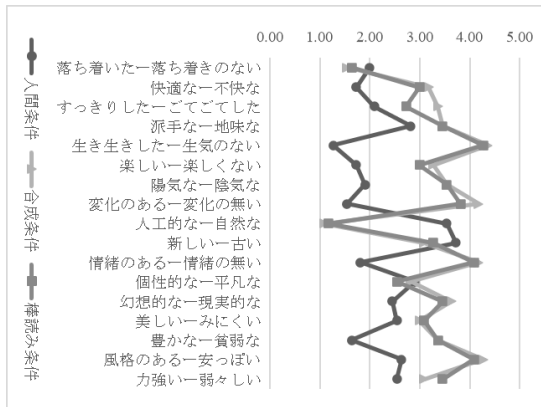


図 3. 設問「子どもの読み聞かせに適している」の多重比較 (\*\*\*)  
 $p < .001$ )



設問 2 の各音声における 11 人のプロフィール分析と平均値は以下ようになった (図 4) .

図 4 各条件における SD 法の平均値プロフィール



SD 法における因子分析の結果は以下になった. なお, 因子負荷量が 0.3 未満のものは表示していない (表 1) .

表 1 合成条件における因子分析の結果

| 因子負荷   | 因子分析   |       |
|--------|--------|-------|
|        | 1      | 独自性   |
| 楽しい    | 0.924  | 0.147 |
| 快適な    | 0.912  | 0.168 |
| 新しい    | -0.847 | 0.282 |
| 陽気な    | 0.802  | 0.356 |
| 幻想的な   | 0.790  | 0.376 |
| 豊かな    | 0.587  | 0.655 |
| 情緒のある  | 0.535  | 0.714 |
| 美しい    | 0.479  | 0.770 |
| 風格のある  | 0.449  | 0.798 |
| 変化のある  | 0.437  | 0.809 |
| 個性的な   | -0.412 | 0.830 |
| 派手な    | -0.386 | 0.851 |
| 力強い    | 0.384  | 0.852 |
| 生き生きした |        | 0.929 |
| 人工的な   |        | 0.931 |
| すっきりした |        | 0.966 |
| 落ち着いた  |        | 0.990 |

4. 考察

質問紙による評価では、「感情表現が豊かである」という設問において人間条件が、合成条件と棒読み条件よりも有意に高かったことから、人間の音声は合成音声よりも感情表現が豊かと感じられるといえる. 一方で、「音声が聞き取りやすい」という設問における条件間の違いは有意でなかった. これは、積極的には結論付けられないが、音声の聞き取りやすさにおいては、人間の声であっても、機械の声であっても影響しない可能性を示唆する. すなわち、アクセントの違いが人間らしさに帰属しない可能性を示唆する. 「子どもの読み聞かせに適している」という設問では、「感情表現が豊かである」という設問と同じく、人間条件が合成条件と棒読み条件よりも有意に高かったことから、機械と比べて人間の声の方が子どもへの読み聞かせには適していると感じられるといえる.

これらのことから、感情表現場面と子どもの読み聞かせのような場面では、人の声であることに聴取者が重点を置いている一方で、聞き取りやすさについては特に重点を置いていないことが考えられる.

SD 尺度によるプロフィールの結果を考察する. 11 人

それぞれのプロフィールは、ばらつきがあるため、全体的な傾向を読み取ることができなかった。平均値のプロフィールでは、合成条件と棒読み条件の間に大きな差はみられなかった。一方で、人間条件においては「生き生きとした一生気の無い」という項目と「情緒のある一情緒の無い」という項目において差がみられた。このことから、聴取者は「生きているかどうか」、「感情があるかどうか」という点に重点を置いていることが窺える。不気味の谷現象は「生きているかどうか」と「感情があるかどうか」に重点を置いていたにもかかわらず、生きているかの判断ができずに不安に感じると考えられる。月元 (2017) は、不気味の谷現象が人とロボットの類似度のみで評価されていることに言及した上で、不気味の谷現象に心の知覚次元のうち感情能力も関与することを示している。この結果は、本研究の結果を支持するといえる。

合成条件の因子分析の結果、抽出された因子は 1 つであった。正の値の因子負荷量は、「楽しい」、「快適な」、「陽気な」、「幻想的な」、「豊かな」、「情緒のある」、「美しい」、「風格のある」、「変化のある」、「力強い」であった。これらの項目には、肯定的な印象、情緒に関わる芸術的な印象、変化を伴う弾力的な印象が含有している。そのため、抽出した 1 因子を「肯定・芸術・弾力性」と呼ぶ。人は読み聞かせ場面において合成音声聴く際に「肯定・芸術・弾力性」を重視するといえる。よって、合成音声の印象評価には、この項目が妥当といえる。しかし、「生き生きとした」の項目が含まれていなかったことから、人間らしさに関する評価に用いるには吟味が必要といえる。

加えて、結果には記述しなかったが、個人情報伏せた上で実験後の参加者の感想を次に示す。合成音声自体が生活に浸透していること、アクセントを抑えた棒読み条件の方が聞き取りやすい場合があること、SD 項目の中に参加者が音声を評価する際に適切でないと思える項目があったこと、人間の声も人間に近い合成音声だと考えたことなどが挙げられた。

以上のことから、読み聞かせ場面では、現時点で機械よりも人間の声に適しているといえる。機械における「人間らしさ」は、「生きていると感じられるどうか」、「感情があるかどうか」、「肯定・芸術・弾力性」にかかる概念を尋ねることで求められるといえる。そのため、今後の研究では、機械における「人間らしさ」を尋ねるための質問紙の作成を課題としたい。また、本研究の課題として、合成音声を日常で用いる場面を再現

することが欠けていた。今後は観察を通して、合成音声と機械全般の日常的な利用に目を向けつつ、「人間らしさ」が読み聞かせ場面とどう関わっているかを研究していきたい。

## 5. 文献

- 岩宮眞一郎・細野晴雄・福田一昭 (1992) 音環境と景観の相互作用—景観の印象に及ぼす音環境の影響と音環境の印象に及ぼす景観の影響— *The Annals of physiological anthropology*, 11, 51-59.
- 小林まおり・倉片憲治 (2023) 女声と男声のどちらが聞き取りやすいか—日本音響学会誌, 79, 85-93.
- 向直人・今井理絵・福永智子 (2023) 絵本の読み聞かせを目的とした感情を込めた合成音声の生成—*相山女学園大学文化情報学部紀要*, 22, 109-115.
- 無料 e 童話 <https://www.e-douwa.com/index.html> (2024 年 7 月 3 日閲覧).
- Sam Devlin, Raluca Georgescu, Ida Momennejad, Jaroslaw Rzepecki, Evelyn Zuniga, Gavin Costello, Guy Leroy, Ali Shaw, Katja Hofmann. (2021). Navigation Turing Test (NTT): Learning to Evaluate Human-Like Navigation-International Conference on Learning Representations (ICLR 2021).
- 高橋ともみ・田中一晶・小林賢一郎・岡夏樹 (2020) 対話エージェントの合成音声への視覚的/音声的感情表現による人間らしさの強化—*ヒューマンインタフェース学会論文誌*, 22, 305-316.
- 月元敬 (2017) 実在ロボットに対する不気味の谷現象に関する感性評価研究—*日本感性工学会論文誌*, 16, 293-298.
- Xihuai Wang, Ziyi Zhao, Siyu Ren, Shao Zhang, Song Li, Xiaoyu Li, Ziwen Wang, Lin Qiu, Guanglu Wan, Xuezhi Cao, Xunliang Cai, Weinan Zhang. (2025). Audio Turing Test: Benchmarking the Human-Likeness of LLM-based Text-to-Speech - *arXiv:2505.11200*.
- 山本隆太郎・片上大輔 (2020) コミュニケーションツールにおけるユーザの対人恐怖心性とアバターの関係性の調査—*人工知能学会全国大会論文集*, 1P3-GS-7-03.