自発音声コーパスを用いて合成した音声で話すエージェントが会話相手の行動に与える影響*

☆飯塚喬久, 森大毅 (宇都宮大)

1　はじめに

人は音声対話システムを社会的存在と見なさず、音声コマンドで動く単なる機械として扱ってきた。我々は、社会的存在たり得る音声対話システムの実現に、音声合成の立場から迫ろうとしている。これまでの研究で、読み上げ音声コーパスを元にした一般的な合成音声に比べ、自発音声コーパスを元にした合成音声で対話するシステムの方がより社会的な存在と認識されるという仮説を、音声対話システムと人が対話している映像の印象を評価させることで検証してきた [1]。

本稿では、より直接的な検証のため、多人数の被験者に実際にシステムとのインタラクションを行わせ、そこでの行動に現れる違いを比較する。具体的には、対話の円滑さやエージェントのアニマシー知覚に関連する行動指標である反応時間および聞き手反応 [2](相槌、感情表出系感動詞、笑い) の頻度に合成音声が与える影響を調べ、対話参与者の行動の観点から上記仮説を定量的に検証する。

2　合成音声 [1]

本研究では音声対話システムの合成音声を作成する音声コーパスの特性の違い (読み上げ・独話 vs 自発・対話) が会話相手に与える影響を調査することを目的とする。これらの 2 種類の音声コーパスを元に、Tacotron 2 [3] を用いて音声合成器を構築した。読み上げ独話音声コーパスとしては JSUT [4] を使用する。本実験では全てのサブセット約 10 時間分のデータを用いて学習を行う。自発対話音声コーパスとしては UUDB (宇都宮大学パラ言語情報研究向け音声対話データベース) [5] を使用する。本実験では女性話者 1 名、約 18 分のデータを用いる。それぞれのコーパスで訓練した音声合成モデルをそれぞれ JSUT モデル、UUDB モデルと呼ぶ。UUDB モデルは、JSUT モデルを UUDB で fine-tuning することで作成した。

3　対話実験

対話実験ではこれまでの研究 [1] で構築した音声対話システムと同じものを使用した。このシステムは、いくつかのクイズを交えながらユーザと雑談するものであり、今回の実験では「国当てクイズ」の対話シナリオを用いた。ただし、今回の実験では音声認識結果は用いず、次発話の送出、相槌の送出、クイズの正誤判定は Wizard である実験者が裏でボタンを操作して行った。ボタンを押すタイミングは、できるだけ人間同士の会話における発話タイミングに近くなるよう努めた。音響的に全く同じ相槌音声が続くことの不自然さ [6] を回避するため、相槌は用意した 3 音声のうち 1 つをランダムに送出した。

実験は大学生・大学院生 44 人に対して行った。実験条件の割り当ては被験者間計画で行い、JSUT モデルの合成音声のシステムとの対話に 22 人を、UUDB モデルの合成音声のシステムとの対話に 22 人を割り当てた。実験後には後述する質問紙によって会話の質およびエージェントに対する印象を評価させた。

4　被験者の行動指標による評価

Fig. 1 に被験者ごとの反応時間と聞き手反応 (相槌、感情表出系感動詞、笑い) の数の分布を示す。反応時間は、被験者の発話始端時刻からシステムの先行発話終端時刻、または聞き手反応の焦点要素 [7] 終端時刻を引いた時間と定義し測定した。反応時間 (Fig. 1 (a)) の平均は、JSUT モデルのシステムで 1.16 s、UUDB モデルのシステムで 1.02 s だった ($p = 0.004$, Welch t 検定)。この結果から、UUDB モデルの方が反応時間が短く、被験者が人間同士の会話のように発話の時間的制約 [8] を感じている可能性があると言える。このことは研究仮説である「自発音声に基づく合成音声で対話するシステムの方がより社会的な存在と認識される」ことの 1 つの証拠になると考えられる。

相槌の数 (Fig. 1 (b)) の平均は、JSUT モデルのシステムで 7.55 回、UUDB モデルのシステムで 17.45 回だった ($p = 0.023$)。この結果から、特に UUDB モデルのシステムに対し、普通は機械に対してほとんど見られない相槌を打つ行動が多く観察され、「自発音声に基づく合成音声で対話するシステムの方がより社会的存在と認識される」ことを示唆している。

感情表出系感動詞 (Fig. 1 (c)) および笑い (Fig. 1 (d)) の数の平均は、JSUT モデルのシステムでそれぞれ 8.82 回と 6.32 回、UUDB モデルのシステムでそれぞれ 10.00 回と 6.14 回であり、それらの差は統計

*How does a spontaneously-speaking dialog system affect user behavior? by IIZUKA, Takahisa, MORI, Hiroki (Utsunomiya University)

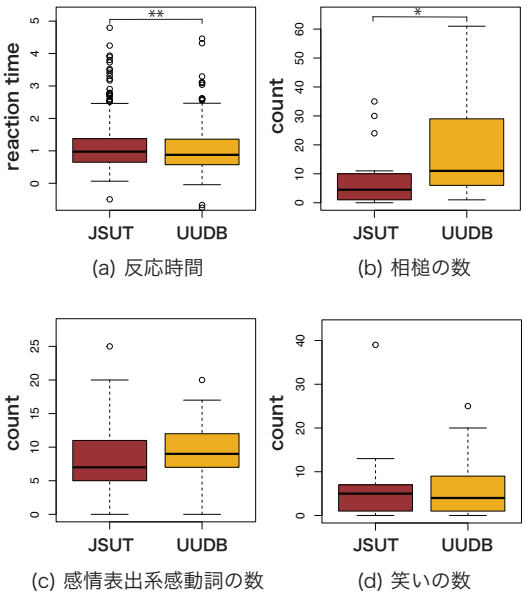

Fig. 1: 被験者の反応時間と聞き手反応の数 (JSUT: JSUT モデルの合成音声のシステムとの対話, UUDB: UUDB モデルの合成音声のシステムとの対話)

的に有意ではなかった。

5 主観評価

主観評価の質問項目とその結果を Table. 1 に示す。「メイちゃんとあなたの会話はどれほど人間同士の会話に近かったですか?」の平均評価値は JSUT モデルのシステムで 3.59、UUDB モデルのシステムで 4.05 であり、後者の方がより人間同士の会話に近い方に分布している傾向があった ($p = 0.062$, Brunner-Munzel 検定)。また、「メイちゃんの声はどのように感じましたか?」の項目の平均評価値は JSUT モデルのシステムで 2.77、UUDB モデルのシステムで 3.27 であり、後者の方がより自然と感じた方に分布している傾向があった ($p = 0.051$)。

被験者の行動指標による評価では反応時間および相槌の数に差があったが、主観評価でも人間同士の会話に近く、音声も自然であると評価される傾向があった。このことから、自発音声を基にした合成音声で対話するシステムに対して、被験者はより社会的な存在と認識し、またそのような相手としてふるまっていると言える。

6 おわりに

本稿では、自発音声を基にした合成音声で対話するシステムが読み上げ音声を基にした合成音声で対話するシステムよりも社会的存在として認識されやすいという仮説を検証するための対話実験について述べた。実験結果から、自発音声を基に合成したシステムの方が、反応時間が短くなり、相槌の数が増えることが分かった。これは、エージェントとの会話を人間同士の会話に近づけたいならば、その合成音声は自発音声の持つ何らかのパラ言語的・非言語的特質を備えなければならない、ということを示す重要な発見であると考える。それがどのような特質かを明らかにすることは、音声対話システムの今後の発展に向けた重要な課題である。

Table 1: 主観評価の質問項目と回答数の分布

	評価					
音声合成モデル	5	4	3	2	1	平均
メイちゃんとあなたの会話はどれほど人間同士の会話に近かったですか? (5: とても近い , 1: 全く近くない)						
JSUT	2	13	4	2	1	3.59
UUDB	6	13	1	2	0	4.05
人に話しかけられたとき、すぐに答えないで黙っていたら申し訳ないと思いますよね。今回メイちゃんに何か聞かれた時、あなたの気持ちに近いのはどれですか? (5: すぐに答えないと申し訳ない , 1: すぐに答えなくても全然申し訳ないと思わない)						
JSUT	7	10	1	2	2	3.82
UUDB	5	12	2	1	2	3.77
またメイちゃんと話してみたいと思いましたか? (5: ぜひまたお話をしてみたい , 1: 二度とお話はしたくない)						
JSUT	13	6	2	1	0	4.41
UUDB	13	7	1	0	1	4.41
メイちゃんにどれくらい好感を抱きましたか? (5: とても好感を抱いた , 1: 全く好感を抱かなかった)						
JSUT	6	9	4	3	0	3.82
UUDB	5	10	3	3	1	3.68
メイちゃんからどれくらい心を感じましたか? (5: とても心を感じた , 1: 全く心を感じなかった)						
JSUT	2	11	6	2	1	3.50
UUDB	4	13	1	4	0	3.77
メイちゃんの声はどのように感じましたか? (5: 自然だった , 1: 不自然だった)						
JSUT	1	7	1	12	1	2.77
UUDB	4	7	3	7	1	3.27

参考文献

[1] 飯塚, 森, 西野, 音講論, pp. 1011–1014, 2021.
[2] Den et al., Proc. LREC 2012, pp. 1332–1337, 2012.
[3] Shen et al., Proc. ICASSP, pp. 4779–4783, 2018.
[4] R. Sonobe, S. Takamichi, and H. Saruwatari, arXiv preprint, 1711.00354, 2017.
[5] Mori et al., Speech Communication, Vol. 53, pp. 36–50, 2011.
[6] H. Mori, Acoustical Science and Technology, Vol. 34, pp. 147–149, 2013.
[7] 小磯, 伝, 日本認知科学第 28 回大会発表論文集, pp. 250–255, 2011.
[8] H. Clark, Speech Communication, Vol. 36, pp 5–13, 2002.