自発音声に基づく合成音声で対話するシステムがユーザに与える影響の調査*

☆飯塚喬久, 森大毅, 西野広直 (宇都宮大)

1 はじめに

ロボットや擬人化エージェントなどの音声対話システムを、人は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う。システムとの対話で相槌やフィラーなどの聞き手反応がほとんど見られないのは、人との対話の時のような協調原理 [1] やタイムプレッシャーを気にする必要がない存在と認識されているのが一因である。本研究は、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するものである。

これまでの音声対話システムの合成音声は、原稿の読み上げ音声を元に訓練したモデルに基づく一般的なテキスト音声合成システムから生成されていた。その合成音声はかなり明瞭性が高く、人間の肉声とはほぼ区別ができないレベルまで性能が向上している。それらの合成音声はユーザに聞き取りやすいようにはっきりと話す点ではすでに十分な性能を発揮しているが、我々が日常会話で発する自発音声とは異なっている。読み上げ音声と異なり、自発音声ははっきり話すことそれ自身を目的とした音声では無いため、発話の意図が達成される範囲でできる限り省力的に話され、読み上げ音声に比べ発音の怠けや不明瞭な発声が多い特徴を持っている。また、読み上げ音声は独話であるため、自発的な対話音声とは発話スタイルが異なっている。

これまでも我々は、音声対話システムの合成音声に感情状態などのパラ言語情報を反映する対話音声合成を研究してきた [2]。我々は、一般的な音声対話システムのような読み上げ音声コーパスを元にした合成音声ではなく、自発音声コーパスを元にした合成音声を用いることで、音声対話システムが単なる機械ではない社会的存在であると認識され、より人間同士の日常会話に近づくと考えているが、これまで自発音声を元にした合成音声の方が優れているという証拠は示されていなかった。

本研究では、一般的な合成音声と比べ、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という仮説に基づき、自発音声と読み上げ音声を元にした合成音声の違いによるユーザへの影響を実験的に調査し、この仮説を検証する。

2 音声対話システム

2.1 合成音声

本研究では音声対話システムが出力する音声合成器を訓練する音声コーパスの特性の違い（読み上げ・独話・声の仕事経験を持つ話者 vs 自発・対話・一般人）がユーザに与える影響を調査することを目的としている。そこで、それら 2 種類の音声コーパスを元に、Tacotron 2 [3] を用いて 2 種類の音声合成器を構築した。

読み上げ独話音声のコーパスとしては JSUT [4] を使用する。JSUT は声による仕事の経験を持つ女性話者 1 名が指定されたテキストを読み上げた音声データのセットである。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。

自発対話音声のコーパスとしては 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。UUDB は大学生同士が 4 コマまんが並べ替えタスクを遂行中に発せられた自然な対話音声のデータセットである。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 Tacotron 2 のモデルをそれぞれ JSUT モデル、UUDB モデルと呼ぶ。UUDB モデルは、JSUT モデルに UUDB の訓練データで Fine-tuning することで得た。

Tacotron 2 への入力は音素記号列であるが、UUDB を訓練データとする場合には、フィラー [6] および感情表出系感動詞 [7] の音声学的変異を表現するために拡張した音素記号を用いた。具体的にはフィラーおよび感情表出系感動詞を構成する母音 (α,ι,υ,ϵ,o) を追加した。文献 [7] ではフィラーと感情表出系感動詞との間にも音響的に異なる性質があることが示されているが、今

*How does a spontaneously-speaking dialog system affect users? by IIZUKA, Takahisa, MORI, Hiroki, and NISHINO, Hironao (Utsunomiya University)

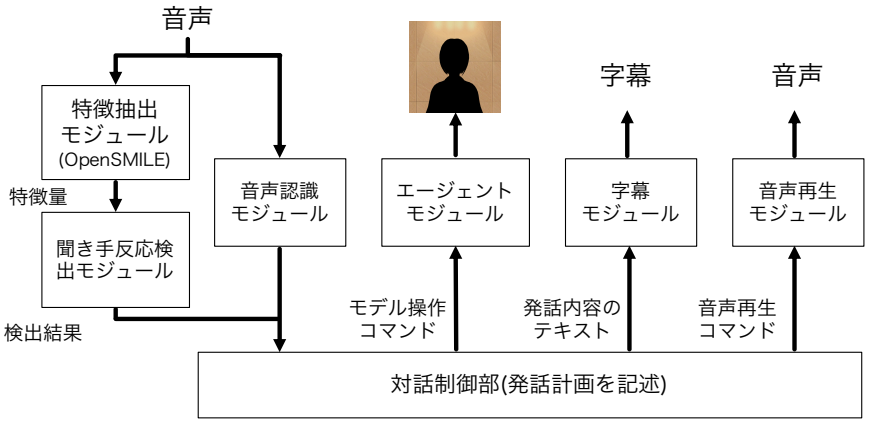

Fig. 1　音声対話システム概要 (森本 [11] 一部改変)

回は通常語彙の母音とだけ区別した。また子音は共通とした。

標本化周波数は 16 kHz、短時間フーリエ変換の条件はフレーム長 50 ms 、フレームシフト 12.5 ms で窓関数にはハン窓を使用した。

Tacotron 2 のモデル構造および最適化の条件は、バッチサイズを除いて NVIDIA の実装 [8] のデフォルトから変更はない。バッチサイズは 32 とした。

波形生成部には Griffin-Lim アルゴリズム [9] を使用した。

2.2　システム概要

研究仮説を検証する実験のため、新たに音声対話システムを構築した。音声対話システム構築には MMDAgent [10] を用いた。MMDAgent は音声認識、音声再生、対話管理、3D モデルのモーション管理などを行えるシステムである。

図 1 に本研究で構築したシステムの概要を示す。今回は MMDAgent の音声合成器は用いず、かわりに JSUT モデルまたは UUDB モデルを用いて予め合成しておいた音声を再生する。また、合成音声による言語情報の伝達を確実にするため、音声の再生と同時に発話内容を字幕で呈示する。

3D エージェントはそれぞれの合成音声から受ける個人性の印象とエージェントの外観から受ける個人性の印象との整合性が音声の評価に影響を与える可能性を排除するためにシルエットとした。

図 1 中の聞き手反応検出モジュール [11] は、音声認識モジュールと並行に動作し、ユーザの相槌やフィラーを即時検出するものである。今回構築した音声対話システムでは、ユーザの相槌が無かった場合に、次発話の前に相槌を少し待つ動作をするようにしている。

2.3　発話計画作成

本研究が想定しているのは、人間と機械との間の表情豊かな会話である。現在の対話システムの技術では、広い話題で破綻なく対話システムと雑談を続けるのは難しく [12]、対話に破綻が生じた場合には、本研究の対象である合成音声以外の要素に対する悪い評価が支配的になってしまう可能性がある。システムが特定の話題に誘導することで、そのような破綻の危険を減らすことができる。今回は、雑談風の会話の中で、もっぱらシステム側が人間に質問したり知識を披露したりするシナリオを考えた。

エージェントの発話計画とユーザのふるまいに関する過去の研究で、人が手で作成した発話計画を使った場合に比べ、会話の中で自然に生じた発話をそのまま発話計画とする方が、相槌などの聞き手反応が多くなることが明らかになっている [13]。そこで、今回の対話システムの発話計画作成においても実際の人間同士の対話を参考にした。具体的なシナリオとしては、システムがユーザと雑談しながらクイズを出して対話するというものにした。クイズを出すことでユーザからのフィードバックが期待でき双方向の対話が実現できる。クイズの内容は答えの候補集合が自明であるような内容とした。これは例えば都道府県を当てるクイズであれば、音声認識の範囲を 47 都道府県に絞ることができるためである。シナリオは次のように作成した。まず、親近性の高い協力者 2 名に対してクイズを出す側と

答える側の役割を与えた。クイズを出す側には、出して欲しいクイズとその答えに関する雑学を覚えさせた。その後、2名にクイズの出題、回答、それに関する雑学を交じえた会話を普段通りにしてもらいその様子を収録した。収録した対話音声は言い淀みやフィラー、感情表出系感動詞、倒置などの文法の破格を含み、断片的な発話となっている。

本研究ではこのような方法で「都道府県クイズ」と「惑星クイズ」の2種類の発話計画を作成し、MMDAgent の FST ファイルとして実装した。

3 印象評価

「一般的な合成音声と比べ、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」という仮説を検証するため、音声対話システムと人が対話している映像を見てもらい、その印象を評価させる実験を行った。

システムとの対話は研究室の大学院生1名 (著者に含まれない) が行い、UUDB モデルの合成音声で都道府県クイズ会話をするセッションと、JSUT モデルの合成音声で惑星クイズ会話をするセッションに、この順で1回ずつ参加させ、音声対話システムの画面と対話者の表情を Zoom で記録することで映像を作成した。

大学生 46 人に対して人とシステムが対話している映像を視聴させ、後述する質問紙によってその印象を5段階 (5: 最も良い、1:最も悪い) で評価させた。実験は被験者間計画で行い、UUDB モデルの合成音声のシステムの評価に 22 人を JSUT モデルの合成音声のシステムの評価に 24 人を割り当てた。質問項目は以下の通りである。

1. システムに対する評価
   - システムの説明は分かりやすかったですか?
   - システムとの会話は一方的でしたか?
   - システムの発話は自然でしたか?
2. ユーザに対する評価
   - システムと会話していたユーザは、どれほどシステムと親しくなったように見えましたか?
3. 対話自体に対する評価
   - どれほど人間同士の会話に近いと思いましたか?
   - あなたはこのシステムとお話ししてみたいと思いましたか?

Table 1　各評価段階 (5: 最も良い、1:最も悪い) の回答数

	評価					
音声合成モデル	**5**	**4**	**3**	**2**	**1**	**平均**
システムの話は分かりやすかったですか?						
UUDB	5	14	1	3	1	3.79
JSUT	10	10	0	2	0	4.27
システムとの会話は一方的でしたか?						
UUDB	0	8	3	9	4	2.62
JSUT	1	7	0	11	3	2.64
システムの発話は自然でしたか?						
UUDB	2	8	4	7	3	2.96
JSUT	1	6	1	11	3	2.59
システムと会話をしていたユーザは、どれほどシステムと親しくなったように見えましたか?						
UUDB	1	9	3	8	3	2.88
JSUT	1	5	11	5	0	3.09
どれほど人間同士の会話に近かったと思いますか?						
UUDB	2	14	4	4	0	3.58
JSUT	2	6	7	5	2	3.05
あなたはこのシステムとお話ししてみたいと思いましたか?						
UUDB	5	9	4	3	3	3.42
JSUT	4	10	4	3	1	3.59

4 実験結果および考察

表1に、各質問項目に対する UUDB モデルの合成音声のシステムおよび JSUT モデルの合成音声のシステムに対する評価を度数分布表で示す。

4.1 システムに対する評価

「システムの説明は分かりやすかったですか」に対する評価の平均値は UUDB モデルで 3.79, JSUT モデルで 4.27 であった ($p = 0.065$)。この差は統計的に有意傾向であるが有意ではなく、より規模を大きくした調査が必要である。JSUT モデルの音声による説明の方が分かりやすい印象を与えたとすれば、JSUT モデルの音声が UUDB モデルの音声に比べ明瞭性が高く、努力的な発声

であったことが作用した可能性がある。

「システムとの会話は一方的でしたか」に対する評価には有意差が無く ($p = 0.991$)、どちらも一方的寄りであるという結果となった。これは合成音声の種類とは関係無くシステム主導のシナリオが原因であると考えられる。

「システムの発話は自然でしたか」に対する評価には有意差はなかった ($p = 0.299$)。

4.2 ユーザに対する評価

「システムと会話していたユーザは、どれほどシステムと親しくなったように見えましたか」に対する評価には有意差はなかった ($p = 0.574$)。JSUT モデルの合成音声で話すシステムの場合と異なり、UUDB モデルの合成音声で話すシステムに対する評価は双峰的であるように見えるが、この原因は現時点では明らかになっていない。

4.3 対話自体に対する評価

「どれほど人間同士の会話に近かったと思いますか」に対する評価の平均値は UUDB モデルで 3.58, JSUT モデルで 3.05 であった ($p = 0.076$)。この差は統計的に有意傾向であるが有意ではない。研究仮説「一般的な合成音声と比べ、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」が正しければ、UUDB モデルの合成音声で話すシステムとの会話の方が人間同士の会話に近いと認識されると考えられる。上に述べた結果はこれを支持するものである可能性があるが、このことは規模を大きくした実験により裏付ける必要がある。

「あなたはこのシステムとお話ししてみたいと思いましたか」に対する評価には有意差はなかった ($p = 0.747$)。

4.4 考察

対話自体に対する質問で「どれほど人間同士の会話に近かったと思いますか」に対する評価に有位傾向が見られた。人間同士の会話に近いということは、社会性のある存在と見なしているという可能性が考えられる。このことから本研究の仮説である社会的存在たり得る音声対話システムの実現に自発音声を元に合成した音声が有効であるという可能性が示唆された。

他の項目についてはほぼ差が無く、合成音声の違いによっては評価が変わらなかったと言える。しかし、本稿での実験は被験者が直接システムと対話したのでは無いため、このような結果になった可能性がある。そのため、今後は対話実験を行う必要があると考える。

5 おわりに

本稿では、自発音声を元にした合成音声で対話するシステムが読み上げ音声を元にした合成音声で対話するシステムよりも社会的な存在と認識されるという仮説を検証するための実験について述べた。実験はシステムと人が対話している映像から受ける印象を評価する形式で行った。実験結果から、自発音声を元にした合成音声を用いることでシステムに対して社会性がある存在と感じられる可能性が示唆された。今後は実際にシステムとの対話実験を行い、合成音声の種類によるユーザへの影響についてさらに調査を行っていく必要がある。

謝辞 本研究は JSPS 科研費 18H04128, 19H01252 の助成を受けたものです。

参考文献

[1] 石崎, 伝, 談話と対話, 東京大学出版会, 2001.
[2] T. Nagata, Speech Communication, Vol. 88, pp. 137–148, 2017.
[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] K. Maekawa and H. Mori, 音声研究, Vol. 21, pp. 53–62. 2018.
[7] H. Mori, Proc. Interspeech 2015, pp. 1309–1313, 2015.
[8] https://github.com/NVIDIA/tacotron2
[9] D. Griffin and J. Lim, IEEE Trans. Acoustics Speech Signal Process., Vol. ASSP-32, pp. 236–242, 1984.
[10] A. Lee, K. Oura and K. Tokuda, Proc. ICASSP, pp. 8382–8385, 2013.
[11] 森本, 森, 人工知能学会研究会資料, SIG-SLUD-B902, pp.99–100, 2019.
[12] 東中 他, 人工知能, Vol. 35, pp. 333–343, 2020.
[13] 高松屋, 森, 音響論, pp. 833–834, 2020.