非流暢な合成音声

モクタリ明子[1] 波多野博顕[2] 新井潤[3]
キャンベル ニック[4] 定延利之[5]
[1]富山県立大学 [2]筑波大学 [3]関西学院大学
[4]ダブリン大学 [5]京都大学
a-mokh@pu-toyama.ac.jp araijun@kwansei.ac.jp hatano.hiroaki.ge@u.tsukuba.ac.jp
nick@tcd.ie sadanobu.toshiyuki.3x@kyoto-u.ac.jp

概要

日常会話音声は，テレビなどで耳にするアナウンサーの発話とは異なり，非流暢性に満ちている．そのことから，合成音声に非流暢性を取り入れると，マシーンコミュニケーションがさらにリアルでインタラクティブになると考えられる．本研究では，まず日常会話音声データを,既存の AI による合成音声システムに学習させた．次に，同じ言語内容の非流暢性を加えた発話とそうでない発話を合成し，調査回答者にどちらがより人間味のある話し方かを尋ねた．その結果，非流暢な合成音声の方を，より人間味のある話し方だと判断する回答者の割合が優位に高かったことが明らかになった．

1 はじめに

私たちが日々やりとりすることばは，非流暢である．定延は，「現実のことば」は，訓練されたアナウンサーがニュース原稿を読み上げるようなものとはかけ離れており，たとえ母語話者であってもその発話はたいてい非流暢であるとしている[1]．日本語教育では，アナウンサーがよどみなく話す日本語を手本にするのではなく，日本語母語話者の非流暢性の規則を学習者に教え，現実の日本語を手本に話すための指導が始まっている[2]．同じように，合成音声にも非流暢性を取り入れたなら，マシーンコミュニケーションがより人間味を帯びたものになるのではないか．

AI の深層学習を用いたディープニューラルネットワーク（以下，DNN）音声合成研究は，2010 年代半ばより飛躍的に発展し，明瞭かつ自然な発話音声の合成が可能になった．さらに感情と連動した声のバリエーションを学習させ，合成音声をより表現豊かにするための研究が進められている[3]．

合成音声は様々な用途で使われるが，「現実のことば」の合成を目指すならば，学習に使用するデータが，演じられたものではない自然発話である必要がある．本研究では，1 名の日本語母語話者から収集した長時間日常会話音声を学習データとして，既存の DNN 音声合成システムに学習させた．次に，同じ言語内容の非流暢性を加えた発話とそうでない発話を合成した．前者には，先行研究で論じられている「現実のことば」に現れる非流暢性を取り入れた．調査回答者に，非流暢性を加えた合成音声とそうでない合成音声を聞かせ，より人間味のある話し方はどちらかを尋ねた．また回答の傾向が個人のどのような属性に影響されるかをみるため，回答者の「性別」「年齢」「育った場所」と回答の関係についての考察も行った．

2 データと学習

本研究で DNN 音声合成の学習に使用したデータは，大規模自然発話コーパスの一部である[i]．本データは,収録開始当時 32 歳であった 1 名の女性日本語母語話者が,日常的にヘッドセットマイクを装着し，様々な対話相手との発話を MD プレーヤーに収集したものである．録音室ではなく，話者の自宅，友人の家，自家用車などで収録されたタスクなしの自発的な発話であった．対話相手が同じ場所にいることもあれば，電話での対話を収録したものもあった．いずれの場合も，対話相手の声はデータに含まれていない．対話相手は「友人」「身内」「子供」「他人」に分類され，ラベル付けされた．

本データは，収録後に複数のトランスクライバーによって文字に書き起こしされている．その際，ト

[i] JST/CREST「表現豊かな発話音声コンピュータ処理システム（研究代表者: キャンベル ニック）」において 2000 年前後に収集された．一般公開はされていない．

ランスクライバーは「意味を成す最も短いかたまり (minimum chunks) 」に発話を区切り，書き起こし作業を行うように指示された．本研究では，それらのうち 32, 535 の chunk を学習に使用した．学習には，文字情報に加えて，対話相手の情報も用いられた．その結果，対話相手に応じた話し方の違いを合成することが可能になった[4].

DNN 音声合成の学習には, 以下の既存のシステムを使用した．テキストからスペクトログラムの予測には Tacotron および Tacotron2[5] を，スペクトログラムから音声を生成するのには WaveGAN と Parallel WaveGAN[6]を用いた．また，合成音声の精度を高めるために，FastPitch を使用した[7].

3 非流暢性

本研究では，2 で学習した合成モデルに、次の 5 種類の非流暢性を取り入れて刺激音を作成・出力した。

(a) フィラー
(b) 語中延伸
(c) 文節末での跳躍的上昇
(d) 文節末での跳躍的上昇＋下降
(e) 文節末の判定詞＋終助詞

(a) は高頻度で観察される非流暢性であり，自然発話には不可欠である．本調査では「えーっと」と「えっと」を使用した．(b) は単語の途中でつっかえた際に，ひとつの音を長く延ばすものである．例えば「41」と言おうとして「よんじゅーーいち」となるようなもので，話者のためらいの発話態度と結びついている場合がある[8]．人間の日本語発話では，つっかえた後，単語の初めに戻ってもう一度言い直す「初頭戻り方式」と，後続音をそのまま発話する「続行方式」があるが[8]，ここでは後者を使用した．(c) は文節末に現れる急激に上昇するイントネーションのことで，「それでね，」の「ね」の部分が急激に高くなるようなものである（下線・上線はイントネーションの動態を表す）．(d) は「わたしがぁ，」のように「が」で跳ね上がったイントネーションが，再び低音に戻るものを指す．(e) は文節末に判定詞「です」と終助詞「ね」が連なって「〜はですね，」となり, その後に次の文節が続いていくものを指す．(c)(d)(e) は，定延が「文節単位のこま切れ諸現象」として説明している非流暢発話である[1]．これら 5 種類の非流暢性はいずれも日本語日常会話音声において広く観察されるものである．

4 調査

調査はオンライン上で実施した．回答者は，調査についての説明を読み，「性別」「年齢」「育った場所」についての質問に答えたあと，合成音声についての印象評価を行うよう求められた．

4.1 刺激音

5 つのペアの刺激音を作成した．どのペアも，同じ言語内容の非流暢性を加えた発話－dis (disfluent) －とそうでない発話－ntl (neutral)－から成っていた．以下，刺激音のリストを示す．非流暢な合成発話には，3 節で説明した非流暢性が 1 つ以上含まれており（下線部），その種類を (a) – (e) で示している．

Q1. dis: 私がぁ (d), 旅行で行きたいところは，えーっと (a), モロッコとぉ (d), えーっと (a), フランスです.
ntl: 私が旅行で行きたいところは，モロッコとフランスです.

Q2. dis: 13 プラス 28 イコールよんじゅーー (b) いちです.
ntl: 13 プラス 28 イコール 41 です.

Q3. dis: それでね (c)，私がね (c)，スーパーに行ったらね (c)，先生がいたんです.
ntl: それで私がスーパーに行ったら，先生がいたんです.

Q4. dis: 北陸はですね (c)(e)，11 月になるとですね (e)，道路からお湯がでるんです.
ntl: 北陸は 11 月になると，道路からお湯がでるんです.

Q5. dis: 浅草寺に初めて行ったのは，えっと (a), にせんー (b) 22 年の 6 月です.
ntl: 浅草寺に初めて行ったのは，2022 年の 6 月です.

4.2 回答者

回答者は，性別・年齢のバランスを考慮して 35 の都道府県から人材派遣会社を通して集められた日本語母語話者 135 名であった. 調査は有償であった.

4.3 方法

調査ページは, Google Forms を用いて作成した（図 1 参照）．回答者は，2 つの合成音声 A・B を聞き，「より人間味のある話し方」だと感じる発話を選択するよう指示された．A と B は同じ言語内容であっ

たが，どちらか 1 つは非流暢性を加えた発話で，もう 1 つはそうでない発話であった（4.1 節参照）．どちらの音声が先に再生されるかは，ランダムに決められた．回答者は，聞こえてくる全ての発話が合成音声であることを知らされていた．回答時間は無制限で，回答者たちは何度も音声を再生したり，前の設問に戻って回答し直したりすることができた．設問は全 5 問 (Q1-Q5) であった．

図 1 調査ページの画面

5 結果

結果を表 1 に示す．設問全 5 問について，非流暢な合成発話を「より人間味のある話し方」だと判断した回答者の割合は 68%であり，チャンスレベルより優位に高かった（二項検定，$p<0.05$）．そのように判断した回答者が最も少ない Q1 においても，その傾向は変わらなかった ($p<0.05$).

表 1 非流暢な合成音声の方が「より人間味のある話し方」だと回答した人の割合 (n=135)

Q1	64 %
Q2	70 %
Q3	73 %
Q4	67 %
Q5	66 %
平均	68 %

次に，「性別」による回答者の割合を表 2 に示す．

表 2 非流暢な合成音声の方が「より人間味のある話し方」だと回答した人の性別における割合

性別	n	平均	Q1	Q2	Q3	Q4	Q5
男性	57	71%	68%	77%	67%	67%	74%
女性	63	67%	65%	64%	78%	70%	60%
無回答	15	63%	47%	73%	73%	60%	60%

1 列目は「性別」，2 列目の「n」は各カテゴリーの回答者数，3 列目の「平均」は設問全 5 問に対して非流暢な合成音声の方を「より人間味がある」と回答した人の割合，4-8 列目の Q1-Q5 は設問ごとにみた同様の割合示している．平均については，性別による回答の開きはみられなかった．設問ごとにみていくと，性別「無回答」において，Q1 は「非流暢な合成音声の方が人間味のある話し方」だと答えた回答者の割合がやや低くなっていた．

次に，年齢別の回答者の割合を表 3 に示す．

表 3 非流暢な合成音声の方が「より人間味のある話し方」だと回答した人の年齢別割合

年齢	n	平均	Q1	Q2	Q3	Q4	Q5
20 代	18	69%	56%	83%	67%	72%	67%
30 代	30	70%	60%	70%	77%	77%	67%
40 代	22	67%	64%	82%	73%	50%	68%
50 代	23	69%	74%	74%	70%	61%	65%
60 代	20	61%	60%	65%	65%	55%	60%
70 代	16	76%	69%	69%	75%	81%	88%
80 代	6	60%	83%	0%	100%	100%	17%

年齢についても，平均には回答の開きがみられなかった．Q2 および Q5 は，80 代において，「非流暢な合成音声の方が人間味のある話し方」だと答えた回答者の割合が著しく低かった．ただし，80 代の回答者数は 6 名と，他のカテゴリーよりも少なかった．

最後に，育った場所における回答の割合を表 4 に示す．全部で 35 都道府県あり，回答者が 1 名のみの地域も多くあった．東京都が 24 名と最も多かった．回答者が 5 名以上いた都道府県と，その回答を以下にまとめる．

表 4 非流暢な合成音声の方が「より人間味のある話し方」だと回答した人の都道府県別割合

場所	n	平均	Q1	Q2	Q3	Q4	Q5
千葉	7	80%	71%	71%	100%	71%	86%
大阪	8	76%	63%	50%	88%	100%	88%
神奈川	12	73%	67%	83%	83%	67%	67%
福岡	6	70%	67%	83%	67%	50%	83%
北海道	7	69%	86%	43%	86%	71%	57%
東京	24	68%	71%	67%	67%	71%	63%
埼玉	14	60%	36%	71%	64%	71%	57%
愛知	11	53%	18%	100%	36%	46%	64%
広島	9	51%	56%	56%	67%	33%	44%

平均をみると，非流暢な合成発話を「より人間味のある話し方」だと判断した回答者の割合が，愛知県と広島県において他よりやや低いことが分かる．また特に愛知県の回答者は，Q1 における逆回答率（非

流暢でない合成音声発話を「より人間味がある話し方」だと判断した回答者の割合）が高く，Q2 との差が顕著であった．Q1 は (a) フィラーと (d) 跳躍的上昇＋下降を含む非流暢な合成発話であった．一方 Q2 は，(b) 語中延伸を含むものであった．非流暢性の種類と，特定の地域における回答の傾向に関連があるか，今後考察を重ねたい．他にも，広島県における Q4 など，逆回答率が比較的高いケースが観察された．

以上のように，「性別」「年齢」「育った場所」における各設問の回答者の割合を取り上げると，逆の傾向が観察される事例があったが，全体的には表 1 に示したように非流暢な合成音声の方が「より人間味のある話し方」だと回答した人の割合が優位に高かったことが分かった．

6 おわりに

本研究では，既存の DNN 音声合成システムに，日本語母語話者 1 名より収集した日常会話音声データを学習させた．その合成モデルを用いて，同じ言語内容の非流暢性を加えた発話とそうでない発話を出力し，どちらがより人間味のある話し方かを問う調査を行った．非流暢な発話には，「フィラー」，「語中延伸」，「文節末での跳躍的上昇」，「文節末での跳躍的上昇＋下降」，「文節末の判定詞＋終助詞」の 5 種類を取り入れた．その結果，非流暢な合成音声発話の方が，「より人間味のある話し方」だと知覚していた回答者の割合が優位に高かったことが明らかになった．

今回取り扱った非流暢性の中には，比較的容易に自然な出力ができるものと，そうでないものとがあった．具体的には，「文節末での跳躍的上昇」，「文節末での跳躍的上昇＋下降」，「文節末の判定詞＋終助詞」が前者であり，「語中延伸」が後者であった．「フィラー」に関しては，「えー」と入力すると，驚きの「えー」が数多く出力されるなど，いずれの場合も学習に用いたデータの性質が大きく影響していた．AI の発展により，合成音声研究は目覚ましい発展を遂げているが，「現実のことば」を合成するためには，学習用データそのものが Observer's Paradox[9] ―録音していると意識することによって，普段とは異なった話し方になること― に陥っていない自然発話である必要がある．また，非流暢性や対話相手に応じた発話様式の変化など，日常会話に散りばめられた現象を取り入れていく必要がある．

謝辞

本研究は JSPS による基盤研究 (S) 「非流暢な発話パターンに関する学際的・実証的研究」（課題号番号: JP20H05630，研究代表者: 定延利之） の助成を受けている．

参考文献

1. 定延利之『文節の文法』，大修館書店，2019.
2. 舩橋瑞貴「第 5 部（言語教育からみた（非）流暢性）のねらいと論文紹介」，定延利之・丸山岳彦・遠藤智子・舩橋瑞貴・林良子・モクタリ明子（編）『流暢と非流暢性』，pp.251-253，ひつじ書房，2024.
3. J. Lorenzo-Trueba et al., "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis", *Speech Communication*, 99, pp. 135-143, 2018.
4. A. Mokhtari, N. Campbell and T. Sadanobu, "Two efforts towards natural speech synthesis: Incorporating disfluency and speaking style change based on the interlocutor", Acoustics 2023 (poster presentation), Sydney, 2023.
5. J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions", *In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.
6. R. Yamamoto, E. Song and J. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram", *IEEE International Conference on Acoustics, In Proceedings of Speech and Signal Processing*, 2000.
7. A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction", *In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.
8. 定延利之「フィラー、とぎれ、延伸と発話態度の結びつき」，『日語偏誤与日語教学研究』7，pp. 3-15，日語偏誤与日語教学学会，2022.
9. W. Labov, *The sociolinguistic patterns*. University of Pennsylvania Press, 1972.