応答発話を聞きやすくするためのフィラー/ポーズの自動挿入に基づく話し言葉変換

平田 遼　　太田 健吾
阿南工業高等専門学校
1134272@st.anan-nct.ac.jp, kengo@anan-nct.ac.jp

1 はじめに

近年, 高齢化が進む中で, 認知症を患う高齢者の増加が問題になっている. 認知能力を維持, 増進するためには音声による他者との会話が有効であることが従来より指摘されている. このことから, 高齢者の話し相手になってくれるような雑談対話システムが求められている [1]. こうした雑談対話システムでは, ユーザがより長く対話を継続したくなるように適切かつ自然な応答を生成することが重要である.

本研究では, 雑談対話システムの応答音声を従来の読み上げ調から話し言葉調に変換することにより応答音声の自然性を向上させる手法を提案する. 特にフィラー（“あー”や“えっと”といった場繋ぎ的に発声される単語）やポーズ（無音による間）に注目し, これらを応答音声の適切な箇所に挿入するモデルを機械学習によって構築する.

2 関連研究

フィラーやポーズが音声の聞き取りやすさや理解しやすさに影響を与えることは，従来より指摘されている [2]. 伊藤ら [3] や Shiwa[4] らは, 音声対話システムや対話ロボットにおいて, 次文を生成するまでの待ち時間にシステムにフィラーを発声させることで, システムが動作していることをユーザに示すことができ, ユーザに安心感を持たせる上で有効であることを示している. また, 吉田ら [5] は音声合成による読み上げにおいて, ポーズがテキストの内容を感覚的, 意味的に捉えやすくする役割を持ち, 音声合成の自然性を高めることを示した. これらの先行研究からも, 音声対話システムを設計する上で, フィラーやポーズを適切に扱うことは重要である. 本研究では, 音声対話システムにおいて, 特に応答文の文中に適切にフィラーやポーズを挿入することで, 自然さや聞き取りやすさを向上させる手法について検討する.

3 提案手法

3.1 学習モデルの構築

以下の手順で, 応答音声の適切な箇所にフィラーやポーズを挿入するための分類器を作成する.

1. 単語列に対して, 各単語の直後に間（ポーズ）やつなぎ語（フィラー）を入れるかどうかを “0” と “1” でラベル付けする.
2. ラベル付けされた系列データを教師データとして学習を行う.

単語の並びをラベル列にすることで系列ラベリング問題として定式化し, モデルの学習には連続的なデータを扱う場合に効力がある機械学習アルゴリズムの RNN/LSTM を使用する. LSTM への入力は 100 次元の分散表現で, ウィキペディア日本語版（jawiki[1]）をコーパスとした Word2Vec[6] によって学習したものを使用し, 隠れ層のサイズは 100 次元とした. ラベル付けの例を表 3.1 に示す.

表 3.1: 単語列のラベリングの例

入力	今日	は	(えー)	良い	天気	だ
単語列	今日	は	F	良い	天気	だ
ラベル列	0	1		0	0	0

3.2 コーパス

学習用データとして日本語話し言葉コーパス（CSJ）に含まれる学会講演（987 講演）を用いる. 学会発表

[1] https://dumps.wikimedia.org/jawiki/latest/jawiki-latest-pages-articles.xml.bz2

では, 発表者は十分に練習がなされているため, 聞き取りやすい発話であることが考えられる. 本研究では, 分類器の学習にはフィラーやポーズ, 言い澱みや言い直しなどを区別せずに全て同一の記号（<F>）として扱った.

3.3 モデル評価

学習したモデルに文章を入力した例を表 3.2 に示す. いくつかの名詞や助詞の後にはフィラーが挿入される確率が他よりも高いという結果が得られたが, フィラーが挿入される確率が 50%を超えることはなかった. これは話者によっては同じ単語でもフィラーの有無が異なることや同じ話者であっても同じ単語の後にフィラーを必ずしも含まないことが原因である. そこで, 応答音声の生成の際には閾値としてある値を定め, フィラーが挿入される確率が閾値を超えていればフィラーを挿入するようにした. 今回は閾値を 0.20 とした.

また, 学習に用いたコーパス内で出現頻度が低い単語の後にフィラーが挿入されていると, その単語の後にフィラーが挿入されやすくなる. 特に名詞の場合, 名詞の後にフィラーが挿入される確率は一般的に低くなるため, 閾値を他よりも高くするなどの対策が考えられる.

表 3.2: モデルに文章を入力した例

単語	品詞	フィラーが挿入される確率
フィラー	名詞	0.1935505
が	助詞	0.2918799
挿入	名詞	0.1229687
さ	動詞	0.1197949
れる	動詞	0.1194425
確率	名詞	0.1196922
を	助詞	0.1347695
出力	名詞	0.1193897

4 評価実験

学習したモデルによって生成された応答音声の聞きやすさを確認するため評価実験を行った.

4.1 実験概要

実験に使用する文章は CSJ から抜粋したもの（学会講演とインタビュー）と著者が考案したものとで計 10 個用意し, 各文章それぞれについて 3 つのパターンを用意する.

1. 生成元となる文章（フィラーなし）
2. 適当な位置にランダムにフィラーを挿入した文章
3. モデルを使用してフィラーを挿入した文章

表 4.1 に上記の例を示す. 音声の出力には音声合成ソフトである VOICEROID ＋結月ゆかり ex[2]を使用し, <F>記号（フィラー/ポーズを意味する）の部分に 0.5 秒のポーズを挿入するようにした. 11 人の被験者に各パターンの聞きやすさと自然さを 5 段階で評価してもらい, フィラー（ポーズ）の有無や位置によってどのように聞きやすさが変化するかを確認した.

表 4.3: 実験に使用する文章の内訳

文章の種類	文章数
CSJ（インタビュー）から抜粋	4
CSJ（学会講演）から抜粋	3
著者考案	3

4.2 実験結果

各パターンの聞きやすさと自然さの平均値を表 4.4 に示す.

表 4.4: 各パターンの評価結果

	聞きやすさ	自然さ
原文（フィラーなし）	3.256	3.439
ランダム	1.805	2.146
提案手法	3.634	3.866

ポーズを含まない文章とランダムにポーズを挿入した文章とを比較すると違いが顕著に見られた. 適当な位置にポーズが挿入されているよりもポーズを含まない方が聞きやすく, またポーズを含まない文章とモデルを使用してポーズを挿入した文章とを比較すると, モデルを使用した方がより聞きやすく自然であるという結果となった. 応答音声の発話はポーズを全く含まない場合でも聞きにくいということはなく, 長文でない限りは聞き取りが可能であるが, ポーズを適切な位置に含むことでより聞きやすくなることが実験からわかった.

[2]https://www.ah-soft.com/voiceroid/yukari_ex/

表 4.1: フィラーを挿入した文章の例

原文	フィラーが挿入された文章とそうでない文章でどのように聞きやすさが変化するかの実験です
ランダム	フィラーが挿入さ<F>れた<F>文章とそう<F>でない文章で<F>どのよう<F>に聞きやす<F>さが変化<F>するか<F>の実験です
提案手法	フィラーが<F>挿入された文章と<F>そうでない文章で<F>どのように<F>聞きやすさが<F>変化するかの実験です

表 4.2: 読み上げた文章の例

文章番号	
2	明日の天気をお知らせします明日の鳴門は平均気温１４度で断続的に雨が降るでしょう
5	これでどんな風に交互作用が出てるのかっていうのを分かり易くする為にこちらに棒グラフを用意しているんですけれども
7	次に音源に注目いたしまして先程定義しました距離との関係について考えたいと思います

次に文章ごとの平均値を抜粋して表 4.5〜4.6 に示す.

表 4.5: 文章ごとの平均値（聞きやすさ）

文章番号	原文	ランダム	モデル
1	3.636	1.727	4.000
2	2.727	1.545	4.091
7	3.000	1.429	3.286
9	3.143	1.429	3.143
10	3.143	1.571	4.000

表 4.6: 文章ごとの平均値（自然さ）

文章番号	原文	ランダム	モデル
1	3.909	1.818	4.545
2	2.727	1.909	4.455
7	3.286	1.714	3.429
9	3.571	2.286	3.429
10	3.143	2.286	3.857

先述したように, モデルを使用してポーズを挿入した文章がもっとも聞き取りやすく, 次いでポーズを含まない文章となっている. しかし, 読み上げる文章によってはモデルを使用してもあまり聞きやすくなっていないものがみられた. このことからポーズの挿入によって応答音声の聞きやすさが必ずしも向上することはなく, ポーズの挿入が行われる元の文章の生成が重要であると考えられる. 今回はポーズの挿入のみを行ったが, フィラーを挿入する場合, フィラーの種類は直前の母音に影響されやすいことが知られている. したがって, 文脈を考慮したフィラーを選択することで自然性を高めることができると考えられる. また, 通常の音声合成器でフィラーの合成を行うと, 抑揚の不自然なフィラーが生成されてしまう場合がある. フィラーに特化した音声合成器を用いることでさらに聞きやすさや自然さを改善できる可能性がある [7].

5 おわりに

本研究では, 応答音声にポーズを挿入することによる聞きやすさ向上の手法を提案した. 日本語話し言葉コーパス（CSJ）の学会講演をコーパスに用い, 系列ラベリング問題として学習モデルを構築した. 評価実験では, ポーズを含まない原文に対しモデルを使用してポーズを挿入した文章とランダムにポーズを挿入した文章を作成し, それぞれの聞きやすさと自然さの評価を行った. 実験の結果, モデルを使用したポーズの挿入が音声の聞きやすさの向上に有効である場合が多いが, ポーズが挿入される原文によっては聞きやすさが向上しにくい場合もあることが確認された. 今回はCSJ に含まれる学会講演を使用し学習を行ったが, 他のコーパスを使用した場合の検証実験や学習モデルの評価方法の見直しが今後の課題として挙げられる.

謝辞

本研究の一部は, 総務省戦略的情報通信研究開発推進制度（SCOPE）の委託を受けて実施した.

参考文献

[1] 比企野純太, 中野有紀子, 安田清他, 会話エージェントを利用した認知症患者のためのコミュニケーション支援 情報処理学会全国大会講演 論文集. vol.2011. no.1. pp.195196. 2011.

[2] M.Somiya, K.Kobayashi, H.Nishizaki and Y.Sekiguchi, ”The Effect of Filled Pauses in a Lecture Speech on Impressive Evaluation of Listeners”, Proc. of INTERSPEECH2007, pp.2673-2676, 2007.

[3] 伊藤敏彦, 峯松信明, 中川聖一, 間投詞の働きの分析とシステム応答生成における間投詞の利用と評価, 日本音響学会誌, vol.55, no.5, pp.333-342, 1999.

[4] Toshiyuki Shiwa, Takayuki Kanda, Michita Imaia, Hiroshi Ishiguro and Norihiro Higata, ”How quickly should a communication robot respond?”, I nternational Journal of Social Robotics, vol. 1, no. 2, pp. 141-155, 2009.

[5] 吉田有里, 奥平康弘, 田村直良, 音声合成による朗読システムに関する研究 情報科学技術フォーラム講演論文集, vol.8, no.2, pp.377-380. 2009.

[6] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” Proceedings of the International Conference on Neural Information Processing Systems, pp.31113119, 2013.

[7] Jordi Adell, David Escudero and Antonio Bonafonte, ”Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence”, Speech Communication, vol.54, no.3, pp.459-476, 2012.