4Q-06

合成音声へのポーズ自動挿入に向けた自由発話の分析

穴沢 早紀† 内田 ゆず† 青木 直史‡

北海学園大学† 北海道大学‡

1．はじめに

近年，様々なテキスト音声合成システムが開発されている．その中でも個人の特徴を再現する音声合成システムでは同じ内容でも単に言語情報のみを伝えるのではなく，個々人のポーズや抑揚などの非言語情報の再現が重要となる．従来の音声合成システムは特に書き言葉のような聞き手に意図が伝わりやすい構文的なテキストの再生を目的としていることが多い．

こうしたシステムは事前に音声を録音し，モデルデータを作成する必要がある．その際，所定のテキストを読み上げる形式が一般的だ．一方で，非流暢的な自由発話を録音，再現するシステムは多くない．朗読よりも会話のような自由発話の方が本人の特徴をより捉えられることができるため，本研究では個人の自由発話から話し言葉を収集しそのポーズ位置に着目して分析する．

話し言葉においてポーズは言語的な区切りとは無関係な位置に現れることがあり，言語的なまとまりになっているとは限らない．日本語話し言葉コーパスに収録されている講演音声上でポーズが節境界[1]に対応しているのは全体のポーズ数に対して５０％である一方，３５％のポーズ位置には言語的役割を確認することはできない[2]．このようにポーズの生起要因は多様である．

2．ポーズ位置の解析

2.1 解析の手順

自由発話の音声を解析するにあたり，用意した２つの物語と１枚の写真を被験者がアドリブで説明した音声を録音した．被験者は２０代の男性２人，女性３人の計５人である．

録音内容を書き起こし，形態素解析器MeCab[3]を用いて形態素解析した．その後，節境界ラベルを第一著者が付与した．節境界の種類は日本語話し言葉コーパスに用いられている節境界ラベルを使用した．この節境界ラベルでは節境界直後の切れ目の大きさから「絶対境界」「強境界」「弱境界」と大きく３つに分類される．絶対境界とは発話の完全な区切り目にあたる部分でいわゆる文末にあたる表現である．強境界は「〜ですが」のような強い文末表現ではないが構造的に大きな切れ目となる節境界である．また弱境界は「〜なら」のような強境界ほど強い切れ目ではないが内容次第では稀に切り目となり得る境界であることを指す．例を表１に示す．ここで「sp」をポーズ，「[ ]」を絶対境界「<>」弱境界とする．

表１ データ例

開始時間	終了時間	発話内容	発話時間
53.0968	54.2467	しかし<接続詞>どの壺も	1.149887
54.2467	55.3332	sp	1.086440
55.3332	56.7086	口を下にして<テ節>	1.375465
56.7086	57.2249	sp	0.516326
57.2249	58.8684	底を上にしています[文末]	1.643447

音声のポーズ位置の決定はVoice Label Editor[4]を用い，第一著者が人手で行なった．なお，ポーズの基準は日本語話し言葉コーパスで設定されている200ms（0.2秒）以上のポーズが生じている位置，あるいは文法的な文末形式に続いて50ms(0.05秒)以上のポーズが生じている位置とする．また，語断片や言い直しなどの，発話誤りは除外した．

2.2 結果

ポーズ箇所に出現した品詞を被験者ごとに図１に示す．ポーズ位置を品詞別で見ても大きな差はなく品詞のみの比較では個人差を特定するのは難しいことが明らかになった．

全体に共通する特徴として，物語の説明をした際，「『待ってて』sp と言った」のように本文

Analysis of Spontaneous Speech for Automatic Pause Insertion in Synthesized Speech
†Saki Anazawa, Yuzu Uchida・Hokkai-Gakuen University
‡Naofumi Aoki・Hokkaido University

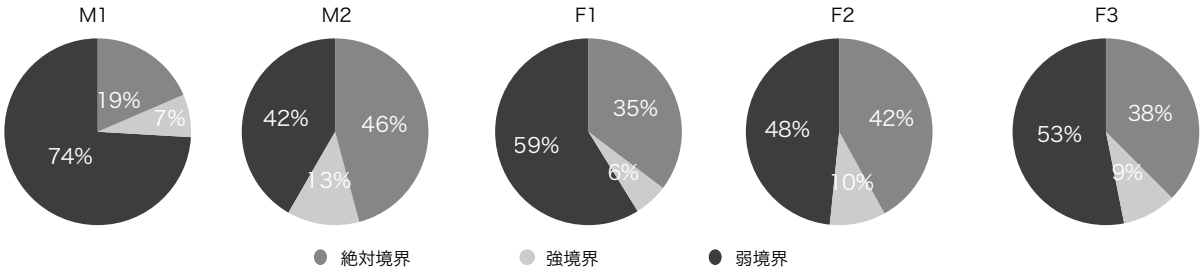

図２　節境界

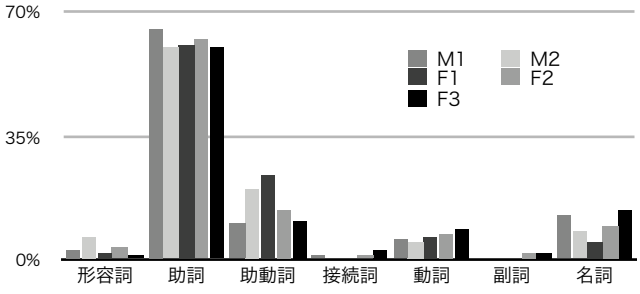

図１　ポーズ位置と品詞

中のセリフやナレーションを抜粋した内容の直後にポーズが入る箇所が多く，その直後の助詞にはポーズが挿入されないことも見受けられた．

発話時に文節中など不自然にポーズが入っている箇所が存在したが，その後に話す内容について考察している間であると考えられる．

５名分のデータについて，節境界ラベルの絶対境界，強境界，弱境界の割合を図２に，単語数とポーズ数を表２に示す．表からM2のポーズ割合が他よりも少ないことが見て取れる．この被験者は学習塾の講師をしており，日常的に子供に対して指導を行なっている．説明に慣れている人は単語数に対してポーズ数が少なく，まとまった説明が可能であると考えられる．また，表と図からM1，F1の単語数が多く弱境界の割合が高いことがわかる．これは以下に示すように，発話の際に強い切れ目をあまり用いない話し方をしているためだと考えられる．

F1「しかし<接続詞>どの壺も sp 口を下にして
<テ節> sp 底を上にしています[文末]」

F3「口が下に置いてあって<テ節>底が上になっている状態で sp 並んでいる[文末]」

３．おわりに

本研究では，個人の特徴を再現するテキスト音声合成システム構築のために複数の被験者のポーズ位置の特徴を分析した．まず，５人の被験者の

表１　単語数とポーズ数の関係

	M1	M2	F1	F2	F3
単語数	530	405	620	443	404
ポーズ数	126	60	129	101	87
ポーズ割合	23.8%	14.8%	20.8%	22.8%	21.5%

自由発話の節境界とポーズ位置の関係を分析し，品詞のみでは個人差を特定することは困難だが，ポーズの割合から説明にポーズ位置における弱境界の出現割合が高ければ単語数が増加する傾向があることを明らかにした．

今後は，節境界において個人の特徴があることを元にポーズ挿入モデルを作成し，実証実験を行う予定である．

謝辞

本研究の一部は公益財団法人セコム科学技術振興財団の助成を受けたものである．

参考文献

[1] 丸山岳彦, 高梨克也, 内元清貴: 第 5 章 節単位情報, 国立国語研究 所報告書 124 日本語話し言葉コーパスの構築法，pp.255–322．国立国語研究所 (2006).

[2]西光 雅弘, 高梨 克也, 河原 達也: 係り受けとポーズ・フィラーの情報を用いた話し言葉の段階的チャンキング,情報処理学会研究報告書,Vol.2005, No.127, pp.247-252(2005).

[3]Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237 (2004).

[4]青木 直史, 須藤 健次: 音韻ラベリングのための音声波形エディタの開発,電気・情報関係学会北海道支部連合大会(2016).