様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 １７ 日現在

機関番号：９４３０５
研究種目：若手研究
研究期間：2020～2021
課題番号：２０Ｋ１９９０３

研究課題名（和文）対話型AIのための音声と身体表現の同時生成に基づく自然なインタラクションの実現

研究課題名（英文）A study on a response generation method based on simultaneous generation of speech and physical expression for conversational AI

研究代表者

千葉　祐弥（Chiba, Yuya）

日本電信電話株式会社ＮＴＴコミュニケーション科学基礎研究所・協創情報研究部・研究員

研究者番号：３０７８０９３６

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：近年対話システムの分野において盛んに研究が行われているニューラルベース応答生成において，ユーザ発話の言語的情報と韻律情報を考慮する音声応答モデルを検討した．提案手法では，ベースラインよりも自然音声に近いF0系列が得られることを確認した．さらに，表情制御信号を扱えるように提案手法を拡張したマルチモーダル応答生成モデルを検討した．実験により，入力情報として複数のモダリティを考慮することでモデルの性能が向上できる可能性を示唆する結果を得た．加えて、マルチモーダル情報を利用した応答タイミング推定モデルを提案した．全体で6件の国内学会・研究会発表，4件の国際会議発表，1件の特許出願を行った．

研究成果の学術的意義や社会的意義
本課題では，近年盛んに研究されているニューラルベースの応答生成技術が言語情報だけでなく韻律や表情といった非言語情報も扱えること，また人間のコミュニケーションにおける社会的な現象を考慮できる可能性があることを示した．加えて，そのような非言語ベースの応答生成において効果的にモデルを学習するためのデータ拡張手法，自然なタイミング・間での応答を実現する応答タイミング推定手法も提案し，それぞれ一定の効果が得られた．これらの検討より，非言語情報を取り入れた対話システムの応答生成研究における有益な知見を提供できたと考える．本研究の成果は今後ますます重要性を増す対話システムの自然性の向上に寄与するものである．

研究成果の概要（英文）：This study constructed a spoken response generation method using linguistic and prosodic information of the user's utterance based on the neural conversational model, which is actively studied for dialogue systems. Our experiments confirmed that the proposed method can produce F0 sequences that are closer to natural speech than the baseline. Then, our research group expanded the spoken response generation model to a multimodal response generation model, that can consider the facial expression control signals. Experimental results suggested that the performance of the model can be improved by considering multimodal information. Additionally, we also proposed a response timing estimation model based on the dialogue context encoder and the continuous LSTM. We have presented six papers at domestic conferences and workshops, four papers at international conferences, and applied for one patent.

研究分野：対話システム

キーワード：音声対話システム　マルチモーダル情報処理　応答生成

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年，Amazon Alexa や対話ロボットなどの対話型 AI が社会に浸透しつつある．しかしながら，このような対話システムが違和感なくユーザと会話を行い，人間的なインタラクションを行うためには未だ多くの課題が残されている．問題の一つとして，現在の対話システムが姿形を持たず文字言語に大きく依存したインタラクションを行うため，表情や視線の動き・抑揚・間(ま)といった発話の非言語情報を対話の文脈に合わせて適切に扱えていないということが挙げられる．これは例えば，ユーザが高揚して会話している場合であっても対話システムは画一的で一本調子に応答してしまうといった，対話の噛み合わなさを感じさせる要因となっている．対話システムが対話相手と関係を築くことができるような社会的な存在になるためには，人間のように感情表現や身体表現を用いた違和感のないインタラクションができる必要がある．

２．研究の目的

本研究課題では，大規模データから学習されるニューラル応答生成モデルを様々な非言語情報が扱えるように拡張したマルチストリーム応答生成モデルを提案する．提案モデルは，申請者がこれまでに収録した大規模マルチモーダル雑談対話データベースを用いることで，人間同士の会話から文脈にふさわしい発話内容(文字言語)と表情や感情・視線などの非言語表現を自動で学習できるため，人の会話に近い表現豊かな応答生成システムの構築に有用である．

３．研究の方法

1 年目は韻律情報を用いた基礎モデルの構築・評価を行う．ベースのシステムとして，まずは言語情報に基づく一般的な応答生成モデルを学習する．モデルは Attention 機構を備えた Seq2Seq モデルとする．ネットワークの事前学習には Twitter から収集されたリプライデータ 200 万件を用い，申請者が保有するマルチモーダル雑談対話データに含まれる約 10 万応答対でファインチューニングする．この際に生成された応答の品質を確認し，従来と同等程度の性能が出るように調整する．話し言葉の言語生成に問題がある場合はフィラーや言い直しなどの話し言葉特有の言語現象を学習データから可能な限り取り除く．実装には深層学習モデルの開発ツールである PyTorch を利用する．その後，構築したベースモデルを非言語情報のうち音声合成の分野で効果が確認されている韻律情報を導入した応答生成モデルに拡張する．モデルの学習には言語情報と非言語情報の時間的対応が取れたデータが大量に必要であるため，申請者が保有するマルチモーダル雑談対話コーパスを用いる．このとき，非言語情報と言語情報の生成シンボルに関して別々の損失関数を用いるマルチタスク学習を行う．モデルの構造に関しても，それぞれの情報に関して個別の BLSTM を利用するネットワークを検証する．1 年目の最後には対話実験に基づく応答評価を行う．応答評価では，ベースモデルから生成された発話を既存手法に則って音声合成器に入力した場合と，提案モデルで発話とともに生成された韻律制御ラベルを用いて音声合成を行う場合を比較する．この際，オープンドメインの対話システムの応答性能の評価に一般的に利用される，自然性，ユーザ満足度，エンゲージメントの度合いに関する 5 段階評価を実施する．

2 年目は提案する応答生成モデルをさらに表情・その他非言語情報へと拡張し，マルチストリーム応答生成モデルを構築する．非言語情報として表情・視線・間(ま)・ジェスチャなどの対話システムにおいて有用なものを選別する．従来の研究では，特に表情などは発話全体に対して単一のラベルが付与されるのが一般的である．そこで本研究では，発話全体のラベルを用いた場合と，発話中の変動を提案手法の枠組みで考慮した場合を比較する．最後に，対話実験による総合的な応答評価を行う．ここでは前年度構築した韻律情報に基づくモデルと，マルチストリーム応答生成モデルを比較する．この時，考慮する非言語情報の組み合わせを変えることで，それぞれの非言語情報がユーザ評価に与える効果も同時に検証する．

４．研究成果

<一年目>

<u>Encoder-Decoder モデルに基づく音声応答システムの構築</u>

まず，自然発話音声を対象とした応答生成モデルを学習するためのデータ拡張手法を検討した．この検討では，Twitter から収集されたツイート・リプライ対に対して，フィラー挿入を行うモデルを学習した．提案モデルは目標コーパスである Spontaneous Multimodal One-on-one Chat-

表 1 フィラー挿入モデルの客観評価実験の結果

モデル	予測条件	W	挿入位置の評価 適合率	再現率	F 値	種類の評価 適合率	再現率	F 値	挿入位置と種類の評価 適合率	再現率	F 値
CRF and N-gram	最尤条件	1000	0.575	**0.516**	0.544	0.243	0.174	0.150	0.011	0.044	0.018
		5000	0.582	0.515	0.546	0.281	0.178	0.154	0.011	0.044	0.018
		10000	0.664	0.507	**0.575**	0.293	0.176	0.151	0.068	0.043	0.023
		15484	0.601	0.486	0.537	0.301	0.176	0.151	0.159	0.050	0.036
	ランダム条件	1000	0.317	0.353	0.334	0.176	0.173	0.170	0.049	0.050	0.049
		5000	0.300	0.354	0.325	0.179	0.175	0.171	0.043	0.046	0.044
		10000	0.405	0.419	0.412	0.187	0.181	0.178	0.066	0.061	0.062
		15484	0.458	0.443	0.450	0.177	0.173	0.170	0.075	0.063	0.067
BLSTM_POS			0.631	0.453	0.527	-	-	-	-	-	-
BLSTM_CAT			-	-	-	**0.424**	**0.395**	**0.399**	-	-	-
CRF (最尤, W=10000) and BLSTM_CAT			-	-	-	-	-	-	0.217	**0.207**	0.208
Cascade (再学習なし)			0.574	0.450	0.505	-	-	-	0.314	0.200	0.201
Cascade (出力層のみ再学習)			0.701	0.357	0.473	-	-	-	0.315	0.203	0.240
Cascade (全層を再学習)			0.661	0.338	0.447	-	-	-	0.345	0.197	**0.249**
Simultaneous			**0.808**	0.190	0.308	-	-	-	**0.471**	0.112	0.141

talk (SMOC)のフィラー挿入位置を学習する．ここでは，Cascade モデルと Simultaneous モデルの 2 つのモデルを提案した．結果を表 1 に示す．結果より，提案モデルがベースラインと比較して F 値ベースで高い性能を示すことを確かめた．特に，挿入位置と種類の両方を考慮して評価した場合，フィラーの位置と種類を推定する BLSTM を結合し，対象コーパスで再学習したモデルが最も高い性能であった．

続いて，ユーザ発話の言語的情報と韻律情報を入力する音声応答モデルの検討を行った．提案モデルは Encoder-Decoder 型のモデルである．ユーザ発話の単語系列とその平均対数 F0 系列を入力とし，応答発話とその韻律を制御する差分 F0 コンテキスト系列を出力するモデルを 2 つ(同時モデル，連結モデル)提案した．客観評価実験の結果を表 2 に示す．対数 F0 は連結モデルでベースラインの合成音声よりもスコアが高かった．したがって，この手法では，ベースラインよりも自然音声に近い F0 系列が得られていると示唆される．一方で，同時モデルはベースラインの音声よりも客観的な指標ではスコアが低かった．同時モデルに比べて連結モデルは学習が容易であることや，連結モデルにおいて韻律生成の Decoder に BLSTM を使用できることが要因であると考えられる．また，対数 F0 の RMSE 自体は，発話内容から直接単語ごとの差分 F0 コンテキストを出力する場合が最も低かった．続いて，提案手法の聴感上の優劣を測るため，客観評価を実施した．実験結果を図 1 に示す．1 標本 t 検定では全ての手法間の比較において有意差が見られた．したがって，差分 F0 コンテキストを用いることでより対話に近い韻律の音声が生成できていることが確かめられた．また，連結モデルの比較から，ユーザ発話の言語・韻律情報を入力することでより対話として適切な韻律の音声が生成できることが示された．

Objective evaluation of language generation

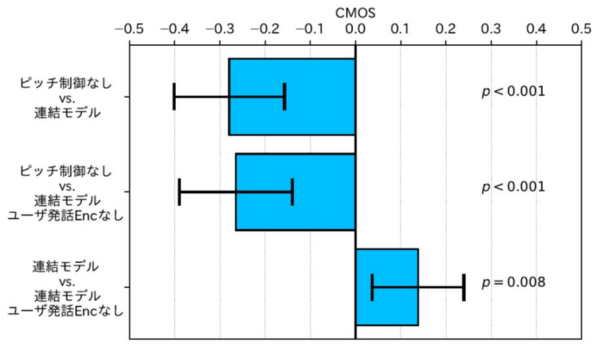
Model	Modality	PPL*	BLEU-4	Dist-2
Unified model	L	258.16	1.61	21.31
	P	**97.60**	0.93	0.22
	L+P	257.09	1.70	**21.33**
Separated model (Language generation)	L	255.38	**1.78**	20.08

Objective evaluation of prosody generation

Model	Modality	ΔLF0 MSE*	LF0 RMSE* [cent]
Unified model	L	0.0615	438.9
	P	0.0606	442.5
	L+P	0.0618	433.8
Separated model (ΔF0 context generation)	L	0.0501	407.2
	P	0.0503	407.8
	L+P	**0.0492**	406.2
	w/o Encoder	0.0505	**400.0**
Baseline	–	–	433.3

表 2 提案する音声応答モデルの客観評価結果: L は言語情報を入力に用いた場合の結果，A は韻律情報を入力に用いた場合の結果を示す．*は teacher-forcing によって得られた単語系列によって評価したことを示す．

図 1 提案する音声応答モデルの主観評価実験の結果: 正のスコアは前者のシステムが，負のスコアは後者のシステムがより選好されたことを示す．



マルチモーダル情報を利用した応答タイミング推定

文献 (Roddy and Harte, 2020)のモデルに対して，対話コンテキストエンコーダを導入したモデルを提案した．これによって，先行研究で問題であった，推定時にシステムの応答内容が決まっていなければならない問題を緩和する．モデルの性能は，テストデータに対する正解の応答タイミングと推定されたタイミングの差である Mean Absolute Error (MAE)によって評価した．結果を表 3 に示す．結果より，提案モデルに導入した対話コンテキストエンコーダは将来のシステム発話の発話意図を間接的に表現するものであるため，単体の性能は先行研究に及ばないものの，画像情報を組み合わせることで先行発話と同等の性能が得られることが示された．

表 3 提案モデルを含む手法間の MAE の比較．RTNet は従来手法 (Roddy and Harte, 2020)を示す．RAND は応答タイミングをランダムに決めた場合の結果である．

Method	Encoder	MAE (Avg. ± SE)
RTNet	Response Encoder	0.595 ± 0.015
+ visual	Response Encoder	**0.524 ± 0.028**
Proposed	Context Encoder	0.668 ± 0.006
+visual	Context Encoder	0.601 ± 0.026
Inference LSTM	w/o	0.686 ± 0.011
+visual	w/o	0.638 ± 0.004
RAND	–	1.219

<2 年目>

Encoder-Decoder モデルに基づくマルチモーダル応答システムの構築

昨年検討した音声応答モデルに対して表情制御信号を導入したマルチモーダル応答生成モデルを検討した．検討したモデルは Attention 機構を有する Encoder-Decoder 型のモデルである．検討手法は，ユーザ発話の単語系列と韻律・表情特徴量を順次入力する．韻律・表情特徴量は対応する単語区間の平均対数 F0，平均 AU である．その後，デコーダは最終時刻のエンコーダ内部状態を内部状態の初期値として韻律・表情制御信号の生成を始める．この時，デコーダには応答生成部から受け取ったシステム応答文の単語系列を順次入力する．これによって，システム応答の各単語に対応する韻律・表情制御信号を得る．韻律・表情制御信号は先行研究と同様のものを用いる．すなわち，韻律制御信号は単語ごとの差分 F0 コンテキスト，表情制御信号は単語ごとの平均 AU である．提案モデルの学習時には，人間同士の対話音声における発話と，あらかじめ生成した同一発話内容の音声合成との間で計算された差分 F0 コンテキストを用いる．ネットワークから生成された差分 F0 コンテキストを用いることで，自然発話に近づくように補正されたシステム応答音声が生成できる．

客観評価実験の結果を表 4 に示す． 表 4 において，L，A，V はそれぞれ入力における言語特徴量，韻律特徴量，表情特徴量を表す．また，ST はシングルタスク学習条件，MT はマルチタスク学習条件の結果を示す．表より，韻律制御信号においては，シングルタスク学習条件では言語と表情特徴量を用いた場合，マルチタスク学習条件ではすべての特徴量を用いた場合に最も性能が高かった．また，表情制御信号においては，シングルタスク学習条件では言語と表情特徴量を用いた場合，マルチタスク学習条件では韻律特徴量と表情特徴量を用いた場合に最も性能が高かった．したがって，入力情報として複数のモダリティを考慮することでモデルの性能が向上できることが示唆される．また，いずれの条件においても，最も性能の高い条件では表情特徴量が含まれていた．

続いて，マルチタスク学習条件とシングルタスク学習条件を比較した．結果から，マルチタスク学習条件においてすべての入力を用いることで，韻律制御信号・表情制御信号のいずれにおいてもシングルタスク学習条件を上回る性能が得られた．このことから，マルチモーダル情報を用いたマルチタスク学習はエージェントの非言語行動の生成に関して有用であることが示唆される．しかしながら，生成タスクにおいては必ずしも客観評価性能が高いモデルの出力が主観的にも高い評価を得るとは限らない．そのため，今後は実際の生成例を主観的に評価する必要がある．

表 4 マルチモーダル応答生成モデルのテストセットに対する MSE (平均±標準誤差)．Δ LF0 は韻律制御信号，AU は表情制御信号である．L，A，V はエンコーダ入力における言語・韻律・表情特徴量を示す．ST，MT はそれぞれシングルタスク学習条件，マルチタスク学習条件の結果を表す．太字は同一の出力条件における最小の MSE である．また，w/o Enc. はエンコーダなしの条件である．

input\output	Δ LF0 (ST)	AU (ST)	Δ LF0 (MT)	AU (MT)
L+A+V	0.0534 ± 0.0002	0.2649 ± 0.0003	$\mathbf{0.0527 \pm 0.0001}$	0.2366 ± 0.0015
L+A	0.0530 ± 0.0001	0.2801 ± 0.0004	0.0530 ± 0.0003	0.2423 ± 0.0024
L+V	$\mathbf{0.0529 \pm 0.0004}$	$\mathbf{0.2628 \pm 0.0008}$	0.0538 ± 0.0007	0.2335 ± 0.0014
A+V	0.0541 ± 0.0004	0.2635 ± 0.0018	0.0534 ± 0.0004	$\mathbf{0.2206 \pm 0.0003}$
L	0.0533 ± 0.0003	0.2776 ± 0.0024	0.0531 ± 0.0002	0.2441 ± 0.0008
A	0.0541 ± 0.0003	0.2775 ± 0.0009	0.0537 ± 0.0004	0.2333 ± 0.0004
V	0.0537 ± 0.0003	0.2661 ± 0.0015	0.0544 ± 0.0003	0.2212 ± 0.0003
w/o Enc.	0.0547 ± 0.0004	0.2792 ± 0.0011	0.0542 ± 0.0007	0.2328 ± 0.0022

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計10件（うち招待講演　0件／うち国際学会　4件）

１．発表者名 Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, Akinori Ito
２．発表標題 Filler Prediction Based on Bidirectional LSTM for Generation of Natural Response of Spoken Dialog
３．学会等名 IEEE GCCE（国際学会）
４．発表年 2020年

１．発表者名 山崎善啓，千葉祐弥，能勢隆，伊藤彰則
２．発表標題 雑談コーパスを用いた双方向LSTMに基づくフィラー予測の検討
３．学会等名 音響学会秋季研究発表会
４．発表年 2020年

１．発表者名 山崎善啓，千葉祐弥，能勢隆，伊藤彰則
２．発表標題 言語・韻律情報の同時モデル化に基づく音声応答生成の検討
３．学会等名 人工知能学会 言語・音声理解と対話処理研究会
４．発表年 2020年

１．発表者名 矢作凌大，千葉祐弥，伊藤彰則
２．発表標題 先行発話を利用したマルチモーダル応答タイミング推定
３．学会等名 人工知能学会 言語・音声理解と対話処理研究会
４．発表年 2020年

１．発表者名
千葉祐弥，伊藤彰則

２．発表標題
対話者間の親密さに基づく言語・非言語的対話行動の分析

３．学会等名
人工知能学会 言語・音声理解と対話処理研究会

４．発表年
2020年

１．発表者名
山崎善啓，千葉祐弥，能勢隆，伊藤彰則

２．発表標題
言語・F0 特徴量系列を考慮したニューラル音声応答生成の検討

３．学会等名
音響学会春季研究発表会

４．発表年
2021年

１．発表者名
Ryota Yahagi, Yuya Chiba, Takashi Nose, Akinori Ito

２．発表標題
Multimodal dialogue response timing estimation using dialogue context encoder

３．学会等名
IWSDS（国際学会）

４．発表年
2021年

１．発表者名
Yuya Chiba, Yoshihiro Yamazaki, Akinori Ito

２．発表標題
Speaker intimacy in chat-talks: Analysis and recognition based on verbal and non-verbal information

３．学会等名
SemDial（国際学会）

４．発表年
2021年

１．発表者名 Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, Akinori Ito
２．発表標題 Neural spoken-response generation using prosodic and linguistic context for conversational systems
３．学会等名 INTERSPEECH（国際学会）
４．発表年 2021年

１．発表者名 渡辺稜哉，千葉祐弥，能勢隆，伊藤彰則
２．発表標題 マルチモーダル情報に基づくシステム応答の韻律・表情制御信号の生成に関する検討
３．学会等名 人工知能学会 言語・音声理解と対話処理研究会
４．発表年 2021年

〔図書〕　計0件

〔出願〕　計1件

産業財産権の名称 音声対話システムのための区分的韻律制御技術	発明者 山崎善啓，千葉祐弥，能勢隆，伊藤彰則	権利者 同左
産業財産権の種類、番号 特許、2021183018	出願年 2021年	国内・外国の別 国内

〔取得〕　計0件

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関