会話によるニュース記事伝達のための間の調整

高津 弘明[1]　横山 勝矢[1]　本田 裕[2]　藤江 真也[1,3]　林 良彦[1]　小林 哲則[1]

早稲田大学[1]　本田技術研究所[2]　千葉工業大学[3]

{takatsu,katsuya}@pcl.cs.waseda.ac.jp, Hiroshi_01_Honda@n.t.rd.honda.co.jp,
shinya.fujie@p.chibakoudai.jp, yshk.hayashi@aoni.waseda.jp, koba@waseda.jp

1　はじめに

会話によるニュース記事伝達において，割り込みを許容しながら快適なリズムで会話を進行させるための間の調整について検討する．

音声メディアは，伝え手による情報提供と，受け手の情報取得が同時に行われ（送受信の同時性），かつ提供した情報は瞬時に消え去る（揮発性）という特徴を持つ [1]．このため，受け手が取得できる情報の量と質は，伝え手と受け手の相互行為の時間構造，すなわち，どのようなやりとりを，どのような時間制約に従って行うかに大きく依存する．

本研究では，相互行為の時間構造に影響を与える要素の一つとして，話し手の「間」に注目する．話し手が入れる間が十分でない場合，聞き手は話の途中で質問したいことがあっても割り込む隙がないため質問を躊躇してしまい，相互行為は生じない．また，情報を理解するための時間がないため内容が頭に入らないという問題もある．一方で，不必要に間を空けてしまうと会話のテンポ，リズムが失われ，相互行為は活性化しない．このように，ニュース記事のようなまとまった量の情報を会話で伝える場合，相互行為の円滑化・活性化において，間の適切な制御は，会話の質を向上させるための本質的な課題と言える．

従来の音声対話システム研究において，漫才を対象にしたもの [2] など一部を除いて，間の問題はほとんど扱われて来なかった．システム発話の工夫によって相互行為を活性化させようとする観点では，聞き手からの相槌やうなずきなどを誘発する研究が行われているが，これも韻律制御に留まっており，間の制御までは扱っていない [3]．そこで，本研究では我々が情報伝達のために開発した即応性に富む会話システム [4] を基礎として，ユーザーが理解しやすく，発話中でも割り込みやすい間の実現を目指す．

本稿では，まず，2 章で人が理解を促す話し方をした際の間の取り方について調査した結果を述べるとともに，この間を実現する手法を提案する．次に，3 章で人がシステムの発話中に質問を割りこませるのに必要な時間について調査した結果を述べる．そして，4 章で以上の結果を踏まえて間の調整を行った音声と行わなかった音声を比較し，理解しやすさと質問しやすさの観点で評価した結果について報告する．

表 1: データセットの統計

	発話内文節間の間	発話間の間
データ数	5,327	507
最小値（秒）	0.000	1.172
平均値（秒）	0.210	2.657
最大値（秒）	2.680	4.283

2　理解を促すための間

従来の文単位の読み上げ音声合成器は，まとまった量の情報を会話で伝えることを想定した間の取り方になっていない．そこで，本研究ではニュース記事から発話シナリオを作成し，会話で相手に理解させることを強く意識させた上で声優に発話させることで間の入れ方の教師となるデータセットを作成した．発話間の間と発話内の文節間の間を推定するモデルを設計し，声優の間の入れ方を学習させることで，聴きやすく，割り込みやすい間を実現する．

2.1　ポーズ長タグ付きコーパスの作成

テクノロジー系のニュース記事 100 個を人手で要約・口語化し，発話のシナリオを作成した．このシナリオを女性声優に発話させ，その発話音声を収録した．収録にあたり，話者に対して以下の点に注意するように指示した．

- 話を聞いている相手に内容を伝える（理解させる）つもりになって発話すること
- 重要な箇所は強調し，あまり重要でない箇所はさらっと伝えるなど，メリハリのある話し方を意識すること
- 発話中でも割り込んで質問がくる可能性があり，質問がくると予想される箇所では不自然でない程度に間を入れること

収録した音声から文節間の間および発話間の間に関する時間情報を算出し，ポーズ長タグ付きコーパスを作成した．ここで，文節区切りには JUMAN++(Ver.1.02)[1] と KNP(Ver.4.18)[2] の結果を採用した．データセットの統計を表 1 に示す．

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++
[2] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

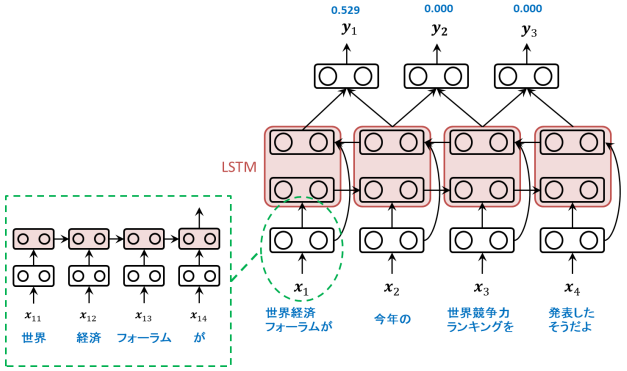

図 1: 発話内文節間ポーズ推定モデル

表 2: 発話内文節間ポーズ推定モデルの結果（RMSE）

追加素性	単語 BoW	文字 BoW	単語 RNN	文字 RNN
	0.252	0.260	0.251	0.233
品詞	0.234	0.240	0.231	0.224
活用形	0.239	0.242	0.230	0.230
活用型	0.245	0.255	0.237	0.229
カテゴリ	0.253	0.261	0.249	0.236
ドメイン	0.252	0.259	0.248	0.233
固有表現	0.250	0.260	0.250	0.236
親密度	0.263	0.264	0.250	0.243
難易度	0.260	0.262	0.249	0.242
品詞+活用形+活用型				**0.219**

2.2 発話内文節間の間の推定

情報の塊を会話で伝える場合，複数の発話に分けて情報を小出しにして伝える．ここでは，一発話内の文節間の間を人の発話の間に合わせることが目的である．モデルには双方向 LSTM を用いており，文節間の間は前後の文節の LSTM の状態に基づいて決まる（図 1）．モデルの構造として，文節を構成する単語や文字の Bag-of-Words（BoW）を入力として与えたときと，これらを LSTM でリカレントに埋め込んだときの平均二乗誤差平方根（RMSE）の比較を行った．さらに，単語や文字の基本素性の他に JUMAN++や KNP を適用して得られる品詞や活用形，活用型，カテゴリ，ドメイン，固有表現の他，『基本語データベース』[3] の単語親密度および『日本語教育語彙表 ver1.0』[4] の語彙難易度を追加素性として使用し，有効な素性の組み合わせについて検討した．なお，間の推定結果が読点「、」の与え方に左右されないように，入力文から「、」を取り除いた．実験では，埋め込み層および隠れ層の次元を 100 次元に設定し，活性化関数には tanh，最適化アルゴリズムには Adam を使用した．

10 分割交差検定により評価した結果を表 2 に示す．実験の結果から，文節の文字をリカレントに入力したときの結果が総じて良く，追加素性として品詞と活用形および活用型の情報を与えたときに最も RMSE が小さな値を示した．さらに，推定値と正解の誤差のヒストグラム（図 2）を確認したところ，0.2 秒未満の誤差の割合が 77%であった．

[3] https://hon.gakken.jp/reference/special/jiten/kihongo_db/index.html

[4] http://jhlee.sakura.ne.jp/JEV.html

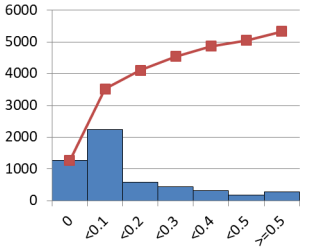

図 2: 文節間ポーズの誤差

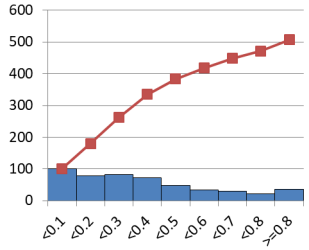

図 3: 発話間ポーズの誤差

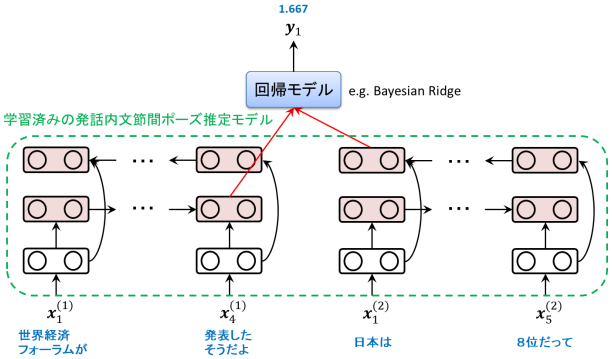

図 4: 発話間ポーズ推定モデル

2.3 発話間の間の推定

ここでは，発話間の間を人の発話の間に合わせることが目的である．発話間の間の役割として，発話内容を咀嚼・理解させることや質問誘発の他に，話題変更の合図が考えられる．そのため，現在の発話内容だけでなく，次の発話内容も考慮して間を決定する必要がある．そこで，現在の発話と次の発話の情報を用いて発話間の間を推定する（図 4）．ここで，特徴量としては 2.2 節で学習したモデルの LSTM の最後の隠れ層の値を用いた．予備実験の結果，現在の発話の順方向の最後の隠れ層と次の発話の逆方向の最後の隠れ層の組み合わせが最も良い結果を示した．また，scikit-learn[5] の回帰モデルを一通り比較した結果，BayesianRidge モデルが最も良い結果を示した．

10 分割交差検定で評価したところ，RMSE は 0.450 であった．間が前後の発話の長さに影響されるという研究報告もあり [5]，前後の発話の文字数，単語数，文節数を素性に加えて比較したところ，RMSE は文字数を加えた場合 0.440，単語数を加えた場合 0.436，文節数を加えた場合 0.437 を示し，発話長の有効性が確認された．さらに，単語数を加えたモデルにおいて，推定値と正解の誤差のヒストグラム（図 3）を確認したところ，0.5 秒未満の誤差の割合が 75%であった．

2.4 間の推定結果の例

10 分割交差検定で得られた間の推定結果の例を図 5 に示す．ここで，実際に声優が発話したときの間を赤色，システムの推定結果を青色，AITalk[6]（話者：のぞみ）によって与えられたデフォルトの間を緑色で示している．AITalk では，間は一瞬のポーズ，短めのポーズ，長めのポーズの 3 種類に分類される．これらの値

[5] http://scikit-learn.org/stable/

[6] https://www.ai-j.jp/about/

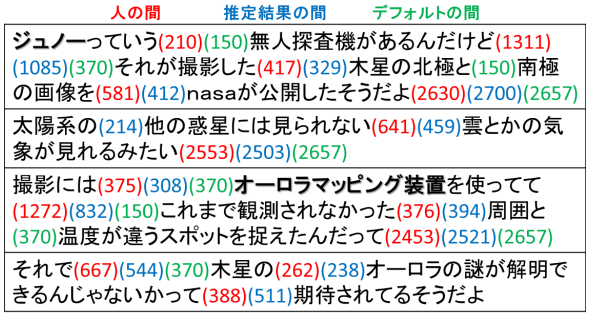

図 5: 間の推定結果 [ミリ秒]（100 ミリ秒以上の間）

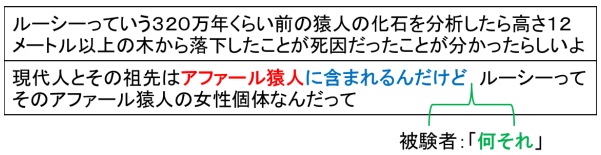

図 6: 質問に必要な間の調査実験における質問の例

は設定によって変えられるが，ここではデフォルト値（150 ミリ秒，370 ミリ秒，800 ミリ秒）を用いた．また，発話間の間のデフォルト値はデータセットから得られた発話間の間の平均値を与えた．

結果を比べてみると，提案モデルで推定した値の方が人の話し方に近い間の取り方になっていることが分かる．また，音声収録において指示した「質問が予測される箇所で不自然でない程度に間を入れる」という部分に関して，実際に声優は 1 発話目の「あるんだけど」の後，および 3 発話目の「使ってて」の後に十分な間を空けており，システムもそれを再現できている．なお，今回の実験では「、」を明示的に与えずに合成を行っているため，AITalk のデフォルトの合成結果には長めのポーズが出現しなかった．例えば，「オーロラマッピング装置」について「何それ」と聞き手が質問したくなったとする．このとき，デフォルトの間だと 150 ミリ秒しか間がないので質問を挟むことができないが，推定値だと 800 ミリ秒以上の間が空いているので，3 章で説明する実験結果（図 9）を参照すると，90%程度の割合で質問をカバーできることが分かる．

3 質問を挟みやすくするための間

実際に人がシステムの発話中に割り込んで質問する際，質問するまでにどれくらいの時間を必要とするかについて調査する実験を行った．

3.1 実験設定

難しい単語を含む重文の発話を対象に「何それ」と質問させ，質問に要する時間を計測した．例えば，図 6 のようなシステム発話があったとき，被験者は「アファール猿人」について「に含まれるんだけど」の後に，「何それ」と質問する．実験は 11 人の被験者に対して行い，合計 16 個の質問対象語（e.g. アファール猿人）を含む 14 個のニューストピックについて会話させた．ここで，実験の順序効果をなくすためにトピックの順番は被験者ごとにランダムに入れ替えて実施した．また，実験では，なるべく実際の会話と利用状況を似せるために，被験者の相槌や感想などの発話も許容するとともに，上記質問に対する回答も提示した．

3.2 実験結果

質問までの平均時間が早かった順に被験者を A から K まで番号付けし，質問するまでにかかった時間を被験者ごとに箱ひげ図で表したものを図 7 に示す．この結果から質問に必要な時間は被験者ごとに大きく異なることが分かる．次に，質問までの平均時間が早かった順に質問対象語を a から p まで番号付けし，質問するまでにかかった時間を質問対象語ごとに箱ひげ図で表したものを図 8 に示す．この結果から文脈依存性はそれほど大きくないことが読み取れる．次に，ポーズ時間と質問カバー率の関係を図 9 に示す．この結果から 0.7 秒のポーズを入れることで 80%の質問をカバーできることが分かる．しかしながら，被験者ごとに確認してみると，A,B,C,E のように 0.7 秒のポーズで 100%の質問をカバーできる被験者も入れば，K のように 44%しかカバーできない被験者もいる．

4 主観評価

音声合成器として AITalk（話者：のぞみ）を使用し，間の調整を行ったときと行わなかったときでどちらが「質問しやすかったか」「相槌を打ちやすかったか」「頭に入りやすかったか」について評価を行った．

4.1 間の調整方法

評価対象の発話シナリオに対して，以下の手順で間の調整を行った．

(1) 2 章で説明した手法を用いて，理解を促す間の取り方に調整
(2) 3 章の知見に基づき，質問が予想される箇所（知名度が低い用語を含む連用節の後）において，上記推定結果の間が 0.7 秒未満の場合，0.7 秒まで延長

4.2 実験設定

実験は 5 人の被験者に対して 5 トピックずつ実施した．「質問しやすかったか」の評価では，3 章で行った実験と同様に，ある用語について指定された箇所で「何それ」と質問し，質問しやすかった方を選択する．「相槌を打ちやすかったか」の評価では，対話中に「うん」や「へー」などの相槌を打ち，相槌を打ちやすかった方を選択する．「頭に入りやすかったか」の評価では，二つの音声を聴き比べ，頭に入りやすい間の取り方であった方を選択する．なお，どちらとも言えない場合は「どちらとも言えない」を選択する．ここで，比較する二つの音声について，どちらが間の調整を行ったものか分からないようにするために，ランダムに順番を入れ替えて実験を行った．

4.3 実験結果

それぞれの評価項目に対する実験結果を図 10，図 11，図 12 に示す．

質問しやすさに関する評価では，大半が推定値の間の方が良いと評価した．これは，2.4 節での分析からも

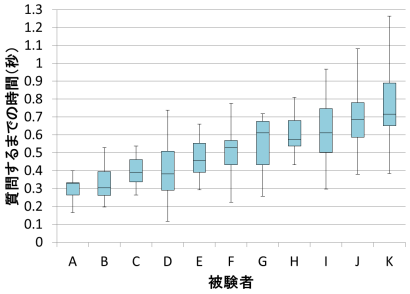

図 7: 質問に要する時間 (被験者ごと)

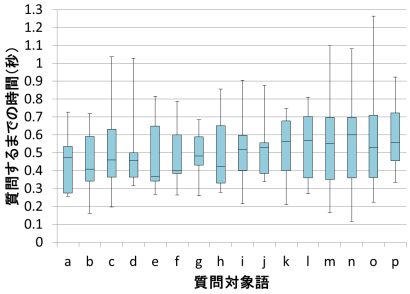

図 8: 質問に要する時間 (質問対象語ごと)

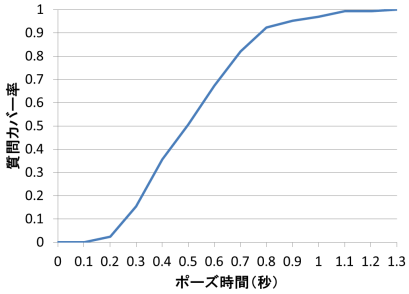

図 9: ポーズ時間と質問カバー率の関係

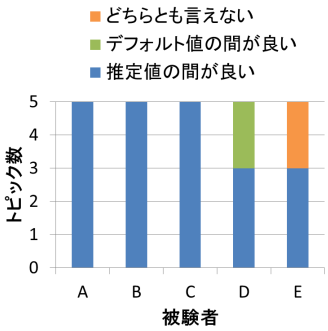

図 10: 質問しやすかった

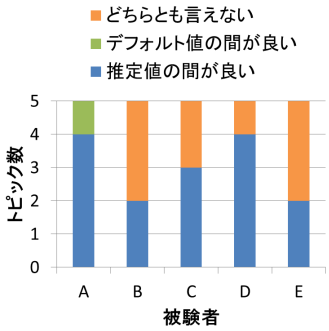

図 11: 相槌を打ちやすかった

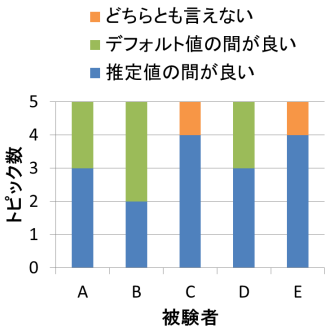

図 12: 頭に入りやすかった

分かる通り，デフォルトの音声では質問箇所で十分な間がないことに起因する．その一方で，被験者 D は質問箇所ですぐに質問しており，デフォルトの間でもシステム発話と衝突することなく質問できていた．図 7 が示すように，人によって質問するタイミングが大きく異なるため，今後はユーザーごとに間の取り方を調整する方法について検討する必要がある．

相槌の打ちやすさに関する評価では，推定値の方が良いという評価が得られた．しかしながら，「どちらとも言えない」という評価も比較的多かった．これは，「うん」のような一瞬で終わる相槌の場合，それほど長い間は必要なく，デフォルトの間でも十分に相槌を打つことができたためだと考えられる．

頭に入りやすかったかどうかに関する評価では，デフォルトの間の方が良いという被験者もいたが，過半数において推定値の方が良いという結果が得られた．今回の会話タスクでは，被験者は実験に集中していたため，システムの発話内容を理解するのにそれほど多くの時間と労力を必要としなかったと考えられる．しかしながら，運転しながらや料理しながらといった利用状況では良いとされる間の取り方が変わる可能性があるため，今後は多重タスク下での間の取り方についても検討していきたい．

5 おわりに

ニュース記事のような情報の塊を話し手が主導となって伝えるタイプの会話において，相互行為の時間構造は話し手が支配することになる．聞き手は話し手が支配する時間構造の中で “隙”（間）を見つけて興味や理解状態を話し手にフィードバックしながら，欲しい情報を獲得していく．しかしながら，この間の取り方が不適切であると，質問したいタイミングで質問できない，内容が頭に入ってこないという問題が発生する．そこで本研究では，相手に理解させることを強く意識した上で発話したときの人の話し方に間を合わせることで聴きやすい間を実現しつつ，質問が予想される箇所では必要十分な長さの間を設けることで，会話のリズムを損なうことなく割り込みやすい間の実現を試みた．音声合成器として AITalk を用いて，間の調整を行ったときと行わなかったときでどちらが「質問しやすかったか」「頭に入りやすかったか」主観評価を行ったところ，どちらについても調整を行った間の方が良いという結果が得られた．

今後は，個人ごとの間の調整や多重タスク下での間の取り方について検討するとともに，ユーザーとの相互行為を円滑化・活性化させる，間以外の要素にも着目し，抑揚の付け方や話速といった韻律情報を総合的に制御できる仕組みについても検討していく [6]．

参考文献

[1] 畠弘巳: “話しことばの特徴 - 冗長性をめぐって -”, 国文学解釈と鑑賞, Vol.52, No.7, pp.22-34, 1987.
[2] 川嶋宏彰, スコギンズリーバイ, 松山隆司: “漫才の動的 構造の分析—間の合った発話タイミング制御を目指して”, ヒューマンインタフェース学会, Vol.9, No.3, pp.379-390, 2007.
[3] 翠輝久, 水上悦雄, 志賀芳則, 川本真一, 河井恒, 中村哲: “ユーザの相づち・うなずきを喚起する音声対話システム”, 電子情報通信学会論文誌 A, Vol.95, No.1, pp.16-26, 2012.
[4] 高津弘明, 福岡維新, 藤江真也, 林良彦, 小林哲則: “意図性の異なる多様な情報行動を可能とする音声対話システム”, 人工知能学会論文誌, Vol.33, No.1, 2018.
[5] 鴨井一人, 山本知仁, 三宅美博: “文章発話におけるポーズ長とその前後の発話長の作用関係”, ヒューマンインタフェースシンポジウム 2011 論文集, pp.357-364, 2011.
[6] I.Fukuoka, K.Iwata, and T.Kobayashi: “Prosody Control of Utterance Sequence for Information Delivering”, INTERSPEECH, pp.774-778, 2017.
[7] 中村敏枝: “コミュニケーションにおける「間」の感性情報心理学”, 音声研究, Vol.13, No.1, pp.40-52, 2009.