4P-7

神経力学モデルの引き込みによる相槌タイミングの予測

佐野 正太郎 † 日下 航 ‡ 尾形 哲也 ‡ 高橋 徹 ‡ 奥乃 博 ‡

† 京都大学 工学部情報学科 ‡ 京都大学大学院 情報学研究科 知能情報学専攻

1. はじめに

近年, ロボット対話システムや音声対話エージェントなど, 人間と機械がコミュニケーションを行うことのできるユーザインタフェースの実現をめざして各所で研究が行われている. しかし, 現在のシステムでは, ユーザとシステムが一律な決まりのもとで交互に発話することだけが仮定されている. 実際の人間どうしの対話では, 自然なタイミングでの相槌や話者交代が行われており, 聞き手側の話の理解状況が話し手にフィードバックされている. 対話システムにおいても, 相槌のようなフィードバックが得られれば, ユーザにとってより使いやすいシステムになると考えられる.

相槌を扱うシステムの研究は, これまでにも様々に行われている. 西らは話し手の一定長のポーズを基に相槌のタイミング決定を行っている [1]. また,Ward らは話し手の低ピッチ領域を合図に相槌を打つシステムを提案している [2].

多くの従来研究では, 発話が終了する段階に相槌が集中することに着目しており, ポーズや低ピッチ領域を合図として, 発話終了を検出した上で相槌のタイミングを判定している. しかし, 岡登らが示すように, 相槌は発話終了よりやや早い段階から被せるように発生し始める傾向にある. したがって, 発話終了の検出に基づく手法では, 早いタイミングで相槌を打つことができない [3]. 岡登らはこれを指摘した上で, HMM による発話終了時点の予測を行う相槌タイミングの決定システムを提案した.

また, 相槌の出現は発話終了時点に限られるわけではない. 神谷らの分析では, ポーズを含む区間に打たれた相槌は全体の約半数ほど出現しており [4], 確かに相槌は発話終了時点に集中する傾向がみられるが, 例外となる相槌も数多く存在することがわかる.

本稿では Multiple Timescale Recurrent Neural Network (MTRNN)[5] による相槌タイミング決定システムを提案する. このシステムにより, 予測に基づき, かつ発生時点を限定しないタイミング決定が可能となる.

2. タイミング予測モデル

MTRNN は現在の状態を入力に, 次ステップの状態を出力する予測器である. 階層構造を持ったネットワークであり, 入出力を司る *IO* ノード, *IO* ノードより遅れて変化する Cf ノード, 更に遅れて変化する *Cs* ノードを持つ. 内部状態を階層的に変化させることで, 複雑な時系列予測が可能となる. 本研究で実験に用いた MTRNN は, 特徴量に対応する 5 個の *IO* ノード ($IO_{parameter}$), 相槌タイミングに対応する 1 個の *IO* ノード ($IO_{backchannel}$), 15 個の Cf ノード, 8 個の *Cs* ノードから構成される. $IO_{parameter}$ を除く各ノードは同種のノードへのフィードバックループを持つほか, Cf は *IO*, *Cs* と相互に結合されている. また, $IO_{parameter}$ には外部から特徴量が入力される. 時刻 t におけるノード i の値 $y_{t,i}$ は以下のように求める.

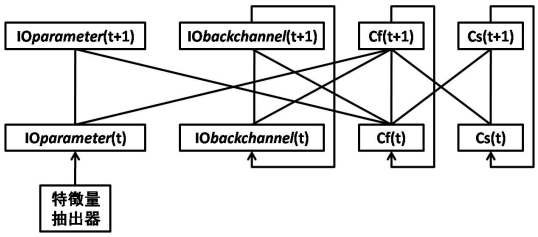

図 1: MTRNN による時系列生成モデル

$$y_{t,i} = \frac{1}{1+exp(-u_{t,i})} \tag{1}$$

$$u_{t,i} = \begin{cases} 0 & \text{t=0} \\ (1-\frac{1}{\tau_i})u_{t-1,i} + \frac{1}{\tau_i}\sum_j \omega_{ij} x_{t,j} & \text{otherwise} \end{cases} \tag{2}$$

$$x_{t,j} = \begin{cases} f_j(t) & \text{j} \in IO_{parameter} \\ y_{t-1,j} & \text{otherwise} \end{cases} \tag{3}$$

ここで, $f_i(t)$ は時刻 t における特徴量 i の値, τ_i はノード i の時定数, ω_{ij} はノード j からノード i への結合重みである. なお, 時定数 τ_i は *IO* で 2, Cf で 5, *Cs* で 70 とした. 時定数が大きいほどノードの状態は緩やかに変化する. また, 重みの学習は Back Propagation Through Time 法 (BPTT 法) によって行った.

3. 評価実験

上記で述べたモデルについて, 相槌タイミング決定の評価実験を行った. 実験ではポスターセッションのコーパスを対象とし, プレゼンターの韻律と視覚情報から聞き手 1 人の相槌タイミングを判断した. コーパス中に出現する相槌タイミングと, MTRNN の予測した相槌タイミングを比較し, その性能を再現率と適合率で評価した.

3.1 実験概要

韻律情報として音声パワーと F0 を用いた. また, 視覚情報として, 視線が聞き手に向けられるタイミング, 視線がポスターに向けられるタイミング, 頷くタイミングをプレゼンタの行動から抽出した. F0 の検出は自己相関関数を用いて行い, 分析の周期は 62.5ms とした. また視覚情報のデータはアノテーションツール iCorpusStudio[6] を用いてあらかじめラベル付けされていたものを用いた. これらのデータはラベルの段階では離散的なデータである. したがって, 次に示す式で特長量 i に対する連続的な時系列 $f_i(t)$ を定義した.

$$f_i(t) = \sum_{n=1}^{N_i} exp\{-\frac{(t-t_{i,n})^2}{2}\} \tag{4}$$

$t_{i,n}$ は特徴量 i が対象とする動作が起こった時刻, N_i はその総数である. なお, 学習用に用いる相槌タイミングの

Prediction of Back Channel Timing using Neurodynamical Model: *Shotaro Sano (Kyoto Univ.), Tetsuya Ogata (Kyoto Univ), Wataru Hinoshita (Kyoto Univ), Toru Takahashi (Kyoto Univ.), and Hiroshi G. Okuno (Kyoto Univ.)*

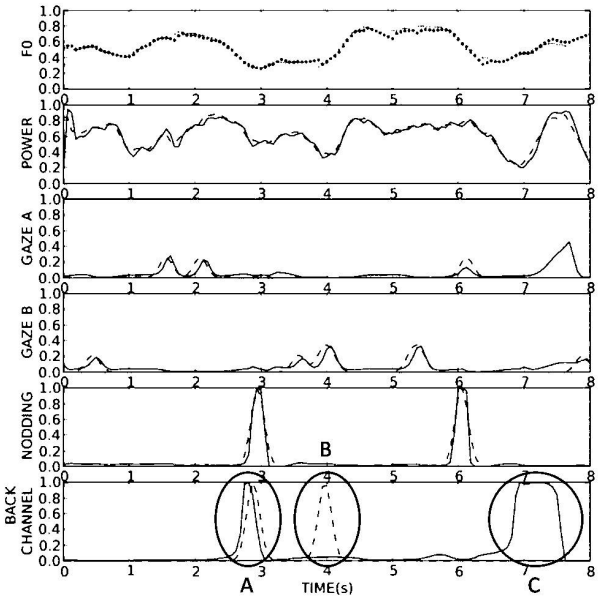

図 2: IO ノードの予測結果

A:そうですね/
(0.2)
A:まあ/若干それを間違えたりは [しますけど/
B: [ああ::/
(0.2)
A:[ま/それを別にピッチだけを見てるんではなくて/
B:[はあ/
(0.3)
A:まあ/他のその/

図 3: 対話ログ

時系列も同様に定義した.

実験用にコーパス中の聞き手の相槌を含む部分区間 (5 秒から 15 秒ほど) を計 44 個用意した. うち 22 個を学習用データとして用い, BPTT 法によってネットワークの結合重みを決定した. 残りの 22 個は評価用データとし, 2 章で述べた手法に従い相槌タイミング時系列を生成した. 学習用データ 22 個中には合計で 43 個の相槌が, 評価用データ 22 個中には合計で 35 個の相槌が, 合計で 78 個の相槌が含まれていた. また, それらの相槌のうち 47 個は話し手の発話終了点を含む区間で打たれた相槌, 残る 31 個はそれ以外の時点で打たれた相槌であった.

3.2 結果

評価用データに対する生成結果の一例を図 2 に示す. また, その対話ログを図 3 に示す. 図 2 において,1 段目から 5 段目は特徴量の時系列,6 段目は相槌タイミングの時系列である. 破線は現実に観測されたデータを表し, 実線は MTRNN が生成した時系列を表す (ただし F0 は実測値を十字点で, 生成値を丸点で表している).

6 段目において破線では 2 つのピークがたっており, これらの極大値が実際に相槌が打たれたタイミングである. 実線でも 2 つのピークが立っており, A の部分では破線のピークと重なっている. つまり, このピークに対応する相槌に関しては, 実測のタイミングと MTRNN が挿入したもののタイミングが同時であったといえる. 一方で, B の破線ピークに対しては, 実線のピークが立っておらず, 実測された相槌を MTRNN が生成できなかったとわかる. また, C の実線ピークに対しては, 相槌の実測はなく, ここではコーパスになかった相槌が挿入されていた.

ピークは極大値に到達する 0.5 秒以上前から立ち始めていることに注目しておきたい. これは暗に, MTRNN が相槌タイミングの到来を, その少し前から予測していることを示している. ここに, 1 章で挙げた相槌タイミング決定問題における予測の必要性に対して, MTRNN の有効性が確認できる.

実測データの相槌を正解の相槌と定め, 再現率と適合率によって, 本手法の性能を定量的に評価した. MTRNN の挿入した相槌のタイミングと, 実測された相槌のタイミングが前後 0.25 秒以内で一致していれば, その相槌を正解と判断した. 学習用データにおいては再現率 34.9%, 適合率 31.3%, F 値 32.9%の評価が得られた. 一方で, 評価用データにおいては再現率 22.9%, 適合率 17.8%, F 値は 19.6%であった.

相槌のタイミングに関しては絶対的な正解の基準は無いため, コーパスに無いタイミングで予測された相槌であっても不自然なものであるとは限らない. 例えば, 図 2 において, C の実線ピークに対応する相槌は, コーパス上に存在しなかったものである. しかし, この領域は対話ログにおいて, 3 回目のポーズが現れる箇所と重複しており, また, 特徴量の F0 が低くなる領域とも重なっている. したがって, この位置での相槌は適切であったとも考えられる. このようなデータも総合して評価するため, 今後は, 主観的な評価を行う必要がある.

最後に, 正解した相槌のうち話し手の発話終了点で打たれた相槌は 14 個で, 発話終了点における全ての相槌 47 個のうち 29.8%の相槌を再現していた. 一方で, その他の点において正解した相槌は 9 個であり, もとの 31 個の相槌のうち 29.0%の相槌を再現していた. いずれの点においても同様の性能が示されており, 本手法が発話終了点以外の相槌にも有効であることが示された.

4. おわりに

本稿では, 視聴覚情報を基にした相槌の挿入を,MTRNN を用いて行った. この手法で, 相槌の発生を, そのしばらく前から予測できることが示された. また, 相槌の発生箇所を, 発話終了点などに限定せず挿入できることが確認された.

今後は, 提案手法をロボット対話システムに実装し, 実環境において提案手法が有効であるかを検討すると同時に, 性能の主観的な評価を行う予定である.

謝辞 本研究の一部は JST さきがけ, 科研費基盤 (B), 科研費学術創成の支援を受けた.

参考文献

[1] 西宏之, 小島順次 : キーワードネットワークを用いた電話取り次ぎ対話処理, 信学技報, SP88-30(1988)
[2] Ward, N.: In Japanese a low pitch means ”backchannel feedback please”, 情報処理学会音声研究会報告, SLP-11-2(1996)
[3] 岡登洋平, 加藤佳司, 山本幹雄, 板坂秀一 : 韻律情報を用いた相槌タイミングの挿入, 情報処理学会論文誌, Vol.40 No.2(1999)
[4] 神谷優喜, 大野誠寛, 松原茂樹, 柏岡秀紀 : 同調対話システムにおける相槌挿入タイミング, 言語処理学会第 16 回年次大会発表論文集 (2010)
[5] Y. Yamashita and J. Tani: ”Emergence of Functional Hierarchy in a Multiple Timescale Neural Network Model: a Humanoid Robot Experiment,” PLoS Comput. Biol., vol.4, no.11(2008).
[6] 來嶋 宏幸, 坊農 真弓, 角 康之, 西田 豊明 : マルチモーダルインタラクション分析のためのコーパス環境構築, 情報処理学会研究報告（ヒューマンコンピュータインタラクション）, Vol.2007, No.99(2007)