エントレインメント現象を用いた 音声認識に適した発話様式誘導に関する調査*

☆杉山 昂太郎, 戸田 智基, Graham Neubig, Sakriani Sakti, 中村 哲 (奈良先端大)

1 はじめに

音声対話システムにおいて，前段の処理である音声認識の結果は後段の処理に影響を与えるため，高精度な音声認識処理の実現は本質的に重要である．従来，音響モデル構築に使用される大規模コーパスは形式的な独話である場合が多く，我々人間が日常的に発話している自由で柔軟な発話様式とは異なる．そのため，システム設計者の意図した発話様式や発話文が得られない場合，音声認識失敗のリスクとなり得る．

本研究では，利用者側に働きかけるという観点から，エントレインメント現象を用いて利用者の発話様式を誘導する手法の確立を目指す．それに向けて，本稿では，日本語音声における音響的なエントレインメント現象に着目し，利用者から音声認識に適した発話様式を引き出す発話様式誘導に関する調査を行う．

2 音声対話とエントレインメント現象

エントレインメント現象とは，特徴的な生命現象の一つで，生体リズムが相互に同期化する現象を指す [1]．人間とコンピュータあるいはロボット間のインタラクションにおいても同様の同期化現象が確認されており [2]，システムから発せられた状況的振る舞いが，人間にとって合理的であるならば，それは人間に模倣されることが示されている [3]．

音声コミュニケーションにおけるエントレインメントとして，音声の言語的な特徴量や音響的な特徴量に着目した研究がいくつか行われている．言語的特徴のエントレインメントに関しては，語彙，言語スタイル，文法構造に変化が生じることが報告されている．また，音響的特徴に関しては，基本周波数，強度，声質などの特徴量で引き込みが起こることが報告されている [4]．しかしながら，定量的な議論は未だ限定的なものであり，特に音響エントレインメントに関しては，日本語における調査はあまりなされていない．

そこで我々は，日本語音声における音響エントレインメントについて，実験的評価を交えて調査した [5]．その結果，文献 [4] と同様に，日本語音声対話においても，基本周波数，強度，ジッター，シマー，発話速度の 5 つの特徴量に対してエントレインメント現象が生じることを確認した．

3 音声認識を考慮した発話様式誘導

音声対話システムにおいて，利用者の発話様式をシステムが受理可能な発話様式へと誘導することで，高い音声認識成功率を引き出すアプローチを提案する．自然な発話様式誘導を実現するためには，エントレインメント現象を利用する．発話者はエントレインメントにより対話相手の言語的・音響的特徴に近い発話様式に引き込まれるため，システムの応答音声として高い音声認識率が得られる発話様式を用いることで，利用者の発話も音声認識に適した発話様式へと引き込むことができると期待される（Fig. 1）．

本稿では，提案法の実現可能性を調査するために，人同士の対話を対象として，音声認識率を考慮した発話様式誘導実験を行う．2 名の対話において，内 1 名をシステム側とみなし，高い音声認識率が得られる発話様式により発声する．その際に，利用者側とみなすもう 1 名に対して，エントレインメントによる発話様式誘導が生じ，音声認識率が向上するか否かについて調査する．その際に，システム側の発話者は，高い音声認識率が得られる発話様式を行う必要がある．そのような発話様式を意図的に引き出すために，発話者は発声時に音声認識システムを使用して，正しい認識結果が得られるように注意しながら発話を行う．この手順により，実際に所望の発話様式が得られるか否かについても調査する．

4 実験的評価

4.1 実験条件

実験題材として，地図情報提供者と情報追従者の 2 名による地図課題対話 [6] を用いた．各話者は小窓で隔てられた別室にて対面して対話を行った（Fig. 2）．話者はそれぞれヘッドセットマイクとイヤホンを装着し，互いの発話を頼りに地図課題タスクを行った．被

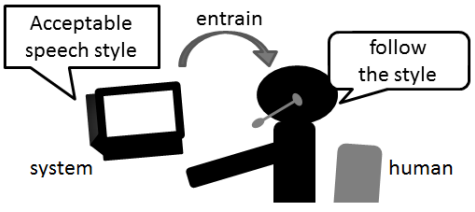

Fig. 1 Overview of the proposed method

*An examination of speech style manipulation using entrainment phenomenon for speech recognition system, by SUGIYAMA, Koutaro, TODA, Tomoki, SAKTI, Sakriani, NEUBIG, Graham, NAKAMURA, Satoshi (NAIST)

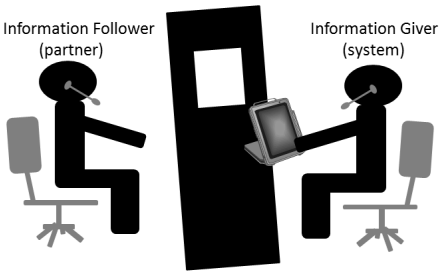

Fig. 2　Recording conditions

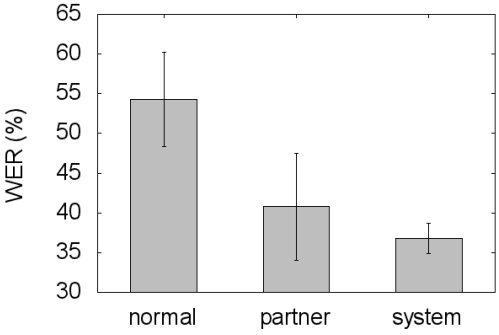

Fig. 3　Word error rate

験者は，情報提供者 4 名，情報追従者 4 名の計 8 名の 20 代男性とした．各被験者は，条件を変えて 2 回の対話収録を行った．1 回目の対話収録では，情報提供者側は iPad 用の音声認識アプリケーションである Dragon Dictation を使用して発話を行った．その際に，音声認識が成功しやすい発話を心がけるよう指示した．一方，情報追従者は音声認識システムを使用せずに発話を行った．発話様式についても，一切の指示は与えなかった．また，2 回目の対話収録では，情報提供者，情報追従者ともに，音声認識システムを使用せず，発話様式に関する指示も与えずに対話を行った．平均対話時間は 1 回目が 8 分 32 秒，2 回目が 6 分 59 秒であった．

1 回目の対話収録における音声認識を意識した発話（system 群）とその対話相手の発話（partner 群），2 回目の対話収録における制約のない発話（normal 群）の 3 種類の音声に対して，単語誤り率による音声認識精度を評価した．収録時と同様に，Dragon Dictation を用いて認識処理を行った．

4.2　実験結果

3 群それぞれの平均単語誤り率を Fig.3 に示す．system 群の発話は，normal 群の発話と比較し，単語誤り率が大幅に低減しており，3 群の中で最も低い単語誤り率が得られる．このことから，認識結果を確認しながら認識が成功しやすい発話を心掛ける手順により，高い認識精度が得られる発話様式を意図的に引き出せることが分かる．一方で、その対話相手である partner 群の発話に対する単語誤り率についても，system 群の発話には 4 %程度及ばないものの，normal 群の発話と比較すると大幅に低減される．以上の結果から、片方の話者が高い認識精度が得られる発話様式により発話を行うことで，その対話相手の発話様式をエントレインメントにより誘導することが可能であり，音声認識率の高い発話を引き出せることが分かる．

5　おわりに

本稿では，音声対話システムにおいて音声認識精度を改善する手法として，エントレインメント現象を用いた発話様式誘導の提案と，その実現可能性に関する調査を行った．人同士の対話による実験結果から，片方の発話者が高い音声認識精度が得られる発話様式を用いることで，対話相手の発話様式を誘導することができ，より高い音声認識精度が得られる発話を引き出せることを明らかにした．今後は，高い音声認識精度が得られる発話様式を音声合成システムでモデル化して対話システムに導入し，システムと人との対話において本手法の有効性を検証する予定である．

謝辞　本研究は，（独）情報通信研究機構の委託研究「知識・言語グリッドに基づくアジア医療交流支援システムの研究開発」の一環として実施した．

参考文献

[1] 渡辺 富夫．成人間コミュニケーションにおけるエントレインメント（音声ー体動同期現象）の分析. 情報処理学会論文誌, Vol. 26(2), pp. 272–277, 1985.

[2] 飯尾 尊優 他. 語彙の引き込み:ロボットは人間の語彙を引き込めるか？情報処理学会論文誌, Vol. 51, pp. 277–289, 2010.

[3] 駒込 大輔 他. Robotmeme: 模倣による人ーロボットの周辺的相互適応．ヒューマンインタフェース学会論文誌, Vol. 10(1), pp. 47–57, 2008.

[4] Rivka Levitan, Julia Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of Interspeech*, 2011.

[5] 杉山 昂太郎 他．音声認識のためのエントレインメント現象を用いた発話様式誘導．電子情報通信学会技術研究報告, 2014.

[6] 市川 熹 他．日本語地図課題対話コーパス．音声研究, Vol. 4(2), pp. 4–15, 2000.