Vol.2016-SLP-112 No.15
2016/7/30

ユーザフレンドリィな音声対話システム実現のための ユーザ話速および発話内容に基づく システム話速制御手法の検討

三原 寛哉[1]　李 晃伸[1]

概要：よりユーザフレンドリィな音声対話システム実現のため，本研究では，ユーザ話速，およびシステムの発話内容に応じてシステム話速を制御する手法を提案する．まず，ユーザ話速と好まれるシステム話速の関連性について調査した．結果，話速が速いユーザは速いシステム話速を，遅いユーザは遅いシステム話速を好むことが示された．この実験の知見を基に，リアルタイムにユーザ話速に応じてシステム話速を制御するシステムを構築した．話速の制御手法についてはユーザごとの平均話速，発話ごとの話速を用いる方法の他に発話内容に依存して変化させる方法，ならびにデフォルト話速から徐々に数ターンかけて話速を変化させる方法についても実験を行った．評価実験を行った結果，固定話速のシステムはユーザによらず安定した評価を得られること，発話単位の同調は好まれないこと，ユーザの平均話速へは徐々に同調することで評価が高くなること，発話内容に依存した話速変化が好まれることなどがわかった．

1. はじめに

近年，音声認識や音声対話の技術が向上し，Apple の siri, NNT ドコモのしゃべってコンシェルやハンズフリーのカーナビゲーションシステムなどの音声対話システムを利用したアプリケーションが急速に普及しつつある．このようなアプリケーションの増加に伴って，よりユーザフレンドリィな音声対話システムが求められている．

一般的に，人間同士の対面コミュニケーションでは対話の引き込みと呼ばれる現象が発生することが知られている [1], [2]．これは対面でコミュニケーションを行っている二者が，発話内容以外の情報（例えば声の大きさ，高さ，テンション，発話速度，交代潜時〈一方が話し終わってからもう一方が話し始めるまでの時間〉などのことで，以降ノンバーバル情報と呼ぶ）をお互いに同調させる現象のことであり，これによってコミュニケーションが円滑になることが知られている．

そこで本研究では，ノンバーバル情報の中でも特に取得が容易な発話速度（以降，話速と呼ぶ）に着目し，音声対話システムにおいて，ユーザの特性や発話内容を考慮してシステム話速を制御すれば，よりユーザフレンドリィな音声対話システムが構築できるとし，まずユーザ話速から好まれるシステム話速の関係性を調査する．そして，実験により得られた知見を用いて，リアルタイムにユーザの特性および発話内容を考慮してシステム話速を制御するシステムを構築し，評価実験を行う．

[1] 名古屋工業大学大学院　工学研究科
Graduate School of Engineering, Nagoya Institute of Technology

2. 音声対話における話速と引き込み

本節では，対話における話速に関連する研究についていくつか取り上げ，音声対話システムとの関連について述べる．

2.1 人間同士の対話における引き込み現象

人間同士の対話における引き込みと呼ばれる現象の有無について調査した研究として，小松ら [3] の，人間同士のコミュニケーションにおいて相手の話速に同調するような話速の引き込み現象が観察されるかを確認する研究が存在する．小松らは人が容易に取得できる話速に着目し，人間同士の対話状況において話速に関する引き込み現象が観察されるのかどうか確認する実験を行った．

各被験者の単独話速を基準として，被験者が相手の話速に自分の話速を同調させているかを，対話話速測定実験で獲得した計 90 発話について調査すると，63 %が，相手の話速に同調するように自らの話速を変化させていた．よって，話速の引き込み現象はかなりの頻度で観察される現象だと分析している．この結果より，対話コミュニケーショ

ンを行っている人間には，相手の話速に自分の話速を合わせようとする傾向があると結論づけている．

音声対話システムにおいても，人間とシステムの間でも引き込み現象を発生させることでより親しみのある対話が実現できると考えられる．

2.2 人間同士の対話現象を模倣する音声対話システムの評価

人間同士の対話現象を模倣した音声対話システムに対し被験者が抱く印象について調査した西村ら [4] の研究が存在する．この研究では人間同士の対話現象を模倣する音声対話システムを構築するために，実験システムに，あいづち，復唱，共同補間，オーバーラップ，バージインの 5 種類を実装し評価実験を行った．

実験の結果，被験者のほとんどがあいづちに対して親しみを感じた．また，復唱に関しては被験者の好みに応じて適応する必要があることがわかった．共同補間については，対話システムとユーザの間には現れにくく，オーバーラップ応答については頻度が高過ぎるとユーザに嫌われるが，音声認識率が高いときは，オーバーラップ頻度が高い対話が好まれる傾向になった．バージインについては，システムの誤認識に対してすぐに対処出来る点でユーザに好評であった． 以上を総合すると，人間同士の対話現象を模倣して応答することが可能な音声対話システムは，ユーザに親しみを感じさせる点において有効であるが，システムの音声認識精度に大きく依存するといえる．

2.3 話速同調と共感度の関係性

話速同調と共感の度合いとの関連性について調査した Bo Xiao ら [5] の研究が存在する．この研究では，セラピストとその患者のカウンセリングの様子を録音したコーパスからセラピストと患者の話速を抽出し，発話ごとのセラピストと患者との間に存在する話速の差の平均や話速の変化量の差の平均を算出し，それらと共感度に相関があるかどうか調べた．

実験結果からはやや負の相関があることが示されている．つまり，セラピストと患者の話速には差がないほうがよいことや，セラピストの話速の変化量と患者の話速の変化量にも差がないほうがよいことも確認できた．すなわち，患者の話速がある程度一定であると仮定すると，セラピストは初めから話速を患者に同調するのではなく，徐々に同調したほうがより共感が得られることがわかった．

そこで本研究において，システムが徐々にユーザの話速に同調するように話速を制御すればよりユーザフレンドリィな音声対話システムが構築できると考えられる．

3. 目標とする音声対話システム

前節で取り上げた従来研究から示唆されるように，音声対話システムに適したノンバーバル情報の扱いについては，

(1) 模倣や同調が話者を引き込む効果がある
(2) 音声認識精度に依存しない手法がロバストである
(3) 同調の速度や変化量にも着目すべきである

といった点が挙げられる．このことから，本研究では話速に着目し，これを適切に制御することで，よりユーザフレンドリィな音声対話システムを構築する．

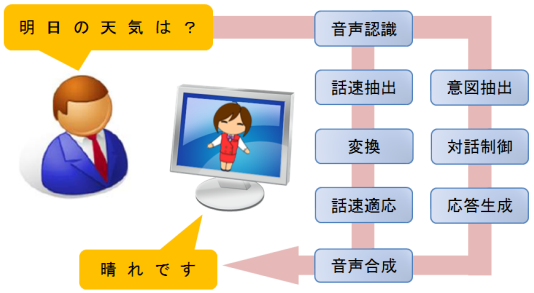

図 1 目標とする音声対話システムの構成図

従来の音声対話システムの音声認識モジュールでは，認識した音声を文字列として認識するのみである．目標とする音声対話システム（図 1）では，音声を認識する際に，ユーザの特性（ここでは話速）を取得する．そして取得した話速を用いて適切なシステム話速を算出し，音声合成の際に利用する．

4. ユーザ話速とシステム話速の関連性の調査

ユーザの特性に同調するシステム構築の前に，システム話速とユーザ話速の関係がユーザフレンドリィさにどのような影響を及ぼすのか調査する．本節では，この実験の詳細について述べる．

4.1 実験条件

音声系の研究室に従事する大学生および大学院生計 12 名に対し実験を行う．ディスプレイに等身大のキャラクタを表示し，被験者と観察者 2 名以外のいない静かな屋内を実験環境として設定する．実験には，観察者の思い通りにシステムに返答させることができる Wizard of Oz (WOZ) 法 [6] を採用し，MMDAgent[7] を用いて対話システムを構築した．実験条件としては，システムが出力する音声の話速を 5 種類用意した．それぞれ，とても遅い話速（0.7 倍），やや遅い話速（0.9 倍），標準話速（1.0 倍），やや速い話速（1.1 倍），とても速い話速（1.3 倍）とする．この倍率は MMDAgent に用いられている音声合成エンジン OpenJTalk[8] の標準倍率を 1.0 としたときの倍率である．また，この実験では，モーラ数 [mora] を発話時間 [sec] で割った値を話速 [mora/sec] として定義する．

実験に用いる対話のタスクとしては，ユーザとシステムのやりとりの複雑さを変化させるため，以下の 3 種類を用意した．

U:ユーザ，S:システム
U「こんにちは」
S「こんにちは，今日も寒いですね」
S「今朝は何時に起きたのですか」
S「夜は何時に寝たのですか」
S「もう少し，寝たほうがいいですよ」
S「ありがとうございました」
（U の発話内容は自由応答のため省略する）

図 2 一問一答タスクにおける対話シナリオの例

図 3 実験システムの外観

- 聞き取りタスク
  被験者はシステムが読み上げる内容を聞くだけで，対話は行わない．
- 一問一答タスク
  システムがユーザに質問を行い，ユーザはそれに対し「はい/いいえ」で返答を行う．
- 会話タスク
  一問一答タスクのようにユーザの質問を行うだけではなく，システムとの会話を楽しんでもらうことを目指す．

なお，一問一答タスクと会話タスクに関しては，被験者に初めに「こんにちは」とシステムに話しかけてもらい，5 ターン前後の対話を行う．また，上記の 3 種類のタスクについてそれぞれ 3 通りの対話を用意し，タスクごとに 5 回の対話を計 15 回行う．そのとき，前述の 5 種類のシステム話速倍率をランダムで変更し，全ての話速が一度ずつ使用されるよう設定する．また，一問一答タスクにおける対話シナリオの例を図 2 に，実験システムの外観を図 3 に示す．

評価については，被験者は実験後，1 回の実験ごとにそのとき用いた音声対話システムを「話しやすかったか」，「聞き取りやすかったかたか」，「使いやすかったか」，「親切だったか」，「自然な対話だったか」，「対話に引き込まれたか」，「イライラしたか」，「総合評価」の 8 つの項目について 5 段階評価を行う（5 が最も高い評価）．

4.2 ユーザ話速とシステム話速の関連性

実験から得られた音声データより，被験者ごとに全発話を平均した値を算出した．それらを基に被験者 12 名を話速が遅い人のグループ (〜6.14[mora/sec]) (以下グループ A)，話速が普通の人のグループ (6.14〜6.46[mora/sec])(以下グループ B)，話速が速い人のグループ (6.46〜[mora/sec])(以下グループ C) の 3 つに分類し分析を行った．

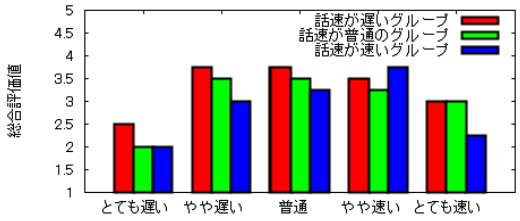

図 4 会話タスクでのユーザ話速ごとのシステム話速の総合評価値の平均

会話タスクにおけるシステム話速ごとの総合評価の値の平均のグラフを図 4 に示す．

図 4 より会話タスクでは，グループ A の総合評価値が最大となるシステム話速は「やや遅い」〜「普通」のところに存在し，グループ B の総合評価値が最大となるシステム話速は「やや遅い」〜「普通」のところに存在し，グループ C の総合評価値が最大となるシステム話速は「やや速い」ところに存在している．別の観点からは，グループ A は「やや遅い」以外の項目でも比較的大きい値を示すが，グループ C は「やや速い」以外の項目では値は小さくなっている．

これらの実験結果から，話速の遅い人は遅いシステム話速を，話速の速い人は速いシステム話速を好む傾向があることがわかる．また，話速の遅い人は自分の話速とシステム話速のずれに寛容であるが，話速の速い人は自分の話速とシステム話速のずれに不寛容であるといえる．

また，タスクごとの差に関しては，一問一答タスクは会話タスクと同じく，話速の遅い被験者は遅いシステム話速を，速い被験者は速いシステム話速を好む傾向があった．また，聞き取りタスクにおいては，ユーザ話速とシステム話速の関係は薄く，被験者の好みなど別の要因が評価に影響したと考えられる．

4.3 発話内容によるユーザ話速の変化

次に，実験で収集した音声データから得られた被験者の話速の推移の一例を図 5 に示す．被験者の話速が徐々に速くなっていき，6 発話目では急上昇する傾向があることがわかる．この傾向から，被験者の話速が徐々に速くなる原因はシステムへの慣れにあるとした。また，6 発話目で話速が急上昇したのは，「ありがとうございました」などの定型的な言葉を発話したためであるといえる．また，システム話速が「とても遅い話速」の場合，発話ごとに被験者の話速が大きく変化している．このことから，発話ごとに適

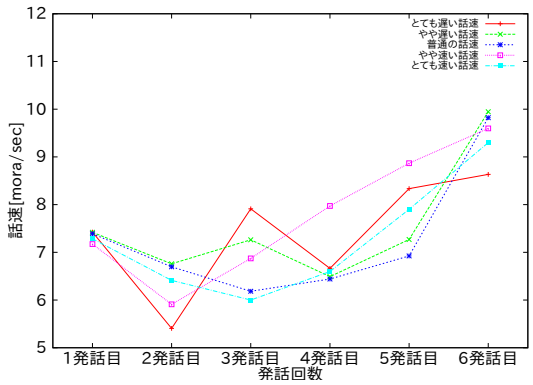

図 5 会話タスクでの被験者の話速推移の一例

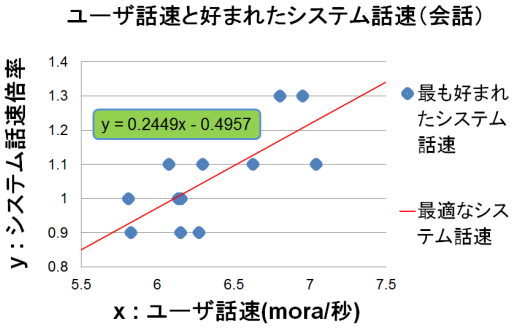

図 6 ユーザの話速から最適なシステム話速倍率への変換関数

切な話速が異なることが示唆された．

5. ユーザ話速に同調する音声対話システム

前節の結果を基に，ユーザ話速に対して適切なシステム話速で応答するシステム，およびシステムの発話内容によってシステム話速を制御するシステムを構築し，比較評価実験を行う．

5.1 ユーザ話速から好まれるシステム話速倍率への変換式

本研究では，図 6 に示すように被験者の話速とその被験者が好んだシステム話速倍率の関係が一次式であると仮定し，最小二乗法によりユーザの話速から好まれるシステム話速の倍率へと変換する関数を作成した．x がユーザの話速，y がシステム話速倍率を表す．

$$y = 0.2449x - 0.4957 \quad (1)$$

5.2 システム構成

提案するシステムの構成を図 7 に示す．まず，発話者の話速を取得する．本研究で構築する音声対話システムは，音声認識部に Julius[9] を用いている．Julius では，認識結果においてその単語や音素，あるいは HMM の状態がそれぞれ入力音声のどの区間にマッチしたのかを知ることができ，より正確なアラインメントを求めるために，認識中の近似を含む情報は用いずに，認識が終わった後に得られた認識結果の単語列に対して，改めて forced alignment を

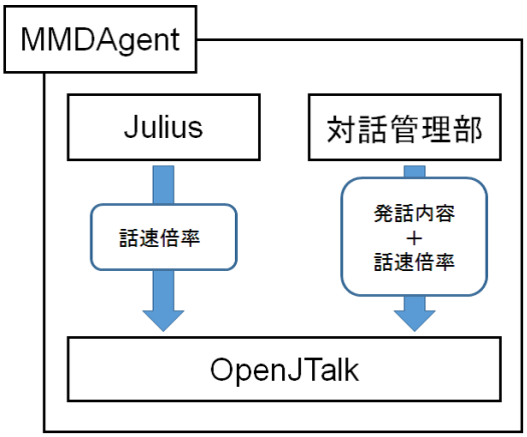

図 7 ユーザ話速と発話内容に基づき話速を制御するシステムの構成

行うことができる．この出力結果よりまず発話フレーム長を取得する．次に音素の出力結果から母音，長音，促音，撥音を認識し mora 数をカウントし，第 4 節で述べた話速の定義に従ってユーザ話速を算出する．次に，式 1 を用いて，ユーザ話速からシステム話速倍率へ変換する．さらに，求められたシステム話速倍率を OpenJTalk へと送る．OpenJTalk 側ではメッセージを受け取ると，合成音声の倍率を受け取った倍率に上書きする．

発話内容によってシステム話速を制御するシステムにおいては，MMDAgent の対話管理部において，発話内容とそれに対応する話速倍率を人手で記述し，OpenJTalk にメッセージを送ることで話速を制御する．

6. 予備実験

対話実験を行ったところ，ユーザ話速に同調するシステムに関して，システム話速が速くなりすぎた，もしくは直前のシステム話速と比べて変化量が大きくなりすぎたことが原因によって低評価となった．そこで，聴取実験を行い，違和感なく対話できるシステム話速の範囲に関して以下の 3 つを調査した．

- システム話速の上限下限
- 1 ターンごとのシステム話速の変化量
- 1 ターン中のシステム話速の変化量

6.1 実験設定

システム話速の上限下限の調査方法に関しては，MMDAgent の対話管理部において，話速倍率を 0.5 から 1.5 まで 0.1 刻みで作成した対話シナリオを用い，被験者は違和感のない範囲を回答する．1 ターンごとのシステム話速の変化量の調査方法に関してはシステムがまず「こんにちは」と発話し，次に被験者が「こんにちは」と発話すると，システムが「今日もいい天気ですね」と返答するので，システムの 1 発話目と 2 発話目の話速の変化量の許容範囲を被験者が回答する．1 ターン中のシステム話速の変化量の調査方法に関しては，「こんにちは」と「今日もいい天気です

ね」という発話内容に話速の変化を生じさせ，その変化量を徐々に増加させる．そして許容できる 1 ターン中の話速の変化量の上限を被験者が回答する．

6.2 実験結果

被験者 7 名に対し聴取実験を行った．その結果，話速倍率の上限に関しては，1.3 倍と回答した被験者が 4 人，1.2 倍と回答した被験者が 2 人，1.1 倍と回答した被験者が 1 人となり，下限は，0.8 倍と回答した被験者が 5 人，0.9 倍と回答した被験者が 2 人となった．1 ターンごとのシステム話速の変化量の上限に関しては，0.3 と回答した被験者が 2 人，0.4 と回答した被験者が 4 人，0.5 と回答した被験者が 1 人となった．1 ターン中のシステム話速の変化量の上限に関しては，0.3 と回答した被験者が 4 人，0.4 と回答した被験者が 3 人となった．

以上の実験結果より，システム話速の許容範囲に関しては中央値を上限下限に設定する．また，1 ターンごとのシステム話速の変化量の上限，1 ターン中のシステム話速の変化量の上限に関しては，中央値の 0.1 小さい値を採用する．そしてこれらの制約を全てのシステムに適用する．

7. 評価実験

7.1 比較するシステム

- システム A（固定話速）
  話速をデフォルトの標準話速 (1.0 倍) から変更しない標準システム．
- システム B（ユーザ＋直前の発話に同調）
  ユーザが発話するたびに式 1 を適用し，直前の発話に同調するシステム．
- システム C（ユーザの平均話速に同調）
  前章の結果を基に，ユーザ発話における対話の始まりからの累積 mora 数と累積フレーム長から，ユーザの平均話速を算出し, それをシステムの最適話速倍率に変換し同調するシステム．発話が蓄積されるに従って一定のシステム話速へ収束する．
- システム D（ユーザの平均話速＋標準話速から同調）
  システムの 1 発話目は標準話速，2 発話目が標準話速とユーザの平均話速の平均値, 3 発話目以降はユーザの平均話速に完全に同調するシステム．
- システム E（発話内容）
  システム話速の倍率を，「こんにちは」等の定型的な発話内容は 1.2 倍，質問文を 0.9 倍，その他を 1.0 倍と発話内容に対するシステム話速倍率を手動で設定したシステム．

7.2 実験条件

音声系の研究室に従事する大学生および大学院生計 20 名に対し実験を行う．ディスプレイに等身大のキャラクタを表示し，被験者と観察者 2 名以外のいない静かな屋内を実験環境として設定する．実験には，観察者の思い通りにシステムに返答させることができる Wizard of Oz (WOZ) 法を採用し，MMDAgent を用いて対話システムを構築した．用いた実験システムは前述の 5 システムである．タスク設定としては，「はい/いいえ」で答えられるような一問一答タスクでは，被験者本来の話速を観測しづらいため，第 4 章の実験とは異なり被験者が自由に応答できる会話タスクのみに絞った．被験者には初めに「こんにちは」とシステムに話しかけてもらい，6 ターン前後の対話を行う．また，対話内容は全て同じものを用いると事前に対話内容が被験者にわかってしまい，システム話速に関係なく慣れによって被験者の話速が速くなってしまう可能性があるため，システムごとに異なる対話シナリオを用意した．対話シナリオの例を図 8 に示す．

そして，被験者は実験後，1 回の実験ごとにそのとき用いた音声対話システムを第 4 章の実験と同じ評価項目について 5 段階評価を行う．

U:ユーザ，S:システム
U「こんにちは」
S「こんにちは，あなたの出身地はどこですか」
U「一人暮らしですか，実家暮らしですか」
（U が一人暮らしの場合）
S「自炊はちゃんとしていますか」
S「そうなんですか．一人で寂しかったりしませんか」
S「なるほど，ありがとうございました
（U が実家暮らしの場合）
S「通学時間はどれくらいかかっていますか」
S「そうなんですか．一人暮らししたいと思ったことはありませんか」
S「なるほど，ありがとうございました」
（U の発話内容は自由応答のため省略する）

図 8 対話シナリオの例

表 1 実験の評価アンケート結果の平均値

項目	A	B	C	D	E
話しやすかったか	3.75	3.70	3.70	3.95	3.95
聞き取りやすかったか	4.05	3.80	3.50	3.74	3.80
使いやすかったか	3.65	3.50	3.50	3.63	3.75
親切だったか	3.45	3.35	3.55	3.84	3.75
自然な対話だったか	3.60	4.10	3.30	4.05	4.10
対話に引き込まれたか	3.25	3.20	3.30	3.37	3.50
イライラしたか	4.05	3.65	3.85	4.32	4.05
総合評価	3.60	3.50	3.60	3.79	3.90

7.3 実験結果

表 1 にシステムごとのアンケート結果の平均値を示す（イライラしたかの項目だけ値を反転させている）．実験より得られた結果から，それぞれのシステムについて考察を行う．

システム A（固定話速）は「聞き取りやすかったか」という項目で最も評価が高く，話速を固定することによる聞き取りやすさが高評価につながった．それ以外の項目は平

均的な値であった．システム B（ユーザ＋直前の発話に同調）は，「イライラしたか」という項目で評価が低く，それ以外の項目は平均的であった．このシステムでは直前のシステム話速との変化量が大きくなる傾向があり，それがイライラにつながったと考えられる．システム C（ユーザの平均話速に同調）は，「自然な対話だったか」という項目で評価が低く，その他の項目では平均かやや低めの値となった．第 4 節の実験結果より，定型的な発話内容において被験者の話速が速くなるという知見があった．今回の実験では被験者の「こんにちは」という発話で対話を開始しており，被験者本来の話速より速くなった可能性があり，その話速に同調したせいで評価が低くなったと考えられる．最初から同調するよりは，システム D（ユーザ平均話速＋標準話速から同調）のように，標準話速から徐々に同調させていく方がよいことがわかった．システム E（発話内容）は，全体的に評価が高く，特に「自然な対話だったか」「総合評価」で高い値となった．今回手動で発話内容に対して話速を設定したが，より精密なモデルを構築すればさらにユーザフレンドリィな音声対話システムを構築できると考えられる．

8. むすび

本研究では，よりユーザフレンドリィな音声対話システムを構築するために，話速に着目し，まず，ユーザ話速とシステム話速の関係性を調査した．実験結果より，話速の遅い被験者は遅いシステム話速を，速い被験者は速いシステム話速を好むことが判明した．また，「ありがとうございました」等の定型的な発話内容については，被験者の話速が速くなり，発話内容によって最適なシステム話速が変化することが示唆された．さらに本研究では，これらの知見を基にシステムを構築し比較評価実験を行った．実験結果より，徐々にユーザ話速に同調することや，発話内容によってシステム話速を制御することによってよりユーザフレンドリィな音声対話システムを構築できることがわかった．一方，同調のさせ方によっては，固定話速の方が評価が高くなることから，対話中に内容と関係なく話速を変化すべきではないことが示唆された．

今後の課題としては，徐々にユーザ話速に同調する手法と発話内容に応じてシステム話速を制御する手法を組み合わせたシステムの構築等が挙げられる．

参考文献

[1] Condon, S.W and Sander, L.W, “Neonate movement in synchronized with adult speech, ”Interaction participation and language acquisition, Science, Vol.183, p.99–101 1974.
[2] 渡辺富夫，“身体性コミュニケーションにおける引き込みと身体性”，ベビーサイエンス，Vol.2, pp.4–12，2003
[3] 小松孝徳，森川幸治，“ 人間と人工物との対話コミュニケーションにおける発話速度の引き込み現象 ”，情報処理学会研究報告-知能と複雑系，Vol.2004, No.105, pp.71—78, 2004.
[4] 西村良太，“ 人間同士の対話現象を組み入れた音声対話システムの研究 ”，豊橋技術大学博士論文，2010-9.
[5] Bo Xiao, Zac E. Imel, David C. Atkins, Panayiotis G. Georgiou, Shrikanth S. Narayanan, “ Analyzing Speech Rate Entrainment and Its Relation to Therapist Empathy in Drug Addiction Counseling ”, INTERSPEECH, 2015.
[6] N. M. Fraser and G. N. Gilbert,“ Simulating Speech Systems, ”Computer Speech and Language, Vol. 5, No. 1, pp.8199, 1991.
[7] 大浦圭一郎，山本大介，内匠逸，李晃伸，徳田恵一，“ キャンパスの公共空間におけるユーザ参加型双方向音声案内デジタルサイネージシステム ”，人工知能学会誌, Vol.28, No.1, pp.60-67, 2013.
[8] 大浦圭一郎，酒向慎司，徳田恵一，“ 日本語テキスト音声合成システム OpenJTalk ”，日本音響学会講演論文集，vol.1, 2-7-6, pp.343-344, 2010.
[9] 河原達也，李晃伸，“ 連続音声認識ソフトウェア Julius，人工知能学会誌 ”，Vol.20, No.1, pp.41—49, 2005.