音声対話システムにおける発話内容を考慮した応答タイミングの評価

吉田 奎一郎[*1]　大村 卓矢[*1]　山本 知仁[*1]

Evaluation of Response Timing in Dialogue System Considering Context of Utterance

Keiichiro Yoshida[*1], Takuya Ohmura[*1] and Tomohito Yamamoto[*1]

Abstract – The speech dialogue systems such as Apple's "Siri" have been widespread nowadays. However they sometimes give unnatural impression to users. One approach to solve this problem is to understand the relation between verbal and non-verbal information in human communication. Therefor we have focused on the utterance rhythm as one of important non-verbal information and we have analyzed appropriate utterance timing of the system so far. As a result, about 600msec after utterance is proper timing in Japanese greeting. However in the actual human dialogue, overlap (the response latency becomes negative) occurs when the speakers are intimate, or the dialogue becomes liven up. It is better to consider this phenomenon to realize more natural communication with the dialogue system. Therefore, in this research, to analyze the proper utterance timing including overlap in the system, we have analyzed the impression of the timing in four different dialogue contents. As a result, the proper utterance timing is different depending on the context of the dialogue.

Keywords: Dialogues system, Response timing, Overlap, Context of utterance

1. はじめに

これまで音声対話システムは，観光地の音声ガイダンスやコミュニケーションロボットなど，さまざまな領域で利用されてきている．さらに近年では，Apple の「Siri」や Yahoo Japan の「音声アシスト」のようなモバイル端末で動作するシステムも実用化され，その有用性が広く認知された．しかし，現在利用されている音声対話システムと人との対話では，ユーザがシステムに対して違和感を持つことがある．その理由として，人同士の対話には発話リズムや韻律情報などの非言語情報が存在しており，上記のシステムでは，これらが十分に考慮されていないことが挙げられる．対面のコミュニケーションにおいて非言語情報は全体の 65%[1]または 93%[2]の情報量をもつともいわれ，話者間の円滑な情報伝達に重要な役割を果たしている．音声対話システムは，今後も様々な場面で利用されるインターフェースとなりうる．このような背景を考えると，人同士の対話における非言語情報を解析し，その結果を利用してより自然な対話を実現させていくことは，これからも重要であると考えられる．

これまでに，われわれは非言語情報の一つである発話リズムに着目して研究を行ってきた．その先行研究において，発話タイミングを制御する音声対話システムの構築を行い，正の反応潜時における適切な発話タイミングの評価を行った[3]．一方で，人同士の対話では反応潜時が負になり，相手の発話中に割り込むオーバーラップが生じることがある．オーバーラップは，話者間に親しさがあるときや，話が盛り上がっているとき，また確認を行うような対話のときに起きやすいとされている[4]．より自然な音声対話システムを構築するためには，このようなオーバーラップについても考慮していく必要があると考えられる．そこで，オーバーラップしながら応答する機能を音声対話システムに実装し，1 発話からなる対話において，オーバーラップを含む複数の発話タイミングがどのような印象を話者に与えるかを調査した[5]．その結果，発話の前半に割り込むオーバーラップはあまり良い印象を与えないが，文末付近で割り込むオーバーラップは相対的に良い印象を与えるということが示唆された．これは，オーバーラップが起こる位置によって応答の印象が大きく変わることを意味している．

このような結果に加えて，オーバーラップを含む発話タイミングは対話内容や文脈などにも大きく依存することが考えられる．しかし，これらの要素を考慮した発話タイミングの評価は，これまで十分には行われてきていない．そこで本研究では，予備的な研究[6]において行った 3 発話からなる 2 種類の対話に加え，さらに対話を 2 種類用意し，計 4 種類の対話の内容によって発話タイミングの印象がどのように変化するのかを調査する．

2. 実験手法

2.1 実験概要

本研究では，3 発話からなる 4 種類の対話においてオーバーラップを含む発話タイミングの評価実験を行い，対話内容の違いによる適切な発話タイミングについて調査する．実験で使用する対話内容には，天気を尋ねる対

*1: 金沢工業大学大学院　工学研究科　情報工学専攻
*1: Graduate Program in Information and Computer Engineering, Graduate School of Engineering, Kanazawa Institute of Technology.

話，漫才の対話，観光相談の対話，目的地までの道のりを尋ねる対話を用意した．本研究では対話の内容を，応答までの適切な時間が異なることを想定したものにし，いくつかの発話タイミングを提示することで，どのようなタイミングが適切かを調査する．

表 1　天気を尋ねる対話
Table 1 Dialogue about the weather

Dialogue corpus		Tag
Subject	：コンサート楽しみだね.	s
System	：はい，楽しみです.	s^aa
Subject	：でもちょっと天気が不安かな.	s
System	：調べてみますね.	s^cc
Subject	：明日の天気はどう？	qw
System	：明日の天気は晴れです.	s^aa

表 2　漫才の対話
Table 2 Dialogue about the MANZAI

Dialogue corpus		Tag
Subject	：もうすぐ金沢に新幹線が来るね.	s
System	：そうですね.	s^aa
Subject	：東京までどれぐらいか知ってる？	qw
System	：3 時間ぐらいですか？	qo^d
Subject	：3 秒かな.	s
System	：そんなあほな.	s

表 3　観光相談の対話
Table 3 Dialogue about the tourism

Dialogue corpus		Tag
Subject	：観光に行きたいね.	s
System	：どの辺に行きたいですか？	qo
Subject	：とりあえず西のほうかな.	s
System	：京都とかどうですか？	qo^d
Subject	：京都の～･･･	s
System	：お寺巡りはどうでしょう？	qo^d

表 4　目的地までの対話
Table 4 Dialogue about the destination

Dialogue corpus		Tag
Subject	：ハチ公までもうすぐ？	qo^d
System	：もうすぐです.	s^aa
Subject	：次の交差点を右だっけ？	qo^d
System	：今調べています.	s^cc
Subject	：早く急いで！	s^co
System	：次の交差点を左です.	s^co

また，このような実験に用いる対話の内容については，さまざまな可能性が考えられる．本研究では，まず応答に対する時間的な条件が異なることが予想される対話を用意したが，今後さまざまな対話内容を使用していくことを考えると，発話単位タグを用いてタグ付けを行い，文を内容的，および形式的に区別していくことが望ましい．発話単位タグには，最も適切なタグを一つだけ付与するものや，該当する全てのタグを付与するものなどがある．本研究では，対話文にはさまざまな趣意があると考え，非タスク指向型や同調型対話で比較的使用されていると考えられる MRDA[7]を用いてタグ付けを行った．実験に用いた 4 種類の対話内容を表 1 から表 4 に示す．

実験は，オープンソースソフトウェアである汎用大語彙連続音声認識エンジン「Julius[8]」を用いて構築した音声対話システムを使用して行った．発話は表 1 から表 4 の対話内容を元に被験者，システムの順に発話を行っていく．発話タイミングを評価する部分は 3 発話目であり，1,2 発話目は文脈をつくるための対話としている．そのため，1,2 発話目の発話タイミングは，先行研究[3]で最も良い評価を得ていた固定値のタイミング (600msec) を設定した．3 発話目において，オーバーラップのタイミングを厳密に制御することは非常に難しい．そのため，まず被験者の発話を各 5 回ずつあらかじめ録音しておき，平均発話長を算出して，そこからオーバーラップさせる時間を求めた．次に，求めた時間分だけ Windows の Sleep 関数を用いて待機させることによって，オーバーラップするタイミングを実現した．

実験において，3 発話目のタイミングに使用する 6 つのタイミングを以下に示す．式において *PD* (Pause Duration) は反応潜時長を表している．

Timing 1 ： *PD* = - 400 *msec*
Timing 2 ： *PD* = - 200 *msec*
Timing 3 ： *PD* = 0 *msec*
Timing 4 ： *PD* = 400 *msec*
Timing 5 ： *PD* = 600 *msec*
Timing 6 ： *PD* = 800 *msec*

本実験では，これらの発話タイミングの評価を行うが，実際に音声対話システムが，実験中に反応潜時長を制御できているかについては実験後に検証する必要がある．そのために，ボイスレコーダ (Roland 社：R-09HR) を使用して実験中における音声を全て録音するようにした．また，システム側のデータ (発話長，反応潜時長の設定値，対話の合計時間，合計時間から逆算した反応潜時長) をログとしてシステムに保存し，これらのデータもシステムの動作検証に用いた．

2.2　実験手順

実験は以下の手順で行った．

手順 1：被験者に実験目的と評価用紙の記入方法，実験の流れ，発話タイミングついての説明を行う．

手順 2：発話－応答の練習を行う．
手順 3：被験者は評価用紙の「練習」の記入欄に記入を行う．
手順 4：3 発話からなる連続対話を 1 つの Timing 試行として，被験者は実験者の指示に従って発話を行っていき，3 発話目のタイミング評価を行う．発話は先行の Timing と後行の Timing を交互に 2 回ずつ行う．
手順 5：被験者は評価用紙の記入欄に記入を行う．
手順 6：手順 4 から 5 までを 1 セットとして，これを全ての Timing の組み合わせ分(15 回)行う．

6 種類の発話タイミングの評価は一対比較法(中屋の変法)を用いて行った．評価項目としては，「自然か」，「好きか」，「親しみやすいか」，「話しやすいか」，「活発か」，「丁寧か」，「速いか」の 7 つの項目を用意し，5 段階で評価させるようにした．項目中，「速いか」は被験者が発話タイミングの違いを正しく認識しているかどうかを調査する目的で用意し，それ以外の項目は，被験者の印象を調査する目的で用意した．

対話内容，Timing の選択は，順序効果が出ないように，ランダムに行った．また，システムが応答直後に次の発話を行った場合，システム側でログデータが正確に取得できないことがあったため，1 発話毎に約 1500msec 間隔を空けるようにした．音声認識に失敗した場合，先行と後行の Timing の組み合わせをもう一度行うようにした．

図 1 に実験環境を示す. 実験には健常な 20 代の男子大学生 10 名が参加した(平均年齢 22.5 歳)．被験者は，図 1 に示すようにモニターと向かい合うように椅子に座った．その後，ワイヤレスヘッドセット (Audio Technica 社：PRO9HEW/P) を装着し，モニターに表示される指示に従って発話を行った．このとき，発話を行う速度は被験者にとって自然な速度で行うように指示した．また，視覚情報が実験結果に影響することを考慮して，実験者と被験者の間にカーテンを設置した．

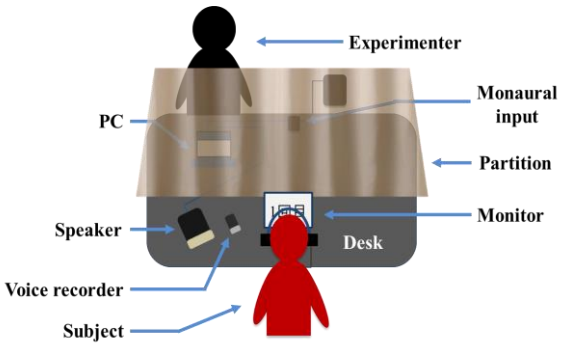

図 1　実験環境
Fig.1 Experiment environment.

3.　実験結果

表 5 から表 8 に，実際に計測された各タイミングにおける反応潜時長の平均値と標準偏差を示す．平均値をみると，実際に制御しようとした反応潜時長と比べて誤差が生じているのがわかる．これらの時間的なずれは，システムの認識に要する時間に揺らぎがあることや，発話データの個人差によるところが多く，現状では 0 にすることは難しい．ただし，各タイミング間の差はほぼ確保できており，被験者も (図 2 から図 5 の“fast”の結果より) その速さを適切に認識しているため，以下にこれらの反応潜時長で実験が行われた元での評価の結果について述べる．

天気を尋ねる対話における評価の結果を図 2 に，漫才の対話における評価の結果を図 3 に，観光相談の対話における評価の結果を図 4 に，目的地までの道を尋ねる対話における評価を図 5 にそれぞれ示す．まず，天気を尋ねる対話の結果についてであるが，全ての項目において主効果に統計的な有意差がみられた($p < .05$)．その中で Timing4 が最も良い評価を得ていることが図よりわかる．一方, オーバーラップする Timing1,2 は総じて悪い評価を得ていることがわかる．また，正の反応潜時長である Timing5 については比較的良い評価を得る傾向があるこ

表 5　天気を尋ねる対話における平均反応潜時長と標準偏差
Table 5 Mean and S.D. of pause duration of the dialogue about the weather

	-400msec	-200msec	0msec	400msec	600msec	800msec
Mean	-381.1	-189.3	-11.23	423.9	607.6	804.3
S.D.	119.1	127.7	113.5	19.7	31.31	32.79

表 6　漫才の対話における平均反応潜時長と標準偏差
Table 6 Mean and S.D. of pause duration of the dialogue about the MANZAI

	-400msec	-200msec	0msec	400msec	600msec	800msec
Mean	-371.0	-159.1	19.13	433.1	622.3	831.0
S.D.	105.5	126.7	97.93	73.21	110.0	43.98

表 7　観光相談の対話における平均反応潜時長と標準偏差
Table 7 Mean and S.D. of pause duration of the dialogue about the tourism

	-400msec	-200msec	0msec	400msec	600msec	800msec
Mean	-329.9	-175.0	2.360	408.6	604.7	798.6
S.D.	165.4	142.4	146.9	34.50	48.54	41.84

表 8　目的地までの対話における平均反応潜時長と標準偏差
Table 8 Mean and S.D. of pause duration of the dialogue about the destination

	-400msec	-200msec	0msec	400msec	600msec	800msec
Mean	-454.8	-254.2	-73.79	420.8	636.5	829.1
S.D.	112.9	112.9	98.48	24.23	34.66	39.38

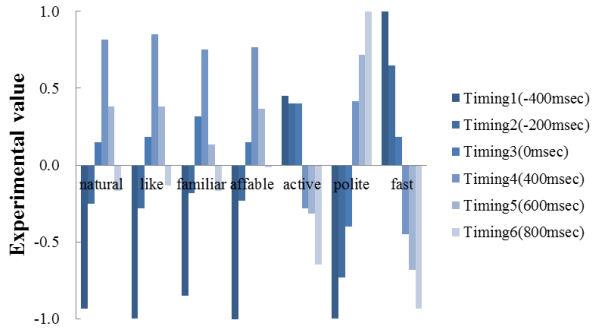

図 2　天気を尋ねる対話における印象評価の結果
Fig.2 Result of pairwise comparison in the dialogue about the weather

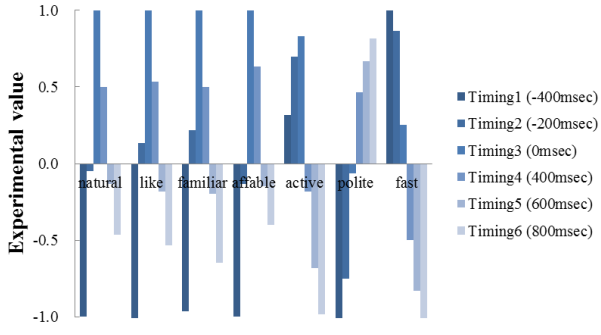

図 3　漫才の対話における印象評価の結果
Fig.3 Result of pairwise comparison in the dialogue about the MANZAI

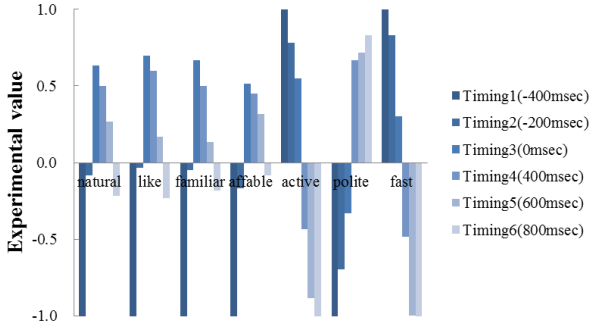

図 4　観光相談の対話における印象評価の結果
Fig.4 Result of pairwise comparison in the dialogue about the tourism

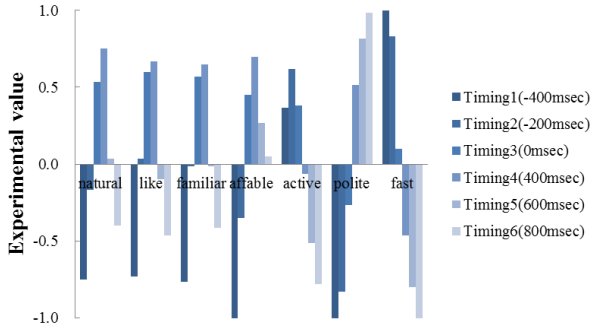

図 5　目的地までの対話における印象評価の結果
Fig.5 Result of pairwise comparison in the dialogue about the destination

とが図より見てとれる．

次に，漫才の対話の結果についてであるが，まず全ての項目において主効果に統計的な有意差がみられた($p < .05$)．その中で最も良い評価を得ていたのは Timing3 であった．発話の前半でオーバーラップする Timing1,Timing2 に関しては，Timing1 は総じて悪い評価を得ていることがわかる．一方で Timing2 は「好きか」，「親しみやすいか」の項目において良い評価を得る傾向がみられた．

続いて，観光相談の対話の結果についてであるが，全ての項目において主効果に統計的な有意差がみられた($p < .05$)．その中で最も良い評価を得ていたのは Timing3 であった．また，Timing4 も同程度の良い評価を得ていることが図よりわかる．一方，発話の前半や途中でオーバーラップする Timing1,Timing2 は総じて悪い評価を得ている．しかし，天気を尋ねる対話と比較すると，若干ではあるが良い評価を得ていることがわかる．

最後に，目的地までの道を尋ねる対話の結果についてであるが，全ての項目において主効果に統計的な有意差が見られた($p < .05$)．その中で，最も良い評価を得ていたのは Timing4 であった．また，Timing3 も同程度の良い評価を得ていることがわかる．天気を尋ねる対話と比較すると，発話の途中でオーバーラップする Timing2 の評価が良くなっていることが図よりわかる．一方で，正の反応潜時長である Timing6 は比較的悪い評価を得る傾向にあることがわかる．

また, 以上の 4 種類の対話内容に共通した結果として，被験者はオーバーラップするタイミングを基本的には活発と感じている一方で，丁寧とは感じない傾向があることが，図より見てとれる．

4.　考察

本研究では，3 発話からなる対話を 4 種類用意し，対話の内容によって，オーバーラップを含む音声対話システムの発話タイミングの印象がどのように変化するかを調査した．結果として，発話内容によって適切な発話タイミングが異なることが明らかになった．発話内容や文脈によって適切な応答タイミングが変化するということは，直感的には理解できていることであるが，より自然な音声対話システムの構築のためには，本研究のように定量的にどの程度の値が適切であるのかを示すことが，まずは重要であると考えられる．

まず，天気を尋ねる対話においては，反応潜時長が 400msec 前後から 600msec 程度の比較的遅い応答タイミングが相対的に良い評価を得る傾向にあった．先行研究[9]において，新しい情報や重要な情報を扱う対話ではオーバーラップはあまり起きないことが示唆されている．対話の 3 発話目は新しい情報を扱うものであったため，結果として，正の反応潜時長が良い評価を得ることにな

ったと考えられる．

次に，漫才の対話においては，反応潜時長が 0msec 前後の発話終了直後に応答を行う比較的速い応答タイミングが相対的に良い評価を得る傾向にあった．また，反応潜時長が-200msec 程度の発話の途中でオーバーラップするタイミングに関しては，「好きか」，「親しみやすいか」の項目で良い評価を得る傾向にあった．漫才では，あらかじめどのような対話を行うかは共有されており，ツッコミが行われる部分を理解していると考えられる．そのため，発話を遮られる印象が薄く，文末付近でのオーバーラップが良い評価を得ることになったと考えられる．

3 つ目の対話内容である観光相談の対話においては，0msec 前後の発話終了直後に応答を行う，比較的速い応答タイミングが相対的に良い評価を得る傾向にあった．観光相談の対話における 3 発話目は言い淀むような対話となっており，それに対してシステムが応答を返すものとなっている．このような対話はオーバーラップ発話における「相手発話継続」とされており，相手の発話を理解しているという意思表明であることが示唆されている[10]．そのため，発話を遮る印象が薄れ，文末付近でオーバーラップするような負の反応潜時長が良い評価を得ることになったと推測される．

4 つ目の対話内容である目的地までの対話においては，0msec 前後から 400msec 程度の比較的速い応答タイミングが良い評価を得る傾向にあることがわかった．また，-200msec 程度の発話の途中でオーバーラップするタイミングは，「好きか」の項目において若干ではあるが良い評価を得る傾向にあった．この対話ではその内容から切羽詰っているような状況であることがわかる．このため，反応潜時長が 0msec から 400msec にかけての比較的速い応答タイミングが好まれたと考えられる．しかし，扱っている情報が道順という重要な情報であるため，オーバーラップの評価が悪くなったと考えられる．

先行研究で述べられているように，オーバーラップは対話が盛り上がっているときや，確認を行うときに起きやすいことが知られている．今回の実験では，漫才の対話や観光相談の対話で，発話終了直後に応答を返す発話タイミングが適切であることが明らかになった．漫才の対話では，適切なタイミングでツッコミを入れるためにあらかじめどのような対話が行われるかが理解され，発話のタイミング自体が対話の面白さを決めている．また，観光相談の対話では，求める情報は 2 発話目までで明確になっており，3 発話目では相手の要求を補完するような形の発話が行われている．これらの対話の共通点として，2 発話目までに文脈の共有が進んでおり，3 発話目に求める情報をある程度推測できていることが挙げられる．一方，天気を尋ねる対話では，「コンサート」や「明日」，「天気」という情報は共有されているが，実際に求めている情報を共有するのは 3 発話目である．そのため，文脈の共有度が低く 3 発話目の内容を推測できないため，発話を遮る印象が強くなったと考えられる．また，目的地までの対話では，2 発話目までで文脈の共有がされているものの，扱っている情報が重要なものであるため，オーバーラップの印象が悪くなったと推測される．これらのことを考えると，オーバーラップを含めた応答タイミングの印象は，話者間でどの程度文脈を共有し，その後の発話内容をどの程度推測できるかということに大きな影響を受けていることが示唆される．

今後，このようなオーバーラップを含めた発話タイミングを考慮した音声対話システムを構築していくためには，さらにさまざまな対話状況における発話タイミングの構造を明らかにしていく必要がある，しかし，実際には対話の種類の数が膨大になるため，ただ数を増やしていくのは非現実的である．そこで今後は，今回用いたようなタグを付与してあるコーパスデータを用いて，対話をある程度分類した上で，それぞれの対話の状況においてどのような発話タイミングが適切であるか調べていく必要がある．

また，今回の実験では反応潜時長が十分に制御できない場合があった．先行研究[3]では，「こんにちは」という単純な発話の時に 20msec 程度の誤差で反応潜時長を制御できることが明らかになったが，今回のように複数の発話が続き，また発話内容がやや複雑になると個人差の影響もあり，その制御が難しくなることが明らかになった．今後，このような誤差をどのように抑えることができるかについても，さらに検討していく必要があると考える．

5. まとめ

本研究では，3 発話からなる対話を 4 種類用意し，対話の内容によって，オーバーラップを含む音声対話システムの発話タイミングの印象がどのように変化するかを調査した．結果として，発話内容によって適切な発話タイミングが異なることが明らかになった．

今後，さまざまな対話内容や文脈において，オーバーラップを含む発話タイミングが与える印象をより明確にしていき，実験で得られた結果を元に，発話タイミングを生成できるモデルの構築を行っていく．

6. 参考文献

[1] Birdwhistell, R.: Kinesics and context : Essays on body motion communication; Philadelphia, University of Pennsylvania Press (1970).

[2] Mehrabian, A., Ferris, S.R.: Inference of attitudes from nonverbal communication in two channels; Journal of Consulting Psychology, Vol.31,pp. 248-252 (1967).

[3] 小林，大村，山本: 音声対話システムにおける適切な発話タイミング生成に関する考察; ヒューマン

インタフェース学会研究報告集，Vol.15，No.9，pp.23-28 (2013).
[4] 西村，北岡，中川: 対話における韻律変化・タイミングのモデル化と音声対話システムへの適用; 人工知能学会，言語・音声理解と対話処理研究会，SIG-SLUD-A602-7，pp.37-42 (2006).
[5] 大村，小林，山本: 音声対話システムとの質問－応答型対話におけるオーバーラップを含む発話タイミングの評価; 平成 26 年度電気関係学会北陸支部連合大会，pp.27 (2014).
[6] 大村，吉田，山本: 音声対話システムにおけるオーバーラップを含む発話タイミングの評価; ヒューマンインタフェース学会研究報告集, Vol.16, No.8, pp.29-32 (2014).
[7] Dhillon, R., Bhagat, S., Carvey, H., Shriberg, E.: Meeting Recorder Project:Dialog Act Labeling Guide; ICSI Technical Report TR-04-002 , International Computer Science Institute (2004).
[8] Julius: http://Julius.sourceforge.jp/
[9] 藤原，伊藤，荒木: タスク指向対話における相互の対話意図を考慮した対話リズムの分析; 人工知能学会，言語・音声理解と対話処理研究会，SIG-SLUD-A701，pp.45-50 (2007).
[10] 榎本，土屋: オーバーラップ発話の評定方法とその基礎統計：日本語地図課題対話を通して; 電子情報通信学会技術研究報告, 言語理解とコミュニケーション，Vol.99，No.524，pp.13-18 (1999).