ヒューマンインタフェース 65－7
音声言語情報処理 10－7
（1996. 2. 29）

韻律パターンの認識を用いた相槌挿入とその評価

○岡登洋平 加藤佳司 山本幹雄 板橋秀一

筑波大学 電子・情報工学系 〒305 茨城県つくば市 天王台 1-1-1

概要

人間と同様に相槌を挿入することが機械との対話においても可能であれば、話しやすさの改善につながる可能性がある。本稿では相槌の挿入される発話の性質、相槌挿入の実装上の問題点、試作した相槌挿入システムについてについて述べる。まず相槌が挿入される発話の性質を調べ、発話長、発話速度、F_0と関連があることを確かめた。また知覚実験の結果を用い、同一対話による相槌の挿入傾向の定量的評価と、タイミングについて要求される精度、対話システムによる実装の際の問題点(発話終了の予測、必要とされる精度)についての考察を行なった。最後に、以上の検討に基づいた韻律パターンの認識による相槌挿入システムを試作し、評価を行なった。

Prosodic pattern recognition for insertion of interjectory responses and its evaluation.

Yohei OKATO Keiji KATO Mikio YAMAMOTO Shuichi ITAHASHI

University of Tsukuba Institute of Information Sciences and Electronics,

1-1-1 Tennodai, Tsukuba, Ibaraki, 305 Japan

abstract

User's comfortableness in man-machine speech dialog environment may improve if speech dialog systems insert correct aizuchi(interjectory responses) against user's utterances. This paper describes nature of aizuchi, required conditions for aizuchi insertion systems and evaluation of an implemented system. At first, we analyze human-human speech dialog corpus and show that utterances inserted aizuchi have special properties of duration, speech and F_0. We also investigate utterances which aizuchi could be inserted and relation to utterance delay and dialog disfluency by perceptual experiments. Since the experiments indicate severe condition on insertion timing of aizuchi, an aizuchi insersion system has to predict end of target utterance. Lastly, we propose a mechanism for aizuchi insersion system and carry out its evaluation.

1. はじめに

従来の音声対話システムでは、システムはユーザの発話が終了するまで、一切、応答を行なうことができない。このために、ユーザは話していて、システムの理解状況がどうなのか、全処理を終えるまでわからず、システムの応答性に問題があることが指摘されている [1]。人間同士の実際の対話では、話し手は聞き手の状態を把握しつつ、対話を行なっている。その伝達は相槌など音声によるもの、顔の表情、ジェスチャーなどがあるが、本稿では音声による相槌などの応答をシステムでも行なうことを考える。

相槌を挿入するには、対話システムが発話をどの程度の理解しているかなどの要因も重要である考えられるが、本稿では、表層的な情報として、韻律パターンを用いた相槌の挿入を考える。

本稿ではまず、相槌が挿入される場合の音声パターンの特徴を調べる。次に相槌の評価のために、相槌がどのような発話に入りうるか、またどの程度の遅れが許容されるかについての、知覚実験を行なう。次に対話システムによって相槌を挿入する際に問題となる、予測処理と精度についてまとめ、最後に韻律パターンを用いた相槌挿入システムを試作し、評価を行なう。

表 1: 相槌の呼応先とそれ以外の発話の韻律的性質

	オペレータ		注文者	
(句音声単位)	相槌あり	相槌なし	相槌あり	相槌なし
発話長 (秒)	1.06	0.88	0.96	0.60
発話速度 (モーラ／秒)	8.0	8.0	6.4	7.0
最大 F0(正規化)	0.44	0.76	0.95	0.59
最小 F0(正規化)	−1.99	−1.08	−1.52	−0.66
(やりとり単位)	相槌あり	相槌なし	相槌あり	相槌なし
発話長 (秒)	3.19	1.59	1.50	0.86
発話速度 (モーラ／秒)	7.4	7.9	6.1	7.0
最大 F0(正規化)	1.09	0.90	0.83	0.68
最小 F0(正規化)	−2.50	−1.51	−1.71	−0.79

2. 対話の分析

まず相槌が挿入された場合と、それ以外の発話の違いを検討する。次に相槌の挿入されたタイミングについて検討する。分析に用いた対話は筑波大で収録したテレフォンショッピングをタスクとした5対話、合計約17分である [2]。この対話は二人の話者（注文者とオペレータ）が顧客情報、買い物のリスト、支払い方法などの情報をオペレータが主導的に、非対面で伝えるタスクである。オペレータ役は固定されており、事前に準備を行なっている。話者数はオペレータ1名、注文者3名である。

対話の流れはあらかじめ計画されており、マニュアルに従って、以下の順にオペレータが誘導する。

(1) 顧客情報の取得
(2) 注文情報の取得
(3) 情報の確認、

対話は2種類の長さの単位で分析を行なう。一つは文節あるいは、句点、読点、言い淀みなどによるポーズで区切られた「句音声」単位である (総数:1251)。もう一つはそれをつなぎ合わせたもので、ほぼ文に相当する「やりとり単位」である。ただし、発話の途中で相槌が挿入されたり、割り込まれた場合などは割り込みや相槌の呼応先の句音声で区切る (総数:720)。この例を図1に示す。

これは相槌が挿入される位置の局所的性質とその発話全体の性質を見ることを目的としている。相槌はオペレータ57個、注文者88個、計145個観察された。

また、この5対話は注文者がオペレータの要求する事項を伝えるというほぼ同様の話題展開を示す。そのため、注文者とオペレータでは発話の性質が異なるので、オペレータと注文者の発話はそれぞれ別に扱う。分析を行なったパラメータは F_0 、各発話単位の時間長、モーラ数、発話速度である。なお、F_0 は対数を取ってから、各話者ごとに平均0、分散1に正規化した。まず各パラメータの統計的性質を表1に示す。

開始時間	終了時間	話者	句音声単位	やりとり単位
142.63	143.29	A	商品名	
143.29	144.68	A	スチールラック	
145.00	145.68	A	品番	
145.68	146.47	A	シーエイチの	
146.47	147.50	A	3 5 7	
147.39	147.72	B	はい (相槌)	
148.06	149.07	A	注文コード	
149.07	150.37	A	8 3 3 5	
150.38	151.30	A	8 2 5	

図 1: 分析の単位

2.1 相槌が挿入された発話の特徴

表1によると、相槌が挿入された場合とそうでない場合の違いは、やりとり単位の方では、発話長、発話速度に見られる F_0 はどちらの話者でも共通した違いが見られる。また、句音声単位の場合でも最小 F_0 には大きな違いがある。これらの各事項を以下に検討する。

まず、相槌が挿入された発話はかなり長めであることがわかる。やりとり単位では、発話の長さはオペレータと注文者ではほぼ2倍異なるが、挿入された句音声の長さはそれほどは変わらない。そこでやりとり単位の発話長の分布を見ると、図2のようになった。ここで発話が t 秒続いたとき、この発話に相槌が挿入される割合は以下の式のようになる [9]。

$$t\text{秒続いた発話に相槌が挿入される割合} = \frac{\text{発話長が } t \text{ 秒よりも長い発話に挿入された相槌総数}}{\text{発話長が } t \text{ 秒よりも長い発話の総数}}$$

この発話長と相槌の挿入される割合を、やりとり単位についてについて示したものが図3である。これを見ると、発話時間が長くなるにつれて、相槌が挿入されて終了して

いる割合が増えていることがわかる。同様の分析を最低 F_0 についても行なった(図4,5)。これらは相槌全体の機能の一つを表すものであり、相槌を挿入する条件の一つであると考えられる。

また発話速度はやりとり単位では相槌が入る場合は遅くなっている。これは高木[3]から、この速い発話と遅い発話はそれぞれ文末に、1) キーワード的な語がある場合、2) 文末に現れる定型表現のそれぞれの場合に対応しており、オペレータと注文者の立場の違いを表していると考えられる。また最大 F_0 の傾向の違いもこの理由と思われる。

なお、相槌が挿入された発話において、発話継続時間と発話速度の相関係数は、句音声単位で -0.07、やりとり単位で 0.15 と小さく、それらが独立して起こる現象であることを示している。

2.2 相槌が挿入されたタイミングの分析

次に相槌が挿入されるタイミングについて調べた。話者が交替する時のポーズ長を相槌の場合と、全体の場合の分布を図4に示す。相槌が挿入されるタイミングは発話終了から平均 0.07 秒、標準偏差が 0.24 秒であった。

発話終了から相槌が挿入される時間は、ここで調べた韻律的な情報との間の相関はほとんど見られず、むしろ発話の最後に「～ので」「～ます」のような発話の切れ目を表す定型的な表現が来たときに、相槌が速く挿入される傾向がある(表2)。これはそれらの発話には実質的な情報がほとんど含まれていないためであると考えられる。また今回の対話ではオペレータの発話に対する注文者の相槌の方が速い傾向が見られたが、これは話者の違いによる影響と考えられる。

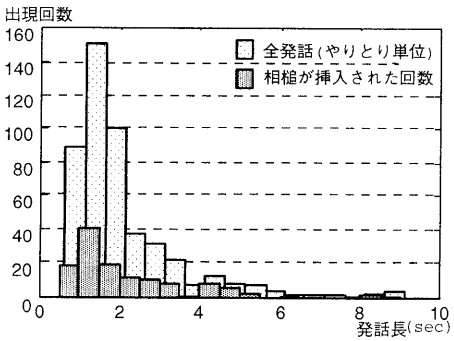

図2: やりとり単位の発話長の全体の分布と相槌の前の分布

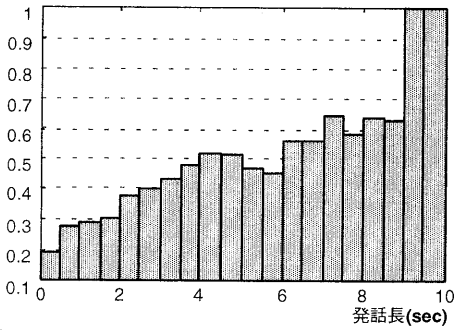

図3: 発話長と発話が相槌によって終了した割合の関係

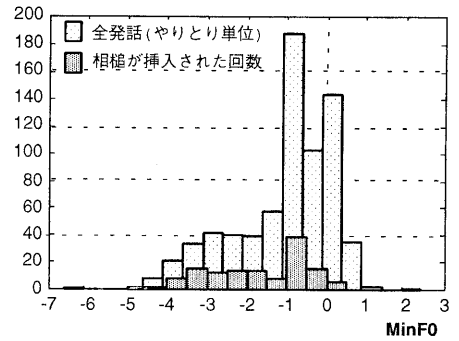

図4: 全発話と相槌の前の発話の $F_0(\min)$ の分布

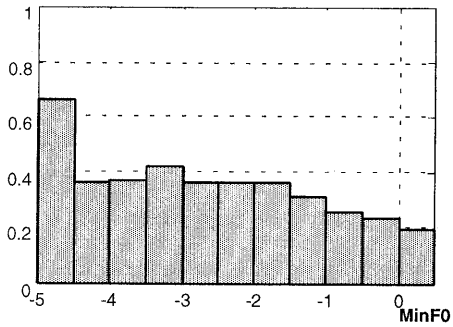

図5: 最小 F_0 と発話が相槌によって終了した割合の関係

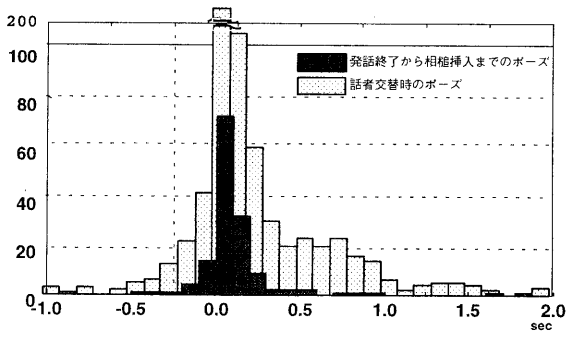

図6: 話者交替時のポーズ長

表2: 発話の終端の語と挿入タイミングの関係

文末の語 \ 単位 (秒)	オペレータ		注文者	
	MEAN	SD	MEAN	SD
名詞、 名詞＋1文字の助詞	0.15	0.31	0.06	0.20
その他定型表現	−0.06	0.20	−0.09	0.14

表3: 最低 F_0 の出現時間と、発話終了時間の関係

	相槌前		それ以外	
	MEAN	SD	MEAN	SD
句音声単位	0.20	0.22	0.23	0.29
やりとり単位	0.70	1.44	0.34	0.66

以上まとめると、相槌が挿入された発話では発話長と F_0 に影響がある。発話速度にも影響が見られるが発話の内容によって性質が異なる。相槌のタイミングは文末の表現などの決まりきっていて、情報の少ないところでは早く挿入される傾向がある。

3. 相槌の挿入とタイミングに関する知覚実験

相槌はもともと入っても入らなくても良いと考えられるので、対話システムなどで扱う上で、どの程度の相槌を入れるべきであるのか検討する必要がある。また図4に示したように相槌は発話終了から処理に使える時間が短いので、対話システムで用いる場合には、どの程度のタイミングであれば適切であるか明らかにする必要がある。

そこで、これらの点について明らかにするため、知覚実験を行なった。実験は相槌の挿入しうる発話を探すものと、相槌を始めとする発話の遅れに対するものである。実験に参加した被験者は大学生を中心とした男性22名、女性12名である。

3.1 実験1: 相槌が挿入されうる発話の検出

次に、相槌が対話中のどのような発話が相槌をより多く挿入される傾向があるのか、また、実際の対話中での相槌と比べてどうかという点について知覚実験を行なった。

まずテレフォンショッピングの対話を編集し、その中の相槌と判定したものを全て除去する。次にこの対話を被験者に聞かせ、相槌を挿入したいと思った時点でボタンを押してもらう。ボタンを押すと、被験者が聞いている対話中に、注文者側の声で「はい」という相槌が実際に挿入され、押した時間が記録される。対話は2通り (tsu1103,tsu1208) を準備し、注文者の性別が被験者と同じものを利用した。実験後集計し、相槌でない部分 (注文者が発話権をとる場面や、Yes-No-Questionの部分に多い) を削除した。各発話の挿入タイミングの標準偏差は平均すると0.26秒程度であった。

実験で得た相槌のうち、同一箇所に2人以下のデータしかない点を除いて、実対話で出現したもの、実験でのみ出現したもの、どちらにも出現したものと分けると表3のようになった。知覚実験中、同一の発話に対して、相槌を挿入した被験者数の分布は図7のようになった。この結果から、相槌がほぼ確実に入る発話が存在することが明らかになった。そのような発話は、オペレータが行なう長い確認の発話に対して、その文の終りに挿入されたものであった。

表4: 実対話中と知覚実験での相槌の数

	tsu1103(男性)	tsu1208(女性)
実対話のみ	4	8
知覚実験のみ	9	2
両方に出現	16	22
実験回数	22	12

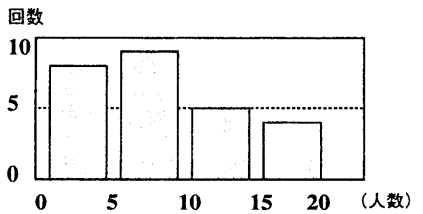

図7: 同一発話に対して相槌を挿入した人数

3.2 実験2: 相槌のタイミングの評価

次に相槌を挿入するタイミングについての実験を行ない、相槌がどれだけ遅れると不自然になるかというのを調べた。

実験の方法は以下の通りである。まず対話の編集を行ない、話者が交替した場面の句音声単位の発話間のポーズを **300msec, 600msec, 900msec** のいずれかに固定した対話を作成する。

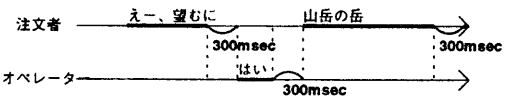

図8: ポーズを300msecにした対話

被験者にはその情報を知らせずにその対話を聞かせ、聞いている途中で遅いあるいは速いと違和感をを感じた場合にはボタンを押してもらう。ボタンは速い、やや速い、やや遅い、遅いとの4通りがあり、ボタンを押すとその時点の時間と内容が記録される。この実験を一人3回行なう。ただし2回目には、無修正の元の対話を聞かせた。被験者は前述した34人である。

実験の結果、特に相槌周辺で違和感を感じた場合が多く、まとめると以下のような結果を得た。

- ポーズ長 **300msec** に固定
  相槌の部分で違和感は少なく、それ以外の部分でやや速いと感じている。全体的に違和感を感じた点は少ない。

- ポーズ長 **600msec** に固定
  発話終了から相槌が挿入されるまでの時間が遅いと感じている。それ以外の部分では違和感は少ない。全体的に違和感を感じた点は少ない。

- ポーズ長 **900msec** に固定
  発話終了から相槌が挿入されるまでと、相槌が挿入されてから、次の発話が始まるまでのいずれにおいても遅いと感じている。それ以外の部分では比較的、違和感を感じていない。全体的に違和感を感じた部分が多い。

この結果、相槌の評価基準としては600msecでは遅く、相槌としての発話は発話終了から300msec程度で応答する必要があると考えられる。これらの結果は図6の相槌の挿入される時間分布の急激に小さくなる部分と対応していると考えられる。ただし実験後に行なったアンケートでは300msecの相槌でもやや遅いと感じられている傾向がある。

本実験では1対話平均4分強の長さがあり、被験者はそれぞれの対話のリズムやテンポに適応して、対話を聞いていると考えられるが、実際に対話に参加する直接評価と、今回の実験のような間接評価の違い[8]についても考慮する必要があると考えられる。

4. 相槌挿入の問題点

対話システムによって相槌を挿入するには、1）相槌の挿入される発話の検出、2）相槌を挿入するタイミングの検出、について考慮する必要がある。

相槌がどのような場合に、挿入されるかは、状況や文脈などの情報や、発話者の意図にも依存すると考えられるが、2.1の分析から、表層的な情報を使うこともできると考えられる。

また、相槌は、発話の切れ目や終了とほぼ同時に挿入される[5, 6]。相槌のタイミングは発話の内容や相槌の機能により、若干異なると考えられるが、対話システムによって実対話と同様なタイミングで相槌を挿入するには、発話が終了する前に発話の終了を予測し、相槌を挿入する必要がある[7]。このためには、処理はユーザの発話と時間同期で処理される必要がある。また、どれだけ早い時点で予測を行なわなければならないかは、システムの処理時間に依存する。一般に予測の時間が伸びるほど、予測の精度が落ちると考えられるため、相槌挿入の処理時間と挿入タイミングの精度についても検討する必要がある。このことについては、前に述べた知覚実験の結果を使うことができると考えられる。

5. 韻律パターンの認識による相槌挿入

相槌の挿入される発話は韻律的にも特徴があることは前に述べた。そこで、本稿では、相槌を挿入する発話と、相槌のタイミングの検出を韻律的な特徴から抽出し、実際に相槌を挿入するシステムを提案する。

4章で述べた1）の相槌の挿入される発話の検出は表層的な情報として、2章で述べた発話長などを用いる。また、以下で述べる韻律テンプレートに現在の発話と同様のものがあるとき、相槌が挿入できる場合であるとする。他にも対話の文脈情報などを利用できると考えられる。

また、2）の相槌のタイミング検出は実際の発話に現われたピッチパターンをテンプレートとしたテンプレートマッチングを用いて、類似した発話を検出し、テンプレートの終了を検出した時点で相槌を挿入する。このテンプレートの大きさは2章の分析で用いた句音声単位程度を想定する。この手法の利点としては、実際に現われた発話に対してはどのような発話でも相槌挿入のタイミングを検出することができることである。またテンプレートとして用いる発話の長さを適当に変えることで予測時間を容易に変えることができる。以下に、ここで使用した手法について説明し、次節で評価を行なう。

なお、入力パラメータには対数をとった後で話者ごとに平均、分散1になるように正規化した F_0 とパワーを用いた。

用いたテンプレートとして、自己ループ付きの状態遷移モデルを用いる(図9)。各状態は実際のパターンの何フレームかに相当する。各状態は対応する音声フレームの平均を代表として持ち、各状態の滞在時間にはPoisson分布を仮定した制限を行なった。またパターン全体としての概形が歪みすぎないように、現在いる状態の一つ前の状態までの伸縮に合わせ、割合を同じに保つように、現在の状態の滞在できる時間長を修正を行なう。また400msec以下には縮まないように制限する。

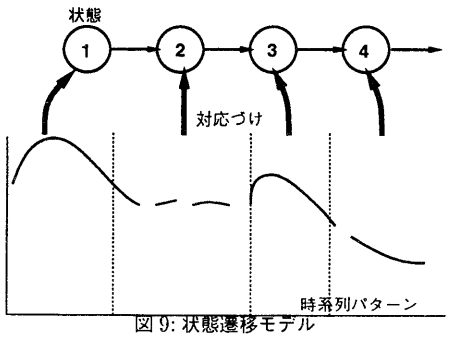

図9: 状態遷移モデル

テンプレートの学習にはDPを使った方法を用いる。学習の順序は以下の通りである。

1) まず各状態に均等に有声フレームを振り分ける。
2) 各状態ごとに割り当てられたフレームの平均で各状態の代表パターンを作成する。
3) 現在のテンプレートで、学習サンプルを認識して、状態ごとの学習テンプレートを新たに割り当てる。
4) 3)で得た学習サンプルの認識のスコアが収束したら学習終了、そうでなければ2)に戻る。

なお学習に用いるデータは、相槌の挿入された発話の韻律パターンから、想定したシステムの処理時間（予測時間）を引いたところまでを用いる。

認識は入力の韻律パターンと時間同期で、終端フリーのDPマッチングを行なう。各状態の滞在時間には前に述べた制限がかかっている。相槌の挿入とタイミングの判断は、以下のように行なう。まず時間フレームごとに、それぞれの相槌のタイミングを検出するテンプレートについて、状態iから次の状態に遷移するスコア(S_i)を計算する。最終状態からの遷移スコア(S_{LAST})はモデルが終了する場合のスコアである。このとき、相槌を挿入すると判断するのは以下の3条件を満たした時とする。

1) S_{LAST}がモデルの全状態中で一番スコアが大きい
2) スコアS_{LAST}がしきい値よりも大きい
3) 最終状態にモデルの定めた滞在時間だけ留まっている

6. 評価実験

5章で述べた相槌挿入手法の評価を行なう。

まず、3.1節で行なった知覚実験で相槌が挿入された発話38個をそれぞれを相槌が挿入される発話のテンプレートとして使用する。テンプレートには予測時間200msecのものと、400msecのものを作成した。入力パラメータには対数をとった後で話者ごとに平均、分散1になるように正規化したF_0とパワーを用いた。

次にこれを用いて2対話に出現する句音声とマッチングを行ない、相槌が挿入される発話とそのタイミングの検出を行なった。スコアのしきい値としてはそれぞれに対して-0.50と-0.30の2種類で行なった。このようにして2対話に出現した発話とマッチングを取った。マッチングの結果、ほとんどのテンプレートとマッチしなかったものについては除き、残りのものについて実験結果を表5に示す。各欄の値は各テンプレートの性能の平均である

句検出率はそのテンプレートが発話全体のどれだけとマッチしたかであり、テンプレートの精度を示す。また被相槌挿入率は検出したテンプレートにどれだけ挿入されたかを示している。タイミングの検出は知覚実験3.2の結果から、発話終了に対して$-100msec$以上$300msec$以下とした。このとき、どれだけ正しくタイミングを検出できたか(タイミング検出率)と、タイミング検出の精度(正解数/システムの検出したタイミング)を示す。これらは実対話で相槌が挿入された発話を正解とした。また実際に相槌が入っていない発話の発話終了に対して同様の評価をおこなった。その結果を発話終了検出率に示す。実験の結果、予測時間の長いものの方が精度で劣り、しきい値を厳しく取ったものの方がテンプレート自体は良いものが選択されていることを示している。

表5: 評価実験

予測時間	200msec		400msec	
スコアのしきい値	-0.50	-0.30	-0.50	-0.30
テンプレート数	22	10	23	9
被相槌挿入率(%)	23	32	21	26
タイミング検出率(%)	69	77	77	75
タイミング精度(%)	41	49	42	47
発話終了検出率(%)	36	44	33	40

7. まとめ

相槌が挿入される発話が韻律的にどのような特徴があるか調べ、韻律情報を用いた相槌挿入システムを作成した。

さらにモデルの精度向上を行ない、相槌の挿入位置の決定できるようにすることと、対話システムを用いた評価を行なうことが今後の課題である。

謝辞

御助言をいただいている、電子技術総合研究所知能情報部音声研究室の皆様に感謝致します。

参考文献

[1] 小高、天野：" 音声対話システムの応答生成と対話性について", 音講論, 1-7-17(1994-3)

[2] 上田直子、高木一幸、板橋秀一："テレフォンショッピング対話の収録と分析", 音講論, 2-Q-21(1995-3)

[3] K. Takagi, S. Itahashi: "Temporal Characteristics of Utterances Units and Topic" Structure of Spoken Dialogs, *IEICE Trans.Inf.& Syst.,Vol.E78-D,No.3 Mar. 1993*

[4] S.K. メイナード："会話分析", くろしお出版,1993

[5] 小坂："あいづちを中心とした会話音声の呼応関係の分析", 信学技報 SP87-107

[6] 小磯、堀内、土屋、市川："下位発話単位の音声的特徴と「あいづち」との関連について", SIG-J-9501-2

[7] 渡辺、佐藤、八木、井宮、市川："音声対話理解システム構想 -CHIBA-", 10th Symposium on HUMAN INTERFACE, Oct., 1994

[8] 西宏之、北井幹雄: "蓄積型音声対話システムにおける発話促進要因の分析と評価", 音声言語情報処理 95-SLP-5, Feb.,1995.

[9] 西宏之、五味和洋、小島順治: "音声対話における確率的発声終了検出法", 信学論 D Vol.J70-D No.11 pp.2108–2114, Nov., 1987.