韻律情報を用いた相槌の挿入

岡登洋平[†] 加藤佳司[††]
山本幹雄[†††] 板橋秀一[†††]

機械とユーザの対話において，機械が人間と同様に相槌を打つことが可能であれば，ユーザの話しやすさの改善につながる．本研究では，話し手の発話間にポーズの出現とほぼ同時に相槌が打たれる場合を対象として，システムによる相槌挿入を試みた．システムが適切なタイミングで相槌を打つにはポーズを検出するよりも早く相槌の挿入判定を行う必要がある．そこで本稿では話し手の発話から抽出した韻律情報を用いて，予測的に相槌の挿入判定を行う手法について検討した．まず対象としたテレフォンショッピングをタスクとした対話について音声を分析し，聞き手の相槌が韻律的に特徴のある話し手の発話箇所で打たれていることを示した．次に相槌音声を消去した対話を聞かせ，相槌の箇所を人間が判定する実験を行ったところ，実際に出現した相槌の76%は実験でも検出され，発話長が長い場合に相槌を打つと判定した被験者が多いことが明らかになった．さらに相槌を打つタイミングについて，対話の分析と知覚実験を行った．この結果，相槌は発話中のポーズ開始から0.3秒以内に打つ必要があることが明らかになった．そこでテンプレートを用いた韻律パターンの認識による相槌タイミングの検出方法を提案し，相槌判定のための予測時間を変えて相槌挿入判定とタイミングの検出実験を行ったところ，予測時間0.1秒のとき84%，予測時間0.4秒のとき72%のタイミング正解率を得た．また予測時間0.1秒のとき得られたシステムの応答を人間が評価したところ，抽出箇所の74%は自然な発声箇所であると判定された．

Giving 'Aizuchi' Using Prosodic Information

YOHEI OKATO,[†] KEIJI KATO,[††] MIKIO YAMAMOTO[†††]
and SHUICHI ITAHASHI[†††]

A user's degree of comfort in a man-machine spoken dialog environment is likely to improve, if spoken dialog systems can provide correct 'Aizuchi' responses to the user's utterances. This hypothesis was evaluated using a dialog corpus that relates to telephone shopping tasks, and contains 'Aizuchi' responses near the end of a speaker's utterance. The evaluation also requires a dialog system capable of detecting 'Aizuchi' timing before the end of the utterance. To this end, therefore, a method is proposed which uses prosodic information to guide correct 'Aizuchi' responses. A preliminary prosodic analysis of our utterances confirmed that an 'Aizuchi' indeed relates to the duration, speaking rate and minimum F0 of an utterance. Next, using dialogs from which 'Aizuchi' responses were previously removed, an experiment was carried out to spontaneously prompt such responses from human subjects. Results show that subjects were able to match about 80% of the 'Aizuchi' responses contained in the original dialogs, and that many subjects tended to do so during long utterances. Then, a dialog analysis was performed to investigate 'Aizuchi' timing, results of which indicate that the system should give an 'Aizuchi' within 0.3 seconds of the end of the speaker's utterance. By comparison, in an 'Aizuchi'-prompting experiment based on prosodic pattern recognition, the system achieved 84% with no 0.1-second prediction of end of utterance and 72% with 0.4-second prediction. Finally, human perceptual evaluation of the timing of system detection, yielded an accuracy of 74% which lends support to the naturalness of 'Aizuchi'response given by the system.

1. はじめに

人間と機械が音声を用いて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，音声認識/合成の研究が各研究機関で行われている．それら音声処理技術の向上と計算機の能力向上により，対象を限定すれば実時間に近い速度で音声認識

† 筑波大学大学院博士課程工学研究科
Graduate School of Engineering, University of Tsukuba
†† アルパイン株式会社
Alpine Electronics, Inc.
††† 筑波大学電子・情報工学系
Institute of Information Sciences and Electronics, University of Tsukuba

を行い，応答を返すことが可能になった．これを基盤として，音声を用いて対話的に問題解決を行う音声対話システムの構築が各研究機関で試みられている[1)~3)]．しかしながら，従来の一般的な音声対話システムはシステムとユーザが交互に発話することを仮定し，時間的に単純な発話管理を行っている[4)]．このため，ユーザは自由なタイミングで発話を行うことができず，システムもタイミング良く応答を返すことが難しい．

実際の人間同士の対話では，話し手は聞き手の状態を把握しつつ対話を行っている．相槌などの音声的応答，表情，うなずき，ジェスチャなどの視覚的情報により，話し手は発話を中断することなく，聞き手の状態を知ることができる．

人間と対話システムとの対話においても，相槌のような応答を扱うことができれば，より応答性が高く使いやすいシステムが構築できると期待される．ユーザは自由なタイミングで発話できるため，より効率的な対話が実現できる[5)]．一方タイミングの良いシステムの応答により，ユーザが発話中にシステムの理解状況を知ることができれば，システムとの対話の快適性はより改善される[6)]．そこで筆者らは，これらの聞き手の応答のうち音声的なフィードバックである「相槌」に焦点を当て，音声対話システムが相槌を行うことによる応答性の向上について検討してきた[7),8)]．

相槌の機能には「話しを続けて」というシグナル，内容理解を示す，聞き手の判断などがある[9)]．相槌が打たれる状況は会話分析や言語学的な研究から以下のような知見が得られている．すなわち相槌は平均すると話し手の発話 14–26 音節おきに打たれる傾向がある[10)]．また話し手の発話のうち，短いポーズが存在する韻律句でそのポーズの出現に合わせて相槌が打たれる傾向があり，語彙的にも特有の表現が見られる[9),11)]．さらに相槌が打たれる発話は韻律句末に特有の韻律パターンがあることが指摘されている[12)~14)]．これらは相槌が聞き手の理解の程度のみならず，構文的・韻律的な情報など談話上の様々な要素と密接に関係することを示している．

これまで相槌を扱うシステムに関する研究として，西ら[15)]は電話取り次ぎの場面において，一定長のポーズから相槌を打つタイミングを決定している．また Watanabe[16)]は一方の話者が一定のリズムで発話を続けるような発話について，発話の ON/OFF パターンを用いた相槌の挿入モデルを提案した．

これらの研究で扱う発話の性質はそれぞれ異なる．タスクや話者の違いにより，打たれる相槌のタイミングも変化する[17)]．発話中のポーズ検出から相槌のタイミングを判定する手法は，一般的に発話には音韻的な無音区間が含まれるため，発話が限定される．さらにポーズ検出の必要があることから，早いタイミングで相槌を打つことができない．対話では発話権が頻繁に移動するため，発話のリズムは 1 人で話す場合ほど一定ではない．そこで本稿ではオペレータ主導であり，対話の形式が限定されるものの，対話全体に相槌が出現するテレフォンショッピングをタスクとし，出現する相槌について検討した．本タスクの相槌は話し手の発話間にポーズの出現とほぼ同時に打たれる傾向がある．このためシステムが適切なタイミングで相槌を打つには，ポーズの検出よりも早く相槌の挿入判定を行うことが必要な場合がある．そこで対話中，相槌が話し手の発話の終了とほぼ同時に打たれる場合について，韻律情報による相槌挿入箇所の判定について検討する．韻律パターンをモデル化し，アクセントや句境界を推定する方法はすでに提案されているが[18),19)]，本稿では相槌が打たれる発話の同定と相槌のタイミングの予測にモデルを用いる．システムを評価する場合，相槌が聞き手個人やその理解の程度に依存する点，適切な相槌タイミングの判定などの問題がある．そこで相槌を打つべき適切な相槌の箇所やそのタイミングについて，対話の分析と知覚実験により検討する．

また扱う対象が自由会話で本研究とやや異なるものの，同時期に韻律情報を用いた相槌の挿入を行う試みがなされている[20)]．本稿では適切な相槌の条件について，人間が第三者として相槌箇所を判定する相槌挿入実験や知覚実験から実際に検討している点が異なる．また本研究のデータについて比較実験を行った．

本文は以下の構成をとる．まず 2 章で本稿で用いた音声資料について説明する．次に 3 章で，発話の韻律的情報と相槌の関係について分析する．また，対話のどこで相槌を打つか人間が判定する実験を行い，相槌が打たれやすい発話について検討する．さらに 4 章では相槌を打つ適切なタイミングについて，対話の分析と相槌のタイミングを変えた対話を聞かせることによる知覚実験により検討する．5 章では韻律パターンを用いた相槌挿入手法について述べ，6 章では本手法を用いて対話中の相槌を打つべき箇所の判定実験を行う．

2. 分析に使用した対話

2.1 音声資料

相槌の分析と実験には，文部省重点領域研究「音声対話」の対話音声コーパス[21)]から，テレフォンショッピングをタスクとした 5 対話[22)]（tsu1103, tsu1107, tsu1204, tsu1208, tsu1210），合計 17 分を用いた．

これらの対話は電話による買い物の状況を想定した電話オペレータと注文者の非対面の対話である．話題の進行は原則として，あらかじめ設定された手順に従いオペレータ主導で行われる．電話オペレータ役の話者は固定された1話者で，注文者は3話者である．電話オペレータはタスク内容について，事前に十分訓練しているが，注文者は特に準備していない．

2.2 分析単位

対話中の発話分析の最小単位は，文節あるいは，句点，読点，言い淀みなどに相当するポーズで区切られた「句音声」を単位[23]とした（総数：1251，平均1.4秒，標準偏差1.5秒）．ただし発話の途中で相槌が打たれたり，割り込まれたりした場合は相槌・割込みの照応する文節の末尾で区切った．

会話分析において，相槌には様々な定義があるが，本稿ではメイナード[9]の「対話中の話し手と聞き手が明らかな状況における，発話権の移動をともなわない聞き手の発話のうち，発話内容自体には意味を持たないもの」を相槌と定義する．すなわち発話権交代に無関係な聞き手の無意味発話のみを対象とし，“Yes”の意味を持つ「はい」という応答，話者交替直後の「はい」や，復唱は相槌としない．分析した対話中，打たれた相槌はすべて「はい」という発話で，電話オペレータ側57個，注文者側88個の発話，合計145個存在した．それぞれの相槌の照応先は人手で判定した．相槌の照応先は，相槌の打たれたタイミングから，0.4秒以上前に発話を開始した時間的に最も近い話し手の句音声とすべて一致した．

タスクの性質から聞き手の理解が曖昧な発話などは含まず，照応先の判定で迷うことはなかった．聞き手の相槌箇所で区切った句音声のうち，14カ所はポーズを挟まず音声的に連続していた．それら連続した句音声の後続箇所に対して複数の相槌が照応することはなかった．また同一の句音声に複数の相槌が照応することはなかった．

本稿で用いた対話の書き起こし例を**図1**に示す．書き起こし中，中括弧（{はい}）で囲まれた箇所は聞き手の発話を示し，下線は発話が相槌であることを示している．また斜線（／）で区切られた単位は句音声を示す．本タスクでは，1）電話オペレータの問合せ（顧客情報，注文情報），2）電話オペレータによる注文内容の確認，などで相槌が打たれていた．

3. 相槌が打たれやすい発話の分析

2章で示した音声資料について，相槌が打たれやすい発話の性質を検討した．まず，相槌が打たれた発話と，そうでない発話の韻律的な特徴を比較した．次に人間による相槌挿入実験を行い，オペレータの発話に対し注文者が相槌を打ちうる箇所を抽出し，その傾向を分析した．

注文者	商品名は{はい}／ウッドラック．
オペレータ	ウッドラック
注文者	はい．
オペレータ	品番，お願いします．
注文者	エックスシーの{はい}／二
オペレータ	エックスシーの，二
	注文コード，お願いします
注文者	八三三四の{はい}／九三四．
オペレータ	八三三四の九三四
注文者	はい．
オペレータ	サイズコード，お願いします．
注文者	六十です，{六十}／エーの方で．
オペレータ	はい．
	こちらの方，数量は
注文者	二つ，お願いします

＿＿は相槌，「／」は句音声境界を示す．

図1 対話の書き起こし
Fig. 1 Dialog transcription.

表1 相槌の呼応先とそれ以外の発話の韻律的性質
Table 1 Prosodic analysis of utterances with and without ‘Aizuchi’.

	電話オペレータ		注文者	
	相槌あり	相槌なし	相槌あり	相槌なし
発話長（秒）				
平均	3.19	1.59	1.50	0.86
標準偏差	2.93	1.52	0.76	0.90
発話速度（※）				
平均	8.0	8.0	6.4	7.0
標準偏差	1.6	1.7	1.4	2.0
F_0（正規化）				
平均	−0.56	0.02	−0.10	−0.07
標準偏差	0.53	0.63	0.49	0.74
最大	0.44	0.76	0.95	0.59
最小	−1.99	−1.08	−1.52	−0.66
Power（正規化）				
平均	−0.44	−0.49	−0.10	−0.82
標準偏差	0.49	0.70	0.39	0.77
最大	0.58	0.67	0.85	0.53
データ数	88	660	57	441

※1秒あたりのモーラ数

3.1 相槌が打たれた発話の韻律的傾向

まず相槌が打たれた発話の韻律的な傾向を調べた（**表1**）．韻律的特徴として，句音声単位の基本周波数（F_0），短時間パワー，発話速度の平均・標準偏差・最大値・最小値を電話オペレータ，注文者それぞれについて求めた．F_0とパワーは対数をとり，その後，全区間の音声を用いて各話者ごとに平均値0，分散1に正規化した．韻律的特徴のうち，発話時間は句音声単位ではやや短すぎると考え，句音声をつなぎ合わせ，書

き言葉における文に相当する単位ごとに分析した．ただし，発話の途中で相槌が挿入されたり，割り込まれた場合は，相槌・割込みの照応する句音声で区切った．

表1では，F_0 の最小値，発話長は電話オペレータと注文者で共通した傾向を示し，相槌が入る場合はそれ以外に比べ，F_0 の最小値はより小さく，また発話長はより長くなる傾向を示した．

実際のデータを見ると，句音声はほぼ1つの F_0 の起伏を表している．F_0 が最小値をとるのは句音声末付近であり，平均すると発話終了の0.25秒前に F_0 が平均値から20%低下する．この F_0 の下降の程度や，句音声の長さが相槌が打たれる割合に影響を与え，同様に F_0 が低いつなぎ語などと区別できると考えられる．

3.2 相槌挿入実験

テレフォンショッピングタスクのように対話の進行がほぼ固定されている場合でも，聞き手の個人差や理解の程度の違いにより相槌が打たれる場合とそうでない場合がありうる．それらによる相槌挿入箇所の違いを定量的に分析するために，同一の音声資料に対して人間が相槌箇所を判定する実験を行った．

実験に用いた音声資料はテレフォンショッピング対話の電話オペレータの相槌部分をポーズに置き換えたものである．被験者に加工した対話を聴取させ，相槌を打つべきであると判断した箇所でボタンを押すように指示した．ボタンが押されると,「はい」という音声で被験者に応答を返し，システムはその時刻を記録した．実験中，被験者は対話を一時中断することができるが，対話を繰り返し聞き判定することはできない．

用いた音声資料は tsu1103（注文者男性），tsu1208（注文者女性）の2対話である．実験では注文者の性別が被験者と同じ対話を利用し，男性22人，女性12人が相槌挿入実験を行った．実験後，人手により相槌と発話の対応関係を調べた．この際，句音声末尾から0.5秒を目安に相槌の照応先を定めた．本実験で打たれた相槌のうち，該当する句音声がない箇所，注文者が発話権をとる場面や,「はい」「いいえ」などの応答を求めた質問の応答部分など，本稿で定義する相槌と異なると判断したものは削除した．また得られた相槌箇所のうち，同一箇所に相槌を打つと判定した被験者が2人以下の箇所はデータが十分でないため削除した．模擬対話と本相槌挿入実験で得た相槌箇所を，A）模擬対話でのみ出現したもの，B）本相槌挿入実験でのみ出現したもの，C）どちらにも出現したもの，と分けると**表2**のようになった．元の対話で打たれた相槌のうち，tsu1103（男性）で80%（$= 16/(16+4)$），tsu1208（女性）で73%の相槌は，コーパス中の相槌が相槌挿入実験でも検出された．

表2 模擬対話と相槌挿入実験の相槌位置の比較
Table 2 Comparative results of 'Aizuchi' location: simulated dialogs versus human experiments.

	tsu1103（男性）	tsu1208（女性）
模擬対話のみ	4	8
相槌挿入実験のみ	11	2
両方に出現	16	22

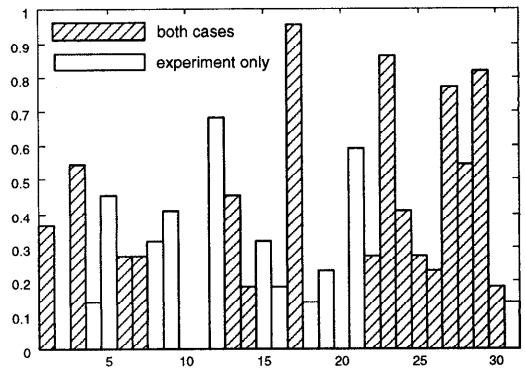


図2 同一発話に対して相槌を挿入した被験者の割合
Fig. 2 Histogram of subject agreements for every 'Aizuchi' (simulated dialogs and human experiments).

特に音声資料 tsu1103 について，実際の対話中の相槌と本実験の結果を比較した．**図2**に同一の句音声に対して相槌を打つと判定した被験者の割合を示す．横軸は元の対話で相槌が打たれた句音声と，相槌挿入実験で被験者が相槌を打つと判断した句音声を時間順に並べたものである．縦軸は相槌を打つと判断した被験者の割合を示している．白い棒で示されている句音声は元の対話では出現しなかった相槌箇所を示している．

特にオペレータが行う長い確認の発話（図2中の相槌17，21，23）に対して，相槌が打たれた割合が高い．被験者の相槌は平均すると対話中の相槌と37%一致し，一方，元の対話の相槌は被験者の打つ相槌と平均67%一致している．これらは6章における評価実験で被験者が相槌挿入システムの代わりをした場合の検出率・精度に相当する．

3.3 考　　察

相槌は，発話時間が長く，F_0 の最小値が小さい発話で多く見られた．

相槌挿入実験では，コーパス中対話でオペレータ発話に対して注文者が相槌を打った個所は，男性の場合80%，女性の場合73%が実験でも検出された．本実験では被験者は対話に部分的に参加している立場であり，元の対話における実際の聞き手の理解の程度を完

全に把握しているわけではない．また相槌を除いた後のポーズが相槌を促進している可能性もあり，注意が必要である．しかし，長い確認調の文の終わりで，発話が継続されている部分では実際の対話と高い一致を見た．これらの箇所は，発話の長さ，語彙などから，ポーズが後続することが示されていること，確認的な文脈の発話であり，協調的な対話を進めるうえで聞き手は相槌が求められていることが考えられる．

4. 相槌が打たれるタイミング

相槌は話し手の発話中のポーズなどの韻律句末に合わせて打たれる傾向がある．しかし，話し手はその後も発話を続けるので，相槌を返すタイミングが重要になる[24]．そこで対話システムによる相槌として，許容されるタイミングの範囲について検討した．まず，音声資料から対話中の相槌の応答時間を調べた．次にポーズ長を加工した音声資料を用い，知覚実験により相槌がどの程度遅れると不自然に感じられるか検討した．

4.1 タイミングの分析

相槌は話し手の発話中のポーズなどの韻律句末に合わせて打たれる傾向がある．先行するオペレータ・注文者の発話の句音声末から，聞き手の相槌が打たれるまでのポーズ長の分布は**図3**のようになった．ただしポーズ長が負の場合は，話し手の句音声発話の終了以前に聞き手が相槌を開始していることを表す．図3によると，相槌に先行するポーズ長は全体の90%が−0.1秒から+0.3秒の範囲に存在した．

4.2 知覚実験

次に相槌のタイミングとして，どの程度の遅れが許容されるか，知覚実験により検討した．知覚実験では対話の相槌・話者交替が起きた箇所のポーズ長を加工した音声資料を作成し，被験者に聴取させてタイミングの自然さを調べた．

知覚実験に用いる音声資料は，聞き手の相槌および，相槌以外の発話による話者交代の箇所のポーズ長を一定（**0.3秒，0.6秒，0.9秒，1.2秒**の条件）に固定した対話である．**図4**はポーズ長を0.3秒としたときの対話の時間構成を示している．この音声資料を被験者に聴取させ，対話中の発話タイミングとして，不自然に感じた箇所を検出させた．タイミングが不自然に感じられた場合，「早い」，「やや早い」，「やや遅い」，「遅い」の4種類からボタンにより選択させ，システムはその内容と時刻を記録した．実験は1人の被験者について，以上の4条件の中から適当に選んだ3条件について，それぞれ異なる対話を加工した音声資料を用いて行った．被験者のタイミングの感覚を戻すため，2回目はつねに加工を行っていない無修正の対話を聞かせた．被験者は大学生・大学院生で，各条件のポーズ長について12〜13人ずつ行った．実験後，被験者が不自然と感じた箇所を相槌部分とそれ以外に分けて集計した．その際に「早い」，「やや早い」は早いタイミング，「遅い」，「やや遅い」は遅いタイミングとしてまとめた．結果を**表3**に示す．それぞれ実験回数で正規化し，1対話あたりの不自然に感じられた箇所数の平均を表した．表3から以下のような傾向が見られる．

- ポーズ長**0.3秒**に固定
  相槌部分の違和感は少なく，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箇所は少ない．
- ポーズ長**0.6秒**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．それ以外の部分で違和感は少ない．全体的に違和感を感じた箇所は少ない．
- ポーズ長**0.9，1.2秒**に固定
  相槌の箇所，それ以外の箇所の両方で遅いと感じ

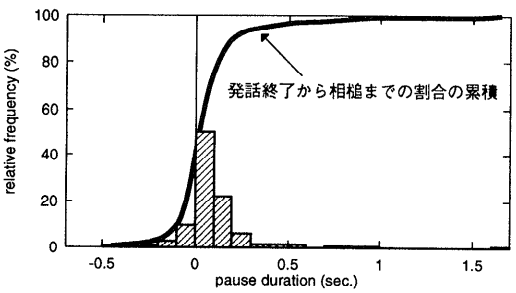


図3 句音声末から相槌までのポーズ長
Fig. 3 Distribution of duration between the end of utterance and the start of 'Aizuchi'.

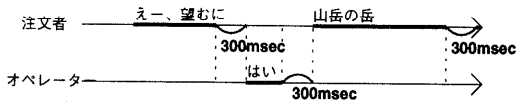


図4 話者交替時のポーズ長の加工（ポーズ長を0.3秒に固定した場合）
Fig. 4 Processing of dialog's pause duration (0.3 s each).

表3 知覚実験におけるポーズ長固定の対話に対する違和感を感じた平均箇所数
Table 3 Normalized rating of number of unnatural locations in dialogs processed with fixed pauses.

ポーズ	相槌		それ以外の発話		実験回数
	速い	遅い	速い	遅い	
0.3	0.3	0.5	1.5	0.4	13
0.6	0.0	3.4	1.8	1.7	12
0.9	0.4	7.8	0.6	4.6	13
1.2	0.0	5.8	0.8	5.8	12

ている．全体的に違和感を感じた部分が多い．

相槌が打たれるまでのポーズ長を0.6秒にした場合，0.3秒の場合に比べ相槌が遅く感じられた句音声は6.8倍増加した．この結果は図3において，相槌が打たれるタイミングの割合が累積93%に達し，相槌の打たれる割合が急激に減少する部分と対応している．

4.3 考　　察

今回用いた音声資料では，93%の発話について相槌が句音声末から0.3秒以内に打たれている．知覚実験の結果も，相槌が遅く感じられた句音声数は，ポーズ長を0.6秒にした場合，0.3秒の場合に比べ6.8倍増加した．このため，対話システムによる相槌挿入の実装では，発話終了後0.6秒では遅く，0.3秒程度で行う必要がある．そこで5章で行うシステムによる相槌タイミングの検出では，図3の相槌が打たれるまでのタイミング分布と合わせ，句音声末 −0.1秒から，+0.3秒を適切な相槌のタイミングであると判定する．

本稿では韻律的な情報についてのみ検討したが，ポーズ長と相槌のタイミングの関係については，句音声内の語彙や文脈，話者の親密さによる違いも検討する必要がある．また知覚実験は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についても考慮する必要がある[25]．

5. 韻律情報による相槌挿入

3章の分析結果から韻律情報は相槌挿入の判定に有用であると考えられる．本章では韻律情報から対話システムが相槌を打つタイミングを検出する手法について検討する．4章の検討から，相槌を打つ場合，そのタイミングが重要である．相槌は句音声の終了とほぼ同時に打たれるが，このタイミングで相槌を打つには句音声の終了を検出してからでは遅く，途中までの音声から相槌を打つタイミングを予測する必要がある．

そこで本稿では，句音声の開始から相槌が打たれるまでの発話をあらかじめモデル化しておき，モデル上の句音声の終端から相槌のタイミングを得る．システムは入力音声の韻律パターンとモデルを照合し，モデル上の現在の位置と照合スコアから相槌の判定を行う．モデルは句音声の韻律パターンを複数の状態を用いて表しており，モデルの先頭の状態が句音声の開始，モデル終端が句音声の終わりを表す．各状態は一定時間の韻律パターンを示す．**図5**は正規化した F_0 を4状態で表した例である．システムは入力パターンがモデルの終端まで到達したと判断したとき，相槌を打つと判定する．モデルの各状態は発話速度の違いを考慮し伸縮を許しているが，歪みが大きくなりすぎないように，各状態の伸縮はPoisson分布を仮定した滞在フレーム数の制約をペナルティとして加えている．

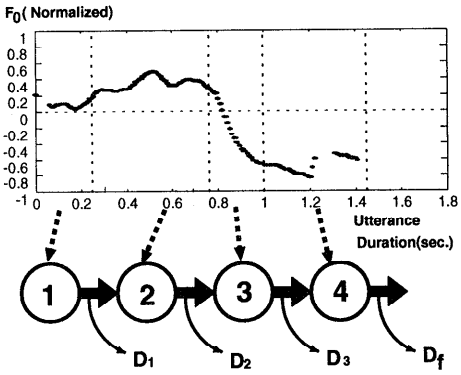


図5　状態遷移モデル
Fig. 5　State transition model.

具体的なモデルと入力パターンの照合方法は以下のとおりである．時刻 t の入力音声フレーム x_t とモデルの k 番目の状態 m_k の距離はそれぞれのベクトル間のユークリッド距離（式(1)）とする．また状態 k に対応する入力パターンのフレーム数 s_k とモデル作成時の学習パターンの対応フレーム数の平均 d_k から，状態 k の滞在ペナルティ $P_{dur}(d_k, s_k)$ を式(2)のように定義する．ここで $Po(m,x)$ は平均・分散が m のポアソン分布を示す．

時刻 T までの入力パターンと状態 K までのモデルの距離 $D_{K,T}$ は状態と入力パターンの距離の合計と，各状態の滞在フレーム数 $d_k\,(k=1\ldots K)$ に対するペナルティ $P_{dur}(d_k, s_k)$ に重み w_{dur} をかけて加えたもので定義する．$1 \le k_t \le K$ かつ $k_t \le k_{t+1}$ の制約条件のもとで，$k_t\,(t=1\ldots T)$ に関して入力パターンの各フレーム x_t と状態 m_k の対応は式(3)を最小化したものである．式(3)の第1項はDPマッチングにより効率的に求めることができる．ただし有声でない入力音声フレームについては滞在フレーム数のみ考慮する．滞在フレーム数の重み w_{dur} は実験的に定める．

$$d(m_k, x_t) = ||m_k - x_t||^2 \tag{1}$$

$$P_{dur}(d_k, s_k) = \log Po(d_k, d_k) - \log Po(d_k, s_k) \tag{2}$$

$$D_{K,T} = \min_{\{k_t\}} \frac{1}{T} \sum_{t=1}^{T} d(m_{k_t}, x_t) + w_{dur} \cdot \frac{1}{K} \sum_{s=1}^{K} P_{dur}(d_k, s_k) \tag{3}$$

相槌のタイミングは本モデルの終端に相当する．相

槌を打つ判定は各時刻フレームごとに行い，以下の条件を満たすとき相槌を打つタイミングであると判定する．

1) 距離 $D_{K,T}$ の値がしきい値より小さいこと．
2) 距離 $D_{K,T}$ が $k < K$ の途中の状態までの距離 $D_{k,T}$ よりも小さいこと．
3) モデルの最終状態の平均滞在フレーム数よりも，入力パターンの対応するフレームの方が長いこと．

条件 1) は入力パターンと照合されたモデルの類似性を表している．条件 2) は入力パターンがモデルの途中までより最終状態までとより良く照合されることを表し，最終状態まで到達していることを示す．これはDPマッチングの計算過程で得られる値をそのまま利用することができる．条件 3) はモデルの終端に到達したことを判定するものである．

モデルのパラメータはViterbi Alignmentを反復的に行い，学習パターンとモデル間の距離が最小になるように決定する．学習の手順は以下のとおりである．まず学習パターンの有声フレームをモデルの状態数で時間方向に均等分割し，状態とパターンの対応を求める．各状態に対応する学習パターンの平均・分散を求め，パラメータの初期値とする．次にViterbiアルゴリズムにより，学習パターンと状態の対応を再計算し，得られた対応関係についてモデルのパラメータを求め直す．この手順を学習パターンとモデル間の距離が減少する間，反復する．状態 k の平均滞在フレーム数 d_k はこのときの分割の結果に従って決定する．

6. 評価実験

5章で述べた手法について相槌挿入の決定・タイミング検出の評価実験を行った．実験では注文者の発話に対するオペレータの相槌を対象としてコーパスの対話中で打たれた相槌とシステムが抽出した相槌箇所の一致の程度を調べ，4章の相槌タイミングの検討を考慮したシステムのタイミング抽出の精度について検討した．相槌は不適切に打たれた場合，相槌がない場合よりもシステムの対話性を低下させる恐れがある[25]．そこでシステムと対話の相槌の一致が高いものについて，抽出された相槌箇所の性質を調べた．最後にWard[20]の方法と比較し，考察した．

6.1 実験方法

5章で述べた手法を適用するにあたり，韻律を表すパラメータとして，F_0 と短時間パワーを用いた．F_0 の抽出には信号処理パッケージESPS[26]に含まれるプログラム `get_f0` を使用し，分析のフレーム周期は10 msとした．得られたパラメータは対数をとった後，あらかじめ対話全体で話者ごとに平均0，分散1に正規化した．また，モデルは各4状態とし，句音声ごとに1つのモデルを作成した．モデルの学習には実際に相槌が打たれた句音声の韻律パターンを用いた．

相槌の検出はあらかじめ切り出された句音声の開始からフレーム同期で行い，システムが相槌のタイミングを検出するか，句音声末に到達するまで続けた．検出した時刻があらかじめ設定した時間の範囲内にあれば正解，そうでなければ誤りとした．

相槌タイミングの正解は4章の検討から，句音声末の0.1秒前から，0.3秒後までの範囲とした．また実際にシステムが相槌を打つにはシステムの処理時間を考慮する必要があり，処理時間の分だけ早く予測的に判定を行う必要がある．提案する手法ではモデルの終端が相槌のタイミングを表している．そのため句音声末から想定する予測時間の分だけ前の音声区間までをモデルの学習に用いた．

3.2節の検討から，対話中に打たれる相槌の位置は聞き手の違いによらず，ある程度一致する．そこで分析に使用したコーパスの注文者の発話に対するオペレータの相槌について，コーパス中に含まれる相槌とシステム判定した結果を比較した．評価の尺度はコーパスに含まれる相槌箇所に占めるシステムの相槌箇所の割合（検出率），およびシステムの相槌箇所に占めるコーパスに含まれる相槌箇所の割合（精度）を用いた．またシステムが検出した相槌のタイミングについて，相槌の有無にかかわらず，句音声末が適切に検出できている割合（タイミング正解率）を調べた．

予備実験として，モデル学習用の対話に含まれる句音声に対して相槌の検出を行い，本実験に用いるモデルの選択および，パラメータ（予測時間，判定のしきい値，滞在ペナルティの重み）の範囲を設定した．

予備実験では，予測時間が長くなると，精度，タイミング正解率が低下する傾向が見られた．得られたモデルを検討すると，全般的にモデルの後半で F_0 が大きく低下する．また滞在ペナルティの重みを増すと検出率が下がり，精度が増す傾向があった．精度が高いモデルではモデルの後半で F_0 が低下するがパワーはそれほど低下しないか，むしろ上昇している場合が目立った．これらの性質が相槌が打たれやすく，それ以外の場合と識別的な特徴であると考えられる．

6.2 実　　験

モデル作成と異なる話者の発話に対して相槌検出実験を行い，性能を調べた．評価には注文者側の発話を用い，2人の話者（4対話）について，一方をモデルの作成用，他方をテスト用とし，その逆と合わせて2

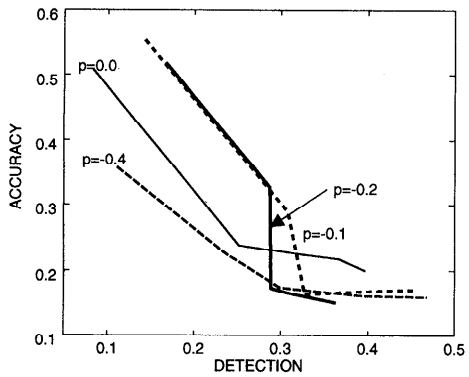


図6 相槌検出実験（検出率と精度の関係）
Fig. 6 'Aizuchi' detection experiment (detection rate vs. accuracy).

表4 相槌検出実験（タイミング正解率）
Table 4 'Aizuchi' detection experiment (correct timing rate).

予測時間（秒）	0.0	−0.1	−0.2	−0.4
タイミング正解率（%）	88	84	78	72

回の実験を行った．データは全体で15分で，377個の句音声，45個の相槌が含まれる．相槌の判定には予備実験で高い性能を示したテンプレート上位3個を使用した．実験では予測時間を0.0秒から0.4秒まで変化させ，それぞれ相槌判定のしきい値を変えて検出率・精度・タイミング正解率を求めた．

得られた検出率と精度の関係を**図6**，タイミング正解率を**表4**に示す．予測時間が0.1秒のとき，精度が最大の58%となり，このとき検出率は15%となった．予測時間が0.2秒のとき，検出率17%に対して精度52%，タイミング正解率78%となった．また予測時間が0.4秒のときに検出率が最大となり，このとき検出率46%，精度16%となった．また予測が短いほど，タイミング正解率は向上した．

6.3 システムの打った相槌の分析

相槌は不適切に打たれた場合，相槌が打たれない場合よりもシステムの対話性を低下させる恐れがある[25]．そこでコーパス中の対話とシステムの判定結果で相槌箇所の一致が高いものについて，人間が聴取により評価した．被験者はオリジナルの対話とシステムが判定したタイミングに従った相槌を左右分割した音声によりヘッドホンで聴取する．システムの相槌時刻は，あらかじめ紙で伝えてあり，システムの相槌それぞれについて（A）適切な相槌，(B）相槌のタイミングが早く不適切，(C）相槌のタイミングが遅く不適切，(D）相槌の定義にあてはまらないが自然な発話，(E）それ以外の不適切な発話，の中から選択させた．使用したシステムの相槌は図6に示したもののうち，最も高い精度が得られた場合（高精度モデル，予測時間0.1秒，検出率15%，精度58%）と，最も高い検出率が得られた場合（高検出率モデル，予測時間0.4秒，検出率46%，精度16%）の場合について，同一の2対話を用いた．システムの相槌の音声として，オペレータの「はい」という同一音声を利用した．評価は5人の大学院生が行った．実験中，被験者は任意のタイミングで対話を止め，再度聴取することができる．回答の割合を対話ごとに平均した結果を**表5**に示す．

表5 システムの生成した相槌に対する人間の評価
Table 5 Human evaluation of system's 'Aizuchi'.

割合（%）	A	B	C	D	E
高精度モデル	46	10	6	28	8
高検出率モデル	36	17	9	24	14

システムが相槌を生成した箇所で被験者の過半数が（A）の適切とした箇所のうち，元の対話で相槌が打たれていない箇所は1対話平均3カ所存在した．また（D）はその後発話交替があった箇所である．この箇所に「はい」という発話が入ることは不自然ではないと考えられる．(A）と（D）を合計すると，高精度モデルでは74%となる．

一方（B),(C）のタイミングや（E）の相槌として不適切とされた箇所は実際にシステムを構築する際に問題となる．これらについて定性的に性質を調べたところ，タイミングでは実質的な応答を表す内容語に後続する付属語列と重なって相槌が打たれる傾向が見られた．また不適切とされた発話には句音声の先頭のつなぎ語など意味内容に乏しい発話が見られた．また今回相槌を打つ頻度は対象としていないが，短い句音声の箇所で相槌が時間的に連続してしまい，不適切に感じられる箇所が存在した．

またシステムが検出した句音声末のタイミングについて調べた．**図7**は高精度モデルと予測時間が異なるモデル（図6の各予測時間の最左端）を示す．横軸はシステムが検出した句音声末タイミングを示し（負の値は句音声末，以前のタイミングを検出したことを示す），縦軸はその相対頻度を示す．図中,「Aizuchi」はテストセットの相槌の時間分布を示している．**表6**にそれぞれの平均，標準偏差を示す．予測時間が短いほど，標準偏差が小さくなっていることが分かる．

6.4 比較実験

比較実験として，同一データに対してWard[20]の方法を適用した．比較手法は会話から抽出したルールに基づく手法である．Wardの方法では相槌の適切な範囲を実際の相槌前後0.5秒としており，それに従った．

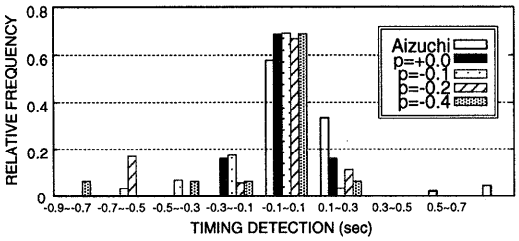


図7 システムの検出した句音声末タイミングの分布（全体）
Fig. 7 'Aizuchi' timing distribution of the system detection.

表6 システムの検出した句音声末のタイミングの分布
Table 6 Timing distribution of the end of utterance.

予測	平均（秒）	標準偏差（秒）
+0.0	−0.05	0.19
−0.1	−0.15	0.25
−0.2	−0.21	0.31
−0.4	−0.24	0.36
相槌	0.10	0.25

今回の実験データに適用した結果，検出率57%，精度10%となった．精度の違いは比較手法が対象としたものと会話のスタイルが異なり，パワーの扱いや状態間の変化など，モデルの違いが影響したと思われる．

6.5 考　　察

システムを用いた相槌箇所判定により，韻律情報のみから注文者の相槌箇所を句音声が終了する0.2秒前に推定した場合，テストデータに含まれる相槌に対して17%の検出率，52%の精度を得た．この結果は韻律情報の有効性を示している．また相槌を打つうえで重要なタイミングの推定は80%程度であり，判定に韻律情報を利用することが有用である．予測時間が0秒の結果（図6の細線）は，0.1秒の結果（図6の太線）よりも性能が低下した．これは相槌を打つのに必要な情報が多くの場合，発話終了よりも以前に示されていることを示唆している．

このシステム応答を人間が評価したところ，相槌として適切な箇所と発話交替で問題ない箇所を合計すると74%に達した．不適切な相槌を打たないようにするには韻律パターンにリスク・信頼度を表す指標を付けたり，相槌を打つのに不適切な箇所のモデル化などが考えらえれる．

実際に実時間システムを構築するには実時間で入力を正規化する必要がある．このためには対話の始め部分でF_0，パワーの平均値を推定しておく方法が考えられる．また今回対象としたコーパスには出現しなかったが，相槌が1つの句音声中に2つ以上出現する場合を扱うには複数の句音声モデルの接続が考えられる．

7. ま と め

本稿では韻律情報による相槌の挿入について検討した．テレフォンショッピングをタスクとした対話を用い，まず対話の分析を通して韻律情報が相槌挿入と関連があることを確認した．次に対話システムによって相槌のような応答を行う場合，ポーズの検出によらないタイミング検出法の必要性を述べ，韻律テンプレートを用いて相槌のタイミングを検出する手法について述べた．また相槌として許容されるタイミングの範囲について，対話の分析と知覚実験の結果から検討した．このタイミングの範囲を正解として，提案手法の評価実験を行ったところ，予測時間が0.1秒のとき，精度が最大の58%となり，このとき検出率は15%となった．このシステムが生成した発話の74%は適切な相槌，あるいは話者交替時の発声として許容される発話だった．今後，不適切な相槌タイミングを抽出しないように，詳細な検討を行う必要がある．また対話システムを用いた評価を行うことが今後の課題である．

謝辞 ご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いたします．なお，本研究の一部は筑波大学ベンチャービジネスラボラトリーの援助を受けた．

参 考 文 献

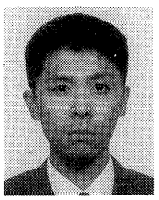
1) Hemphill, C.T., Godfrey, J.J. and Doddington, G.R.: The ATIS spoken language systems pilot corpus, *Proc. 3rd DARPA Speech and Natural Language Workshop*, pp.91–95 (Jan. 1990).
2) Itou, K., Hayamizu, S., Tanaka, K. and Tanaka, H.: System Design, Data Collection and Evaluation of a Speech Dialogue System, *IEICE*, Vol.E-76D, No.1, pp.121–127 (1993).
3) Takebayashi, Y., et al.: A Real-Time Speech Dialogue System Using Spontaneous Speech Understanding, *IEICE*, Vol.E76-D, No.1, pp.112–120 (1993).
4) Lee, C.H.: Stochastic Modeling in Speech Dialogue System Design, *Proc. ISSD-93*, pp.161–168 (1993).
5) 菊池英明，工藤育男，小林哲則，白井克彦：音声インタフェースにおける発話権管理による割り込みへの対処，信学論，Vol.J77-D-II, No.8, pp.1502–1511 (1994).
6) 小高俊之，天野明男：音声対話システムのシステム応答作成と対話性について，平成6年春季日本音響学会講演論文集，pp.33–34 (1995).
7) 岡登洋平，加藤佳司，山本幹雄，板橋秀一：韻律

パターンの認識を用いた相槌挿入とその評価，情報処理学会音声言語研究会報告，SLP-10-7 (1996).
8) Okato, Y., Kato, K., Yamamoto, M. and Itahashi, S.: Insertion of Interjectory Response Based on Prosodic Information, *Proc. Interactive Voice Technology for Telecommunications Applications (IVTTA-96)* (1996).
9) メイナード，泉子.K.: 会話分析，日英語対照研究シリーズ2，くろしお出版 (1993).
10) 水谷 修（編）：話しことばの表現（あいづちと応答），講座日本語の表現（3），筑摩書房 (1983).
11) 小磯花絵，堀内靖雄，土屋 俊．市川 熹：先行発話断片の終端部分に存在する次発話者に関する言語的・韻律的要素について，信学技報，NLC95-72, pp.25–30 (1996).
12) 杉藤美代子：ポーズとイントネーション，談話行動の諸相，国立国語研究所報告，92，三省堂 (1987).
13) 小磯花絵，堀内靖雄，土屋 俊，市川 熹：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，人工知能学会研究会資料，SIG-J-9501-2 (1995).
14) 堀内靖雄，小磯花絵，土屋 俊，市川 熹：自発的音声対話における話者交代の制御に関わる発話末の統語的・韻律的特徴，情報処理学会音声言語研究会報告，SLP-10-9, pp.45–50 (1996).
15) 西 宏之，小島順治：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).
16) Watanabe, T.: A Voice Reaction System with a Visualized Response Equivalent to Nodding, *Advances in Human Factors/Ergonomics*, 12A (1989).
17) 小坂直敏：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 (1987).
18) 高橋 敏，松永昭一：統計的韻律モデルによる連続音声の句境界検出，信学技報，SP90-71 (1990).
19) 吉村 隆，速水 悟，田中和世：モーラ単位HMMの接続による単語アクセントパターン識別，信学技報，SP92-104 (1992).
20) Ward, N.: In Japanese a low pitch means "backchannel feedback please", 情報処理学会音声言語研究会報告，SLP-11-2 (1996).
21) 平成6年度文部省科学研究費補助金重点領域研究：音声・言語・概念の統合的処理による対話の理解と生成に関する研究，対話音声コーパス，Vol.1–2 (1995, 1996).
22) 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，平成7年春季日本音響学会講演論文集，2-Q-21, pp.337–338 (1995).
23) 高木一幸，保浦直子，板橋秀一：対話音声中の発話単位の時間関係，平成5年春季日本音響学会講演論文集，pp.239–240 (1993).
24) 渡辺直也，佐藤伸二，八木建司，井宮 淳，市川 熹：音声対話理解システム構想—CHIBA，*10th Symposium on HUMAN INTERFACE*, (Oct. 1994).
25) 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，情報処理学会音声言語研究会報告，95-SLP-5 (Feb. 1995).
26) Mannual of ESPS Program Version 5.0, Entropic Research Laboratory (1993).

(平成10年 6 月1日受付)
(平成10年12月7日採録)

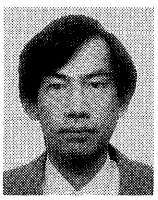
岡登 洋平

1994年筑波大学第三学群情報学類卒業．現在同大学大学院博士課程工学研究科に在学．音声認識・音声対話システムの研究に従事．日本音響学会会員．



加藤 佳司

1996年筑波大学第三学群情報学類卒業．1998年同大学大学院修士課程理工学研究科修了．現在アルパイン（株）勤務．在学時は音声認識・音声対話システムの研究に従事．日本音響学会会員．



山本 幹雄（正会員）

1986年豊橋技術科学大学大学院情報工学専攻修了．同年（株）沖テクノシステムズラボラトリ入社．1988年豊橋技術科学大学情報工学系教務職員．1992年同助手．1995年筑波大学電子・情報工学系講師．1998年同助教授．博士（工学）．音声・言語処理の研究に従事．電子情報通信学会．人工知能学会，言語処理学会，ACL等会員．

板橋 秀一（正会員）

1964年東北大学工学部通信学科卒業．1970年同大学大学院工学研究科博士課程単位取得退学．同年東北大学電気通信研究所助手．1972年通産省工業技術院電子技術総合研究所入所．1977～1978年ストックホルム王立工科大学客員研究員．1982年筑波大学電子・情報工学系助教授．1987年より教授．工学博士．音声・言語・画像の処理・理解に関する研究に従事．